UNIVERSIDADE FEDERAL DE GOIÁS
ESCOLA DE AGRONOMIA
PROGRAMA DE PÓS-GRADUAÇÃO EM
GENÉTICA E MELHORAMENTO DE PLANTAS

RECURSOS GENÔMICOS PARA A CROADINHA (*Mouriri elliptica* Mart. - MELASTOMATACEAE)

JULIANA BORGES PEREIRA BRITO

Setembro – 2025

UNIVERSIDADE FEDERAL DE GOIÁS
ESCOLA DE AGRONOMIA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES

E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do material bibliográfico

[ ] Dissertação [X] Tese [ ] Outro*:_____

*No caso de mestrado/doutorado profissional, indique o formato do Trabalho de Conclusão de Curso, permitido no documento de área, correspondente ao programa de pós-graduação, orientado pela legislação vigente da CAPES.

Exemplos: Estudo de caso ou Revisão sistemática ou outros formatos.

2. Nome completo do autor

Juliana Borges Pereira Brito

3. Título do trabalho

RECURSOS GENÔMICOS PARA A CROADINHA (Mouriri elliptica Mart. - MELASTOMATACEAE)

4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)

Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
a) consulta ao(à) autor(a) e ao(à) orientador(a);
b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
O documento não será disponibilizado durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Juliana Borges Pereira Brito**, **Discente**, em 10/11/2025, às 14:52, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Thannya Nascimento Soares**, **Professora do Magistério Superior**, em 11/11/2025, às 09:34, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5774486** e o código CRC **6F6E8009**.

JULIANA BORGES PEREIRA BRITO

RECURSOS GENÔMICOS PARA A CROADINHA (*Mouriri elliptica* Mart. - MELASTOMATACEAE)

Tese apresentada ao Programa de Pós-Graduação em Genética e Melhoramento de Plantas, da Escola de Agronomia, da Universidade Federal de Goiás (UFG), como requisito para obtenção do título de Doutora em Genética e Melhoramento de Plantas.
Área de concentração: Genética e Melhoramento de Plantas
Linha de pesquisa: Conservação e Melhoramento de Espécies do Cerrado

Orientadora:

Prof.ª Dr.ª Thannya Nascimento Soares

Co-orientadora:

Prof.ª Dr.ª Adriana Maria Antunes Taquary

GOIÂNIA, GO — BRASIL

2025

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Brito, Juliana Borges Pereira
RECURSOS GENÔMICOS PARA A CROADINHA (Mouriri elliptica Mart. - MELASTOMATACEAE) [manuscrito] / Juliana Borges Pereira Brito. - 2025.
102 f.

Orientador: Profa. Dra. Thannya Nascimento Soares; co orientadora Dra. Adriana Maria Antunes Taquary.
Tese (Doutorado) - Universidade Federal de Goiás, Escola de Agronomia (EA), Programa de Pós-graduação em Genética e Melhoramento de Plantas, Goiânia, 2025.
Apêndice.

1. barcoding. 2. bioinformática. 3. Cerrado. 4. Croadinha. 5. diversidade genética. I. Soares, Thannya Nascimento , orient. II. Título.

CDU 575

UNIVERSIDADE FEDERAL DE GOIÁS

ESCOLA DE AGRONOMIA

ATA DE DEFESA DE TESE

Ata Nº PPGGMP/067/2025 da sessão de Defesa de Tese de **Juliana Borges Pereira Brito** que confere o título de Doutora em Genética e Melhoramento de Plantas, na área de concentração em Genética e Melhoramento de Plantas.

Aos vinte e cinco dias do mês de setembro de dois mil e vinte e cinco, a partir das quatorze horas, por meio de videoconferência, realizou-se a sessão pública de Defesa de Tese intitulada "RECURSOS GENÔMICOS PARA A CROADINHA (Mouriri elliptica Mart. - MELASTOMATACEAE)". Os trabalhos foram instalados pela Orientadora, Professora Doutora Thannya Nascimento Soares (PPGGMP/UFG), com a participação dos demais membros da Banca Examinadora: Doutor Marco Aurélio Caldas de Pinho Pessoa Filho (Embrapa Recursos Genéticos e Biotecnologia), membro titular externo; Professor Doutor Lázaro José Chaves (PPGGMP/UFG), membro titular interno, Professora Doutora Thainara Policarpo Mendes (DAA/IFG), membro titular externo e Professora Doutora Renata de Oliveira Dias (ICB/UFG), membro titular externo. Durante a arguição os membros da banca não fizeram sugestão de alteração do título do trabalho. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Tese tendo sido a candidata aprovada pelos seus membros. Proclamados os resultados pela Professora Thannya Nascimento Soares, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, aos vinte e cinco dias do mês de setembro do ano de dois mil e vinte e cinco.

TÍTULO SUGERIDO PELA BANCA

Documento assinado eletronicamente por **MARCO AURELIO CALDAS DE PINHO PESSOA FILHO**, **Usuário Externo**, em 10/11/2025, às 08:10, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Thainara Policarpo Mendes**, **Usuário Externo**, em 10/11/2025, às 09:14, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Lazaro Jose Chaves**, **Usuário Externo**, em 10/11/2025, às 11:57, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Thannya Nascimento Soares**, **Professora do Magistério Superior**, em 11/11/2025, às 09:34, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Renata De Oliveira Dias**, **Professora do Magistério Superior**, em 12/11/2025, às 08:48, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5774462** e o código CRC **29C40375**.

Dos medos nascem as coragens; e das dúvidas, as certezas. Os sonhos anunciam outra realidade possível e os delírios, outra razão". *Eduardo Galeano*

Aos meus pais, José Maria e Edir, e às minhas filhas, Isadora e Helena, por todo amor e inspiração.
DEDICO

AGRADECIMENTOS

Agradeço primeiramente ao Programa de Pós-Graduação em Genética e Melhoramento de Plantas (PPGGMP) e à Universidade Federal de Goiás (UFG) por proporcionarem um ambiente acadêmico de qualidade e os recursos necessários para a realização desta pesquisa.

Agradeço ao apoio financeiro das instituições de fomento à pesquisa, fundamentais para a realização deste trabalho. Este estudo foi desenvolvido no contexto do Instituto Nacional de Ciência e Tecnologia em Ecologia, Evolução e Conservação da Biodiversidade (INCT–EECBio), apoiado pelo MCTIC/CNPq (processo nº 465610/2014-5), pela Fundação de Amparo à Pesquisa do Estado de Goiás (FAPEG, processo nº 201810267000023) e pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, processo nº 88887.136301/2017-00).

À minha orientadora, Prof.ª Dr.ª Thannya Nascimento Soares, e à minha coorientadora, Prof.ª Dr.ª Adriana Maria Antunes Taquary, expresso minha mais profunda gratidão. Obrigada por acreditar em mim, pela paciência, pelos ensinamentos e pelo apoio em todas as fases desse processo. Suas orientações fizeram toda a diferença. Agradeço especialmente à professora Thannya por sua sensibilidade e compreensão diante da minha condição de mãe solo, professora e pesquisadora, equilibrando exigências acadêmicas com empatia e incentivo constante. Seu exemplo de dedicação e humanidade foi fundamental para que eu pudesse avançar neste trabalho com confiança e motivação.

Aos professores e colegas que contribuíram para a construção deste trabalho, em especial aos membros do grupo EuGeM, pela parceria na realização de eventos e trocas de conhecimento, meu muito obrigada.

Aos amigos que estiveram ao meu lado e me ajudaram a seguir em frente, Daniel Lucas, Edson Júnior, Thainara, Lígia, Marla, Wagner, Ana Beatriz, Kássia, Cynthia e Cármem, cada um, à sua maneira, foi essencial nesta caminhada.

Agradeço à minha família, que sempre esteve ao meu lado. Aos meus pais, José Maria e Edir, obrigada por todo apoio, amor e incentivo. Ver minha mãe vencer o câncer me ensinou o verdadeiro significado de força.

Às minhas filhas, Isadora e Helena, minha razão maior. Vocês são meu porto seguro e minha inspiração diária. Cada passo que dou é por vocês e para vocês.

Agradeço também a Deus, por me sustentar nos momentos em que minhas forças pareciam acabar. Durante esse processo, enfrentei momentos muito difíceis: passei por uma cirurgia, chorei muitas vezes, vivi o luto pela perda da minha avó Helena, além de enfrentar o fim de um ciclo seguido de um importante recomeço em minha vida pessoal. Mas em nenhum momento pensei em desistir.

Esta tese é fruto de muita resistência, superação e amor. Obrigada a todos que, de alguma forma, fizeram parte dessa história.

Sumário

RESUMO

BRITO, J. B. P. **Recursos Genômicos para a Croadinha (*Mouriri elliptica* Mart. - MELASTOMATACEAE).** 2025. Tese (Doutorado em Genética e Melhoramento de Plantas) – Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2025.[1]

A família Melastomataceae, possui cerca de 5700 espécies, apresenta importância ecológica e evolutiva, contudo carece de recursos genômicos disponíveis. Essa lacuna dificulta avanços em estudos filogenéticos e na genética da conservação de suas espécies. Diante desse cenário, esta tese estabeleceu como objetivo central gerar e analisar dados genômicos de *Mouriri elliptica*. Buscamos com isso construir uma base sólida de conhecimento para a família e fornecer subsídios para investigações genéticas e evolutivas futuras. Os resultados desta tese estão divididos em três capítulos. No primeiro, realizamos uma revisão da literatura sobre tecnologias de sequenciamento de ácidos nucleicos, abordando desde o método de Sanger até as abordagens de segunda e terceira geração, como as baseadas em síntese e em leitura de molécula única. Também são discutidos os princípios de montagem e anotação genômica, a importância da bioinformática para análise de grandes volumes de dados e o papel da genômica na conservação de recursos genéticos vegetais. O segundo capítulo descreve o sequenciamento e caracterização do genoma cloroplastidial de *M. elliptica*, que apresentou 156.791 pares de bases, estrutura circular típica e quatro regiões principais: LSC (86.943 pb), SSC (17.234 pb) e duas regiões invertidas (26.307 pb cada). Foram identificados 79 genes codificadores de proteínas, 4 genes de rRNA e 30 genes de tRNA. A análise filogenética posicionou a espécie próxima ao gênero *Memecylon*, confirmando sua inclusão na subfamília Olisbeoideae. Os *primers* que desenvolvemos para os genes *MatK* e *rbcL* obtiveram sucesso nas análises *in silico*, para aplicação potencial como DNA barcodes, contribuindo para a identificação molecular dentro da família. O terceiro capítulo apresenta a montagem parcial do genoma nuclear, composta por 47.075 *scaffolds*, com N50 de 26.418 pb e tamanho estimado de 354 Mb. A avaliação de completude via BUSCO indicou 90,8% de ortólogos completos. Foram identificados 6.602 loci de microssatélites e 119.655 RNAs não codificantes, incluindo rRNAs, tRNAs, miRNAs e snoRNAs. A partir desses dados, foram desenvolvidos primers para amplificação de microssatélites em sistemas de PCR multiplex. Esta tese disponibiliza um conjunto de recursos genômicos para *M. elliptica*, o que constitui uma contribuição para o avanço do conhecimento genômico do gênero. A caracterização dos genomas cloroplastidial e nuclear amplia o entendimento sobre a diversidade genômica e a história evolutiva da espécie, enquanto os marcadores moleculares desenvolvidos oferecem ferramentas aplicáveis a estudos de diversidade, estrutura populacional e filogenia, com potencial de uso em estratégias futuras de conservação e manejo.

Palavras-chave: *barcoding;* bioinformática*;* Cerrado; Croadinha; diversidade genética; genoma cloroplastidial; genoma nuclear; microssatélites; PCR multiplex.

---

[1] Orientadora: Profª. Drª. Thannya Nascimento Soares. ICB/UFG
Co-orientadora: Prof.ª Dr.ª Adriana Maria Antunes Taquary. ICB/UFG

ABSTRACT

BRITO, J. B. P. **Genomic Resources for Croadinha (*Mouriri elliptica* Martius - MELASTOMATACEAE).** 2025. Thesis (Doctorate in Genetics and Plant Breeding) – Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2025.[1]

The Melastomataceae family, which has about 5,700 species, is ecologically and evolutionarily important, but lacks available genomic resources. This gap hinders advances in phylogenetic studies and in the conservation genetics of its species. Given this scenario, this thesis established as its central objective the generation and analysis of genomic data from *Mouriri elliptica*. Our aim is to build a solid knowledge base for the family and provide support for future genetic and evolutionary research. The results of this thesis are divided into three chapters. In the first, we review the literature on nucleic acid sequencing technologies, covering everything from the Sanger method to second- and third-generation approaches, such as those based on synthesis and single-molecule reading. We also discuss the principles of genome assembly and annotation, the importance of bioinformatics for analyzing large volumes of data, and the role of genomics in the conservation of plant genetic resources. The second chapter describes the sequencing and characterization of the chloroplast genome of *M. elliptica*, which had 156,791 base pairs, a typical circular structure, and four main regions: LSC (86,943 bp), SSC (17,234 bp), and two inverted regions (26,307 bp each). Seventy-nine protein-coding genes, four rRNA genes, and 30 tRNA genes were identified. Phylogenetic analysis placed the species close to the genus Memecylon, confirming its inclusion in the subfamily Olisbeoideae. The primers we developed for the *MatK* and *rbcL* genes were successful in *in silico* analyses for potential application as DNA barcodes, contributing to molecular identification within the family. The third chapter presents the partial assembly of the nuclear genome, consisting of 47,075 scaffolds, with an N50 of 26,418 bp and an estimated size of 354 Mb. The completeness assessment via BUSCO indicated 90.8% complete orthologs. A total of 6,602 microsatellite loci and 119,655 non-coding RNAs were identified, including rRNAs, tRNAs, miRNAs, and snoRNAs. Based on these data, primers were developed for microsatellite amplification in multiplex PCR systems. This thesis provides a set of genomic resources for *M. elliptica*, which contributes to the advancement of genomic knowledge of the genus. The characterization of the chloroplast and nuclear genomes broadens the understanding of the genomic diversity and evolutionary history of the species, while the molecular markers developed offer tools applicable to studies of diversity, population structure, and phylogeny, with potential for use in future conservation and management strategies.

Keywords: barcoding; bioinformatics; Cerrado; Croadinha; genetic diversity; chloroplast genome; nuclear genome; microsatellites; multiplex PCR.

---

[1] Advisor: Profª. Drª. Thannya Nascimento Soares. ICB/UFG
Co-advisor: Prof.ª Dr.ª Adriana Maria Antunes Taquary. ICB/UFG

1 INTRODUÇÃO

O conhecimento da biodiversidade vegetal é crucial para a compreensão dos ecossistemas e para a formulação de estratégias de conservação. Em escala global, a biodiversidade está sofrendo forte pressão devido a fatores antrópicos, tais como, a fragmentação e destruição de habitats, a exploração desenfreada de recursos e as mudanças climáticas (IPBES, 2019). Dessa forma, a estabilidade ecossistêmica e a manutenção dos serviços ambientais essenciais ao bem-estar humano são ameaçadas (Díaz et al., 2019). Por essa razão, a conservação da diversidade genética, elemento central da biodiversidade, torna-se uma prioridade inadiável para a resiliência das espécies (Frankham & Ballou, 2004; Hoban et al., 2023).

No contexto brasileiro, o Cerrado destaca-se como um bioma de grande relevância ecológica, reconhecido como um *hotspot* mundial devido à sua grande variedade de espécies endêmicas (Klink & Machado, 2005; Myers et al., 2000). A flora do Cerrado também possui papel de sustentação para populações tradicionais, que fazem uso de seus recursos em diversas formas, como, alimentares, paisagísticos e medicinais (Castro Oliveira & Viveiro, 2013; Antônio et al., 2024). Contudo, a intensa expansão agropecuária, associada à mineração e urbanização, tem provocado a fragmentação das áreas nativas, demandando o fortalecimento das bases biológicas para a formulação de estratégias de conservação eficazes (Klink & Machado, 2005; Souza, Telles & Diniz-Filho, 2016a).

Assim, a caracterização genômica de espécies nativas revela-se como uma ferramenta importante. Informações genômicas provêm os subsídios necessários para decifrar a organização, a função dos genes e a história evolutiva das espécies (Stein, 2001; Yandell & Ence, 2012). Estes dados são importantes para quantificar a diversidade intraespecífica, identificar unidades evolutivas e monitorar os impactos da fragmentação de habitats, fornecendo base para o manejo adaptativo e a conservação (Allendorf, Hohenlohe & Luikart, 2010; Hoban et al., 2023). A caracterização genômica de espécies nativas é fundamental para entender sua diversidade e evolução. Os dados genômicos permitem compreender a organização e a função dos genes, além de revelar aspectos

evolutivos das espécies (Stein, 2001; Yandell & Ence, 2012). Os dados são cruciais para quantificar a diversidade intraespecífica, identificar unidades evolutivas significativas e populações prioritárias para conservação, monitorar os impactos da fragmentação e projetar estratégias de manejo adaptativo frente às mudanças climáticas (Allendorf, Hohenlohe & Luikart, 2010; Funk et al., 2012; Hoban et al., 2023).

A família Melastomataceae destaca-se no Cerrado pela grande diversidade de espécies e relevância ecológica, além de seu potencial farmacológico (Albuquerque et al., 2013). Com cerca de 5.700 espécies distribuídas globalmente em regiões tropicais e subtropicais, a família possui no Brasil um de seus principais centros de diversidade, contando com aproximadamente 1.436 espécies (1.520 espécies segundo a Flora do Brasil (2025), sendo 1.007 endêmicas), das quais 66,25% são endêmicas, concentrando-se majoritariamente nos biomas Amazônia, Mata Atlântica e Cerrado (Stevens & Davis, 2005; Wink et al., 2024). Suas espécies desempenham papéis ecológicos cruciais, atuando como recursos alimentares para a fauna, contribuindo para processos de polinização e dispersão de sementes, e participando da ciclagem de nutrientes (Clausing & Renner, 2001). Adicionalmente, muitas espécies apresentam compostos bioativos com propriedades antioxidantes, anti-inflamatórias e antimicrobianas, conferindo-lhes notável potencial biotecnológico (Reginato et al., 2016).

Mouriri elliptica Mart. (Melastomataceae) popularmente conhecida como croada, croadinha, ou coroa-de-frade, é uma árvore de pequeno porte ou arbusto, atingindo até 5 metros de altura. Apresenta tronco com casca fissurada e ramos novos achatados e sulcados. Suas folhas são opostas, de forma elíptica a oblongo-elíptica (característica que dá nome à espécie), com base arredondada e textura cartácea a subcoriácea. As flores são brancas, solitárias ou aos pares, com estames de anteras poricidas e conectivo ventralmente biapendiculado. Seus frutos são bagas globosas, glabras, de coloração vinácea a negra quando maduros, com polpa suculenta e doce, contendo numerosas sementes pequenas. Sua distribuição é restrita e endêmica ao bioma Cerrado. A espécie é típica e abundante nas formações savânicas (Cerrado *stricto sensu* e Cerradão), onde é considerada um elemento fundamental do estrato arbustivo-arbóreo (Völtz & Goldenberg, 2020). Neste bioma, a espécie demonstra notável adaptação a solos pobres e ácidos, exibindo características morfofisiológicas típicas de plantas savânicas, como folhas coriáceas e sistema radicular robusto (Furquim et al., 2018). Ecologicamente, *M. elliptica* é reconhecida como uma espécie pioneira, importante na

sucessão secundária, e serve como fonte de recursos para a fauna frugívora, auxiliando na dispersão e manutenção das teias tróficas (Assis et al., 2016; Silveira et al., 2014) (Figura 1.1).

Figura 1.1 *Mouriri elliptica* Mart. (Melastomataceae), conhecida popularmente como croadinha ou coroa-de-frade.

Além de seu papel ecológico, *M. elliptica* possui significativa importância etnobotânica e econômica. Seus frutos saborosos são consumidos in natura e apreciados pelas populações locais, possuindo potencial para exploração alimentícia e produção de polpas (Assis et al., 2016). Suas folhas são tradicionalmente empregadas na medicina popular para o tratamento de afecções como úlceras e inflamações, uso este corroborado por estudos fitoquímicos que identificaram a presença de compostos fenólicos com atividade antioxidante e anti-inflamatória (Machado, Aquino & Neves, 2014; Silveira et al., 2014; Vasconcelos et al., 2010).

Apesar de sua relevância ecológica e socioeconômica, *M. elliptica* ainda é pouco estudada sob os aspectos genético e genômico. Uma análise criteriosa das bases de

dados públicas confirma a ausência completa de sequências genômicas ou plastidiais montadas, bem como de estudos aprofundados com marcadores moleculares hipervariáveis, como microssatélites (SSRs). Esta lacuna no conhecimento impede uma compreensão de aspectos fundamentais de sua biologia, como sua diversidade genética intra e interpopulacional, sua estruturação filogeográfica, seu potencial adaptativo e seu histórico de domesticação. A carência de tais informações constitui um obstáculo direto à elaboração de estratégias eficazes de conservação, manejo sustentável e melhoramento genético (Ekblom & Galindo, 2010; Freeland, Kirk & Petersen, 2012).

As novas tecnologias de sequenciamento permitem superar essa limitação. A montagem *de novo* de genomas, tanto nucleares quanto organelares, aliada a abordagens bioinformáticas, permite a caracterização de sua estrutura genômica e o desenvolvimento *in silico* de marcadores genéticos customizados (Formenti et al., 2022; Goodwin, McPherson & McCombie, 2016; Nagarajan & Pop, 2013). Técnicas como o SSR-Seq, que utilizam conjuntos de dados genômicos para o desenvolvimento de marcadores microssatélites, permitem a identificação de centenas de loci polimórficos de forma rápida e eficiente (Šarhanová et al., 2018; Zalapa et al., 2012). A subsequente multiplexagem desses marcadores viabiliza a genotipagem de larga escala de forma economicamente viável (Oliveira et al., 2021). Além disso, o uso de regiões padrão para DNA barcoding, como os genes *MatK* e *rbcL* do cloroplasto e a região nuclear *ITS*, é aplicado para identificação de espécies e para pesquisas sobre sua evolução e distribuição geográfica (Hebert et al., 2003; Kress & Erickson, 2007).

Diante deste contexto, o objetivo principal deste trabalho foi disponibilizar recursos genômicos para *Mouriri elliptica* Mart. Os objetivos específicos incluíram:

• Montar e anotar o genoma cloroplastidial completo, identificando genes, regiões regulatórias e elementos repetitivos;

• Montar e caracterizar um rascunho do genoma nuclear;

• Desenhar *primers* para amplificação conjunta (PCR multiplex) e detecção via sequenciamento de amplicons de regiões microssatélites (SSR-Seq) e regiões potencialmente úteis como DNA barcode.

Os resultados obtidos poderão servir como base para estudos filogenéticos em Melastomataceae, e os dados genômicos gerados podem contribuir para a compreensão

da diversidade genética de *M. elliptica*, auxiliando futuras iniciativas de conservação e manejo sustentável.

2 REVISÃO DA LITERATURA: TECNOLOGIAS DE SEQUENCIAMENTO DE ÁCIDO NÚCLEICO E APLICAÇÕES PARA A CONSERVAÇÃO DE RECURSOS GENÉTICOS

2.1 SEQUENCIAMENTO DE ÁCIDOS NUCLEICOS

O sequenciamento de DNA permite obter informações detalhadas sobre a ordem dos nucleotídeos (adenina, timina, citosina e guanina) em uma molécula de ácido nucleico é realizada através do sequenciamento de DNA. Este processo fornece acesso ao material genético de organismos de qualquer origem que contenham unidades funcionais de hereditariedade com valor real ou potencial (Sanger, Nicklen & Coulson, 1977; CBD, 1992; Frankham & Ballou, 2004).

O sequenciamento se divide em três etapas principais: preparo da biblioteca, leitura e análise de dados. A primeira etapa é o preparo da amostra, onde a molécula de DNA (ou RNA, após conversão em DNA complementar) é fragmentada e adaptadores são ligados, o que resulta na biblioteca de sequenciamento. Na etapa seguinte, a leitura dos fragmentos ocorre na plataforma. O processo combina reações químicas e métodos de detecção (físicos, ópticos ou elétricos) para determinar a ordem precisa dos nucleotídeos. Finalmente, a montagem e a análise da sequência são realizadas utilizando ferramentas de bioinformática para alinhar e montar os fragmentos de leitura (*reads*, sequências curtas geradas pelas plataformas) com leituras sobrepostas, buscando formar sequências contíguas (Goodwin, McPherson & McCombie, 2016; Nagarajan & Pop, 2013).

Por muitos anos, o sequenciamento foi predominantemente realizado utilizando o método de terminação de cadeia, desenvolvido por Sanger e seus colaboradores (Sanger, Nicklen & Coulson, 1977). Contudo, o surgimento das novas tecnologias, o Sequenciamento de Nova Geração (NGS), transformou a genômica na década de 2000. As tecnologias NGS reduziram drasticamente o custo por base e o tempo de execução, permitindo ainda trabalhar com menos material genético (Goodwin, McPherson & McCombie, 2016; Metzker, 2010).

2.2 SEQUENCIAMENTO DE SEGUNDA GERAÇÃO

Um dos primeiros exemplos da tecnologia NGS foi o sequenciador 454 Life Sciences, desenvolvido em 2004. Essa plataforma utilizava o princípio do pirosequenciamento, um método que detecta a liberação de pirofosfato durante a adição de um nucleotídeo à fita de DNA em síntese. A liberação de pirofosfato desencadeia uma cascata de reações enzimáticas. A luz emitida por essa reação é detectada e quantificada. A intensidade da luz emitida é diretamente proporcional ao número de nucleotídeos incorporados sequencialmente (em casos de homopolímeros), permitindo identificar qual nucleotídeo foi adicionado (Ronaghi, Uhlén & Nyrén, 1998).

O processo de sequenciamento na plataforma 454 envolvia a amplificação por PCR em emulsão, onde cada fragmento de DNA era amplificado dentro de microesferas (*beads*). Essas *beads* eram depositadas em uma placa de micro-poços (*Picotiter*), onde o sequenciamento propriamente dito ocorria. Durante o sequenciamento, a detecção do sinal de luz emitido pela liberação de pirofosfato em tempo real permitia o registro da incorporação de cada nucleotídeo (Margulies et al., 2005).

A empresa Solexa, posteriormente adquirida pela Illumina, desenvolveu a plataforma de sequenciamento de segunda geração que viria a dominar o mercado. A principal diferença para o 454 está na etapa de amplificação, que utiliza PCR em ponte (*bridge PCR*) em uma superfície sólida (Goodwin, McPherson & McCombie, 2016). Além disso, a plataforma Illumina utiliza índices (*barcodes*) que permitem a amostragem múltipla (*multiplexação*) de diversas amostras em uma mesma corrida, aumentando a eficiência do processo. O 454 teve sucesso inicial, mas a limitação no comprimento dos fragmentos e os custos mais altos permitiram que a tecnologia Illumina dominasse o mercado (Mardis, 2008; Metzker, 2010).

O sequenciamento na plataforma Illumina envolve o preparo de bibliotecas, onde o DNA é fragmentado aleatoriamente e adaptadores são ligados às extremidades. Essa etapa é essencial para a clusterização, onde os fragmentos de DNA são amplificados por PCR em ponte, gerando grupos clonais (*clusters*) que amplificam o sinal de fluorescência. Os fragmentos com adaptadores hibridizam-se a oligonucleotídeos complementares ancorados na superfície de uma *flow cell*. Após a fixação, os fragmentos se curvam, formando pontes de amplificação com *primers* adjacentes, gerando os *clusters*

de fragmentos clonados (Goodwin, McPherson & McCombie, 2016; Mardis, 2008; Metzker, 2010).

Após a clusterização, o sequenciamento ocorre por meio do método de terminação reversível. Nesse método, nucleotídeos fluorescentes (com cores distintas ou método de detecção de cor única) e DNA polimerase são adicionados à *flow cell*. Os nucleotídeos possuem um grupo bloqueador que interrompe a síntese da fita após a incorporação de apenas um nucleotídeo por ciclo. A fluorescência emitida é capturada em uma imagem para identificar qual nucleotídeo foi adicionado. Em seguida, o grupo bloqueador é removido quimicamente, e o processo é repetido até que a leitura dos *clusters* seja finalizada (Mardis, 2008; Metzker, 2010).

A Illumina utiliza comumente a estratégia de sequenciamento pareado (*paired-end*), na qual ambas as extremidades de cada fragmento de DNA são sequenciadas, produzindo fragmentos de leitura aos pares. Essa abordagem é vantajosa, pois melhora a montagem (fornecendo informações sobre o fragmento original), facilita a detecção de inserções e deleções e permite a identificação de rearranjos genômicos (Buermans & Dunnen, 2014).

Os sequenciadores de segunda geração (Illumina) geram um grande volume de sequências e possibilitam o sequenciamento de genomas inteiros em dias ou horas. No entanto, o principal desafio na montagem de fragmentos de leitura curtos (*short reads*) é lidar com regiões de sequências repetitivas, o que frequentemente leva à fragmentação do genoma montado. Para refinar a montagem e minimizar erros na chamada de bases, é crucial aumentar a cobertura (*depth*) do sequenciamento, ou seja, o número médio de vezes que uma determinada base do genoma é sequenciada. A tecnologia Illumina apresenta baixas taxas de erros (principalmente substituições de nucleotídeos) em comparação com outras plataformas NGS (Goodwin, McPherson & McCombie, 2016).

2.3 SEQUENCIAMENTO DE TERCEIRA GERAÇÃO (TGS)

As tecnologias de sequenciamento de terceira geração (TGS) se baseiam no sequenciamento de moléculas únicas (SMRT), o que dispensa a necessidade de amplificação por PCR. Essas plataformas incluem a tecnologia PacBio SMRT, que

detecta a liberação de luz durante a incorporação do nucleotídeo, e a tecnologia de Nanoporos (*Nanopore Sequencing* - ONT), da Oxford Nanopore. A ONT lê o DNA enquanto ele atravessa um nanoporo. As alterações na corrente elétrica são detectadas e usadas para identificar a sequência de nucleotídeos (Goodwin, McPherson & McCombie, 2016; Jain et al., 2018).

O TGS oferece a vantagem de gerar fragmentos de leitura longos (*long-reads*), que facilitam a montagem *de novo* e permitem a detecção direta de haplótipos. Os fragmentos de leitura longos conseguem ultrapassar as regiões repetitivas que fragmentam as montagens baseadas em fragmentos de leitura curtos. A detecção direta de haplótipos refere-se à capacidade de ler alelos em um cromossomo em uma única leitura contínua, permitindo diferenciar e sequenciar os dois conjuntos de alelos (haplótipos) de um organismo. Além disso, as TGS apresentam menor necessidade de material de partida e portabilidade do equipamento (Goodwin, McPherson & McCombie, 2016; Jain et al., 2018).

As tecnologias de sequenciamento de alta capacidade (NGS/TGS) marcaram uma mudança na genômica, pois reduziram o custo e aumentaram drasticamente a escala da geração de dados. Como revisado por Goodwin, McPherson & McCombie (2016), cada plataforma possui características operacionais distintas que as tornam mais ou menos adequadas para diferentes objetivos experimentais. A Tabela 2.1 sintetiza as principais características das plataformas de sequenciamento, baseando-se na comparação apresentada por esses autores, com a adição de informações atualizadas sobre a tecnologia de nanoporo.

Tabela 2.1 Comparativo das plataformas de sequenciamento de DNA de diferentes gerações, adaptado de Goodwin et al. (2016).

Característica	1ª Geração (Sanger)	2ª Geração (NGS - *Short-Read*)	3ª Geração (TGS - *Long-Read*)
Princípio de Sequenciamento	Eletroforese capilar com terminação de cadeia (didesoxi)	Sequenciamento por Síntese (SBS) com imagens de fluorescência (Illumina) ou detecção de pirofosfato (454)	Leitura direta de bases através de nanoporos (ONT) ou detecção de luz (PacBio SMRT)
Throughput por corrida	Baixa (~0,001−0,1 Gb)	Alta a Muito Alta (0,3 Gb - 6 Tb, dependendo do equipamento)	Variável (0,1 - 50 Gb ou mais, dependendo do dispositivo e tempo de corrida)
Comprimento da Leitura	Longo (~500−1000 pb)	Curto (50−600 pb)	Muito Longo (>10 kb, até centenas de kb / >2 Mb reportados)
Precisão por Base	Muito Alta (>99,99%, Q30+)	Alta (>99,9%, Q30+ para a maioria das plataformas)	Moderada-Alta* (~99%/Q20+ após correção; crua ~95−98%)
Tempo de Corrida	Horas (por amostra)	1−6 dias	Minutos a Dias (0,5−48h, tempo real)
Aplicação Ideal	Validação, clonagem, genes únicos	Ressequenciamento, transcriptômica, genotipagem (GBS, RAD-seq)	Montagem *de novo*, faseamento de haplótipos, detecção de variações estruturais, epigenética
Custo Relativo (por base)	Alto	Muito Baixo	Moderado (em queda rápida)
Necessidade de PCR	Sim	Sim (Amplificação clonal obrigatória)	Não (Sequenciamento de molécula única - SMRT)
Portabilidade	Não	Não (equipamentos de bancada)	Sim (MinION é um dispositivo USB)

Independentemente da plataforma escolhida, a etapa de sequenciamento gera um volume massivo de dados brutos na forma de fragmentos de leitura, cujo processamento, montagem e análise dependem criticamente de ferramentas de bioinformática (Ejigu & Jung, 2020; Stadtländer, 2018).

2.4 GENOMAS DE CLOROPLASTO: UMA FERRAMENTA-CHAVE EM FILOGENÔMICA E CONSERVAÇÃO

O DNA do cloroplasto (cpDNA), ou plastoma, é uma molécula de DNA extranuclear de herança predominantemente uniparental (geralmente materna em angiospermas). Sua estrutura é tipicamente circular e altamente conservada na maioria das plantas, com um tamanho que varia entre 120 e160 kpb (Shinozaki et al., 1986). A organização do plastoma se caracteriza por duas regiões de cópia única (LSC e SSC) separadas por duas repetições invertidas (IRA e IRB). O plastoma é relevante para a genômica e evolução por suas características intrínsecas, como a herança uniparental, que simplifica a análise de linhagens maternas para estudos de filogeografia, e sua taxa de evolução lenta, que confere maior estabilidade para resolver relações filogenéticas em níveis taxonômicos mais altos. Além disso, sua estrutura conservada facilita a comparação e o alinhamento de genomas entre espécies distantes (Ohyama et al., 1986; Shinozaki et al., 1986). Historicamente, o sequenciamento do plastoma foi pioneiro na genômica de organelas, com a publicação do genoma completo do tabaco (*Nicotiana tabacum*) e da hepática (*Marchantia polymorpha*) ainda na década de 1980, utilizando o sequenciamento de primeira geração (Sanger) (Shinozaki et al., 1986; Ohyama et al., 1986).

O advento das tecnologias NGS mudou a forma como o plastoma é estudado. Anteriormente, as análises de diversidade e filogenia em plantas dependiam da amplificação e sequenciamento de regiões específicas (genes únicos ou regiões intergênicas). Atualmente, a rotina envolve a montagem do plastoma completo, mesmo a partir do DNA genômico total, utilizando plataformas de segunda geração (Illumina) ou de terceira geração (PacBio/ONT) (Goodwin, McPherson & McCombie, 2016). A capacidade de sequenciar o genoma completo do cloroplasto possibilita a Filogenômica, uma abordagem que emprega o genoma de organelas na íntegra para inferências evolutivas. A análise de centenas de genes do plastoma, em contraste com a limitação de poucos marcadores, aumenta significativamente a resolução e a robustez das hipóteses filogenéticas. Essa abordagem é crucial para esclarecer complexos de espécies ou divergências recentes, onde marcadores moleculares curtos não oferecem resolução suficiente. Adicionalmente**,** é útil para a Filogeografia e Fluxo Gênico, ao mapear a distribuição geográfica de haplótipos com alta precisão, o que permite a compreensão da

história demográfica, eventos de colonização e padrões de dispersão em espécies (Freeland, Kirk & Petersen, 2011).

No contexto da conservação de recursos genéticos, a análise do plastoma é crucial para a identificação de Unidades de Conservação (UCs), pois as informações detalhadas sobre a estrutura genética e a conectividade (ou isolamento) entre populações guiam a correta delimitação dessas unidades (Funk et al., 2012). Além disso, o plastoma é importante para o rastreamento materno em estudos de hibridização e introgressão, pois sua herança estritamente materna permite rastrear com precisão o parental feminino, um dado essencial para o manejo de populações híbridas naturais ou em programas de melhoramento (Frankham, Ballou & Briscoe, 2002). Assim, o sequenciamento do plastoma, graças ao NGS, se consolidou como um componente importante da Genômica da Conservação (Allendorf, Hohenlohe & Luikart, 2010).

2.5 BIOINFORMÁTICA: PRINCÍPIOS GERAIS DA GENÔMICA

A bioinformática é um campo multidisciplinar que integra biologia, ciência da computação, estatística e matemática com o objetivo de analisar e interpretar dados biológicos em larga escala (Luscombe et al., 2004). Impulsionada pelos avanços contínuos nas tecnologias de sequenciamento e pelo aumento exponencial de dados genômicos, a bioinformática se consolidou a partir da década de 1980 como uma resposta à necessidade (Giani et al., 2020; Staden, 1980).

A bioinformática, que aplica procedimentos e técnicas computacionais e estatísticas, tornou-se fundamental para a compreensão dos dados biológicos em diversas áreas do conhecimento. Por meio do desenvolvimento de ferramentas avançadas, ela possibilita a obtenção, o processamento, o armazenamento, a distribuição, a análise e a interpretação de informações biológicas de alta dimensão. Ao permitir o acesso eficiente a bancos de dados, a bioinformática acelera a investigação em setores como agronomia, biotecnologia e medicina (Borém, 2001; Iquebal et al., 2015). Em particular, ela desempenha um papel central na anotação genômica, fornecendo os métodos necessários para identificar genes, prever sua função e analisar sua organização e evolução (Mount, 2004; Yandell & Ence, 2012).

O Brasil teve um papel pioneiro na aplicação dessa ciência com o Laboratório de Bioinformática da Unicamp. Este laboratório foi responsável pela montagem, em 2000, do genoma da *Xylella fastidiosa*, bactéria causadora do amarelinho-da-laranja e o primeiro organismo sequenciado integralmente no país (Simpson et al., 2000).

Com o acúmulo global de dados de sequência gerados em laboratórios, tornou-se essencial organizar essas informações em bases acessíveis e detalhadas, tanto para evitar redundâncias nas pesquisas quanto para democratizar o acesso por um maior número de cientistas. Os bancos de dados armazenam sequências de DNA e RNA, genomas inteiros, sequências de proteínas, estruturas tridimensionais e diversos produtos da era genômica. O armazenamento e a compreensão desses dados representam um desafio contínuo, mas de grande importância para a genômica (Bairoch & Apweiler, 2000; Stadtländer, 2018).

O Centro Nacional para Informação Biotecnológica (NCBI) é um dos principais recursos globais, atuando como um centro de agregação de dados genômicos. No NCBI, a Taxonomia do NCBI é um recurso hierárquico e único que organiza nomes de organismos de todos os domínios dos seres vivos. Embora esta taxonomia se baseie nos nomes válidos e atuais de acordo com as autoridades e códigos de nomenclatura, deve-se sempre consultar a literatura científica relevante para a informação taxonômica mais robusta (Schoch et al., 2020). O principal banco de dados presente no NCBI é o GenBank, que faz parte da *International Nucleotide Sequence Database Collaboration* (INSDC). O GenBank armazena uma vasta gama de informações, incluindo sequências nucleotídicas (DNA e RNA), genomas completos e a organização taxonômica associada a cada registro (Sayers et al., 2022). O NCBI também oferece o banco de dados RefSeq, que visa disponibilizar sequências de referência de alta qualidade para genomas, transcrições e proteínas (O'Leary et al., 2016).

Diversas ferramentas de bioinformática permitem o acesso e a análise desses bancos de dados. Entre elas, o BLAST (*Basic Local Alignment Search Tool*) é a mais utilizada. O BLAST realiza a busca por similaridade de uma sequência-alvo em um banco de dados, identificando correspondências curtas e iniciando alinhamentos com base em regiões chamadas *hot spots*. A ferramenta fornece alinhamentos e estatísticas detalhadas sobre a qualidade das comparações (Ye, McGinnis & Madden, 2006) (Figura 2.1).

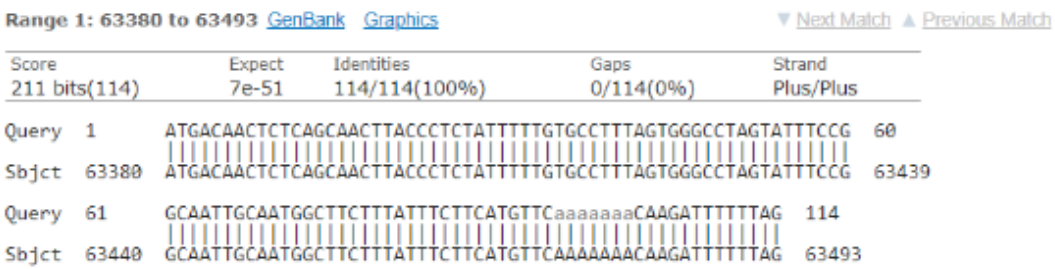

Figura 2.1 Resultado utilizando o programa BLAST para a busca de similaridade. O Query representa o segmento de DNA sequenciado, e apresenta 100% de homologia com o gene psbI (Sbjct).

Apesar da importância do GenBank, outros bancos de dados como o *European Nucleotide Archive* (ENA) e o *DNA Data Bank of Japan* (DDBJ) também disponibilizam sequências nucleotídicas (Karsch-Mizrachi et al., 2018). O UniProt é uma base de dados essencial que combina sequências curadas manualmente (*UniProtKB/Swiss-Prot*) e sequências anotadas automaticamente (*UniProtKB/TrEMBL*), representando uma fonte abrangente de informações proteicas (The UniProt Consortium, 2021). Outro recurso fundamental com foco na anotação de proteínas é o Consórcio InterPro, que integra diversos bancos de dados (como Pfam, PROSITE, CATH-Gene3D e outros) para fornecer informações sobre famílias proteicas, domínios e locais funcionais (Blum et al., 2021b). Existem ainda bancos de dados especializados, como NONCODE, Dfam, Pseudogene.org e miRBase, que fornecem dados específicos para a anotação de RNAs não codificantes, elementos transponíveis, pseudogenes e microRNAs, respectivamente (Ejigu & Jung, 2020; Harrison et al., 2005).

Dados produzidos pelo NGS apresentam um grande volume, e exigem processos intensivos de computação. Nesse contexto, os *pipelines* de bioinformática são um conjunto de algoritmos e ferramentas de programas que processam esses dados em uma ordem definida. Eles são pacotes abrangentes que exploram relevantes informações que estão disponíveis a partir da predição de genes e de outros elementos genômicos, sendo dessa forma, capazes de processar grandes quantidades de dados de sequenciamento em diversos ambientes e bancos de dados (Cantarel et al., 2008).

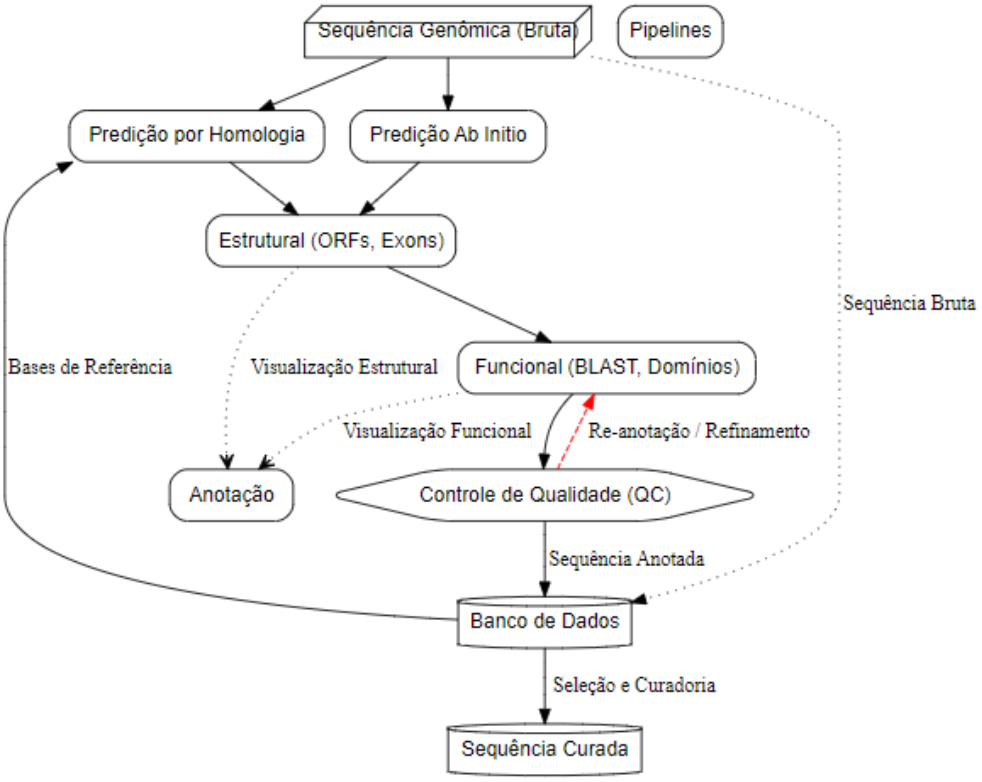

Figura 2.2 Fluxograma com as etapas relacionadas a anotação genômica baseado em Ejigu & Jung, 2020.

2.6 MONTAGEM DE GENOMAS

Após o sequenciamento e a geração dos fragmentos de leitura (*reads*), inicia-se a etapa fundamental de montagem do genoma. Esse processo, realizado por meio de ferramentas de bioinformática, busca reconstruir a sequência completa do genoma a partir dos fragmentos sequenciados (Mardis, 2008; Pop, Salzberg & Shumway, 2002). A confiabilidade da montagem depende de fatores como o tamanho dos *contigs*, a qualidade dos dados e a cobertura do sequenciamento (Deschamps & Llaca, 2016; Nagarajan & Pop, 2013).

Inicialmente, as leituras passam por uma análise de qualidade, em que bases com baixo índice de confiança são removidas. O Phred *score* é uma escala logarítmica que indica a probabilidade de erro na identificação de cada base (por exemplo, Phred 20 corresponde a 1% de chance de erro, enquanto Phred 30 indica 0,1%) (Andrews, 2010; Bolger, Lohse & Usadel, 2014). Quanto maiores os valores de Phred, maior a confiança na atribuição correta de cada nucleotídeo. Programas como FastQC e Trimmomatic realizam essa filtragem e limpeza das sequências.

Na etapa seguinte, os fragmentos de leitura são combinados para formar *contigs*, que correspondem a sequências contínuas reconstruídas a partir de sobreposições entre as leituras. A união de *contigs* gera os *scaffolds*, estruturas mais longas compostas por *contigs* conectados por regiões não sequenciadas, conhecidas como gaps. Esses trechos ausentes são estimados pelos algoritmos com base nos dados disponíveis. A cobertura de sequenciamento, número médio de vezes que cada base é lida, influencia diretamente a precisão da montagem e a redução de erros, embora outros fatores, como repetitividade do genoma e qualidade dos dados, também sejam determinantes (Miller, Koren & Sutton, 2010; Schatz, Delcher & Salzberg, 2010).

A montagem pode ser realizada *de novo* ou guiada por genoma de referência. No primeiro caso, não há necessidade de genoma previamente conhecido e a estratégia é mais utilizada para espécies não-modelo. Quando há disponibilidade de genomas completos de espécies próximas, a montagem guiada por referência reduz a necessidade de cobertura extensa e recursos computacionais, embora possa dificultar a detecção de variações estruturais complexas, como inversões, translocações e grandes inserções/deleções (Ejigu & Jung, 2020; Miller, Koren & Sutton, 2010).

Para montagens *de novo*, muitos programas utilizam grafos de Bruijn, nos quais cada fragmento de leitura é dividido em subfragmentos de tamanho k (k-mers). Cada nó do grafo representa uma sequência de k-1 nucleotídeos, e os diferentes caminhos possíveis refletem sobreposições entre os k-mers. Bolhas e ramificações surgem devido a erros de sequenciamento ou regiões repetitivas, e os algoritmos selecionam os caminhos mais consistentes considerando a profundidade de cobertura e a qualidade das leituras (Schatz, Delcher & Salzberg, 2010; Zimin et al., 2013).

A avaliação da montagem é realizada por meio de métricas que verificam contiguidade, integridade e completude do genoma. QUAST (Quality Assessment Tool for Genome Assemblies) fornece estatísticas detalhadas sobre *contigs* e *scaffolds* (Gurevich et al., 2013). Já a ferramenta BUSCO (Benchmarking Universal Single-Copy Orthologs) avalia a completude do espaço gênico amostrado, verificando a presença de genes ortólogos universais e altamente conservados, o que indica a cobertura do genoma em termos de conteúdo funcional (Simão et al., 2015; Ejigu & Jung, 2020).

2.7 ANOTAÇÃO DE GENOMAS

Concluída a montagem do genoma, a próxima etapa é a anotação genômica. Este processo visa identificar e localizar todos os elementos genômicos, tanto codificantes quanto não codificantes, dentro da sequência consenso (Stein, 2001). A anotação genômica é fundamental para converter uma sequência bruta em conhecimento biológico funcional e pode ser dividida em duas categorias principais: anotação estrutural e anotação funcional (Eilbeck et al., 2009).

2.7.1 Anotação Estrutural

O objetivo da anotação estrutural é localizar os elementos físicos do genoma, incluindo genes codificadores de proteínas, RNAs não codificantes, regiões regulatórias como promotores e *enhancers*, sítios de ligação de ribossomos e sequências repetitivas, como os microssatélites. Estes últimos consistem em repetições curtas de 1 a 6 pares de bases distribuídas ao longo do genoma e são importantes como marcadores genéticos para estudos de diversidade, estrutura populacional e mapeamento genético (Beier et al., 2017; Yandell & Ence, 2012). Os microssatélites podem ser identificados por programas como MISA e QDD (Oliveira et al., 2021).

A identificação dos elementos estruturais pode ser realizada por métodos *ab initio*, que usam algoritmos para prever genes e regiões repetitivas com base em características intrínsecas das sequências, ou por métodos de similaridade, que comparam sequências com bancos de dados previamente anotados (Altschul et al., 1990; Besemer, Lomsadze & Borodovsky, 2001; Rodrigues, 2023; Stanke & Morgenstern, 2005). Ferramentas como AUGUSTUS e GeneMark realizam predições *ab initio*, enquanto o MAKER integra predições *ab initio* e evidências de similaridade para genomas eucarióticos, e o Prokka permite anotação rápida de genomas bacterianos (Cantarel et al., 2008; Seemann, 2014). O BUSCO (e outras métricas de completude) avalia a qualidade da anotação estrutural, verificando a presença de genes ortólogos (Simão et al., 2015).

2.7.2 Anotação Funcional

O objetivo central da anotação funcional é atribuir um significado biológico aos elementos genômicos. Este processo mapeia genes e outros elementos para termos ontológicos específicos, associando-os a vias metabólicas, processos celulares e funções moleculares, construindo um panorama funcional integrado do genoma (Gabaldón & Koonin, 2013; Yandell & Ence, 2012).

Para atingir esse objetivo, a anotação funcional emprega uma estratégia convergente de múltiplas abordagens computacionais. Inicialmente, ferramentas de varredura contra bancos de dados de domínios conservados, como o consórcio InterPro, são cruciais para identificar motivos funcionais e classificar proteínas em famílias evolutivas (Blum et al., 2021). De forma complementar, a busca por similaridade de sequência contra bancos de dados públicos curados, como UniProt para sequências proteicas e GO (*Gene Ontology*) para termos funcionais, permite inferir funções biológicas com base em ortólogos previamente caracterizados (O'Leary et al., 2016; The UniProt Consortium, 2021).

O cerne da anotação moderna reside na integração sistemática dessas evidências diversas. *Pipelines* abrangentes, como o MAKER, são especificamente projetados para combinar as predições estruturais (*ab initio*), as evidências funcionais baseadas em homologia e os dados de expressão (*RNA-seq*), aumentando exponencialmente a confiabilidade das predições e permitindo a geração de anotações funcionais detalhadas e precisas (Cantarel et al., 2008; Eilbeck et al., 2009).

A qualidade da anotação funcional é avaliada por métricas distintas, incluindo: a proporção dos genes anotados que receberam uma atribuição funcional (ou seja, taxa de sucesso); a riqueza e o nível de detalhe dessa atribuição (ex: cobertura por termos GO); e a consistência das predições entre as diferentes fontes de evidência (Cantarel et al., 2008).

Finalmente, a anotação funcional abrange mais do que a catalogação de genes. Ao considerar elementos não codificantes e regiões regulatórias, ela contribui para a compreensão de mecanismos complexos. Por exemplo, a localização de microssatélites em regiões promotoras pode sugerir um papel potencial na modulação da expressão gênica e na geração de variabilidade fenotípica (Oliveira et al., 2021; Šarhanová et al.,

2018). Dessa forma, a anotação funcional fornece a camada interpretativa essencial que transforma dados brutos de sequência em conhecimento biológico aplicável, servindo como base fundamental para pesquisas subsequentes em genômica comparativa, genética de populações e biologia molecular (Allendorf, Hohenlohe & Luikart, 2010; Ekblom & Galindo, 2011).

2.8 DESAFIOS E PERSPECTIVAS FUTURAS NA ANOTAÇÃO GENÔMICA

Desde que o genoma de *Haemophilus influenzae* foi completamente anotado em 1995 (Fleischmann et al., 1995), um marco que inaugurou a era da genômica comparativa, os métodos de anotação genômica têm avançado significativamente. No entanto, apesar dos progressos notáveis nas tecnologias de sequenciamento e montagem genômica, a anotação estrutural e funcional precisa e abrangente continua sendo um dos principais desafios da genômica moderna (Yandell & Ence, 2012). A complexidade dos genomas, a presença de elementos repetitivos, a diversidade de mecanismos regulatórios e a dificuldade em identificar elementos genômicos não codificantes e suas funções são apenas alguns dos obstáculos que os bioinformatas enfrentam (Eilbeck et al., 2009; Stein, 2001). A superação desses desafios é fundamental para a plena realização do potencial da genômica na compreensão da biologia e na resolução de problemas práticos.

Embora existam diversas ferramentas online voltadas para a anotação automatizada, a identificação de genes em eucariotos, com sua intrincada arquitetura genômica e seus complexos mecanismos de *splicing* alternativo, é consideravelmente mais complexa do que em procariontes. A predição de genes *de novo*, ou seja, a identificação de genes sem o auxílio de informações de homologia, ainda enfrenta dificuldades substanciais, e mesmo os algoritmos mais sofisticados frequentemente resultam na fusão ou fragmentação incorreta de genes. Em genomas eucariotos fragmentados, essas imprecisões tendem a superestimar o número total de genes e dificultam a identificação precisa dos códons de início e parada (Brent, 2005; Mathé et al., 2002). A anotação precisa de genomas eucarióticos é crucial para a compreensão da biologia desses organismos e para o desenvolvimento de aplicações biotecnológicas. No entanto, é importante reconhecer que a anotação *de novo* é um problema inerentemente

difícil, e que a combinação de diferentes abordagens e a validação experimental são fundamentais para garantir a precisão dos resultados (Brent, 2005).

Além disso, a percepção de que a identificação de genes bacterianos é um problema praticamente resolvido, com taxas de sensibilidade próximas de 99%, tem desacelerado o desenvolvimento de novas ferramentas *ab initio* para procariontes. Atualmente, muitas ferramentas ainda classificam centenas ou milhares de genes como "proteínas hipotéticas" por não apresentarem similaridade com genes previamente conhecidos. Embora algumas dessas sequências possam, de fato, codificar proteínas funcionais, muitas representam falsos positivos. A anotação incorreta desses elementos pode comprometer análises funcionais subsequentes (Saha & Battle, 2018). Com as abordagens atuais, é difícil refutar a funcionalidade de um *open reading frame* (ORF), o que contribui para a permanência de proteínas hipotéticas nos bancos de dados por longos períodos, representando um desafio contínuo para a anotação genômica e a compreensão funcional dos genes (Brunet et al., 2019). A resolução desse problema requer o desenvolvimento de novas abordagens que combinem informações genômicas, transcriptômicas e proteômicas, e que incorporem o conhecimento da biologia celular e da fisiologia dos organismos (Brunet et al., 2019; Saha & Battle, 2018).

Novas abordagens vêm sendo propostas para a anotação genômica, incorporando múltiplas dimensões na caracterização funcional dos genes. O sequenciamento de RNA (RNA-seq), por exemplo, permite não apenas quantificar transcritos, mas também identificar isoformas e padrões de expressão gênica, integrando informações do transcriptoma ao processo de anotação (Depuydt, Rybel, De & Vandepoele, 2023; Reed et al., 2006). A análise integrada de dados genômicos e transcriptômicos permite uma anotação mais precisa e completa dos genes. Nesse contexto, propõe-se uma evolução do conceito de anotação genômica: enquanto a anotação unidimensional se refere à simples identificação de genes e suas funções, a bidimensional considera as interações e componentes celulares; a tridimensional incorpora informações sobre empacotamento e localização celular; e a quarta dimensão refere-se às mudanças evolutivas adaptativas (Ejigu & Jung, 2020; Reed et al., 2006). Embora apenas as duas primeiras dimensões sejam atualmente mais acessíveis, as demais devem ganhar relevância com o avanço das tecnologias. A anotação multidimensional representa o futuro da anotação genômica, permitindo uma compreensão mais completa da função dos genes e de sua regulação. No entanto, é importante reconhecer que a

anotação multidimensional exige a integração de diferentes tipos de dados e a utilização de ferramentas computacionais sofisticadas, o que representa um desafio significativo para a comunidade científica (Ejigu & Jung, 2020).

O aprendizado de máquina (*Machine Learning*) tem se mostrado promissor na integração de dados heterogêneos para melhorar a acurácia da anotação gênica. Algoritmos de aprendizado profundo, como redes neurais convolucionais e redes recorrentes, buscam padrões complexos em grandes volumes de dados para prever a estrutura e função de proteínas, identificar elementos regulatórios e classificar genes (Senior et al., 2020). Apesar dos desafios, como a necessidade de grandes conjuntos de dados de treinamento e a especificidade de modelos para diferentes organismos, o aprendizado de máquina tende a desempenhar um papel cada vez mais relevante à medida que os bancos de dados genômicos se expandem (Ejigu & Jung, 2020; Mahood, Kruse & Moghe, 2020). É fundamental que esses modelos sejam validados experimentalmente e usados com cautela, pois podem propagar vieses presentes nos dados de treinamento, gerando predições enganosas, especialmente em espécies não modelo (Rodrigues, 2023; Whalen, Schreiber & Noble, 2022).

Os algoritmos automatizados de anotação funcional geralmente se baseiam em ortólogos de organismos modelo, os quais nem sempre são os filogeneticamente mais próximos, dificultando a identificação de genes específicos de determinadas linhagens. Por isso, a curadoria manual ainda é indispensável para alcançar modelos gênicos mais precisos. Métodos semiautomáticos, que combinam predições independentes para gerar consensos, têm se mostrado eficazes em unir a velocidade da anotação automatizada com a confiabilidade da curadoria manual (Liu, Ma & Goryanin, 2013). A combinação de métodos automatizados e curadoria manual representa uma abordagem promissora para a anotação genômica. Essa abordagem é particularmente relevante para a anotação de genomas de espécies nativas do Cerrado, como a *M. elliptica*, que apresentam características genéticas únicas e que não são bem representadas nos organismos modelo.

Na era da metagenômica e do *big data*, a anotação genômica continua sendo um desafio crítico, exigindo *pipelines* confiáveis e bases de conhecimento de alta qualidade (Danchin, 2003; Whalen, Schreiber & Noble, 2022). Novas ferramentas, como MEGAnnotator2, têm simplificado a montagem e anotação de genomas microbianos, reduzindo ambiguidades e atendendo às exigências de submissão ao NCBI (Lugli et al., 2023). Abordagens integradas, como a Mantis, combinam informações de múltiplas

fontes por meio de consenso e mineração de texto, alcançando alta cobertura e precisão, ao mesmo tempo em que mantêm flexibilidade e reprodutibilidade (Queirós et al., 2020). Revisões recentes destacam a importância de boas práticas bioinformáticas, sensibilidade, eficiência computacional e reprodutibilidade na anotação de metagenomas, reforçando a necessidade de consolidar pipelines sólidos e confiáveis (Pérez-Llano et al., 2021). Esses avanços tornam possível realizar uma anotação mais precisa e validada experimentalmente em espécies não modelo, como *Mouriri elliptica*, garantindo resultados confiáveis e comparáveis em estudos genômicos modernos (Whalen, Schreiber & Noble, 2022).

2.9 CONSERVAÇÃO DE RECURSOS GENÉTICOS

A conservação de recursos genéticos é um pilar fundamental para a manutenção da biodiversidade e a garantia da sustentabilidade dos ecossistemas (Frankham & Ballou, 2004). A crescente conscientização sobre a importância da diversidade genética para a adaptação das espécies às mudanças ambientais e para a segurança alimentar global tem impulsionado o desenvolvimento contínuo de estratégias e tecnologias para a conservação *in situ* (no habitat natural) e *ex situ* (em bancos de germoplasma, jardins botânicos e coleções) (Hoban et al., 2023). No entanto, a implementação efetiva dessas estratégias requer um profundo conhecimento da estrutura genética das populações e das ameaças que as afetam (Frankham, Briscoe & Ballou, 2003; Hoban et al., 2023).

As tecnologias de Sequenciamento de Nova Geração (NGS) consolidaram-se como ferramentas poderosas, desempenhando um papel crucial na análise detalhada da diversidade genética em populações naturais e cultivadas (Allendorf, Hohenlohe & Luikart, 2010; Andrews et al., 2016). As informações geradas por essas tecnologias são essenciais para a tomada de decisões informadas em relação à conservação, permitindo a identificação de populações geneticamente distintas, a avaliação do impacto da fragmentação do habitat e o desenvolvimento de estratégias de manejo mais eficazes (Funk et al., 2012). Dessa forma, o NGS representa um avanço transformativo na conservação genética, permitindo análises mais rápidas, precisas e abrangentes do que as abordagens tradicionais.

A utilização de marcadores moleculares, como microssatélites (Simple Sequence Repeats - SSR) e polimorfismos de nucleotídeo único (Single Nucleotide Polymorphisms - SNPs), consolidou-se no processo de avaliação da diversidade genética e na identificação de populações geneticamente distintas (Andrews et al., 2016). Esses marcadores, distribuídos ao longo do genoma, permitem a identificação de variações genéticas entre indivíduos e populações, fornecendo informações importantes sobre a estrutura genética, a história evolutiva e o potencial adaptativo das espécies (Freeland, Kirk & Petersen, 2012). A análise desses marcadores pode auxiliar na formulação de estratégias de conservação mais eficazes, permitindo a identificação de áreas prioritárias e a seleção de indivíduos para programas de reprodução *ex situ* (Hoban et al., 2023). Contudo, a escolha dos marcadores moleculares adequados depende das características genéticas da espécie em estudo e dos objetivos específicos da análise (Oliveira et al., 2021).

A análise de dados genômicos em larga escala, obtidos por meio do Sequenciamento do Genoma Total (*Whole-Genome Sequencing* - WGS), tem se tornado mais acessível e possui o potencial de fornecer informações detalhadas sobre a diversidade genética e a adaptação das espécies (Andrews et al., 2016; Ekblom & Galindo, 2010). O WGS permite a identificação de variações genéticas distribuídas em todo o genoma, abrangendo genes, regiões regulatórias e elementos não codificantes, fornecendo uma visão abrangente da arquitetura genética. Essa abordagem tem sido aplicada, por exemplo, na detecção de genes associados à adaptação a diferentes ambientes, na avaliação do impacto de gargalos populacionais e na identificação de áreas com alta diversidade genética a serem priorizadas (Hoban et al., 2023; Ekblom & Galindo, 2010).

Apesar de seu potencial, o WGS ainda apresenta custo elevado e demanda significativa de recursos computacionais para o processamento e análise dos vastos volumes de dados gerados, o que pode limitar sua aplicação em estudos de populações de espécies não modelo. Apesar de seu potencial, o WGS ainda apresenta custo elevado e demanda significativa de recursos computacionais... Nesse contexto, metodologias mais direcionadas, como o SSR-Seq (Sequenciamento de Microssatélites por Amplicons), surgem como alternativas eficientes e de alto rendimento (*high-throughput*). O SSR-Seq combina a multiplexação de *primers* (uso de múltiplos pares em uma única reação para

amplificar diversos *loci* microssatélites) com plataformas de Sequenciamento de Nova Geração (NGS), como a Illumina (Šarhanová et al., 2018).

O SSR-Seq permite a genotipagem robusta de microssatélites a um custo reduzido. Essa abordagem é particularmente útil para estudos de diversidade genética, estrutura populacional e conservação em organismos para os quais ainda não há genomas de referência disponíveis (Šarhanová et al., 2018). Em plantas nativas de biomas ameaçados, como o Cerrado, o SSR-Seq tem se mostrado uma alternativa eficiente para estudos de diversidade genética. Com essa metodologia, é possível identificar populações geneticamente distintas e avaliar os efeitos da fragmentação do habitat, fornecendo subsídios para estratégias de conservação *in situ* e *ex situ* (de Souza, Telles & Diniz-Filho, 2016a). Embora não substitua o WGS, o SSR-Seq apresenta excelente custo-benefício em espécies não modelo, constituindo ferramenta estratégica para apoiar programas de manejo e conservação no Cerrado (Andrews et al., 2016; Hoban et al., 2023).

2.10 REFERÊNCIAS

ALLENDORF, F. W.; HOHENLOHE, P. A.; LUIKART, G. Genomics and the future of conservation genetics. **Nature Reviews Genetics**, v. 11, n. 10, p. 697–709, 2010.

ALTSCHUL, S. F.; GISH, W.; MILLER, W.; MYERS, E. W.; LIPMAN, D. J. Basic local alignment search tool. **Journal of Molecular Biology**, v. 215, n. 3, p. 403–410, 1990.

ANDREWS, K. R.; GOOD, J. M.; MILLER, M. R.; LUIKART, G.; HOHENLOHE, P. A. Harnessing the power of RADseq for ecological and evolutionary genomics. **Nature Reviews Genetics**, v. 17, n. 2, p. 81–92, 2016.

ANDREWS, S. FastQC: a quality control tool for high throughput sequence data. 2010. Disponível em: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

ANTÔNIO, M. et al. Conservação genética de populações de melhoramento: Revisão de literatura sob a ótica das espécies do bioma Cerrado. **LUMEN ET VIRTUS**, v. 15, n. 39, p. 2845–2860, 2024.

BAIROCH, A.; APWEILER, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. **Nucleic Acids Research**, v. 28, n. 1, p. 45–48, 2000.

BENSON, D. A. et al. GenBank. **Nucleic Acids Research**, v. 41, n. Database issue, p. D36–D42, 2013.

BESEMER, J.; LOMSADZE, A.; BORODOVSKY, M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding

sequence motifs in regulatory regions. **Nucleic Acids Research**, v. 29, n. 12, p. 2607–2618, 2001.

BIRNEY, E. et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. **Nature**, v. 447, n. 7146, p. 799–816, 2007.

BLUM, M. et al. The InterPro protein families and domains database: 20 years on. **Nucleic Acids Research**, v. 49, n. D1, p. D344–D354, 2021.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114–2120, 2014.

BORÉM, A.; SANTOS, F. R. **Genética**. 3. ed. Viçosa: Editora UFV, 2017.

BRENT, M. R. Genome annotation past, present, and future: how to define an ORF at each locus. **Genome Research**, v. 15, n. 12, p. 1777–1786, 2005.

BRUNET, M. A. et al. OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. **Nucleic Acids Research**, v. 47, n. D1, p. D403–D410, 2019.

BUERMANS, H. P. J.; DEN DUNNEN, J. T. Next generation sequencing technology: advances and applications. **Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease**, v. 1842, n. 10, p. 1932–1941, 2014.

CANTAREL, B. L. et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. **Genome Research**, v. 18, n. 1, p. 188–196, 2008.

CBD. **Convention on biological diversity: text and annexes**. Montreal: Secretariat of the Convention on Biological Diversity, 1992.

DANCHIN, A. Genomes and evolution. **Current Issues in Molecular Biology**, v. 5, n. 4, p. 133-144, 2003.

DEPUYDT, T.; DE RYBEL, B.; VANDEPOELE, K. Charting plant gene functions in the multi-omics and single-cell era. **Trends in Plant Science**, v. 28, n. 10, p. 1091-1099, 2023.

DESCHAMPS, S.; LLACA, V. Strategies for sequence assembly of plant genomes. In: EDWARDS, D. (Ed.). **Plant bioinformatics: methods and protocols**. New York: Humana Press, 2016. p. 1-279.

EILBECK, K.; MOORE, B.; HOLT, C.; YANDELL, M. Quantitative measures for the management and comparison of annotated genomes. **BMC Bioinformatics**, v. 10, n. 1, p. 67, 2009.

EJIGU, G. F.; JUNG, J. Review on the computational genome annotation of sequences obtained by next-generation sequencing. **Biology**, v. 9, n. 9, p. 295, 2020.

EKBLOM, R.; GALINDO, J. Applications of next generation sequencing in molecular ecology of non-model organisms. **Heredity**, v. 107, n. 1, p. 1–15, 2011.

FLEISCHMANN, R. D. et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. **Science**, v. 269, n. 5223, p. 496–512, 1995.

FRANKHAM, R.; BALLOU, J. D.; BRISCOE, D. A. **Introduction to conservation genetics**. Cambridge: Cambridge University Press, 2002.

FREELAND, J. R.; KIRK, H.; PETERSEN, S. **Molecular ecology**. 2nd ed. Chichester: Wiley-Blackwell, 2011.

FUNK, W. C.; MCKAY, J. K.; HOHENLOHE, P. A.; ALLENDORF, F. W. Harnessing genomics for delineating conservation units. **Trends in Ecology & Evolution**, v. 27, n. 9, p. 489–496, 2012.

GABALDÓN, T.; KOONIN, E. V. Functional and evolutionary implications of gene orthology. **Nature Reviews Genetics**, v. 14, n. 5, p. 360–366, 2013.

GIANI, A. M. et al. Long walk to genomics: history and current approaches to genome sequencing and assembly. **Computational and Structural Biotechnology Journal**, v. 18, p. 9–19, 2020.

GOODWIN, S.; MCPHERSON, J. D.; MCCOMBIE, W. R. Coming of age: ten years of next-generation sequencing technologies. **Nature Reviews Genetics**, v. 17, n. 6, p. 333–351, 2016.

GUREVICH, A.; SAVELIEV, V.; VYSHI, N.; TESLER, G. QUAST: quality assessment tool for genome assemblies. **Bioinformatics**, v. 29, n. 8, p. 1072–1075, 2013.

HARRISON, P. M. et al. Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. **Nucleic Acids Research**, v. 33, n. 8, p. 2374–2383, 2005.

HOBAN, S. et al. Genetic diversity goals and targets have improved, but remain insufficient for clear implementation of the post-2020 global biodiversity framework. **Conservation Genetics**, v. 24, n. 2, p. 181–191, 2023.

HUNTER, S. et al. InterPro: the integrative protein signature database. **Nucleic Acids Research**, v. 37, n. suppl_1, p. D211–D215, 2009.

Illumina. **Sequencing and array-based solutions**. Disponível em: https://www.illumina.com. Acesso em: 24 maio 2025.

INTERPRO. **InterPro documentation**. Disponível em: https://interpro-documentation.readthedocs.io/. Acesso em: 12 ago. 2025.

IQUEBAL, M. A. et al. Applications of bioinformatics in plant and agriculture. In: BAHADUR, B. et al. (Ed.). **Plant omics: the omics of plant science**. New Delhi: Springer, 2015. p. 755–789.

IŠTVÁNEK, J. et al. Genome assembly and annotation for red clover (Trifolium pratense; Fabaceae). **American Journal of Botany**, v. 101, n. 2, p. 327–337, 2014.

JAIN, M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. **Nature Biotechnology**, v. 36, n. 4, p. 338–345, 2018.

JOSÉ BORGES DE SOUZA, U.; PIRES DE CAMPOS TELLES, M.; ALEXANDRE FELIZOLA DINIZ-FILHO, J. Tendências da literatura científica sobre genética de populações de plantas do Cerrado. **Hoehnea**, v. 43, n. 3, p. 461–477, 2016.

KARSCH-MIZRACHI, I. et al. The international nucleotide sequence database collaboration. **Nucleic Acids Research**, v. 46, n. D1, p. D48–D51, 2018.

LI, H. D.; OMENN, G. S.; GUAN, Y. A proteogenomic approach to understand splice isoform functions through sequence and expression-based computational modeling. **Briefings in Bioinformatics**, v. 17, n. 6, p. 1024–1031, 2016.

LIU, Z.; MA, H.; GORYANIN, I. A semi-automated genome annotation comparison and integration scheme. **BMC Bioinformatics**, v. 14, n. Suppl 2, p. S6, 2013.

LUGLI, G. A. *et al.* MEGAnnotator2: a pipeline for the assembly and annotation of microbial genomes. **Microbiome Res Rep 2023;2:15.**, v. 2, n. 2, p. N/A-N/A, 30 abr. 2023.

LUO, R. et al. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. **GigaScience**, v. 1, n. 1, p. 18, 2012.

LUSCOMBE, N. M. et al. Genomic analysis of regulatory network dynamics reveals large topological changes. **Nature**, v. 431, n. 7006, p. 308–312, 2004.

MAHOOD, E. H.; KRUSE, L. H.; MOGHE, G. D. Machine learning: a powerful tool for gene function prediction in plants. **Applications in Plant Sciences**, v. 8, n. 7, p. e11376, 2020.

MARDIS, E. R. Next-generation DNA sequencing methods. **Annual Review of Genomics and Human Genetics**, v. 9, p. 387–402, 2008.

MARGULIES, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. **Nature**, v. 437, n. 7057, p. 376–380, 2005.

MATHÉ, C. et al. Current methods of gene prediction, their strengths and weaknesses. **Nucleic Acids Research**, v. 30, n. 19, p. 4103–4117, 2002.

METZKER, M. L. Sequencing technologies - the next generation. **Nature Reviews Genetics**, v. 11, n. 1, p. 31–46, 2010.

MILLER, J. R.; KOREN, S.; SUTTON, G. Assembly algorithms for next-generation sequencing data. **Genomics**, v. 95, n. 6, p. 315–327, 2010.

MOUNT, D. W. **Bioinformatics: sequence and genome analysis**. 2nd ed. Cold Spring Harbor: Cold Spring Harbor Laboratory Press, 2004.

NAGARAJAN, N.; POP, M. Sequence assembly demystified. **Nature Reviews Genetics**, v. 14, n. 3, p. 157–167, 2013.

OHYAMA, K. et al. Chloroplast gene organization deduced from complete sequence of liverwort Marchantia polymorpha chloroplast DNA. **Nature**, v. 322, n. 6079, p. 572–574, 1986.

O'LEARY, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. **Nucleic Acids Research**, v. 44, n. D1, p. D733–D745, 2016.

OLIVEIRA, A. J. et al. Principais marcadores moleculares. **Research, Society and Development**, v. 10, n. 15, p. e562101523633, 2021.

PÉREZ-LLANO, Y. *et al.* Metagenomic Tools for Taxonomic and Functional Annotation. **Metagenomics and Microbial Ecology**, p. 21–44, 15 nov. 2021.

POP, M.; PHILLIPPY, A.; DELCHER, A. L.; SALZBERG, S. L. Comparative genome assembly. **Briefings in Bioinformatics**, v. 5, n. 3, p. 237–248, 2004.

QUEIRÓS, P. *et al.* Mantis: flexible and consensus-driven genome annotation, 3 nov. 2020. Disponível em: <http://biorxiv.org/lookup/doi/10.1101/2020.11.02.360933>

REED, J. L. et al. Systems approach to refining genome annotation. **Proceedings of the National Academy of Sciences**, v. 103, n. 46, p. 17480–17484, 2006.

RODRIGUES, D. L. N. Desafios na padronização da anotação genômica. **BIOINFO**, v. 3, n. 1, p. 05-10, 2023.

RONAGHI, M.; UHLÉN, M.; NYRÉN, P. A sequencing method based on real-time pyrophosphate. **Science**, v. 281, n. 5375, p. 363–365, 1998.

SAFONOVA, Y.; BANKEVICH, A.; PEVZNER, P. A. dipSPAdes: assembler for highly polymorphic diploid genomes. **Journal of Computational Biology**, v. 22, n. 6, p. 528–545, 2015.

SAHA, A.; BATTLE, A. False positives in trans-eQTL and co-expression analyses arising from RNA-sequencing alignment errors. **F1000Research**, v. 7, p. 1860, 2018.

SANGER, F.; NICKLEN, S.; COULSON, A. R. DNA sequencing with chain-terminating inhibitors. **Proceedings of the National Academy of Sciences**, v. 74, n. 12, p. 5463–5467, 1977.

ŠARHANOVÁ, P. et al. SSR-seq: Genotyping of microsatellites using next-generation sequencing reveals higher level of polymorphism as compared to traditional fragment size scoring. **Ecology and Evolution**, v. 8, n. 22, p. 10817–10833, 2018.

SAYERS, E. W. et al. Database resources of the National Center for Biotechnology Information. **Nucleic Acids Research**, v. 50, n. D1, p. D20–D26, 2022.

SCHATZ, M. C.; DELCHER, A. L.; SALZBERG, S. L. Assembly of large genomes using second-generation sequencing. **Genome Research**, v. 20, n. 9, p. 1165–1173, 2010.

SCHNOES, A. M. et al. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. **PLoS Computational Biology**, v. 5, n. 12, p. e1000605, 2009.

SCHOCH, C. L. et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. **Database**, v. 2020, p. baaa062, 2020.

SEEMANN, T. Prokka: rapid prokaryotic genome annotation. **Bioinformatics**, v. 30, n. 14, p. 2068–2069, 2014.

SENIOR, A. W. et al. Improved protein structure prediction using potentials from deep learning. **Nature**, v. 577, n. 7792, p. 706–710, 2020.

SHINOZAKI, K. et al. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. **The EMBO Journal**, v. 5, n. 9, p. 2043–2049, 1986.

SIMÃO, F. A. et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. **Bioinformatics**, v. 31, n. 19, p. 3210–3212, 2015.

SIMPSON, A. J. G. et al. The genome sequence of the plant pathogen Xylella fastidiosa. **Nature**, v. 406, n. 6792, p. 151–159, 2000.

SIMPSON, J. T. et al. ABySS: a parallel assembler for short read sequence data. **Genome Research**, v. 19, n. 6, p. 1117–1123, 2009.

STADEN, R. A new computer method for the storage and manipulation of DNA gel reading data. **Nucleic Acids Research**, v. 8, n. 16, p. 3673–3694, 1980.

STADTLÄNDER, C. T. K.-H. **Next-generation sequencing data analysis**. New York: CRC Press, 2017.

STANKE, M.; MORGENSTERN, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. **Nucleic Acids Research**, v. 33, n. suppl_2, p. W465–W467, 2005.

STEIN, L. Genome annotation: from sequence to biology. **Nature Reviews Genetics**, v. 2, n. 7, p. 493–503, 2001.

The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. **Nucleic Acids Research**, v. 49, n. D1, p. D480–D489, 2021.

WATSON, J. D.; CRICK, F. H. C. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. **Nature**, v. 171, n. 4356, p. 737–738, 1953.

WHALEN, S.; SCHREIBER, J.; NOBLE, W. S. Navigating the pitfalls of applying machine learning in genomics. **Nature Reviews Genetics**, v. 23, n. 3, p. 169-181, 2022.

YANDELL, M.; ENCE, D. A beginner's guide to eukaryotic genome annotation. **Nature Reviews Genetics**, v. 13, n. 5, p. 329–342, 2012.

YE, J.; MCGINNIS, S.; MADDEN, T. L. BLAST: improvements for better sequence analysis. **Nucleic Acids Research**, v. 34, n. Web Server issue, p. W6–W9, 2006.

ZIMIN, A. V. et al. The MaSuRCA genome assembler. **Bioinformatics**, v. 29, n. 21, p. 2669–2677, 2013.

CAPÍTULO 2

GENOMIC ANALYSIS OF THE CHLOROPLAST OF *MOURIRI ELLIPTICA* MARTIUS (MELASTOMATACEAE)

Juliana Borges Pereira Brito[1,2]; Adriana Maria Antunes[1,2]; Lázaro José Chaves[1]; Mariana Pires de Campos Telles[2]; Wagner Nunes Ribeiro[1,2]; Thainara Policarpo Mendes[1,2]; Thannya Nascimento Soares[1,2]

Capítulo elaborado para ser enviado ao periódico científico Funcional & Integrative Genomics

[1]Graduate Program in Genetics and Plant Breeding, Federal University of Goiás, Goiânia, Brazil.
[2]Laboratory of Genetics and Biodiversity, Federal University of Goiás, Goiânia, Brazil.
*Corresponding authors: juliana.freitas@seduc.go.gov.br and tnsoares@ufg.br

3 GENOMIC ANALYSIS OF THE CHLOROPLAST OF *MOURIRI ELLIPTICA* MARTIUS (MELASTOMATACEAE)

Juliana Borges Pereira Brito[1,2]; Adriana Maria Antunes[1,2]; Lázaro José Chaves[1]; Mariana Pires de Campos Telles[2]; Wagner Nunes Ribeiro[1,2]; Thainara Policarpo Mendes[1,2]; Thannya Nascimento Soares[1,2]

[1]Graduate Program in Genetics and Plant Breeding, Federal University of Goiás, Goiânia, Brazil.
[2]Laboratory of Genetics and Biodiversity, Federal University of Goiás, Goiânia, Brazil.
*Corresponding authors: juliana.freitas@seduc.go.gov.br and tnsoares@ufg.br

ABSTRACT

Mouriri elliptica Martius (Melastomataceae) is notable for its medicinal and ornamental properties. In this study, the chloroplast genome of *M. elliptica* was sequenced and analyzed to investigate its structural features and phylogenetic relationships, as well as to develop primers for the *MatK* and *rbcL* genes with potential application in DNA barcoding. Young leaves of *M. elliptica* were collected, followed by DNA extraction and sequencing using the Illumina MiSeq platform. The assembled chloroplast genome is 156,791 base pairs in length and exhibits the typical quadripartite structure found in angiosperms, with LSC, SSC, and IR regions measuring 86,943 bp, 17,234 bp, and 26,307 bp, respectively. A total of 105 single-copy genes essential for photosynthesis and 20 duplicated genes in the IR regions were identified. Analysis of inverted repeat (IR) regions revealed variations in IR length and gene positioning. Phylogenetic reconstruction placed *M. elliptica* in close relationship with the genus *Memecylon*. The primers developed for the *MatK* and *rbcL* genes proved to be effective in *in silico* analyses for plant barcoding, highlighting their potential in species identification. These findings emphasize the genetic diversity and evolutionary patterns of chloroplast genomes within Melastomataceae, proving valuable insights for taxonomic delimitation strategies.

Keywords: chloroplast genome, Cerrado biome, genetic conservation, DNA barcoding, Melastomataceae

4 INTRODUCTION

The field of plant genomics has entered a new era characterized by the routine generation of chromosome-scale, gap-free genome assemblies. This transition, driven by

long-read sequencing technologies and advanced bioinformatic tools, has shifted the focus from obtaining a single reference genome to constructing *pan-genomes* that capture the full spectrum of genetic diversity within species (Song et al., 2019).

Within this genomic revolution, organellar genomes, particularly chloroplast genomes, remain invaluable molecular resources. Due to their high copy number, conserved structure, uniparental inheritance, and moderate evolutionary rate, chloroplast genomes provide robust markers for evolutionary, taxonomic, and ecological studies (Ruhlman & Jansen, 2014). Moreover, they have become fundamental in the development of molecular tools such as DNA barcodes, which facilitate accurate species identification and biodiversity monitoring (Hollingsworth et al., 2011).

These genomic advancements have enabled the comprehensive analysis of diverse plant taxa, including several medicinal and timber species native to the Cerrado biome, such as *Pterodon emarginatus* Vogel and *Pterodon pubescens* Benth (Brito et al., 2023), *Dipteryx alata* Vogel (Antunes et al., 2020), and *Stryphnodendron adstringens* Mart. (Souza et al., 2019). Within the Melastomataceae family, chloroplast genomes have been sequenced for various genera, including *Allomaieta*, *Allomorphia*, and *Anerincleistus*. The genus *Phyllagathis* Blume is particularly well represented, with approximately 40 complete chloroplast genomes available (Weng et al., 2021), reflecting a growing dataset that has advanced phylogenetic and evolutionary studies in the family.

According to the Brazilian Biodiversity Information System (SiBBr), *Mouriri elliptica* Mart. is an endemic species of the Cerrado, occurring in the states of Tocantins, Bahia, Maranhão, Piauí, Goiás, Mato Grosso, Mato Grosso do Sul, and Minas Gerais (Völtz & Goldenberg, 2020). *M. elliptica* is a small tree recognized for its sweet, traditionally consumed fruits and its medicinal properties, particularly in the treatment of kidney disorders. Its leaves are used in traditional medicine for treating ulcers, and the species is recommended for ornamental purposes and cultivation in home orchards (Assis et al., 2016; Moleiro et al., 2009). Additionally, it serves as a vital nectar source for honey production in apiculture (Pott & Pott, 1994).

This study aims to sequence, assemble, and annotate the complete chloroplast genome of *M. elliptica*, thereby addressing the current genomic void for the genus *Mouriri*. By doing so, our work will contribute to a broader understanding of chloroplast genome evolution and diversity within the Melastomataceae. Furthermore,

the characterized genome will provide a foundation for developing robust molecular tools, such as species-specific primers for DNA barcoding, with direct applications in accurate species identification, informing genetic conservation strategies, and supporting comprehensive biodiversity assessments in the threatened Cerrado biome.

4.2 MATERIALS AND METHODS

4.2.1 Plant Sampling, DNA Extraction, and Sequencing

Leaf material from a mature individual of *Mouriri elliptica* was collected in 2023 from an orchard planted in 1992 at the School of Agronomy of the Federal University of Goiás (UFG), Goiânia, Goiás, Brazil (latitude: –16.5991637; longitude: –49.2792997). The specimen (voucher UFG 51952) was deposited in the UFG Herbarium. Young, healthy leaves were selected for DNA extraction following the cetyltrimethylammonium bromide (CTAB) method described by Doyle and Doyle (1987). DNA quality and integrity were evaluated on a 1% agarose gel, and concentration was determined using a Qubit® fluorometer (Invitrogen). The genomic library was prepared using the Nextera DNA Flex kit (Illumina) according to the manufacturer's protocol. Library size distribution and quality were verified on an Agilent 2100 Bioanalyzer and quantified with Qubit prior to denaturation. Sequencing was performed on the Illumina MiSeq platform using the MiSeq Reagent Kit v3 (600 cycles; paired-end reads, 2 × 300 bp).

4.2.2 Sequence Quality Assessment and Genome Assembly

Raw read quality was evaluated using FastQC (v0.11.9), which generated detailed quality metrics. Adapter trimming and low-quality base removal were conducted with Trimmomatic (v0.39), applying a minimum Phred score of 20 and a minimum read length of 36 bp (Bolger et al., 2014). *De novo* assembly of the chloroplast genome was performed using GetOrganelle (v1.7.7.0) with default parameters optimized for organellar genomes (Jin et al., 2020).

4.2.3 Gene Annotation and Microsatellite Region Identification

Coding sequences (CDSs), ribosomal RNAs (rRNAs), and transfer RNAs (tRNAs) were annotated using GeSeq and manually curated to ensure accuracy (Tillich et al., 2017). The presence of tRNAs and rRNAs was validated using Aragorn, tRNAscan-SE, and RNAmmer (Lagesen et al., 2007; Lowe & Eddy, 1997). Circular genome maps were generated using OGDRAW (Lohse, Drechsel & Bock, 2007).

Chloroplast microsatellites (SSRs) were identified using MISA (v2.1) (Beier et al., 2017), which detects and categorizes SSRs based on predefined parameters. The following thresholds were used: mononucleotide repeats with ≥10 bases (10 repeats), dinucleotides with ≥10 bases (5 repeats), trinucleotides and tetranucleotides with ≥12 bases (4 and 3 repeats, respectively), pentanucleotides with ≥15 bases (3 repeats), and hexanucleotides with ≥18 bases (3 repeats). These criteria are commonly used in chloroplast genome studies to investigate genetic variation. Long repeat sequences (forward, reverse, complementary, and palindromic) were detected using REPuter, which reports their length and genomic position, contributing to insights on genome organization and stability.

4.2.4 Comparative Chloroplast Genome Analyses

Comparative analyses included the chloroplast genome of *M. elliptica* and 42 additional Melastomataceae species retrieved from GenBank. Nucleotide diversity was assessed using a sliding window approach in DNAsp (v6.12.03), with a window size of 600 bp and a step size of 25 bp. Gene sequences were extracted using Geneious, aligned with MAFFT (v7.505), and refined using Gblocks (v0.91b) to remove poorly aligned regions (Rozas et al., 2017). This process yielded 53 genes that were common to all analyzed genomes.

Concatenated gene sequences were analyzed in MEGA X (Kumar et al., 2018) to construct a phylogenetic tree using the Maximum Likelihood method. The Tamura-Nei model was selected based on model selection tests in MEGA, as it provided the best fit to the data according to statistical criteria. Node support was assessed with

1,000 bootstrap replicates, and the resulting tree was edited in FigTree (v1.4.4) Rambaut, 2018).

To facilitate phylogenetic interpretation, Table 3.1 lists 42 species from several tribes within Melastomataceae and closely related families. Species were selected based on their representation within the family and the availability of complete chloroplast genomes in GenBank. Two species from the tribe Myrteae (*Psidium guajava* [KY635879.1] and *Eugenia uniflora* [KY392761.1]) were included as outgroups. This selection is supported by prior phylogenetic evidence positioning Myrtaceae as a sister group to Melastomataceae within the order Myrtales. The evolutionary proximity of these groups provides a robust root for phylogenetic inference, enhancing the reliability of the resulting tree topology. This organization also clarifies the dataset used and underscores the relevance of each sequence for comparative and phylogenetic analyses

Table 3.1 *Melastomataceae* species used in the phylogenetic analysis, including their respective subfamilies and tribes, and the corresponding chloroplast gene sequence accessions from GenBank. The classification adopted here follows the most recent taxonomic framework proposed by Penneys (2022).

Species Name	Subfamily / Tribe	GenBank Accession(s)
Melastoma malabathricum	Melastomatoideae / Melastomateae	OQ595235.1
Melastoma candidum	Melastomatoideae / Melastomateae	NC_034716.1
Melastoma dodecandrum	Melastomatoideae / Melastomateae	NC_042821.1
Heterotis rotundifolia	Melastomatoideae / Melastomateae	NC_050999.1
Osbeckia stellata	Melastomatoideae/ Melastomateae	NC_046486.1
Pterogastra divaricata	Melastomatoideae/ Melastomateae	NC_031885.1
Tibouchina semidecandra	Melastomatoideae / Melastomateae	NC_053325.1
Tibouchina urvilleana	Melastomatoideae / Melastomateae	NC_043810.1
Heterocentron elegans	Melastomatoideae / Melastomateae	NC_051000.1
Tibouchina longifólia	Melastomatoideae / Melastomateae	NC_031889.1
Nepsera aquática	Melastomatoideae / Marcetieae	NC_031883.1
Rhexia virginica	Melastomatoideae / Rhexieae	NC_031886.1
Microlicia cogniauxiana	Melastomatoideae / Lavoisiereae	NC_043792.1
Rhynchanthera bracteata	Melastomatoideae / Lavoisiereae	NC_031887.1

Merianthera pulchra	Melastomatoideae / Pyramieae	NC_031881.1
Triolena amazônica	Melastomatoideae / Trioleneae	NC_031890.1
Blakea schlimii	Melastomatoideae / Pyxidantheae	NC_031877.1
Dalenia sarawakensis	Melastomatoideae / Dissochaeteae	OL813723.1
Opisthocentra clidemioides	Melastomatoideae / Sonerileae	NC_031884.1
Phyllagathis elliptica	Melastomatoideae / Sonerileae	NC_068148.1
Anerincleistus quintuplinervis	Melastomatoideae / Sonerileae	OL813724.1
Anerincleistus sertulifer	Melastomatoideae / Sonerileae	MK994888.1
Barthea barthei	Melastomatoideae / Sonerileae	NC_035661.1
Sonerila cantonensis	Melastomatoideae / Sonerileae	NC_068144.1
Sonerila nervulosa	Melastomatoideae / Sonerileae	OL813716.1
Tigridiopalma longmenensis	Melastomatoideae / Sonerileae	NC_058962.1
Tigridiopalma magnifica	Melastomatoideae / Sonerileae	NC_036021.1
Plagiopetalum esquirolii	Melastomatoideae / Sonerileae	NC_068152.1
Sonerila parviflora	Melastomatoideae / Sonerileae	MK994900.1
Sonerila borneensis	Melastomatoideae / Sonerileae	MK994893.1
Sonerila velutina	Melastomatoideae / Sonerileae	MK994892.1
Allomaieta villosa	Melastomatoideae / Cyphostyleae	NC_031875.1
Bertolonia acuminata	Melastomatoideae / Bertolonieae	NC_031876.1
Miconia dodecandra	Melastomatoideae / Miconieae	NC_031882.1
Eriocnema fulva	Melastomatoideae / Eriocnemea	NC_031878.1
Graffenrieda moritziana	Melastomatoideae / Merianieae	NC_031879.1
Salpinga maranonensis	Melastomatoideae / Merianieae	NC_031888.1
Henriettea barkeri	Melastomatoideae / Henrietteeae	NC_031880.1
Memecylon pauciflorum	Olisbeoideae / -	NC_043809.1
Memecylon ligustrifolium	Olisbeoideae / -	MK994913.1
Oxyspora paniculata	Melastomatoideae / Sonerileae	OL813721.1
Mouriri emarginata	Olisbeoideae / -	SRA ERR5034830

4.2.5 Genome Characterization

Expansion and contraction of the chloroplast inverted repeat (IR) regions were examined using IRscope, a tool for visualizing and comparing IR boundaries. For comparative IR analyses, a subset of 20 complete chloroplast genomes was selected from the original dataset, prioritizing species phylogenetically related to *M. elliptica* (subfamily Olisbeoideae). This selection aimed to explore structural patterns and evolutionary relationships. Genetic diversity was also assessed using DNAsp (v6.12.03), estimating nucleotide diversity (π), haplotype diversity, and theta (θ) values (Rozas et al., 2017).

4.2.6 Primer Design for DNA Barcoding

The design of primers for the *MatK* and *rbcL* chloroplast genes was guided by a combined analysis of nucleotide diversity and evolutionary conservation. Initially, nucleotide diversity (π) was calculated across the entire chloroplast genome using DNAsp (v6.12.03) to identify hypervariable regions suitable for species discrimination. Regions exhibiting intermediate to high variability ($\pi > 0.02$) were prioritized for primer design, as they offer optimal polymorphism for distinguishing closely related species while maintaining reliable amplification efficiency.

Specifically, for the *MatK* gene, primers were designed to amplify two variable domains: the first spanning positions 79 to 203 and the second from 299 to 498 of the reference sequence. For the *rbcL* gene, three informative regions were targeted: positions 69–226, 103–239, and 116–253. This strategy leverages the higher evolutionary rate of *MatK* for fine-scale discrimination and the conserved nature of *rbcL* for broader taxonomic comparisons, while ensuring the amplified fragments are within the optimal size range for sequencing.

Subsequently, *in silico* validation of the designed primers was performed using the R package openPrimeR (Hildebrand et al., 2017). This analysis assessed critical quality parameters, including melting temperature (Tm ≈ 50°C), GC content (40–60%), absence of secondary structures (hairpins, self-dimers, and cross-dimers), and target coverage specificity. Primers failing to meet these thresholds were discarded. Finally, all

selected primers were optimized for compatibility with Illumina short-read sequencing platforms, ensuring amplicon sizes (125–200 bp) suitable for generating overlapping paired-end reads.

4.3 RESULTS

4.3.1 Chloroplast Genome Structure

In this study, we conducted a comprehensive analysis of the chloroplast genome of *Mouriri elliptica*. The genome has a total length of 156,791 base pairs (bp) and exhibits the typical circular structure of chloroplast genomes. It is organized into four main regions: the Large Single Copy (LSC) region of 86,943 bp, the Small Single Copy (SSC) region with 17,234 bp, and two Inverted Repeat regions (IRa and IRb), each comprising 26,307 bp (Figure 3.1).

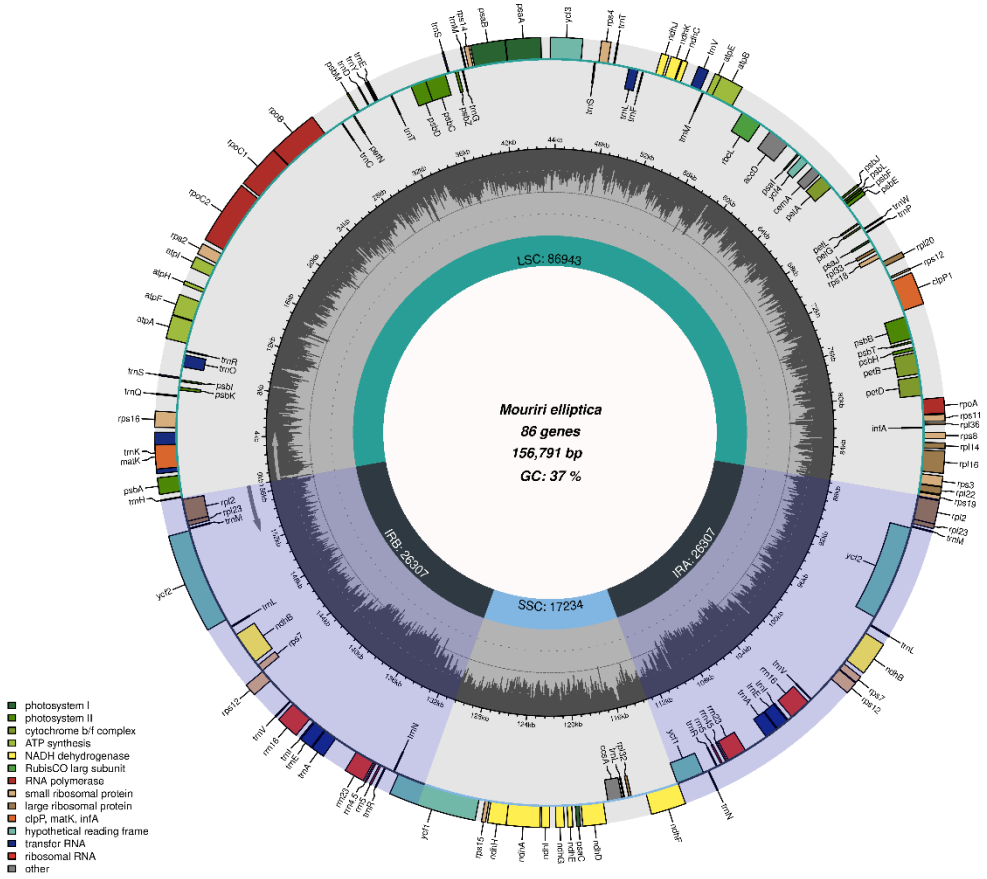

Figure 3.1 Circular map of the *Mouriri elliptica* chloroplast genome. Genes shown on the outside of the circle are transcribed clockwise, while those on the inside are transcribed counterclockwise. The map highlights the LSC, SSC, and IR regions. The inner gray ring represents the GC content, and the colored boxes denote gene functional categories.

4.3.2 Genome Annotation

Genome annotation revealed a variety of essential genes, including 79 protein-coding genes, 4 ribosomal RNA (rRNA) genes, and 30 transfer RNA (tRNA) genes. Among the protein-coding genes, those associated with photosynthesis are prominent, including 14 genes for *photosystem II*, 5 *photosystem I*, and 6 *ATP synthase* genes. Additionally, 21 genes involved in ribosomal protein synthesis and 11 *NADH dehydrogenase* genes were identified (Table 3.2).

Table 3.2 Annotated genes in the chloroplast genome of *Mouriri elliptica*. The table lists gene categories, including subunits of ATP synthase, NADH dehydrogenase, cytochrome b/f complex, photosystems I and II, and other functional genes.

Gene Group	Genes
ATP synthase subunits	*atpA, atpB, atpE, atpF, atpH, atpI*
NADH dehydrogenase subunits	*ndhA, ndhB*, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*
Cytochrome b/f complex subunits	*petA, petB, petD, petG, petL, petN*
Photosystem I subunits	*psaA, psaB, psaC, psaI, psaJ*
Photosystem II subunits	*psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbT, psbZ*
Large ribosomal subunit	*rpl2*, rpl14, rpl16, rpl20, rpl22, rpl23*, rpl32, rpl33, rpl36*
Small ribosomal subunit	*rps2, rps3, rps4, rps7*, rps8, rps11, rps12*, rps14, rps15, rps16, rps18, rps19**
DNA-dependent RNA polymerase	*rpoA, rpoB, rpoC1, rpoC2*

Rubisco large subunit	*RbcL*
Cytochrome c-type synthesis	*CcsA*
Envelope membrane protein	*CemA*
Maturase	*MatK*
Protease	*clpP1*
Acetyl-CoA carboxylase subunit	*AccD*
Translation initiation factor	*InfA*
Conserved ORFs	*ycf1, ycf2*, ycf3, ycf4*
Ribosomal RNA genes	*rrn16*, rrn23*, rrn4.5*, rrn5**
Other tRNA genes	*trnH-GUG, trnK-UUU, trnQ-UUG, trnS-GCU, trnO-CUA, trnR-UCU, trnC-GCA, trnD-GUC, trnY-GUA, trnE-UUC*, trnT-GGU, trnS-UGA, trnG-GCC, trnM-CAU*, trnS-GGA, trnT-UGU, trnL-UAA*, trnF-GAA, trnV-UAC*, trnW-CCA, trnP-UGG, trnV-GAC, trnL-CAA, trnI-GAU*, trnA-UGC*, trnR-ACG*, trnN-GUU**

*Genes marked with an asterisk (**) are duplicated in the inverted repeat (IR) regions of the chloroplast genome.

4.3.3 Intron Analysis

Intron analysis revealed that *ndhA* contains the longest intron (1,050 bp), while *rpl2* and *ndhB* have the shortest (676 bp each). In total, nine genes contain introns, with *ycf3* and *clpP1* each harboring two introns, suggesting increased regulatory complexity. The *rpl2* and *ndhB* genes are duplicated due to their location within the inverted repeat (IR) regions (Figure 3.2).

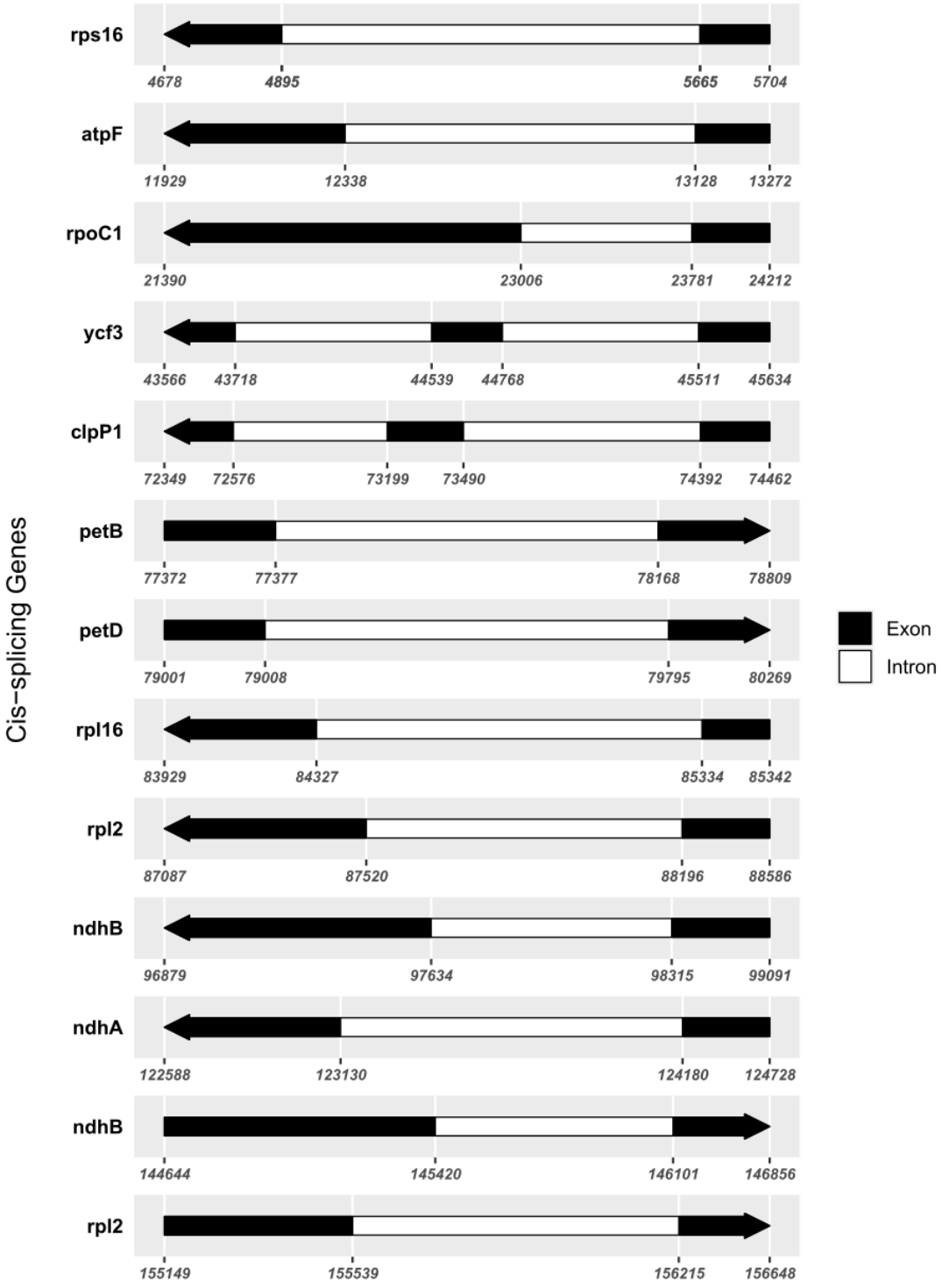

Figure 3.2 Schematic representation of intron-containing genes in the *Mouriri elliptica* plastid genome. Arrows indicate transcription direction. Exons are shown as black boxes and introns as white boxes. Numbers below indicate nucleotide positions.

4.3.4 IR Region Variation

Comparative analysis of IR regions across Melastomataceae chloroplast genomes showed variation in both length and gene positioning. In *M. elliptica*, IRs span 26,307 bp, while in *Microlicia cogniauxiana* they reach 26,667 bp. These differences reflect the dynamics of IR expansion and contraction. Chloroplast genome size varies from 153,311 bp in *Salpinga maranonensis* to 157,216 bp in *Miconia dodecandra*. Gene position shifts, especially of *rps19* and *ndhF*, highlight structural plasticity and underscore the evolutionary significance of junctions JLB, JSB, JSA, and JLA, which correspond, respectively, to the boundaries between the Large Single Copy (LSC) and IRb (JLB), IRb and Small Single Copy (SSC) (JSB), SSC and IRa (JSA), and IRa and LSC (JLA) regions of the chloroplast genome (Figure 3.3).

Inverted Repeats

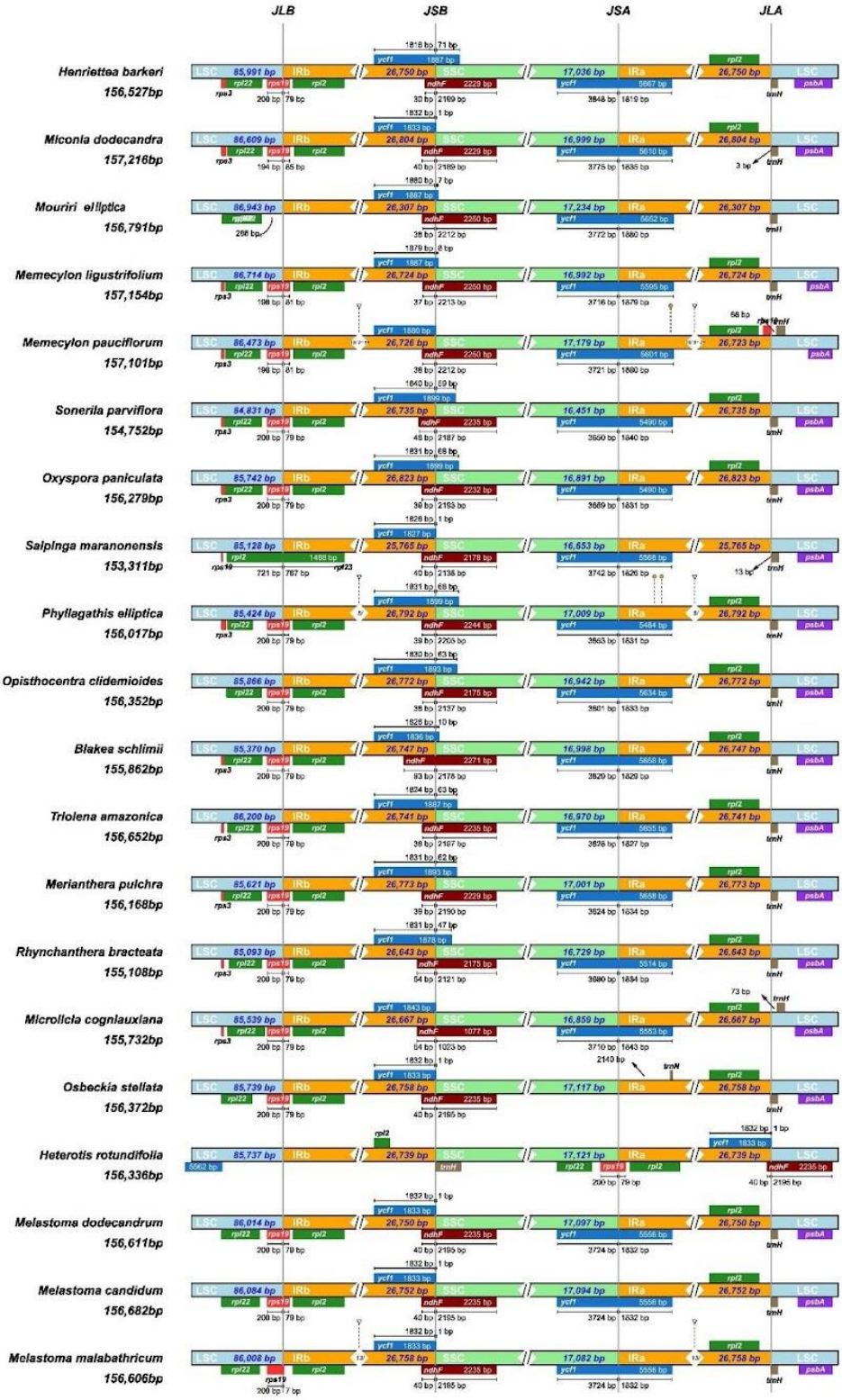

Figure 3.3 Comparison of IR boundaries in Melastomataceae chloroplast genomes, including *Mouriri elliptica*. Arrows represent gene direction; numbers show region lengths (bp). IR junction variation across species is illustrated at JLB, JSB, JSA, and JLA.

4.3.5 Tandem Repeats and SSRs

The nucleotide diversity (π) across the chloroplast genome of *Mouriri elliptica* was first assessed using a sliding window analysis to detect intra-genomic variation. The π values ranged from 0.00054 to 0.09508, revealing the coexistence of highly conserved and hypervariable regions. Peaks of elevated nucleotide diversity were primarily located in intergenic spacers and within the *ycf1* gene.

To further evaluate the phylogenetic informativeness of these hypervariable regions, a comparative analysis was performed across 42 *Melastomataceae* chloroplast genomes. Nucleotide diversity calculated from multiple sequence alignments of the core DNA barcode genes showed contrasting patterns: *rbcL* was highly conserved (π = 0.02070), whereas *MatK* exhibited substantially higher variability (π = 0.05461). The distribution of these hypervariable regions was consistent with patterns reported for other angiosperm chloroplast genomes (Figure 3.4).

4.3.6 Nucleotide Diversity

The nucleotide diversity (Pi) across the assembled chloroplast genome of *Mouriri elliptica* was calculated using a sliding window analysis. The Pi values exhibited considerable variation, ranging from 0.00054 to 0.09508, indicating the presence of both highly conserved and hypervariable regions. Distinct peaks of high nucleotide diversity were identified in specific regions, particularly in intergenic spacers and the *ycf1* gene. To validate the utility of highly variable regions for molecular markers, the core DNA barcode genes were analyzed across 42 Melastomataceae species. The nucleotide diversity (π) calculated from multiple sequence alignments revealed a strong disparity: *rbcL* was highly conserved (π = 0.02070), while *MatK* was significantly more variable (π = 0.05461). The location of the hypervariable regions is consistent with patterns observed in other angiosperm chloroplast genomes (Figure 3.4).

.

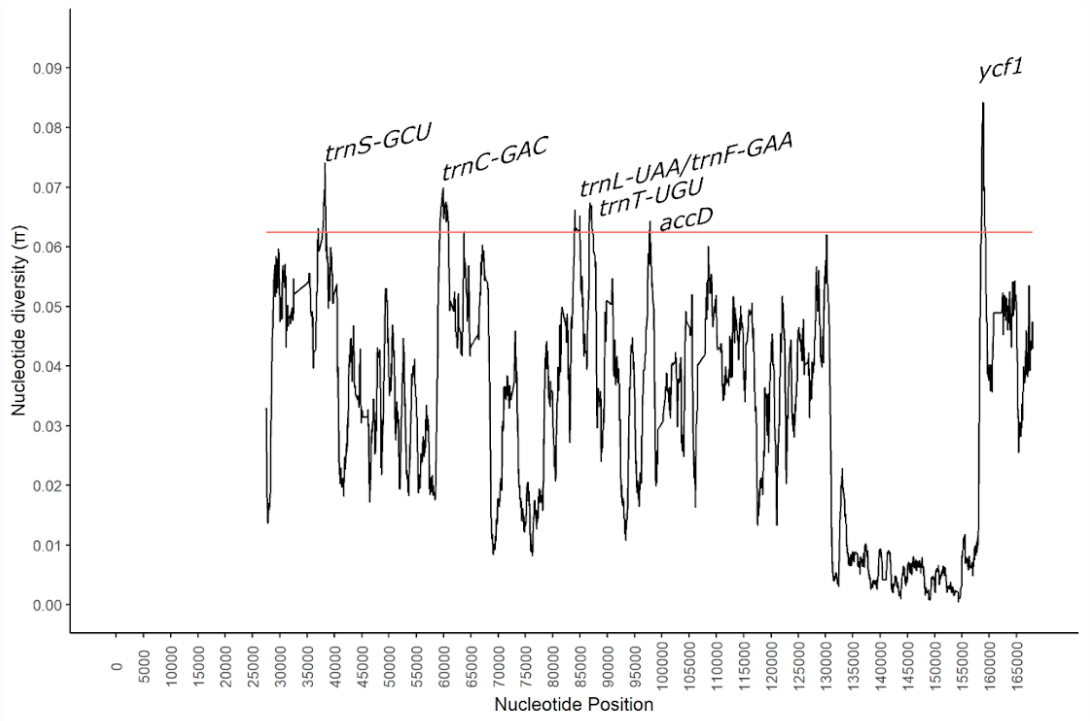

Figure 3.4 Nucleotide diversity (π) across the chloroplast genomes of Melastomataceae species. The solid line represents the median π value. Regions with π values above the median are considered potentially polymorphic and may serve as candidate loci for studies of genetic diversity, population structure, and phylogeography

4.3.7 Phylogenetic Analysis

The phylogenetic analysis revealed a strongly supported clade uniting *Mouriri elliptica* and *M. emarginata* (bootstrap = 100%). Although a complete chloroplast genome for *M. emarginata* is not yet available, partial chloroplast sequences were reconstructed in this study using raw reads retrieved from the Sequence Read Archive (SRA) to enable its inclusion in the phylogenetic inference. The genus *Memecylon* was recovered as sister to *Mouriri*, also with full support, reinforcing the placement of *M. elliptica* within the subfamily Olisbeoideae. Notably, *Heterocentron elegans* was found nested within *Tibouchina*, indicating a close phylogenetic relationship (Figure 3.5).

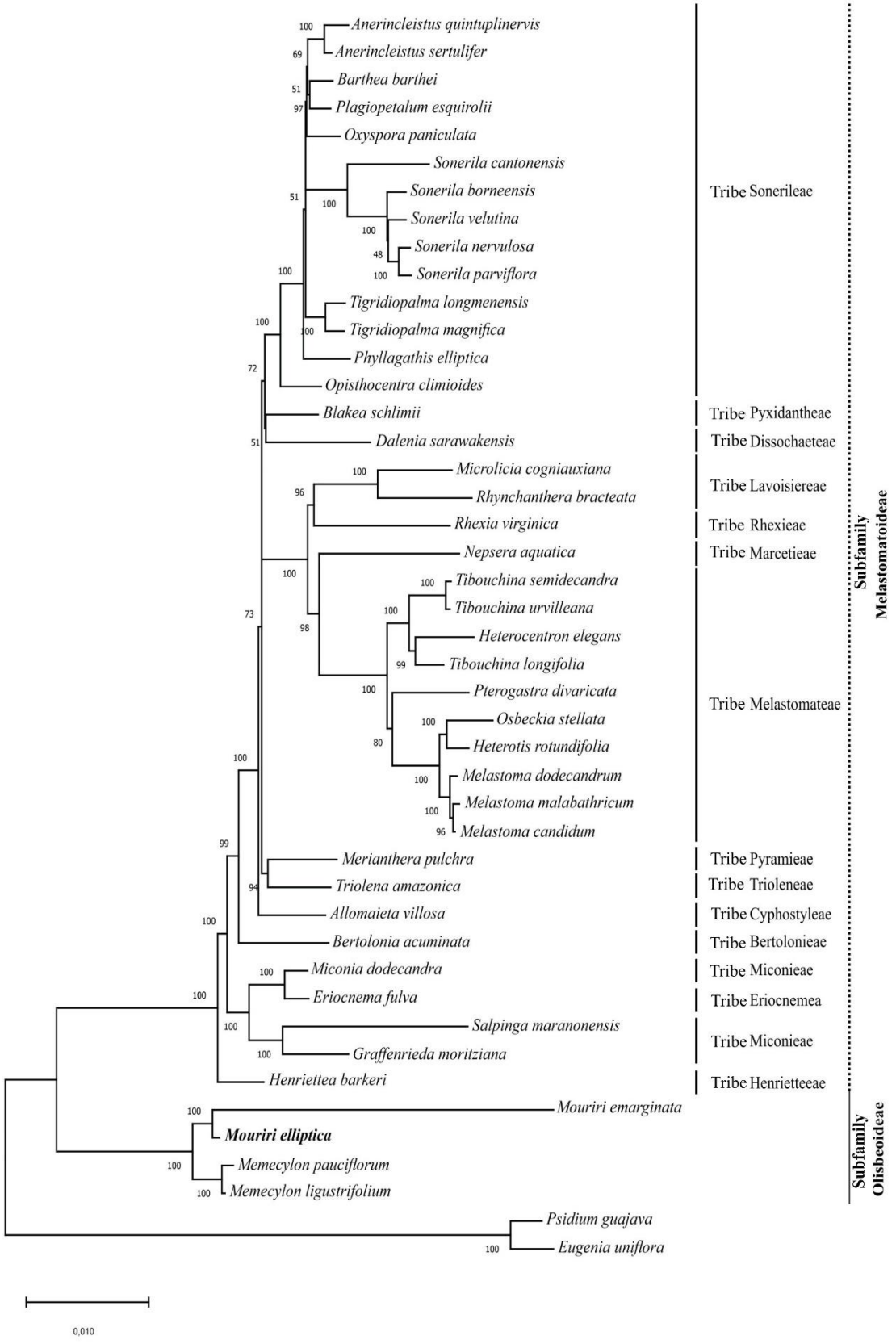

Figure 3.5 Phylogenetic tree of Melastomataceae species based on molecular data. Node values indicate bootstrap support.

4.3.8 Primer Design and *In silico* Validation

Primers for *MatK* and *rbcL* were designed and validated *in silico*, ensuring no self-dimers, cross-dimers, or stable secondary structures that could impair PCR efficiency. Amplicon size was optimized for compatibility with short-read sequencing platforms (e.g., Illumina MiSeq), ensuring the generation of high-quality, overlapping paired-end reads for accurate sequence assembly, which is a standard approach for high-throughput DNA barcoding studies. Primer details, including direction, binding position, amplicon length, Tm, GC content, and sequence, are comprehensively listed in Table 3.3.

Table 3.3 Primers designed for DNA barcoding of *Mouriri elliptica*.

Gene	Binding Position (Gene)	Forward Primer (5'→3')	Reverse Primer (5'→3')	Product Size (bp)	Tm (°C)
Mat K	79-203	TGAAGAAGCTCGAAACAAGA	ATCAGAATCGGATGAATCGG	125	50.21
Mat K	299-498	CATTTGCAGTCATTGTGGAA	TTGCGCCAAGATTTCTAGAT	200	49.99
rbcL	116-253	CAGGTACATGCGAAGAAATG	GTAGACCATTATCTCGGCAA	138	50.23
rbcL	69-226	TAGCAATGGAGTCCTGAAC	TAGTAAGAGATTGGGCCGAG	158	50.21
rbcL	103-239	GATATCTTTGCAGCATTCCG	TAACGATCAAGGCTGGTAAG	137	49.99

4.4 DISCUSSION

The chloroplast genome of *Mouriri elliptica* is 156,791 bp in length and exhibits the typical circular quadripartite structure, consisting of a large single-copy (LSC) region of 86,943 bp, a small single-copy (SSC) region of 17,234 bp, and two inverted repeats (IRs) of 26,307 bp each. This organization is a common and highly conserved feature of plastid genomes across angiosperms (Ruhlman & Jansen, 2014). The genome size and structure of *M. elliptica* are consistent with those reported for other Melastomataceae species. For instance, *Melastoma candidum* D. Don possesses a 156,682 bp genome (Ng et al., 2017), and *Osbeckia stellata* Buch.-Ham. ex D.Don has a 156,372 bp plastome (Liang et al., 2019). A broader analysis of 16 Melastomataceae

species by Reginato et al. (2016) found genome sizes ranging from 153,311 bp (*Salpinga maranonensis*) to 157,216 bp (*Miconia dodecandra*), with a mean of 155,806 bp, further underscoring the structural conservation within the family.

Variations in IR length among species reflect the dynamic processes of IR expansion and contraction, which are key drivers of chloroplast genome evolution. For example, *Microlicia cogniauxiana* has longer IRs (26,667 bp) than *M. elliptica*, while *Salpinga maranonensis* has shorter ones (25,765 bp) (Reginato et al., 2016; Zhang, X.-F. et al., 2021). Although these variations are relatively minor, they can influence the positioning of genes at the LSC/IR/SSC boundaries (JLB, JSB, JSA, JLA), with potential impacts on gene expression and regulation. The IRs in *M. elliptica* are stable and contain the typical suite of duplicated genes, unlike the rare but notable cases of complete IR loss reported in some lineages of Papilionoideae (Fabaceae) (Sabir et al., 2014; Wang, Wang & Yi, 2022). The overall structural conservation, including the maintenance of the IRs, reinforces the evolutionary stability of the chloroplast genome in Melastomataceae.

Genome annotation revealed a total of 113 unique genes, comprising 79 protein-coding genes, 30 transfer RNAs (tRNAs), and 4 ribosomal RNAs (rRNAs). Gene duplication within the IRs resulted in a total of 20 genes being duplicated, including 6 protein-coding genes (*ndhB, rpl2, rpl23, rps7, rps12, ycf2*), 7 tRNAs, and 4 rRNAs. This gene content is highly conserved and aligns with reports from other Melastomataceae species. For instance, *Melastoma dodecandrum* Lour. was reported to have 85 protein-coding genes, 38 tRNAs, and 8 rRNAs (Zheng et al., 2019), and Reginato et al. (2016) reported a common set of 84 protein-coding genes, 37 tRNAs, and 8 rRNAs (including duplicates) across the family. The duplication of genes like *rpl2* and *ndhB* in the IR regions is a recurrent feature associated with structural conservation and genome stability (Zhang, H. et al., 2021).

Intron analysis revealed that *ndhA* contains the largest intron (1,050 bp), while *rpl2* and *ndhB* harbor the smallest (676 bp each). The presence of introns in genes such as *ndhA* and *rpl2* indicates a potential for complex regulatory mechanisms, as also observed in *Melastoma candidum* (Ng et al., 2017). We identified 51 microsatellites (SSRs) and 24 longer repeat sequences. While chloroplast SSRs have historically been used as molecular markers (Provan, Powell & Hollingsworth, 2001), their utility has been surpassed by the analysis of entire plastome sequences and the identification of highly variable regions. The repeats identified here, particularly the 24 long repeats (13

palindromic and 11 forward), are a common feature of plastomes and can provide insights into genome organization and evolutionary mechanisms such as intramolecular recombination (Ahmad et al., 2022; Shaw et al., 2007). However, their functional and evolutionary significance in *M. elliptica* requires further investigation.

Nucleotide diversity (π) analysis across the Melastomataceae chloroplast genomes revealed considerable variation (π = 0.00054 to 0.09508), with distinct peaks of high diversity in intergenic spacers and the *ycf1* gene. The *ycf1* gene is a well-documented hotspot for nucleotide substitutions and indels and is often proposed as a super-barcode region for plant phylogenetics and species discrimination (Dong et al., 2015). The analysis of the core barcode genes confirmed this pattern: *rbcL* was highly conserved (π = 0.02070), while *MatK* was significantly more variable (π = 0.05461)(Hollingsworth et al., 2009a). These hypervariable regions are prime candidates for developing robust molecular markers for population genetics, phylogeography, and species delimitation within the Melastomataceae.

The phylogenetic analysis based on 53 concatenated chloroplast genes yielded a strongly supported phylogeny that clearly separates the subfamilies Olisbeoideae and Melastomatoideae, corroborating the most recent taxonomic framework for the family (Penneys et al., 2022). *Mouriri elliptica* and *M. emarginata* were recovered as sister species with maximum support (bootstrap = 100%). This clade was grouped with species of *Memecylon*, also with full support, confirming the close phylogenetic relationship between *Mouriri* and *Memecylon* within Olisbeoideae. This relationship is consistent with earlier phylogenetic studies based on a few plastid markers (Clausing & Renner, 2001) and reinforces the taxonomic placement of the genus.

An unexpected yet strongly supported relationship (bootstrap = 100%) was observed between *Tibouchina* and *Heterocentron elegans*, with the latter nested within *Tibouchina*. This result suggests that *Tibouchina* may not be monophyletic and highlights the complex evolutionary history within the tribe Melastomateae. Similar intergeneric relationships have been suggested in previous molecular studies (Michelangeli et al., 2013; Zhang et al., 2020), indicating that the current morphology-based taxonomy of some groups may not reflect evolutionary lineages accurately. Furthermore, the placement of *Eriocnema fulva* (tribe Eriocnemeae) as sister to *Miconia dodecandra* (tribe Miconieae) underscores the need for a comprehensive phylogenetic

reassessment of the family using large genomic datasets to resolve these complex relationships.

The primers designed for variable regions of the *MatK* and *rbcL* genes followed a strategic approach to leverage the higher variability of *MatK* for fine-scale discrimination and the conservation of *rbcL* for broader taxonomic comparisons. The *in silico* validation confirmed their optimal properties for PCR amplification, including appropriate melting temperature, GC content, and absence of secondary structures. This strategy is well-established for developing effective DNA barcodes (Hollingsworth et al., 2009b; Zheng et al., 2019).

While *in silico* results are promising, empirical validation through PCR amplification on *M. elliptica* and related species is an essential next step to confirm their utility for species identification, phylogenetic studies, and conservation efforts within the Melastomataceae. Overall, the data presented here provide the first complete chloroplast genome for the genus *Mouriri* and significantly enhance our understanding of chloroplast genome evolution and phylogenetic relationships in the Melastomataceae family. This work provides a solid genomic foundation for future taxonomic revisions, biogeographic studies, and conservation strategies, particularly for species in the threatened Cerrado biome.

4.5 CONCLUSION

This study presents the first complete chloroplast genome of *Mouriri elliptica*, expanding genomic resources for the Melastomataceae. The plastome exhibits the typical quadripartite structure of angiosperms and a conserved set of genes essential for photosynthesis and plastid function. Variations observed in inverted repeat regions and intron content reflect subtle evolutionary adjustments within the family.

The primers designed for *MatK* and *rbcL* demonstrated potential for DNA barcoding applications, supporting accurate species identification and biodiversity assessments. Phylogenetic analyses positioned *M. elliptica* as sister to *M. emarginata* and closely related to *Memecylon*, highlighting the need for further systematic revision within the subfamily Olisbeoideae.

Overall, this work fills a key genomic gap for *Mouriri*, strengthens the phylogenetic framework of Melastomataceae, and provides molecular tools for future evolutionary and conservation studies.

Conflict of Interest Statement: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Funding agencies had no role in the design of the study, in the collection, analysis, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

Acknowledgments: We thank INCT EECBio, the support from the Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES) through the PDPG Strategic Partnerships in States III Program (PDPG-FAPIII), Grant #88887.966073/2024-00. This work was developed under the project "Sustainable Development: Working Groups towards the 2030 Agenda" (CAPES Grant #88887.798673/2022-00), the Graduate Program in Genetics and Plant Breeding (PPGGMP), and the Federal University of Goiás (UFG) for financial support and research scholarships.

4.6 REFERENCES

AHMAD, W. et al. Complete chloroplast genome sequencing and comparative analysis of threatened dragon trees Dracaena serrulata and Dracaena cinnabari. **Scientific Reports**, v. 12, n. 1, p. 16787, 2022.

ANTUNES, A. M. et al. The chloroplast genome sequence of Dipteryx alata Vog. (Fabaceae: Papilionoideae): genomic features and comparative analysis with other legume genomes. **Brazilian Journal of Botany**, v. 43, n. 2, p. 271–282, 2020.

ASSIS, E. S. et al. Dissimilarity between *Mouriri elliptica* (Mart.) plants cultivated in vitro and in situ through anatomic parameters. **Genetics and Molecular Research**, v. 15, n. 4, p. gmr15049072, 2016.

BEIER, S. et al. MISA-web: a web server for microsatellite prediction. **Bioinformatics**, v. 33, n. 16, p. 2583–2585, 2017.

BRITO, J. B. P. et al. Complete Chloroplast Genomes of Pterodon emarginatus Vogel and Pterodon pubescens Benth: Comparative and Phylogenetic Analyses. **Current Genomics**, v. 24, n. 4, p. 236–249, 2023.

CLAUSING, G.; RENNER, S. S. Molecular phylogenetics of Melastomataceae and Memecylaceae: implications for character evolution. **American Journal of Botany**, v. 88, n. 3, p. 486–498, 2001.

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

DOYLE, J. J.; DOYLE, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. **Phytochemical Bulletin**, v. 19, n. 1, p. 11–15, 1987.

HOLLINGSWORTH, P. M. et al. A DNA barcode for land plants. **Proceedings of the National Academy of Sciences**, v. 106, n. 31, p. 12794–12797, 2009.

JIN, J. J. et al. GetOrganelle: a fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. **Genome Biology**, v. 21, n. 1, p. 241, 2020.

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. **Molecular Biology and Evolution, 35**(6), 1547–1549. https://doi.org/10.1093/molbev/msy096

LAGESEN, K. et al. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. **Nucleic Acids Research**, v. 35, n. 9, p. 3100–3108, 2007.

LIANG, R. et al. Characterization of the chloroplast genome of Osbeckia stellata (Melastomataceae). **Mitochondrial DNA Part B: Resources**, v. 4, n. 1, p. 1136–1137, 2019.

LOHSE, M.; DRECHSEL, O.; BOCK, R. OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. **Current Genetics**, v. 52, n. 5–6, p. 267–274, 2007.

LOWE, T. M.; EDDY, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. **Nucleic Acids Research**, v. 25, n. 5, p. 955–964, 1997.

MICHELANGELI, F. A. et al. Phylogenetic relationships and distribution of New World Melastomeae (Melastomataceae). **Botanical Journal of the Linnean Society**, v. 171, n. 1, p. 38–60, 2013.

MOLEIRO, F. C. et al. *Mouriri elliptica*: validation of gastroprotective, healing and anti-Helicobacter pylori effects. **Journal of Ethnopharmacology**, v. 123, n. 3, p. 359–368, 2009.

NG, W. L.; TAN, S. G. The complete chloroplast genome sequence of Melastoma candidum (Melastomataceae). **Mitochondrial DNA Part B: Resources**, v. 2, n. 2, p. 783–784, 2017.

PENNEYS, D. S. et al. A new Melastomataceae classification informed by molecular phylogenetics and morphology. In: GOLDENBERG, R.; MICHELANGELI, F. A.; ALMEDA, F. (Ed.). **Systematics, Evolution, and Ecology of Melastomataceae**. Cham: Springer International Publishing, 2022. p. 109–165.

POTT, A.; POTT, V. J. **Plantas do Pantanal**. Brasília, DF: Embrapa-SPI, 1994.

PROVAN, J.; POWELL, W.; HOLLINGSWORTH, P. M. Chloroplast microsatellites: new tools for studies in plant ecology and evolution. **Trends in Ecology & Evolution**, v. 16, n. 3, p. 142–147, 2001.

Rambaut, A. (2018). FigTree v1.4.4. [Computer program]. Available from: http://tree.bio.ed.ac.uk/software/figtree/

REGINATO, M. et al. The first complete plastid genomes of Melastomataceae are highly structurally conserved. **PeerJ**, v. 4, p. e2715, 2016.

ROZAS, J. et al. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. **Molecular Biology and Evolution**, v. 34, n. 12, p. 3299–3302, 2017.

RUHLMAN, T. A.; JANSEN, R. K. The plastid genomes of flowering plants. In: **Molecular Plant Taxonomy: Methods and Protocols**. New York: Humana Press, 2014. p. 3–38.

SABIR, J. et al. Evolutionary and biotechnology implications of plastid genome variation in the inverted-repeat-lacking clade of legumes. **Plant Biotechnology Journal**, v. 12, n. 6, p. 743–754, 2014.

SHAW, J. et al. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. **American Journal of Botany**, v. 94, n. 3, p. 275–288, 2007.

SONG, Y. et al. Comparative Chloroplast Genomes of Sorghum Species: Sequence Divergence and Phylogenetic Relationships. **BioMed Research International**, v. 2019, p. 2935741, 2019.

SOUZA, U. J. B. de et al. The complete chloroplast genome of Stryphnodendron adstringens (Leguminosae - Caesalpinioideae): comparative analysis with related Mimosoid species. **Scientific Reports**, v. 9, n. 1, p. 14206, 2019.

TILLICH, M. et al. GeSeq - versatile and accurate annotation of organelle genomes. **Nucleic Acids Research**, v. 45, n. W1, p. W6–W11, 2017.

VÖLTZ, R. R.; GOLDENBERG, R. Mouriri in Flora e Funga do Brasil. Jardim Botânico do Rio de Janeiro, 2024. Disponível em: https://floradobrasil.jbrj.gov.br/FB10000. Acesso em: 12 ago. 2025.

WANG, Z. X.; WANG, D. J.; YI, T. S. Does IR-loss promote plastome structural variation and sequence evolution? **Frontiers in Plant Science**, v. 13, p. 1016097, 2022.

WENG, A.-F. et al. The complete chloroplast genome sequence of Phyllagathis hainanensis (Melastomataceae) and phylogenetic analysis. **Mitochondrial DNA Part B: Resources**, v. 6, n. 3, p. 1178–1180, 2021.

ZHANG, H. et al. Transcriptome based high-throughput SSRs and SNPs discovery in the medicinal plant Lagenaria siceraria. **BIOCELL**, v. 45, n. 2, p. 371–386, 2021.

ZHANG, X. et al. The complete chloroplast genomes of Heterotis rotundifolia and Heterocentron elegans (Melastomataceae). **Mitochondrial DNA Part B: Resources**, v. 5, n. 3, p. 3094–3095, 2020.

ZHANG, X.-F. et al. Comparative analysis of chloroplast genome structure and molecular dating in Myrtales. **BMC Plant Biology**, v. 21, n. 1, p. 219, 2021.

ZHENG, X. et al. Structure and features of the complete chloroplast genome of Melastoma dodecandrum. **Physiology and Molecular Biology of Plants**, v. 25, n. 4, p. 1043–1055, 2019.

CAPÍTULO 3

PRIMER SYNTHESIS AND PARTIAL GENOME ASSEMBLY FOR GENOTYPING BY AMPLICON SEQUENCING IN *Mouriri elliptica* Mart. (MELASTOMATACEAE)

Juliana Borges Pereira Brito[1,2]; Adriana Maria Antunes[1,2]; Lázaro José Chaves[1]; Mariana Pires de Campos Telles[2]; Marla Arianne Almeida Silva[2]; Thannya Nascimento Soares[1,2]

[1]Graduate Program in Genetics and Plant Breeding, Federal University of Goiás, Goiânia, Brazil.
[2]Laboratory of Genetics and Biodiversity, Federal University of Goiás, Goiânia, Brazil.
*Corresponding authors: juliana.freitas@seduc.go.gov.br and tnsoares@ufg.br

5 PRIMER SYNTHESIS AND PARTIAL GENOME ASSEMBLY FOR GENOTYPING BY AMPLICON SEQUENCING IN *Mouriri elliptica* Mart. (MELASTOMATACEAE)

Juliana Borges Pereira Brito[1,2]; Adriana Maria Antunes[1,2]; Lázaro José Chaves[1]; Mariana Pires de Campos Telles[2]; Marla Arianne Almeida Silva[2]; Thannya Nascimento Soares[1,2]

[1]Graduate Program in Genetics and Plant Breeding, Federal University of Goiás, Goiânia, Brazil.
[2]Laboratory of Genetics and Biodiversity, Federal University of Goiás, Goiânia, Brazil.
*Corresponding authors: juliana.freitas@seduc.go.gov.br and tnsoares@ufg.br

ABSTRACT

The molecular characterization of plant genetic resources is crucial for biodiversity conservation and the sustainable use of native species. However, many non-model plants from the Brazilian Cerrado remain genetically unexplored, limiting the development of molecular tools for their conservation and management. In this context, this study aimed to generate genomic resources for *Mouriri elliptica* Mart. (Melastomataceae), an ecologically important Cerrado species, through the partial assembly and annotation of ITS nuclear genome and the recovery of organellar sequences. Genome skimming data were used to develop and validate primers for microsatellite loci and standard DNA barcode genes. The assembly yielded 47,075 scaffolds (N50 = 26,418 bp) and demonstrated high completeness (90.8% complete orthologs, BUSCO). We identified 6,602 microsatellite regions, and primer pairs were designed for all loci; a stringent filtering process selected 65 high-quality candidates for microsatellite genotyping. In addition, chloroplast (*MatK* and *rbcL*) and nuclear ribosomal (ITS) barcode regions were recovered, and primers were designed for each (*MatK*: 2 pairs; *rbcL*: 3 pairs; ITS: 1 pair). The PCR multiplex systems were validated *in silico* using OpenPrimeR, showing high predicted compatibility and efficiency. These molecular markers provide valuable tools for assessing genetic diversity, population structure, and phylogeography of *M. elliptica*, supporting biodiversity conservation efforts and the sustainable use of Cerrado plant resources.

Keywords: Bioinformatics; Cerrado; Conservation genetics; Molecular markers; SSR discovery.

5.1 INTRODUCTION

The growing demand for conservation strategies and the sustainable use of biodiversity has driven the need for genetic characterization of native species, particularly

in threatened biomes such as the Brazilian Cerrado (Mittermeier et al., 2004). Recognized as a biodiversity hotspot, the Cerrado harbors a wide range of endemic and ecologically important species, many of which face serious threats due to deforestation, agricultural expansion, and climate change (Klink & Machado, 2005; Myers et al., 2000).

Recent advances in next-generation sequencing techologies (NGS) technologies have enabled comprehensive genetic studies of non-model species, providing unprecedented opportunities for genomic resource development (Lu et al., 2025). These approaches have proven particularly valuable for plant species from biodiverse yet threatened ecosystems like the Cerrado (Alvarez et al., 2025).

Mouriri elliptica Mart. (Melastomataceae), commonly known as "croadinha" or "croada," is a shrub species endemic to the Brazilian Cerrado, with both ecological and medicinal importance (Völtz & Goldenberg, 2024). Ecologically, it plays a key role in maintaining Cerrado biodiversity, serving as a food source for local fauna and contributing to the regeneration of degraded areas (Moleiro et al., 2009). Medicinally, *M. elliptica* has been used in traditional medicine for treating various conditions, including inflammation, ulcers, and infections (Moleiro et al., 2009). Despite its relevance, no genomic studies or molecular marker systems have been developed specifically for *M. elliptica*, hindering its genetic characterization and the implementation of effective conservation strategies.

In this context, molecular markers serve as essential tools for understanding genetic diversity, population structure, and adaptive potential of native plant species (Allendorf, Hohenlohe & Luikart, 2010). Microsatellites, or simple sequence repeats (SSRs), are DNA sequences composed of short, tandemly repeated nucleotide motifs, widely distributed across eukaryotic genomes (Ellegren, 2004). Their abundance, high polymorphism, and codominant inheritance make them particularly valuable for genetic studies, including analyses of genetic diversity, population structure, genetic mapping, and kinship identification (Mittal & Dubey, 2009). The development of NGS-based methodologies has greatly enhanced the efficiency of microsatellite marker discovery (Zalapa et al., 2012). In this study, we not only leveraged NGS for *in silico* marker development from genome skimming data but also adopted a genotyping-by-sequencing approach for microsatellites (SSR-GBS). This method, as demonstrated in pioneering studies (Tibihika et al., 2019; Kariuki et al., 2021; Vartia et al., 2016), allows for highly multiplexed genotyping by sequencing SSR amplicons directly, offering key advantages

over traditional capillary electrophoresis. These advantages include higher throughput, the ability to discriminate alleles by both fragment length and nucleotide composition, and the generation of easily comparable digital data. This integrated pipeline represents a powerful strategy for comprehensive population genomic studies in non-model species like *M. elliptica*.

In parallel, DNA barcoding, based on the analysis of conserved gene regions, mainly plastid genes such as maturase K (*MatK*) and ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit (*rbcL*), and the nuclear Internal Transcribed Spacer (ITS) region, has emerged as an effective tool for species identification and discovery of new taxa (Hebert et al., 2003; Kress & Erickson, 2007). These standard barcode regions can be efficiently analyzed using NGS platforms, facilitating their integration with microsatellite-based approaches (Chen et al., 2010; Hollingsworth, 2011). The integration of multiplex PCR, which allows simultaneous amplification of multiple loci in a single reaction, with NGS-based marker development represents an efficient and cost-effective strategy for large-scale genetic studies (Edwards & Gibbs, 1994; Kreer et al., 2019). This approach enables comprehensive genetic assessments that are particularly valuable for conservation-oriented research.

In this study, we employed genome skimming to sequence, assemble, and annotate the partial nuclear genome of *M. elliptica*, in order to design primers for multiplex PCR systems for microsatellite and putative DNA barcode regions. Our goal was to provide genetic tools for future populations genetic analysis in *M. elliptica*, thereby contributing to both *in situ* and *ex situ* conservation efforts. Bioinformatic tools were applied for genome assembly, gene prediction, SSR identification, and primer design for both nuclear and plastid targets (*MatK, rbcL)* and ITS. The selected primers were organized into pools and evaluated for multiplex compatibility and efficiency using the OpenPrimeR software (Kreer et al., 2019). T his work establishes a methodological framework and genomic resources for *M. elliptica*, which may guide similar approaches in other native Cerrado plant species.

5.2 MATERIALS AND METHODS

5.2.1 Plant Material, DNA Sequencing

In 2023, leaf samples from a mature individual of *Mouriri elliptica* Mart. were collected at the School of Agronomy, Federal University of Goiás (UFG), in Goiânia, Goiás, Brazil (latitude: –16.5991637; longitude: –49.2792997). The plant material was transported to the laboratory under conditions that ensured DNA preservation, and the leaves were stored at –80°C until processing. This specimen was selected due to its accessibility and the availability of young, healthy tissues, minimizing disturbance to natural populations. Because no genomic data were previously available for this species, a single individual was sampled to generate a reference genomic resource and to identify candidate molecular markers. Although this approach does not allow for direct inferences about population-level diversity, the markers developed here can be applied and validated in future studies using multiple individuals and populations to assess genetic diversity and structure in *M. elliptica*. The voucher specimen was deposited in the UFG Herbarium under accession number UFG 51952.

Genomic DNA was isolated from young, healthy leaves of *Mouriri elliptica* following the cetyltrimethylammonium bromide (CTAB) protocol originally proposed by Doyle and Doyle (1987). The integrity of the extracted DNA was verified on 1% agarose gels, and its concentration was determined with a Qubit® fluorometer (Invitrogen). Library preparation followed the Nextera DNA Flex (Illumina) protocol, and sequencing was carried out on an Illumina MiSeq platform using the MiSeq v3 kit (2 × 300 bp, 600 cycles), yielding approximately 6.9 million paired-end reads. This low-coverage genome skimming strategy was adopted to enable the identification of molecular markers suitable for genotyping-by-sequencing applications.

Sequence quality was assessed with FastQC v0.11.9 (Andrews, 2010), which indicated a progressive reduction in base quality toward the ends of reads. To improve downstream assembly performance, reads were preprocessed using Trimmomatic v0.39 (Bolger, Lohse & Usadel, 2014) for adapter removal and trimming of low-quality bases, employing default settings. Multiple assemblers were subsequently evaluated to determine the optimal strategy for *de novo* genome reconstruction.

.

5.2.2 Genome Assembly Assessment and Completeness Analysis

De novo assembly was performed using SPAdes v3.15.4 (Bankevich et al., 2012), SOAPdenovo2 (Luo et al., 2012), ABySS v2.3.5 (Simpson et al., 2009), and MaSuRCA v4.0.9 (Zimin et al., 2013). These assemblers were selected due to their common use in plant genome assembly from short-read data. SPAdes was run with default parameters optimized for eukaryotic genomes. SOAPdenovo2 was run with k-mer size 31, ABySS with k-mer size 51 following manual recommendations, and MaSuRCA with default parameters, adjusting only PE=2 300 150 to match read length and insert size standard deviation.

Under this light preprocessing, all assemblers were compared using default settings. SPAdes v3.15.4 achieved the best balance among assembly metrics: producing fewer fragmented contigs (≥1,000 bp) than SOAPdenovo2, the longest maximum contig, and good overall integrity, despite a slightly lower N50 than SOAPdenovo2. This combination suggested a more continuous assembly with lower redundancy, desirable for reference purposes. Additionally, SPAdes is widely recognized for its robustness in assembling genomes from paired-end reads, supporting its selection for subsequent steps.

To investigate the effect of stricter preprocessing, SPAdes was rerun after more aggressive trimming with Trimmomatic v0.39, using the parameters: ILLUMINACLIP:NexteraPE.fa:2:20:10 HEADCROP:21 SLIDINGWINDOW:4:20 CROP:263. These settings targeted Nextera adapter removal (ILLUMINACLIP), minimum base quality enforcement (SLIDINGWINDOW), and removal of low-quality bases at the start and end of reads (CROP and HEADCROP).

The genome size of *Mouriri elliptica* was estimated *in silico* using k-mer frequency analysis. Raw reads were analyzed with GenomeScope (Vurture et al., 2017) and KmerGenie (Chikhi & Medvedev, 2014) to obtain independent estimates of genome size and assess data quality. These estimates were used as references for evaluating assembly completeness and guiding downstream analyses.

The resulting assemblies from all tested tools and parameters were compared based on standard metrics (number of contigs ≥ 1000 bp, N50, L50, total size, and largest contig) to select the best one for downstream analyses. Genome completeness was

evaluated using BUSCO (Benchmarking Universal Single-Copy Orthologs) v5.6.1 (Simão et al., 2015). This tool detects conserved single-copy orthologs within a taxonomic group to assess genome assembly integrity. We used the eudicots_odb10 dataset, containing 2,326 universal orthologs for eudicots, consistent with the taxonomic placement of *M. elliptica*. BUSCO was run in euk_genome_met mode with the MetaEuk gene predictor, optimized for eukaryotic genome annotation.

Assemblies of *Melastoma malabathricum* subsp. *normale* (GCA_029816415.1), *Melastoma candidum* (GCA_023653495.1), *Melastoma dodecandrum* (GCA_029817715.1), *Melastoma sanguineum* (GCA_029817735.1), and *Mouriri elliptica* were analyzed. The inclusion of *Melastoma* species was based on their phylogenetic proximity to *Mouriri*, enabling more accurate genome completeness comparisons. Only these Melastomataceae species were available in the NCBI database at the time. The results of this analysis are presented in the Results section.

5.2.3 Genome Annotation Pipeline

The annotation of the draft nuclear genome was performed using an integrated pipeline to identify repetitive elements, non-coding RNAs, and protein-coding genes. First, repetitive sequences were masked using RepeatMasker v4.1.2 with a *de novo*-generated repeat library from RepeatModeler v2.0.4. Subsequently, non-coding RNAs (tRNAs, miRNAs, snoRNAs, snRNAs) were identified using tRNAscan-SE v2.0.9 and Infernal v1.1.4 with Rfam models. Finally, protein-coding genes were predicted on the repeat-masked genome using AUGUSTUS v3.5.0.

Known repeats were annotated using RepeatMasker v4.1.2 (Tarailo-Graovac & Chen, 2009) with the eudicot repeat library from Dfam/RepBase, run with default parameters and the Crossmatch search engine. A *de novo* analysis was performed using RepeatModeler v2.0.4 (Flynn et al., 2020) to generate a *M. elliptica*-specific repeat library, which was then applied in a second RepeatMasker run for more comprehensive masking.

The tRNAs were identified with tRNAscan-SE v2.0.9 (Lowe & Eddy, 1997) in default mode with sensitivity optimized for eukaryotic genomes. Genomic coordinates of predicted tRNAs were extracted for further analysis. Additional non-coding RNAs

(ncRNAs), including miRNAs, snoRNAs, and snRNAs, were identified using Infernal v1.1.4 (Nawrocki & Eddy, 2013) with covariance models from Rfam v14.10 (Kalvari et al., 2021). The cmscan command was used with the --tblout option to generate tabular reports, considering only hits with e-value ≤ 1e–6 as reliable.

Gene prediction was performed using AUGUSTUS v3.5.0 (Stanke & Morgenstern, 2005) with the *Arabidopsis thaliana* training set, selected for its widespread use and high-quality training data. Repeat-masked sequences from RepeatMasker were used as input to avoid interference from repetitive regions.

5.2.4 Primers Design and PCR Multiplex Systems

The microsatellite regions were identified using QDD v3.1 (Meglécz et al., 2014) which detects SSRs in DNA sequences and designs specific primers. *M. elliptica* genomic sequences were scanned for microsatellite regions with 2–6 nucleotide repeat motifs, using default QDD parameters except for increasing the minimum flanking sequence length on each side of the repeat to 200 bp. An initial set of 5,134 primers was generated. A rigorous, multi-step filtering pipeline was applied within QDD using the following sequential criteria: minimum sequence length of 80 bp, flanking region length on each side ≥ 200 bp, and minimum repeat numbers 10 for dinucleotides, 8 for trinucleotides, and 6 for tetra- and pentanucleotides. Additional filters included: GC content (40-60%), melting temperature (Tm, 55-65°C), and potential for secondary structure formation (self-dimerization $\Delta G > -6$ kcal/mol, hairpin formation $\Delta G > -3$ kcal/mol). Furthermore, trinucleotide motifs were excluded from the final selection to prioritize more variable markers, and only one primer pair per contig was retained to avoid marker clustering.

For DNA barcoding, primers were designed for *MatK* and *rbcL* genes in a previous study and for the ITS gene in this study. Consensus sequences for these regions were obtained from our genome skimming data and aligned with reference sequences from Melastomataceae species available in GenBank using using the MUSCLE alignment algorithm (Edgar, 2004) implemented in MEGA X (Kumar et al., 2018). Using Primer3 (Untergasser et al., 2012), six primer pairs were designed: three for *rbcL*, two for *MatK*, and one for ITS. Criteria included primer length of 18–23 nucleotides, melting

temperature (Tm) of 48–62 °C, GC content of 40–60%, and absence of secondary structures. Primer3 was run with default parameters, adjusting only Tm and GC ranges to the specified values.

Selected primers were assessed for multiplexing feasibility using openPrimeR v1.0 (Kreer et al., 2019). The software was configured with the following parameters to optimize amplification efficiency and primer compatibility in multiplex reactions: primer length 18–23 nucleotides, Tm between 55 °C and 60 °C, GC content 40–60%, dimer formation free energy (ΔG) > –6 kcal/mol, and secondary structure formation ΔG > –3 kcal/mol. These criteria ensured specificity and efficient amplification in multiplex PCR.

After the initial selection, all primer pairs were manually re-examined to confirm their quality and potential for multiplex amplification. Candidates that did not meet the expected parameters of GC content (40–60%) or melting temperature (55–60 °C) were discarded. Primers showing low coverage of the target region or a high probability of forming stable secondary structures or dimers (ΔG values more negative than –3 kcal/mol and –6 kcal/mol, respectively) were also removed. To further refine the set, *in silico* verification was performed in OpenPrimeR, which identified and excluded primers with possible cross-dimer interactions. Only those combinations that satisfied all thermodynamic and specificity criteria were kept for the final multiplex configurations, ensuring their reliability for future microsatellite genotyping and DNA barcoding analyses (Figure 4.1).

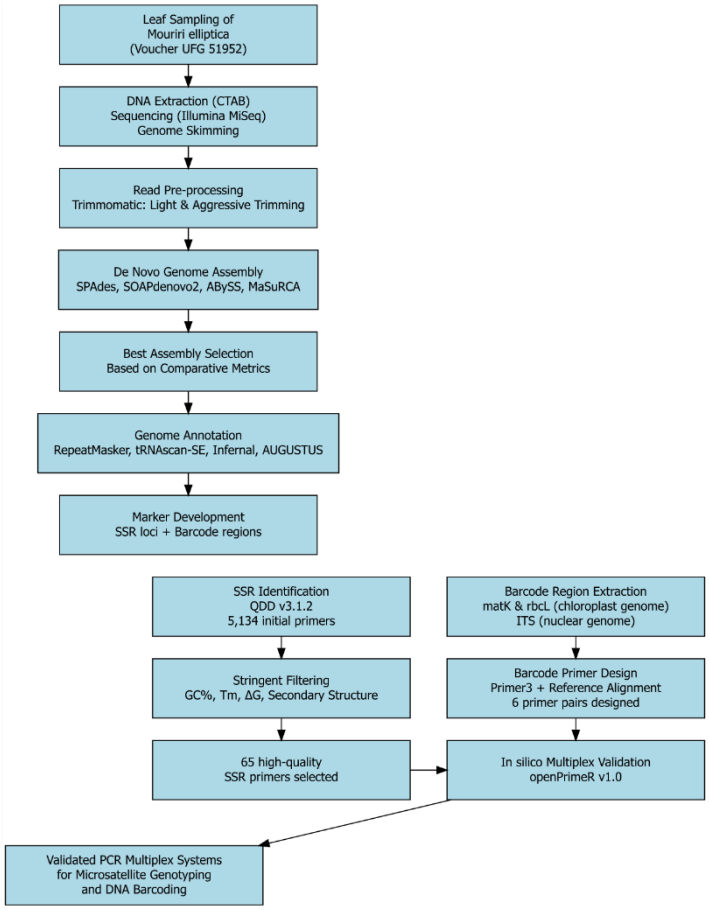

Figure 4.1 Workflow of the methodological steps for obtaining multiplex PCR systems for microsatellite loci and DNA barcoding in *Mouriri elliptica*. The process encompasses genome skimming, assembly, annotation, and the pipeline for the identification and *in silico* validation of primers for both simple sequence repeats (SSRs) and plastid DNA barcoding regions.

5.3 RESULTS

5.3.1 Nuclear Genome Assembly

The *de novo* assembly of the nuclear genome of *Mouriri elliptica* Mart. was performed using multiple assemblers and preprocessing strategies (Table 4.1). Among the tested approaches, SPAdes with aggressive trimming ("SPAdes (end)" in Table 4.1) showed the best performance in terms of continuity and completeness, and was therefore selected for all downstream analyses. The final assembly showed moderate continuity typical of short-read Illumina assemblies and an estimated total genome size of approximately 354 Mb, with genome coverage averaging 10.5×.

The genome size of *Mouriri elliptica* was estimated *in silico* using two k-mer–based approaches. GenomeScope estimated a size of approximately 283 Mb, while KmerGenie produced an estimate of 318.7 Mb. The slight difference between the two values suggests that the final assembly may marginally exceed the true genome size. Among the assemblies tested, the version generated with SPAdes showed the best balance between contiguity and completeness and was therefore selected for subsequent analyses.

Table 4.1 Comparison of assemblers for the nuclear genome assembly of *Mouriri elliptica*.

Assembler	#Contigs ≥1000 bp	N50 (bp)	L50	Total Size (bp)	Largest Contig (bp)	GC (%)
SPAdes (end)	67,212	26,418	6,110	354,000,000	812,730	44.75
SPAdes (light trim)	54,315	1,137	42,880	161,643,069	56,956	44.96
SOAPdenovo	67,212	2,277	24,371	230,117,124	372,124	44.75
ABySS (k=51)	117,799	4,048	15,407	261,800,000	52,152	-
MaSuRCA	42	697	207	88,140	2,402	

5.3.2 Genome Completeness Assessment (BUSCO)

The BUSCO completeness evaluation of five Melastomataceae genome assemblies, using the plant ortholog database (n = 2326), revealed high percentages of complete genes for all analyzed species (Figure 4.2).

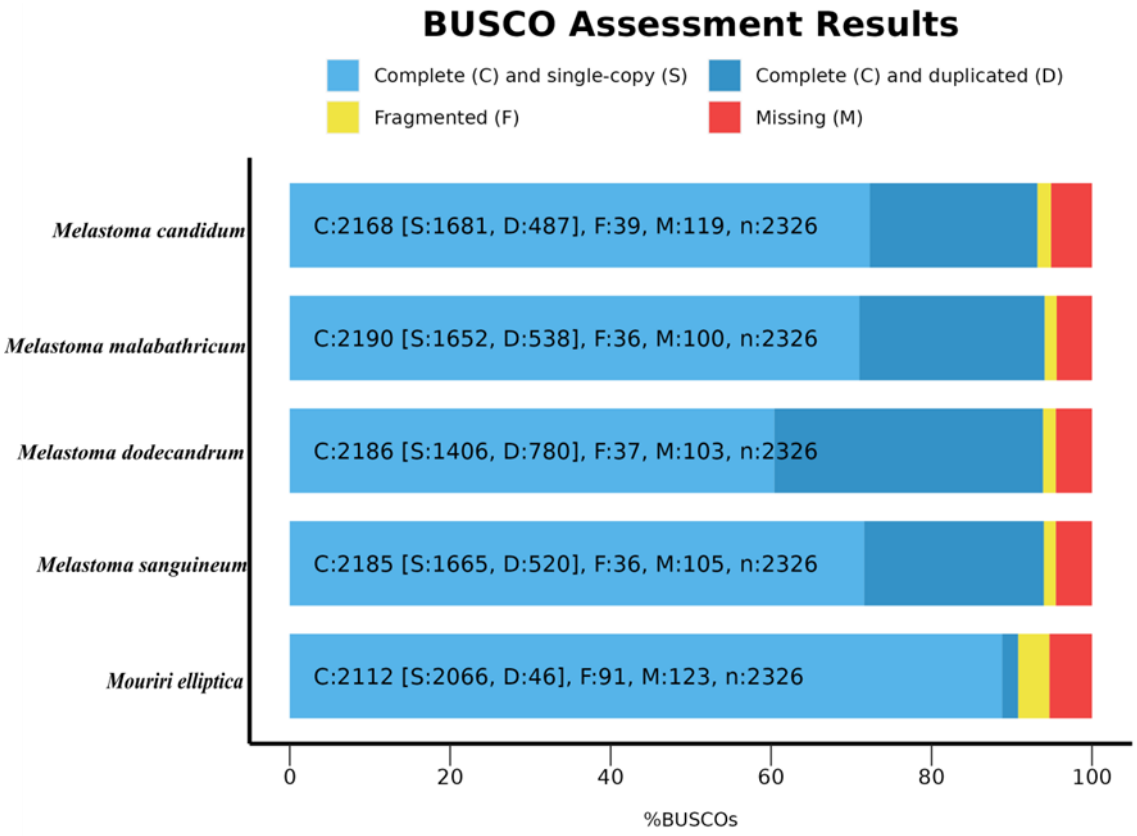

Figure 4.2 BUSCO completeness assessment for five *Melastomataceae* genomes using BUSCO v5.3.2 with the *eudicots_odb10* dataset. GenBank accession numbers: *Melastoma malabathricum* subsp. *normale* (GCA_029816415.1); *Melastoma candidum* (GCA_023653495.1); *Melastoma dodecandrum* (GCA_029817715.1); *Melastoma sanguineum* (GCA_029817735.1); and *Mouriri elliptica* (this study).

Despite the relatively low sequencing coverage, the nuclear genome of *M. elliptica* exhibited a satisfactory level of completeness, with 2,112 complete BUSCO genes identified, corresponding to 90.8% of the expected genes for the *eudicots_odb10* dataset. BUSCO completeness values ranged from 90.8% in *M. elliptica* (C: 2112) to 94.1% in *Melastoma malabathricum* (C: 2190). Among the complete BUSCO genes in *M. elliptica*, 2,066 (88.8%) were single-copy, while 46 (2.0%) were duplicated. Additionally, 91 fragmented genes (3.9%) were detected. The proportion of complete single-copy genes was notably lower in *Melastoma dodecandrum* (60.4%) due to a higher number of duplicated genes (D: 780). The number of fragmented genes was low, ranging from 36 to 91 among species, whereas missing genes represented between 4.3% (*M. malabathricum*) and 5.3% (*M. elliptica*) of the total (Table 4.2).

Table 4.2 Quality metrics for the *de novo* genome assembly of *Mouriri elliptica* under different filtering (No filter, >500 bp, and >1000 bp).

Metric	No filter	>500 bp	>1000 bp
Number of scaffolds	260711	159158	47075
Total scaffold length (bp)	472249253	427025333	354276744
N50 (bp)	12707	17135	26418
L50	6110	4574	2864
Largest scaffold (bp)	812730	812730	812730
Number of contigs	260731	159178	47095
Average scaffold length (bp)	1811	2683	7526

*Filter applied to the final assembly used for annotation.

5.3.3 Genome Annotation for Repetitive Elements, tRNA, and Non-Coding RNAs

Repetitive element analysis of the nuclear genome of *M. elliptica* revealed that approximately 19.4% of the genome content corresponds to interspersed repeats, with a predominance of LTR elements (4.49%), LINEs (1.66%), and DNA transposons (1.04%). Among retrotransposons, Ty1/Copia and Gypsy/DIRS1 were the most abundant. Additionally, approximately 12.06% of the genome was classified as repetitive but could not be assigned to a specific family, highlighting the presence of poorly characterized repetitive regions. No occurrences of simple repeats, low-complexity regions, or satellite sequences were detected using the employed library. Complete data are presented in Figure 4.3.

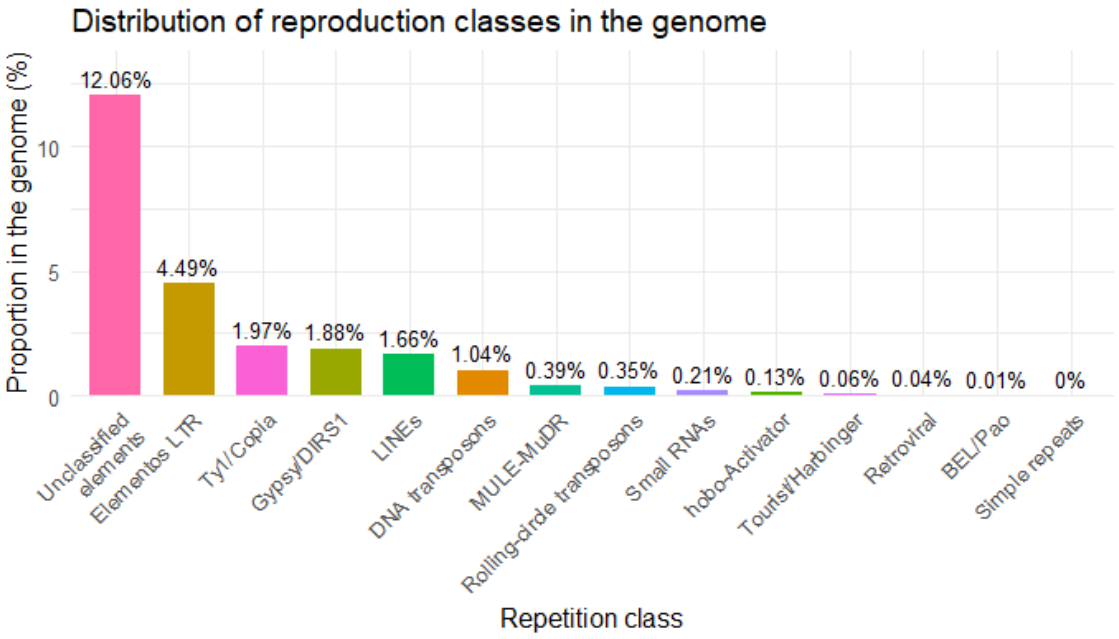

Figure 4.3 Repetitive elements identified in the nuclear genome of *Mouriri elliptica*.

Although Figure 4.3 shows no simple repeats detected by the genome-wide repeat annotation, microsatellite regions (SSRs) were identified and primers designed using a targeted approach with QDD v3.1. Genome-wide tools such as RepeatMasker primarily detect larger or more abundant repetitive elements, whereas QDD specifically targets short tandem repeats suitable for marker development.

Analysis of the nuclear genome of *M. elliptica* using tRNAscan-SE resulted in the annotation of 2,772 tRNA loci distributed across the scaffolds. Identified tRNAs corresponded to various amino acids, including codons for leucine (Leu), glycine (Gly), arginine (Arg), isoleucine (Ile), among others. Most tRNAs exhibited scores above 50, indicating high prediction reliability. Some loci were classified as pseudogenes or unidentified tRNAs (Undet), with scores below 40, which is common in highly divergent genomic regions or those with structural mutations. The presence of introns was detected in several functional tRNAs, particularly in codons such as TGT, CGA, and TTT.

The partial nuclear genome analysis of *M. elliptica* identified 119,655 non-coding RNAs (ncRNAs), distributed across 947 distinct functional classes. Among these, ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) were the most abundant, while most classes were represented by only a few copies each.

The predominant class was rRNAs, representing approximately 66.9% of the total, followed by tRNAs (15.4%) and misc_RNAs. Among small nuclear and related RNAs, 107 miRNAs and 2,063 snoRNAs were detected. Of these, only one miRNA and one snoRNA could be directly recognized by identifiers in the generated analysis files (Figure 4.4).

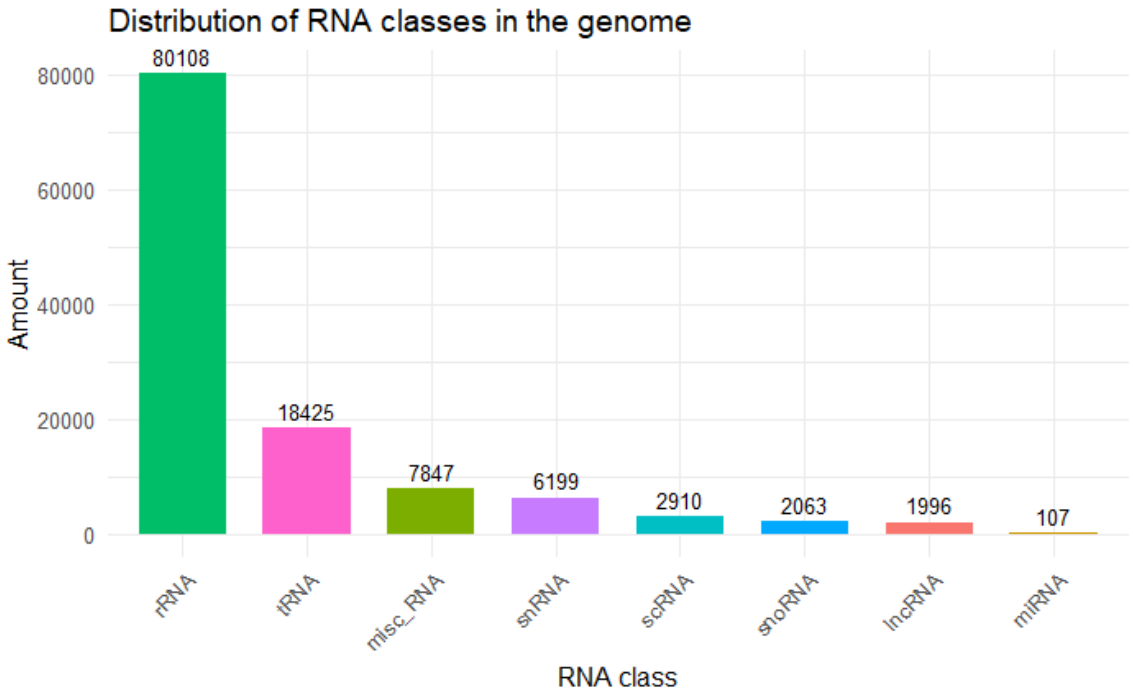

Figure 4.4 Percentage distribution of ncRNA classes identified in the nuclear genome of *Mouriri elliptica*.

5.3.4 Selection and Molecular Markers and Multiplex Design

Microsatellite analysis using QDD identified short tandem repeat regions in the nuclear genome of *Mouriri elliptica*. A total of 6,254 sequences were submitted for analysis (Pipe 2), of which 5,359 sequences contained detectable microsatellites (Pipe 3).

Primers were successfully designed, and after rigorous filtering based on GC content, melting temperature (Tm), secondary structure formation (ΔG > -3 kcal/mol), and dimer formation potential (ΔG > -6 kcal/mol), 65 high-quality primer pairs were selected for microsatellite loci.

For DNA barcoding, primer pairs were designed for the plastid genes *MatK* and *rbcL*, and the nuclear ITS region, resulting in six primer pairs: two for *MatK*, three for *rbcL*, and one for ITS.

The final set of 71 primers (65 SSRs + 6 barcodes) was analyzed with openPrimeR, which grouped them into three optimal PCR multiplex systems based on compatible melting temperatures and absence of cross-dimers. The characteristics of primers used in each multiplex assay are presented in tables 4.3, 4.4, and 4.5.

Table 4.3 Characteristics of primers used in multiplex assay 01 for microsatellite genotyping and DNA barcoding

Identifier	Forward sequence (5'-3')	Reverse sequence (5'-3')	Tm (°C)	Motif	N_ rep	Product size (bp)
Mel_gbs01	TTTGACCAAGCAATAACCGA	AGAAAGAGCCAACGTTAACT	49.91	AG	20	137
Mel_gbs02	AGCAGGTAAGACTCATTGAC	CTTCTCAGGCCTTTGTTGA	49.53	AC	16	199
Mel_gbs03	TCGAACAACAACAACAACAG	AATACGATCTGACATGGCAA	49.66	AG	19	157
Mel_gbs04	TTTGAGTCCAGCAAAGTTTG	GCAACGATGAAGACACAATC	49.93	AG	18	168
Mel_gbs05	TAAATCCGAGTAATCCACCG	CGGAAATTAATCTGCTGACG	50.06	AG	16	189
Mel_gbs06	TCCATTGTTATAGCCAACGG	AGGGATGATATGGGTGAAGA	50.41	AG	18	154
Mel_gbs07	GAACTGTGGGAGATGTGAAT	CACGCATCACTTCTTACAAC	49.13	AG	16	153
Mel_*rbcL*1	CAGGTACATGCGAAGAAATG	GTAGACCATTATCTCGGCAA	50.23	-	-	138
Mel_*MatK*1	TGAAGAAGCTCGAAACAAGA	ATCAGAATCGGATGAATCGG	50.21	-	-	125
Mel_*MatK*2	CATTTGCAGTCATTGTGGAA	TTGCGCCAAGATTTCTAGAT	49.99	-	-	200

Table 4.4 Characteristics of primers used in multiplex assay 02 for microsatellite genotyping and DNA barcoding

Identifier	**Forward sequence (5'-3')**	**Reverse sequence (5'-3')**	**Tm (°C)**	**Motif**	**N_rep**	**PCR product size**
Mel_gbs08	GGTTGTGAGGAATAAGCTCA	CCACCACATCAACGATACTA	49.91	AG	16	161
Mel_gbs09	CAACTGATAAAGGAGTCGCT	CAAGCACAATGAAGTTCGTT	49.53	AG	19	164
Mel_gbs10	AATGTTTGTACAGGAAGCCA	ACTGTAACACTTTGATGCCT	49.66	AC	24	162
Mel_gbs11	TGATGACCCTCTTGTTATGG	TCTCCTGTAACCAACAAGTG	49.93	AG	16	187
Mel_gbs12	TGTACATAGCGAGGAACAAC	CCTTGCTTCTTTCAATCCAC	50.06	ACACC	7	196
Mel_gbs13	GGTGTGTGTGTGTAAGAGAA	GTCCGGTTGATCCTTAAAGT	50.41	AG	18	152
Mel_gbs14	AGTTGACTCTAAATGCCTGG	GCATAAGCATGACAAGTTGT	49.13	AC	16	196
Mel_gbs15	TTCCTGGTCATAGCATGAAG	GAGAAGTATGTAGTCCGTCG	50.23	AG	19	164
Mel_*rbcL*2	TAGCAAATGGAGTCCTGAAC	TAGTAAAAGATTGGGCCGAG	50.21	-	-	158
Mel_*rbcL*3	GATATCTTTGCAGCATTCCG	TAACGATCAAGGCTGGTAAG	49.99	-	-	137

Table 4.5 Characteristics of primers used in multiplex assay 03 for microsatellite genotyping and DNA barcoding

Identifier	Forward sequence (5'-3')	Reverse sequence (5'-3')	Tm (°C)	Motif	N_rep	Product size (bp)
Mel_gbs16	GACAGAACACTGACTAAGCA	GCATCATCACGCAATCTTAG	49.96	AC	17	196
Mel_gbs17	ATTTCAGGTTTGTGCTTTGG	CTTTGAATGAGCCACTTTGG	49.58	ACTC	8	171
Mel_gbs18	CGGCAGAATACATGTGAATG	TCCAAGGATGTAACGAAAGG	49.92	AG	22	196
Mel_gbs19	TTTGATCGTAGAAGGAGTGC	CCATTTGTCTCTCTGCTCAT	50.40	AG	20	159
Mel_gbs20	AATCCATGGTTAGTGTGCTT	GACCGACTGATTAATTGGGT	49.76	AG	19	197
Mel_gbs21	GGAGTGGAATCGTAACAAGT	TCTCAGACTGCTTATTCACG	50.12	AG	16	157

Identifier	Forward sequence (5'-3')	Reverse sequence (5'-3')	Tm (°C)	Motif	N_rep	Product size (bp)
Mel_gbs22	ACCTTTCTTGCTTCCTCATT	CGCAACAACTTGTACATTCA	49.26	AG	22	196
Mel_gbs23	CTTGCCACCTGTATTGTTTC	ATTATGATGCTGTCGTGTCC	50.04	AC	19	191
Mel_gbs24	GAAAGAGAGAGGCTTGTCAA	AAAGGGTAAAGGTTGGAGTC	50.33	AG	16	179
Mel_its1	ACTTTCAACAACGGATCTCT	ATACTCTGAGGTGCAATGTG	49.36	-	-	125

5.4 DISCUSSION

The genome assembly of *Mouriri elliptica* Mart., obtained using SPAdes with aggressive trimming, resulted in a total size of 354 Mb — a value higher than the *in silico* genome size estimates from GenomeScope (~283 Mb) and KmerGenie (~318.7 Mb). This discrepancy likely reflects the inclusion of repetitive regions or redundant fragments, a common occurrence in *de novo* assemblies based on short reads (Chikhi & Medvedev, 2014; Vurture et al., 2017). Such expansion can also result from assembly artifacts, including technical duplications and fragmented contigs (Bankevich et al., 2012). Despite these challenges, the final SPAdes assembly exhibited superior contiguity (higher N50) and fewer scaffolds than other assemblers tested, demonstrating its suitability for subsequent genomic analyses. With a nuclear genome coverage of 10.5×, the sequencing depth was adequate for an exploratory study of a non-model species such as *M. elliptica* (Chikhi & Medvedev, 2014; Vurture et al., 2017). Overall, considering the evaluated parameters and the intrinsic limitations of short-read assemblies, the resulting dataset is considered robust and appropriate for the objectives of this study.

BUSCO analysis revealed a completeness of 90.8% of the expected conserved genes for eudicots, with 88.8% of these present as single-copy orthologs and 2.0% as duplicates. These values indicate satisfactory recovery of core genes, reflecting a functionally representative assembly with minimal duplication artifacts (Manni et al., 2021). The proportion of fragmented genes (3.9%) suggests that assembly continuity could still be improved, for example, through hybrid scaffolding with long-read data. Although a completeness value of 90.8% is not exceptionally high compared with well-assembled genomes (>95%), it remains satisfactory given the limited coverage and

exploratory nature of this work. As noted by Simão et al. (2015), BUSCO provides a biologically grounded metric that complements purely technical measures such as N50.

In *Melastoma dodecandrum* Lour., the elevated number of duplicated BUSCOs may represent real biological duplications or assembly artifacts arising from unresolved haplotypes. The BUSCO manual (Manni et al., 2021) recommends that duplication levels be interpreted in light of both biological and technical factors. Conversely, the slightly lower completeness of *M. elliptica*, with a higher proportion of fragmented and missing genes, may reflect greater genomic complexity or lower read quality, which can compromise contig integrity (Gurevich et al., 2013). Taken together, these results validate the methodological choices of this study and emphasize the need for cautious interpretation of gene duplications and fragmentation events.

Analysis of repetitive elements indicated that approximately 19.4% of the *Mouriri elliptica* Mart. genome consists of repeats. This estimate is likely conservative due to the partial assembly and the limitations of homology-based tools such as RepeatMasker (Smit, Hubley & Green, 2013–2015), which may miss lineage-specific or highly diverged elements. Long terminal repeat (LTR) retrotransposons of the Ty1/Copia and Ty3/Gypsy superfamilies were the most abundant, representing 1.95% and 2.00% of the genome, respectively. The relatively high proportion of unclassified repeats (~12%) suggests the presence of species-specific elements not yet represented in public databases. Similar patterns, where Copia and Gypsy elements dominate the repetitive landscape, have been observed in *Spinacia oleracea* L. (Li et al., 2019) and members of the Orobanchaceae family (Piednoël et al., 2012).

Annotation of non-coding RNAs revealed 2,772 tRNA loci and a total of 119,655 ncRNA elements, predominantly rRNAs and tRNAs, consistent with observations in other eudicot species such as *Eucalyptus grandis* W. Hill ex Maiden (2,185 tRNAs; Myburg et al., 2014) and *Populus trichocarpa* Torr. & A. Gray ex Hook. (2,452 tRNAs; Tuskan et al., 2006). Within the Melastomataceae, *Melastoma candidum* D. Don displayed a comparable pattern, with most annotated ncRNAs being structural (Zhong et al., 2023). Regulatory ncRNAs, including miRNAs and lncRNAs, were less represented, as expected in initial genome assemblies. These results provide a conservative but informative baseline for the development of microsatellite markers and other genomic studies in *M. elliptica*. Future work integrating species-specific repeat

libraries, *de novo* SSR detection, and long-read sequencing could further refine repeat content estimates and enhance genome annotation.

The proportion of repetitive elements can vary considerably among plants, reflecting their evolutionary histories and the dynamics of expansion and deletion of transposable elements (Flavell, 1986). The relatively high proportion of "unclassified" repeats (~12%) suggests the presence of lineage-specific elements in *Mouriri* or repetitive families not yet represented in public databases. This underscores the importance of *de novo* genomic studies for non-model species, particularly those from underexplored biomes such as the Brazilian Cerrado (Argentin et al., 2023).

The abundance of tRNA loci (2,772) detected in *M. elliptica* is comparable to that of other eudicots, such as *Eucalyptus grandis* W. Hill ex Maiden (2,185) and *Populus trichocarpa* Torr. & A. Gray ex Hook. (2,452) (Myburg et al., 2014; Tuskan et al., 2006). This pattern is consistent with species possessing larger genomes and higher repetitive content. tRNAs are essential in protein synthesis, and their abundance may be adapted to the translational demands of each species (Du et al., 2017). Annotation of non-coding RNAs (ncRNAs) revealed 119,655 elements, predominantly rRNAs (66.9%) and tRNAs (15.4%). Comparable results have been observed in other Melastomataceae genomes. In *Melastoma candidum* D. Don, 1,818 ncRNAs were annotated—mostly tRNAs (699) and rRNAs (233), along with snRNAs (698) and miRNAs (188). Similarly, *Barthea barthei* (H. Lév.) C. Y. Wu ex C. Chen presented 1,560 ncRNAs, including 737 tRNAs, 240 rRNAs, 348 snRNAs, and 235 miRNAs. These findings confirm that early-stage assemblies typically recover structural ncRNAs, while regulatory types such as miRNAs and lncRNAs are less represented, consistent with our observations for *M. elliptica* Mart.

Functional annotation using BLASTx produced no significant hits for ncRNAs, which was expected since BLASTx compares translated sequences against protein databases (e.g., UniProt/Swiss-Prot), and ncRNAs are non-coding by nature (Bairoch & Apweiler, 2000; Camacho et al., 2009). For ncRNA annotation and function prediction, specialized tools, such as covariance model, based searches and databases like Rfam, are more appropriate (Nawrocki & Eddy, 2013; Kalvari et al., 2021). To further explore ncRNA functionality in *M. elliptica*, complementary approaches such as *de novo* prediction, differential expression analyses across tissues, and functional assays (e.g., silencing or knockdown) are recommended, as demonstrated for small RNAs and lncRNAs in plants (Liu, Wang & Chua, 2015). To date, few studies in Melastomataceae

have characterized ncRNA repertoires in detail; for instance, the *Melastoma candidum* D. Don genome reported 1,818 ncRNAs (Zhong et al., 2023), highlighting both the relevance and scarcity of such analyses in this family.

The microsatellite primer pairs designed here show strong potential for applications in genetic diversity and conservation studies of *M. elliptica* through genotyping-by-sequencing. This approach enables the simultaneous analysis of multiple SSR loci and DNA barcode regions, allowing a comprehensive assessment of genetic variability (Tibihika et al., 2019; Yang et al., 2018). Microsatellites are abundant, highly polymorphic, and amenable to PCR amplification, making them particularly useful in population genetic studies (Meglécz et al., 2014). Primer design and *in silico* validation, including multiplex checks, followed established high-throughput marker development workflows. Similar strategies have been employed in *Eugenia klotzschiana* O. Berg for SSR discovery and validation using QDD.

In addition, DNA barcoding primers for *MatK*, *rbcL*, and ITS regions will be genotyped together with the SSR markers in the same sequencing runs, facilitating both species identification and phylogenetic analyses within Melastomataceae. This integrated strategy is particularly valuable for taxa like *M. elliptica*, whose identification in the field can be challenging in the absence of reproductive structures. Barcoding remains an efficient method for species discrimination, and the combination of *rbcL* + *MatK* is widely accepted as the standard plant barcode (Hebert et al., 2003; Kress & Erickson, 2007). The ITS region serves as a complementary nuclear marker, often enhancing resolution among closely related taxa (Chen et al., 2010; Hollingsworth, 2011). Consequently, the primer sets developed in this study constitute promising tools for both ecological applications and phylogenetic investigations within the genus *Mouriri* Aubl.

This study provided the first genomic and molecular framework for *Mouriri elliptica*, an endemic species of the Brazilian Cerrado. Despite the low sequencing coverage, the assembly achieved satisfactory completeness and offered valuable information about the genome structure of the species. The identification of microsatellite regions and the design of multiplexed primers integrated with DNA barcode markers established a practical and cost-effective genotyping strategy.

These results expand the available genetic resources for the Melastomataceae family and demonstrate the feasibility of developing genomic tools for non-model

species. The approaches proposed here can support future studies on population genetics, phylogeography, and conservation. Further efforts should aim to improve genome completeness and validate the designed markers experimentally across natural populations.

Conflict of Interest Statement: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Funding agencies had no role in the design of the study, in the collection, analysis, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

Acknowledgments: We thank INCT EECBio, the support from the Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES) through the PDPG Strategic Partnerships in States III Program (PDPG-FAPIII), Grant #88887.966073/2024-00. This work was developed under the project "Sustainable Development: Working Groups towards the 2030 Agenda" (CAPES Grant #88887.798673/2022-00), the Graduate Program in Genetics and Plant Breeding (PPGGMP), and the Federal University of Goiás (UFG) for financial support and research scholarships.

5.5 REFERENCES

ALLENDORF, F. W.; HOHENLOHE, P. A.; LUIKART, G. Genomics and the future of conservation genetics. **Nature Reviews Genetics 2010 11:10**, v. 11, n. 10, p. 697–709, 17 set. 2010.

ALVAREZ, F. *et al.* Tree species hyperdominance and rarity in the South American Cerrado. **Communications Biology**, v. 8, n. 1, p. 1–9, 1 dez. 2025.

ANDREWS, S. (2010). FastQC: A quality control tool for high throughput sequence data. .

ARGENTIN, J. *et al.* Comparative analysis of repeat content in plant genomes, large and small. **Frontiers in Plant Science**, v. 14, p. 1103035, 14 jul. 2023.

BAIROCH, A.; APWEILER, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. **Nucleic Acids Research**, v. 28, n. 1, p. 45–48, 1 jan. 2000.

BANKEVICH, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. **Journal of computational biology : a journal of computational molecular cell biology**, v. 19, n. 5, p. 455–477, 1 maio 2012.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics (Oxford, England)**, v. 30, n. 15, p. 2114–2120, 1 ago. 2014.

CAMACHO, C. *et al.* BLAST+: Architecture and applications. **BMC Bioinformatics**, v. 10, n. 1, p. 1–9, 15 dez. 2009.

CHEN, S. *et al.* Validation of the ITS2 Region as a Novel DNA Barcode for Identifying Medicinal Plant Species. **PLOS ONE**, v. 5, n. 1, p. e8613, 7 jan. 2010.

CHIKHI, R.; & MEDVEDEV, P. Informed and automated k-mer size selection for genome assembly. **Bioinformatics**, v. 30, n. 1, p. 31–37, 1 jan. 2014.

DIAS TIBIHIKA, P. *et al.* Application of microsatellite genotyping by sequencing (SSR-GBS) to measure genetic diversity of the East African Oreochromis niloticus. **Conservation Genetics**, v. 20, p. 357–372, 2019.

DOYLE, J. *et al.* A rapid DNA isolation procedure for small amounts of fresh leaf tissue. 1987.

DU, M. Z. *et al.* Co-adaption of tRNA gene copy number and amino acid usage influences translation rates in three life domains. **DNA Research**, v. 24, n. 6, p. 623–633, 1 dez. 2017.

EDWARDS, M. C.; GIBBS, R. A. Multiplex PCR: advantages, development, and applications. **PCR methods and applications**, v. 3, n. 4, 1994.

ELLEGREN, H. Microsatellites: simple sequences with complex evolution. **Nature Reviews Genetics 2004 5:6**, v. 5, n. 6, p. 435–445, jun. 2004.

FLAVELL, R. B. Repetitive DNA and chromosome evolution in plants. **Philosophical transactions of the Royal Society of London. Series B, Biological sciences**, v. 312, n. 1154, p. 227–242, 1986.

FLYNN, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. **Proceedings of the National Academy of Sciences of the United States of America**, v. 117, n. 17, p. 9451–9457, 28 abr. 2020.

GÖTZ, S. *et al.* High-throughput functional annotation and data mining with the Blast2GO suite. **Nucleic Acids Research**, v. 36, n. 10, p. 3420–3435, 1 jun. 2008.

GUREVICH, A. *et al.* QUAST: Quality assessment tool for genome assemblies. **Bioinformatics**, v. 29, n. 8, p. 1072–1075, 15 abr. 2013.

HEBERT, P. D. N. *et al.* Biological identifications through DNA barcodes. **Proceedings. Biological sciences**, v. 270, n. 1512, p. 313–321, 7 fev. 2003.

HOLLINGSWORTH, P. M. Refining the DNA barcode for land plants. **Proceedings of the National Academy of Sciences**, v. 108, n. 49, p. 19451–19452, 6 dez. 2011.

KALVARI, I. *et al.* Rfam 14: Expanded coverage of metagenomic, viral and microRNA families. **Nucleic Acids Research**, v. 49, n. D1, p. D192–D200, 8 jan. 2021.

KARIUKI, J. *et al.* Application of microsatellite genotyping by amplicon sequencing for delimitation of African tilapiine species relevant for aquaculture. **Aquaculture**, v. 537, p. 736501, 15 maio 2021.

KLINK, C. A.; MACHADO, R. B. A conservação do Cerrado brasileiro. [s.l: s.n.].

KREER, C. *et al.* openPrimeR for multiplex amplification of highly diverse templates. 2019.

KRESS, W. J.; ERICKSON, D. L. A Two-Locus Global DNA Barcode for Land Plants: The Coding *rbcL* Gene Complements the Non-Coding trnH-psbA Spacer Region. 2007.

KUMAR, S. *et al.* MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. **Molecular Biology and Evolution**, v. 35, n. 6, p. 1547, 1 jun. 2018.

LI, S. F. *et al.* The landscape of transposable elements and satellite DNAs in the genome of a dioecious plant spinach (Spinacia oleracea L.). **Mobile DNA**, v. 10, n. 1, 18 jan. 2019.

LIU, J.; WANG, H.; CHUA, N. H. Long noncoding RNA transcriptome of plants. **Plant Biotechnology Journal**, v. 13, n. 3, p. 319–328, 1 abr. 2015.

LOWE, T. M.; EDDY, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. **Nucleic Acids Research**, v. 25, n. 5, p. 955, 3 mar. 1997.

LU, Y. *et al.* Advances in Whole Genome Sequencing: Methods, Tools, and Applications in Population Genomics. **International Journal of Molecular Sciences**, v. 26, n. 1, p. 372, 1 jan. 2025.

LUO, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. 2012.

MANNI, M.; BERKELEY, MATTHEW R; *et al.* BUSCO: Assessing Genomic Data Quality and Beyond. 2021.

MANNI, M.; BERKELEY, MATTHEW R.; *et al.* BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. **Molecular Biology and Evolution**, v. 38, n. 10, p. 4647–4654, 1 out. 2021.

MEGLÉCZ, E. *et al.* QDD version 3.1: A user-friendly computer program for microsatellite selection and primer design revisited: Experimental validation of variables determining genotyping success rate. **Molecular Ecology Resources**, v. 14, n. 6, p. 1302–1313, 1 nov. 2014.

MITTAL, N.; DUBEY, A. Microsatellite markers - a new practice of DNA based markers in molecular genetics. **Pharmacognosy Reviews**, 2009.

MITTERMEIER, R. A. *et al.* Hotspots revisited: earth's biologically wealthiest and most threatened ecosystems. p. 85, 2004.

MOLEIRO, F. C. *et al.* Mouririelliptica: validation of gastroprotective, healing and anti-Helicobacter pylori effects. **Journal of Ethnopharmacology**, v. 123, n. 3, p. 359–368, 25 jun. 2009.

MYBURG, A. A. *et al.* The genome of Eucalyptus grandis. **Nature 2014 510:7505**, v. 510, n. 7505, p. 356–362, 11 jun. 2014.

MYERS, N. *et al.* Biodiversity hotspots for conservation priorities. **Nature**, v. 403, n. 6772, p. 853–858, 24 fev. 2000.

NAWROCKI, E. P.; EDDY, S. R. Infernal 1.1: 100-fold faster RNA homology searches. **Bioinformatics**, v. 29, n. 22, p. 2933–2935, 15 nov. 2013.

PIEDNOËL, M. *et al.* Next-Generation Sequencing Reveals the Impact of Repetitive DNA Across Phylogenetically Closely Related Genomes of Orobanchaceae. **Molecular Biology and Evolution**, v. 29, n. 11, p. 3601–3611, 1 nov. 2012.

QIU, L. *et al.* Exploiting EST databases for the development and characterization of EST-SSR markers in castor bean (Ricinus communis L.). **BMC Plant Biology**, v. 10, n. 1, p. 1–10, 16 dez. 2010.

SIMÃO, F. A. *et al.* BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. **Bioinformatics (Oxford, England)**, v. 31, n. 19, p. 3210–3212, 1 out. 2015.

SIMPSON, J. T. *et al.* ABySS: A parallel assembler for short read sequence data. **Genome Research**, v. 19, n. 6, p. 1117, jun. 2009.

STANKE, M.; MORGENSTERN, B. AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints. **Nucleic Acids Research**, v. 33, n. SUPPL. 2, jul. 2005.

TARAILO-GRAOVAC, M.; CHEN, N. Using RepeatMasker to identify repetitive elements in genomic sequences. **Current Protocols in Bioinformatics**, v. Chapter 4, n. SUPPL. 25, 2009.

TUSKAN, G. A. *et al.* The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). **Science**, v. 313, n. 5793, p. 1596–1604, 15 set. 2006.

UNTERGASSER, A. *et al.* Primer3—new capabilities and interfaces. **Nucleic Acids Research**, v. 40, n. 15, p. e115–e115, 1 ago. 2012.

VARTIA, S. *et al.* A novel method of microsatellite genotyping-by-sequencing using individual combinatorial barcoding. **Royal Society Open Science**, v. 3, n. 1, 1 jan. 2016.

VÖLTZ, R. R.;v GOLDENBERG, R. Mouriri in Flora e Funga do Brasil. Jardim Botânico do Rio de Janeiro.

VURTURE, G. W. *et al.* GenomeScope: Fast reference-free genome profiling from short reads. **Bioinformatics**, v. 33, n. 14, p. 2202–2204, 15 jul. 2017.

YANG, M. *et al.* Transcriptome Analysis and Microsatellite Markers Development of a Traditional Chinese Medicinal Herb Halenia elliptica D. Don (Gentianaceae). **Evolutionary Bioinformatics**, v. 14, 24 jul. 2018.

ZALAPA, J. E. *et al.* Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. **American Journal of Botany**, v. 99, n. 2, p. 193–208, 1 fev. 2012.

ZHONG, Y. *et al.* Chromosomal-level genome assembly of Melastoma candidum provides insights into trichome evolution. **Frontiers in Plant Science**, v. 14, p. 1126319, 27 jan. 2023.

ZIMIN, A. V. *et al.* The MaSuRCA genome assembler. **Bioinformatics**, v. 29, n. 21, p. 2669, 11 nov. 2013.

6 CONSIDERAÇÕES FINAIS

A jornada exploratória que culmina nesta tese representa um avanço significativo em nossa compreensão dos recursos genômicos de *Mouriri elliptica*, uma espécie nativa do Cerrado brasileiro com um valor ecológico e etnobotânico notável. Ao longo deste trabalho, desvendamos facetas importantes dos genomas cloroplastidial e nuclear da croadinha, lançando luz sobre sua história evolutiva, sua diversidade genética e seu potencial para aplicações biotecnológicas.

A montagem e anotação do genoma cloroplastidial de *M. elliptica* revelaram uma arquitetura genômica conservada, típica das angiospermas, com a identificação de genes essenciais para a fotossíntese e o metabolismo energético. A análise filogenética confirmou a posição de *M. elliptica* dentro da subfamília Olisbeoideae, estreitando sua relação com o gênero *Memecylon*. Esses resultados fornecem dados para estudos comparativos e filogenéticos em Melastomataceae, contribuindo para uma melhor compreensão da evolução desta família diversificada.

A geração de um rascunho do genoma nuclear de *M. elliptica*, embora ainda em um estágio inicial, representou um marco importante, considerando a carência de informações genômicas sobre a espécie. A anotação *in silico* revelou um alto nível de completude, com a identificação de uma vasta gama de genes codificadores, elementos repetitivos e RNAs não codificantes. A identificação de milhares de loci de microssatélites (SSRs) abriu novas perspectivas para o desenvolvimento de marcadores moleculares de alta resolução, com potencial para estudos de diversidade genética, estrutura populacional e mapeamento genético.

O desenvolvimento *in silico* de sistemas multiplex de PCR para loci de microssatélites e genes de barcoding representa um avanço estratégico para a caracterização da diversidade genética e a identificação de *M. elliptica* e outras espécies relacionadas. Embora a validação experimental desses marcadores não tenha sido possível no escopo deste trabalho, a análise *in silico* com o OpenPrimeR demonstrou a alta compatibilidade e eficiência predita dos primers, indicando o potencial desses sistemas para estudos futuros.

É importante reconhecer as limitações inerentes a este estudo. A montagem do genoma nuclear, baseada exclusivamente em dados de *short reads* da plataforma

Illumina, resultou em uma continuidade moderada, com um N50 relativamente baixo. A cobertura genômica limitada também pode ter afetado a qualidade da montagem e a precisão da anotação gênica. A ausência de validação experimental dos marcadores moleculares impede a avaliação do seu polimorfismo e da sua aplicabilidade em estudos de diversidade genética em populações naturais de *M. elliptica*.

Apesar dessas limitações, os resultados desta tese fornecem uma base sólida para futuras investigações sobre a genética e a genômica de *M. elliptica*. Em estudos futuros, recomenda-se a utilização de tecnologias de sequenciamento de terceira geração, como Oxford Nanopore e PacBio HiFi, para a obtenção de *reads* longos, que facilitam a montagem de genomas mais contíguos e completos. A combinação de dados de *short reads* e *long reads* poderá resultar em uma montagem genômica de alta qualidade, com resolução de regiões complexas e repetitivas. A validação experimental dos marcadores moleculares desenvolvidos neste estudo é essencial para confirmar sua eficiência e especificidade, e para avaliar o nível de polimorfismo dos loci microssatélites em diferentes populações de *M. elliptica*.

Além disso, sugere-se a realização de estudos transcriptômicos e proteômicos para a anotação funcional do genoma de *M. elliptica*, visando à identificação de genes relacionados à adaptação ao ambiente do Cerrado, à produção de metabólitos secundários de interesse farmacológico e a outras características relevantes para a espécie. A análise comparativa do genoma de *M. elliptica* com os genomas de outras espécies de Melastomataceae e de outros biomas brasileiros poderá fornecer respostas sobre a evolução genômica e a diversificação das plantas.

Em conclusão, esta tese representa uma contribuição original para o conhecimento dos recursos genômicos de *Mouriri elliptica* e para a conservação da biodiversidade do Cerrado. As ferramentas e os dados gerados neste trabalho poderão subsidiar futuras iniciativas de conservação *in situ* e *ex situ*, além de orientar programas de melhoramento genético e bioprospecção. Espera-se que esta tese inspire novas pesquisas sobre a genética e a genômica de espécies nativas do Cerrado, contribuindo para a proteção deste bioma ameaçado e para o uso sustentável de seus recursos genéticos.

7 REFERÊNCIAS

ALBUQUERQUE, L. B. D.; AQUINO, F. D. G.; COSTA, L. C.; MIRANDA, Z.; SOUSA, S. R. Espécies de Melastomataceae Juss. com potencial para restauração ecológica de mata ripária no cerrado., 2013. Acesso em: 12/ago./2025.

ALLENDORF, F. W.; HOHENLOHE, P. A.; LUIKART, G. Genomics and the future of conservation genetics. **Nature Reviews Genetics 2010 11:10**, v. 11, n. 10, p. 697–709, 2010. Nature Publishing Group. Disponível em: <https://www.nature.com/articles/nrg2844>. Acesso em: 21/ago./2024.

ASSIS, E. S.; RUBIO NETO, A.; CABRAL, P. D. S.; et al. Dissimilarity between *Mouriri elliptica* (Mart.) plants cultivated in vitro and in situ through anatomic parameters. **Genetics and Molecular Research**, v. 15, n. 4, 2016. Fundacao de Pesquisas Cientificas de Ribeirao Preto. Acesso em: 21/ago./2024.

CASTRO OLIVEIRA, H. W.; VIVEIRO, A. A. Cerrado e Plantas Medicinais: Algumas Reflexões sobre o Uso e a Conservação. **Embedded Systems and Applications**, v. 5, n. 3, 2013. Pro Reitoria de Pesquisa, Pos Graduacao e Inovacao - UFF. Acesso em: 20/ago./2024.

CLAUSING, G.; RENNER, S. S. Molecular phylogenetics of Melastomataceae and Memecylaceae: implications for character evolution. **American-Eurasian journal of botany**, v. 88, n. 3, p. 486–498, 2001. Botanical Society of America Inc. Acesso em: 21/ago./2024.

DÍAZ, S.; SETTELE, J.; BRONDÍZIO, E. S.; et al. Pervasive human-driven decline of life on Earth points to the need for transformative change. **Science**, v. 366, n. 6471, 2019. American Association for the Advancement of Science. Acesso em: 20/ago./2024.

EDWARDS, M. C.; GIBBS, R. A. Multiplex PCR: advantages, development, and applications. **PCR methods and applications**, v. 3, n. 4, 1994. Cold Spring Harbor Laboratory Press. Acesso em: 22/ago./2024.

FORMENTI, G.; THEISSINGER, K.; FERNANDES, C.; et al. **The era of reference genomes in conservation genomics**. 2022.

FURQUIM, L. C.; SANTOS, M. P. DOS; ANDRADE, C. A. O. DE; OLIVEIRA, L. A. DE; EVANGELISTA, A. W. P. Relação entre plantas nativas do Cerrado e água. **Científic@ - Multidisciplinary Journal**, v. 5, n. 2, p. 146–156, 2018. Associacao Educativa Evangelica.

GUIZARD, S. *et al.* Deep landscape update of dispersed and tandem repeats in the genome model of the red jungle fowl, Gallus gallus, using a series of de novo investigating tools. **BMC Genomics**, v. 17, p. 659, 2016. DOI: 10.1186/s12864-016-3015-5.

HEBERT, P. D. N.; CYWINSKA, A.; BALL, S. L.; DEWAARD, J. R. Biological identifications through DNA barcodes. **Proceedings. Biological sciences**, v. 270, n.

1512, p. 313–321, 2003. Proc Biol Sci. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/12614582/>. Acesso em: 22/ago./2024.

HOBAN, S.; BRUFORD, M. W.; DA SILVA, J. M.; et al. Genetic diversity goals and targets have improved, but remain insufficient for clear implementation of the post-2020 global biodiversity framework. **Conservation Genetics**, v. 24, n. 2, p. 181–191, 2023. Springer Science and Business Media B.V. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/36683963/>. Acesso em: 29/ago./2025.

JOSÉ BORGES DE SOUZA, U.; PIRES DE CAMPOS TELLES, M.; ALEXANDRE FELIZOLA DINIZ-FILHO, J. Tendências da literatura científica sobre genética de populações de plantas do Cerrado. **Hoehnea**, v. 43, n. 3, p. 461–477, 2016. Disponível em: <http://dx.>. Acesso em: 20/ago./2024.

KLINK, C. A.; MACHADO, R. B. **A conservação do Cerrado brasileiro**. 2005.

KRESS, W. J.; ERICKSON, D. L. A Two-Locus Global DNA Barcode for Land Plants: The Coding *rbcL* Gene Complements the Non-Coding trnH-psbA Spacer Region. 2007. Disponível em: <http://www.rbgkew.org.uk/barcod->. Acesso em: 22/ago./2024.

MACHADO, N. G.; AQUINO, B. G.; NEVES, G. A. P. C. Espécies nativas de plantas frutíferas em uma área de Cerrado em Mato Grosso, Brasil. **Revista Monografias Ambientais**, v. 13, n. 3, 2014. Universidad Federal de Santa Maria.

Ministério do Meio Ambiente — Ministério do Meio Ambiente e Mudança do Clima. .Disponível em: <https://www.gov.br/mma/pt-br>. Acesso em: 10/ago./2025.

MYERS, N.; MITTERMELER, R. A.; MITTERMELER, C. G.; DA FONSECA, G. A. B.; KENT, J. Biodiversity hotspots for conservation priorities. **Nature**, v. 403, n. 6772, p. 853–858, 2000. Nature Publishing Group. Disponível em: <https://www.nature.com/articles/35002501>. Acesso em: 12/ago./2025.

National Center for Biotechnology Information. .Disponível em: <https://www.ncbi.nlm.nih.gov/>. Acesso em: 29/ago./2025.

REGINATO, M.; NEUBIG, K. M.; MAJURE, L. C.; MICHELANGELI, F. A. The first complete plastid genomes of Melastomataceae are highly structurally conserved. **PeerJ**, v. 2016, n. 11, 2016. PeerJ Inc. Acesso em: 30/jan./2025.

SILVEIRA, M. R. S. D.; ALVES, R.; ARAGÃO, F. A. S. D.; FIGUEIREDO, R. W.; FREITAS, S. L. DE A. ESTUDO DE GENÓTIPOS DE PUÇÁ 'COROA DE FRADE' DA VEGETAÇÃO LITORÂNEA DE BEBERIBE-CE. **Revista Caatinga**, 2014. Acesso em: 20/ago./2024.

STEVENS, P. F.; DAVIS, H. M. The angiosperm phylogeny Website - a tool for reference and teaching in a time of change. **Proceedings of the American Society for Information Science and Technology**, v. 42, n. 1, 2005. Wiley. Disponível em: <https://www.ecosia.org/>. Acesso em: 12/ago./2025.

TAMBOSI, L. R. Estratégias espaciais baseadas em ecologia de paisagens para a otimização dos esforços de restauração. 2014. São Paulo: Universidade de São Paulo.

Disponível em: <http://www.teses.usp.br/teses/disponiveis/41/41134/tde-29052014-112453/>. Acesso em: 10/ago./2025.

VASCONCELOS, J. M.; CARDOSO, T. V.; SALES, J. DE F.; et al. Dormancy breaking methods in croada (*Mouriri elliptica* Mart) seeds. **Ciencia e Agrotecnologia**, v. 34, n. 5, p. 1199–1204, 2010. Federal University of Lavras. Acesso em: 20/ago./2024.

VÖLTZ, R. R. ;; GOLDENBERG, R. Mouriri in Flora e Funga do Brasil. Jardim Botânico do Rio de Janeiro.

WINK, J. G.; GOLDENBERG, R.; LIMA, L. C. P.; et al. Floristic inventory of Melastomataceae of the Iguaçu National Park, Paraná, Brazil. **Rodriguésia**, v. 75, 2024. Instituto de Pesquisas Jardim Botânico do Rio de Janeiro. Disponível em: <https://doaj.org/article/ff3b0383801d4b38bd7fc58d5353b875>. Acesso em: 12/ago./2025.

ZALAPA, J. E.; CUEVAS, H.; ZHU, H.; et al. Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. **American Journal of Botany**, v. 99, n. 2, p. 193–208, 2012. John Wiley & Sons, Ltd. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.3732/ajb.1100394>. Acesso em: 22/ago./2024.

APÊNDICES

Apêndice A – Resultado das repetições dispersas no genoma de cloroplasto.

QUERY	SUBJECT	TIPO	INÍCIO QUERY	INÍCIO SUBJECT	SCORE	E-VALUE
26307	86943	P	26307	130484	0	0.00E+00
48	65715	P	48	65715	0	8.73E-20
52	39849	D	52	42073	-2	4.07E-18
44	39857	D	44	42081	-1	2.95E-15
40	100745	D	40	123166	0	5.72E-15
40	123166	P	40	142949	0	5.72E-15
43	80316	P	43	80316	-3	2.98E-11
31	29907	P	31	29907	-1	1.39E-07
33	69618	P	33	69699	-2	4.45E-07
30	7958	P	30	46575	-1	5.40E-07
32	7956	D	32	36652	-2	1.67E-06
31	67375	D	31	67386	-2	6.27E-06
30	36533	P	30	36534	-2	2.35E-05
32	44801	D	32	123164	-3	5.02E-05
31	10148	P	31	10189	-3	1.82E-04
31	44811	D	31	100753	-3	1.82E-04
31	44811	D	31	123174	-3	1.82E-04
31	44811	P	31	142950	-3	1.82E-04
30	32395	P	30	32399	-3	6.57E-04
30	36654	P	30	46575	-3	6.57E-04
30	44294	P	30	63275	-3	6.57E-04
30	44803	D	30	100745	-3	6.57E-04
30	44803	P	30	142959	-3	6.57E-04
30	115773	D	30	115797	-3	6.57E-04

LEGENDA:
P = REPETIÇÃO PALINDRÔMICA (PALINDROMIC REPEAT)
D = REPETIÇÃO DISPERSA (DISPERSED REPEAT)
F = REPETIÇÃO DIRETA (FORWARD REPEAT)
C = REPETIÇÃO COMPLEMENTAR (COMPLEMENT REPEAT)
R = REPETIÇÃO REVERSA (REVERSE REPEAT)

Apêndice B – Repetições em tandem no genoma cloroplastidial de *Mouriri elliptica*

Início	Fim	Comp	Cópias	%matches	%indels	Sequência repetida
170	204	12	2.7	86	1.15	ATGAAAAAAGAAT
185	223	16	2.3	91	0.77	AAAAAAGAATATAAAAA
8154	8196	17	2.5	76	1.34	ATTAAATTAAAATTAGA
8440	8464	13	1.9	100	1.00	TAATTAATAATTA
8466	8493	13	2.2	100	1.26	AAATATTAGTTAA
16036	16068	17	1.9	100	1.21	AATTTATTTATCTTATA
32431	32461	15	2.1	93	1.31	AATAAAAATATATGT
33480	33504	13	1.9	100	0.87	AAATAAAAAAATC
33494	33535	20	2.1	82	1.18	AATAAAAAAACAATAAGAAG
36851	36876	13	2.0	100	1.83	CCAATCTATGTAT
44079	44103	12	2.1	100	1.22	TTTTCTTCATAT
44377	44401	12	2.1	100	1.94	GACTTCTATGAA
61289	61323	16	2.1	94	1.19	ATTTGATTTCTATTTTT
67381	67417	11	3.4	100	1.60	TTTGACTTTAA
68909	68935	12	2.2	100	1.58	TATGACATATTT
69045	69072	14	2.0	100	1.38	TCTTATTTATAATC
70136	70170	18	1.9	100	1.52	TTCTATATTTTCTAGTTG
77247	77284	18	2.1	95	1.68	TTTAATACCAATACTAAAG
84561	84592	16	2.0	93	1.44	TCTTATCTTTGCTTTG
85362	85390	14	2.1	100	1.18	TAATCATTTTTTTA
86624	86671	24	2.0	95	1.88	CTCTATAGGGGTTCGTCCTTCTCA
101094	101140	14	3.3	79	1.04	TTTTATTTTTATTA
109816	109841	13	2.0	100	1.55	CTCTATCTATCCA
115569	115603	18	1.9	88	1.08	AAAATGAAATAAAAATAA
115778	115827	24	2.1	100	1.29	AAAATATTATTACTATAAATAACT

Início	Fim	Comp	Cópias	%matches	%indels	Sequência repetida
130016	130073	12	4.4	85	1.01	TTTTTCTTCTCT
130016	130075	27	2.2	84	0.99	TTTTTCTTCTCTTTTGTATTCTCCTCT
130037	130087	24	2.1	82	0.88	TTCTCTTTTTTCTTCTCATTTTCT
133894	133919	13	2.0	100	1.55	TGGATAGATAGAG

APÊNDICE C – Microssatélites no genoma cloroplastidial de *Mouriri elliptica*

ID	TIPO SSR	MOTIVO	TAMANHO	INÍCIO	FIM
1	C	(A)10GAAAT(A)10	25	199	223
2	P1	(A)10	10	1975	1984
3	P1	(T)11	11	5926	5936
4	P1	(T)10	10	7092	7101
5	P1	(T)10	10	7548	7557
6	P1	(T)13	13	7737	7749
7	P1	(A)10	10	8257	8266
8	P1	(T)11	11	9897	9907
9	P1	(T)15	15	10215	10229
10	P1	(A)12	12	13318	13329
11	P1	(A)10	10	13877	13886
12	P1	(A)10	10	14770	14779
13	C	(T)14(A)11	25	16811	16835
14	P1	(T)10	10	23205	23214
15	P1	(T)10	10	26806	26815
16	P1	(T)11	11	29147	29157
17	P1	(T)11	11	29833	29843
18	P1	(T)10	10	30449	30458
19	P1	(T)10	10	33093	33102
20	P1	(T)11	11	45357	45367
21	P1	(A)11	11	46195	46205
22	P1	(A)10	10	47643	47652
23	C	(T)10GAGTTA(T)11	27	48496	48522
24	P1	(T)12	12	50355	50366
25	P1	(T)10	10	52032	52041

ID	TIPO SSR	MOTIVO	TAMANHO	INÍCIO	FIM
26	P1	(T)11	11	59139	59149
27	P1	(A)12	12	59576	59587
28	P1	(T)11	11	61186	61196
29	P1	(T)11	11	62741	62751
30	P1	(T)10	10	65284	65293
31	P1	(T)11	11	65637	65647
32	P1	(A)10	10	67573	67582
33	P1	(T)11	11	70214	70224
34	P1	(T)10	10	71418	71427
35	P1	(A)10	10	72631	72640
36	P1	(T)10	10	72908	72917
37	P1	(T)11	11	73966	73976
38	P1	(A)10	10	80471	80480
39	P1	(T)10	10	83194	83203
40	P1	(A)11	11	83832	83842
41	P1	(A)10	10	110145	110154
42	P1	(A)10	10	113046	113055
43	P1	(T)11	11	115671	115681
44	P1	(A)11	11	126447	126457
45	P1	(T)11	11	127897	127907
46	P1	(T)13	13	129161	129173
47	P1	(T)10	10	129376	129385
48	P1	(T)11	11	129535	129545
49	P1	(T)10	10	130068	130077
50	P1	(T)10	10	130680	130689

ID	TIPO SSR	MOTIVO	TAMANHO	INÍCIO	FIM
51	P1	(T)10	10	133581	133590