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

RECURSOS GENÔMICOS DE *Eugenia dysenterica* (Mart.) DC. - GENOMA CLOROPLASTIDIAL E TRANSCRITOMA DE REFERÊNCIA

STELA BARROS RIBEIRO

Orientador:
Prof. Dr. Alexandre Siqueira Guedes Coelho

Coorientadora:
Prof.ª Dr.ª Mariana Pires de Campos Telles

Goiânia - GO
Brasil

Junho – 2021

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

2. Nome completo do autor

Stela Barros Ribeiro

3. Título do trabalho

Recursos Genômicos de *Eugenia dysenterica* (Mart.) DC. - Genoma Cloroplastidial e Transcritoma de Referência

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

Concorda com a liberação total do documento [ ] SIM [X] 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 **STELA BARROS RIBEIRO**, **Discente**, em 06/01/2022, às 13:15, 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 **Alexandre Siqueira Guedes Coelho**, **Professor do Magistério Superior**, em 17/01/2022, às 08:20, 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 **2614125** e o código CRC **7E985498**.

Referência: Processo nº 23070.032663/2021-06 SEI nº 2614125

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

2. Nome completo do autor

Stela Barros Ribeiro

3. Título do trabalho

Recursos Genômicos de *Eugenia dysenterica* (Mart.) DC. - Genoma Cloroplastidial e Transcritoma de Referência

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 **Stela Barros Ribeiro**, **Usuário Externo**, em 26/08/2024, às 16:51, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Referência: Processo nº 23070.032663/2021-06 SEI nº 4773235

STELA BARROS RIBEIRO

RECURSOS GENÔMICOS DE *Eugenia dysenterica* (Mart.) DC. - GEMONA CLOROPLASTIDIAL E TRANSCRITOMA DE REFERÊNCIA

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 a obtenção do título de Doutora em Genética e Melhoramento de Plantas.
Área de concentração: Genética e Melhoramento de Plantas
Linha de pesquisa: Genética e Genômica de Plantas

Orientador:
Prof. Dr. Alexandre Siqueira Guedes Coelho

Coorientadora:
Prof.ª Dr.ª Mariana Pires de Campos Telles

Goiânia, GO – Brasil
2021

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

RIBEIRO, STELA BARROS
Recursos Genômicos de Eugenia dysenterica (Mart.) DC - Genoma cloroplastidial e transcritoma de referência [manuscrito] / STELA BARROS RIBEIRO. - 2021.
92 f.

Orientador: Prof. Dr. Alexandre Siqueira Guedes Coelho; co orientadora Mariana Pires de Campos Telles.
Tese (Doutorado) - Universidade Federal de Goiás, Escola de Agronomia (EA), Programa de Pós-graduação em Genética e Melhoramento de Plantas, Goiânia, 2021.
Bibliografia. Apêndice.
Inclui fotografias, gráfico, tabelas.

1. Cerrado. 2. Myrtaceae. 3. NGS. 4. Plastoma. 5. Transcritoma. I. Coelho, Alexandre Siqueira Guedes, orient. II. Título.

CDU 575

UNIVERSIDADE FEDERAL DE GOIÁS

ESCOLA DE AGRONOMIA

ATA DE DEFESA DE TESE

Ata Nº 104/2021 da sessão de Defesa de Tese de **Stela Barros Ribeiro** que confere o título de Doutora em Genética e Melhoramento de Plantas, na área de concentração em Genética e Melhoramento de Plantas.

Aos trinta dias do mês de junho do ano de dois mil e vinte e um, a partir das oito horas, por videoconferência, realizou-se a sessão pública de Defesa de Tese intitulada "Recursos Genômicos de *Eugenia dysenterica* (Mart.) DC.". Os trabalhos foram instalados pelo Orientador e Presidente da Banca Examinadora, **Prof. Alexandre Siqueira Guedes Coelho - EA/UFG,** com a participação dos demais membros da Banca Examinadora: **Profa. Mariana Pires de Campos Telles - ICB/UFG**, coorientadora; **Dra. Tereza Cristina de Oliveira Borba - Embrapa Arroz e Feijão**, membro titular externo; **Dra. Adriana Maria Antunes Taquary - EA/UFG**, membro titular externo; **Dra. Cíntia Pelegrineti Targueta de Azevedo Brito - ICB/UFG**, membro titular externo; **Dra. Isabela Pavanelli de Souza - Embrapa Arroz e Feijão**, membro titular externo. Durante a arguição, os membros da banca fizeram sugestão de alteração do título do trabalho. Após a arguição, 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 pelo Presidente da Banca Examinadora, **Prof. Alexandre Siqueira Guedes Coelho**, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, aos trinta dias do mês de junho do ano de dois mil e vinte e um.

TÍTULO SUGERIDO PELA BANCA

Recursos Genômicos de *Eugenia dysenterica* (Mart.) DC. - Genoma Cloroplastidial e Transcritoma de Referência

Documento assinado eletronicamente por **Alexandre Siqueira Guedes Coelho**, **Professor do Magistério Superior**, em 30/06/2021, às 13:01, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Tereza Cristina de Oliveira Borba**, **Usuário Externo**, em 30/06/2021, às 13:03, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Cintia Pelegrineti Targueta de Azevedo Brito**, **Usuário Externo**, em 30/06/2021, às 13:03, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Adriana Maria Antunes Taquary**, **Usuário Externo**, em 30/06/2021, às 13:05, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **ISABELA PAVANELLI DE SOUZA**, **Usuário Externo**, em 30/06/2021, às 15:24, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Mariana Pires De Campos Telles**, **Professor do Magistério Superior**, em 06/01/2022, às 13:49, 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 **2158904** e o código CRC **D2AB103F**.

Referência: Processo nº 23070.032663/2021-06 SEI nº 2158904

À minha filha, Maria Rita,
que ressignificou a minha vida,

DEDICO.

AGRADECIMENTOS

Este trabalho foi realizado no contexto do Grupo de Trabalho em Genética e Genômica do INCT EECBio (FAPEG/CNPQ) que gostaria de agradecer o apoio financeiro e logístico. Também agradeço à UFG pela estrutura e apoio prestados, e à CAPES pela concessão da bolsa de estudos.

Agradeço ao Programa de Pós-Graduação em Genética e Melhoramento de Plantas (PGMP/UFG), em especial aos coordenadores da atual gestão Dr. Sérgio Tadeu Sibov e Drª. Marcela Pedroso Mendes Resende, bem como aos secretários, pelo acompanhamento e assistência sempre feitos com excelência.

A todos os professores do PGMP, agradeço pela dedicação e trabalho excepcionais. Em especial, agradeço à Professora Rita Ganga, pelos momentos de aprendizado, descontração e carinho e por sempre transmitir sentimentos bons e uma energia extremamente leve.

Agradeço ao Professor Alexandre Siqueira, por ter me acolhido no LGGP, pela orientação e por esses nove anos de convivência, trabalho e aprendizado, essenciais à minha formação.

À Professora Mariana Telles pela coorientação e também pelo carinho, acolhimento ecompreensão, desde o primeiro momento em que nos vimos, e por ser para mim um exemplo pessoal e profissional.

Agradeço aos membros da minha banca de qualificação Dra. Adriana Maria Antunes e Dra. Thannya Nascimento Soares, pelas contribuições que sem dúvidas fizeram a diferença na finalização destre trabalho.

Aos membros titulares da banca examinadora, Dra. Adriana Maria Antunes, Dra. Cintia Targueta, Dra. Isabela Pavanelli e Dra. Tereza Borba pela disponibilidade e contribuições feitas para este trabalho. Também, agradeço aos membros suplentes, Dra. Ramila Braga e Dr. Rhewter Nunes por terem se disponibilizado a participar da banca.

À Ludmila Bandeira, por não ter me deixado desistir na única vez em que eu cogitei essa idéia, e por sempre estar ao meu lado durante todas as etapas de experimentação e análises laboratoriais.

À minha querida amiga Sâmella, a quem tive o prazer de conhecer durante o doutorado e que tem me ensinado muito sobre força, determinação e resiliência.

Às amigas Giorgia e Sabrina, que também tive o prazer de conhecer através do PGMP e que me inspiram como profissionais e, agora, como mães. O apoio de vocês fez toda a diferença na minha caminhada até aqui.

Agradeço especialmente ao querido amigo Ueric José, por fazer parte da minha história na UFG, desde a época em que eu nem imaginada que chegaria tão longe...obrigada por tantos momentos incríveis, e principalmente por ter sido essencial para a elucidação de boa parte deste trabalho.

Aos amigos Suzy, Guilherme, Mariane, Lucas e Kellen pelos momentos incríveis que passamos juntos, e por ocuparem lugares especiais em meu coração.

Às amigas de longa data e eternas, Vanuza, Mayane e Débora por todos os momentos vividos nestes mais de 10 anos de amizade, compartilhando alegrias, tristezas, vitórias e derrotas, sempre apoiando e alegrando umas às outras.

Aos meus amigos Ivone, Isabela e Rhewter por terem sido as pessoas que mais marcaram a minha formação no PGMP. Obrigada por acreditarem em mim, por me darem força, pela amizade, por todos os momentos vividos e pelos que ainda virão. Minhas melhores lembranças remetem a vocês.

Agradeço à querida amiga Lanusse, por ter confiado em mim, permitindo que eu vivesse a experiência que me trouxe certeza do meu amor pela docência.

Agradeço ao meu companheiro de vida, Jakson, por me proporcionar amor, compreensão, respeito e companheirismo, sem medidas. Por ser a pessoa que mais me apoiou durante o doutorado, fazendo inúmeros esforços para que eu mantivesse a calma e persistisse. Por ser a pessoa que mais me apoia em tudo, por enxergar o melhor de mim e principalmente, por me proporcionar uma família com a qual eu sempre sonhei.

Agradeço à minha pequena luz do sol, Maria Rita, que veio para me mostrar o que realmente é importante, e que a minha força como mulher é infinitamente maior do que poderia imaginar. Que logo ela possa ler esse texto e saber que todos os dias, os sorrisos e carinhos dela me fazem ainda mais forte e feliz para continuar a caminhada.

Agradeço à minha mãe Dona Raimunda, por ser meu alicerce, ao meu pai, Sr, João Batista e à minha irmã Stefany, por serem presentes em todos os meus momentos. Obrigada por me proporcionar o melhor de vocês.

Agradeço a Deus, pela minha saúde mental e física, mantidas ao longo de toda a minha jornada na pós-graduação e reforçadas na fase de conclusão deste trabalho, em que me senti privilegiada por conseguir, sendo uma sobrevivente em tempos tão difíceis.

SUMÁRIO

RESUMO

BARROS-RIBEIRO, S. **Recursos Genômicos de *Eugenia dysenterica* (Mart.) DC - Genoma cloroplastidial e transcritoma de referência**. 2021. 92 f. Tese (Doutorado em Genética e Melhoramento de Plantas) – Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2021.[1]

Eugenia dysenterica (Mart.) DC., popularmente conhecida como cagaiteira é uma espécie nativa do Cerrado brasileiro que apresenta diversos potenciais de utilização, seja como alimento, medicinal, madeira para construções ou ornamental. Também apresenta importância ecológica por fornecer alimento a espécies da fauna local. Diversos estudos voltados à caracterização da diversidade, estrutura genética e atuação de processos microevolutivos dentro das populações nativas e de uma coleção de germoplasma de *E. dysenterica* tem sido realizados nas últimas décadas, utilizando principalmente marcadores moleculares. A caracterização genômica de espécies vegetais nativas, como a cagaiteira, se constitui em um passo importante para o desenvolvimento de estratégias eficientes de conservação e melhoramento e isso tem se tornado possível graças aos constantes avanços das tecnologias de sequenciamento de genomas e ferramentas para analise de dados genômicos. O objetivo deste estudo foi realizar a montagem e a caracterização do genoma cloroplastidial, bem como obter um transcritoma de referência para *E. dysenterica*, utilizando dados de sequenciamento de alto rendimento. Para a montagem do plastoma, foram obtidas amostras de DNA a partir de folhas de indivíduos adultos, sequenciadas utilizando a plataforma Illumina MiSeq. Para o transcritoma, amostras de RNA total foram extraídas de folhas e plântulas e sequenciadas utilizando a plataforma Illumina HiSeq 4000. O genoma cloroplastidial de *E. dysenterica* apresenta 158.560 pb de tamanho e está organizado em estrutura quadripartida, comum para as plantas terrestres descritas na literatura. Foram identificados 112 diferentes genes cloroplastidiais, dos quais 78 codificam proteínas, 30 codificam tRNAs e 4 codificam rRNAs. Também foram identificadas 78 regiões SSR, das quais a maioria são mononucleotídeos com motivo de repetição A/T. A maioria dos SSR identificados estão localizados em regiões intergênicas da porção LSC do cloroplasto. A diversidade nucleotidica média estimada entre o plastoma de *E. dysenterica* e outras quatro espécies do mesmo gênero (*E. brasiliensis*, *E. selloi*, *E. pyriformis* e *E. uniflora*) foi de 0,0064, variando de 0,0000 a 0,0315. A taxa de substituição Ka/Ks média entre essas espécies comparadas foi de 0,1907. A análise filogenética confirmou que *E. dysenterica* está intimamente relacionada a outras espécies de Myrtaceae, com fortes valores de *bootstrap* (87,7% a 100%). A caracterização do transcritoma de *E. dysenterica*, utilizando a abordagem de RNAseq, é inédita dentre as espécies nativas do Cerrado e possibilitou a identificação de 171.070 transcritos, de 43.605 genes, anotados com base em diferentes bancos dados de referência. Além disso, também foram identificados 636.269 SNPs que deverão ser validados em futuros estudos. Os dados obtidos neste trabalho constituem importantes recursos genômicos para a espécie, podendo ser utilizados no desenvolvimento de estratégias eficientes para cultivo e conservação de *E. dysenterica*, além de servir como subsídio para estudos com outras espécies vegetais arbóreas não-modelo.

Palavras-chave: Cerrado, Myrtaceae, NGS, Plastoma, Transcritoma

[1] Orientador: Prof. Dr. Alexandre Siqueira Guedes Coelho – EA/UFG.
Coorientadora: Prof.ª Dr.ª Mariana Pires de Campos Telles – ICB/UFG.

ABSTRACT

BARROS-RIBEIRO, S. **Genomic Resources of *Eugenia dysenterica* (Mart.) DC - Chloroplast genome and reference transcriptome**. 2021. 92 l. Thesis (Doctor of Science in Genetics and Plant Breeding) – Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2021.[1]

Eugenia dysenterica (Mart.) DC., popularly known as cagaiteira, is a native species of the Brazilian Cerrado that has several potential uses, whether as food, medicinal, wood for construction or ornamental. It also has ecological importance for providing food for species of local fauna. Several studies had already characterized the diversity, genetic structure and performance of microevolutionary processes within native populations and a collection of *E. dysenterica* germplasm have been carried out in recent decades, using mainly molecular markers. The genomic characterization of native plant species, such as cagaiteira, is an important step towards the development of efficient conservation and plant breeding strategies and this has been made possible thanks to the constant advances in sequencing technologies and tools for data analysis. The objective of this study was to assemble and characterize the chloroplast genome, as well as obtain a reference transcriptome for *E. dysenterica*, using high-throughput sequencing data. For plastome assembly, DNA samples were obtained from adult individual leaves, sequenced using the Illumina MiSeq platform. For the transcriptome assembly, total RNA samples were extracted from leaves and seedlings and sequenced using the Illumina HiSeq 4000 platform. The chloroplast genome of *E. dysenterica* is 158,560 bp in size and is organized in a quadripartite structure, common for terrestrial plants described in the literature. The chloroplast genome of *E. dysenterica* contains a total of 112 different genes, including 78 protein-coding genes, 30 transfer RNA genes, and 4 ribosomal RNA genes. A total of 78 SSR regions were identified, of which the majority are mononucleotides with an A/T repeat motif. Most of the identified SSR are located in intergenic regions of the LSC portion of the chloroplast. The estimated mean nucleotide diversity among the *E. dysenterica* plastome and four other species of the same genus was 0.0064, ranging from 0.0000 to 0.0315. The Ka/Ks substitution rate among these compared species was 0.1907. Phylogenetic analysis confirmed that *E. dysenterica* is closely related to other *Myrtaceae* species, with strong bootstrap values (87.7% to 100%). The characterization of the *E. dysenterica* transcriptome, using the RNAseq approach, is unprecedented among native Cerrado species and allowed the identification of 171,070 transcripts from 43,605 genes, annotated based on different reference databases. In addition, 636,269 putative SNPs were also identified that should be validated in future studies. The information generated in this study constitute important genomic resources for the species, which can be used in the development of efficient strategies for cultivation and conservation of E. dysenterica, in addition to serving as a subsidy for studies with other non-model tree plant species.

Keywords: Cerrado, Myrtaceae, NGS, Plastome, Transcritome

[1] Advisor: Prof. Dr. Alexandre Siqueira Guedes Coelho – EA/UFG.
Co-Advidor : Prof.ª Dr.ª Mariana Pires de Campos Telles – ICB/UFG.

1 INTRODUÇÃO

A família *Myrtaceae* abrange mais de 6.000 espécies de arbustos e árvores, com ocorrência predominante na América do Sul. As espécies são classificadas em 144 gêneros, dos quais alguns têm importância econômica, como *Eucalyptus*, *Melaleuca*, *Leptospermum*, *Syzygium* e *Psidium* (Thornhill et al., 2015). A importância econômica destes gêneros se deve ao aproveitamento da madeira, à produção de frutos consumidos por seres humanos e pela capacidade que apresentam de produzir diversos compostos de interesse farmacológico (antifúngicos, antioxidantes e anti-inflamatórios). Além disso, várias espécies desempenham funções ecológicas em suas áreas de ocorrência por servirem de forragem e alimentos para animais (Guzman et al., 2014; Rodrigues et al., 2020; WCSP, 2021).

Graças às suas diversas aplicações, o interesse na utilização de espécies de *Myrtaceae* em estudos evolutivos e ecológicos tem crescido, fato que pode ser observado com o aumento do número de sequências genômicas publicadas e depositadas em banco de dados públicos (NCBI, 2021). As sequências incluem genomas nucleares (Myburg et al., 2014; Izuno et al., 2016; Thrimawithana et al., 2019), genomas cloroplastidiais (Bayly et al., 2013; Eguiluz et al., 2017; Machado et al., 2017; Schuster et al., 2018; Rodrigues et al., 2020; Zhang et al., 2021) e transcritomas (Guzman et al., 2014; Tobias et al., 2018; Soewarto et al., 2019; Veto et al., 2020).

Apesar do aumento na quantidade de informações, ao considerarmos a diversidade encontrada em *Myrtaceae*, para a maioria das espécies ainda existe pouca ou nenhuma informação genética disponível. Estas espécies de plantas, também chamadas de não-modelo, constituem parte dos recursos genéticos vegetais que por sua vez fornecem matéria prima para estudos de conservação e melhoramento de plantas (Nass et al., 2012; Unamba et al., 2015). A enorme quantidade de dados obtidos com tecnologias de sequenciamento de nova geração (*Next Generation Sequencing* – NGS), associada à utilização de ferramentas de bioinformática para sua análise e interpretação, tem facilitado o desenvolvimento de estudos genômicos em plantas nativas não-modelo, muitas das quais possuem diversas aplicações, têm importância ecológica e vêm sofrendo risco de extinção (Alonso-Herrada et al., 2016; Basanti et al., 2017).

A espécie *Eugenia dysenterica* (Mart.) DC., popularmente conhecida como cagaiteira, pertence à família *Myrtaceae*. Trata-se de uma árvore endêmica do Cerrado brasileiro, considerada uma das espécies prioritárias para conservação e usos atual e potencial, devido às suas diversas possibilidades de utilização (Vieira et al., 2016). Assim como diversas outras espécies nativas da região, a cagaiteira é utilizada pelas populações locais como fonte de alimento e renda, principalmente devido à utilização de seus frutos *in natura* ou processados. A espécie também é utilizada para extração de cortiça e ornamentação, devido à sua beleza, realçada principalmente no período de floração. Além disso, a cagaiteira é uma espécie melífera e constitui fonte de alimento para vários animais da fauna local, o que reforça sua importância ecológica. Seus compostos químicos, como vitaminas A e C, folatos e terpenos, vêm sendo estudados e demonstram diversas propriedades medicinais, como antifúngica e antidiarreica (Cardoso et al., 2011; Souza et al., 2018).

Estudos acerca da variabilidade fenotípica, da estrutura genética, do sistema de cruzamento e história demográfica em populações naturais de *E. dysenterica* tem sido desenvolvidos nas últimas décadas. Estes estudos contam, principalmente com análises de caracteres quantitativos e marcadores moleculares, como aloenzimas e microssatélites (SSRs) além de sequências de regiões específicas de DNAs nucleares e cloroplastidiais (Rodrigues et al., 2016; Lima et al., 2017; Boaventura-Novaes et al., 2018 a, b). Dados de NGS, obtidos com a plataforma Illumina MiSeq foram utilizados por Barros-Ribeiro (2016) e Nunes et al. (2015) para obtenção de um *draft-assembly* do genoma de *E. dysenterica*. A partir desses dados foi feita uma caracterização parcial do genoma, com relação à quantidade e estrutura de genes putativos, abundância de elementos repetitivos e identificação de SNPs.

Os estudos já realizados constituem ferramentas de grande importância para *E. dysenterica*, pois permitiram avanços no conhecimento acerca da variabilidade genética e dos processos microevolutivos que atuam sobre as poluções naturais da espécie, porém, diversas outras questões biológicas ainda precisam ser respondidas para que seja possível desenvolver estratégias eficientes de manejo, conservação e melhoramento genético da espécie. Assim, a integração entre diferentes tipos de dados obtidos a partir das ciências ômicas pode auxiliar no estabelecimento de modelos fundamentais para se entender a evolução, o desenvolvimento e a adaptabilidade de diversos organismos (Antunes et al., 2021b).

Além do DNA nuclear, a caracterização genômica de uma espécie vegetal também pode ser realizada a partir de DNAs cloroplastidiais. Esse material genético possui características como pouca ou nenhuma recombinação, baixas taxas de substituições e herança predominantemente uniparental, tornando-se uma valiosa fonte para estudos de filogenia, filogeografia, genética de populações e identificação de espécies. As publicações baseadas em genomas de cloroplasto têm aumentado, devido ao constante desenvolvimento das plataformas de sequenciamento e ferramentas de bioinformática, contribuindo para a realização desse tipo de estudo dentro da família *Myrtaceae* e com outras espécies nativas do Cerrado (Souza et al., 2019; Antunes et al., 2020a; Nunes et al., 2020, Rodrigues et al., 2020).

Uma outra estratégia de análise utilizada para caracterizar genomas vegetais é a transcritômica, que consiste no sequenciamento e análise de todos os RNAs presentes em uma amostra biológica. A aplicação das tecnologias de NGS à transcritômica, comumente chamada de RNAseq, possibilita a obtenção das sequências expressas em um determinado tecido e condição da planta com elevada sensibilidade, devido à grande profundidade de sequenciamento tipicamente utilizada nestes estudos. Os dados obtidos com essa estratégia permitem identificar novos genes e transcritos, quantificar níveis de expressão gênica e identificar SNPs. Assim, a transcritômica pode ser utilizada tanto para se iniciar a caracterização genômica de uma espécie para a qual ainda não se tem informações disponíveis, quanto para se complementar as informações obtidas a partir de montagens parciais de genomas (Wang et al., 2009; Strickler et al., 2012; Mosa et al., 2017).

Neste contexto, o objetivo deste trabalho foi a geração de recursos genômicos para *E. dysenterica,* através da caracterização do seu genoma cloroplastidial e da obtenção de um transcritoma de referência para a espécie.

2 REVISÃO DE LITERATURA

2.1 CARACTERÍSTICAS GERAIS DOS GENOMAS E TRANSCRITOMAS DE ESPÉCIES VEGETAIS

Parte da diversidade encontrada nos genomas dos vegetais superiores se deve à variação em caracteres como o tamanho do genoma e o número de cromossomos (Figura 1). Considerando o grupo das angiospermas, que compreende a maior diversidade de espécies do reino Viridiplantae, os tamanhos dos genomas variam de 0,063 a 148,8 Gb. O número de cromossomos também apresenta ampla variação, sendo atualmente o menor número observado em *Brachyscome dichromosomatica* (2n = 4) e o maior número observado em *Ophioglossum reticulatum* (2n = 1.400) (Wang et al., 2015; Stuessy et al., 2019; Antunes et al., 2020c).

O aumento do número de cromossomos em vegetais pode ser relacionado a eventos como a poliploidização, que envolve a duplicação de genes, cromossomos ou de genomas como um todo (evento conhecido como *Whole Genome Duplication* – WGS). Esses eventos acometem plantas com muita frequência, sendo muitas vezes relacionados à evolução da diversidade morfológica e fisiológica desses organismos. Como consequências imediatas da poliploidização, são comuns alterações na sequência de DNA, como o silenciamento e a perda de genes; a duplicação ou perda de fragmentos cromossômicos; a expressão tendenciosa de genes homólogos e a ativação de elementos transponíveis (TEs). Evolutivamente, a poliploidia é seguida por uma perda massiva de sequências redundantes, de modo que um estado semelhante a um pseudo-diploide pode ser restaurado nos organismos (Figura 2) (Jiao et al., 2011; Wendel et al., 2018; Qiao et al., 2019; Stuessy et al., 2019).

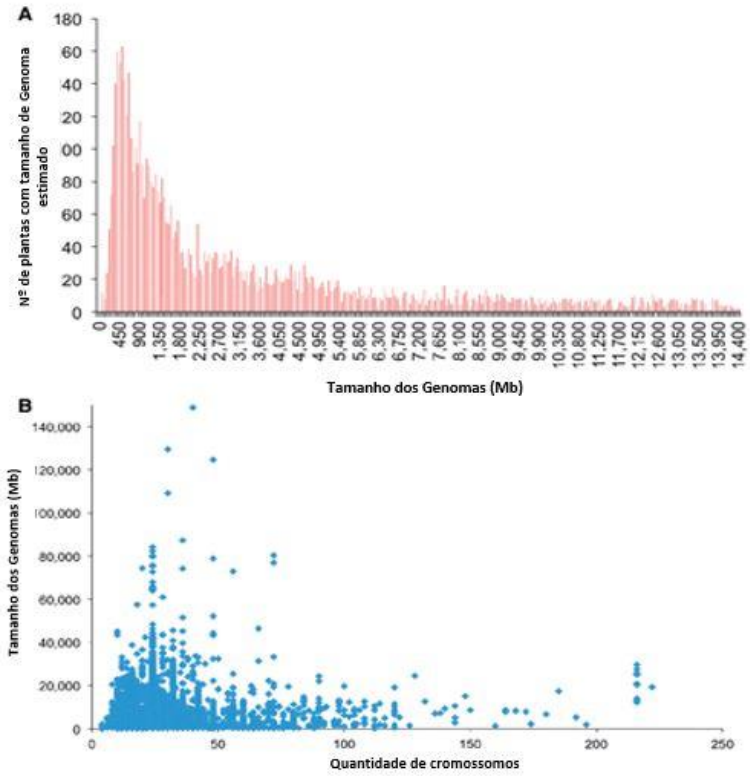

Figura 1. (A) Distribuição dos tamanhos dos genomas de plantas segundo informações obtidas no *Kew Plant DNA C-values Database.* Pode-se observar que a maioria das espécies cujos dados genômicos estão disponíveis no banco de dados tem seus genomas com tamanhos menores que 20.000 Mb. (B) Relação entre o número de cromossomos e o tamanho dos genomas de plantas (Mb). Observa-se que genomas maiores não estão associados a um maior número de cromossomos. Adaptado de Michael (2014).

Outro aspecto relacionado à diversidade observada nos genomas de plantas é a presença de grandes porções de elementos repetitivos. Atualmente, sabe-se que essas sequências de DNA compreendem a maior parte dos genomas de plantas (entre 20 e 85%, em espécies como *Arabidopsis thaliana* e *Helianthus annuus*, respectivamente) e influenciam o tamanho, o metabolismo, a organização e a evolução desses genomas. Por muito tempo as funções destes elementos foram subestimadas, mas graças ao desenvolvimento de estudos de caracterização genômica, foi possível observar que eles podem afetar a diversidade genética, a duplicação de genes e a estabilidade do genoma, além de serem os principais constituintes das regiões centroméricas e teloméricas dos cromossomos eucarióticos (Pisupati et al., 2018; Sahebi et al., 2018; Antunes et al., 2020b).

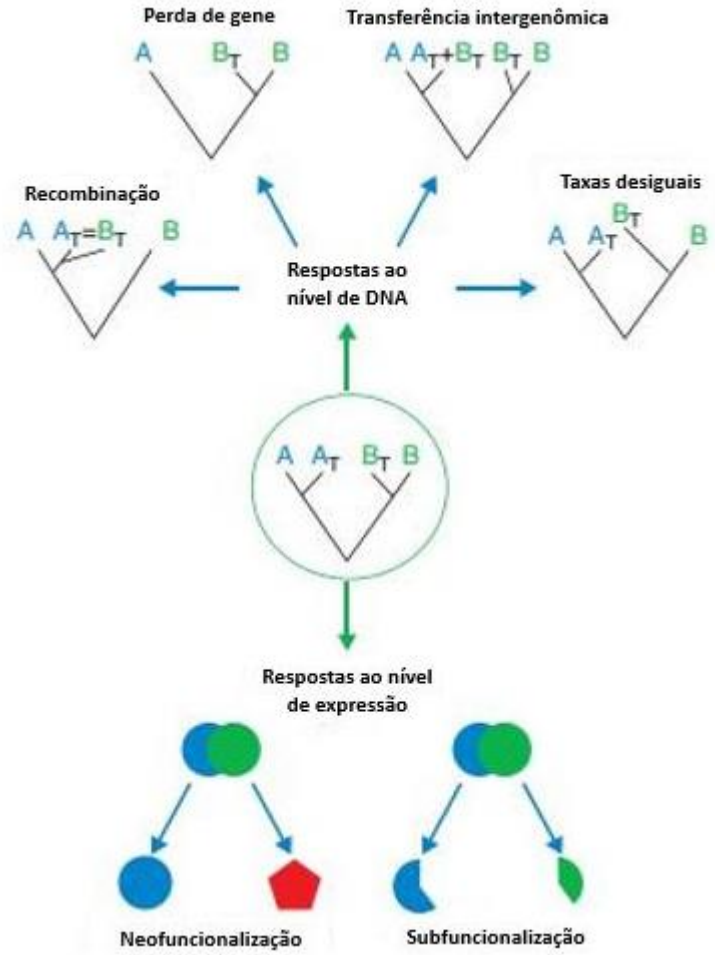

Figura 2. A figura ilustra um genoma alopoliploide hipotético derivado de dois genomas diploides progenitores (A e B). A hibridização e a duplicação destes genomas desencadeiam processos que têm efeitos a curto e longo prazo, tanto ao nível da sequência de DNA (parte superior), quanto ao nível da expressão gênica (parte inferior), respectivamente. Adaptado de Wendel et al. (2016).

Os elementos repetitivos do DNA podem ser classificados em dois grandes grupos com base na forma com que eles se organizam nos genomas. O primeiro grupo inclui as repetições em tandem, cuja principal característica é o fato das cópias estarem dispostas sequencialmente, uma ao lado da outra, sendo encontrado em todo o genoma, com maiores concentrações observadas em regiões de centrômeros e telômeros. Esse grupo é representado por DNAs satélites (satDNA), sequências de repetições simples (SSR) e de DNAs ribossômicos (rDNA) (Bagshaw, 2017; Pisupati et al., 2018).

Os satDNAs consistem em repetições não codificantes com tamanhos que ultrapassam 100 nucleotídeos. Estes elementos podem representar cerca de 30% do DNA de algumas plantas e são os principais componentes da heterocromatina. Já os rDNAs constituem regiões codificantes presentes em elevado número de cópias para atender à alta

demanda celular por ribossomos (Garcia et al., 2014; Mehrotra & Goyal, 2014; Garrido-Ramos, 2015).

As repetições SSR são compostas por motivos de até 6 pb e sequências que pode variar de 10 a 100 unidades de repetição, sendo encontrados com frequências altamente variáveis em diferentes espécies, tanto em regiões codificantes quanto não codificantes. Estas repetições desempenham um papel importante na evolução dos genomas por apresentarem altas taxas de mutação, que resultam em diferenças no número de unidades de repetição. As elevadas taxas de mutação das repetições SSR podem ser explicadas por eventos como erros durante processos de recombinação (*crossing-over* desigual), o deslizamento (*slippage*) da DNA polimerase durante o processo de replicação e erros no mecanismo de reparo do DNA (Amos et al., 2008; Bhargava & Fuentes, 2010; Bagshaw, 2017).

Essas repetições também apresentam como característica a presença de altos níveis de polimorfismo, graças às diferenças de tamanho destes fragmentos, causadas pelo número variável de unidades de repetição. O polimorfismo encontrado nas regiões SSR faz destes locos ferramentas amplamente utilizadas no desenvolvimento de marcadores moleculares aplicados a estudos de mapeamento genético, genética de populações e biologia da conservação para diversos organismos (Amos et al., 2008; Bhargava & Fuentes, 2010; Bagshaw, 2017).

O segundo grupo de DNAs repetitivos engloba os elementos transponíveis (TEs) que se encontram dispersos por todo o genoma. Este grupo é subdividido em duas classes: a Classe I, dos retrotransposons; e a Classe II, dos transposons de DNA, que se diferenciam pelo mecanismo de transposição dos seus elementos. Os retrotransposons utilizam um RNA intermediário cujo transcrito é utilizado como molde para um novo DNA, produzido através da atividade de uma transcriptase reversa, processo também conhecido como copia-e-cola (*copy-and-paste*). Já os transposons de DNA tipicamente utilizam uma transposase para excisar um DNA e inseri-lo em outros locais, processo também conhecido como recorta-e-cola (*cut-and-paste*). Esta classe inclui ainda os hélitrons, que têm um mecanismo diferenciado de transposição, conhecido como círculo-rolante (Pisupati et al., 2018; Sahebi et al., 2018; Antunes et al., 2020b).

Devido à sua capacidade de se movimentar dentro dos genomas, os TEs podem apresentar diversos efeitos, como a duplicação, deleção, mudanças na expressão ou função de genes, podendo também alterar o tamanho dos genomas. Esses efeitos vão depender tanto

estrutura quanto do local onde essa repetição está inserida. Os TEs são geralmente silenciados durante o desenvolvimento da planta e sua atividade é muitas vezes considerada como uma resposta adaptativa do genoma diante de situações de estresse biótico ou abiótico, o que sugere que eles também podem desempenhar um papel benéfico na evolução desses genomas. Algumas das alterações causadas pelos TEs podem contribuir para o surgimento de novos fenótipos, muitos dos quais acabaram se tornando atrativos à utilização humana, como em algumas variedades de uva e tomate (Lisch et al., 2013; Krasileva, 2019).

A informação genética das plantas pode ser encontrada tanto no núcleo das células quanto em organelas (mitocôndrias e cloroplastos), sendo possível ainda encontrar fragmentos de DNAs organelares inseridos de forma independente no DNA nuclear desses organismos. A organização das regiões gênicas também se constitui em uma característica importante dos genomas de plantas. De acordo com a literatura, essas estruturas se apresentam bastante compactadas e distribuídas mais ou menos ao acaso, em quantidades variando entre 20.000 e 50.000 genes por genoma, com densidade variando entre 1 e 38 genes para cada 100 kb, em espécies como *Arabidopsis*, arroz, sorgo, uva e melão (Garcia-Mas et al., 2012; Alonso-Herrada et al., 2016; McCormick et al., 2018, Flavell., 2021).

Evolutivamente, alguns genes podem ser duplicados em situações de erro durante a replicação ou recombinações do material genético, o que resulta, após milhares de anos, em famílias gênicas, que podem ter tamanhos variáveis entre espécies e desempenhar funções importantes relacionadas à adaptação ou à especiação (Martinez, 2011; Lópes-Flores, 2012; Guo, 2013). Os genes e suas localizações genômicas demonstram alta conservação de função e colinearidade, porém, novos genes sempre surgem a partir de mutações, duplicações, permutas e eventos de transferência horizontal. Ao longo do processo evolutivo, todos esses processos deixam evidências de seu acontecimento, através de traços característicos que podem ser detectados na sequência de DNA (Alberts et al., 2010; Hou et al., 2019).

Graças ao desenvolvimento de estudos de caracterização genômica, sabe-se que uma grande fração dos genes de eucariotos é transcrita, porém a grande maioria desses transcritos (~75%) não codificam proteínas, são os chamados RNAs não-codificantes (ncRNAs). Contudo, estes RNAs, que incluem diferentes classes como ribossômicos (rRNA); transportadores (tRNA); pequenos RNAs nucleares (snRNA) e microRNAs, podem desempenhar papeis importantes nos genomas de plantas, relacionados ao seu

desenvolvimento, adaptação e respostas a condições de estresse (Deng et al., 2017; Hou et al., 2019).

Considerando o material genético ao nível organelar, uma grande quantidade de genomas cloroplastidiais já foi caracterizada em espécies modelo, cultivadas, medicinais e em espécies não-modelo. Estes genomas exibem uma estrutura geral altamente conservada: estrutura circular com tamanho que varia entre 118 e 180 kb, organizado em quatro porções, incluindo duas cópias de uma repetição invertida (IR), separadas por duas regiões de cópia única, sendo uma região grande (LSC) e uma região pequena (SSC) (Sablock et al., 2016; Shen et al., 2017b; Guyeux et al., 2019).

A organização das regiões LSC e SSC é citada pelos cientistas como um processo dinâmico com muitos relatos de ocorrência de expansão e contração. Por outro lado, a organização das IRs é menos dinâmica quando comparada às demais, apresentando maior conservação entre as angiospermas (Sablock et al., 2016, Antunes et al., 2020a). A maioria dos genes de cloroplasto, assim como os genes nucleares, também se apresentam conservados em número e estrutura (em geral observa-se uma variação de 70 a 99 genes codificadores de proteínas), sendo as poucas variações observadas com maior frequência em regiões intergênicas (Xu et al., 2015).

Outra característica dos genomas cloroplastidiais são as baixas taxas de mutação entre os seus genes, provavelmente devido a características como o tipo de herança, os eficientes mecanismos de reparo e a baixa ocorrência de fusão e fissão dos plastídios, eventos observados com maior frequência em genomas nucleares e mitocondriais. Tais características fazem dos cloroplastos ferramentas muito úteis para estudos filogenéticos e evolutivos em plantas, sendo suas regiões gênicas utilizadas para o desenvolvimento de marcadores moleculares principalmente devido à sua conservação (Wicke et al., 2011; Sablock et al., 2016).

A caracterização preliminar dos genomas das plantas se constitui em um passo importante dos estudos na área de Biologia Vegetal, especialmente no contexto das chamadas “ciências ômicas”, pois permite aos pesquisadores avaliar a viabilidade financeira e computacional dos projetos de sequenciamento completo destes genomas. A análise de dados genômicos e sua integração a sistemas biológicos como um todo auxilia no estabelecimento de modelos fundamentais para se entender a evolução, o desenvolvimento e a adaptabilidade desses organismos (Antunes et al., 2021b).

Uma das abordagens utilizadas na caracterização de genomas é a transcritômica, na qual estuda-se o conjunto de todos os RNAs de uma amostra biológica. A aplicação das tecnologias de sequenciamento de nova geração (NGS) à transcritômica, comumente chamada de RNAseq possibilita a obtenção das sequências expressas em um determinado tecido e condição da planta, devido à grande profundidade do sequenciamento. Os dados obtidos podem ser usados para se identificar novos genes e transcritos, quantificar níveis de expressão gênica, identificar SNPs e *splicings* alternativos, informações que podem complementar estudos de montagem e anotação de genomas (Wang et al., 2009; Strickler et al., 2012; Mosa et al., 2017).

As informações baseadas em RNAseq também são úteis para estudos de espécies não-modelo, que são aquelas para as quais há pouca ou nenhuma informação genética disponível, uma vez que o foco do sequenciamento nesse caso é restrito às regiões codificantes. Isso pode diminuir os custos e tornar a montagem do transcritoma mais fácil do que uma montagem de genoma, devido à menor ocorrência de elementos repetitivos nessas regiões (Strickler et al., 2012).

2.2 CONSIDERAÇÕES GERAIS SOBRE A OBTENÇÃO DE DADOS GENÔMICOS PARA ESPÉCIES VEGETAIS

2.2.1 Repositórios de dados genômicos de plantas

O compartilhamento de dados genômicos com a comunidade científica é de extrema importância, pois permite que pesquisadores do mundo todo tenham conhecimento do que está sendo feito, ampliando as possibilidades de estudo e colaboração. Com o rápido desenvolvimento da Bioinformática, os bancos de dados de sequências genômicas de plantas evoluíram de simples plataformas de armazenamento para plataformas que disponibilizam todo um conjunto de ferramentas de análise, auxiliando em projetos de sequenciamento e ressequenciamento do genoma de diversas espécies (Chen et al., 2018).

A plataforma *Phytozome* (https://phytozome-next.jgi.doe.gov), mantida pelo Departamento de Energia (DOE) dos EUA, disponibiliza (a partir da versão 12.1.6) 93 genomas montados e anotados de 82 espécies de Viridiplantae (dados coletados até junho de 2021). Além das sequências, as ferramentas acopladas ao *Phytozome* possibilitam que o usuário anote famílias de genes de plantas e estude a evolução dessas famílias (Goodstein et al., 2012; Chen et al., 2018).

O projeto *10.000 Plant Genomes*, ou 10KP (https://db.cngb.org/10kp/), é um consórcio que tem como objetivo sequenciar o genoma de mais de 10.000 espécies de plantas e algas até o ano de 2022 e está sendo realizado pelo BGI (*Beijing Genomics Institute*) em parceria com o *China National GeneBank* (CNGB) (Cheng et al., 2018). Até o momento da coleta destas informações (junho de 2021), o banco inclui dados de DNA cloroplastidial e nuclear, reunindo mais de 1.000 amostras. Já o *Open Green Genomes Initiative* (OGG), financiado pelo *Joint Genome Institute* (https://jgi.doe.gov/), se concentra em gerar dados para 35 espécies representem todas as principais linhagens evolutivas das plantas terrestres. Nesse banco de dados estão disponíveis sequências de genomas e transcritomas de espécies de diferentes famílias, incluindo *Myrtaceae*, e gêneros de interesse econômico como *Eucalyptus* e *Corymbia* (Li & Harkess, 2018).

Considerando plantas medicinais, o projeto *Medicinal Plant Genomics* (http://medicinalplantgenomics.msu.edu/species_list.shtml) já disponibilizou sequências genômicas de 14 espécies, incluindo dados de transcritomas. Este banco de dados tem o objetivo de fornecer recursos para a comunidade que trabalha com metabólitos produzidos pelas plantas e suas aplicações à saúde humana.

Outra ferramente interessante é o banco de dados *plaBiPD* (https://www.plabipd.de/). A partir dele é possível acessar artigos de genomas de plantas já publicados e projetos de sequenciamento que estão em andamento, sendo as informações organizadas em linha do tempo ou por filogenia. Diversos outros bancos de dados estão disponíveis com atualizações constantes, tanto para dados genômicos quanto para informações sobre proteínas e vias metabólicas, observando-se uma tendência ao desenvolvimento e realização de estudos mais abrangentes, incluindo, por exemplo, análises de genômica comparativa (Chen et al., 2018).

2.2.2 Integridade e sequenciamento das amostras

A preparação cuidadosa das amostras a serem sequenciadas é de extrema importância, pois a quantidade e integridade do material genético estão diretamente relacionadas à qualidade da montagem dos genomas e transcritomas. A integridade do DNA ou RNA pode ser visualizada por eletroforese em gel de agarose convencional ou utilizando equipamentos como Bioanalyzer, sendo esta segunda opção uma forma de se obter estimativas mais precisas acerca da distribuição dos tamanhos dos fragmentos nas amostras (Agilent, 2021).

A pureza do material genético também deve ser alta, para se evitar a presença de contaminações que atrapalhem a preparação das bibliotecas de sequenciamento. Neste contexto, espectrofotômetros, como o NanoDrop (Thermo Fisher Scientific, 2021a), são amplamente utilizados, sendo importante se observar que a razão entre os valores de absorbância a 260 nm e 280 nm (A_{260}/A_{280}) das amostras de DNA ou RNA purificados devem estar entre 1,8 e 2,0, enquanto aqueles da razão A_{260}/A_{230} devem estar entre 2,0 e 2,2. A quantificação, por sua vez, deve ser realizada utilizando-se equipamentos baseados na mensuração de fluorescência, como o Qubit (Thermo Fisher Scientific, 2021b) (Li & Harkess, 2018).

As plataformas de sequenciamento de nova geração (NGS) da Illumina, até o momento, continuam sendo as mais utilizadas para o sequenciamento de genomas e transcritomas de plantas. Elas são capazes de produzir um grande volume de dados, com baixo custo e com uma baixa taxa de erros. Por exemplo, atualmente, a análise de uma canaleta de sequenciamento coma plataforma Illumina HiSeq4000 tem custo aproximado de US$ 2.800 e produz de 250 a 400 milhões de *reads* de 150 nucleotídeos, no modo de sequenciamento de ambas as extremidades dos fragmentos (*paired-ends*), totalizando de 75 a 120 bilhões de bases (Gb) sequenciadas. Levando-se em consideração uma planta cujo genoma tenha cerca de 1 Gb de tamanho, este volume de dados de sequenciamento corresponde a uma cobertura de até 120X, em uma única canaleta, o que permite uma montagem de qualidade (Illumina, 2021).

A principal desvantagem do uso das plataformas Illumina continua sendo o tamanho dos *reads*, que atingem no máximo 250 nucleotídeos nas plataformas NovaSeq e 300 nucleotídeos na plataforma MiSeq. O sequenciamento Oxford Nanopore e o sequenciamento em Tempo Real de Molécula Única (SMRT) PacBio são atualmente as plataformas mais populares no mercado para produção de *reads* longos. Ambas geram sequências com dezenas a centenas de milhares de pares de base, porém apresentam uma taxa de erro relativamente alta quando comparada àquela tipicamente obtida em dados Illumina (Li et al., 2017).Uma abordagem que vem sendo bastante utilizada consiste na utilização de montagens híbridas, em que são utilizados dados obtidos a partir de diferentes plataformas de sequenciamento para se montar a sequência genômica com maior contiguidade e precisão (Zimin et al., 2017; Kyriakidou et al., 2018; Xing et al., 2019).

O tamanho dos genomas pode ser estimado antes da montagem por meio das técnicas de citometria de fluxo ou utilizando-se a análise da distribuição de *k*-meros, feita

com o auxílio de ferramentas de Bioinformática. A partir dos *reads* obtidos pelo sequenciamento é possível se estimar o volume de dados em regiões de cópia única e então dividi-lo pela cobertura estimada de sequenciamento. Por este método, para obtenção de estimativas acuradas, recomenda-se utilizar uma cobertura média de pelo menos 30X (Li & Harkess, 2018; Antunes et al., 2020a).

A análise da distribuição de frequências dos *k*-meros também pode ser utilizada para estimar o nível de heterozigosidade e a fração do genoma correspondente às regiões repetitivas. A diferença dos valores de cobertura de sequenciamento de regiões heterozigóticas e de regiões homozigóticas do genoma é evidenciada graficamente através da presença de um pico intermediário formado próximo ao pico que indica a cobertura média de sequenciamento do genoma (Figura 3). Os *k*-meros de regiões repetitivas, por outro lado, serão representados em um pico maior do que aquele que indica cobertura média de sequenciamento, assim como demonstrado na Figura 3 (Li & Harkess, 2018).

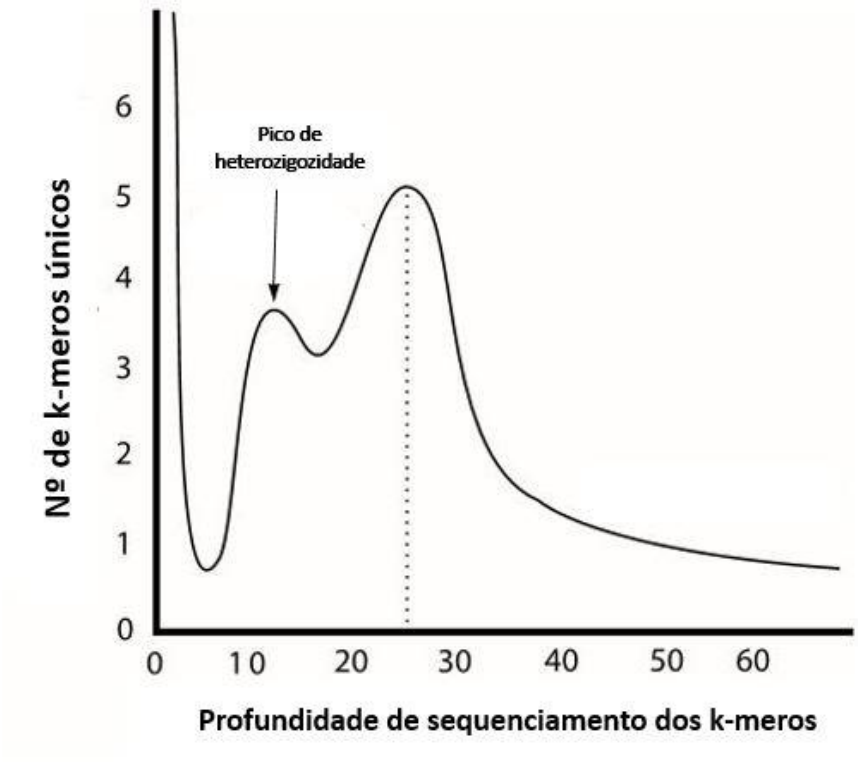

Figura 3. Distribuição de frequência de *k*-meros evidenciando a cobertura (profundidade) de sequenciamento das regiões homozigóticas (pico mais alto) e a cobertura de sequenciamento das regiões heterozigóticas do genoma (pico intermediário – mais baixo). Adaptado de Li & Harkess (2018).

2.2.4 Montagem e anotação de genomas e transcritomas

Desde a publicação do primeiro *assembly* de um genoma nuclear vegetal (The Arabidopsis Genome Initiative, 2000), tanto as tecnologias de sequenciamento quanto os recursos computacionais para a montagem de genomas de plantas vêm evoluindo de forma acelerada, desencadeando uma explosão de recursos para as ciências das plantas. Houve uma redução do tempo de obtenção dos dados de sequenciamento (os dados do primeiro projeto foram obtidos ao longo de dez anos, enquanto hoje seriam obtidos em uma semana). Os custos também têm sido reduzidos (o primeiro projeto teve um custo aproximado de US$ 100 milhões e se o mesmo fosse realizado atualmente, o custo aproximado seria de US$ 1.000). Tais fatores têm possibilitado que até mesmo os pequenos laboratórios realizem esse tipo de pesquisa, aumentando a quantidade de informações genômicas para espécies não-modelo, nativas de diferentes regiões do planeta (Alonso-Herrada et al., 2016; Bolger et al., 2017; Jiao & Schneeberger, 2017).

Considerando-se as espécies vegetais não-modelo, a abordagem mais utilizada para obtenção de sequências de genomas e transcritomas é o emprego do método conhecido por *assembly de novo*, que consiste em uma montagem sem utilizar uma sequência de referência. Entretanto, essa técnica pode resultar em montagens e anotações fragmentadas, pois exige grandes conjuntos de dados e alta capacidade de análises computacionais. Vários programas de montagem *de novo* como ABySS, SPAdes (para montagem de genomas), Velvet e Trinity (para montagem de transcritomas) têm sido utilizados para espécies como *Larix sibirica* (12,34 Gb), *Picea glauca* (20 Gb) e *Acer platanoides* (Birol et al., 2013; Kuzmin et al., 2019; Madritsch et al., 2021).

Cumpre ressaltar que as características típicas dos genomas de plantas citadas até aqui representam um desafio para a etapa de montagem. Elevados níveis de ploidia, por exemplo, implicam na presença de haplótipos adicionais, tanto em espécies alopoliploides, quanto em espécies autopoliploides, o que dificulta a montagem dessas regiões. A heterozigosidade dos genomas também constitui um desafio para os projetos de montagem, pois regiões heterozigóticas complicam a estrutura dos grafos produzidos durante o *assembly* e tornam a fase de obtenção dos haplótipos de mais difícil resolução resultando em um número excessivo de *contigs*. Optar pela utilização de tecnologias de sequenciamento que permitam a obtenção de *reads* longos é uma das formas de se minimizar esse problema (Olsson et al., 2016; Li et al., 2018).

A presença de altas taxas de elementos repetitivos também dificulta a montagem de genomas (Li & Harkess, 2018). No caso dos TEs, já foi observado que o comprimento de alguns desses elementos pode ultrapassar 10 kb, o que pode fragmentar a montagem fragmentada. A caracterização das regiões repetitivas em genomas vegetais pode ser realizada antes da montage, utilizando programas como o RepeatExplorer (Novák et al., 2013) ou após a montagem como programas como o RepeatMasker (Hubley & Smit, 2021), que mascaram essas repetições de forma a facilitar a montagem e identificação de genes (Li & Harkess, 2018; Kalendar et al., 2018).

Após a obtenção do *assembly*, torna-se necessária a anotação estrutural dos genes obtidos, o que inclui informações como tamanho médio dos genes e dos transcritos, a densidade de genes no genoma, quantidade de transcritos por gene, quantidade de éxons e íntrons, bem como tamanho médio dessas regiões (Pisupati et al., 2018, Sahebi et al., 2018).

A anotação estrutural em organismos eucariotos de forma geral não é um processo fácil, principalmente devido à complexidade da estrutura dos genes e a grande proporção de elementos repetitivos nesses genomas, contudo é possível melhorar a anotação através de montagens parciais de genomas, alinhando a elas dados obtidos pelo sequenciamento de transcritomas (Olsson et al., 2016; Bolger et al., 2017). Outras informações necessárias à anotação estrutural incluem a análise do conteúdo GC, a análise de similaridade e divergência com outras espécies e a identificação de possíveis marcadores moleculares (SSR e SNPs) (Garg & Jain, 2013). As ferramentas computacionais Augustus e MAKER-P (Stanke & Morgenstern, 2005; Campbell et al., 2014) são bastante utilizadas para anotação estrutural de genes em plantas e fazem uso de modelos de predição, além dos próprios dados de sequências de transcritos e de proteínas que podem ser adicionados durante o processo para se identificar estruturas gênicas (Campbell et al., 2014; Barros-Ribeiro, 2016; Hoff et al., 2016).

O passo posterior à anotação estrutural consiste em atribuir função biológica aos genes por meio da anotação funcional. Apesar da existência de grandes bancos de dados em que os pesquisadores têm depositado inúmeras informações sobre as funções de genes de plantas, essa etapa ainda é um desafio devido à grande porcentagem de genes não conservados em diversas espécies, cujas funções ainda não foram determinadas (Bolger et al., 2017).

O banco de dados mais utilizado para se realizar a anotação funcional de genes é o *Gene Ontology* (GO) que fornece termos definidos para se caracterizar produtos gênicos

em função de três domínios: “Processo Biológico” (série de eventos ou processos moleculares nos quais os genes estão envolvidos), “Componente Celular” (descreve a localização do produto gênico em nível celular) e “Função Molecular” (tarefas ou habilidades que o produto gênico desempenha).

A ferramenta BLAST (*Basic Local Alignment Search Tool*) também é muito utilizada nesta etapa (Johnson et al., 2008) O BLAST busca regiões de similaridade local entre sequências, através da comparação entre nucleotídeos ou proteínas presentes em seu banco de. Essa ferramenta pode ser usada para inferir relações funcionais e evolutivas entre sequências, bem como ajudar a identificar membros de famílias de genes. Outro recurso que vem sendo bastante utilizado é a Enciclopédia de Genes e Genomas de Kyoto (KEGG, http://www.kegg.jp/), cuja análise está voltada para a caracterização da função molecular de grupos de genes ortólogos, que são definidos para fins de anotação (Bolger et al., 2017).

Pode-se dizer que uma das maiores contribuições dos projetos de análise de genomas de plantas é a descoberta e a caracterização de seu conteúdo gênico, compartilhado entre muitos vegetais superiores. Esse tipo de informação é obtido tipicamente por meio da comparação estrutural e funcional entre genomas, utilizando-se uma abordagem conhecida como Genômica Comparativa (Allonso-Herrada et al., 2016; Li & Harkess, 2018; Silva-Junior et al., 2018).

2.3 CARACTERIZAÇÃO GENÔMICA DE ESPÉCIES VEGETAIS ARBÓREAS NÃO-MODELO

2.3.1 Aspectos gerais

O desenvolvimento das ferramentas e estratégias para sequenciamento de genomas tem possibilitado a obtenção de informações acerca da organização e das funções dos componentes genômicos das mais variadas espécies de plantas. Os avanços ocorreram primeiramente em espécies modelo, seguidas pelas espécies de interesse econômico (grandes culturas e algumas espécies florestais importantes) e, recentemente, pelas espécies não-modelo. O interesse em uma espécie vegetal não-modelo, em geral, se justifica pela sua capacidade de adaptação a condições ambientais extremas, relacionadas a respostas que essas espécies apresentam diante de condições de estresses bióticos e abióticos, ou ainda devido à sua capacidade de produzir metabólitos secundários específicos (Unamba et al., 2015; Silva-Junior et al., 2018).

Como exemplo de espécies vegetais arbóreas não-mdelo que já tiveram amostras de seus genomas caracterizados, podemos citar a espécie *Symphonia globulifera* L.f. (*Clusiaceae*). Popularmente conhecida como 'Oanani', esta árvore teve seu genoma parcialmente sequenciado por meio da combinação de *reads* obtidos por duas tecnologias de sequenciamento (Ilumina e 454) resultando em um *assembly* de 565 *scaffolds* (Tabela 1). A montagem obtida cobriu 67,5% do tamanho estimado para o genoma (1,522 Gb) (Olsson et al., 2016).

Dados do transcritoma de *S. globulifera* também foram utilizados para caracterização do genoma, sendo alinhados à montagem previamente obtida, contribuindo para a melhoria da qualidade da anotação estrutural. Foram preditos 1.046 genes putativos e identificados 923 SNPs de alta qualidade. Também foram identificadas 1.523 regiões contendo microssatélites, das quais 23 foram validadas como marcadores e mais de 15 foram utilizadas com sucesso na análise genética de quatro populações de *S. globulifera* da América do Sul (Brasil e Guiana Francesa) e África (Camarões e Ilha de São Tomé) (Olsson et al., 2016).

Olsson et al. (2016) reforçam que mesmo diante da fragmentação do *assembly* obtido de *S. globulifera*, devido à baixa cobertura de sequenciamento (11X), foi possível caracterizar a densidade de genes na espécie bem como obter marcadores genéticos confiáveis para serem utilizados como ferramentas em outros estudos. *S. globulifera* é uma árvore nativa das florestas da África, México e Brasil. Sua madeira é bastante apreciada e assim como outras espécies do gênero *Symphonia*, possui propriedades medicinais já exploradas pelas populações dos locais onde é encontrada.

Finch et al. (2019) utilizaram sondas construídas a partir de dados de transcritoma para enriquecer bibliotecas genômicas de cinco espécies do gênero *Cedrela* (*Meliaceae*), com o objetivo de identificar SNPs de alta confiança que auxiliem na identificação correta de algumas espécies. A partir da incorporação de 52.181 transcritos à anotação da sequência genômica de *C. odorata*, popularmente conhecida como 'Cedro-rosa', previamente obtida, foi possível anotar seguramente 9.598 genes e detectar 119.020 SNPs. Além do conjunto de SNPs e do transcritoma de referência obtidos para a espécie *C. odorata* (Tabela 1), o estudo também obteve resultados satisfatórios para a transferibilidade de sondas para enriquecimento de genomas nucleares.

Ainda no estudo feito por Finch et al. (2019), também foram caracterizados os genomas cloroplastidiais de outras espécies da família *Meliaceae*, algumas com valor

econômico e ecológico bem estabelecidos, como o ‘Mogno americano’ (*Swietenia mahagoni* L.) e a ‘Taúva’ (*Guarea guidonia* L.) (Finch et al., 2019). O gênero *Cedrela* possui espécies distribuídas principalmente no México e Argentina das quais algumas já estão em risco de extinção devido à exploração e exportação ilegais de sua madeira.

O Ipê-Rosa (*Handroanthus impetiginosus* Mart.) é a espécie arbórea nativa mais explorada no Brasil, devido à qualidade de sua madeira e à sua produção de grandes quantidades de quinoides, metabólitos que possuem ações antitumorais e antibióticas já documentadas. O genoma da espécie foi caracterizado por Silva-Junior et al. (2018) utilizando dados de alta qualidade provenientes da plataforma Illumina HiSeq 2500 (Tabela 1). A montagem conseguiu representar 90,4% (503,7 Mb) do genoma da espécie, estimado em 557 Mb de tamanho. O N50 obtido com essa montagem foi de 81.316 pb organizados em 13.206 *scaffolds*. A anotação dos genes de Ipê-Rosa foi realizada com um *pipeline* que combina transcritos montados a partir de dados de RNAseq e alinhamentos de sequências de proteínas ao *assembly* obtido. Foram identificados 28.603 genes e 35.479 transcritos, com uma média de 1,12 transcritos por gene (Silva-Junior et al., 2018).

O Baru *(Dipteryx alata* (2n = 16, *Fabaceae*)) é até o momento, uma das únicas espécies nativas do Cerrado brasileiro a ter seu genoma nuclear parcialmente sequenciado. Popularmente conhecida como baru, a espécie possui diversos usos para as populações locais: alimentação, forragem, recuperação de áreas degradadas, paisagismo e extração de madeira, além de apresentar propriedades medicinais que também estão em estudo (Ferreira et al., 2018; Antunes et al., 2020b; Antunes et al., 2020c). Os dados foram obtidos com a plataforma Illumina MiSeq (Tabela 1) e a montagem do genoma reportou 275.707 *scaffolds* (N50 = 1.598 pb) totalizando 355 Mb, o que corresponde a cerca de 44% do tamanho do genoma da espécie, estimado em 807 Mb. A cobertura média do *assembly* foi de 12,8X.

Com relação ao conteúdo repetitivo de *Dipteryx alata*, foram identificadas 21.981 regiões microssatélites e 421.701 elementos transponíveis (TEs), representando 39,29% do genoma da espécie (Antunes et al., 2020b; Antunes et al., 2020c). A partir desses dados, foram selecionados 120 pares de *primers* para o desenvolvimento de marcadores microssatélites e outros 100 pares de *primers* para o desenvolvimento de marcadores moleculares baseados no polimorfismo de TEs. As informações sobre o genoma nuclear de *D. alata* se constituem em uma importante ferramenta para desenvolvimento de estratégias de conservação e uso apropriados, auxiliando também possíveis programas de melhoramento da espécie, que ainda não possui apelo comercial em grande escala.

Estudos que visam a caracterização de genomas nucleares de outras espécies do Cerrado estão em desenvolvimento, entre eles, podemos citar a caracterização parcial do genoma de Cagaiteira (*E. dysenterica)*, realizada por Nunes (2015) e Barros-Ribeiro (2016). Considerando genomas cloroplastidiais de espécies nativas do Cerrado, podemos citar a do próprio Baru, além das sequências de Barbatimão (*Stryphnodendron adstringens*) e Pequi (*Caryocar brasiliense* Camb.) (Antunes et al., 2020a; Souza et al., 2019; Nunes et al., 2020).

2.3.2 A família *Myrtaceae*

A família *Myrtaceae*, à qual pertence a espécie em estudo neste trabalho (*Eugenia dysenterica*) abrange mais de 6.000 espécies de arbustos e árvores, classificadas em 144 gêneros dos quais vários apresentam apelo econômico: *Eucalyptus*, *Melaleuca*, *Leptospermum*, *Syzygium* e *Psidium*. Isso se deve à produção de frutos consumíveis pelos seres humanos, pela qualidade da madeira e pela capacidade de produzirem diversos compostos medicinais (antifúngicos, antioxidantes e anti-inflamatórios). Além disso, várias espécies desempenham funções ecológicas importantes em suas áreas de ocorrência, por servirem como forragem e alimento para animais (Guzman et al., 2014; Thornhill et al., 2015; Rodrigues et al., 2020; WCSP, 2021).

O genoma de *Eucalyptus grandis* (2n = 22) foi o primeiro a ser publicado para a família *Myrtaceae* e para a ordem Myrtales. O projeto utilizou dados de sequenciamento Sanger e de RNAseq (Tabela 1) e conseguiu representar 94% do tamanho estimado do genoma (640 Mb). Foram anotados 36.376 genes dos quais 30.341 (84%) estão incluídos em agrupamentos de genes compartilhados com outras linhagens evolutivas de rosídeas. Também foi observado que *E. grandis* tem um grande número de genes codificadores organizados em duplicações em tandem (12.570, 34% do total) (Myburg et al., 2014).

O genoma de *Leptospermum scoparium* (2n = 22) de tamanho estimado em 297 Mb, foi sequenciado usando uma combinação de dados Illumina, de dados de sequenciamento de bibliotecas Hi-C e de mapeamento genético de alta densidade (Tabela 1). Na sequência de referência obtida foram identificados 31.220 genes. Esta foi a primeira versão do genoma de uma espécie de valor cultural reconhecida como um tesouro (*taonga*) pelos indígenas Māori da Nova Zelândia (Thrimawithana et al., 2019).

Considerando-se os trabalhos que foram realizados utilizando sequências de RNA, podemos citar o realizado com *Psidium cattleyanum* (Tabela 1), cuja análise

possibilitou a identificação de genes potencialmente envolvidos na pigmentação dos frutos além de fornecer uma grande quantidade de dados para estudos de espécies nativas dentro da família *Myrtaceae* (Guzman et al., 2020). O transcritoma de *E. uniflora* foi caracterizado por Guzman et al. (2014) e seu genoma cloroplastidial foi caracterizado por Eguiluz et al. (2017). Ambos os estudos foram baseados em sequenciamento Illumina (Tabela 1).

Tabela 1. Dados genômicos obtidos para algumas espécies vegetais arbóreas não-modelo.

Espécie	Família	Tecnologia Utilizada	Sequências utilizadas	Citações
Caryocar brasiliense	*Caryocaraceae*	Illumina	Genoma de cloroplasto	Nunes et al. (2019)
Eucalyptus grandis	*Myrtaceae*	Sanger + RNAseq	Genoma nuclear	Myburg et al. (2014)
Eugenia uniflora	*Myrtaceae*	Illumina	Transcritoma	Guzman et al. (2014)
Cedrela odorata	*Meliaceae*	Illumina	Transcritoma	Finch et al. (2019)
Dypterix alata	*Fabaceae*	Illumina	Genoma nuclear e cloroplastidial	Antunes et al. (2020)b,c
Handroanthus impetiginosus	*Bignoniaceae*	Illumina	Genoma nuclear, transcritos e genoma cloroplastidial	Silva-Júnior et al. (2018) Sobreiro et al. (2020)
Leptospermum scoparium	*Myrtaceae*	Illumina + Hi-C	Genoma nuclear	Thrimawithana et al. (2019)
Psidium cattleyanum	*Myrtaceae*	Illumina	Transcritoma	Rodrigues et al. (2020)
Stryphnodendron adstringens	*Leguminosae*	Illumina	Genoma cloroplastidial	Souza et al (2019)
Symphonia globulifera	*Clusiaceae*	Illumina + 454	Genoma nuclear, transcritos	Olsson et al. (2016)
Swietenia mahagoni	*Meliaceae*	Illumina	Genoma cloroplastidial	Finch et al. (2019)

Genomas cloroplastidiais já caracterizados dentro da família *Myrtaceae* incluem espécies como *Plinia truciflora* (Eguiluz et al., 2017), *Acca sellowiana* (Machado et al., 2017) e *Campomanesia xanthocarpa* (Machado et al., 2020). A partir desse tipo de informação genética, Rodrigues et al. (2020), por exemplo, compararam seis espécies neotropicais pertencentes à família *Myrtaceae* (*Eugenia brasiliensis*, *E. pyriformis*, *E. nitida*, *Myrcianthes pungens*, *Plinia edulis* e *Psidium cattleyanum*) utilizando dados Illumina. Os plastomas sequenciados exibem uma estrutura quadripartida típica, conteúdo gênico e organização altamente conservada entre as espécies analisadas, sendo observadas poucas diferenças no comprimento dos genomas, genes codificadores de proteínas e regiões não codificantes.

2.3.3 *Eugenia dysenterica*

A cagaiteira (*Eugenia dysenterica* (Mart.) DC.) é uma árvore frutífera nativa do Brasil, que ocupa áreas de solos profundos e bem drenados. A espécie pertence à família *Myrtaceae*, que é representada no Cerrado por 14 gêneros e 211 espécies, sendo considerada uma das dez famílias mais representativas desse bioma, contribuindo com cerca de 51% da sua riqueza florística (Chaves & Telles, 2006; Marinotto et al., 2008; Camilo et al., 2013; Grattapaglia et al., 2012).

A cagaiteira possui porte médio com altura que pode variar entre 4 e 10 m, com troncos tortuosos típicos de espécies que ocorrem no Cerrado (Figura 4). As flores são hermafroditas, ocorrendo uma sincronização do florescimento em um curto período, entre os meses de agosto e setembro (Figura 4) (Proença & Gibbs, 1994; Camilo et al., 2013; Almeida Júnior et al., 2014). A frutificação ocorre entre os meses de setembro e outubro e a maturação dos frutos ocorre de maneira relativamente rápida, coincidindo com o início do período chuvoso. A cagaiteira apresenta sistema de cruzamento misto, com predomínio de alogamia, apresentando taxa de fecundação cruzada estimada entre 83,5% e 100,0% (Zucchi et al., 2002; Telles et al., 2003; Rodrigues et al., 2016).

Figura 4. (A) Cagaiteira (*Eugenia dysenterica*) - indivíduo adulto durante o período de floração. (B) Tronco de uma Cagaiteira adulta (c) Frutos da Cagaiteira. Fonte: Acervo pessoal.

A espécie destaca-se dentre as diversas plantas nativas do Cerrado brasileiro que apresentam potencial de utilização em sistemas de produção agrícola, pelo consumo de seus frutos e pelo seu valor ornamental, devido à sua beleza, realçada principalmente na época de florescimento. Silva et al. (2015) concluíram que a cagaiteira possui amplo potencial para desenvolvimento de medicamentos fitoterápicos, porém ainda existe a necessidade de pesquisas a fim de se preencher lacunas no que diz respeito à sua eficácia e segurança, para o posterior desenvolvimento desses produtos (Cardoso et al., 2011; Barros-Ribeiro, 2016).

A diversidade genética bem como a atuação de processos como deriva genética e fluxo gênico em populações naturais e em uma coleção de germoplasma de cagaiteira vêm sendo estudados nas últimas décadas utilizando marcadores aloenzimáticos, RAPDs e SSRs (Telles et al., 2001; Trindade & Chaves, 2005; Zucchi, 2004; Aguiar et al., 2009).

Análises feitas utilizando izoenzimas (Telles et al., 2001) permitiram inferir o nível de endogamia e o sistema reprodutivo predominantes na espécie; o grau de diferenciação interpopulacional e os padrões espaciais associados à divergência genética, informações importantes para o desenvolvimento de programas de conservação. Marcadores RADPs e morfológicos também foram utilizados (Zucchi, 2004; Trindade & Chaves, 2005) para analisar a estrutura genética de subpopulações amostradas em diferentes regiões do estado de Goiás. Análises de estrutura genética populacional para *E. dysenterica* também foram feitas utilizando-se marcadores SSR (Zucchi et al., 2004), primeiramente transferidos de *Eucalyptus* spp e posteriormente, com marcadores desenvolvidos especificamente para a espécie (Telles et al., 2013; Barbosa et al., 2015).

Dentre os resultados obtidos com os trabalhos citados anteriormente, foi observado o comprometimento da estrutura metapopulacional da espécie, pelos indícios de que as populações estudadas sofreram deriva genética devido à ação humana (Zucchi et al., 2004). Também foi observado que a maior parte da variação genética de *E. dysenterica* se encontra dentro de subpopulações (Aguiar et al., 2009).

Os marcadores SSRs associados a caracteres quantitativos também foram utilizados para caracterizar a magnitude e a distribuição da diversidade genética em populações naturais e em uma coleção de germoplasma da espécie, bem como para se estudar seu sistema de cruzamento (Rodrigues et al., 2016; Boaventura-Novaes, 2018a, b). De acordo com Rodrigues et al., 2016, foi possível observar que a coleção de germoplasma de *E. dysenterica* possui um elevado potencial para iniciar um programa de cultivo da espécie, apresentando altas taxas de cruzamento por alogamia e uma elevada diversidade (Rodrigues et al., 2016).

Boaventura-Novaes et al. (2018b) observou que os ambientes de ocorrência natural de *E. dysenterica* amostrados estão em um hotspot de biodiversidade com solo e condições climáticas heterogêneas. Neste estudo também foi observado que a deriva genética está atuando fortemente na diferenciação fenotípica entre as subpopulações analisadas e que *E. dysenterica* possui uma estrutura genética espacial dividida em dois grandes grupos, separados por uma linha que divide o bioma Cerrado em duas regiões.

Alguns estudos baseados em filogeografia e história evolutiva da espécie também foram realizados, utilizando sequências de DNAs nucleares e cloroplastidiais obtidas a partir de regiões específicas (Diniz-filho et al., 2016; Lima et al., 2017). Com estes estudos, foi possível observar que a diversidade genética atual e a estrutura populacional em

E. dysenterica podem ser explicadas por mudanças de alcance geográfico associadas à dinâmica do clima (Diniz-filho et al., 2016). Já os resultados obtidos por Lima et al. (2017) sugerem que a região central do bioma Cerrado é provavelmente o centro de diversidade genética *de E. dysenterica* e que o padrão espacial dessa diversidade pode ser resultado da estabilidade populacional ao longo do Quaternário.

No contexto genômico, a obtenção de um *draft assembly* foi realizada utilizando dados de sequenciamento de alto rendimento (plataforma Illumina MiSeq). A partir desses dados, foi possível obter uma caracterização preliminar do genoma que representou cerca de 56% (~250 Mb) do seu tamanho estimado (~442 Mb, estimativa feita *in silico*), com cerca de 35% da sequência obtida no *assembly* composta por regiões repetitivas. A maior parte das repetições identificadas no *assembly* de *E. dysenterica* foram retrotransposons LTR, assim como observado e descrito para outras espécies, de acordo com a literatura. Também foi possível identificar cerca de 60.000 fragmentos relacionados a genes e 999.016 SNPs putativos, com uma densidade de um SNP a cada ~251 pb. (Nunes, 2015; Barros-Ribeiro, 2016).

3 MATERIAL E MÉTODOS

3.1 SEQUENCIAMENTO DO GENOMA CLOROPLASTIDIAL DE *Eugenia dysenterica*

3.1.1 Obtenção do material genético, sequenciamento e montagem da sequência de referência

O DNA genômico total foi extraído de folhas de árvores adultas coletadas de forma aleatóia na coleção de germoplasma da Escola de Agronomia – Universidade Federal de Goiás (UFG), utilizando-se o protocolo CTAB (Doyle & Doyle, 1987). As amostras de DNA obtidas foram avaliadas quanto à concentração por meio do fluorímetro Qubit (Thermo Fisher, 2021a) e qualidade através de eletroforese horizontal em gel de agarose (1%) e Nanodrop (Thermo Fisher, 2021b). A biblioteca de fragmentos de DNA genômico para sequenciamento foi construída utilizando-se o kit *Nextera® DNA Sample Preparation* (Illumina, 2015a), seguindo o protocolo do fabricante. O sequenciamento foi realizado na plataforma Illumina MiSeq, em modo *paired-end* 2x300 (Illumina, 2015b).

Os *reads* brutos provenientes do sequenciamento foram utilizados na montagem do genoma cloroplastidial pelo *pipeline* Fast-Plast (https://github.com/mrmckain/Fast-Plast) (McKain & Wilson, 2021). Este *pipeline* promove a montagem *de novo* dos genomas através da associação entre a análise de grafos de Brujin e do *software* SPAdes, utilizando durante o processo diversas ferramentas de bioinformática como Trimmomatic, Bowtie2, Blast e R (Figura 5). Os reads brutos foram filtrados quando à sua qualidade, utilizando o Trimmomatic. Após a remoção das sequências de baixa qualidade, os reads foram alinhados ao conjunto de genomas de cloroplasto inteiros, implementados no Fast-Plast, para seleção das sequências de interesse. Este alinhamento foi feito com o Bowtie2.

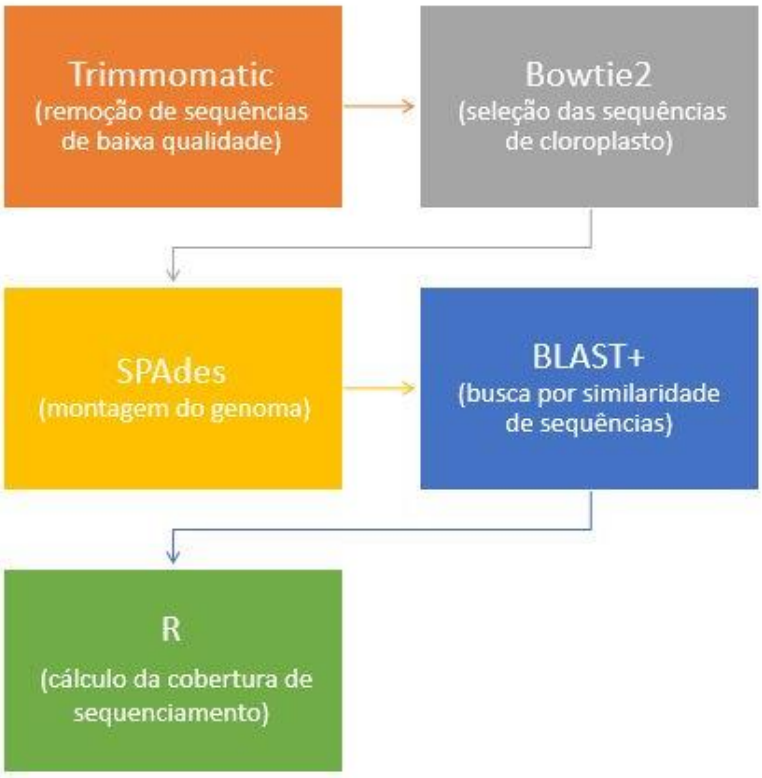

Figura 5. Fluxograma ilustrando as principais etapas do pipiline Fast-Plast, utilizado para a obtenção do genoma cloroplastidial de *Eugenia dysenterica.*

Após a obtenção das sequências de cloroplasto, a montagem do genoma foi feita utilizando o SPAdes. Nesta etapa, o Fast-Plast também utiliza o BLAST+ para confirmar a similaridade da montagem obtida com dados de outros genomas cloroplastidiais. A cobertura de sequenciamento dos reads utilizados na montagem foi feita utilizando o software R, também inmplementado no Fast-Plast (Figura 5).

3.1.2 Anotação dos genes e caracterização de conteúdo repetitivo

Após a montagem do genoma cloroplastidial, a anotação das suas sequências gênicas foi feita utilizando-se os programas *Dual Organellar GenoMe Annotator* (DOGMA) e *GeSeq* (https://chlorobox.mpimp-golm.mpg.de/Alternative-Tools.html) (Wyman et al., 2004). Os códons de iniciação e de terminação, bem como os limites de íntrons/éxons foram inspecionados manualmente. O DOGMA também foi utilizado para anotar as sequências de RNAs de transferência (tRNAs), juntamente com o *software* tRNAscan-SE ver. 2.0. O mapa circular do genoma foi construído com *o Organellar Genome Draw* (OGDRAW) (Greiner

et al., 2019) e a análise de uso de códons foi realizada no servidor web *Bioinformatics* (https://www.bioinformatics.org/sms2/codon_usage.html/).

Foi feita a busca por repetições longas (*Longe Repeats* – LRs) dos tipos *forward*, *reverse*, palindrômicas e complementares) no cloroplasto de *E. dysenterica*, utilizando-se o *software* Reputer (Kurtz, 1999). Para isso foram estabelecidos os seguintes parâmetros: regiões com tamanho mínimo de 30 pb; distância de *Hamming* de 3 e taxa de identidade de pelo menos 90%. Já as sequências de repetição simples (SSRs) foram detectadas utilizando-se a ferramenta MISA (http://pgrc.ipk-gatersleben.de/misa/misa.html) (Beier et al., 2017), utilizando como parâmetros: um mínimo de dez unidades de repetição para mononucleotídeos, cinco unidades de repetição para dinucleotídeos, quatro unidades de repetição para trinucleotídeos e três unidades de repetição para tetra, penta e hexanucleotídeos.

3.1.3 Análise de diversidade nucleotídica e razão Ka/Ks

A diversidade nucleotídica (π) entre o genoma cloroplastial de *E. dysenterica* e de outros quatro genomas cloroplastidiais de espécies do mesmo gênero disponíveis no GeneBank: *E. brasiliensis* (MN095407), *E. selloi* (MN095411), *E. pyriformis* (MN095410) e *E. uniflora* (NC_027744) foi avaliada. Para tanto, foi feito o alinhamento dos genomas inteiros usando a ferramenta MAFFT (Katoh, 2013), disponível no *software* Geneious (Kearse et al., 2012).

As sequências dos 78 genes codificadores de proteínas comuns entre as espécies foram extraídas e alinhadas separadamente pelo MAFFT para se estimar as taxas de substituições não-sinônimas (Ka) e sinônimas (Ks). A razão Ka/Ks para cada um dos genes foi estimada utilizando-se o *software* DnaSP v.6 (Rozas et al., 2017).

3.1.4 Genômica comparativa e análise filogenética

O *software* MVista (http://genome.lbl.gov/vista/mvista/about.shtml) foi utilizado no modo shuffle-LAGAN (Frazer et al., 2004) para se comparar o genoma de cloroplasto obtido para *E. dysenterica* às sequências de cloroplasto completas de outras quatro espécies do mesmo gênero (*E. brasiliensis*, *E. pyriformis*, *E. selloi* e *E. uniflora*). Já

a expansão e contração das regiões IR nos locais de junção entre as cinco espécies analisadas, foram verificadas e visualizadas utilizando-se o *software* IRscope (Amiryousefi et al., 2018).

Para confirmar a posição filogenética e se avaliar a relação de *E. dysenterica* com outras espécies da família *Myrtaceae*, as sequências de 78 genes codificadores de proteínas comuns entre 11 espécies (incluindo *E. dysenterica*) de seis gêneros da família foram obtidas no GenBank (Apêndice 1) e alinhadas com o MAFFT (Katoh, 2013). A tribo Eucalypteae foi utilizada como grupo externo. A filogenia foi reconstruída com base no modelo evolutivo GTR+I+G (*General time reversible* + *I* + *G*), previamente escolhido utilizando-se o JModelTest (http://evomics.org/learning/phylogenetics/jmodeltest/). A reconstrução da filogenia foi baseada no método de Máxima Verossimilhança (*Maximum Likelihood* - ML) com 1000 réplicas de *bootstrap* utilizando-se o *software* MEGA (Hall, 2013). A árvore final foi visualizada no *software* FigTree (http://tree.bio.ed.ac.uk/software/figtree/) (Rambaut, 2007).

3.2 MONTAGEM E ANOTAÇÃO DE UM TRANSCRITOMA DE REFERÊNCIA PARA *Eugenia dysenterica*

3.2.1 Extração e sequenciamento de RNA

Amostras de folhas maduras (que posteriormente deram origem à biblioteca nomeada como FM01) provenientes de árvores adultas da coleção de germoplasma da Universidade Federal de Goiás (UFG) foram coletadas de forma aleatória. Plântulas inteiras, provenientes da semeadura de sementes obtidas a partir dessas mesmas árvores também foram amostradas, das quais duas plântulas foram submetidas a condições de estresse hídrico (que posteriormente deram origem às bibliotecas nomeadas como PS02 e PS03) e três plântulas submetidas a condições favoráveis em termos de suprimento de água (que deram origem às bibliotecas nomeadas como PC01, PC02 e PC03).

A amostragem de plântulas PS e PC foi feita com o objetivo de se obter uma maior diversidade de genes expressos em diferentes condições ambientais. Todas as amostras foram conservadas em nitrogênio líquido logo após a coleta e armazenadas em ultrafreezer (−80 °C) até o momento da extração. O RNA foi extraído utilizando kit *Purelink Plant RNA Reagent* (Thermo Fisher, 2021), seguindo protocolo do fabricante. O controle de qualidade das amostras de RNA extraído foi realizado utilizando-se o fluorímetro Qubit (Thermo Fisher), a eletroforese no equipamento Bioanalyser 2100 (Agilent Technologies) e

em gel de agarose (1%), contendo formaldeído. As amostras de RNA obtidas foram enviadas à empresa BGI GENOMICS para fins de construção das bibliotecas e sequenciamento. Foram construídas seis bibliotecas (Tabela 5), sendo uma construída a partir de RNAs extraídos de folhas maduras (FM01), duas construídas a partir de plântulas submetidas a condições de déficit de água (PS02 e PS03) e três construídas a partir de plântulas submetidas a condições normais de suprimento de água (PC01, PC02 e PC03) (Tabela 5).

O sequenciamento foi realizado na plataforma *HiSeq 4000*, no modo *paired-end* (2 x 100). Os *softwares FastQC* (Andrews, 2010) e *MultiQC* (Ewels et al., 2016) foram utilizados para se visualizar e avaliar a qualidade dos dados brutos obtidos no sequenciamento. A remoção dos adaptadores, bem como das sequências de baixa qualidade foi feita utilizando-se o *software Trimmomatic* (incluído no *software Trinity*).

3.2.2 Montagem e caracterização do transcritoma de referência

A montagem de uma sequência de referência para o transcritoma de *E. dysenterica* foi realizada utilizando-se o *pipeline* Trinity-v2.8.4 (Grabherr et al., 2011). Para isso, todas as seis bibliotecas foram utilizadas. A abordagem utilizada pelo Trinity se baseia no método de montagem *de novo* e seu fluxo de trabalho envolve três módulos: (1) o *Inchworm*, que estende os *reads* em *super-reads*, (2) o *Chrysalis*, que analisa os *super-reads* produzidos pelo *Inchworm* e cria um grafo de Bruijn (*cluster*) para cada transcrito e (3) o *Butterfly*, que extrai as sequências das isoformas de cada *cluster* (Grabherr et al., 2011; Haas et al., 2013).

A predição das regiões codantes a partir dos *contigs* obtidos na montagem foi realizada pelo programa TransDecoder-v5.5.0 (http://transdecoder.github.io) (Hass & Papanicolaou, 2019). O *software* Trinotate-v3.2.1 (https://trinotate.github.io/) (Kitzmiller, 2015) foi utilizado para anotação preliminar das proteínas preditas, utilizando-se como referências as bases de dados *Pfam* (http://pfam.xfam.org) e *SwissProt* (http://www.uniprot.org/uniprot/). As informações obtidas pelo *Trinotate* foram utilizadas para se atribuir termos GO aos transcritos preditos.

3.2.3 Identificação de SNPs no transcritoma de referência

Os *reads* das bibliotecas relativas às plântulas foram alinhados a uma versão modificada do transcritoma de referência em que cada gene foi representado pelo transcrito com a maior ORF. O alinhamento foi realizado pelo *software* BWA (Li & Durbin, 2009). A

identificação e genotipagem dos SNPs foi realizada na plataforma GATK-v 4.1.0.0 (Van der Auwera & O'Connor, 2020).

Seguindo-se as recomendações do *GATK Best Practices*, os *reads* duplicados foram removidos e aos SNPs inicialmente identificados foram aplicados filtros de qualidade em duas etapas. Na primeira etapa só foram mantidos SNPs identificados com base em uma cobertura entre 5 e 500X, e com uma relação QD (*Quality by Depth*) acima de 2. Em uma segunda etapa, com base na análise dos genótipos obtidos para cada um dos cinco indivíduos, só foram mantidos os SNPs que apresentaram uma cobertura de 3X para cada um dos alelos identificados e frequências alélicas entre 0,05 e 0,95.

4 RESULTADOS E DISCUSSÃO

4.1 GENOMA CLOROPLASTIDIAL

4.1.1 Conteúdo e organização do plastoma de *Eugenia dysenterica*

O sequenciamento do DNA total de *E. dysenterica* resultou em 27.208.448 *reads* brutos dos quais 1.047.017 foram utilizados para a montagem do genoma cloroplastidial. A cobertura média de sequenciamento estimada para a montagem foi de 1.177X com um desvio padrão de 278,5X (mediana: 1.215X; mínimo: 57X e máximo: 2.113X) (Figura 6). A montagem resultou em um único *contig* correspondente ao comprimento total do genoma de cloroplasto com tamanho de 158.560 pb (Tabela 2, figura 7). A estrutura geral do plastoma de *E. dysenterica* apresenta organização análoga à observada em outras angiospermas, com quatro regiões típicas sendo as regiões *Large Single Copy* (LSC) e *Small Single Copy* (SSC), com 87.094 pb e 18.650 pb de comprimento, respectivamente, separadas por duas regiões invertidas repetidas (IRa e IRb), que possuem 26.408 pb cada uma (Figura 7).

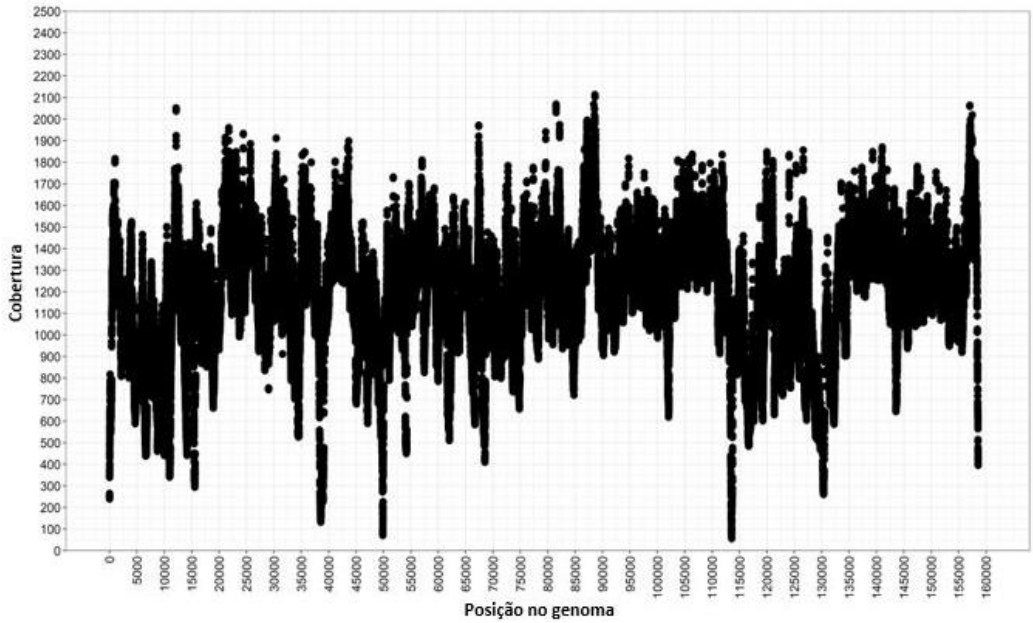

Figura 6. Distribuição da cobertura de sequenciamento ao longo da sequência de referência obtida para o genoma cloroplastidial de *Eugenia dysenterica*.

Tabela 2. Características gerais do genoma cloroplastidial de *Eugenia dysenterica*.

Características	Valores observados
Tamanho total do genoma (pb)	158.560
Tamanho da região LSC (pb)	87.094
Tamanho da região SSC (pb)	18.650
Tamanho de cada região IR (pb)	26.408
Regiões codificantes de proteínas (pb / %)	79.851 / 50,36
Regiões de rRNA (pb / %)	9.056 / 5,71
Regiões de tRNA (pb / %)	2.792 / 1,76
Regiões intrônicas (pb / %)	16.950 / 10,69
Sequências intergênicas (pb / %)	49.911 / 31,48
Número de genes	129
Número de genes codificantes de proteínas diferentes	78
Número de genes de tRNA diferentes	30
Número de rRNA diferentes	4
Número de genes duplicados	17
Número de pseudogenes	1
Conteúdo GC total (%)	36,96
Conteúdo GC em LSC	34,82
Conteúdo GC em SSC	30,73
Conteúdo GC em IRs	42,82

O conteúdo GC das IRs mostrou-se maior (42,8%) quando comparado às demais regiões do cloroplasto (34,8% para LSC e 30,6% para SSC). O conteúdo GC do genoma de cagaiteira é semelhante ao de outras espécies da tribo Myrteae, cujos valores gerais nos plastomos variam de 37 a 39%, com 35-37% na região LSC, 30-35% na região SSC, e 43% para as regiões IR (Gu et al. 2016; Machado et al., 2017). O alto conteúdo de GC nas regiões IR deve-se principalmente presença de genes que codificam para RNAs ribossômico (com conteúdo GC superior a 50%), localizados nesta região.

As sequências codificadoras de proteínas corresponderam a ~50% do genoma cloroplastidial de *E. dysenterica*. As regiões não codificantes, incluindo íntrons e espaçadores intergênicos, corresponderam a ~42% e as sequências que incluem genes de rRNAs e tRNAs compreenderam ~7% do genoma, valores muito semelhantes aos

observados em outras espécies da tribo Myrteae (Gu et al., 2016; Machado, 2017, Rodrigues et al., 2020).

Foram identificados 129 genes no genoma cloroplastidial de *E. dysenterica*, dos quais 112 aparecem em cópia única. Considerando-se somente os genes de cópia única, 78 codificam proteínas, 30 codificam RNAs transportadores (tRNAs) e quatro codificam RNAs ribossômicos (rRNAs) (Tabela 2, figura 7).

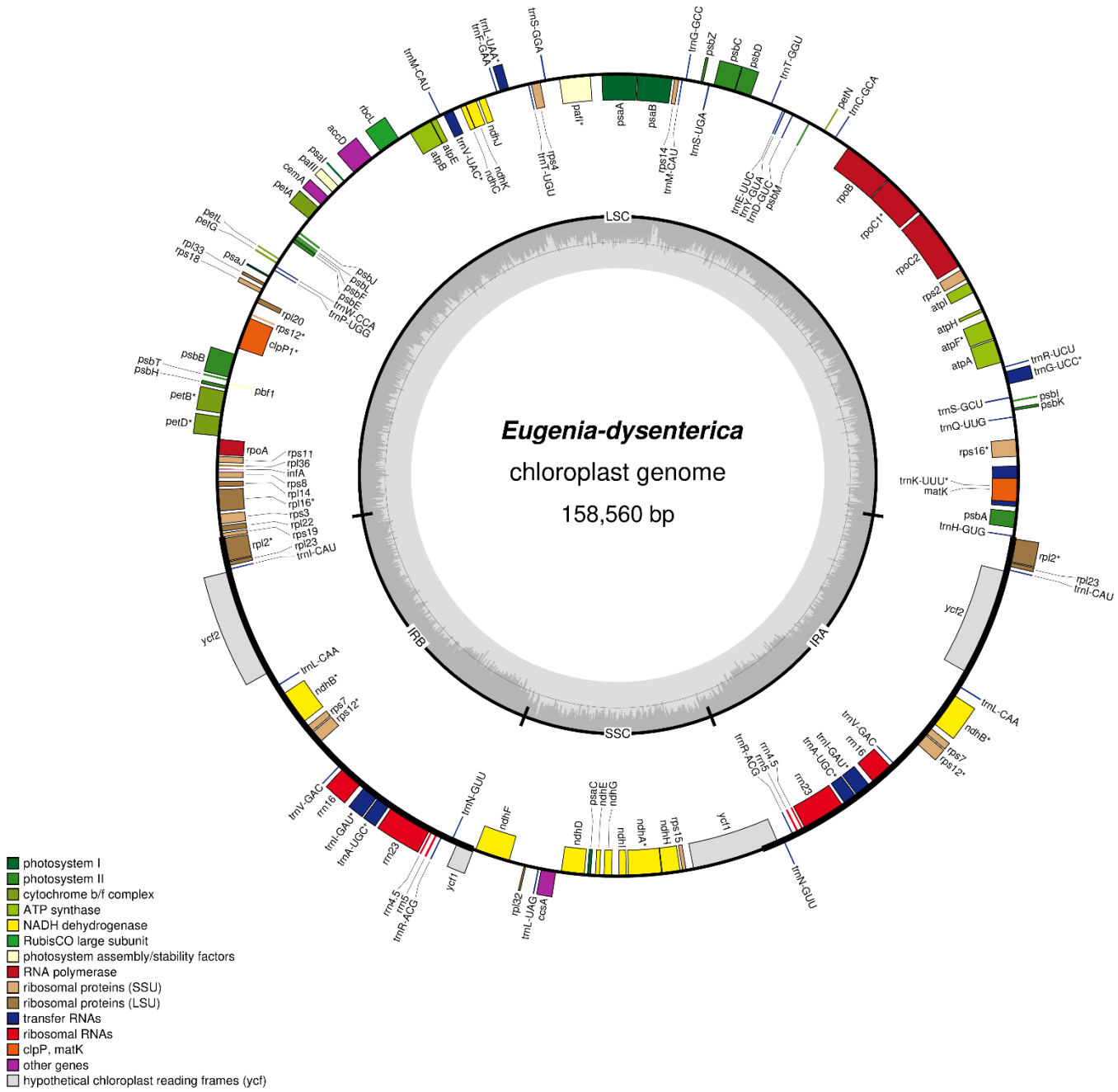

Figura 7. Mapa circular mostrando o conteúdo gênico do genoma cloroplastidial de *Eugenia dysenterica*. As linhas grossas indicam a extensão das repetições invertidas (IRa e IRb). Os genes do lado de fora do mapa são transcritos no sentido horário e os de dentro do mapa são transcritos no sentido anti-horário. Os genes de *E. dysenterica* foram agrupados de acordo com seus grupos funcionais, indicados na legenda com diferentes cores. O cinza mais escuro no círculo interno corresponde ao conteúdo GC, enquanto o cinza mais claro corresponde ao conteúdo AT.

Espécies da família *Myrtaceae* apresentam genomas cloroplastidiais que variam de 157.683 pb em *Eugenia nítida* a 161.071 pb em *Eucalyptus spatulata* (Bayly et al., 2013; Rodrigues et al., 2020). Ao se comparar a estrutura geral dos plastomas de *E. dysenterica* e outras quatro espécies do gênero *Eugenia* (Tabela 3), observou-se que o plastoma de cagaiteira possui o segundo maior genoma, diferenciando-se em 9 pb do genoma de *E. pyriformis*. Contudo, em *E. dysenterica* a região LSC apresentou-se a segunda menor, diferenciando-se em 365 pb quando comparada a *E. uniflora*. Já região SSC em cagaiteira apresentou-se a maior dentre as espécies comparadas, com até 360 pb a mais quando comparada a *E. brasiliensis*.

Tabela 3. Características gerais do genoma cloroplastidial de *Eugenia dysenterica* e de outras quatro espécies do mesmo gênero.

Espécie do gênero *Eugenia*	Tamanho (pb)	LSC (pb)	SSC (pb)	IR (pb)	GC (%)	Nº de proteínas	Nº de RNAs
E. brasiliensis	158.251	87.201	18.290	26.380	36,95	78	34
E. dysenterica	158.560	87.094	18.650	26.408	36,96	78	34
E. pyriformis	158.569	87.189	18.566	26.407	36,90	78	34
E. selloi	157.683	86.436	18.349	26.449	37,04	78	34
E. uniflora	158.445	87.459	18.318	26.334	36,98	77	34

Comparado a outras famílias de angiospermas, especialmente com outras espécies nativas não-modelo, o tamanho do cloroplasto de cagaiteira também se apresenta semelhante ao observado em *Dipteryx alata* (baru) – *Fabaceae* (158.647 pb) e *Handroanthus impetiginosus* (ipê-rosa) – *Bignoniaceae* (159.462 bp) e menor em relação aos de *Caryocar brasiliense* (pequi) - *Caryocaraceae (*165.793 bp) e *Stryphnodendron adstringens* (barbatimão) – *Leguminosae* (162.169 bp) (Souza et al., 2019; Nunes et al., 2020*).* As diferenças de tamanho nos genomas cloroplastidiais podem ser causadas por diferentes fatores, como a redução, expansão ou perda da região IR, perda ou aumento no número de genes e/ou diminuição no comprimento dos íntrons ou das regiões intergênicas (Antunes et al., 2020a).

Considerando-se outras espécies da tribo Myrteae, como *E. brasiliensis* e *Plinia edulis,* o número total de genes identificados bem como as quantidades de genes

codificadores de proteínas e RNAs foram idênticos (Rodrigues et al., 2020). Considerando-se espécies de outras famílias, o conteúdo gênico para cagaiteira foi maior que em barbatimão (111 genes, 77 CDS, 30 tRNA e quatro rRNAs), ipê-rosa (124 genes, 84 CDS, 36 tRNAs e quatro rRNAs) e baru (125 genes, 76 CDS, 29 tRNAs e quatro rRNAs) e menor que em pequi (136 genes, 87 CDS, 37 tRNA genes e oito rRNAs).

Dentre os genes identificados em *E. dysenterica*, 17 são duplicados por estarem localizados na região IR. Considerando-se apenas os genes duplicados, seis codificam proteínas (*rpl2, rpl23, ycf2, ndhB, rps7, yps12,*), sete codificam tRNAs e quatro codificam rRNAs (Tabela 4). O maior gene identificado no plastoma de *E. dysenterica* foi o *ycf2* (6.861 pb) e o menor foi o gene *trnC-GCA* (71 pb) (Tabela 4).

O gene *ycf2* é relatado na literatura como sendo comum a todos os grupos de plantas, e isso se deve a sua possível relevância para o desenvolvimento desses organismos. Wicke (2011) relatou que esse gene tende a ser mais expresso em frutos, informação que pode ser verificada em estudos futuros voltados à análise de expressão de genes específicos em *E. dysenterica*. Machado et al. (2020) sugeriram que o gene *ycf2* apresenta-se como um dos *hotspots* de variabilidade dentro da família *Myrtaceae*, tornando-o um candidato potencial para estudos baseados em DNA *barcode* ao nível de espécies.

A presença de um único pseudogene (*ycf1)* parcial e localizado na borda IRb/SSC (Figuras 5 e 11) do genoma cloroplastidial de *E. dysenterica* é uma característica que também já foi observada em outras espécies de *Myrtaceae* como *E. brasiliensis*, *Plinia eduli* e *Psidium cattleianum*, porém é possível se observar que em outras espécies vegetais foram identificados até quatro pseudogenes (*Caryocar brasiliense, Acca sellowiana, Campomanesia xanthocarpa, Syzygium cumini* e *Eucalyptus* spp.). Estudos sugerem que o gene *ycf1* surgiu nos genomas cloroplastidiais por um evento evolutivo que antecede a ocupação do ambiente terrestre (Wicke et al., 2011).

Foi observado que 18 dos genes identificados em *E. dysenterica* possuem apenas um íntron e três genes (*ycf3, clpP, rps12*) possuem dois íntrons em sua estrutura (Apêndice 2). A maioria dos genes que possuem íntrons (12) estão localizados na região LSC do genoma cloroplastidial e os demais encontram-se distribuídos entre as regiões IR (5), sendo apenas um encontrado na região SSC (*ndhA)* (Tabela 4, Apêndice 2). O gene *trnK-UUU* possui o maior íntron que engloba o gene matK, com 2.526 bp, enquanto o íntron do gene *trnL-UAA* é o menor (508 bp).

Uma pequena variação na quantidade de íntrons nos genes cloroplastidiais de *E. dysenterica* pode estar relacionada aos níveis de expressão desses genes. Em estudos anteriores, observou-se que em plantas, genes com maiores níveis de expressão tendem a ser menos compactos, possuindo mais íntrons ou íntrons mais longos, quando comparados a genes menos expressos (Yang et al., 2009; Das & Bansal, 2019).

Foram identificados um total de 22.934 códons no plastoma de *E. dysenterica* (Apêndice 3) dos quais a maioria são codificadores para o aminoácido leucina (2.421 códons, ~10,5% do número total de códons), sendo o códon UUA observado com maior frequência para esse aminoácido (33.9%). Já os códons que codificam o aminoácido cisteína foram os que ocorreram em menor abundância (276 códons, ~1,2% do total). Além disso, apenas um códon foi identificado para a codificação dos aminoácidos metionina (ATG) e triptofano (TGG).

Os resultados obtidos são bem semelhantes àqueles observados em *S. adstringens*, diferenciando-se apenas na quantidade total de códons identificados (20.986) e no códon de ocorrência mais frequente (TTA em *S. adstringens*). Os códons são estruturas fundamentais no processo de transmissão da informação genética, desempenhando um papel importante nas atividades biológicas do organismo, assim essas informações poderão ajudar no desenvolvimento de programas de otimização de códons (*codon usage*) voltados para estudos com transgenia envolvendo o genoma de cloroplasto (Daniell et al., 2016; Duan et al., 2021).

Tabela 4. Genes presentes no genoma cloroplastidial de *Eugenia dysenterica.*

Categoria	Grupo de genes	Nome dos genes
Self-replication	Large subunit of ribosomal proteins	*rpl2*[1,2], *rpl14*, *rpl16*[1], *rpl20*, *rpl22*, *rpl23*[2], *rpl32*, *rpl33*, *rpl36*
	Small subunit of ribosomal proteins	*rps2*, *rps3*, *rps4*, *rps7*[2], *rps8*, *rps11*, *rps12*[1,2], *rps14*, *rps15*, *rps16*[1], *rps18*, *rps19*
	DNA-dependent RNA polymerase	*rpoA*, *rpoB*, *rpoC1*[1], *rpoC2*
	Ribosomal RNA genes	*rrn4.5*[2], *rrn5*[2], *rrn16*[2], *rrn23*[2]
	Transfer RNA genes	*trnA-UGC*[1,2], *trnC-GCA*, *trnD-GUC*, *trnE-UUC*, *trnF-GAA*, *trnfM-CAU*, *trnG-UCC*[1], *trnG-UCC*, *trnH-GUG*, *trnI-CAU*[2], *trnI-GAU*[1,2], *trnK-UUU*[1], *trnL-CAA*[2], *trnL-UAA*[1], *trnL-UAG*, *trnM-CAU*, *trnN-GUU*[2], *trnP-UGG*, *trnQ-UUG*, *trnR-ACG*[2], *trnR-UCU*, *trnS-GCU*, *trnS-UGA*, *trnS-GGA*, *trnT-UGU*, *trnT-GGU*, *trnV-UAC*[1], *trnV-GAC*[2], *trnW-CCA*, *trnY-GUA*
Photosynthesis	Photosystem I	*psaA*, *psaB*, *psaC*, *psaI*, *psaJ*
	Photosystem II	*psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbM*, *psbN*, *psbT*, *psbZ*
NADH dehydrogenase	NADH dehydrogenase	*ndhA*[1], *ndhB*[1,2], *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhG*, *ndhH*, *ndhI*, *ndhJ*, *ndhK*
	Cytochrome b/f complex	*petA*, *petB*[1], *petD*[1], *petG*, *petL*, *petN*
	ATP synthase	*atpA*, *atpB*, *atpE*, *atpF*[1], *atpH*, *atpI*
	RuBisCo large subunit	*RbcL*
Other genes	Maturase K	*matK*
	Envelope membrane protein	*cemA*
	Subunit of acetyl-CoAcarboxylase	*accD*
	C-type cytochrome synthesis gene	*ccsA*
	Protease	*clpP*[1]
	Conserved hypothetical chloroplast open reading frames	*ycf2*[2], *ycf3*[1], *ycf4*
	Pseudogene	*ycf1*

1 – Genes sem íntrons

2 – Genes completos e duplicados em região IR.

4.1.2 Análise de sequências de repetição

Um total de 33 estruturas repetidas, com comprimentos variando de 30 a 47 pb, foram detectadas no genoma do cloroplasto de *E. dysenterica*. Destas, 17 são repetições *forward*, 13 são repetições palindrômicas e três são repetições *reverse*. Não foram identificadas estruturas de repetições complementares. Quanto aos tamanhos, as repetições *forward* variaram de 30 a 47 pb e as repetições *reverse* foram de 30 a 38 pb. Já as repetições palindrômicas variaram de 30 a 45 pb. Considerando-se a localização das repetições identificadas, a maioria delas (14) estão na região IR do cloroplasto, oito estão na região LSC e cinco estão na região SSC.

Duas das repetições *forward* (com 39 e 30 pb de comprimento, respectivamente) foram encontradas no íntron do gene *ycf3* (LSC) e na região entre os genes *rps12* e *trnL-UAA* (LSC/IR). Também foram encontradas repetições *forward* (42 pb) nos íntrons dos genes *ycf3* (LSC); *ndhA* (SSC) (IR/SSC) e na região intergênica *rps12* - *trnV-GAC* (IR) (30 pb) (Apêndice 5). Essas repetições desempenham um papel importante no rearranjo dos plastomas, sendo utilizadas para estudos evolutivos e de genética populacional, através do desenvolvimento de marcadores moleculares (Yi et al., 2013, Sobreiro et al., 2020; Alzahrani et al., 2021).

As repetições encontradas em regiões codantes do genoma de cloroplasto de *E. dysenterica* apresentam pouca variação de tamanho, assim como observado por Asif et al. (2013) em *Syzygium cumini* e *Eucalyptus grandis* (*Myrtaceae*) e em espécies de outras famílias como *Gossypium barbadense*, *Arabidopsis thaliana* e *Nicotiana tabacum* o que sugere que essas estruturas sejam conservadas em plantas angiospermas.

Foram identificadas 78 regiões SSR no cloroplasto de *E. dysenterica* (Apêndice 4), das quais 58 têm como motivos sequências de mononucleotídeos, uma de dinucleotídeo, três de trinucleotídeos e 16 de tetranucleotídeos. Considerando mononucleotídeos, o número máximo de repetições identificadas foi 14, seguido de 12 para di, tri e tetranucleotídeos. O motivo de repetição mais frequente dentre os mononucleotídeos foi A/T (Figura 8), característica que já foi demonstrada em plastomas de outras espécies de plantas (Rodrigues et al., 2020). Dentre os SSR identificados, 49 encontram-se em regiões intergênicas e 15 estão em regiões codantes. Foram identificados 14 SSR dentro de íntrons. Considerando-se a estrutura cloroplastidial, a maioria dos SSR identificados (60) encontram-se na região LSC (Figura 9, Apêndice 4), o que também foi observado em espécies como pequi; ipê-rosa e barbatimão (Souza et al., 2019; Nunes et al., 2019; Sobreiro et al., 2020).

A quantidade de repetições SSR em *E. dysenterica* foi baixa quando comparada a outras espécies, incluindo espécies de *Myrtaceae*, porém, o padrão dessas repetições, no que diz respeito à frequência de motivos de repetição e à localização dessas repetições entre as espécies mostrou-se conservada (Souza et al., 2019, Antunes et al., 2020b; Rodrigues et al., 2020; Sobreiro et al., 2020;). As regiões SSR obtidas neste estudo podem ser testadas quanto à sua aplicação como marcadores moleculares podendo ser utilizados em análises de diversidade genética, filogenia e filogeografia para esta e outras espécies da família *Myrtaceae* (Rabelo et al.; 2011; Soares et al., 2012; Nunes et al., 2020).

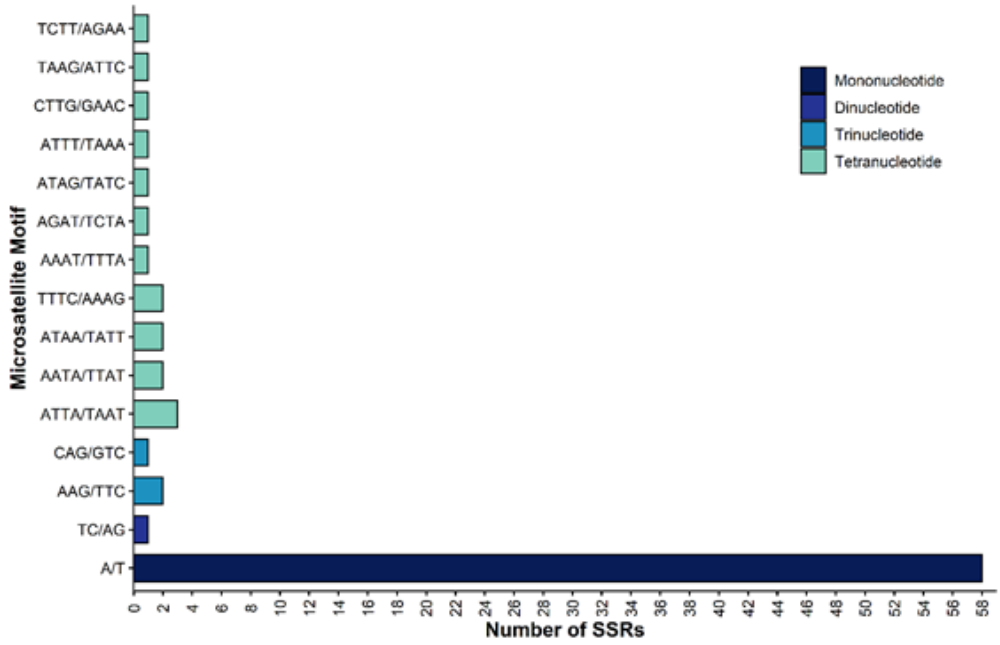

Figura 8. Número e tipo de repetições de sequências simples (SSRs) no genoma cloroplastidial de *Eugenia dysenterica.*

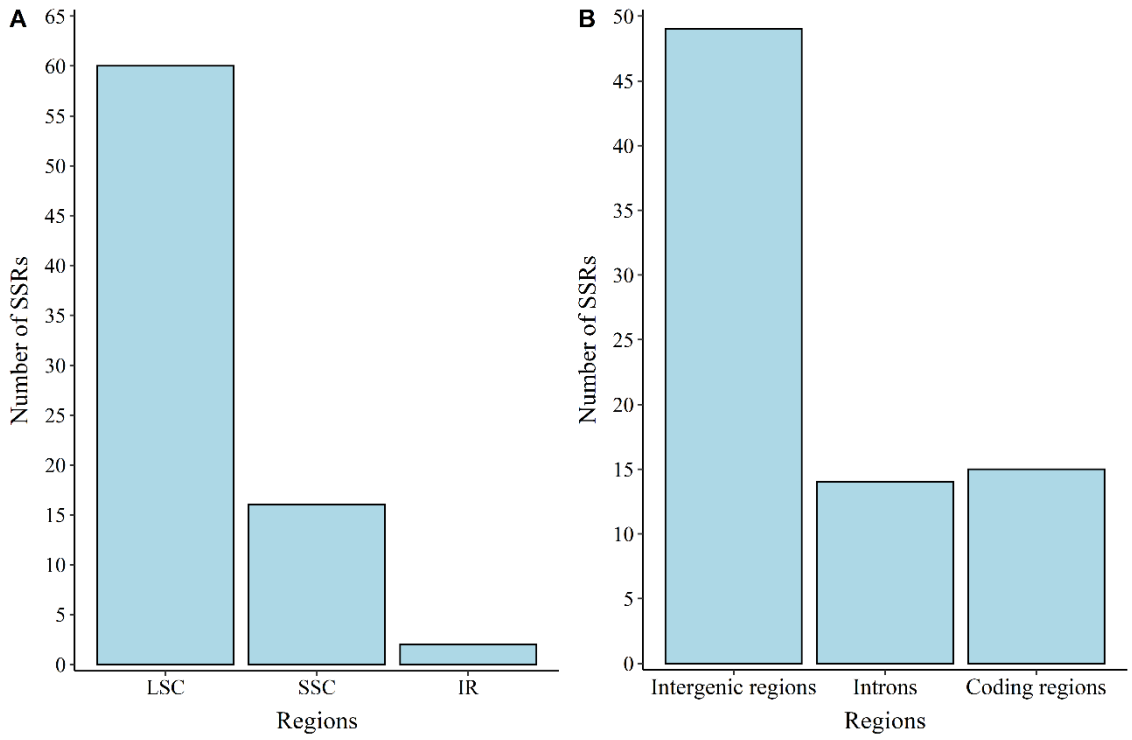

Figura 9. Distribuição das repetições SSR no genoma cloroplastidial de *Eugenia dysenterica*, de acordo com as regiões do cloroplasto (A) e de acordo com as estruturas gênicas (B).

4.1.3 Diversidade nucleotídica e razão entre as taxas de substituições não-sinônimas e sinônimas (Ka/Ks)

A diversidade nucleotídica média estimada (π) entre os cinco genomas de cloroplastos de espécies de *Eugenia* foi de 0,0064, variando de 0,0000 a 0,0315. As seis regiões com maior diversidade foram *trnH-GUG-psbA, trnG-UCC-trnR-UCU, rps16-trnQ-UUG*, *accD-psaL*, *ndhF-rpl32* e *ycf1* (Figura 10). Resultado semelhante foi observado por Sobreiro et al. (2020) em ipê-rosa e por Rodrigues et al. (2020) em diferentes espécies de *Myrtaceae*, onde os genes *accD*, *ndhF* e *ycf1* também apresentaram as maiores taxas de diversidade nucleotídica.

Os locos *trnH-GUG-psbA*, *trnG-UCC-trnR-UCU*, *rps16-trnQ-UUG* e *accD-psaL* estão localizados no espaçador intergênico da região LSC, enquanto os locos *ndhF-rpl32* e *ycf1*estão na região SSC (Figura 10). As estimativas de diversidade nucleotídica das regiões LSC, SSC e IR variaram de 0,0007 a 0,0315, de 0,0033 a 0,0253, e de 0,0000 a 0,0107, com uma média de 0,0082, 0,0109 e 0,0016, respectivamente. Este resultado confirma o que já foi observado para outras espécies de plantas, cujas menores taxas de diversidade nucleotídica encontram-se nos genes das regiões IR do cloroplasto, sugerindo

uma maior conservação dos genes nessas regiões (Antunes et al., 2020a; Machado et al., 2020).

Nos genomas cloroplastidiais de *E. dysenterica* e espécies relacionadas, a ORF hipotética conservada *ycf1* também apresentou a maior taxa de substituições não-sinônimas (Ka) (0,0124), enquanto o gene *rpl32*, que codifica uma proteína relacionada à subunidade maior do ribossomo, teve a maior taxa de substituições sinônimas (Ks) (0,0494) (Apêndice 6).

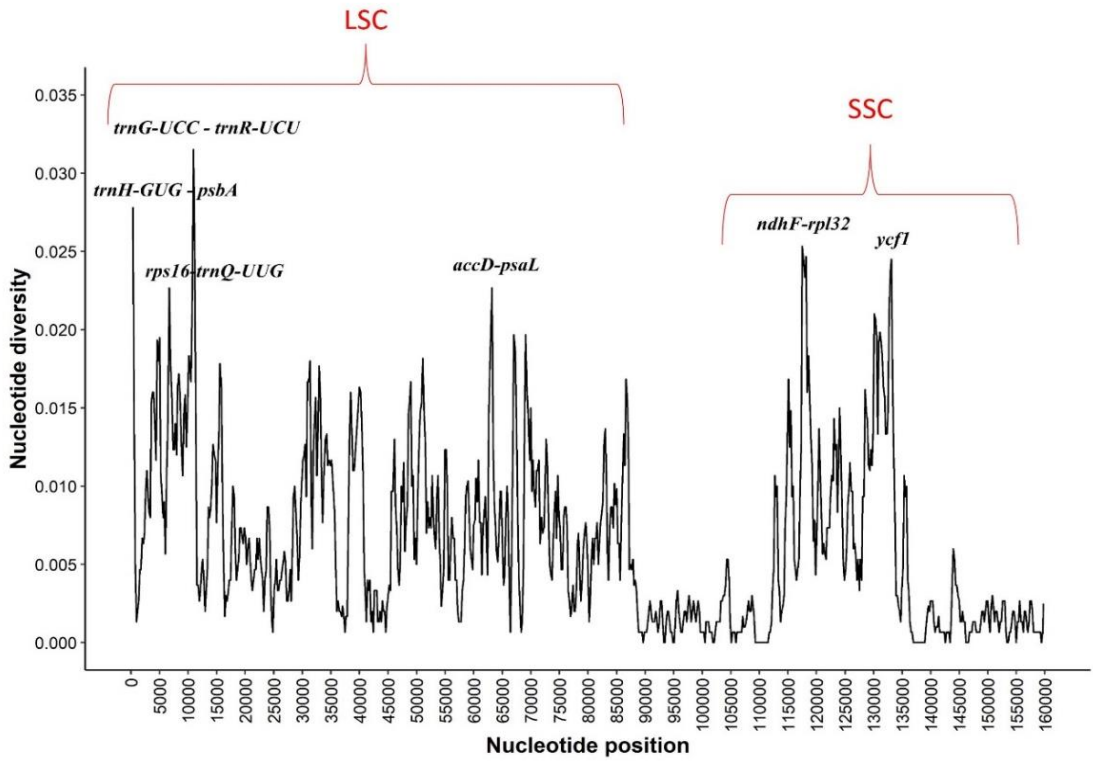

Figura 10. Taxas de diversidade nucleotídica ao longo do genoma de cloroplasto de cinco espécies do gênero *Eugenia*, incluindo *Eugenia dysenterica*. Os quatro primeiros locos (da esquerda para a direita) estão localizados em espaçadores intergênicos das regiões LSC e os dois últimos locos estão nas regiões SSC desses genomas. Os picos mais baixos correspondem a regiões de menor diversidade, localizadas nas IRs.

Em regiões codantes, as substituições não-sinônimas modificam a composição de aminoácidos da proteína codificada, enquanto as substituições sinônimas, ou mutações silenciosas, não alteram essa composição. A razão entre as taxas Ka/Ks pode ser utilizada para se determinar se as forças evolutivas que estão atuando entre os genes são comuns às espécies comparadas (Machado et al., 2017).

A razão entre as taxas de substituição não-sinônimas e sinônimas (Ka/Ks) foi estimada para os 78 genes codificadores de proteínas comuns aos genomas de cloroplasto de *E. dysenterica*, *E. brasiliensis*, *E. pyriformis*, *E. selloi* e *E. uniflora* (77 genes + pseudogene *ycf1*) (Apêndice 6). As razões Ka/Ks das cinco espécies de *Eugenia* variaram de 0 a 1,9329 (média = 0,1907). As razões Ka/Ks mais baixas foram observadas em genes que codificam subunidades do fotossistema I e fotossistema II, o complexo citocromo b/f e a protease clp (Apêndice 6). Foram obtidas estimativas de Ka/Ks iguais a 0 para 36 genes, dos quais dois (*psaC* e *ndhI*) estão localizados na região SSC, três na região IR (*rpl23*, *rps7* e *rps12*) e 31 na região LSC (Figuras 9 e 10).

Valores nulos (Ka/Ks=0) sugerem a atuação de seleção purificadora extremamente forte ocorrendo nesses genes. Estimativas de Ka/Ks inferiores a 1,0 foram obtidas para 76 dos 78 genes codificadores de proteínas (Figura 11), sugerindo novamente atuação de seleção purificadora sobre esses genes, resultando na sua conservação durante a história evolutiva do gênero. As estimativas de Ka/Ks sugerem a presença de seleção positiva em apenas dois dos genes analisados (*rbcL e ndhG)* (Figura 11). Esses genes estão relacionados à subunidade maior da proteína RuBisCO, codificada pelo gene *rbcL* (Ka/Ks = 1,2732) e à enzima NADH desidrogenase, codificada pelo gene *ndhG* (Ka/Ks = 1,9329).

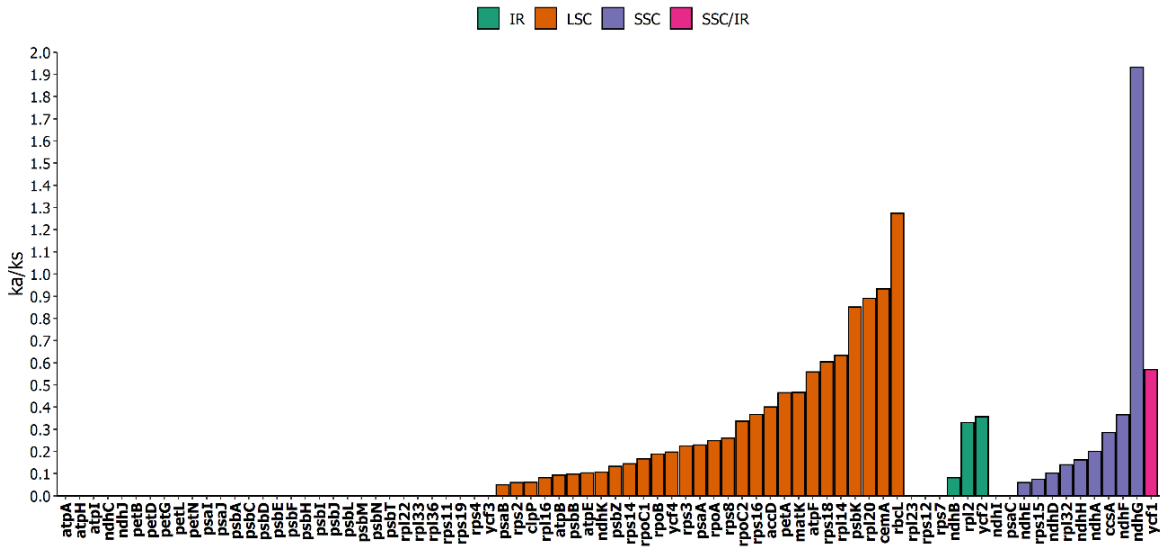

Figura 11. Distribuição das estimativas da razão entre as taxas de substituições não-sinônimas e sinônimas (Ka/Ks) obtidas para os 78 genes (incluindo o pseudogene *ycf1*) codificadores de proteínas comuns aos genomas de cloroplasto de *Eugenia dysenterica*, *Eugenia brasiliensis*, *Eugenia pyriformis*, *Eugenia selloi* e *Eugenia uniflora*.

4.1.4 Genômica comparativa

De modo geral, as características genômicas, como tamanho, estrutura e abundância de genes de cloroplastos são semelhantes nas espécies de *Myrtaceae* descritas anteriormente. Apesar dessa similaridade, a comparação do plastoma de *E.dysenterica* com aqueles das demais espécies do gênero, mostrou que algumas regiões apresentam similaridade inferior. As regiões não codificantes, apresentaram menor conservação, incluindo os pontos de maior divergência entre *E. dysenterica* e as espécies *E. selloi* e *E. uniflora* (Figura 12). Em relação aos genes codificadores de proteínas, os pontos de maior divergência foram observados na região de encontro entre os genes *ycf1* e *ndhF* para todas as espécies (Figura 12).

Com relação as regiões de junção de IRs, a maior diferença estrutural entre o genoma cloroplastidial de *E. dysenterica* e aqueles das demais espécies analisadas está na região de LSC/IRb, onde o gene *rps19* ultrapassa o limite da junção em 31 pb (Figura 12). A incorporação do gene *rps19* nessa região foi observada por Machado et al (2020), porém com comprimentos menores para espécies de *Eugenia* e iguais em espécies do gênero *Plinia* e *Psidium*.

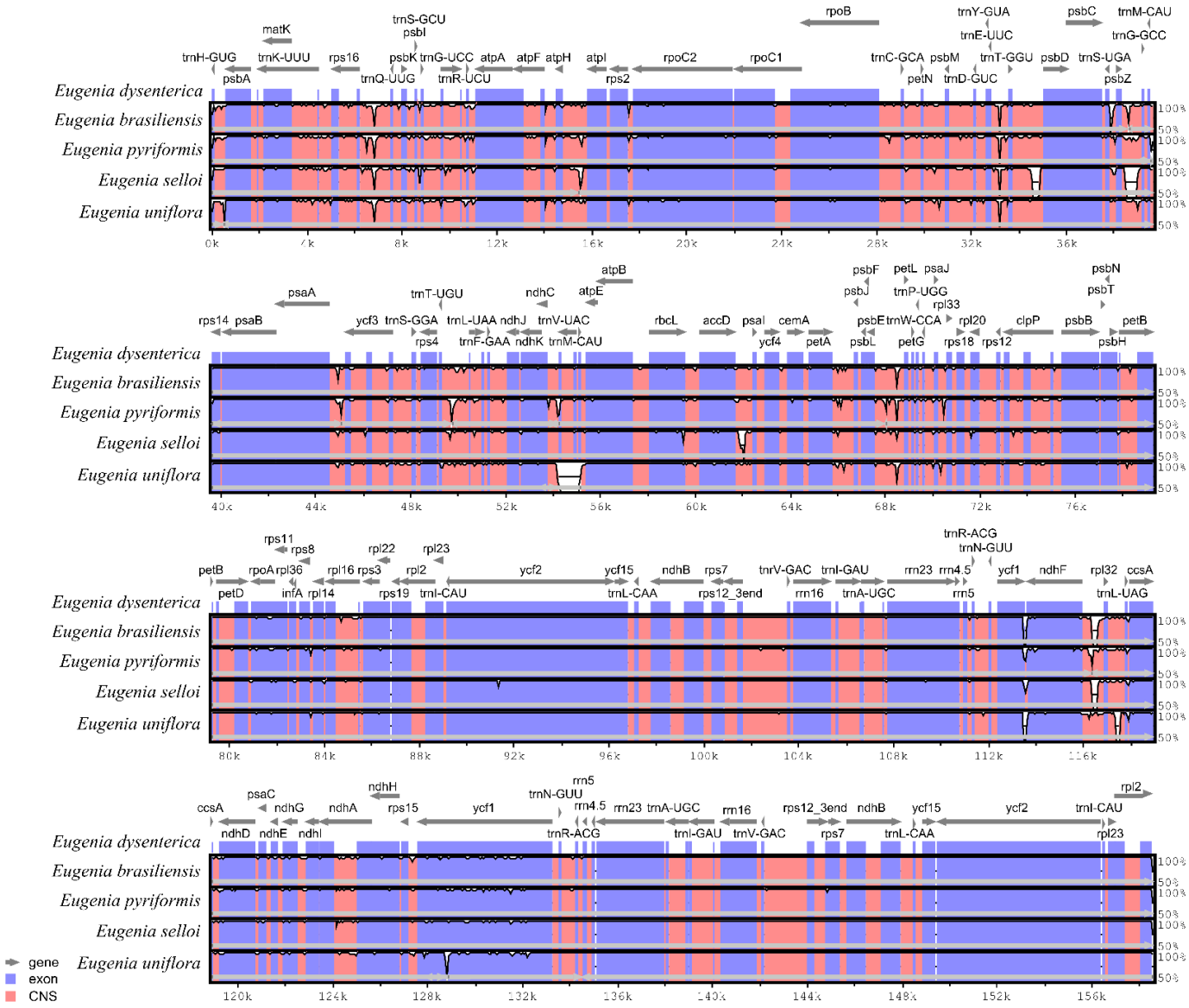

Figura 12. Comparação de estrutura dos genomas cloroplastidiais de *Eugenia dysenterica* e aqueles de outras quatro espécies do gênero *Eugenia*. As regiões codificantes (em azul) apresentam-se mais conservadas do que as regiões não codificantes (em vermelho).

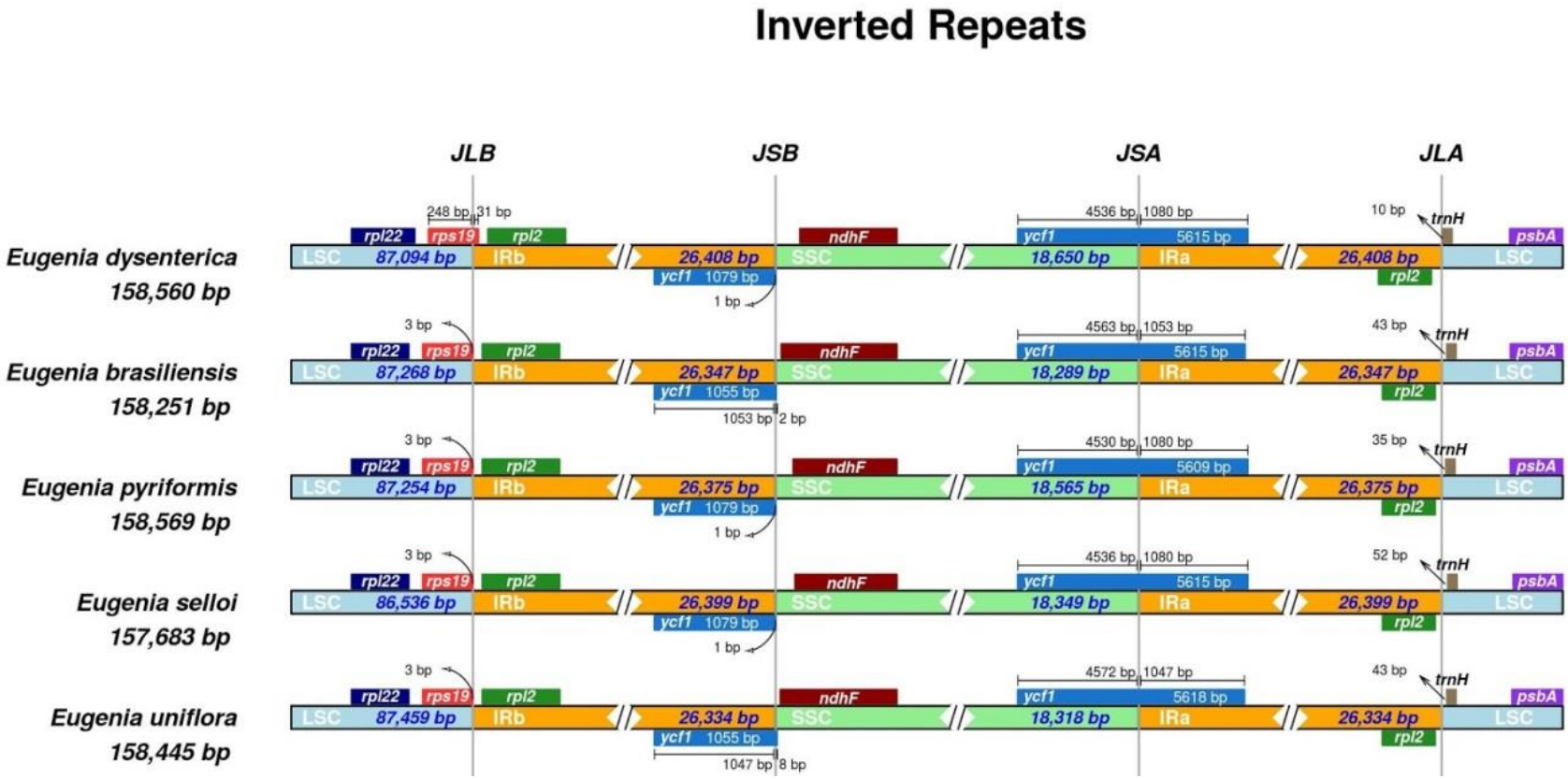

Figura 13. Comparação dos locais de junção entre as regiões LSC (em azul claro), SSC (em verde claro), IRa e IRb (em laranja) entre os genomas de cloroplasto de cinco espécies de *Myrtaceae*: *Eugenia dysenterica*, *Eugenia brasiliensis*, *Eugenia pyriformis*, *Eugenia selloi* e *Eugenia uniflora*. Os códigos JLB (IRb/LSC), JSB (IRb/SSC), JSA (SSC/IRa) e JLA (IRa/LSC) identificam os locais de junção entre as regiões correspondentes nos genomas.

Estudos anteriores sugerem que regiões IRs estão frequentemente sujeitas a expansão, contração ou até mesmo perda completa, eventos que podem ocorrer várias vezes durante a evolução das plantas terrestres, podendo atingir apenas uma ou algumas espécies dentro de um gênero. Segundo Wicke et al. (2011), quando envolvem pequenas quantidades de pares de bases, essas alterações costumam não ter efeitos significativos no tamanho dos genomas cloroplastidiais.

Os limites IRB/SSC foram incorporados no pseudogene *ycf1*, variando de um a oito pb em todas as espécies analisadas (Figura 13). Já o limite SSC/IRa foi incorporado em *ycf1*, com um comprimento de 1047 a 1080 pb (Figura 13). Esses resultados demonstram uma conservação de IR dentro do gênero e pode ser considerado uma das razões para a variação do tamanho do genoma entre as espécies, assim como relatado em estudos anteriores (Machado et al., 2020).

4.1.5 Análise filogenética

A análise filogenética confirmou a monofilia da família *Myrtaceae* e as relações filogenéticas entre as tribos Myrteae, Eucalypteae e Syzygieae dentro dessa família, assim como observado por Eguiluz et al. (2017) e Machado et al. (2020) (Figura 14). Os dados cloroplastidiais permitiram identificar a posição filogenética de Eugenia dysenterica como sendo grupo irmão ao de todas as outras espécies do gênero amostrados (Figura 14). Também foi possivel observar que *E. dysenterica* tem uma relação de parentesco mais próxima com *Plinia trunciflora* e relações mais distantes com as espécies do grupo externo (Tribo Eucalypteae), conforme esperado, baseando-se em observações já relatadas na literatura (Bayly et al., 2013; Liang et al., 2021).

De maneira geral, os resultados mostram que *E. dysenterica* está intimamente relacionada a todas as espécies analisadas, com fortes valores de *bootstrap* (87,7% a 100%) (Figura 14). A análise ainda demonstrou que os oito taxons do grupo interno (*Syzygium forestii*, *Rhodomyrtus tomentosa*, *Plinia trunciflora*, *Eugenia dysenterica*, *E. selloi*, *E. pyriformis*, *E. brasiliensi*s e *E. uniflora*) foram divididos em clados que estão de acordo com os propostos para *Myrtaceae*. Os dados obtidos somam informações importantes para o desenvolvimento de estudos evolutivos e filogenéticos, porém, sequências de plastomas que abordem outras tribos de *Myrtaceae* serão importantes para resolver as questões filogenéticas que ainda se encontram pendentes nesta família (Li et al., 2021).

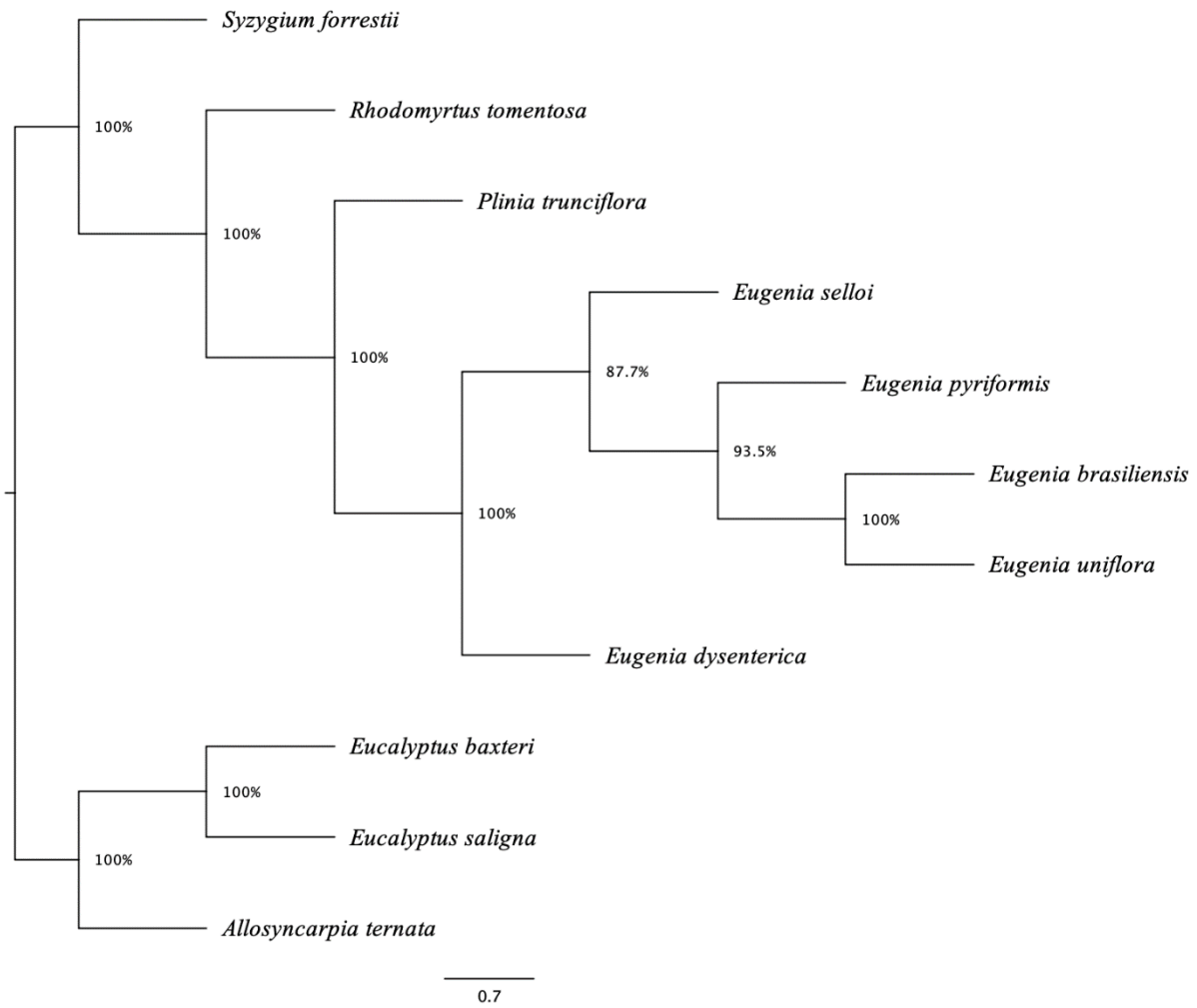

Figura 14. Árvore filogenética obtida pelo método de máxima verossililhança (ML) a partir da análise de 78 genes cloroplastidiais codificadores de proteínas comuns às espécies incluídas. Foram utilizadas 11 espécies, incluindo *E. dysenterica*, representando três das tribos que compõem a família *Myrtaceae* (Myrteae, Eucalypteae e Syzygieae.) Os números representam os valores *bootstrap* para cada nó. À direita estão indicadas as espécies utilizadas na comparação, bem como as tribos, diferenciadas pelas barras coloridas.

4.2 TRANSCRITOMA DE REFERÊNCIA

4.2.1 Sequenciamento e montagem do transcritoma de referência de *Eugenia dysenterica*

O sequenciamento gerou um total de 664 milhões de pares de *reads* com 100 pb, totalizando 132,80 Gb de sequências (Tabela 5). De acordo com estatísticas fornecidas pelo MultiQC (Ewels et al., 2016), a maioria dos *reads* (aproximadamente 70%) foram duplicados durante o sequenciamento e aproximadamente 30% dos *reads* apresentaram-se em cópia única. Com a montagem do transcritoma foram obtidos 171.070 transcritos e 43.605 genes, com média de 3,9 isoformas por gene. O tamanho médio dos transcritos foi de 1.737 pb e o valor N50 obtido com a montagem foi de 2.288 pb (Tabela 6).

Tabela 5. Estatísticas descritivas dos dados brutos de sequenciamento utilizados na montagem do transcritoma de referência de *Eugenia dysenterica*.

Biblioteca	Nº de pares de *reads* (milhões)	Volume de dados (Gb)
FM01	329,6	65,92
PC01	65,0	13,00
PC02	73,1	14,62
PC03	71,5	14,30
PS02	59,2	11,84
PS03	65,6	13,12
TOTAL	664,0	132,80

FM: Folha madura, PC: Plântulas submetidas a condições favoráveis em termos de suprimento de água, PS: Plântulas submetidas a condições de estresse hídrico.

O número de transcritos obtidos para *E. dysenterica* foi maior que os encontrados para outras três espécies da família *Myrtaceae* (*Arillastrum gummiferum* – 117.839, *Syzygium longifolium* - *89.782* e *Tristaniopsis glauca* - 108.823) e menor que os valores encontrados para *Eugenia uniflora* e *Psidium cattleyanum* (304.425 e 301.058, respectivamente). Por outro lado, o número de transcritos identificados neste trabalho assemelha-se aos encontrados para *Melaleuca quinquenervia* e *Syzygium samarangens* (192.557 e 134.199, respectivamente) (Tabela 7) (Guzman et al., 2014; Chen et al., 2017; Hsieh et al., 2018; Soewarto et al., 2019; Veto et al., 2020).

O número de transcritos identificados neste estudo também foi menor que o observado por Barros-Ribeiro (2016), em uma amostra do genoma nuclear de *E. dysenterica* (228.510 transcritos). Esse resultado já era esperado, levando em consideração que o tipo de dado e a estratégia utilizada neste trabalho (RNAseq) permitem a obtenção de sequências menos fragmentadas que aquelas obtidas através da montagem de genomas utilizando apenas *short-reads* de DNA nuclear.

As regiões codantes (CDS) apresentaram os maiores comprimentos, podendo atingir até 5.500 pb, seguidas das regiões UTR-3' com comprimentos próximos a 4.000 pb, sendo que a maioria delas contém cerca de 500 pb (Figura 15). Já as regiões UTR-5' anotadas possuem comprimentos de até 3.000 pb.

Tabela 6. Estatísticas descritivas do transcritoma de referência de *Eugenia dysenterica*, obtido na análise com o Trinity. Foram considerados para compor o transcritoma de referência os *contigs* maiores do que 500 pb.

Parâmetros	Valores observados
Número total de *contigs*	171.070
Tamanho médio dos *contigs* (pb)	1.737
Tamanho total do *assembly* (Mb)	68,60
N50 (pb)	2.288
Número de genes	43.605
Número de transcritos	171.070
GC%	45,07

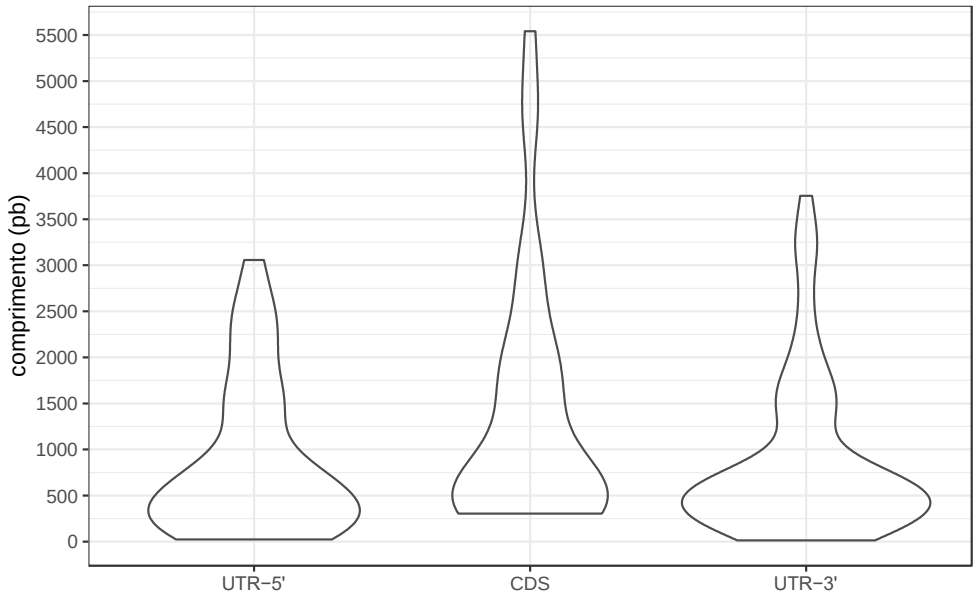

Figura 15. Distribuição dos tamanhos das regiões anotadas como *UTR-5'* (*5'-Untranslated Region*), *CDS* (*Coding DNA Sequence*) e *UTR*-3' (*3'-Untranslated Region*) no transcritoma de referência obtido para *Eugenia dysenterica*.

Tabela 7. Aspectos gerais dos transcritomas de *Eugenia dysenterica* e outras espécies da família *Myrtaceae*.

Espécie	Nº de transcritos	Tamanho médio dos transcritos (pb)	Nº de genes	Conteúdo GC(%)	Citações
Arillastrum gummiferum	117. 839	843	84.919	45,86	Soewarto et al. (2019)
Eugenia dysenterica	171.070	1737	43.605	45,07	Barros-Ribeiro (2021) – este estudo
Eugenia uniflora	304.425	702	72.742	-	Guzman et al. (2014)
Melaleuca quinquenervia	192.557	1474	-	43,45	Hsieh et al. (2018)
Psidium cattleyanum	301.058	889	282.768	49,17	Veto et al. (2020)
Syzygium longifolium	89.782	867	64.716	46,37	Soewarto et al. (2019)
Syzygium samarangense	134.199	1506	54.536	49,93	Chen et al. (2017)
Tristaniopsis glauca	108.823	876	76.982	44,45	Soewarto et al. (2019)

Considerando outras espécies de plantas não-modelo, a quantidade de transcritos identificados em *E. dysenterica* foi menor que o observado em espécies *Kaya grandifoliola* - *Meliaceae* (116.289) e *Arundo donax* - *Poaceae* (111.749), tendo, contudo, o número de genes identificados no transcritoma de *E. dysenterica* (45.605) sido muito semelhante ao encontrado para esta última espécie (45.821) (Fu et al., 2016; Silva-Junior et al., 2018, Soares, 2019). Aspectos como tamanho médio dos transcritos e conteúdo GC apresentam-se semelhantes com pequenas variações entre *E. dysenterica* e demais espécies de plantas citadas (Tabela 7).

A variação observada na quantidade de transcritos e genes entre as espécies citadas pode estar relacionada tanto à ocorrência de *splicings* alternativos, quanto ao tipo de amostragem feita para cada estudo, levando-se em consideração que em todos os trabalhos comparados, a tecnologia de sequenciamento utilizada foi basicamente a mesma (Tabela 7). Já as diferenças em numero de transcritos ligadas ao tipo de amostra utilizada, podem ser justificadas pelo fato de que determinados genes podem ou não ser transcritos. Os transcritos, por sua vez, podem se expressar de diferentes formas, dependendo do tecido amostrado, estágio de desenvolvimento e condições de estresse aos quais a planta foi submetida (Feng et al., 2019; Soewarto et al., 2019).

A partir de 2.326 genes ortólogos conservados entre *E. dysenterica* e outras espécies de plantas eudicotiledôneas, identificados com o software BUSCO, 2.245 tiveram suas sequências completamente montadas no transcritoma de referência, o que corresponde a mais de 90% dos ortólogos identificados (Tabela 8). Também foi possível observar que a maioria dos genes completos (2.034 genes) aparecem em mais de uma cópia e apenas 211 aparecem em cópias únicas no transcritoma obtido (Tabela 8).

Segundo Seppey et al. (2019), as altas taxas de duplicação observadas em montagem de transcritomas podem ser causadas pela presença de diferentes haplótipos, eventos de duplicação recente de todo o genoma ou artefatos técnicos que devem ser investigados. Neste caso especificamente, tal resultado é explicado pela presença dos múltiplos transcritos identificados para cada um dos genes, no transcritoma de referência.

Tabela 8. Resultados da análise de completitude realizada com o *software* BUSCO (*Benchmarking Universal Single-Copy Orthologs*), relativa à identificação de 2.326 genes ortólogos conservados em eudicotiledôneas no transcritoma de referência obtido para *Eugenia dysenterica*.

Genes completos identificados		2.245 (96,5%)
	Em cópia única	211 (9,1%)
	Duplicados	2.034 (87,4%)
Genes fragmentados		40 (1,7%)
Genes não identificados		41 (1,8%)

4.2.2 Anotação funcional do transcritoma de referência

A anotação funcional dos transcritos obtidos para *E. dysenterica* foi feita com o *software* Trinotate, a partir da busca por sequências de proteínas similares nos bancos de dados *Pfam* e *SwissProt* (http://www.uniprot.org/uniprot/). As sequências que apresentaram similaridade foram anotadas com base nos domínios estabelecidos pelo *Gene Onthology* (GO) (Figuras 15, 16 e 17): Componente Celular (CC), Função molecular (FM) e Processo Biológico (PB).

Considerando a anotação para o domínio CC (Figura 16), que abrange locais da célula onde os produtos gênicos são ativados, é possível se observar que a maior parte dos transcritos anotados pertencem às categorias ontológicas relacionadas à estrutura anatômica intercelular (GO: *Intercellular Anatomical Structure*, 15.694 transcritos), citoplasma (GO: *Cytoplasm*, 12.303 transcritos), composição de membrana (GO: *Membrane*, 8.415 transcritos) e núcleo (GO: *Nucleus*, 6.841 transcritos). Ainda no domínio CC, os transcritos menos frequentes foram relacionados à composição de organelas como lisossomos e à membrana nuclear (158 e 195 transcritos, respectivamente).

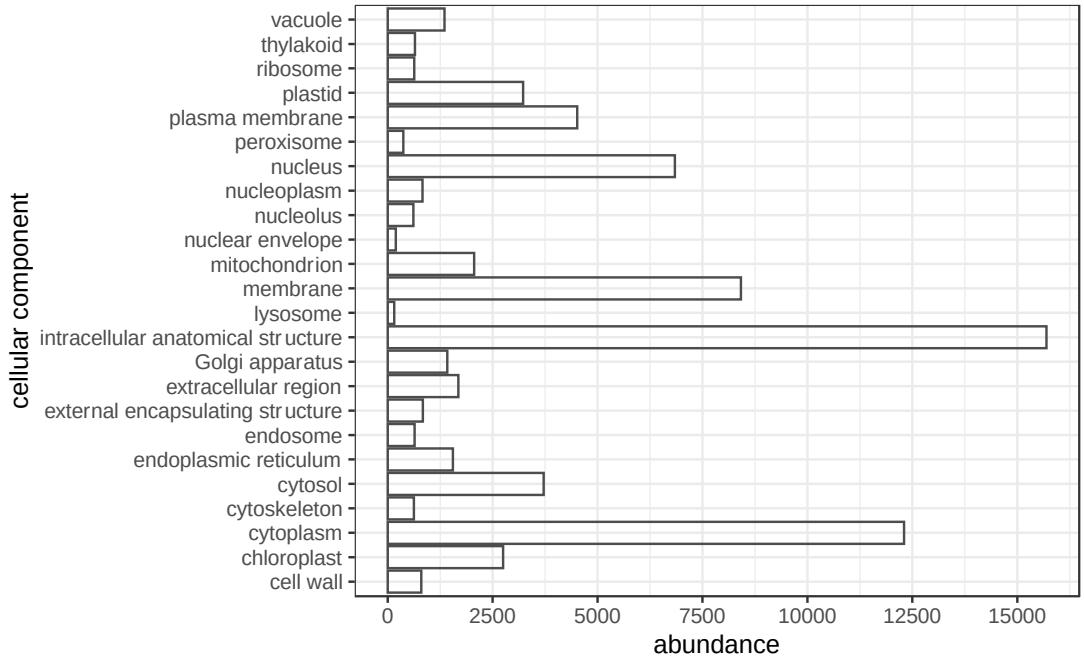

Figura 16. Classificação GO dos transcritos preditos no transcritoma de referência de *Eugenia dysenterica*, considerando sua abundância dentro do domínio Componente Celular (CC).

Para o domínio FM (Figura 17), que representa a atividade bioquímica do produto gênico, a maior parte dos transcritos foram anotados nas classes *Binding* e *Catalytic Activity* (14.526 e 11.618 transcritos, respectivamente). Já para o domínio PB (Figura 18), que define como o produto gênico está contribuindo para o funcionamento do organismo, a maioria dos transcritos foram anotados nos GOs: *Cellular Process* e *Metabolic Process* (16.364 e 13.689 transcritos, respectivamente), enquanto processos relacionados às classes *Cell-Signaling* e *Fruit Ripenig* reuniram a menor quantidade de anotações (112 e 59 transcritos, respectivamente).

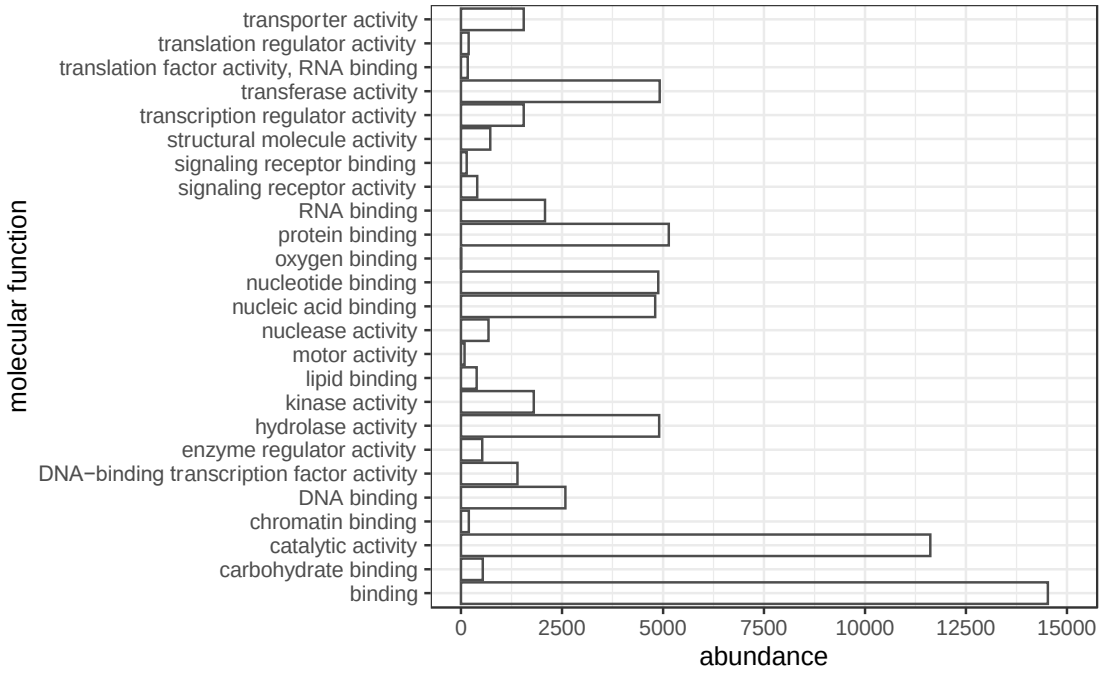

Figura 17. Classificação GO dos transcritos preditos no transcritoma de referência de *Eugenia dysenterica*, considerando sua abundância dentro do domínio Função Molecular (FM).

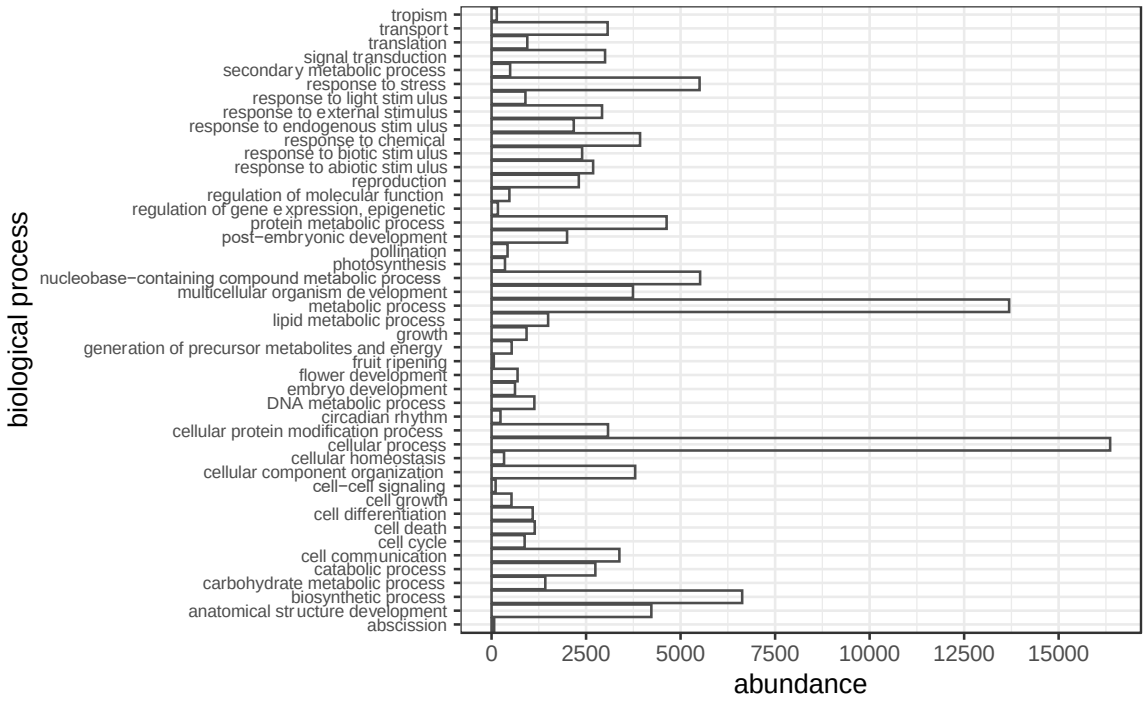

Figura 18. Classificação GO dos transcritos preditos no transcritoma de referência de *Eugenia dysenterica*, considerando sua abundância dentro do domínio Processo Biológico (PB).

4.2.3 Identificação de SNPs no transcritoma de *E. dysenterica*

O alinhamento dos *reads* das bibliotecas construídas com base nas amostras de RNA dos cinco indivíduos permitiu a identificação de um total de 1.147.895 SNPs iniciais, que, após a aplicação dos filtros de controle de qualidade resultaram na obtenção de 636.269 SNPs (Tabela 9). Estes SNPs deverão ser validados em trabalhos futuros e poderão ser utilizados para o desenvolvimento de marcadores moleculares voltados para estudos com *E. dysenterica*, em escala genômica.

A densidade estimada de SNPs/gene foi de 19,37. Cerca de 75% dos genes presentes no transcritoma de referência apresentaram pelos menos um SNP ao longo dos seus transcritos, resultado que corrobora com o observado por Nunes (2015), ao identificar ~999.000 SNPs genômicos em *E. dysenterica*, verificando sua maior frequência de ocorrência em regiões gênicas. A relação de substituições do tipo transição sobre aquela do tipo transversão (Ts/Tv) foi de 1,91, dentro da faixa esperada para SNPs em regiões codantes. Esse valor também se assemelha ao observado no trabalho citado anteriormente, utilizando dados de genoma de *E. dysenterica* (Nunes, 2015).

Tabela 9. Número e distribuição dos SNPs identificados no transcritoma de *Eugenia dysenterica*.

Parâmetros	Valores observados
Número total de genes	43.605
Número de genes com SNPs	32.853
% de genes com SNPs	75,34
Nº total de SNPs identificados	636.269
Nº de SNPs identificados/gene	19,37
Número de transições (Ts)	368.799
Número de transversões (Tv)	193.431
Relação Ts/Tv	1,91

A relação Ts/Tv identificada neste estudo assemelha-se a observada para quatro diferentes espécies de árvores neotropicais cujos transcritomas foram caracterizados por Broussau et al. (2014), sendo que em *E. dysenterica* mostrou-se um pouco maior. A densidade de SNPs observada aqui também se mostrou maior que a observada no mesmo

estudo citado acima, onde os autores reforçam que essa característica pode ter relação direta com o rigor dos critérios de filtragem aplicados (Broussau et al., 2014).

Uma avaliação completa da diversidade genética das espécies de árvores tropicais, como é o caso de *E. dysenterica* requer grandes quantidades de dados genômicos e marcadores moleculares, nesse contexto os SNPs identificados aqui, após serem validados, servirão como recurso fundamental para o desenvolvimento desse tipo de estudo.

5 CONCLUSÃO

O genoma cloroplastidial de *E. dysenterica* apresenta uma estrutura geral análoga a outras plantas terrestres descritas na literatura, tendo também seu conteúdo gênico conservado em maioria.

Foram identificadas 33 repetições sendo a maioria palindrômicas, além de 78 regiões SSR que poderão ser testadas quando à sua aplicação como marcadores moleculares.

A análise filogenética realizada neste trabalho corrobora com as informações descritas na literatura, reforçando a relação próxima de *E. dysenterica* com outras espécies da família *Myrtaceae* e fornece subsídio para desenvolvimento de estudos filogenéticos utilizando outras espécies do gênero *Eugenia*.

A caracterização do genoma cloroplastidial da cagaiteira fornece informações que poderão ser utilizadas em estudos de genômica populacional, filogeografia e filogenética, e constitui o quarto genoma cloroplastidial de uma espécie nativa do Cerrado a ser sequenciado, o que também contribui para a elucidação de estudos voltados à conservação da flora endêmica desse bioma.

A caracterização do transcritoma de *E. dysenterica*, utilizando a abordagem de RNAseq, é inédita dentre as espécies nativas do Cerrado e possibilitou a identificação de 171.070 transcritos, de 43.605 genes, anotados com base em diferentes bancos dados de referência. Também foram identificados 636.269 SNPs que poderão ser validados em estudos futuros.

Os recursos genômicos obtidos com este trabalho se constituem em um passo importante para o desenvolvimento de estratégias eficientes de cultivo e conservação de *E. dysenterica*, além de servir como subsídio para estudos de outras espécies vegetais não-modelo, o que tem sido possibilitado pelo contínuo desenvolvimento das tecnologias de sequenciamento e ferramentas de análise de dados.

6 REFERÊNCIAS BIBLIOGRÁFICAS

ALMEIDA JÚNIOR, E. B. DE; CHAVES, L. J.; SOARES, T. N. Caracterização genética de uma coleção de germoplasma de cagaiteira, uma espécie nativa do cerrado. **Bragantia**, v. 73, n. 3, p. 246–252, 2014.

ALONSO-HERRADA, J.; URRUTIA, I.; ESCOBAR-FEREGRINO, T. et al. Sequencing of Non-model Plants for Understanding the Physiological Responses in Plants. **Plant Genomics**, 2016.

AMIRYOUSEFI, A.; HYVÖNEN, J.; POCZAI, P. IRscope: an online program to visualize the junction sites of chloroplast genomes. **Bioinformatics**, v. 34, n. 17, p. 3030-3031, 2018.

AMOS, W.; FLINT, J.; XU, X. Heterozygosity increases microsatellite mutation rate, linking it to demographic history. **BMC genetics**, v. 9, p. 72, 2008.

ANTUNES, A. M.; SOARES, T, N.; TARGUETA, C. P. 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, 2020a.

ANTUNES, A. M.; NUNES, R.; NOVAES, E.; COELHO, A. S. G.; Large number of repetitive elements in the draft genome assembly of *Dipteryx alata* (Fabaceae). **Genetics and Molecular Research**, v. 19, n. 2, 2020b.

ANTUNES, A. M.; Targueta, C. P.; Castro, A. A. et al. Genome size and chromosome number of *Dipteryx alata* (Leguminosae): A model candidate for comparative genomics in papilionoideae. **Genetics and Molecular Research**, v. 19, n. 3, p. 1–6, 2020c.

ANDREWS, S.: FastQC. A quality control tool for high throughput sequence data. Disponível em: < http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. Acesso em: 17 fev. 2021.

ASIF, H.; KHAN, A.; IQBAL, A. et al. The chloroplast genome sequence of *Syzygium cumini* (L.) and its relationship with other angiosperms. **Tree genetics & genomes**, v. 9, n. 3, p. 867-877, 2013.

BAGSHAW, A.T. M. Functional mechanisms of microsatellite DNA in eukaryotic genomes. **Genome biology and evolution**, v. 9, n. 9, p. 2428-2443, 2017.

BARROS-RIBEIRO, S. **Sequenciamento e caracterização parcial do genoma de cagaiteira (*Eugenia dysenterica* DC.).** 2016. 77 f. Dissertação (Mestrado em Genética e Melhoramento de Plantas) – Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2016.

BASANTANI, M. K.; GUPTAA, D.;MEHROTRA, R. et al. An update on bioinformatics resources for plant genomics research. **Current Plant Biology**, v. 11, p. 33-40, 2017.

BAYLY, M.J.; RIGAULT, P.; SPOKEVICIUS, A.; et al. Chloroplast genome analysis of Australian eucalypts- *Eucalyptus, Corymbia, Angophora, Allosyncarpia and Stockwellia* (*Myrtaceae*). **Mol Phylogenet Evol**. v. 69, n. p. 3704-16, 2013.

BEIER, S.; THIEL, T.; MÜNCH, T.; SCHOLZ, U. MASCHER, M. MISA-web: a web server for microsatellite prediction, **Bioinformatics**, v. 33, n. 16, p. 2583–2585, 2017.

BHARGAVA, A.; FUENTES, F. F. Mutational dynamics of microsatellites Molecular Biotechnology. **Molecular biotechnology**, v. 44, n. 3, p. 250-266, 2010.

AGILENT. **Bioanalyser system**. Disponível em: https://www.agilent.com/en/product/automated-electrophoresis/bioanalyzer-systems/bioanalyzer-instrument. Acesso em 01 mar 2021.

BIROL, I.; RAYMOND, A., JACKMAN, S. D. ET al. Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. **Bioinformatics**, v. 29, n. 12, p. 1492–1497, 2013.

BOAVENTURA-NOVAES, C. R. D.; NOVAES, E.; MOTA, E. E. S.; et al. Genetic drift and uniform selection shape evolution of most traits in *Eugenia dysenterica* DC. (*Myrtaceae*). **Tree Genetics & Genomes**. v. 14, p. 76-76, 2018a

BOAVENTURA-NOVAES, C. R. D.; MOTA, E. E. S.; NOVAES, E. S; et al. Structure of the phenotypic variability of fruit and seed traits in natural populations of *Eugenia dysenterica* DC. (*Myrtaceae*). **Revista brasileira de fruticultura** (ONLINE), v. 40, p. e, 2018b.

BOLGER, M. E.; ARSOVA, B.; USADEL, B. Plant genome and transcriptome annotations: From misconceptions to simple solutions. **Briefings in Bioinformatics**, v. 19, n. 3, p. 437–449, 2018.

CAMPBELL, M. S.; LAW, M., HOLT, C.; et al. MAKER-P: A Tool Kit for the Rapid Creation, Management , and Quality Control of Plant. **Plant physiology**. v. 164, n. February, p. 513–524, 2014.

CAMILO, Y. M. V.; SOUZA, E. R. B. D.; VERA, R.; NAVES, R. V. Fenologia, produção e precocidade de plantas de *Eugenia dysenterica* visando melhoramento genético. **Revista de Ciências Agrárias**, v.36, v.2, p.192-198, jan. 2013.

CARDOSO, L. D. M. et al. Cagaita (*Eugenia dysenterica DC*.) of the Cerrado of Minas Gerais, Brazil: Physical and chemical characterization, carotenoids and vitamins. **Food Research International**, v. 44, n. 7, p. 2151–2154, 2011.

CHAVES, L. J.; TELLES, M. P. C. de. **Cagaita**. In: Frutas Nativas da Região CentroOeste do Brasil. 1.ed. Brasília: Embrapa, 2006, cap. 7, p. 120-134

CHEN, F., HAO, Y., YIN.; et al. Transcriptome of wax apple (*Syzygium samarangense*) provides insights into nitric oxide-induced delays of postharvest cottony softening. **Acta**

Physiologiae Plantarum, v. 39, n. 12, p. 273, 2017.

CHEN, F., DONG, W., ZHANG, J.; et al. The sequenced angiosperm genomes and genome databases. **Frontiers in plant science**, v. 9, p. 418, 2018.

DANIELL, H.; LIN, C.S.; YU, M.; CHANG, W. J. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. **Genome biology**, v. 17, n. 1, p. 1-29, 2016.

DAS, S. & BANSAL, M. Variation of gene expression in plants is influenced by gene architecture and structural properties of promoters. **PloS one**, v. 14, n. 3, p. e0212678, 2019.

DENG, P., LIU, S., NIE, X.; et al. Conservation analysis of long non-coding RNAs in plants. **Science China Life Sciences**, v. 61, n. 2, p. 190-198, 2018.

DINIZ-FILHO, J. A. F.; BARBOSA, A. C. O. F.; COLLEVATTI, R. G.; et al. Spatial autocorrelation analysis and ecological niche modelling allows inference of range dynamics driving the population genetic structure of a Neotropical savanna tree. **Journal of Biogeography**, v. 43, p. 167-177, 2016.

DOYLE, J.J.; DOYLE J.L. Isolation of plant DNA from fresh tissue. **Focus**, v.12, p.13-15, 1987.

DUAN, H., ZHANG, Q.; WANG, C.; et al. Analysis of codon usage patterns of the chloroplast genome in *Delphinium grandiflorum* L. reveals a preference for AT-ending codons as a result of major selection constraints. **PeerJ**, v. 9, p. e10787, 2021.

EGUILUZ, M.; RODRIGUES, N. F.; GUZMAN, F.; YUYAMA, P.; MARGIS, R. The chloroplast genome sequence from *Eugenia uniflora*, a *Myrtaceae* from Neotropics. **Plant Systematics and Evolution**, v. 303, n. 9, p. 1199-1212, 2017.

EWELS, P.; MAGNUSSON, M.; LUNDIN, S.; KÄLLER, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. **Bioinformatics**, v. 32, n. 19, p. 3047-3048, 2016.

FENG, Jia-Wu et al. Plant ISOform sequencing database (PISO): a comprehensive repertory of full-length transcripts in plants. **Plant biotechnology journal**, v. 17, n. 6, p. 1001, 2019.

FLAVELL, R. B. Perspective: 50 years of plant chromosome biology. **Plant Physiology**, v. 185, n. 3, p. 731-753, 2021.

FINCH, K. N.; JONES, F. A.; CRONN, R. C. Genomic resources for the Neotropical tree genus *Cedrela* (Meliaceae) and its relatives. **BMC Genomics**, v. 20, n. 1, p. 1–17, 2019.

FU, Y., POLI, M., SABLOK, G.; et al., Dissection of early transcriptional responses to water stress in *Arundo donax* L. by unigene-based RNA-seq. **Biotechnology for biofuels**, v. 9, n. 1, p. 1-18, 2016.

GARCIA-MAS, J. et al. The genome of melon (*Cucumis melo* L.). **Proceedings of the**

National Academy of Sciences, v. 109, n. 29, p. 11872–11877, 2012.

GARCIA, S. et al. Plant rDNA database: update and new features**. Database: the journal of biological databases and curation**, v. 2014, p. 1–7, 2014.

GARG, R.; JAIN, M. RNA-Seq for transcriptome analysis in non-model plants. **In: Legume Genomics**. Humana Press, Totowa, NJ. p. 43-58. 2013

GARRIDO-RAMOS, M. A. Satellite DNA in plants: more than just rubbish. **Cytogenetic and Genome Research**, v. 146, n. 2, p. 153-170, 2015.

GRABHERR, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. **Nature biotechnology**, v. 29, n. 7, p. 644, 2011.

GRATTAPAGLIA, D. et al. Progress in *Myrtaceae* genetics and genomics: *Eucalyptus* as the pivotal genus. **Tree Genetics and Genomes**, v. 8, p. 463-508, abr. 2012.

GREINER S, LEHWARK P, BOCK R. OrganellarGenomeDRAW (OGDRAW) version 1.3. 1: expanded toolkit for the graphical visualization of organellar genomes. **Nucleic Acids Research** 47: W59-W64, 2019.

GU, C., TEMBROCK, L. R., JOHNSON, N. G., SIMMONS, M. P., & WU, Z. The complete plastid genome of *Lagerstroemia fauriei* and loss of rpl2 intron from Lagerstroemia (Lythraceae). **PLoS One**, v. 11, n. 3, p. e0150752, 2016.

GUO, Y. L. Gene family evolution in green plants with emphasis on the origination and evolution of *Arabidopsis thaliana* genes. Plant Journal, v. 73, n. 6, p. 941–951, 2013.

GUZMAN, F., KULCHESKI, F. R., TURCHETTO-ZOLET, A. C., & MARGIS, R. De novo assembly of *Eugenia uniflora* L. transcriptome and identification of genes from the terpenoid biosynthesis pathway. **Plant Science**, v. 229, p. 238-246, 2014.

GUYEUX, C. et al. Evaluation of chloroplast genome annotation tools and application to analysis of the evolution of coffee species. **PLoS ONE**, v. 14, n. 6, p. 1–20, 2019.

HAAS, B. J. et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. **Nature protocols**, v. 8, n. 8, p. 1494-1512, 2013.

HAAS, B J.; PAPANICOLAOU, A. **TransDecoder 5.5.0**. Disonível em: https://github.com/TransDecoder/TransDecoder/wiki. Acesso em 20 jun 2021.

HOU, J., LU, D., MASON, A. S., LI, B HSIEH et al. Non-coding RNAs and transposable elements in plant genomes: emergence, regulatory mechanisms and roles in plant development and stress responses. **Planta**, v. 250, n. 1, p. 23-40, 2019

HUANG Y, YANG Z, HUANG S, AN W, LI J, ZHENG X. Comprehensive Analysis of *Rhodomyrtus tomentosa* Chloroplast Genome. **Plantas** (Basileia). 4 de abril de 2019; 8 (4): 89. doi: 10.3390 / plants8040089. PMID: 30987338; PMCID: PMC6524380.

ILLUMINA. High performance, low cost genomics- Hiseq 4000. Disponivel em:

https://www.illumina.com/systems/sequencing-platforms/hiseq-3000-4000.html. Acesso em 25 abr. 2021a.

ILLUMINA. Focused power on the MiSeq System.Illumina. Disponível em: https://www.illumina.com/systems/sequencing-platforms/miseq.html. Acesso em 25 abr. 2021b.

IZUNO, A. et al. Genome sequencing of *Metrosideros polymorpha* (*Myrtaceae*), a dominant species in various habitats in the Hawaiian Islands with remarkable phenotypic variations. **Journal of plant research**, v. 129, n. 4, p. 727-736, 2016.

JIAO, W. B.; SCHNEEBERGER, K. The impact of third generation genomic technologies on plant genome assembly. **Current Opinion in Plant Biology**, v. 36, p. 64–70, 2017.

JOHNSON, M.; ZARETSKAYA, I.; RAYTSELIS, Y.; et al., NCBI BLAST: a better web interface, **Nucleic Acids Research**, v. 36, Issue suppl_2, p. W5 – W9, 2008.

KEARSE M, MOIR R, WILSON A, et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. **Bioinformatics** 28:1647-1649, 2012.

KALENDAR, R., AMENOV, A., & DANIYAROV, A. Use of retrotransposon-derived genetic markers to analyse genomic variability in plants. **Functional Plant Biology**, v. 46, n. 1, p. 15-29, 2019.

KATOH S. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution 30:772-780, 2013.

KITZMILLER, A. **Trinotate workflow example on Odyssey**. Disponivel em: https://informatics.fas.harvard.edu/trinotate-workflow-example-on-odyssey.html. Acesso em 20 jun, 2021.

KRASILEVA, Ksenia V. The role of transposable elements and DNA damage repair mechanisms in gene duplications and gene fusions in plant genomes. **Current opinion in plant biology**, v. 48, p. 18-25, 2019.

KURTZ S.; SCHLEIERMACHER C. REPuter: fast computation of maximal repeats in complete genomes. **Bioinformatics** 15:426-427, 1999.

KUZMIN, D. A. et al. Stepwise large genome assembly approach: A case of Siberian larch (*Larix sibirica* Ledeb). **BMC Bioinformatics**, v. 20, n. Suppl 1, 2019.

KYRIAKIDOU, M., TAI, H. H., ANGLIN.; et al., Current strategies of polyploid plant genome sequence assembly. **Frontiers in plant science**, v. 9, p. 1660, 2018.

LANGMEAD B, SALZBERG SL. Fast gapped-read alignment with Bowtie2. **Nature Methods** 9:357–359, 2012

LI, F. W.; HARKESS, A. A guide to sequence your favorite plant genomes. **Applications in Plant Sciences**, v. 6, n. 3, p. 1–7, 2018.

LI, P., GUO, W., LEI, K., & JI, L. Characterization of the complete chloroplast genome of *Syzygium nervosum*. **Mitochondrial DNA**, v. 6, n. 3, p. 1014-1015, 2021.

LIMA, JACQUELINE S.; TELLES, MARIANA P. C.; CHAVES, LÁZARO J.; et al., Demographic stability and high historical connectivity explain the diversity of a savanna tree species in the Quaternary. **Annals of Botany** (Print), v. 1, p. mcw257, 2017.

LISCH, D. How important are transposons for plant evolution? Nature Reviews Genetics, v. 14, n. 1, p. 49–61, 2013.

LÓPEZ-FLORES, I.; GARRIDO-RAMOS, M. A. The repetitive DNA content of eukaryotic genomes. Repetitive DNA, v. 7, p. 1–28, 2012.

MACHADO L. O, VIEIRA L. D, STEFENON V. M, OLIVEIRA P. F, SOUZA E. M, GUERRA M. P, NODARI R. O. Phylogenomic relationship of feijoa (*Acca sellowiana* (O.Berg) Burret) with other *Myrtaceae* based on complete chloroplast genome sequences. **Genetica**. V. 145, n. 2, p. 163-174, 2017.

MACHADO, L. D. O., VIEIRA, N. L.; STEFENON, V. M; et al. Molecular relationships of *Campomanesia xanthocarpa* within *Myrtaceae* based on the complete plastome sequence and on the plastid ycf2 gene. **Genetics and molecular biology**, v. 43, n. 2, 2020.

MCCORMICK, R. F.; et al. The Sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. **The Plant Journal**, v. 93, n. 2, p. 338-354, 2018.

MADRITSCH, S.; BURG, A.; SEHR, E. M. Comparing de novo transcriptome assembly tools in di- and autotetraploid non-model plant species. **BMC Bioinformatics**, v. 22, n. 1, p. 1–17, 2021.

MARA, Y. et al. Phenology, production and precocity of *Eugenia dysenterica*. **Plants aiming breeding**. n. 1985, 2013.

MARTINEZ, M. Plant protein-coding gene families: Emerging bioinformatics approaches. **Trends in Plant Science**, v. 16, n. 10, p. 558–567, 2011.

MAZUTI SILVA, S. M. et al. *Eugenia dysenterica Mart*. Ex Dc. (Cagaita): Planta brasileira com potencial terapêutico. Infarma - **Ciências Farmacêuticas**, v. 27, n. 1, p. 49, 2015.

MCKAIN, M.; WILSON, M. **Fast-Plast: Rapid de novo assembly and finishing for whole chloroplast genomes.** Disponível em: https://github.com/mrmckain/Fast-Plast. Acesso em: 25 mai. 2021.

MCCORMICK, R. F. et al. The *Sorghum bicolor* reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. **Plant Journal**, v. 93, n. 2, p. 338–354, 2018.

MEHROTRA, S.; GOYAL, V. Repetitive Sequences in Plant Nuclear DNA: Types, Distribution, Evolution and Function. **Genomics, proteomics & bioinformatics**, v. 12, n. 4, p. 164–171, 2014.

MICHAEL, Todd P. Plant genome size variation: bloating and purging DNA. **Briefings in functional genomics**, v. 13, n. 4, p. 308-317, 2014.

MOSA, K. A.; ISMAIL, A.; HELMY, M. Omics and system biology approaches in plant stress research. In: **Plant stress tolerance**. Springer, Cham, 2017. p. 21-34.

MYBURG, A. A. et al. The genome of *Eucalyptus grandis*. **Nature**, v. 509, n. 7505, p. 356–62, 2014.

NASS, L.L; SIGRIST, M. S.; RIBEIRO, C. S. C.; ReifschneideR, F. G. B. Genetic resources: the basis for sustainable and competitive plant breeding. **Crop Breeding and Applied Biotechnology** S2: p. 75-86, 2012.

Novák, P; Neumann, P.; J.; Pech, J.; Steinhaisl, J. Macas, J. RepeatExplorer: A Galaxy-based web server for genome- wide characterization of eukaryotic repetitive elements from next- generation sequence reads. **Bioinformatics**, v. 29, p. 792–793, 2013

NUNES R**. Identificação e caracterização de SNPs no genoma de *Eugenia dysenterica* DC. (Myrtaceae)**., 2015. 84 f. Dissertação (Mestrado em Genética e Melhoramento de Plantas) – Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2015.

NUNES R.; SOUZA, U. J. B. D.; TARGUETA C. P.; et al. Complete chloroplast genome sequence of *Caryocar brasiliense* Camb. (*Caryocaraceae*) and comparative analysis brings new insights into the plastome evolution of Malpighiales. **Genetics and Molecular Biology**.v. 43, n. 2. 2020.

OLSSON, S. et al. Development of genomic tools in a widespread tropical tree, *Symphonia globulifera* L.f.: a new low-coverage draft genome, SNP and SSR markers. **Molecular Ecology Resources**, v. 17, n. 4, p. 614–630, 2017.

PISUPATI, R.; VERGARA, D.; KANE, N. C. Diversity and evolution of the repetitive genomic content in Cannabis sativa. **BMC genomics**, v. 19, n. 1, p. 1-9, 2018.

PROENÇA, C. E. B.; GIBBS, P. E. Reproductive biology of eight sympatric Myrtaceae from Central Brazil. **New Phytologist**, v. 126, n. 2, p. 343–354, 1994.

QIAO, X. et al. Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. **Genome Biology**, v. 20, n. 1, p. 1–23, 2019.

RABELO S. G.; TEIXEIRA C. F; TELLES, M.P.C.; COLLEVATTI R.G. Development and characterization of microsatellite markers for Lychnophora ericoides, an endangered Cerrado shrub species. **Conservation Genetic Resourses** v. 3. p. 741-74, 2011.

RODRIGUES, EDUARDO B.; Collevatti, Rosane G.; CHAVES, LÁZARO J.; MOREIRA, LUCAS R.; TELLES, MARIANA P. C. Mating system and pollen dispersal in *Eugenia dysenterica* (Myrtaceae) germplasm collection: tools for conservation and domestication. **Genetica ('s-Gravenhage)**, v. 1, p. 141-146, 2016.

RODRIGUES NF, BALBINOTT N, PAIM I, GUZMAN F, MARGIS R. Comparative analysis of the complete chloroplast genomes from six Neotropical species of *Myrteae* (*Myrtaceae*). **Genetics and Molecular Biology** 43(2), 2020.

ROZAS J, FERRER-MATA A, SÁNCHEZ-DELBARRIO JC, et al. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Datasets. **Mol Biol Evol** 34:3299-3302, 2017.

SABLOK, G. et al. Chloroplast genomics: Expanding resources for an evolutionary conserved miniature molecule with enigmatic applications. **Current Plant Biology**, v. 7–8, p. 34–38, 2016.

SAHEBI, M., HANAFI, M. M., VAN WIJNEN, A. J.; et al. Contribution of transposable elements in the plant's genome. **Gene**, v. 665, p. 155-166, 2018.

SCHUSTER, T. M. et al. Chloroplast variation is incongruent with classification of the Australian bloodwood eucalypts (genus *Corymbia*, family Myrtaceae). **PloS one**, v. 13, n. 4, p. e0195034, 2018. Seppey et al., 2019.

SHEN, X. et al. Complete Chloroplast Genome Sequence and Phylogenetic Analysis of the Medicinal Plant *Artemisia annua*. **Molecules**, v. 22, n. 8, 2017.a

SHEN, C. et al. Identification and analysis of genes associated with the synthesis of bioactive constituents in *Dendrobium officinale* using RNA-Seq. **Scientific Reports**, v. 7, n. 1, p. 1-11, 2017.

SILVA-JUNIOR, O. B. et al. Genome assembly of the Pink Ipê (*Handroanthus impetiginosus*, Bignoniaceae), a highly valued, ecologically keystone Neotropical timber forest tree. **GigaScience**, v. 7, n. 1, p. 1–16, 2018.

SOARES T. N; MELO, D.B; RESENDE L.V. VIANELLO R.P., CHAVES L.J; COLLEVATTI R. G.; TELLES M.P. C . Development of microsatellite markers for the neotropical tree species Dipteryx alata (Fabaceae). **Am J Bot** 99:e72-e7, 2012.

SOARES, S.D. Montagem e análise do transcritoma do mogno africano (*Khaya grandifoliola* C. DC.).)., 2019. Tese (Doutorado em Genética e Biologia Molecular) – Instituto de Ciências Biológicas, Universidade Federal de Goiás.

SOBREIRO, M. B., VIEIRA, L. D., NUNES, R.; et al. Chloroplast genome assembly of *Handroanthus impetiginosus*: comparative analysis and molecular evolution in Bignoniaceae. **Planta**, v. 252, n. 5, p. 1-16, 2020.

SOEWARTO, J. et al. Transcriptome data from three endemic Myrtaceae species from New Caledonia displaying contrasting responses to myrtle rust (*Austropuccinia psidii*). **Data in brief**, v. 22, p. 794-811, 2019.

SOUZA, A. M. et al. Traditional uses, Phytochemistry, and antimicrobial activities of Eugenia species–a review. **Planta medica**, v. 84, n. 17, p. 1232-1248, 2018.

SOUZA U.J.B.; NUNES R.; TARGUETA C.P.; DINIZ-FILHO J. A. F, TELLES M.PD. C. The complete chloroplast genome of *Stryphnodendron adstringens* (*Legu.minosae - Caesalpinioideae*): Comparative analysis with related mimosoid species. **Scientific Reports** 9:1-12, 2019.

STANKE M, MORGENSTERN B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. **Nucleic Acids Res**,, v 1, n. 33. doi:

10.1093/nar/gki458. 2005.

STUESSY, T.; WEISS-SCHNEEWEISS, H. What drives polyploidization in plants? **New Phytologist**, v. 223, n. 4, p. 1690–1692, 2019.

STRICKLER, S. R.; BOMBARELY, A.; MUELLER, L. A. Designing a transcriptome next-generation sequencing project for a nonmodel plant species1. **American Journal of Botany**, v. 99, n. 2, p. 257-266, 2012.

TELLES, M. P. C; SILVA, R. S. M.; CHAVES, L. J.; COELHO, A. S. G.; FILHO, J. A. F. D. Divergência entre subpopulações de cagaiteira (Eugenia dysenterica) em resposta a padrões edáficos e distribuição espacial. Pesquisa Agropecuária Brasileira, Brasília, v. 36, n. 11, nov. 2001

TELLES, M. P. C; COELHO, A. S. G.; CHAVES, L. J.; FILHO, J. A. F. D.; VALVA, F. D. Genetic diversity and population structure of Eugenia dysenterica DC. ("cagaiteira" – Myrtaceae) in Central Brazil: Spatial analysis and implications for conservation and management. Conservation Genetics, v. 4, n. 6, p. 685-695, nov. 2003.

TELLES, M. P. C. SILVA, J. B.; RESENDE, L. V.; VIANELLO, R. P.; CHAVES, L. J.; SOARES, T. N.; COLLEVATTI, R. G. Development and characterization of new microsatellites for Eugenia dysenterica DC (Myrtaceae). Genetics and Molecular Research, v. 12, n. 3, p. 3124–3127, FEV. 2013.

THERMO FISHER SCIENTIFIC. Qubit Fluorometric Quantification. Disponível em: https://www.thermofisher.com/br/en/home/industrial/spectroscopy-elemental-isotope-analysis/molecular-spectroscopy/fluorometers/qubit.html. Acesso em 25 abr. 2021a.

THERMO FISCHER SCIENTIFIC. NanoDrop™ Lite Spectrophotometer. Disponível em: thermofisher.com/order/catalog/product/ND-LITE-PR#/ND-LITE-PR. Acesso em 25 abr. 2021b.

THERMO FISHER SCIENTIFIC. Purelink Plant RNA Reagent. Disponivel em: https://www.thermofisher.com/order/catalog/product/12322012#/12322012. Acesso em 09 mar. 2021c.

THRIMAWITHANA, A. H. et al. A whole genome assembly of *Leptospermum scoparium* (Myrtaceae) for mānuka research**. New Zealand Journal of Crop and Horticultural Science**, v. 47, n. 4, p. 233-260, 2019.

THORNHILL, A. H. et al. Interpreting the modern distribution of Myrtaceae using a dated molecular phylogeny. **Molecular Phylogenetics and Evolution**, v. 93, p. 29-43, 2015.

TOBIAS, P. A., GUEST, D. I., KÜLHEIM, C., PARK, R. F. De novo transcriptome study identifies candidate genes involved in resistance to *Austropuccinia psidii* (myrtle rust) in *Syzygium luehmannii* (Riberry). **Phytopathology**, v. 108, n. 5, p. 627-640, 2018.

UNAMBA, C. I. N; NAG, A; SHARMA, R. K. Next generation sequencing technologies: the doorway to the unexplored genomics of non-model plants. **Frontiers in plant science**, v. 6, p. 1074, 2015.

VETÖ, N. M. et al. Transcriptomics analysis of *Psidium cattleyanum* sabine (Myrtaceae) unveil potential genes involved in fruit pigmentation. **Genetics and Molecular Biology**, v. 43, n. 2, p. 1–11, 2020.

VIEIRA, R. F.; CAMILLO, J.; CORADIN, L. **Espécies nativas da flora brasileira de valor econômico atual ou potencial: plantas para o futuro: Região Centro-Oeste**. Disponível em: https://www.embrapa.br/busca-de-publicacoes/-/publicacao/1073295/especies-nativas-da-flora-brasileira-de-valor-economico-atual-ou-potencial-plantas-para-o-futuro-regiao-centro-oeste . Acesso em 14 de jun. 2021.

WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-Seq: a revolutionary tool for transcriptomics in Western Equatoria State. **Nature Reviews Genetics**, v. 10, n. 1, p. 57, 2009.

WANG, J.; LIU, J.; KANG, M. Quantitative testing of the methodology for genome size estimation in plants using flow cytometry: a case study of the Primulina genus. **Frontiers in plant science**, v. 6, p. 354, 2015.

WCSP, Work Cheklist of Selectec Plant Families. Disponível em: https://wcsp.science.kew.org/qsearch.do .Acesso em: 6 jun. 2021.

WENDEL, J. F., JACKSON, S. A., MEYERS, B. C., WING, R. A. Evolution of plant genome architecture. **Genome biology**, v. 17, n. 1, p. 1-14, 2016.

WENDEL, J. F., LISCH, D., HU, G., MASON, A. S. The long and short of doubling down: polyploidy, epigenetics, and the temporal dynamics of genome fractionation. **Current opinion in genetics & development**, v. 49, p. 1-7, 2018.

WICKE, S. et al. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. **Plant molecular biology**, v. 76, n. 3, p. 273-297, 2011.

WYMAN, S. K.; JANSEN, R. K.; BOORE, J. L. Automatic annotation of organellar genomes with DOGMA. **Bioinformatics**, v. 20, n. 17, p. 3252-3255, 2004.

XIAO-MING, Z. et al. Inferring the evolutionary mechanism of the chloroplast genome size by comparing whole-chloroplast genome sequences in seed plants. Scientific Reports, v. 7, n. 1, p. 1–10, 2017.

XING, Y. et al. Hybrid de novo genome assembly of Chinese chestnut (*Castanea mollissima*). **GigaScience**, v. 8, n. 9, p. giz112, 2019.

XU, J. H. et al. Dynamics of chloroplast genomes in green plants. **Genomics**, v. 106, n. 4, p. 221-231, 2015.

YANG, H. In plants, expression breadth and expression level distinctly and non-linearly correlate with gene structure. **Biology direct**, v. 4, n. 1, p. 1-15, 2009.

ZHANG, X. F. et al. Comparative analysis of chloroplast genome structure and molecular dating in Myrtales. **BMC plant biology**, v. 21, n. 1, p. 1-19, 2021.

ZIMIN, A. V. et al. Hybrid assembly of the large and highly repetitive genome of *Aegilops*

tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. **Genome research**, v. 27, n. 5, p. 787-792, 2017.

TRINDADE, M. G; CHAVES, L. J. Genetic structure of natural Eugenia dysenterica DC (Myrtaceae) populations in northeastern Goias, Brazil, accessed by morphological traits and RAPD markers. **Genetics and Molecular Biology**, v. 28, n. 3, p. 407–413, fev. 2005.

ZUCCHI, M.I. **Análise da estrutura genética de Eugenia dysenterica DC utilizando marcadores RAPD e SSR**. Tese (Doutorado) – Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba, nov. 2002.

ZUCCHI, M. I.; PINHEIRO, J. B.; AGUIAR, A. V.; CHAVES, L. J.; COELHO, A. S. G.; VENCOVSKY, R. Padrão espacial de divergência em populações de *Eugenia dysenterica* DC. utilizando marcadores microssatélites. **Floresta e Ambiente**, v. 11, n. 1, p. 29–38, ago. 2004.

Apêndice 1. Números de acesso aos genomas de cloroplasto de espécies da família *Myrtaceae* amostradas para comparação com o genoma cloroplastidial de *Eugenia dysenterica*.

Tribo	Espécie	Nº de acesso NCBI
Eucalypteae	*Allosyncarpia ternata*	NC_022413.1
	Eucalyptus baxteri	NC_022382.1
	Eucalyptus saligna	NC_022397.1
Myrteae	*Eugenia brasiliensis*	MT900596.1
	Eugenia uniflora	NC_027744.1
	Eugenia pyriformis	MN095410.1
	Eugenia selloi	MN095411.1
	Plinia trunciflora	NC_034801.1
	Rhodomyrtus tomentosa	NC_043848.1
Syzygiae	*Syzygium forrestii*	MK102721.1

Apêndice 2. Genes com íntrons, identificados no genoma do cloroplasto de *Eugenia dysenterica*, incluindo seus comprimentos (pb), bem como os comprimentos dos exons.

Gene	Localização	Exon I (pb)	Intron I (pb)	Exon II (pb)	Intron II (pb)	Exon III (PB)
rps16	LSC	39	872	207	-	-
atpF	LSC	147	740	411	-	-
rpoC1	LSC	451	729	1,619	-	-
petB	LSC	6	778	642	-	-
petD	LSC	9	752	474	-	-
rpl16	LSC	399	1.002	9	-	-
rpl2	IR	390	664	435	-	-
ndhB	IR	777	681	756	-	-
ndhA	SSC	552	1.056	537	-	-
ycf3	LSC	126	763	228	727	153
clpP	LSC	69	858	291	616	228
rps12	IR	114	-	210	567	27
trnK-UUU	LSC	37	2.526	35	-	-
trnG-UCC	LSC	23	755	49	-	-
trnL-UAA	LSC	35	508	48	-	-
trnV-UAC	LSC	37	596	39	-	-
trnA-UGC	IR	38	803	35	-	-
trnI-GAU	IR	37	957	35	-	-

Apêndice 3. Análise de uso de códons em 78 genes codificadores de proteínas diferentes, incluindo o pseudogene *ycf1*, identificados no genoma de cloroplasto de *Eugenia dysenterica.* RSCU - *Relative synonymous codon usage*.

Aminoácido	Codon	Frequência	RSCU
Ala	GCU	573	1,84
	GCC	194	0,62
	GCA	341	1,09
	GCG	138	0,44
Arg	CGU	309	1,35
	CGC	84	0,37
	CGA	330	1,44
	CGG	82	0,36
	AGA	412	1,8
	AGG	160	0,7
Asn	AAU	848	1,56
	AAC	239	0,44
Asp	GAU	731	1,63
	GAC	166	0,37
Cys	UGU	204	1,48
	UGC	72	0,52
Gln	CAA	631	1,56
	CAG	176	0,44
Glu	GAA	915	1,52
	GAG	292	0,48
Gly	GGU	535	1,36
	GGC	155	0,39
	GGA	637	1,62
	GGG	243	0,62
His	CAU	420	1,56
	CAC	119	0,44
Ile	AUU	990	1,5
	AUC	377	0,57
	AUA	609	0,92
Leu	UUA	823	2,04
	UUG	464	1,15
	CUU	508	1,26
	CUC	166	0,41
	CUA	323	0,8
	CUG	137	0,34
Lys	AAA	887	1,51
	AAG	285	0,49
Met	AUG	544	1

Aminoácido	Codon	Frequência	RSCU
Phe	UUU	893	1,35
	UUU	893	1,35
	UUC	427	0,65
Pro	CCU	383	1,62
	CCC	184	0,78
	CCA	273	1,15
	CCG	106	0,45
Ser	UCU	502	1,76
	UCC	280	0,98
	UCA	325	1,14
	UCG	158	0,55
	AGU	352	1,23
	AGC	99	0,35
Thr	ACU	483	1,6
	ACC	229	0,76
	ACA	358	1,19
	ACG	134	0,45
Trp	UGG	406	1
Tyr	UAU	698	1,6
	UAC	175	0,4
Val	GUU	455	1,44
	GUC	147	0,47
	GUA	486	1,54
	GUG	174	0,55
End	UGA	19	0,65
	UAA	50	1,7
	UAG	19	0,65

Apêndice 4. Regiões SSR identificadas no genoma cloroplastidial de *Eugenia dysenterica*.

No. SSR	SSR	Tamanho	Início	Fim	Localização
1	$(CAG)_4$	12	1159	1170	LSC (psbA)
2	$(T)_{11}$	11	2451	2461	LSC (matK)
3	$(A)_{11}$	11	3744	3754	LSC (intron – trnK-UUU)
4	$(A)_{11}$	11	4499	4509	LSC
5	$(A)_{10}$	10	4735	4744	LSC
6	$(AGAT)_3$	12	4848	4859	LSC
7	$(A)_{10}$	10	6513	6522	LSC
8	$(T)_{12}$	12	7877	7888	LSC
9	$(A)_{13}$	13	7953	7965	LSC
10	$(T)_{10}$	10	8359	8368	LSC
11	$(T)_{12}$	12	8490	8501	LSC
12	$(A)_{11}$	11	8729	8739	LSC
13	$(AAAT)_3$	12	8758	8769	LSC
14	$(T)_{10}$	10	9103	9112	LSC
15	$(A)_{10}$	10	10726	10735	LSC
16	$(ATTA)_3$	12	11045	11056	LSC
17	$(ATTT)_3$	12	11093	11104	LSC
18	$(T)_{10}$	10	13218	13227	LSC (intron - atpF)
19	$(A)_{10}$	10	14754	14763	LSC
20	$(T)_{13}$	13	15508	15520	LSC
21	$(T)_{14}$	14	15549	15562	LSC
22	$(T)_{11}$	11	17532	17542	LSC
23	$(A)_{10}$	10	17543	17552	LSC
24	$(T)_{11}$	11	19763	19773	LSC (rpoC2)
25	$(T)_{10}$	10	22359	22368	LSC (rpoC1)
26	$(T)_{10}$	10	27424	27433	LSC (rpoB)
27	$(CTTG)_3$	12	29214	29225	LSC

No. SSR	SSR	Tamanho	Início	Fim	Localização
28	$(A)_{10}$	10	31564	31573	LSC
29	$(A)_{11}$	11	31968	31978	LSC
30	$(ATTA)_3$	12	33356	33367	LSC
31	$(T)_{10}$	10	34178	34187	LSC
32	$(TAAG)_3$	12	45852	45863	LSC (intron –ycf3)
33	$(T)_{14}$	14	46971	46984	LSC (intron – ycf3)
34	$(A)_{11}$	11	47073	47083	LSC (intron – ycf3)
35	$(T)_{14}$	14	51109	51122	LSC
36	$(T)_{11}$	11	53355	53365	LSC (ndhK)
37	$(T)_{10}$	10	54098	54107	LSC
38	$(T)_{10}$	10	54358	54367	LSC (intron – trnV-UAC)
39	$(T)_{10}$	10	57274	57283	LSC (atpB)
40	$(T)_{10}$	10	57749	57758	LSC
41	$(A)_{10}$	10	59960	59969	LSC
42	$(T)_{11}$	11	63112	63122	LSC (ycf4)
43	$(TCTT)_3$	12	63550	63561	LSC
44	$(TC)_5$	10	63933	63942	LSC (cemA)
45	$(A)_{10}$	10	67990	67999	LSC
46	$(T)_{10}$	10	69233	69242	LSC
47	$(T)_{11}$	11	70291	70301	LSC
48	$(ATAA)_3$	12	71311	71322	LSC (rps18)
49	$(T)_{11}$	11	72585	72595	LSC
50	$(T)_{12}$	12	73549	73560	LSC (intron – clpP)
51	$(A)_{10}$	10	73729	73738	LSC (intron – clpP)
52	$(T)_{10}$	10	73812	73821	LSC (intron – clpP)
53	$(T)_{10}$	10	74251	74260	LSC (intron – clpP)
54	$(A)_{10}$	10	74680	74689	LSC (intron – clpP)
55	$(T)_{11}$	11	75409	75419	LSC

No. SSR	SSR	Tamanho	Início	Fim	Localização
56	$(TTTC)_3$	12	77860	77871	LSC
57	$(A)_{13}$	13	83981	83993	LSC
58	$(TTTC)_3$	12	85222	85233	LSC (intron – rpl16)
59	$(A)_{10}$	10	85535	85544	LSC
60	$(A)_{10}$	10	86338	86347	LSC (rpl22)
61	$(TTC)_4$	12	106394	106405	IRB (intron-trnI-GAU)
62	$(TTAT)_3$	12	113523	113534	SSC
63	$(TATT)_3$	12	113535	113546	SSC
64	$(T)_{10}$	10	113615	113624	SSC
65	$(T)_{10}$	10	113626	113635	SSC
66	$(ATAG)_3$	12	115733	115744	SSC (ndhF)
67	$(T)_{10}$	10	115920	115929	SSC
68	$(A)_{11}$	11	116779	116789	SSC
69	$(A)_{11}$	11	117065	117075	SSC
70	$(A)_{10}$	10	117077	117086	SSC
71	$(A)_{11}$	11	117684	117694	SSC
72	$(A)_{10}$	10	118944	118953	SSC
73	$(AATA)_3$	12	119434	119445	SSC (ndhD)
74	$(A)_{10}$	10	122780	122789	SSC
75	$(T)_{10}$	10	122872	122881	SSC
76	$(TAAT)_3$	12	129289	129300	SSC (ycf1)
77	$(A)_{10}$	10	131696	131705	SSC (ycf1)
78	$(AAG)_4$	12	139251	139259	IRA (intron-trnI-GAU)

Apêndice 5. Repetições *forward*, *reverse* e palindrômicas identificadas no genoma cloroplastidial de *Eugenia dysenterica*.

No.	Tipo	Tamanho (bp)	Localização	Região	Unidade de repetição
1	F	47	IGS (psbE – petL)	LSC	GATTCAAATTTCTTATTATTATTTTAATAATTCAAATTTCTTATTAT
2	F	45	ycf2	IR	TCGATATTGAGGATAGTGACGATATTGAGGATAGTGACGATATTG
3	F	42	Intron (ycf3), intron (ndhA)	LSC/SSC	GTTCCAGAACCGTACGTGAGATTTTCACCTCATACGGCTCCT
4	F	40	IGS (rps12 - trnV-GAC)	IR	ACAGAACCGTACATGAGATTTTCACCTCATACGGCTCCTC
5	F	39	Intron (ycf3), IGS (rps12 - trnV-GAC)	LSC/IR	CCAGAACCGTACGTGAGATTTTCACCTCATACGGCTCCT
6	F	41	psaB, psaA	LSC	ATGCAATAGCCAAATGATGGTGAGCAATAAGTCAACCAT
7	F	31	psaB, psaA	LSC	CAAATGATGGTGAGCAATATCAGTCAACCAT
8	F	31	IGS (trnT-UGU - trnL-UAA)	LSC	ATAGTTATATATAATAAATATATATTTAGTT
9	F	31	ycf2	IR	AGACAAAAAAAGAAGAAACTTGGACAAAAAG
10	F	45	ycf2	IR	ATCAATATCGTCACTATCCTCAATATCGTCACTATCCTCAATATC
11	F	31	ycf2	IR	GTTTTGTCCAAGTTACTTCTCTTTTTGTCCA
12	F	30	Intron (ycf3), IGS (rps12 - trnV-GAC)	LSC/IR	GTGAGATTTTCACCTCATACGGCTCCTCCC
13	F	30	ycf2	IR	ATATCGATATTGAGGATAGTGACGATATTG
14	F	30	IGS (rps12 - trnV-GAC), ndhA (intron)	IR/SSC	ATGAGATTTTCACCTCATACGGCTCCTCGT
15	F	30	IGS (ycf1 – ndhF)	SSC	GTATTTATTAATTATTTATTTATTATTTAT
16	F	30	ycf2	IR	ATCAATATCGTCACTATCCTCAATATCGTC

No.	Tipo	Tamanho (bp)	Localização	Região	Unidade de repetição
17	F	32	trnS-GCU, trnS-UGA	LSC	GAAACGGAAAGAGAGGGATTCGAACCCTCGGT
18	R	34	ndhF	SSC	GAAATAAATGAAACTTAAAAAAAATTCAAATTAA
19	R	38	IGS (ycf1 – ndhF)	SSC	TTTATTAATTATTTATTTATTATTTATTTATTAATTAT
20	R	33	IGS (ycf1 – ndhF)	SSC	GTATTTATTAATTATTTATTTATTATTTATTTA
21	P	45	ycf2	IR	ACGATATTGAGGATAGTGACGATATTGAGGATAGTGACGATATTG
22	P	39	Intron (ycf3), IGS (trnV-GAC – rps12)	LSC/IR	CCAGAACCGTACGTGAGATTTTCACCTCATACGGCTCCT
23	P	45	ycf2	IR	TCGATATTGAGGATAGTGACGATATTGAGGATAGTGACGATATTG
24	P	30	psbD, IGS (ndhC – trnV-UAC)	LSC	GAAAATGATTTATTTGATATTATGGATGAC
25	P	30	trnS-GUC, trnS-GGA	LSC	AACGGAAAGAGAGGGATTCGAACCCTCGGT
26	P	40	Intron (ndhA)	SSC	CCAGAACCGTACGTGAGATTTTCACCTCATACGGCTCCTC
27	P	30	ycf2	IR	ATATCGATATTGAGGATAGTGACGATATTG
28	P	30	ycf2	IR	GTGACGATATTGAGGATAGTGACGATATTG
29	P	30	Intron (ycf3), IGS (rps12 - trnV-GAC)	LSC/IR	GTGAGATTTTCACCTCATACGGCTCCTCCC
30	P	30	Intron (ndhA), IGS (trnV-GAC – rps12)	IR	GTGAGATTTTCACCTCATACGGCTCCTCGA
31	P	31	ycf2	IR	GTTAGACAAAAAAAGAAGAAACTTGGACAAA
32	P	31	ycf2	IR	TTTTTGTCCAAGTTTCTTCTTTTTTTGTCTA
33	P	30	trnS-UGA, trnS-GGA	LSC	AAAGGAGAGAGAGGGATTCGAACCCTCGAT

P: Palindrômica, F: *Forward*, R: *Reverse;* IGS: Espaçadores intergênicos

Apêndice 6. Razão Ka/Ks em genes do genoma de cloroplasto de cinco espécies de *Eugenia* (*E. dysenterica*, *E. brasiliensis*, *E. pyriformis*, *E. selloi* e *E. uniflora*), considerando-se os genes individuais e as regiões do cloroplasto.

Grupos gênicos	Gene	Ka	Ks	Ka/Ks	Região
Large subunit of ribosomal proteins	*rpl14*	0,0029	0,0046	0,6324	LSC
	rpl16	0,0013	0,0162	0,0817	LSC
	rpl20	0,0078	0,0088	0,8904	LSC
	rpl22	0,0041	0,0000	0,0000	LSC
	rpl33	0,0080	0,0000	0,0000	LSC
	rpl36	0,0048	0,0000	0,0000	LSC
	rpl2	0,0007	0,0020	0,3299	IR
	rpl23	0,0019	0,0000	0,0000	IR
	rpl32	0,0069	0,0494	0,1403	SSC
Small subunit of ribosomal proteins	*rps11*	0,0000	0,0150	0,0000	LSC
	rps14	0,0035	0,0240	0,1448	LSC
	rps16	0,0067	0,0181	0,3673	LSC
	rps18	0,0035	0,0057	0,6038	LSC
	rps19	0,0000	0,0128	0,0000	LSC
	rps2	0,0007	0,0123	0,0602	LSC
	rps3	0,0033	0,0147	0,2255	LSC
	rps4	0,0061	0,0000	0,0000	LSC
	rps8	0,0053	0,0205	0,2593	LSC
	rps12	0,0000	0,0000	0,0000	IR
	rps7	0,0000	0,0000	0,0000	IR
	rps15	0,0019	0,0260	0,0747	SSC
DNA-dependent RNA polymerase	*rpoA*	0,0064	0,0257	0,2481	LSC
	rpoB	0,0020	0,0108	0,1889	LSC
	rpoC1	0,0015	0,0091	0,1670	LSC
	rpoC2	0,0038	0,0113	0,3366	LSC
Photosystem I	*psaA*	0,0011	0,0046	0,2298	LSC
	psaB	0,0005	0,0096	0,0491	LSC
	psaI	0,0000	0,0000	0,0000	LSC
	psaJ	0,0000	0,0000	0,0000	LSC
	psaC	0,0000	0,0234	0,0000	SSC
Photosystem II	*psbA*	0,0000	0,0080	0,0000	LSC
	psbB	0,0009	0,0089	0,0972	LSC
	psbC	0,0000	0,0081	0,0000	LSC
	psbD	0,0000	0,0081	0,0000	LSC
	psbE	0,0000	0,0000	0,0000	LSC
	psbF	0,0000	0,0000	0,0000	LSC
	psbH	0,0000	0,0000	0,0000	LSC
	psbI	0,0000	0,0154	0,0000	LSC
	psbJ	0,0117	0,0000	0,0000	LSC
	psbK	0,0085	0,0100	0,8507	LSC

Grupos gênicos	Gene	Ka	Ks	Ka/Ks	Região
	psbL	0,0000	0,0162	0,0000	LSC
	psbM	0,0000	0,0000	0,0000	LSC
	psbN	0,0000	0,0000	0,0000	LSC
	psbT	0,0000	0,0159	0,0000	LSC
	psbZ	0,0029	0,0218	0,1328	LSC
NADH dehydrogenase	*ndhC*	0,0015	0,0000	0,0000	LSC
	ndhJ	0,0000	0,0073	0,0000	LSC
	ndhK	0,0018	0,0173	0,1065	LSC
	ndhB	0,0004	0,0043	0,0812	IR
	ndhF	0,0059	0,0161	0,3646	SSC
	ndhA	0,0020	0,0098	0,1996	SSC
	ndhD	0,0021	0,0207	0,1018	SSC
	ndhE	0,0017	0,0292	0,0593	SSC
	ndhG	0,0061	0,0031	1,9329	SSC
	ndhH	0,0022	0,0136	0,1620	SSC
	ndhI	0,0000	0,0108	0,0000	SSC
Cytochrome b/f complex	*petA*	0,0033	0,0071	0,4647	LSC
	petB	0,0000	0,0064	0,0000	LSC
	petD	0,0000	0,0190	0,0000	LSC
	petG	0,0049	0,0000	0,0000	LSC
	petL	0,0000	0,0000	0,0000	LSC
	petN	0,0000	0,0000	0,0000	LSC
ATP synthase	*atpA*	0,0000	0,0146	0,0000	LSC
	atpB	0,0007	0,0076	0,0939	LSC
	atpE	0,0013	0,0128	0,1029	LSC
	atpF	0,0019	0,0033	0,5589	LSC
	atpH	0,0000	0,0314	0,0000	LSC
	atpI	0,0000	0,0057	0,0000	LSC
RubisCo large subunit	*rbcL*	0,0078	0,0062	1,2732	LSC
Maturase K	*matK*	0,0074	0,0159	0,4657	LSC
Envelope membrane protein	*cemA*	0,0026	0,0028	0,9324	LSC
Subunit of acetyl-CoAcarboxylase	*accD*	0,0043	0,0108	0,3998	LSC
C-type cytochrome synthesis gene	*ccsA*	0,0048	0,0169	0,2851	SSC
Protease	*clpP*	0,0009	0,0146	0,0615	LSC
Conserved hypothetical chloroplast open reading frames	*ycf3*	0,0000	0,0157	0,0000	LSC
	ycf4	0,0019	0,0096	0,1964	LSC
	ycf2	0,0011	0,0029	0,3571	IR
	ycf1	0,0124	0,0219	0,5684	SSC/IR