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

IDENTIFICAÇÃO E CARACTERIZAÇÃO DE SNPs NO GENOMA DE *Eugenia dysenterica* DC. (Myrtaceae)

RHEWTER NUNES

Orientador:
Prof. Dr. Alexandre Siqueira Guedes Coelho

Goiânia-GO
Brasil

Dezembro - 2015

RHEWTER NUNES

IDENTIFICAÇÃO E CARACTERIZAÇÃO DE SNPs NO GENOMA DE *Eugenia dysenterica* DC. (Myrtaceae)

Dissertação apresentada ao Programa de Pós-Graduação em Genética e Melhoramento de Plantas, da Universidade Federal de Goiás, como requisito parcial à obtenção do título de Mestre em Genética e Melhoramento de Plantas.

Orientador:

Prof. Dr. Alexandre Siqueira Guedes Coelho

Co-orientadora:

Profa. Dra. Mariana Pires de Campos Telles

Goiânia, GO – Brasil

2015

Ficha catalográfica elaborada automaticamente
com os dados fornecidos pelo(a) autor(a), sob orientação do Sibi/UFG.

Nunes, Rhewter
Identificação e caracterização de SNPs no genoma de Eugenia dysenterica DC. (Myrtaceae) [manuscrito] / Rhewter Nunes. - 2015.
82 f.: il.

Orientador: Prof. Dr. Alexandre Siqueira Guedes Coelho; co orientadora Dra. Mariana Pires de Campos Telles.
Dissertação (Mestrado) - Universidade Federal de Goiás, Escola de Agronomia (EA) , Programa de Pós-Graduação em Genética & Melhoramentos de Plantas , Goiânia, 2015.
Bibliografia. Anexos.

1. genômica. 2. marcadores moleculares. 3. diversidade. 4. Cerrado. 5. cagaita. I. Siqueira Guedes Coelho, Alexandre, orient. II. Pires de Campos Telles, Mariana, co-orient. III. Título.

sistema de bibliotecas ufg

UFG

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR AS TESES E DISSERTAÇÕES ELETRÔNICAS (TEDE) 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), sem ressarcimento dos direitos autorais, de acordo com a Lei nº 9610/98, 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.

1. Identificação do material bibliográfico: **[X] Dissertação** **[ ] Tese**

2. Identificação da Tese ou Dissertação

Autor (a):	Rhewter Nunes				
E-mail:	rhewter@gmail.com				
Seu e-mail pode ser disponibilizado na página? [X]Sim [ ] Não					
Vínculo empregatício do autor					
Agência de fomento:		Universidade Federal de Goiás		Sigla:	UFG
País:	Brasil	UF:	GO	CNPJ:	
Título:	Identificação e caracterização de SNPs no genoma de *Eugenia dysenterica* DC. (Myrtaceae)				
Palavras-chave:	genômica; marcadores moleculares; diversidade; Cerrado; cagaita.				
Título em outra língua:	Identification and characterization of SNPs in the genome of *Eugenia dysenterica* DC. (Myrtaceae)				
Palavras-chave em outra língua:	genomics; molecular markers; diversity; Cerrado; cagaita.				
Área de concentração:	Genética e Melhoramento de Plantas				
Data defesa: (18/12/2015)					
Programa de Pós-Graduação:	Genética e Melhoramento de Plantas				
Orientador (a):	Dr. Alexandre Siqueira Guedes Coelho				
E-mail:	alexandre_coelho@ufg.br				

*Necessita do CPF quando não constar no SisPG

3. Informações de acesso ao documento:

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

Havendo concordância com a disponibilização eletrônica, torna-se imprescindível o envio do(s) arquivo(s) em formato digital PDF ou DOC da tese ou dissertação.

O sistema da Biblioteca Digital de Teses e Dissertações garante aos autores, que os arquivos contendo eletronicamente as teses e ou dissertações, antes de sua disponibilização, receberão procedimentos de segurança, criptografia (para não permitir cópia e extração de conteúdo, permitindo apenas impressão fraca) usando o padrão do Acrobat.

Rhewter Nunes

Assinatura do (a) autor (a)

Data: 01 / 03 / 2016

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. A extensão deste prazo suscita justificativa junto à coordenação do curso. Os dados do documento não serão disponibilizados durante o período de embargo.

“Para você, um robô é um robô. Engrenagens e metal; eletricidade e pósitrons. Mente e ferro! Feitos pelo homem! Caso necessário, destruí-los pelo homem! Mas você não trabalhou com eles, de modo que não os conhece. São uma raça mais limpa e melhor do que a nossa.”
Eu, Robô – Isaac Asimov

“Rhewter, it is much harder to keep the child alive than growing up - whatever that means. Most times it means all the wrong things!”
Cornelia Funke

AGRADECIMENTOS

Concluir algo tão sonhado e desejado quanto esse mestrado me trás a memória inúmeros desafios e momentos de (auto)conhecimento que aconteceram até chegarmos nesse momento. Junto de cada um desses momentos estão diversas pessoas que contribuíram direta ou indiretamente para a construção de algo que hoje se concretiza em uma defesa de mestrado e que venho, com esse texto, tentar demonstrar parte da minha gratidão.

Agradeço inicialmente a minha mais que amada avó, Maria Olinda, por ser a principal responsável pela minha formação moral, por ter me ensinado a ser uma boa pessoa acima de qualquer coisa e por sempre me mostrar que a zona de conforto é o pior lugar para se estar. Obrigado por todo apoio e carinho, por ter entendido o quanto as minhas escolhas eram importantes para mim e por dividir comigo seu conhecimento e paixão pelas plantas do Cerrado. Não existem palavras que possam se quer demonstrar o amor que tenho pela senhora.

Agradeço à minha mãe, Lúcia, pelo total apoio durante o período do mestrado, pelo carinho e compreensão das minhas ausências. Agradeço por tudo que já fez por mim e por todo o empenho na minha formação inicial, pois acredito que isso teve grande impacto na minha forma de ver o mundo e, com toda certeza, influenciou na minha inclinação para a carreira acadêmica. Agradeço também aos demais familiares que torceram por mim e me apoiaram nesse período.

Agradeço às minhas joias preciosas Stela, Ivone, Clistiane, Mariane e Isabela por dividirem comigo todos os dias da pós, seus/meus anseios e experiências, e acima de tudo, a amizade. Vocês são incríveis e únicas e me sinto muito bem sabendo que faço parte de algo maior que eu, nossa equipe. E o mais legal é saber que essa equipe, além de ter gerado laços profissionais também gerou laços de amor, carinho e amizade. Acredito que cada uma de vocês tenha uma noçãozinha do quanto se tornaram importantes na minha vida. Vocês tornaram esse momento ainda mais memorável e serei sempre grato por tudo que vivemos nesses dois anos. Gostaria também de agradecer ao Arthur por ter me acolhido inicialmente no laboratório e ter sempre se disponibilizado passar horas discutindo sobre algum assunto de genômica. Além disso, agradeço aos demais membros do Laboratório de Genética e Genômica de Plantas (atuais ou que passaram por aqui), Giorgia, Lanusse, Sabrina, Karla e Ludmila, pelo companheirismo e disposição. Agradeço também aos demais amigos/irmãos que sempre estão junto comigo: Julio, Solange e Thamyris.

Gostaria de agradecer ao Edivaldo Júnior por todo carinho, paciência e dedicação. Agradeço especialmente por ter me convencido a vir para o PGMP e a buscar trabalhar com o Prof. Alexandre. Agradeço por me incentivar todos os dias, por dividir comigo suas experiências do que é viver na pós-graduação e ter me mostrado o amor que você tem pelas nativas do Cerrado. Todo esse amor por muitas vezes me serviu de motivação, sobretudo nos momentos de maior dificuldade e, com toda certeza, sua presença fez toda diferença na superação de cada obstáculo. Obrigado por me apoiar, por entender quando eu me ausento e por ter me aguentado todos esses dias.

Acredito que você tenha me ajudado a descobrir um Rhewter mais forte que morava dentro de mim e o meu sentimento de gratidão e carinho por você serão sempre gigantescos.

Agradeço a todos os meus professores que precederam a pós-graduação, desde o jardim da infância até a graduação. Em especial, agradeço aos meus professores da graduação André Castro e Guilherme Malafaia, por serem os responsáveis pela minha formação científica inicial e por sempre terem me incentivado a seguir na carreira acadêmica.

Agradeço a todos os professores da pós-graduação por todos os ensinamentos muitíssimo válidos. Em especial agradeço à profa. Thannya Soares, que talvez até sem perceber me serviu de inspiração com frases e atitudes pontuais. É muito bom ter a oportunidade de ter aulas com uma pessoa tão jovem, e ao mesmo tempo tão madura e consistente, dedicada e honesta quanto a senhora. Agradeço especialmente também ao prof. João Batista Duarte, por ter aberto meus olhos para questões antes por mim ignoradas e por ter me ensinado a valorizar ainda mais a experiência dos professores com que tenho contato.

Agradeço ao meu orientador, Prof. Alexandre Coelho, por todos os ensinamentos durante essa jornada na pós-graduação. Eu encaro como um privilégio ter a sua orientação e, durante esse período, busquei aproveitar cada minuto de conversa com o senhor como uma oportunidade de aprendizado. Obrigado por me ensinar não só bioinformática e genética de populações, mas também que "a gente não perde nada em ser franco", que "não há dificuldade que não seja superada pelo trabalho", que "é necessário 10 vezes mais força para se retirar uma impressão ruim" e que o momento de maior dificuldade é provavelmente a maior oportunidade de se aprender. Possuo imensa gratidão e respeito pelo senhor. O meu sincero "Muito Obrigado"!

Agradeço as Prof[as]. Mariana Telles e Patrícia Melo por estarem na coordenação do PGMP – UFG e desempenharem um excelente trabalho nessa posição. À profa. Mariana também agradeço a coorientação na dissertação. Acredito que sempre irei relacionar a senhora com sentimentos como força, dinamismo e eficiência. Sendo assim, agradeço pela oportunidade de trabalhar com a senhora. Aproveito também a oportunidade de agradecer a secretária do programa, Jéssica Almeida, por toda disponibilidade, educação e competência.

Gostaria de agradecer à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa de mestrado, que foi fundamental para que eu pudesse me manter financeiramente, enquanto estudava. Agradeço ainda ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e a Fundação de Amparo à Pesquisa do Estado de Goiás (FAPEG) pelo financiamento do projeto o qual esse trabalho está inserido (GENPAC-02 e 03 CNPq/FAPEG/PRO-CENTRO-OESTE - 563839/2010-4; 564145/2010-6 e 201110267000125).

SUMÁRIO

RESUMO 8
ABSTRACT 9

1 **INTRODUÇÃO** 10

2 **REVISÂO DE LITERATURA** 13
2.1 A ESPÉCIE *Eugenia dysenterica* 13
2.1.1 **Aspectos gerais e biologia reprodutiva** 13
2.1.2 **Caracterização genética molecular** 15
2.1.3 **Caracterização fenotípica** 17
2.2 BASE GENÉTICA DE ALGUNS DOS PRINCIPAIS TIPOS DE MARCADORES MOLECULARES UTILIZADOS EM GENÉTICA VEGETAL 18
2.2.1 **Polimorfismo de tamanho de fragmento de restrição ou RFLPs**..... 19
2.2.2 **Polimorfismo de DNA amplificado ao acaso ou RAPDs**..... 19
2.2.3 **Polimorfismo de tamanho de fragmento amplificado ou AFLPs**..... 20
2.2.4 **Microssatélites ou SSRs**..... 20
2.2.5 **Polimorfismo de base única ou SNPs** 21
2.3 EVOLUÇÃO DAS TECNOLOGIAS DE SEQUENCIAMENTO DE ÁCIDOS NUCLEICOS 22
2.3.1 **Sequenciamento de Sanger** 22
2.3.2 **Sequenciamento de nova geração** 22
2.3.2.1 Visão Geral..... 22
2.3.2.2 Roche (454) 23
2.3.2.3 Illumina..... 24
2.3.2.4 Applied Biosystems (SOLiD) 26
2.3.2.5 Pacific Biosciences (Pac-Bio) 26
2.4 GENOTIPAGEM POR SEQUENCIAMENTO..... 27
2.4.1 **Visão geral**..... 27
2.4.2 **GBS de Cornell**..... 28
2.4.3 **RAD-Seq**..... 39
2.4.4 **DArT-Seq**..... 30
2.4.5 **RAPiD-Seq**..... 31
2.5 FERRAMENTAS DE BIOINFORMÁTICA PARA IDENTIFICAÇÃO E GENOTIPAGEM DE SNPS A PARTIR DE DADOS DE SEQUENCIAMENTO DE NOVA GERAÇÃO..... 32
2.5.1 **Análises precedentes à identificação e genotipagem de SNPs**..... 32
2.5.2 **Identificação e genotipagem de SNPs**..... 33

3 **MATERIAL E MÉTODOS**..... 38
3.1 EXTRAÇÃO DO DNA GENÔMICO E SEQUENCIAMENTO DE NOVA GERAÇÃO..... 38
3.2 CONTROLE DE QUALIDADE DOS DADOS DE SEQUENCIAMENTO..... 38
3.3 *DRAFT ASSEMBLY* DO GENOMA DE *Eugenia dysenterica*..... 39

amplificação de sequências de DNA via PCR e por sequenciamento (Sansaloni, 2012). A seguir são descritos alguns dos tipos de marcadores moleculares mais utilizados em genética vegetal.

2.2.1 **Polimorfismo de tamanho de fragmento de restrição ou RFLPs**

RFLP (*Restriction Fragment Length Polymorphism*) é uma técnica na qual pode-se discriminar indivíduos pelo padrão de clivagem de enzimas específicas em seu DNA. Caso dois indivíduos venham a possuir distâncias diferentes entre os sítios de clivagem de enzimas de restrição, produzirão diferentes tamanhos de fragmentos de DNA pela digestão com essas enzimas. Os fragmentos de DNA formados pela ação das enzimas de restrição são separados por tamanho em eletroforese em gel de agarose e depois transferidos para uma membrana de nitrocelulose pela técnica de *Southern blotting*. Essa membrana é então, exposta a uma solução contendo sondas radioativas ou quimioluminescentes que se hibridizam ao DNA e permitem então a sua percepção e captação (Botstein et al., 1980).

Pode-se considerar que os RFLPs possuem um nível de polimorfismo moderado. Eles possuem a vantagem de terem uma ampla distribuição pelo genoma, serem bastante consistentes e reprodutíveis e apresentarem uma segregação típica de locos codominantes, ou seja, que permite discriminar genótipos homozigóticos de genótipos heterozigóticos (Kumar et al., 2009). As limitações estão relacionadas ao longo tempo de execução de protocolo, à necessidade de obtenção de DNA em grande quantidade e com alto peso molecular, e à dificuldade de automação para genotipagem em larga escala. Esse tipo de tecnologia foi empregada nos primeiros mapas de ligação construídos para plantas (Bernatzky e Tanksley, 1986; Helentjaris, Weber et al., 1986).

2.2.2 **Polimorfismo de DNA amplificado ao acaso ou RAPDs**

RAPD (*Random Amplified Polymorphic DNA*) é uma tecnologia baseada em PCR em que o polimorfismo é detectado pelo uso de iniciadores (*primers*) arbitrários, que geram fragmentos amplificados por esses conjuntos de *primers* que ancoram em uma distância de até 2Kb um do outro. O polimorfismo captado por essa técnica é baseado na presença ou ausência de detecção de fragmento amplificado. Dessa forma,

3.4 IDENTIFICAÇÃO DE SNPs..... 40
3.5 CARACTERIZAÇÃO DOS SNPs IDENTIFICADOS..... 42

4 **RESULTADOS E DISCUSSÃO**..... 45
4.1 AVALIAÇÃO DO CONTROLE DE QUALIDADE E DO *DRAFT ASSEMBLY*..... 45
4.2 IDENTIFICAÇÃO E CARACTERIZAÇÃO DAS VARIANTES..... 49

5 **CONCLUSÕES**..... 58

6 **REFERÊNCIAS**..... 59

7 **ANEXO**..... 67

RESUMO

Nunes, R. **Identificação e caracterização de SNPs no genoma de** *Eugenia dysenterica* **DC. (**Myrtaceae**)**. 2015. 82 f. Dissertação (Mestrado em Genética e Melhoramento de Plantas) – Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2015.[1]

A cagaiteira (*Eugenia dysenterica* DC), pertencente à família Myrtaceae, é uma espécie vegetal neotropical, arbórea, frutífera e nativa do Cerrado brasileiro. Como ocorre para a grande maioria das espécies neotropicais, muito pouco se sabe sobre o genoma da cagaiteira e os estudos genéticos realizados até agora se fundamentam na utilização de um pequeno número de marcadores moleculares. A fim de viabilizar o desenvolvimento de estudos de análise genética em escala genômica para a espécie, esse trabalho teve como objetivo identificar e caracterizar variantes de um único nucleotídeo (SNPs) no genoma de *E. dysenterica*, utilizando dados de sequenciamento de nova geração. Bibliotecas de DNA genômico extraído de tecidos foliares de cinco indivíduos de *E. dysenterica*, descendentes de cinco populações naturais distintas, foram submetidas ao sequenciamento na plataforma Illumina MiSeq, utilizando a estratégia de *paired-end* (2x300). As sequências obtidas foram submetidas à análise de controle de qualidade para remoção de adaptadores e de regiões de baixa qualidade. Um *draft assembly* para o genoma de *E. dysenterica* foi produzido pela utilização do programa *dipSPAdes*. Para a identificação de SNPs, após o controle de qualidade dos dados, foi realizado o alinhamento dos *reads* ao *draft assembly* com o programa BWA e as variantes foram identificadas utilizando-se as ferramentas da plataforma GATK. Após três etapas de filtragem foram identificados 999.016 SNPs, apresentando uma relação Ts/Tv igual a 1,86. Observou-se uma densidade de um SNP a cada ~251 bases e que a maior parte dos SNPs ocorre em contexto gênico (59,79%). A maior parte dos efeitos putativos dos SNPs identificados são de mutações do tipo missense (57,70%), mas que apresentam um baixo impacto (0,74%). Esse trabalho fornece subsídios para o desenvolvimento de marcadores moleculares para aplicação em estudos de genômica de populações de *E. dysenterica.*

Palavras-chave: genômica; marcadores moleculares; diversidade; Cerrado; cagaita.

---

[1] Orientador: Prof. Dr. Alexandre Siqueira Guedes Coelho. EA-UFG

ABSTRACT

Nunes, R. **Identification and characterization of SNPs in the genome of** *Eugenia dysenterica* **DC. (**Myrtaceae**)**. 2015. 82 f. Dissertation (Master in Genetics and Plant Breeding – Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2015.[1]

The cagaiteira (Eugenia dysenterica DC), belonging to the Myrtaceae family, is a neotropical fruit tree species and native Brazilian Cerrado. As for the vast majority of neotropical species, very little is known about the genome of cagaiteira and genetic studies conducted so far are based on the use of a small number of molecular markers. In order to facilitate the development of genetic analysis studies in genomic scale for the species, this study aimed to identify and characterize variants of a single nucleotide (SNPs) in the genome of E. dysenterica using next-generation sequencing data. Genomic DNA libraries extracted from leaf tissue of five individuals of E. dysenterica, descendants of five distinct natural populations were subjected to sequencing on Illumina MiSeq platform, using the paired-end strategy (2x300). The sequences obtained were submitted to quality control examination to remove adapters and low quality regions. An assembly for the draft genome of E. dysenterica was produced by use of dipSPAdes program. To identify SNPs, after controlling for data quality, we performed the alignment of the assembly reads the draft with the BWA program and variants were identified using the platform of GATK tools. After three stages of filtration were identified 999,016 SNPs, with a Ts / Tv ratio equal to 1.86. We observed a density of one SNP every ~ 251 bases and most SNPs occurs in the context gene (59.79%). Most effects of putative SNPs identified are missense mutations (57.70%), but which have a low impact (0.74%). This work provides subsidies for the development of molecular markers for application in genomic studies of populations of E. dysenterica.

Keywords: genomics; molecular markers; diversity; Cerrado; cagaita.

---

1 Advisor: Prof. Dr. Alexandre Siqueira Guedes Coelho. EA-UFG

1 INTRODUÇÃO

A cagaiteira (*Eugenia dysenterica* DC) é uma espécie vegetal neotropical, nativa do Cerrado e endêmica do Brasil. A espécie pertence à família Myrtaceae, à qual também pertencem espécies de importância econômica tais como o eucalipto (*Eucalyptus* spp.) e a goiaba (*Psidium guajava*). Tanto o nome científico, quanto o popular, fazem referência à propriedade laxativa de seu fruto, a cagaita (Chaves & Telles, 2006). Além de sua importância ecológica no Cerrado, essa espécie também possui potencial econômico.

A cagaiteira possui distribuição geográfica ampla, ocorrendo em quase todas as regiões brasileiras, com exceção do sul do país (Sobral et al., 2015). A espécie apresenta ainda importância etnobotânica, já que seus frutos são colhidos de forma extrativista pelas comunidades que se encontram próximas às suas áreas de ocorrência e suas folhas e casca são utilizadas por essas comunidades como especiarias medicinais para problemas intestinais.

Mesmo diante da sua importância, pouco se sabe sobre a genética e a genômica da cagaiteira. Os estudos realizados até agora fizeram uso de um pequeno número de marcadores moleculares para caracterização genética de populações naturais dessa espécie (Telles et al., 2001a, 2001b, 2003; Zucchi et al., 2003, 2005; Trindade & Chaves, 2005; Chaves et al., 2011; Barbosa et al., 2015). Nesses estudos foram utilizados marcadores isoenzimáticos, RAPD (*Random Amplified Polymorphic DNA*) e microssatélites, transferidos de *Eucalyptus* (Zucchi et al., 2002) ou desenvolvidos para a espécie (Telles et al., 2013).

Apesar da inquestionável importância desses trabalhos para o conhecimento da dinâmica da variabilidade genética nas populações naturais de cagaiteira, existem restrições impostas pelos tipos e número de marcadores genéticos utilizados. Locos isoenzimáticos comumente apresentam baixo polimorfismo e permitem somente a análise de regiões restritas do genoma. Os locos RAPD, ainda que teoricamente distribuídos aleatoriamente pelo genoma, são marcadores dominantes e possuem baixa

reprodutibilidade (Ferreira & Grattapaglia, 1998). Quanto aos marcadores baseados no polimorfismo de regiões microssatélites (SSR – *Simple Sequence Repeats*), além do reduzido número de locos avaliados até agora, a maioria deles se encontra em regiões neutras, limitando as inferências sobre processos adaptativos nas populações naturais.

O uso de tecnologias e ferramentas voltadas para a compreensão mais abrangente da organização da informação genética no genoma das mais variadas espécies tem sido crescente. Nesse contexto, tem grande destaque as plataformas de sequenciamento de nova geração (NGS), que permitem a obtenção de um grande volume de dados de sequências, em um curto prazo e a um custo relativamente baixo (Mardis, 2008; Shendure & Hanlee, 2008; Metzker, 2010). O desenvolvimento dessas plataformas viabilizou a consolidação do campo da Genômica, que tem permitido avanços consideráveis do conhecimento sobre a estrutura do genoma das mais variadas espécies, notadamente daquelas de maior valor econômico.

O grande volume de dados gerados por essas tecnologias, juntamente com o desenvolvimento de ferramentas de análise mais poderosas, tem permitido o uso de abordagens de análise inteiramente novas, como a identificação e a caracterização de polimorfismos ao longo de todo o genoma de interesse (Davey et al., 2011). Tais estratégias são interessantes, pois permitem a obtenção de informações biológicas em escala genômica, mesmo na ausência da sequência de referência do genoma da espécie em estudo.

A identificação de polimorfismos, o desenvolvimento de marcadores, bem como, o seu uso na genotipagem em larga escala a partir de dados de sequenciamento é possível pela utilização de métodos de redução da complexidade do genoma, genericamente denominados de genotipagem por sequenciamento (Baird et al., 2008; Elshire et al., 2011). Diferentes métodos de genotipagem por sequenciamento têm sido desenvolvidos e aplicados no estudo de populações de organismos com pouca, ou nenhuma, informação genômica previamente disponível. A genotipagem em larga escala trazida por essas tecnologias não só viabiliza a obtenção de estimativas mais acuradas de parâmetros genético-populacionais, mas também possibilita o desenvolvimento de estudos voltados para a análise refinada dos efeitos da ação da seleção no genoma da espécie de interesse (Allendorf et al., 2010).

Considerando-se então a importância de se identificar e caracterizar os polimorfismos que ocorrem ao longo do genoma de espécies de interesse, este trabalho foi conduzido com o objetivo identificar e caracterizar variantes de um único

nucleotídeo (SNPs) no genoma de *E. dysenterica*. Serão apresentados os resultados obtidos em tal busca e uma caracterização dos SNPs identificados, potenciais marcadores genéticos moleculares para serem utilizados em estudos de genômica de populações dessa espécie.

2 REVISÃO DE LITERATURA

2.1 A ESPÉCIE *Eugenia dysenterica*

2.1.1 **Aspectos gerais e biologia reprodutiva**

A cagaiteira (*Eugenia dysenterica* DC.) é uma espécie vegetal neotropical nativa do Brasil. É membro da família Myrtaceae, as quais também pertencem espécies de importância econômica tais como o eucalipto (*Eucalyptus sp.*) e a goiaba (*Psidium guajava*). Tanto o nome científico, quanto o popular, faz referência à propriedade laxativa de seu fruto, a cagaita (Chaves & Telles, 2006). Além disso, esta espécie possui como sinonímias os termos: *Myrtus dysenterica* Mart., *Hexachlamys macedoi* Legrand e *Stenocalyx dysentericus* (Mart.) O.Berg (Sobral et al., 2015).

Em aspectos geográficos, a cagaiteira possui ampla distribuição, ocorrendo em quase todas as regiões brasileiras, com exceção do sul do país (SOBRAL et al., 2015). Porém, essa espécie é encontrada com maior frequência nos estados de Goiás, Minas Gerais e Bahia, em regiões de cerrado stricto sensu e cerradões onde, na maioria das vezes, formam densos agrupamentos de indivíduos (Martinotto et al., 2008).

As folhas são dispostas de maneira oposta, com formas ovalares, glabras, aromáticas e caducas no período de floração. Já as flores de Cagaiteira são hermafroditas, completas e autocompatíveis, apresentando um padrão de floração, denominado por Proença & Gibbs, 1994 como “Big Bang”, no qual ocorre uma sincronização do florescimento em um curto período de tempo. Nas galhas, as flores são organizadas de forma solitária ou em trios, suas pétalas têm coloração branca, sendo as abelhas das espécies *Bombus atratus* e *B. morios* as principais responsáveis pela polinização (Silva et al., 2001; Telles et al., 2001a).

A relação floração/frutificação é menor em indivíduos mais jovens. Isso provavelmente acontece por conta do menor estabelecimento dessas plantas uma vez

que o sistema radicular ainda não se desenvolveu completamente, vindo a impedir o aproveitamento da água em camadas mais profundas do solo. A floração ocorre no período da primavera, mas a indução dos botões florais parece estar relacionada com outros fatores que não o regime pluviométrico (Souza et al., 2008).

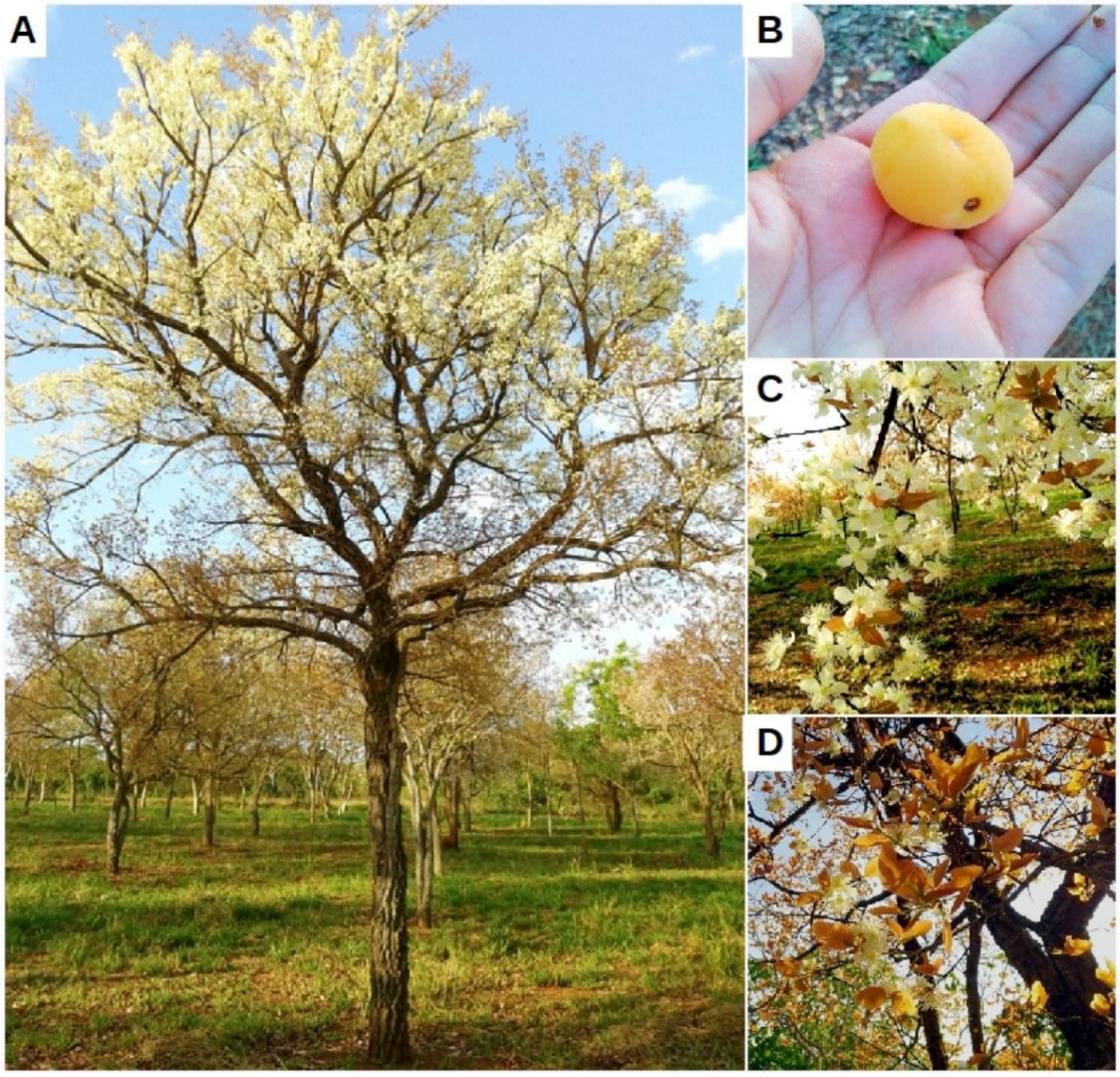

Figura 1. A cagaiteira: A) indivíduo adulto; B) fruto da cagaiteira maduro (cagaita); C) flores e D) folhas jovens. Fotos: Rhewter Nunes, set. 2014.

As flores da cagaiteira são brancas, com tamanho médio de 15 mm, podendo possuir de 60 a 70 estames e de 4 a 8 óvulos. A antese dura em média 6,5 horas ocorrendo preferencialmente no período da manhã (Proença & Gibbs, 1994). A espécie é predominantemente alógama, com autocompatibilidade, não havendo diferença estatisticamente significativa entre a produção de frutos resultantes de autofecundação e fecundação cruzada (Zucchi et al., 2003; Barbosa, 2014). A predominância de alogamia

também foi observada no estudo realizado por Telles et al., (2001a), onde foram analisadas dez populações naturais utilizando marcadores isoenzimáticos. Com esse estudo foi observada um taxa de fecundação cruzada (aparente) de 0,835 demonstrando que a cagaiteira possui um sistema misto de cruzamento.

Considerando o fluxo gênico, as sementes da cagaiteira são dispersas por animais (zoocoria) e não apresentam dormência. Esse fenômeno provavelmente está associado à própria morfologia do fruto que apresenta uma poupa carnosa, podendo então, servir como uma espécie de "recompensa" para um possível dispersor (Sano et al., 1995). Além das espécies de abelhas do gênero *Bombus* (*B. atratus* e *B. morio*), outras espécies de abelhas também são visitantes ocasionais como *Trigona spinipes* e *Melipona sp.*, mas não há indícios de que elas efetivamente polinizem as flores de cagaiteira (Proença & Gibbs, 1994).

2.1.2 **Caracterização genética molecular**

Diversos estudos têm sido realizados com o intuito de compreender os padrões de organização da variabilidade genética em populações de cagaiteira. Telles e colaboradores (2001a) utilizaram oito locos isoenzimáticos juntamente com dados de caracteres morfológicos e demográficos com o objetivo de identificar a relação entre dez populações naturais de cagaiteira e as características geográficas e edáficas das regiões que essas populações ocupam. Esse estudo apontou que a distribuição geográfica das populações é o principal fator determinante da diversidade genética. Já os caracteres fenotípicos (altura da planta, circunferência do colo, circunferência do tronco a 1,30 m, diâmetro de projeção da copa nos sentidos norte-sul e leste-oeste, área basal e área do tronco a 1,30 m do solo) são melhor explicados pelos padrões edáficos e pela distribuição espacial.

Outro estudo utilizando esse mesmo conjunto de locos isoenzimáticos (Telles et al., 2001b) foi realizado com o intuito de avaliar os padrões espaciais da variabilidade genética, mas nesse caso, inferindo os prováveis processos evolutivos que atuam na diferenciação das populações analisadas. Os resultados encontrados nesse estudo indicam que a evolução neutra deve provavelmente ser o fator responsável pela diferenciação genética das populações, através de uma relação entre fluxo gênico a curtas distâncias juntamente com a deriva genética ocorrendo dentro das populações, assim como é esperado em um modelo de isolamento-por-distância. Esse padrão

espacial na distribuição da variabilidade genética em *E. dysenterica* foi mais tarde confirmada por Telles e colaboradores (2003).

Utilizando um conjunto de dez pares de iniciadores transferidos de *Eucaliptus spp.*, Zucchi e colaboradores (2003) analisaram 116 indivíduos de *E. dysenterica* provenientes de dez populações naturais. Além da eficiência no uso de marcadores microssatélites (SSRs) transferidos de *Eucaliptus ssp.* para a análise de populações de cagaiteira, este estudo demonstrou alto nível de estruturação genética entre as populações estudadas (F_{ST} = 0,250) e baixo número de migrantes por geração (N_M= 0,680). Uma nova análise da estrutura genética e do fluxo gênico em *E. dysenterica* foi realizada por Zucchi e colaboradores (2005) utilizando 54 locos marcadores RAPD em três populações naturais da espécie. Novamente foi encontrada forte estruturação genética entre as populações (φ_{ST} = 0,270).

Ainda no contexto do uso de marcadores RAPD na análise de estrutura genética de cagaiteira, Trindade & Chaves (2005) analisaram 13 populações com 12 indivíduos cada uma. Os resultados corroboraram a estruturação genética já observada nos trabalhos anteriores que utilizaram SSRs e RAPDs demonstrando a existência de grande variabilidade genética e fenotípica estruturada entre populações.

Uma caracterização molecular da coleção de germoplasma de cagaiteira da Escola de Agronomia da Universidade Federal de Goiás foi realizada por Rodrigues (2012). Neste estudo, além da caracterização molecular, também foram avaliados o fluxo gênico via pólen e o sistema reprodutivo das cagaiteiras presentes na coleção. Essas análises foram realizadas utilizando um conjunto de oito locos microssatélites (quatro transferidos de *Eucaliptus* ssp. e quatro desenvolvidos para *E. dysenterica*). Com os resultados encontrados foi possível concluir que a coleção estabelecida na EA-UFG representa bem a diversidade presente nas populações naturais de *E. dysenterica* amostradas. A taxa de fecundação cruzada multiloco foi de 0,957. Também foi observada uma taxa de fecundação cruzada entre indivíduos aparentados de 10,9%. A maior parte dos cruzamentos verificados foram biparentais (84,7%), sendo aparentados no grau de meios-irmãos, além de apenas 4,6% das sementes terem sido formadas por autofecundação. Além disso, foi observado um alcance máximo de dispersão de pólen de 224 metros, abrangendo quase toda a área experimental da coleção de germoplasma.

Uma expansão nas análises de estruturação genética de populações de *E. dysenterica* foi realizada por Barbosa e colaboradores (2015). Nesse estudo foram analisados 736 indivíduos de 23 populações utilizando onze locos microssatélites (oito

locos transferidos de *Eucaliptus* ssp. e três desenvolvidos para cagaiteira (Telles et al., 2013). Foi observado um F_{ST} = 0,290 demonstrando alto nível de estruturação genética entre as populações e resultados de F_{ST} par-a-par demonstraram significativa correlação com as distâncias geográficas. Essas análises revelaram um gradiente contínuo de diferenciação entre as populações do noroeste ao sudoeste e não a formação de dois grupos distintos de populações como sugeriam trabalhos anteriores.

Não há relatos na literatura científica de estudos abrangentes relativos ao genoma de *E. dysenterica*. Nesse sentido, a própria família Myrtaceae é pouco representada quando comparada a outras famílias de plantas, tendo apenas um genoma nuclear completamente sequenciado e anotado, o da espécie *Eucalyptus grandis* (Myburg et al., 2014). Por se tratar de uma espécie neotropical, de grande importância ecológica e pouco estudada no sentido genômico, a cagaiteira é uma boa espécie a ter seu genoma sequenciado e caracterizado.

2.1.3 **Caracterização fenotípica**

Diversos trabalhos também têm sido realizados com o intuito de caracterizar fenotipicamente indivíduos de cagaiteira. Aguiar e colaboradores (2012) utilizaram das informações disponíveis sobre a variabilidade genética medida por diferentes marcadores moleculares para relacioná-la com a variação genética quantitativa de caracteres poligênicos. Para isso foi realizado um ensaio de progênies com indivíduos pertencentes a dez populações naturais do sudeste de Goiás nas condições experimentais da coleção de germoplasma da Escola de Agronomia – UFG (blocos completos casualizados).

Foram então estimados, em um período de quatro anos, a herdabilidade ao nível de média de progênies e o coeficiente de variação genética de cada população para dois caracteres: altura da planta e diâmetro do fuste. Correlação linear e regressão múltipla foram utilizadas para fazer a associação entre as variações moleculares e quantitativas aos níveis intra e interpopulacional. Foi observada uma fraca correlação entre as medidas de divergência observadas. Além disso, foi observada uma tendência à correlação linear positiva entre a divergência genética molecular intrapopulacional e a variação genética poligênica, o que sugere a necessidade de se trabalhar com uma ampla gama de marcadores dispersos pelo genoma em estudos dessa natureza (Aguiar et al., 2012).

Almeida-Junior e colaboradores (2014) também fizeram uma caracterização genética quantitativa das populações da coleção de germoplasma de cagaiteira da EA-UFG para outros caracteres. Nesse caso foram avaliados altura da planta, altura da primeira bifurcação, circunferência do caule a 10 cm do solo, projeção média da copa e variáveis das folhas tais como comprimento do limbo, largura do limbo, formato das folhas e comprimento do pecíolo. Os resultados apontaram que os caracteres relacionados ao tamanho da planta, circunferência do caule e diâmetro de projeção da copa não estão estruturados entre as populações, ao passo que, caracteres relacionadas a morfologia da folha estão.

A partir de uma amostragem mais abrangente das populações de cagaiteira presentes no Cerrado, Boaventura-Novaes (2014) realizou uma avaliação genética quantitativa e molecular de caracteres relacionados ao crescimento inicial dessas plantas. Um total de 3.000 sementes foram analisadas utilizando nove locos microssatélites, análise de variância e estimativas de Q_{ST}.

Observou-se significativa estruturação quantitativa entre as populações relativa aos caracteres de comprimento da raiz e taxa de crescimento de altura e diâmetro (Q_{ST} = 0,23 e 0,20). Já quanto aos caracteres de biomassa e emergência de plântulas não foi observada diferença significativa (Q_{ST} < 0,04). A relação $Q_{ST} – F_{ST}$ não foi significativa para 16 dos 18 caracteres em estudo indicando que a deriva genética é o principal contribuinte para a diferenciação dessas populações. Com os materiais utilizados nesse estudo foi estabelecida uma segunda coleção de germoplasma de cagaiteira (conservação *ex situ*) na Escola de Agronomia da UFG. Esta nova coleção possui tamanho efetivo 63 e a caracterização genética realizada sugere uma boa representatividade da variabilidade genética presente nas populações naturais.

2.2 BASE GENÉTICA DE ALGUNS DOS PRINCIPAIS TIPOS DE MARCADORES MOLECULARES UTILIZADOS EM GENÉTICA VEGETAL

Os marcadores moleculares podem ser definidos como o uso do polimorfismo existente em sequências homólogas como uma ferramenta de discriminação de indivíduos (Kumar et al., 2009). Esse polimorfismo pode ser devido a alterações em um único nucleotídeo (SNPs), inserções ou deleções (indels) ou rearranjos (translocações ou inversões). Além disso, esse polimorfismo pode ser detectado de diferentes maneiras: uso de enzimas de restrição, uso de sondas de DNA,

os RAPDs podem ser considerados marcadores com padrão de segregação típica de locos dominantes, sendo incapazes de determinar genótipos heterozigóticos (Schlötterer, 2004).

Os RAPDs apresentam como vantagem a simplicidade de protocolo e a rapidez na obtenção dos resultados. Diferente dos RFLPs, não é necessária uma grande quantidade de DNA, uma vez que os fragmentos são amplificados por PCR e, além disso, preservam a característica de serem amplamente distribuídos pelo genoma dos indivíduos. As limitações estão relacionadas à baixa reprodutibilidade devida à necessidade de padronização experimental para otimizar as condições das reações (Kumar et al., 2009).

2.2.3 **Polimorfismo de tamanho de fragmento amplificado ou AFLPs**

AFLP (*Amplified fragment length polymorphism*) é basicamente um método que mistura procedimentos dos RFLPs com o uso da PCR. Nessa técnica o DNA é clivado utilizando dois tipos de enzimas de restrição: enzimas de corte raro e enzimas de corte frequente. Adaptadores são ligados às extremidades dos fragmentos formados por serem complementares às regiões formadas pela ação das enzimas de restrição. Em seguida, é realizada uma reação de PCR seletiva onde apenas alguns dos fragmentos são amplificados. Os fragmentos têm seus tamanhos aferidos por eletroforese em gel de poliacrilamida e geralmente apresentam um padrão dominante de marcação (Schlötterer, 2004).

As vantagens desse tipo de tecnologia se encontram no fato de esses marcadores terem considerável reprodutibilidade, serem geradas muitas bandas informativas em uma mesma reação e o fato de não serem necessários conhecimentos de sequências previamente para o desenho dos *primers*, uma vez que eles possuem relação com os adaptadores. Como desvantagens, apresentam o fato de serem marcadores dominantes e, com isso, terem uma baixa quantidade de informação por loco, além de os custos com reagentes e equipamentos serem maiores, quando comparados a outros métodos (Kumar et al., 2009).

2.2.4 **Microssatélites ou SSRs**

Os marcadores baseados em polimorfismo de regiões microssatélites (SSRs - *Simple Sequence Repeats*) constituem uma das ferramentas mais populares para análise genética de indivíduos por conta da sua versatilidade e reprodutibilidade. Eles estão relacionados às repetições em tandem de motivos de nucleotídeos que podem variar de uma até seis bases. Os alelos então são inferidos por variação de comprimento da sequência resultante da variação no número de repetições dos motivos (Litt & Luty, 1989).

Essa classe de marcadores apresenta vantagens como uma alta taxa de mutação, aumentando a informação genética obtida para cada loco em análise, uma grande abundância no genoma de eucariotos e o fato de serem marcadores codominantes. No entanto, nem sempre se consegue obter uma abundância de locos que sejam polimórficos, principalmente quando se deseja genotipar em larga escala. Alguns estudos apontam para uma distribuição não uniforme de marcadores microssatélites no genoma (Schlötterer, 2004).

2.2.5 **Polimorfismo de base única ou SNPs**

Os SNPs (*Single-nucleotide polymorphism*) podem ser definidos como o polimorfismo que ocorre pela substituição de um nucleotídeo por outro em uma cadeia de DNA ou RNA. Essa substituição pode ser do tipo transição (substituição de uma purina por outra purina, ou uma pirimidina por outra pirimidina) ou transversão (substituição de uma purina por uma pirimidina, ou vice-versa). A principal vantagem desse tipo de marcador é a ampla distribuição e a alta densidade nos genomas (Schlötterer, 2004).

Uma das aplicações mais comuns do sequenciamento de nova geração tem sido a descoberta e a genotipagem de SNPs em larga escala. Esses marcadores têm sido utilizados para diferentes finalidades como, por exemplo, para a construção de mapas de ligação, análises de diversidade genética, mapas de associação e seleção assistida por marcadores Kumar et al. (2012). Isso tem sido aplicado para diversas espécies vegetais como: berinjela (Solanum melongena), girassol (Helianthus annuus), alcachofra (Cynara cardunculus), Sobreiro (Quercus suber), Capim-Zebra (Miscanthus sinensis), milho (Zea mays), mandioca (Manihot esculenta), panicum (Panicum virgatum) e batata (Solanum tuberosum) (Barchi et al., 2011; Pegadaraju et al., 2013; Scaglione et al.,

2012; Hipp et al., 2014; Slavov et al., 2014; Crossa et al., 2013; Prochnik et al., 2012; Lu et al., 2013; Uitdewilligen et al., 2013).

2.3 EVOLUÇÃO DAS TECNOLOGIAS DE SEQUENCIAMENTO DE ÁCIDOS NUCLEICOS

2.3.1 **Sequenciamento de Sanger**

Em 1977 foi desenvolvido e publicado um dos primeiros métodos de sequenciamento de DNA. O método de sequenciamento de Sanger (1977) foi empregado no primeiro sequenciamento completo de um genoma, o do bacteriófago Phi X174 (Sanger et al., 1978). Esse método consiste basicamente na síntese de uma fita complementar a partir de um molde de DNA utilizando uma mistura de nucleotídeos naturais, 2' - deoxinucleotídeos (dNTPs), e nucleotídeos terminadores modificados, 2', 3' - dideoxinucleotídeos (ddNTPs), que são adicionados à cadeia via DNA polimerase. O equilíbrio entre as reações de síntese e terminação resultam em um conjunto de fragmentos aninhados que diferem entre si em unidades de monofosfato de nucleosídeo. Esses fragmentos são então separados por tamanho utilizando eletroforese em gel de alta resolução e analisados para revelar a sequência de DNA (Metzker, 2005).

Posteriormente, o método de Sanger passou um processo de automatização onde cada ddNTP passou a ser marcado com uma fluorescência de cor específica. No sequenciador, os fragmentos marcados passam através de uma região de detecção em que os fluoróforos são excitados por um laser, produzindo emissões de fluorescência com uma das quatro cores atribuídas a cada um dos quatro ddNTPs. A cor da fluorescência é utilizada na atribuição de uma chamada de base e a ordem dos fragmentos fluorescentes revela a sequência de DNA (Metzker, 2005). Este método conferiu o Prêmio Nobel de química a Frederick Sanger em 1980 e foi rapidamente implementado pela indústria biotecnológica.

2.3.2 **Sequenciamento de nova geração**

2.3.2.1 Visão Geral

O termo “sequenciamento de nova geração”, ou Next Generation Sequencing (NGS), em inglês, é aplicado ao conjunto de tecnologias automáticas de sequenciamento de DNA que surgiram no mercado posteriormente ao método de Sanger, também chamado de sequenciamento de primeira geração (Shendure & Ji, 2008). Essas tecnologias permitiram a obtenção de uma grande quantidade de dados de sequências de DNA por corrida de forma rápida e com custo relativamente baixo (Barzon et al., 2013).

Apesar de possuírem diferenças significativas entre si, como, por exemplo, na forma de obtenção e determinação das sequências de DNA, de uma maneira geral, as plataformas de NGS são caracterizadas pela rápida obtenção de dados (Metzker, 2010). Isso se deve ao fato de essas plataformas utilizarem uma metodologia de sequenciamento paralelo, em que várias sequências são determinadas ao mesmo tempo. Essa característica é um diferencial frente ao método de Sanger em que se era limitado à análise de até 96 fragmentos por vez em um sequenciador de eletroforese (Metzker, 2010).

Outra característica que distingue as plataformas de NGS do método de sequenciamento de Sanger é a forma de obtenção das bibliotecas de fragmentos de DNA. No método de Sanger essas bibliotecas eram produzidas através da clonagem de fragmentos utilizando células vetores, o que no sequenciamento de nova geração são obtidas utilizando um segmento de DNA molde imobilizado em uma superfície sólida (Varshney et al., 2009).

2.3.2.2 Roche (454)

A plataforma 454, desenvolvida pela Roche, foi a primeira tecnologia de NGS disponibilizada no mercado. Ela utiliza como princípio básico para a determinação das bases nitrogenadas o pirosequenciamento que usa moléculas liberadas pela adição de bases específicas no alongamento de uma cadeia para produzir luz, que é então capitada por câmeras fotossensíveis e então determinada que tipo de base foi adicionada (Rothberg & Leamon, 2008).

Em uma reação de alongamento de uma cadeia de DNA *in vitro* são liberados, por vez, um determinado tipo de desoxinucleotídeo que, quando adicionado à cadeia, libera para o meio em que se encontra, uma molécula de pirofosfato. Essa molécula de pirofosfato é convertida em ATP (adenosina trifosfato) por uma enzima

chamada ATP-sulfurilase que será utilizada pela luciferase para oxidar a luciferina e então produzir um sinal de luz (Carvalho & Silva, 2010). A identificação de que determinado tipo de desoxinucleotídeo foi adicionado à cadeia é feita pela captação desse sinal de luz por uma câmera CCD (charge-coupled device) acoplada ao sistema (Shendure & Ji, 2008).

A preparação das amostras é realizada fragmentando-se o DNA aleatoriamente e posteriormente ligando a esses fragmentos adaptadores de sequência específica. Na extremidade 3' dos fragmentos é ligado um adaptador do tipo A e na extremidade 5' um adaptador do tipo B, tendo este último biotina ligada a ponta 5' que será utilizada na seleção dos fragmentos, pois apenas fragmentos A-B participam do restante da reação (Rothberg & Leamon, 2008).

Posteriormente, em uma reação de purificação, os fragmentos A-B são eluídos e ligados a microesferas que estão repletas de sequências complementares ao adaptador B, enquanto que no adaptador A irá anelar um primer para dar início à reação de sequenciamento. As microesferas serão capturadas por micelas formadas pela presença de óleo em solução aquosa e juntamente com os demais reagentes da PCR em emulsão ocorrerá a amplificação dos fragmentos presos às microesferas (Rothberg & Leamon, 2008).

Após realizada a PCR em emulsão, na qual são geradas milhares de cópias a partir de cada fragmento-molde (sequência alvo), as microesferas são encaminhadas para micropoços de uma placa de sílica juntamente com os reagentes para o sequenciamento. Cada micropoço aloca uma única microesfera e em cada um deles existe sistema óptico de leitura acoplado. O sequenciamento ocorre por ciclos e em cada ciclo é liberado para os poços um tipo de desoxinucleotídeo que, se incorporado à cadeia, liberará um sinal luminoso que será capitado pelo sistema (Rothberg & Leamon, 2008).

2.3.2.3 Illumina

No sequenciamento Illumina, assim como no sequenciamento de Sanger, são utilizados nucleotídeos terminadores marcados com diferentes fluoróforos. Nessa metodologia, porém, os fragmentos são imobilizados em uma superfície de vidro para serem clonados, processo conhecido como PCR em fase sólida. Os fragmentos são amplificados ao terem suas extremidades ligadas a oligonucleotídeos presentes nessa

superfície, formando uma estrutura em ponte (*bridge PCR*). Na etapa de amplificação são formados agrupamentos de fragmentos clonados (*clusters*) (Metzker, 2010).

O *base calling* (atribuição das bases na sequência) dependerá da emissão conjunta de sinais de adição de nucleotídeos dos fragmentos do *cluster*. A atribuição de uma nota de qualidade para as bases atribuídas (Phred) possui relação com a qualidade da emissão conjunta de incorporação de nucleotídeos nos fragmentos, fazendo com que os *reads* dessa plataforma tendam a ter a extremidade terminal com menor qualidade, uma vez que a densidade de fragmentos interferirá na captação de sinais mais profundos no *cluster* (Metzker, 2010).

De maneira simplificada o método utilizado nessa plataforma de sequenciamento consiste em: (1) fragmentar o DNA aleatoriamente empregando para isso um método físico de quebra das cadeias nucleotídicas; (2) ligar adaptadores do tipo A e B às extremidades dos fragmentos formados; (3) DNAs fita simples são imobilizados em um suporte sólido onde estão aderidos uma alta densidade de oligonucleotídeos complementares aos adaptadores adicionados aos fragmentos; (4) o oligonucleodídeo do suporte serve como primer para síntese de uma nova fita e a fita original é então descartada; (5) é realizada uma desnaturação da dupla fita e as extremidades livres dos fragmentos se pareiam a oligonucleotídeos vizinhos formando uma estrutura em ponte; (6) são fornecidos os reagentes e ocorre a primeira etapa de PCR a partir do oligoncleotídeo que funciona como um primer; (7) a temperatura é elevada desfazendo a estrutura em ponte e promovendo a desnaturação da dupla fita formada que irão posteriormente parear com outros oligonucleotídeos do suporte criando novas estruturas em ponte; (8) são realizados diversos outros ciclos de amplificação como este até que sejam formados *clusters* de moléculas idênticas ligadas ao suporte; (9) após realizada uma lavagem dos reagentes, são adicionados novos nucleotídeos marcados; (10) toda vez que um determinado nucleotídeo marcado for adicionado será liberado um sinal de luz com uma determinada cor no *cluster* que será captado e utilizado para construir a sequência (Carvalho & Silva, 2010).

2.3.2.4 Applied Biosystems (SOLiD)

A preparação das bibliotecas de sequenciamento na tecnologia SOLiD é muito similar à da plataforma 454, onde são utilizados microrreatores contendo *beads*, o molde de DNA, *primers* e os demais reagentes para a realização de uma PCR em

emulsão. O tipo de sequenciamento empregado nessa plataforma pode ser denominado de “sequenciamento-por-ligação”. Nela, é possível se gerar até 1,4M *reads* do tipo *paired-end*, com cerca de até 75 nucleotídeos cada, a uma taxa de erro de 0,01% Glenn (2011). Essa tecnologia apresenta uma taxa de erro menor em comparação a outras plataformas de NGS, no entanto os *reads* gerados por ela são pequenos, vindo a apresentar uma maior quantidade de *gaps* na etapa de montagem e uma distribuição mais heterogênea de *reads* quanto a sua posição no genoma que o sequenciamento Illumina, por exemplo (Glenn, 2011).

Na plataforma SOLiD, os fragmentos de DNA são gerados e ligados ao adaptador P1, que se liga especificamente a uma microesfera. Dessa forma, o sequenciamento ocorre por hibridização de sondas fluorescentes com o alvo. Inicialmente o *primer* (n) libera as primeiras bases da sequência alvo, para hibridização com a sonda, que ao hibridizar libera um sinal de fluorescência que é lido. Em seguida, as três últimas bases da sonda são removidas. Os ciclos de hibridização são repetidos até que o alvo seja todo coberto (35bp). A sequência fita dupla é desnaturada para uma nova etapa de sequenciamento com os *primers* (n-1), (n-2), (n-3) e (n-4) até que a sequência alvo seja determinada. Todas as combinações de dinucleotídeos são marcadas nas sondas com quatro fluoróforos. Portanto, são necessárias duas leituras de cada base para que o dinucleotídeo da sonda seja removido (Carvalho & Silva, 2010).

2.3.2.5 Pacific Biosciences (PacBio)

Os sequenciadores PacBio utilizam uma tecnologia onde DNA polimerases ficam ancoradas em uma superfície de vidro e estão alocadas no fundo de uma estrutura denominada *zero-mode waveguide* (ZMW). Cada ZMW é câmara pequena o suficiente para que geralmente apenas uma molécula de DNA permaneça em seu interior. A química de sequenciamento se dá pela excitação de nucleotídeos marcados que são adicionados às ZMW no momento de sequenciamento. Ao fundo dessas estruturas existe um laser que irá fazer essa excitação apenas dos fluoróforos incorporados pela DNA polimerase (Metzker, 2010).

Esses sequenciadores possuem a capacidade de gerar cerca de 1 Gb por *smart cell*. Os *reads* gerados por essa tecnologia são maiores quando comparadas aos *reads* Illumina, podendo chegar a mais de 50 Kb. Os *reads* isoladamente possuem uma baixa acurácia (~85%), no entanto, uma mesma molécula de DNA pode ser sequenciada

mais de uma vez. Dessa forma, quando maior a profundidade de sequenciamento, melhor se identifica os erros de sequenciamento e melhora-se a acurácia das leituras de sequenciamento (Eid et al., 2009).

2.4 GENOTIPAGEM POR SEQUENCIAMENTO

2.4.1 **Visão geral**

A grande quantidade de dados gerados pelas plataformas de NGS possibilitou o desenvolvimento de diferentes abordagens de extração de informações dos genomas. Dentre essas abordagens pode-se citar o sequenciamento de representações reduzidas e o sequenciamento de genomas inteiros. Há situações, por exemplo, em que o sequenciamento do genoma inteiro de cada indivíduo de uma população é caro ou mesmo desnecessário. Nesse caso pode-se responder diversas questões biológicas através de polimorfismos presentes em um subconjunto de regiões do genoma. Esse é um caso de sequenciamento de representações reduzidas e para a descoberta desses marcadores é utilizado um método de genotipagem por sequenciamento (Davey et al., 2011).

Para se trabalhar com o sequenciamento (via NGS) de representações reduzidas na busca por marcadores moleculares é necessário dispor de métodos de redução de complexidade do genoma. Para tal podem ser usadas diferentes abordagens tais como uso de enzimas de restrição, RNA-Seq ou sequências de captura (Davey et al., 2011).

Com a utilização de métodos de redução de complexidade de genoma baseados em enzimas de restrição é possível se obter de forma eficiente representações do genoma. A eficiência desse método se torna maior para espécies que possuem o genoma rico em regiões repetitivas, nos quais enzimas sensíveis à metilação podem ser utilizadas como forma de não amostrar essas regiões genômicas que, em geral, são pobres em polimorfismos (Davey et al., 2011). Uma limitação para as plataformas de genotipagem que usam esse método é a necessidade de otimização das enzimas para estudos com populações muito divergentes e com novas espécies (Illumina, 2013).

O sequenciamento de nova geração de cDNA (RNA-Seq), método geralmente aplicado para se medir níveis de expressão gênica (Wang et al., 2009), também pode ser utilizado para se obter SNPs. Os custos para tal procedimento se

aproximam dos valores gastos com os métodos que utilizam enzimas de restrição e possuem uma maior chance de descobrir marcadores relacionados a genes ativos (Chepelev et al., 2009). Porém uma desvantagem do método é a possível dificuldade em se inferir genótipos por conta dos múltiplos transcritos alternativos (Davey et al., 2011).

Caso as sequências das regiões de interesse sejam conhecidas, elas podem ser selecionadas por métodos de captura de sequência como SureSelect, Nimblegen e Raindance. Sequências de oligonucleotídeos previamente desenvolvidas podem se ligar a regiões de interesse, um éxon, por exemplo, e gerar diversas cópias dessas regiões antes do sequenciamento. Esses métodos são altamente precisos e de alta qualidade, mas requerem uma forte interação com as regiões relacionadas, correndo o risco de não se ligarem fortemente em populações que se diferenciem muito da sequência de referência (Davey et al., 2011).

A alta densidade de SNPs gerados por essas tecnologias tem auxiliado em estudos com diferentes aplicações biológicas e agronômicas, trazendo não só informações relacionadas a mapeamento de QTLs, seleção assistida por marcadores e seleção genômica ampla, mas também sobre populações naturais trazendo grandes benefícios para a genética da conservação e a estudos de genética ecológica (Allendorf et al., 2010). Além disso, o desenvolvimento de painéis de SNPs possuem o viés dos marcadores estarem relacionados com a população que foi utilizada para obtenção das variantes (*ascertainment bias*). Sendo assim, é importante o desenvolvimento de conjunto de marcadores funcionais para cada espécie (Davey et al., 2011).

2.4.2 **GbS de Cornell**

GbS (*Genotyping-by-Sequencing*) como o próprio nome sugere, é uma plataforma de genotipagem por sequenciamento que utiliza da tecnologia de NGS para gerar milhões de bibliotecas de *short reads* simultaneamente e melhorar a cobertura de sequenciamento por loco através da redução da complexidade do genoma utilizando enzimas de restrição (Poland e Rife, 2012). Pelo fato de a descoberta e a genotipagem de SNPs ocorrer de forma simultânea nessa plataforma, ela consegue produzir um grande volume de dados genotípicos de forma rápida e com um baixo custo por amostra (Chen et al., 2013).

De maneira simplificada, o protocolo para construção das bibliotecas de GbS descrito por Elshire et al. (2011) pode ser descrito nos seguintes passos: (1)

preparação de uma placa que contenha as amostras de DNA, os adaptadores do tipo barcode e os adaptadores do tipo comum; (2) digestão do DNA utilizando as enzimas de restrição ApeKI; (3) os adaptadores são ligados as extremidades dos fragmentos formados com a adição de enzima T4 (ligase), tampão de ligase e ATP (adenosina trifosfato); (4) inativação da enzima T4; (5) adição dos *primers* e realização de uma PCR para amplificação dos fragmentos ligados a adaptadores que irão formar vários *pools*; (6) lavagem para a retirada dos produtos da PCR e avaliação do tamanho dos fragmentos da biblioteca resultante em um analisador de fragmentos, onde os que não tiverem se ligado aos dois adaptadores não irão para o sequenciamento.

2.4.3 **RAD-Seq**

Outro método que utiliza enzimas de restrição é o RAD-Seq. RAD-tags (*Restriction-site Associated DNA-tags*) são pequenas sequências próximas aos sítios de reconhecimento de enzimas de restrição, que funcionam portanto, como marcadores para essas regiões (Davey & Blaxter, 2010). Eles são o princípio para a plataforma de genotipagem por sequenciamento RAD-Seq, que através do uso de enzimas de restrição, consegue obter uma alta densidade de marcadores SNP (*Single Polymorphism Nucleotide*) em amostras com um grande número de indivíduos (Baird et al., 2008).

Os marcadores RAD-tags foram primeiramente utilizados em estudos com *microarrays* (Miller et al., 2007) e posteriormente adaptados para NGS (Baird et al., 2008). Embora a utilização dessas sequências (RAD-tags) conjunta a *microarrays* permita a exibição paralela de vários marcadores polimórficos para estudos de variabilidade genética e mapeamento de mutações de diversos organismos, esse tipo de aplicação permite analisar apenas uma parte dos polimorfismos segregantes (Baird et al., 2008).

O método RAD-Seq (Baird et al., 2008; Davey & Blaxter, 2010) consiste em primeiramente digerir o DNA genômico utilizando enzimas de restrição (ER) e às extremidades dos fragmentos formados ligar adaptadores do tipo P1 (Figura 1). Esses adaptadores são formados por um adaptador de sequenciamento Illumina, um identificador molecular (MID - *Molecular Identifier*) ou *barcode* de 4 ou 5 pb de comprimento que irá identificar a amostra e uma sequência final correspondente a um sítio de restrição (Davey & Blaxter, 2010).

As sequências ligadas aos adaptadores são posteriormente reunidas e separadas em sequências de tamanho específico. Posteriormente são adicionados adaptadores P2, que possuem as extremidades divergentes, também chamados de adaptadores *Y-shaped*. Em seguida os fragmentos são amplificados em uma reação de PCR com *primers* para adaptadores P1 e P2. Somente os fragmentos que tiverem ambos os adaptadores (P1 e P2) serão amplificados e em seguida as amostras serão encaminhadas para o sequenciamento (Baird et al., 2008).

2.4.4 **DArT-Seq**

O método de genotipagem por sequenciamento DArT-Seq é derivado da tecnologia DArT que já utilizava da redução de complexidade do genoma das amostras a serem genotipadas. O que ocorreu foi uma migração de uma metodologia que empregava hibridização de sondas (microwarrays) para uma que utiliza uma plataforma NGS (Sansaloni, 2012).

A tecnologia DArT consiste em: (1) a criação de um grupo de DNA genômico total resultante da união de todas as amostras; (2) submissão desse material à redução de complexidade genômica via corte com enzimas de restrição; (3) o *pool* de representações genômicas obtidas é clonado dando origem à bibliotecas de fragmentos individuais; (4) os fragmentos das bibliotecas são imobilizados em um microarranjo; (5) o DNA das amostras individuais são digeridos utilizando as mesmas enzimas de restrição aplicadas ao grupo anterior dando origem aos *targets*; (6) os targets obtidos são marcados com fluorescência e posteriormente hibridizados ao microarranjo formado anteriormente; (7) é realizada a lavagem, a digitalização dos slides e as análises de detecção de polimorfismos que se dá via presença ou ausência de hibridização (Huttner et al., 2005; Sansaloni, 2012).

O procedimento de genotipagem por sequenciamento estabelecido para DArT-Seq é muito parecido com o de GBS e difere em pontos como o conjunto de enzimas utilizadas para promover a redução da complexidade do genoma e a forma de tratamento dos dados obtidos após o sequenciamento. O método descrito por Sansaloni e colaboradores (2010) e consiste nos seguintes passos: (1) digestão do DNA genômico utilizando grupos de enzimas de restrição; (2) ligação de adaptadores do tipo barcode e do tipo comum; (3) amplificação dos *targets* que são os fragmentos com ambos os adaptadores; (4) remoção das falhas ou dímeros; (5) geração do *pool* de *targets* e

purificação; (6) geração de *clusters* e sequenciamento em plataforma de NGS (Illlumina); (7) alinhamento das *reads* em um genoma de referência e verificação dos polimorfismos.

2.4.5 **RAPiD-Seq**

Dentre os métodos de genotipagem por sequenciamento que utilizam sequências de captura como estratégia de redução da complexidade de genomas, no contexto da genética de plantas, um têm ganhado destaque: RAPiD-Seq. Com o intuito de simplificar ainda mais o processo de genotipagem a plataforma RAPiD-Seq propõe uma metodologia que dispensa os passos convencionais empregados em outros métodos através de uma técnica que utiliza dois passos de PCR para gerar uma redução da complexidade do genoma e, em seguida, realiza o sequenciamento e a detecção dos polimorfismos (Resende Júnior et al., 2012).

Para se chegar a esse objetivo o método empregado passa pelos seguintes passos: (1) é realizada uma primeira PCR que ocorre de forma independente para cada amostra a ser genotipada. Nessa etapa são utilizados *primers* modificados que contém uma sequência específica que determina o número e a distribuição dos anelamentos pelo genoma; (2) com as amostras derivadas da primeira PCR é realizada então uma outra. Nessa etapa os indivíduos são reunidos para serem sequenciados juntos como forma de se aumentar o rendimento na preparação das amostras; (3) as amostras são então encaminhadas para um sequenciador de nova geração e (4) os dados das sequências são então analisados e as informações genotípicas são extraídas utilizando algoritmos específicos (http://www.rapid-genomics.com/).

Os *primers* empregados nesse método são modificados de forma que em suas sequências se encontre um *barcode* para identificar as amostras quando estas estão agrupadas para o sequenciamento, um *primer* de sequenciamento via NGS, uma sequência degenerada que confere estabilidade ao *primer* e uma sequência específica que permita o anelamento em regiões específicas do genoma. Os diferentes tipos de *primers* permitem uma flexibilidade quanto ao número desejado de locos a serem genotipados levando-se em consideração as necessidades da pesquisa ou a espécie estudada (http://www.rapid-genomics.com/).

2.5 FERRAMENTAS DE BIOINFORMÁTICA PARA IDENTIFICAÇÃO E GENOTIPAGEM DE SNPS A PARTIR DE DADOS DE SEQUENCIAMENTO DE NOVA GERAÇÃO

As tecnologias de sequenciamento de nova geração geraram (e continuam gerando) uma grande quantidade de dados e isso impactou diretamente na forma como se armazena, processa e interpreta essa informação. Nesse sentido, surgiu a necessidade de novos recursos computacionais e humanos que estejam preparados para lidar com essa grande quantidade de dados. Com essa necessidade nasceu o campo da bioinformática, que engloba profissionais que trabalham na interface entre a ciência da computação, a estatística e a biologia (Metzker, 2010).

2.5.1 **Análises precedentes à identificação e genotipagem de SNPs**

Quando se deseja realizar análises de identificação e genotipagem de SNPs para uma determinada espécie alguns recursos são necessários previamente. Um dos recursos mais importantes para se realizar essas análises é uma sequência de referência, que será utilizada para detectar a existência de polimorfismo ao confrontá-la com os *reads* em um alinhamento. No entanto, muitas vezes a espécie em estudo não tem o genoma sequenciado, criando a necessidade de se gerar uma sequência de referência utilizando um programa de montagem *de novo*.

Existem diversos programas disponíveis para se realizar a montagem de um genoma de referência a partir de leituras de sequenciamento (*reads*). Dentre eles, pode-se citar como os mais populares: Velvet (Zerbino & Birney, 2008), SOAP-denovo (Luo et al., 2012), ABySS (Simpson et al., 2009) e MaSuRCA (Zimin et al., 2013). No entanto, ainda que existam diversos trabalhos que fazem comparações entre eles, não existe um consenso de qual vem a ser o melhor montador.

Com essa indecisão sobre qual vem a ser a melhor ferramenta de montagem de genomas, a todo o momento surgem novas ferramentas e novas versões dos programas já existentes. Dentre os montadores mais recentes, o programa dipSPAdes surge como uma alternativa interessante (Safonova et al., 2015).

Esse montador é baseado em um *assembler* de genomas procariotos e organelares chamado SPAdes (*St. Petersburg genome assembler*). O dipSPAdes é considerado o primeiro programa desenvolvido especificamente para montagem de genomas diploides altamente polimórficos. A proposta desse montador é tirar proveito

da divergência existente entre haplomas em regiões repetitivas do genoma, gerando assim, *contigs* mais longos (Safonova et al., 2015).

Montar um genoma diploide pode resultar em dois tipos diferentes de *contigs*, sendo eles os *haplocontigs* e os *consensus contigs*. *Haplocontigs* são os *contigs* que representam um dos dois haplomas individualmente, enquanto que, *consensus contigs* representam um consenso entre os dois haplomas para as regiões ortólogas. Um *consensus contig* não é um bom representante de um dos haplomas uma vez que é um mosaico entre esses dois. Como algumas regiões do genoma são mais polimórficas que outras, os montadores convencionais produzem uma montagem com uma mistura de *haplocontigs* e *consensus contigs* quando tentam montar um genoma altamente polimórfico (Safonova et al., 2015).

Como um montador de genomas diploides altamente polimórficos, o que o dipSPAdes busca fazer é montar separadamente os *haplocontigs* e os *consensus contigs*. Ele utiliza dos grafos *de Bruijn* produzidos pelo SPAdes para mascarar polimorfismos nos *contigs* e tentar produzir uma representação mais abrangente do genoma para ambos os tipos de *contigs*. Os resultados obtidos por Safonova et al. (2015) mostram um incremento na montagem de genomas com maior polimorfismo. Isso torna o dipSPAdes um montador extremamente interessante para ser utilizado para plantas diploides em que se espera um alto polimorfismo, como é o caso da cagaiteira.

Além dos montadores de genoma, outro tipo de programa é extremamente importante para a identificação e genotipagem de SNPs: os alinhadores. Assim como no contexto de montagem de genomas, existem diversos programas disponíveis para esse fim. Dentre os alinhadores, merecem destaque: Bowtie2 (Langmead & Salzberg, 2012), BWA (*Burrows-Wheeler Aligner*) (Li & Durbin, 2009) e Novoalign (Novocraft, 2010). Nesse contexto, também não há uma definição clara de qual vem a ser o melhor programa para esse fim, mas o BWA parece apresentar um bom balanço entre tempo de execução, uso de memória e acurácia (Liu et al., 2013).

2.5.2 **Identificação e genotipagem de SNPs**

As tecnologias de sequenciamento de nova geração tem aumentado cada vez mais a quantidade de sequências de diferentes espécies em nível exponencial. Acompanhando a capacidade de gerar cada vez mais dados de sequenciamento vem sendo o decréscimo no custo por base. No entanto, o desenvolvimento de ferramentas

para organizar e analisar toda essa informação ainda permanece um desafio (Liu et al., 2013). Em se tratando da descoberta e genotipagem de SNPs, os *pipelines* comumente apresentam um alinhador e um *valiant caller* como programas principais.

Existem diversos programas disponíveis para identificação e genotipagem de SNPs por meio de dados de sequenciamento de nova geração. Dentre eles pode-se destacar: FreeBayes, GATK e SAMtools.

Um dos programas mais utilizados em análises de *SNP calling* é o SAMtools (Li et al., 2009), pelo uso da ferramenta *mpileup*. Esse é um programa escrito em linguagem *C* e pode ser utilizado para identificar SNPs e *indels* em uma ou muitas amostras (*multi sample*). Ele usa um modelo revisado de MAQ (*Mapping and Assembly with Quality*) para estimar erros de sequenciamento. Suas recomendações de uso incluem os passos de ordenação, indexação e remoção de duplicatas de PCR dos arquivos de alinhamento, que são os passos comuns de controle de qualidade de mapeamento desse tipo de análise.

O FreeBayes (Garrison & Marth, 2012) é um programa que utiliza abordagem bayesiana para detectar SNPs, *indels* e MNPs (multi-nucleotide polymorphisms) utilizando arquivos de alinhamento. Assim como o SAMtools, esses arquivos passam pelos processos básicos de controle de qualidade de mapeamento dos arquivos BAM (*Binary Alignment/Map*) e pode ser utilizado para múltiplas amostras. Esse programa foi desenvolvido em C/C++, mas também utiliza pacotes em *python*, *perl* e *shell*.

Um programa emergente nesse sentido tem sido o GATK (*Genome Analysis Toolkit*) (DePristo et al., 2011). Ele é um pacote de ferramentas implementadas em Java para análise de alta densidade de dados, desenvolvido pela equipe de *Data Science and Data Engineering* do *Broad Institute*. Segundo o próprio site do programa, essa ferramenta é focada na garantia de qualidade da descoberta e genotipagem de variantes.

O pipeline sugerido pelo manual de melhores práticas do GATK possui três conjuntos de etapas principais: 1) Pré-processamento dos dados, principalmente relacionado ao controle de qualidade de mapeamento; 2) a identificação de variantes e 3) as análises preliminares pós identificação das variantes. Esse conjunto de passos é apenas uma sugestão dos desenvolvedores e pode ser flexibilizado conforme casos específicos (DePristo et al., 2011).

Na etapa de pré-processamento, os *reads* (formato *fastq*) que passaram por controle de qualidade de sequenciamento são alinhados em uma referência (que pode

ser também um *draft-assembly* obtido por montagem *de novo*) gerando um arquivo no formato SAM (*Sequence Alignment/Map*). Esse arquivo de alinhamento é então convertido para uma versão binária (BAM), as informações dos grupos de *reads* são inseridas, os *reads* que iniciam e terminam nos mesmos pontos são marcados como prováveis duplicatas de PCR e então é feita a ordenação dos *reads* no BAM conforme coordenada de alinhamento. A etapa de marcar as duplicatas de PCR é um passo importante, pois esse tipo de artefato pode inflar a ocorrência de um falso alelo, fazendo com que seja atribuída uma nota de qualidade alta para uma falsa variante identificada nesse sítio (DePristo et al., 2011).

Ainda no conjunto de etapas de pré-processamento, um conjunto de *indels* conhecidos são utilizados para fazer um realinhamento nas regiões de ocorrência destes. Esse passo é necessário, pois, muitas vezes, os alinhadores tendem a mapear *reads* com *mismatches* nessas regiões e produzir evidências errôneas de ocorrência de SNPs nesses locais. Em seguida os arquivos BAM passam pelo processo de recalibração dos scores de qualidade das bases sequenciadas (*Base Quality Score Recalibration – BQSR*) (DePristo et al., 2011).

Um grande diferencial do GATK é a existência dessa etapa de recalibração dos scores de qualidade das bases sequenciadas. A chamada de variantes é fortemente dependente da qualidade das bases dos *reads*, porém as estimativas de erro por base são feitas pelos sequenciadores e estão sujeitas a diversas fontes de erro técnico de forma sistemática, levando a super ou subestimativas da qualidade das bases. Dessa forma, BQSR é um processo em que é aplicado métodos de aprendizagem de máquina para modelar os erros de sequenciamento empiricamente e reajustar os valores de qualidade para cada uma das bases, melhorando a precisão das chamadas de variantes.

No processo de recalibração da qualidade das bases são considerados diferentes fatores que são denominados de co-variantes. São eles: 1) a nota de qualidade que é relatada; 2) a posição do sítio de variação no *read* (uma vez que no sequenciamento Illumina, bases da porção terminal do *read* tendem a ter qualidade inferior) e 3) a base atual e a base anterior (efeito gerado pela química de sequenciamento). O BQSR envolve duas etapas principais: primeiramente o programa constrói uma tabela que serve de modelo de co-variação baseada nos dados presentes no arquivo BAM e em um conjunto de variantes conhecidas (que podem ser obtidas com uma rodada de identificação de variantes, seguida de filtros rígidos, caso não se tenha

um conjunto assim para a espécie que se esteja trabalhando), e então, é feito o ajuste na qualidade das bases sequenciadas baseadas no modelo criado (DePristo et al., 2011).

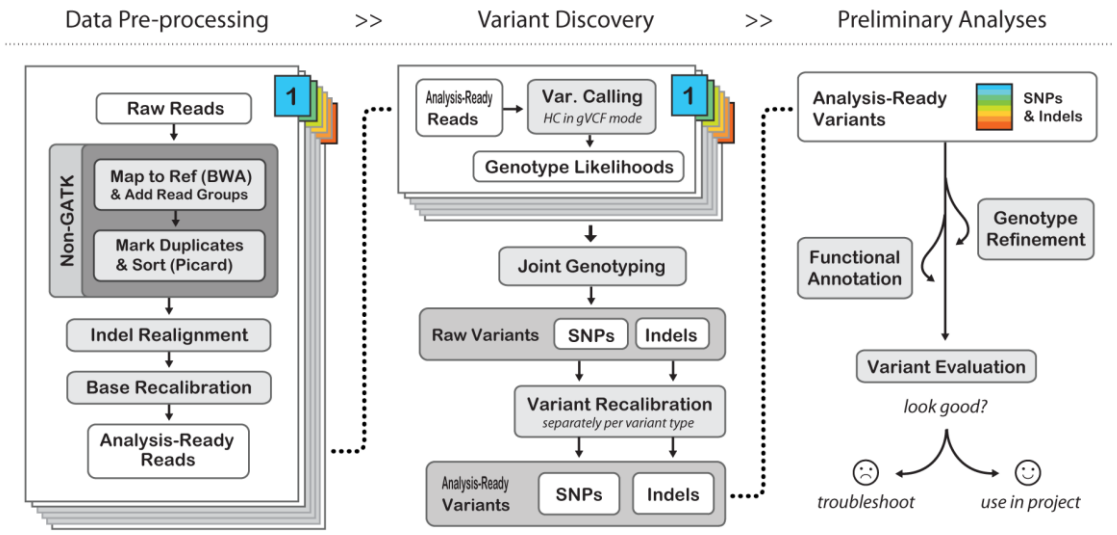

Figura 2. Pipeline de identificação de variantes do GATK. Fonte: https://www.broadinstitute.org/gatk/

Após passarem pelo controle de qualidade de mapeamento e recalibração da qualidade das bases, arquivos BAM podem ser utilizados para identificar as variantes. Para isso o GATK possui uma ferramenta denominada HaplotypeCaller. Esse algoritmo trabalha montando os *reads* para criar haplótipos potenciais, realinhando esses *reads* nesses haplótipos para estimar os que são mais prováveis e depois reprojetando os *reads* de volta para o *assembly*. Isso é utilizado para se estimar a confiança dos SNPs baseado no contraste entre todos os *reads*. Nesse ponto então, pode-se obter diversas estimativas para as variantes, tais como: QD (QualByDepth) - confiança em uma variante dividido pela profundidade das amostras que divergem da referência; FS (FisherStrand) – valor de *phred* escalonado pelo p-valor de um teste Exato de Fisher feito para detectar algum viés gerado por variantes em *reads* com sentidos diferentes; MQ (MappingQuality) – raiz quadrada da qualidade de mapeamento dos *reads* em todas as amostras; e MQRankSum (MappingQualityRankSumTest) – teste de médias Mann-Whitney ranqueado aplicado às variantes heterozigotas (DePristo et al., 2011).

Tanto na etapa de realinhamento de *indels*, quanto na de recalibração dos *scores* de qualidade das bases sequenciadas, é necessário um conjunto de variantes já conhecidos. Esses conjuntos de dados, por muitas vezes, não estão disponíveis para a

espécie em estudo. Uma solução proposta pelos próprios desenvolvedores do GATK, nesse caso, é fazer uma rodada de identificação de variantes com o HaplotypeCaller, separar os SNPs e *indels* do arquivo VCF (*Variant Call Format*) e aplicar filtros rígidos para cada um deles, utilizando-os posteriormente como conjuntos de dados de confiança.

A necessidade de conjuntos de dados previamente conhecidos é um fator que dificulta a execução do GATK e o coloca em uma posição de maior complexidade de utilização comparado a outros *valiant callers*. Entretanto, tem-se observado que essa sofisticação no protocolo de execução traz benefícios para a qualidade da chamada de variantes e da genotipagem. Comparações com outros programas tem demonstrado que o GATK destaca-se por identificar a maior quantidade de variações de forma correta (Liu et al., 2013).

3 MATERIAL E MÉTODOS

3.1 EXTRAÇÃO DO DNA GENÔMICO E SEQUENCIAMENTO DE NOVA GERAÇÃO

Amostras de tecido foliar foram obtidas de cinco indivíduos pertencentes à coleção de germoplasma de cagaiteira da Escola de Agronomia da Universidade Federal de Goiás. Esses indivíduos são oriundos de sementes coletadas nas áreas de ocorrência natural da espécie, sendo pertencentes a diferentes populações naturais estruturadas geneticamente entre si (Rodrigues, 2012). Indivíduos descendentes de populações naturais diferentes foram utilizados buscando-se detectar uma maior quantidade do polimorfismo existente na espécie.

A extração do DNA genômico foi realizada segundo o protocolo CTAB 2% (Doyle & Doyle, 1987). A quantidade e a qualidade do DNA obtido foram avaliadas utilizando-se os equipamentos Qubit (*Life Technologies*) e Bioanalyzer (*Agilent Technologies*). As bibliotecas de DNA genômico para sequenciamento foram construídas utilizando-se o kit *Nextera DNA Sample Preparation* (Illumina) seguindo-se o protocolo estabelecido pelo fabricante. As bibliotecas foram sequenciadas na plataforma Illumina MiSeq. O sequenciamento foi realizado no Laboratório de Genética e Biodiversidade (LGBio), do Instituto de Ciências Biológicas da Universidade Federal de Goiás.

Foram preparadas cinco bibliotecas e realizadas duas corridas de sequenciamento, sendo utilizada um *lane* para cada uma das corridas. A primeira corrida foi feita com um único indivíduo de *E. dysenterica*, enquanto que a outra foi feita em pool com quatro indivíduos utilizando *barcodes* para diferenciá-los posteriormente.

3.2 CONTROLE DE QUALIDADE DOS DADOS DE SEQUENCIAMENTO

Os arquivos em formato *fastq* foram avaliados quanto a qualidade utilizando o programa FastQC (Andrews, 2010). Os módulos "*Per base sequence quality*", "*Per base sequence content*", "*Overrepresented sequences*" e "*Adapter content*" foram considerados para se avaliar as bibliotecas quanto a qualidade das bases sequenciadas, a presença de contaminantes e sequencias super-representadas no sequenciamento e, posteriormente, tomar decisões quanto ao controle de qualidade. Os módulos "*Basic statistics*" e "*Sequence Length distribution*" foram utilizados em uma avaliação geral das características das bibliotecas tais como o número de sequências geradas, a porcentagem de GC e a média do tamanho das sequências.

O controle de qualidade dos dados foi realizado pela utilização do programa Trimmomatic (Bolger et al., 2014). Bases de baixa qualidade identificadas no começo e no final das sequências visualizadas com o FastQC foram retiradas usando as funções HEADCROP e CROP respectivamente. A função SLIDINGWINDOW foi utilizada para escanear os reads em uma janela de quatro bases e cortá-los quando a média de qualidade phred fosse inferior a 15. Para a remoção de adaptadores e outras sequências Illumina foi utilizada a função ILLUMINACLIP.

3.3 ESTIMATIVA DE TAMANHO E DRAFT-ASSEMBLY DO GENOMA DE *Eugenia dysenterica*

Após passarem pelo controle de qualidade, os reads foram utilizados para a montagem de novo do genoma de *E. dysenterica*. A montagem foi realizada pelo programa dipSPAdes (Safonova et al., 2014). Esse assembler foi desenvolvido especificamente para genomas diploides altamente polimórficos, a partir do SPAdes (Nurk et al., 2013), assembler de genomas procariotos. O software faz uso dos grafos de Bruijn e busca obter um consenso entre contigs que representam os dois haplomas de um organismo diploide, produzindo sequências mais contíguas. Além do *assembly*, a estimativa do tamanho do genoma de *E. dysenterica* também foi obtida por esse programa.

O assembly obtido foi avaliado utilizando-se o script assemblathon_stats.pl, desenvolvido por Keith Bradnam (2011). Foram considerados como descritores de qualidade da montagem as métricas: número de contigs, tamanho total dos contigs, porcentagem recuperada do tamanho estimado do genoma, tamanho do maior contig, número de contigs maiores que 10 Kb, tamanho médio dos contigs e N50.

O draft assembly obtido foi submetido à filtragem para remoção de contigs correspondentes aos genomas organelares de cloroplastos e mitocôndrias. Os contigs correspondentes aos genomas organelares foram identificados pela utilização da ferramenta blast (Altschul et al., 1990) contra o banco de dados de sequências de referência (RefSeq) de cloroplastos e de mitocôndrias vegetais do NCBI (National Center for Biotechnology Information). Os contigs com hits com e-values menores ou iguais a 1e-16, query-cover de 80% e identidade de 90% foram retirados do assembly. Além disso foram removidos do arquivo da montagem os contigs que inferiores a 500 bases.

Após a remoção dos contigs relacionados a sequências de cloroplastos e mitocôndrias, o programa RepeatMasker (Smit et al., 2015) foi utilizado para mascarar as regiões de DNA repetitivo utilizando como referência o banco de dados de elementos repetitivos de eudicotiledôneas e o banco de dados de elementos repetitivos construído *de novo*, especificamente para *E. dysenterica* (Ribeiro, 2015, comunicação pessoal) utilizando-se o RepeatModeler (Smit & Hubley, 2015).

3.4 IDENTIFICAÇÃO DE SNPs

A identificação de variantes foi realizada utilizando-se o pacote GATK - The Genome Analysis Toolkit (McKenna et al., 2010). Os reads que passaram pelo controle de qualidade foram mapeados no draft assembly livre de sequências organelares e com regiões repetitivas mascaradas utilizando-se o software BWA (Li & Durbin, 2009). Diversas funções do pacote Picard-Tools (http://broadinstitute.github.io/picard/) foram utilizadas para manipular o arquivo de alinhamento obtido e prepará-lo para uso no GATK (Figura 3).

A função SortSam foi utilizada para converter os arquivos do formato SAM (Sequence Alignment/Map) para o formato BAM (Binary Alignment/Map), bem como, ordená-los por coordenada de alinhamento. A função MarkDuplicates foi utilizada para marcar os reads identificados como prováveis duplicatas de PCR. A função BuildBamIndex foi utilizada para criar os arquivos de índice para cada arquivo BAM. A função AddOrReplaceReadGroups foi utilizada para inserir os dados de identificação nos arquivos de alinhamento, necessários ao processo de recalibração de qualidade das bases realizado pelo GATK.

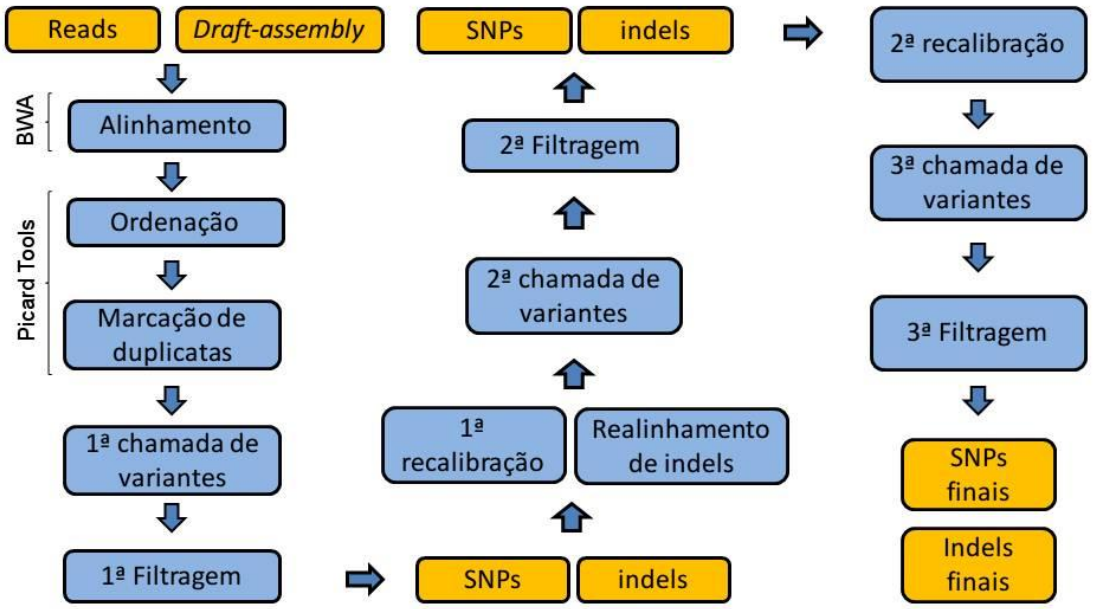

Figura 3. Organograma da análise de identificação de SNPs no genoma de *Eugenia dysenterica* utilizando uma adaptação do *pipeline* estabelecido pelos desenvolvedores do programa GATK. Os retângulos amarelos são arquivos de dados enquanto que os azuis são etapas de processamento.

Os arquivos BAM processados foram analisados em conjunto (multi-sample) para identificação de variantes utilizando a ferramenta HaplotypeCaller do GATK. A partir do arquivo VCF resultante dessa análise foram obtidos SNPs e indels de alta qualidade utilizando-se como filtros: QD < 2,0; FS > 20,0; MQ < 40,0; MQRankSum < -8,0; ReadPosRankSum < -8,0; para SNPs e QD < 2,0; FS > 200,0; ReadPosRankSum < -20,0 para indels. A regiões correspondentes aos indels de alta qualidade foram realinhadas utilizando-se a ferramenta IndelRealigner e os scores de qualidade de todas as bases do alinhamento foram recalibrados utilizando-se a ferramenta BQSR.

Assim como já descrito nesse documento no tópico “Material e métodos”, as siglas dos parâmetros de filtragem do GATK correspondem a: QD (QualByDepth) - confiança em uma variante dividido pela profundidade das amostras que divergem da referência; FS (FisherStrand) – valor de *phred* escalonado pelo p-valor de um teste Exato de Fisher feito para detectar algum viés gerado por variantes em *reads* com sentidos diferentes; MQ (MappingQuality) – raiz quadrada da qualidade de mapeamento dos *reads* em todas as amostras; e MQRankSum (MappingQualityRankSumTest) – teste de médias Mann-Whitney ranqueado aplicado às variantes heterozigotas (DePristo

et al., 2011).

Os arquivos BAM com scores recalibrados foram então submetidos novamente ao processo de identificação de variantes. Os SNPs e indels identificados foram novamente separados e filtrados segundo os critérios já explicitados e utilizados em uma segunda rodada de recalibração e posterior identificação de variantes, quando se observou a estabilização do número de polimorfismos identificados.

3.5 CARACTERIZAÇÃO DOS SNPs IDENTIFICADOS

A caracterização estrutural dos SNPs identificados foi realizada utilizando-se as ferramentas de avaliação do VCFtools (Danecek et al., 2011). Com essas ferramentas foram obtidos os SNPs que passaram nos filtros de qualidade, a taxa de transição/transversão e a identificação de quantos SNPs ocorrem em contexto gênico. Além disso, as ferramentas de seleção de variantes do VCFTools foram utilizadas em conjunto com um script escrito em R desenvolvido por Coelho (2015) para selecionar, dentre os SNPs finais, os que ocorrem isoladamente em um intervalo de 200pb e que apresentam a maior nota de qualidade do contig. Essa seleção foi realizada no intuito de fornecer um conjunto de SNPs recomendados para teste enquanto marcadores moleculares.

A anotação funcional, realizada utilizando-se como referência as predições dos genes no *draft assembly* (Ribeiro, 2015, dados não publicados), foi feita pelo software SnpEff (Cingolani et al., 2012). Os genes que específicos que foram selecionados após a anotação com o SnpEff foram anotados quando a função putativa utilizando a ferramenta blastn.

A anotação feita com o SnpEff foi realizada utilizando termos ontológicos (*Sequence Ontology terms*). Essa anotação é realizada por meio de efeitos putativos que os SNPs podem causar em genes codificadores de proteína. Variantes que afetam genes não-codificantes são anotados e a bio-tipo correspondente é identificado, sempre que a informação está disponível.

Os termos ontológicos utilizados na anotação, que estão relacionados ao efeito putativo dos alelos da variante naquele gene (e, portanto, não corresponde ao número exato de SNPs), são agrupados em impactos putativos baseado no quanto a modificação gerada pode ser deletéria. Esses impactos são: HIGH, MODERATE, LOW e MODIFIER, conforme o quadro abaixo:

Impact	Sequence Ontology Term
HIGH	chromosome_number_variation
HIGH	exon_loss_variant
HIGH	frameshift_variant
HIGH	rare_amino_acid_variant
HIGH	splice_acceptor_variant
HIGH	splice_donor_variant
HIGH	start_lost
HIGH	stop_gained
HIGH	stop_lost
HIGH	transcript_ablation
MODERATE	3_prime_UTR_truncation +exon_loss
MODERATE	5_prime_UTR_truncation +exon_loss_variant
MODERATE	coding_sequence_variant
MODERATE	disruptive_inframe_deletion
MODERATE	disruptive_inframe_insertion
MODERATE	inframe_deletion
MODERATE	inframe_insertion
MODERATE	missense_variant
MODERATE	regulatory_region_ablation
MODERATE	splice_region_variant
MODERATE	TFBS_ablation
LOW	5_prime_UTR_premature start_codon_gain_variant
LOW	initiator_codon_variant
LOW	splice_region_variant
LOW	start_retained
LOW	stop_retained_variant
LOW	synonymous_variant
MODIFIER	3_prime_UTR_variant
MODIFIER	5_prime_UTR_variant
MODIFIER	coding_sequence_variant
MODIFIER	conserved_intergenic_variant
MODIFIER	conserved_intron_variant
MODIFIER	downstream_gene_variant
MODIFIER	exon_variant
MODIFIER	feature_elongation
MODIFIER	feature_truncation
MODIFIER	gene_variant
MODIFIER	intergenic_region
MODIFIER	intragenic_variant
MODIFIER	intron_variant
MODIFIER	mature_miRNA_variant
MODIFIER	miRNA
MODIFIER	NMD_transcript_variant
MODIFIER	non_coding_transcript_exon_variant
MODIFIER	non_coding_transcript_variant
MODIFIER	regulatory_region_amplification
MODIFIER	regulatory_region_variant

MODIFIER	TF_binding_site_variant
MODIFIER	TFBS_amplification
MODIFIER	transcript_amplification
MODIFIER	transcript_variant
MODIFIER	upstream_gene_variant

Quadro 1. Lista de termos de anotação funcional putativa de variantes (Sequence Ontology Terms) e os respectivos grupos de impacto funcional. Disponível em: http://snpeff.sourceforge.net/VCFannotationformat_v1.0.pdf.

4 RESULTADOS E DISCUSSÃO

4.1 AVALIAÇÃO DO CONTROLE DE QUALIDADE E DO DRAFT-ASSEMBLY

Considerando-se as cinco bibliotecas sequenciadas, foi gerado um total de 8,64 Gb, distribuídos em 63.017.960 *reads* do tipo *paired-end*. Após o controle de qualidade, restaram 56.859.932 *reads*, totalizando 5,63 Gb. As figuras 4 e 5 mostram o efeito do controle de qualidade em uma das bibliotecas de sequenciamento. A figura 4 apresenta a distribuição das notas de qualidade das bases para cada posição nos reads, enquanto que a figura 5 apresenta a proporção de cada base para cada posição nos reads, antes e depois do controle de qualidade, respectivamente. É possível se observar nessas figuras o efeito do controle de qualidade, resultando em *reads* com bases que apresentam valores de *phred* superiores a 15 e sem evidências empíricas de sequências contaminantes, como a presença de adaptadores de sequenciamento.

Com esse esforço de sequenciamento foram montados (livres de sequências de organelas) 130.243 *contigs,* que juntos somam 250,5 Mb, com uma cobertura média de ~10X. Além da montagem, o programa dipSPAdes forneceu uma estimativa para o tamanho do genoma de *E. dysenterica* de 262,1Mb. Considerando-se este valor, foram recuperados 95,57% do genoma de *E. dysenterica* no *draft assembly* produzido (Tabela 1).

A estimativa do tamanho de genoma obtida é inferior às encontradas utilizando métodos citométricos para espécies de outros gêneros da família Myrtaceae como o Eucalipto (640 Mb) (Grattapaglia & Bradshaw, 1994) e a goiaba (579 Mb) (Bennett & Leitch, 1997). No entanto, quando se trata das estimativas obtidas para outras espécies do mesmo gênero da cagaiteira (*Eugenia*) o valor obtido encontra-se próximo aos observados até então, como, por exemplo: *E. crenata* (238 Mb), *E. hyemalis* (247 Mb), *E. involucrata* (244 Mb), *E. multicostata* (311 Mb) e *E. uniflora* (244 MB) (Costa et al., 2008).

O valor estimado do tamanho do genoma de *E. dysenterica* obtido *in silico*,

ainda que pouco preciso por ser dependente da relação entre esforço de sequenciamento, cobertura dos *k-mers* e profundidade média, segue um padrão esperado para a família Myrtaceae. Essa família possui espécies com genomas que podem ser considerados de pequenos a médios em comparação com outras angiospermas. Soltis et al. (2003) argumentam que grupos basais dentre as angiospermas tendem a ter genomas menores em comparação com os grupos derivados. Isso sugere que a estimativa obtida *in silico* para o tamanho do genoma de *E. dysenterica* parece ser uma aproximação razoável, estando o seu valor real, provavelmente, próximo ao estimado.

Tabela 1. Estatísticas descritivas do *draft assembly* obtido para o genoma de *E. dysenterica*.

Métrica	Valor obtido
Total de *contigs*	130.268
Total de *contigs* após remoção de sequências organelares	130.243
Comprimento total do *assembly*	250.609.565
Comprimento total do *assembly* após remoção de sequências organelares	250.565.843
Cobertura estimada do genoma total	95,57%
Comprimento do maior *contig*	58.428
Número de contigs maiores que 10 Kb	861
Comprimento médio dos *contigs*	1.924
N50	2.658
Proporção estimada de elementos repetitivos	35,40%

Ainda que o *draft assembly* obtido tenha conseguido recuperar 95,57% do tamanho estimado para o genoma de *E. dysenterica*, ele se encontra, no entanto, ainda muito fragmentado em um grande número de *contigs* quando comparado a outros rascunhos de genomas de plantas publicados recentemente (Peng et al., 2014; Dorn et al., 2015; Nowak et al., 2015; Parween et al., 2015). Isso se deve à aplicação nestes trabalhos de um maior esforço de sequenciamento, da utilização de dados de diferentes plataformas de sequenciamento, além do uso de informações complementares provenientes da análise de outros recursos genômicos como o sequenciamento de BACs, de ESTs e de RNA-Seq. Como neste trabalho o principal objetivo foi o de se identificar e caracterizar SNPs (e não o de se obter uma sequência de referência para o

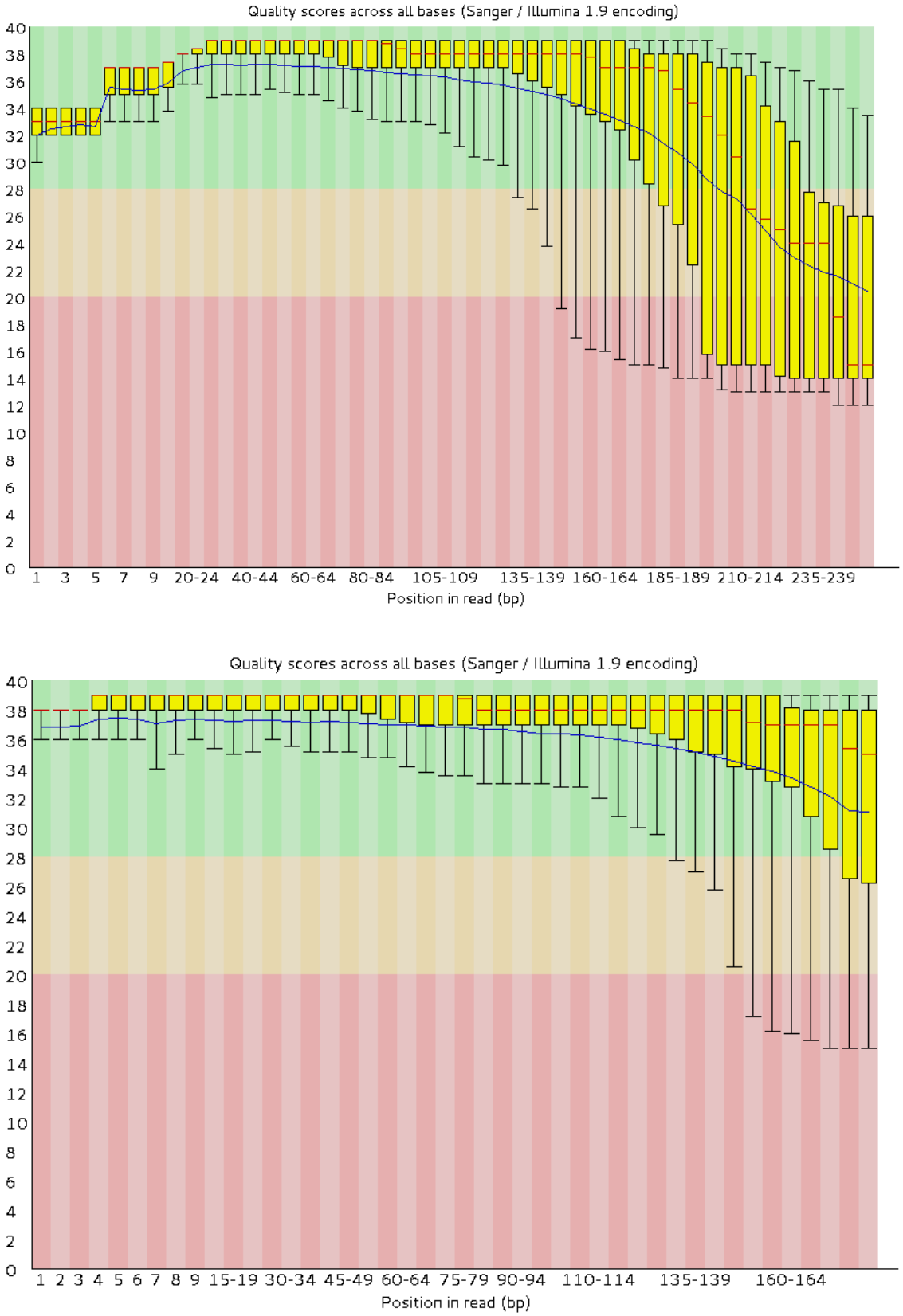

Figura 4. Distribuições das qualidades das bases sequenciadas em uma das bibliotecas de *Eugenia dysenterica* antes e depois do controle de qualidade. No eixo "x" estão as notas de qualidade *phred* e no eixo "y" a posição das bases do *read*.

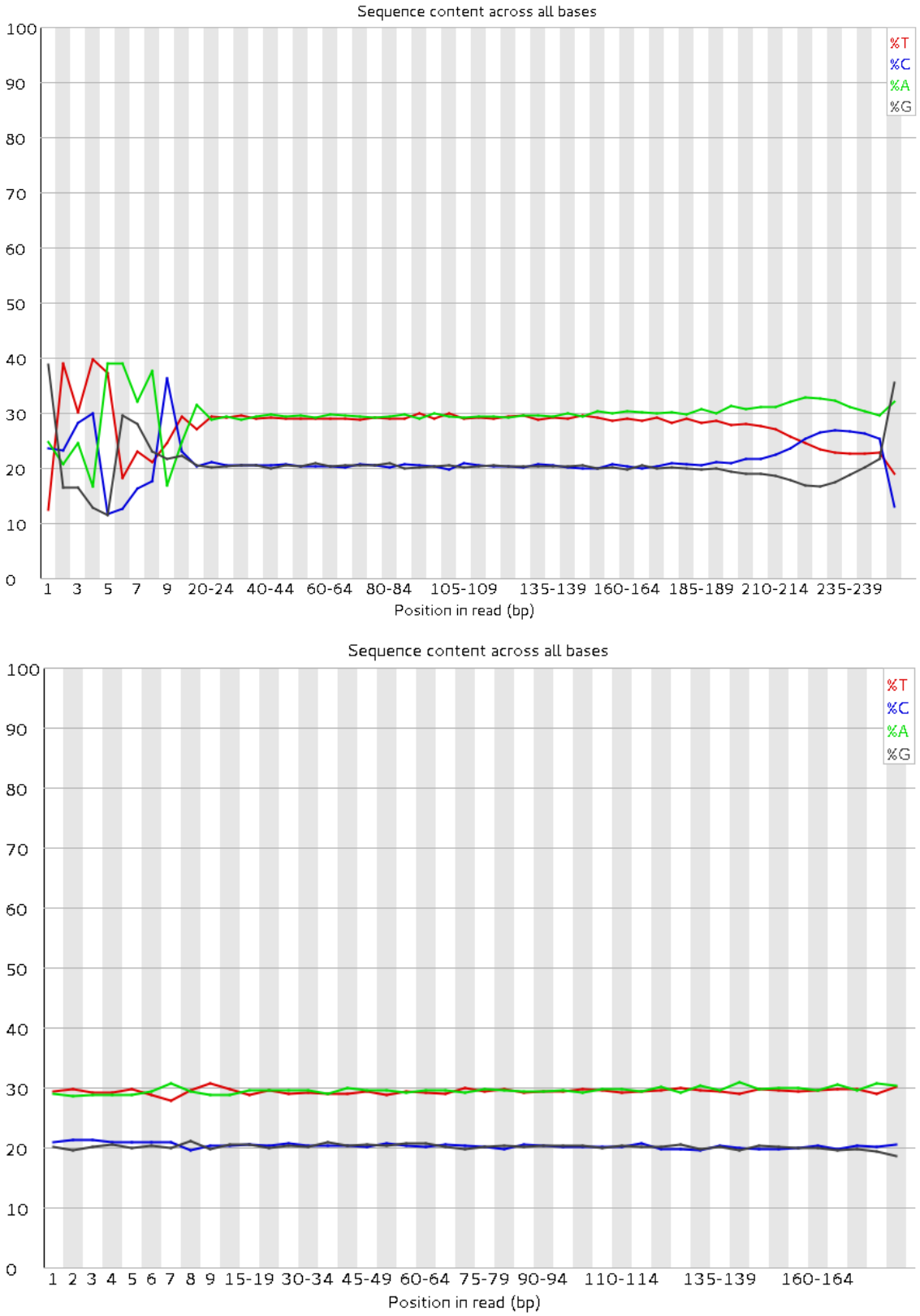

Figura 5. Distribuições do conteúdo de cada base nas sequências de uma das bibliotecas de *Eugenia dysenterica* antes e depois do controle de qualidade. No eixo "x" estão as proporções e no eixo "y" a posição das bases do *read*.

genoma de *E. dysenterica*), o *draft assembly* obtido mostrou-se satisfatório em termos de amostragem do genoma para este fim.

Outros fatores também podem estar relacionados a essa alta fragmentação além da necessidade de aplicação de outros métodos para fechar os *gaps* do genoma. Dentre eles, uma com provável grande influência, é a quantidade de elementos repetitivos no genoma, que dificultam a alocação de sequências conectam outras duas. A busca por elementos repetitivos no draft-assembly obtido revelou uma taxa de 35,40% desses elementos, o que pode ser considerada alta.

4.2 IDENTIFICAÇÃO E CARACTERIZAÇÃO DAS VARIANTES

Uma primeira chamada de variantes foi realizada utilizando-se os *reads* alinhados que passaram pelo controle de qualidade de mapeamento. As variantes que passaram pelos filtros de alta qualidade utilizados foram consideradas como referência, conforme sugestão do manual de melhores práticas do GATK. Nessa primeira de identificação de variantes foram identificados 1.211.226 SNPs e 123.139 indels que, após aplicação dos filtros de qualidade, resultaram em 1.207.591 SNPs e 122.832 indels (Tabelas 2). Um pequeno número de variantes foi eliminado pela aplicação dos filtros de qualidade. Isso pode ser explicado pelo fato de ter sido usado o parâmetro “stand_call_conf” (nota de qualidade mínima para as variantes. Variantes que apresentem notas inferiores ao valor estipulado são marcadas no arquivo vcf como variantes de baixa qualidade (LowQual)) como igual a 30, valor de elevada qualidade, já durante a primeira etapa de identificação de variantes.

Após a primeira etapa de recalibração dos scores de qualidade de sequenciamento e segunda filtragem de qualidade, foram identificados 1.032.676 SNPs e 101.968 indels. Após a segunda etapa de recalibração e terceira filtragem de qualidade, foram identificados 999.016 SNPs e 98.576 indels. A menor queda no número de variantes após a segunda recalibração indica uma tendência de estabilização no número de variantes identificados. Essas variantes foram as consideradas válidas para fins de análises posteriores.

Até o desenvolvimento desse trabalho, não há na literatura nenhuma caracterização de SNPs no genoma de *E. dysenterica.* Nesse trabalho foram identificados pela primeira vez 999.016 SNPs no genoma dessa espécie, sendo 650.554 transições e 348.462 transversões, fornecendo uma estimativa para a razão

transição/transversão (Ts/Tv) de 1,86. Foi observada uma densidade SNPs de um SNP a cada ~251 bases. Nível de polimorfismo que pode ser considerado alto. A razão Ts/Tv é um forte indicativo da qualidade da chamada de variantes, no entanto esse valor parece variar dependendo das condições ambientais em que os organismos estão expostos (Jiang et al., 2014) e o contexto de ocorrência da variação, uma vez que a distribuição dessas mutações não é uniforme ao longo do genoma (Cao et al., 2011).

Tabela 2. Número de SNPs identificados no genoma de *E. dysenterica* em cada uma das etapas do *pipeline* proposto.

Etapa do pipeline	Nº de SNPs	Ts/Tv	Nº de *indels*
Primeira chamada de variantes	1.211.226	1,89	123.139
Primeira filtragem	1.207.591	1,89	122.832
Segunda chamada de variantes (após primeira recalibração)	1.036.384	1,87	102.229
Segunda filtragem	1.032.676	1,86	101.968
Terceira chamada de variantes (após segunda recalibração)	1.002.732	1,86	98.841
Terceira filtragem	999.016	1,86	98.576
Seleção de regiões com um único SNP em 200pb	251.284	1,73	-
Seleção de um único SNP por *contig*	72.633	1,72	-

Observa-se na Figura 6 que a grande maioria dos SNPs (919.142), identificados na primeira chamada de variantes, sobreviveram às duas etapas de recalibração de qualidade das bases e aos filtros de alta qualidade aplicados. Também é possível se observar uma redução no número de SNPs retidos em cada etapa, sendo o menor valor, o do terceiro filtro. Esse resultado sugere uma melhoria gradativa na consistência de chamada de SNPs, como consequência da aplicação dos filtros e do processo de recalibração dos valores de qualidade das bases.

Um total de 597.330 SNPs, correspondendo a 59,79% dos SNPs identificados, ocorrem em contexto de algum dos genes preditos pelo software Augustus. Essa alta quantidade de SNPs relacionados a genes possui uma relação direta com a quantidade de genes preditos nesse *draft assembly*, 60.171 (Ribeiro, comunicação

pessoal). Com essa quantidade de genes preditos, observa-se na montagem utilizada uma densidade de um gene a cada 9.256 bases. Denton et al. (2014) argumentam que pode haver uma superestimativa na quantidade de genes preditos utilizando *draft assemblies*, uma vez que um único gene fragmentado em dois *contigs* diferentes pode ser predito como dois genes separadamente.

A análise de anotação funcional dos SNPs realizada utilizando o *snpEff* gerou um relatório dos efeitos putativos causados por eles no genoma de *E. dysenterica.* Quanto ao tipo de impacto gerado por esses efeitos, a grande maioria (4.282.991, ou 77,30%) apresentaram uma anotação de efeito MODIFIER. Esse tipo de impacto está associado à variação na estrutura dos transcritos que não é possível de ser enquadrada nos demais tipos de impacto (Tabela 3).

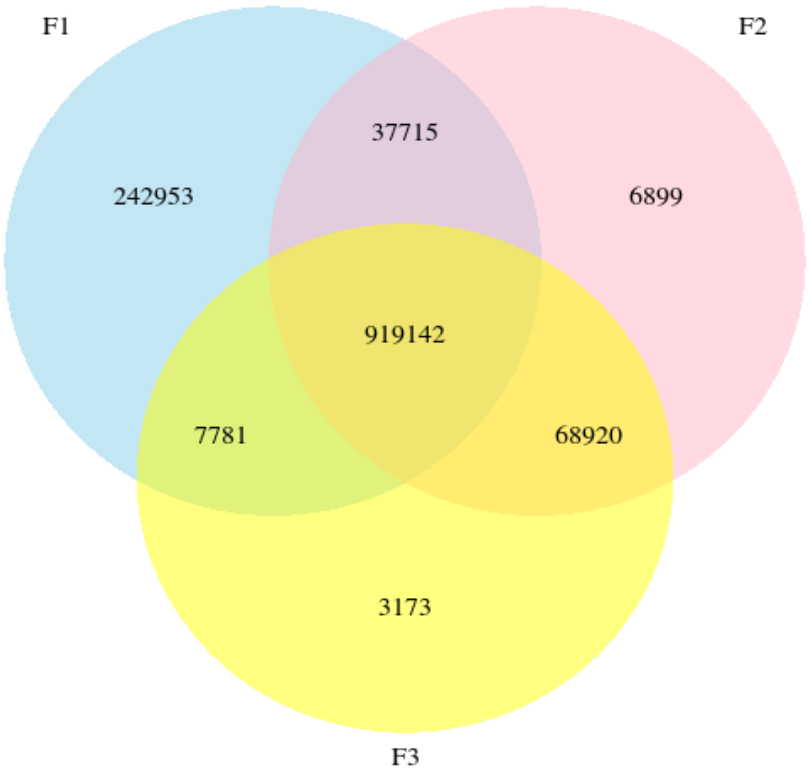

Figura 6. Diagrama de Venn demonstrando o número de SNPs identificados no *draft assembly* do genoma de *E. dysenterica* após cada etapa de filtragem. F1 – filtragem após a primeira chamada de variantes; F2 – filtragem após a primeira recalibração dos *scores* de qualidade das bases e F3 – filtragem após a segunda recalibração dos *scores* de qualidade das bases.

Tabela 3. Número de efeitos por tipo de impacto identificados na anotação funcional de SNPs no genoma de *Eugenia dysenterica.*

Tipo de impacto	Efeitos	Porcentagem
HIGH	41.090	0,74%
LOW	561.350	10,13%
MODERATE	654.795	11,82%
MODIFIER	4.282.991	77,30%

Ainda quanto ao tipo de impacto, pode-se observar que somente uma pequena porção dos efeitos (0,7%) foi enquadrada no nível HIGH, ou seja, que apresentam um alto impacto. Esse resultado era esperado uma vez que os tipos de mudanças que são agrupadas nessa categoria geralmente estão sob forte pressão seletiva. São enquadrados nesse tipo de impacto, termos ontológicos (*Sequence Ontology terms*) relacionados à variação no número de cromossomos, perda um éxon e variações nos sítios de *splicing*, por exemplo.

Quanto à classe funcional em que os efeitos se inserem, a minoria deles se enquadra como *nonsense* (mutações que resultam na criação de um *stop codon* prematuro), ou mutações sem sentido (1,83%). Esses resultados se justificam uma vez que a ocorrência de *stop codon* tem uma maior chance de intervir negativamente no *fitness* do indivíduo e deixá-lo vulnerável à eliminação por seleção natural. Foi observada uma razão entre o número de efeitos *missense* (mutações que resultam na alteração na codificação para um aminoácido) e silenciosos de 1,42. Essa relação foi corroborada por (Fresnedo-Ramírez et al., 2013), onde o valor de 1,44 dessa taxa foi identificado na variedade de pêssego (*Prunus persica* (L.) Batsch), Georgia Belle.

Tabela 4. Número de efeitos por classe funcional identificados na anotação funcional de SNPs no genoma de *Eugenia dysenterica.*

Classe funcional	Efeitos	Porcentagem
MISSENSE	659.223	57,70%
NONSENSE	20.934	1,83%
SILENT	462.229	40,46%

Como as mutações *missense* alteram o aminoácido que será codificado, foi realizada uma análise utilizando a ferramenta blastn, utilizando os primeiros transcritos

dos dez genes que mais apresentaram efeitos dessa classe funcional. Quatro dos dez genes identificados apresentaram um *top hit* para genes similares aos de *Arabidopsis thaliana* relacionados à resistência a doenças. A variação nesse tipo de gene é esperada uma vez que a planta precisa reagir com algum tipo de resposta a hospedeiros e potenciais causadores de doenças. Essa resposta pode vir na forma de uma variedade de metabólitos que inibem a presença do patógeno (Hammond-Kosack & Jones, 1996), o que justifica a necessidade de variação nesse tipo de gene.

O gene predito no genoma de cagaiteira (*g46806*) apresentou *top hit* para um gene de *Eucalyptus grandis* que foi anotado automaticamente como similar ao gene *At4g27190* de *Arabidopsis thaliana.* Segundo o banco de dados TAIR, esse gene tem sido proposto como participante de resposta defensiva e apoptose. As proteínas geradas por esse gene apresentam atividades de ligação de ATP e por se ligarem a outras proteínas. O gene predito *g39932* também apresentou *top hit* para esse mesmo gene, mostrando também uma relação com a resposta a patógenos.

Outro gene predito, *g44286*, apresentou relação com um gene de resistência a doenças. A busca no banco de dados TAIR indicou que o gene de *A. thaliana* que apresentou relação com esse gene predito também possui atividade de apoptose, sendo expresso principalmente no citosol e na membrana plasmática de células-guarda. O gene *g46934* também parece estar relacionado com esse tipo de função estando, porém, localizado no núcleo celular.

Um resultado interessante foi obtido para o gene *g21367*. O *best hit* na análise com o blast para essa sequência foi com o gene de *Eucalyptus grandis* "putative calcium-transporting ATPase 13". Esse gene parece ter função relacionada à catálise do transporte de íons Ca^{2+} por meio da membrana plasmática. Esse resultado sugere uma possível atuação da seleção natural sobre o polimorfismo deste loco, o que é plausível uma vez que os solos sob Cerrado tendem a ser ácidos e ricos em alumínio, substância tóxica para as plantas, que é neutralizada pela presença de cálcio, sendo este componente geralmente encontrado em baixas concentrações nos solos sob Cerrado (Haridasan, 2000).

Outros dois genes (*g45866* e *g18269*) apresentaram similaridade com genes com funções relacionadas ao transporte de substâncias através da membrana celular. Um resultado menos esperado é do gene *g34227* que está relacionado a uma subunidade da DNA-polimerase Zeta. Essa enzima ainda é considerada pouco estudada, mas parece possuir atividade de reparo aos danos causados por radiação (principalmente UV-B).

Mutantes para este gene em *Arabidopsis* são mais sensíveis a esse tipo de radiação. No banco de dados TAIR existem 92 entradas de polimorfismos para esse gene (https://www.arabidopsis.org/servlets/TairObject?type=locus&name=AT1G67500).

Quanto ao número de efeitos por tipo (*Sequence Ontology Term*) destacaram-se as variações de *introns*, seguidas pelas variações na região *upstream* dos genes. Efeitos relacionados ao *splicing* alternativo e à perda de alguma estrutura foram os tipos que apresentaram menor ocorrência. Dentre as estruturas gênicas, como esperado, os *introns* foram as regiões que apresentaram uma maior ocorrência de SNPs, seguidas dos *exons* e das regiões *up* e *downstream* (5 Kb das regiões UTR).

Tabela 5. Resultado da análise de similaridade usando o blastn dos dez genes com maior número de efeitos missense de SNPs identificados no genoma de *Eugenia dysenterica.*

Gene Edy	Blast Hit	Nº de efeitos	E-value	Tamanho da query	% query cover	% identidade	Relação
g46806	*Eucalyptus grandis* disease resistance protein At4g27190-like	150	0	2325	88	83	Resistência a doenças
g44286	*Eucalyptus grandis* probable disease resistance protein At1g61300	142	0	2912	91	76	Resistência a doenças
g21367	*Eucalyptus grandis* putative calcium-transporting ATPase 13, plasma membrane-type	133	0	3921	74	85	Transporte de íons cálcio
g46934	*Eucalyptus grandis* probable disease resistance protein At4g27220	116	0	1265	92	78	Resistência a doenças
g45866	*Eucalyptus grandis* putative ABC transporter C family member 15	111	0	3775	82	87	Transporte de substâncias transmembrana
g37654	*Eucalyptus grandis* microtubule-associated protein futsch	103	0	4799	97	88	Regulação positiva da produção de microtubulina
g39932	*Eucalyptus grandis* disease resistance protein At4g27190-like	102	0	3423	67	90	Resistência a doenças
g34227	*Eucalyptus grandis* DNA polymerase zeta catalytic subunit	92	0	4422	93	90	Tolerância à radiação UV-B
g15962	*Eucalyptus grandis* uncharacterized LOC104455587	89	0	2142	93	88	-
g18269	*Eucalyptus grandis* probable LRR receptor-like serine/threonine-protein kinase At3g47570	88	0	2878	96	87	Receptor de proteína tirosina quinase transmembrana

Tabela 6. Número de efeitos por tipo identificados na anotação funcional de SNPs no genoma de *Eugenia dysenterica.*

Tipo de efeito	Efeitos	Porcentagem
3_prime_UTR_variant	298.609	5,39
5_prime_UTR_premature_start_codon_gain_variant	17.564	0,31
5_prime_UTR_variant	103.832	1,87
downstream_gene_variant	1.042.977	18,82
initiator_codon_variant	198	0
initiator_codon_variant+non_canonical_start_codon	5	0
intergenic_region	322.497	5,82
intron_variant	1.739.659	31,40
missense_variant	636.502	11,49
missense_variant+splice_region_variant	18.293	0,33
splice_acceptor_variant+intron_variant	7.334	0,13
splice_donor_variant+intron_variant	8.590	0,15
splice_region_variant	3.283	0,06
splice_region_variant+initiator_codon_variant	2	0
splice_region_variant+intron_variant	78.078	1,40
splice_region_variant+stop_retained_variant	14	0
splice_region_variant+synonymous_variant	11.427	0,20
start_lost	1.755	0,03
start_lost+splice_region_variant	17	0
stop_gained	19.936	0,37
stop_gained+splice_region_variant	998	0,01
stop_lost	2.426	0,04
stop_lost+splice_region_variant	34	0
stop_retained_variant	709	0,01
synonymous_variant	450.070	8,12
upstream_gene_variant	775.417	14

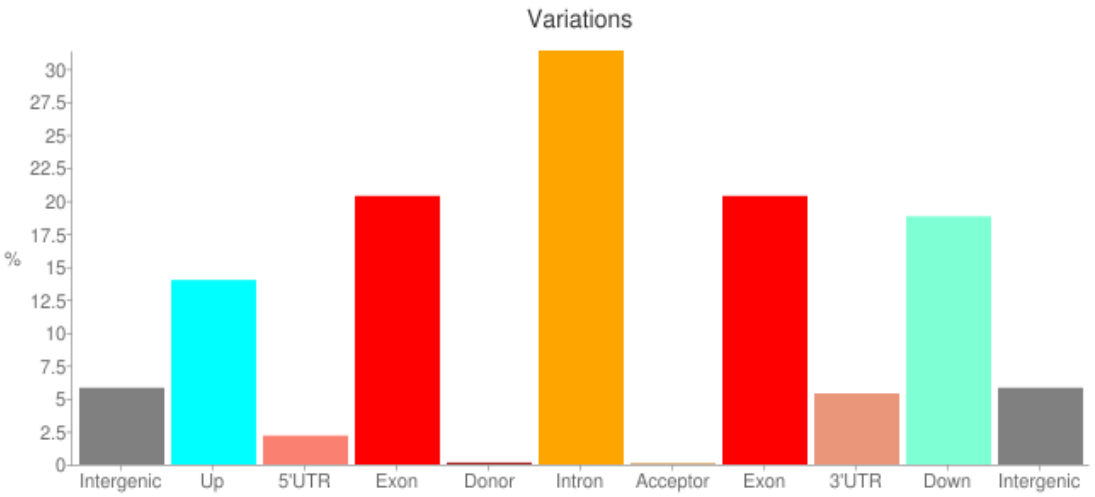

Figura 7. Distribuição das proporções dos efeitos de SNPs identificados no genoma de *Eugenia dysenterica* baseada na estrutura gênica em que esses efeitos ocorrem.

5 CONCLUSÕES

Com os resultados obtidos nesse trabalho pode-se concluir que:

- O genoma de *Eugenia dysenterica* é rico em variantes do tipo SNPs e *indels*;
- Cerca de 60% das variantes identificadas parecem estar relacionadas a regiões gênicas;
- Esse trabalho fornece um *pipeline* otimizado para a identificação de variantes para uma espécie que não se tem um genoma de referência disponível, como é o caso da cagaiteira.
- Esse trabalho fornece subsídios para o desenvolvimento de marcadores moleculares para aplicação em estudos de genômica de populações de *E. dysenterica*. Aqui são recomendados um conjunto de 72.633 variantes candidatas a comporem o primeiro painel de SNPs para essa espécie.

6 REFERÊNCIAS

AGUIAR, A. V. DE; MOURA, N. F.; MOURA, M. F.; et al. Relação entre a variação genética de caracteres quantitativos e marcadores moleculares em subpopulações de cagaiteira (*Eugenia dysenterica* DC). **Revista Brasileira de Fruticultura**, v. 33, n. 1, p. 157–169, 2011.

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

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.

ALTSCHUL, S. F.; GISH, W.; PENNSYLVANIA, T.; PARK, U. Basic Local Alignment Search Tool 2Department of Computer Science. **Journal of Molecular Biology**, London, v. 215, p. 403–410, 1990.

ANDREWS S. (2010). FastQC a Quality Control Tool for High Throughput Sequence Data [Online]. Disponível em: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

BAIRD, N. A.; ETTER, P. D.; ATWOOD, T. S.; CURREY, M. C.; SHIVER, A. L.; LEWIS, Z. A.; SELKER, E. U.; CRESKO, W. A.; JOHNSON, E. A. Rapid SNP discovery and genetic mapping using sequenced RAD markers. **PloS one**, San Francisco, v. 3, n. 10, p. 1-7, 2008

BARBOSA, A. C. D. O. F. et al. Range-wide genetic differentiation of Eugenia dysenterica (Myrtaceae) populations in Brazilian Cerrado. **Biochemical Systematics and Ecology**, v. 59, p. 288–296, 2015.

BARCHI, L.; LANTERI, S.; PORTIS, E.; ACQUADRO, A.; VALÈ, G.; TOPPINO, L.,; ROTINO, G. L. Identification of SNP and SSR markers in eggplant using RAD tag sequencing. **BMC genomics**, London, v. 12, n. 1, p. 304, 2011.

BARZON, L.; LAVEZZO, E.; COSTANZI, G.; FRANCHIN, E.; TOPPO, S.; PALÙ, G. Next-generation sequencing technologies in diagnostic virology. **Journal of Clinical Virology**, Amsterdam v. 58, n. 2, p. 346-350, 2013.

BENNETT, M. D.; LEITCH, I. J. Nuclear DNA amounts in angiosperms - 583 new estimates. **Annals of Botany**, Oxford, v. 80, p. 169-196, 1997.

BERNATZKY, R. & S. D. TANKSLEY. Toward a saturated linkage map in tomato

based on isozymes and random cDNA sequences. **Genetics**, v.112, n.4, p.887-98. 1986.

BOAVENTURA-NOVAES, C. R. D. **VARIAÇÃO GENÉTICA QUANTITATIVA E MOLECULAR EM POPULAÇÕES NATURAIS DE Eugenia dysenterica DC.**, Universidade Federal de Goiás, 2014

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

BOTSTEIN, D.; WHITE, R. L.; SKOLNICK, M.; DAVIS, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. **American journal of human genetics**, v. 32, n. 3, p. 314–31, 1980.

CAO, J.; SCHNEEBERGER, K.; OSSOWSKI, S.; et al. Whole-genome sequencing of multiple Arabidopsis thaliana populations. **Nature Genetics**, v. 43, n. 10, p. 956–963, 2011.

CARVALHO, M. C. C. G.; SILVA, D. C. G. Sequenciamento de DNA de nova geração e suas aplicações na genômica de plantas; Next generation DNA sequencing and its applications in plant genomics. **Ciência rural**, Santa Maria, v. 40, n. 3, p. 735-744, 2010.

CHAVES, L. J.; TELLES, M. P. D. C. Cagaita. In: VIEIRA, R. F. et al. (Eds.). **Frutas Nativas da Região Centro-Oeste do Brasil.** 1. ed. Brasília, DF: Embrapa Recursos Genéticos e Biotecnologia, 2006. p. 119–134.

CHAVES, L. J.; VENCOVSKY, R.; SILVA, R. S. M.; et al. Estimating inbreeding depression in natural plant populations using quantitative and molecular data. **Conservation Genetics**, v. 12, n. 2, p. 569–576, 2011.

CHEN, C.; MITCHELL, S. E.; ELSHIRE, R. J.; BUCKLER, E. S.; EL-KASSABY, Y. A. Mining conifers mega-genome using rapid and efficient multiplexed high-throughput genotyping-by-sequencing (GBS) SNP discovery platform. **Tree Genetics & Genomes**, Heidelberg, v. 9, n. 6, p. 1537-1544, 2013.

CINGOLANI, P.; PLATTS, A.; WANG, L. L.; et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w 1118; iso-2; iso-3. **Fly**, v. 6, n. 2, p. 80–92, 2012.

COSTA, I. R.; DORNELAS, M. C.; FORNI-MARTINS, E. R. Nuclear genome size variation in fleshy-fruited Neotropical Myrtaceae. **Plant Systematics and Evolution**, v. 276, n. 3-4, p. 209–217, 2008.

CROSSA, J.; BEYENE, Y.; KASSA, S.; PÉREZ, P.; HICKEY, J. M.; CHEN, C.; CAMPOS, G. L.; BURGUEÑO, J.; WINDHAUSEN, V. S.; BUCKLER, E. Genomic prediction in maize breeding populations with genotyping-by-sequencing. **G3: Genes, Genomes, Genetics**, Bethesda, v. 3, n. 11, p. 1903-1926, 2013.

DANECEK, P.; AUTON, A.; ABECASIS, G.; et al. The variant call format and

VCFtools. **Bioinformatics**, v. 27, n. 15, p. 2156–2158, 2011.

DAVEY, J. W.; BLAXTER, M. L. RAD-Seq: next-generation population genetics. **Briefings in Functional Genomics**, Oxford, v. 9, n. 5-6, p. 416-423, 2010.

DAVEY, J. W.; HOHENLOHE, P. A.; ETTER, P. D.; BOONE, J. Q.; CATCHEN, J. M.; BLAXTER, M. L. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. **Nature Reviews Genetics**, London, v. 12, n. 7, p. 499-510, 2011.

DENTON, J. F.; LUGO-MARTINEZ, J.; TUCKER, A. E.; SCHRIDER, R. D.; WARREN, W. C.; HAHN, M. W. Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies. **PLoS computational biology**, San Francisco, v. 10, n. 12, p. 1-9, 2014.

DEPRISTO, M. A; BANKS, E.; POPLIN, R.; et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. **Nature genetics**, v. 43, n. 5, p. 491–498, 2011.

DORN, K. M.; FANKHAUSER, J. D.; WYSE, D. L.; MARKS, M. D. A draft genome of field pennycress (Thlaspi arvense) provides tools for the domestication of a new winter biofuel crop. **DNA Research**, v. 22, n. 2, p. 121–131, 2015.

DOYLE, J. J. ; DOYLE, J. L. Isolation of plant DNA from fresh tissue. Focus, v.12, p.13-15, 1990.

EID, J.; FEHR, A.; GRAY, J.; et al. Real-time DNA sequencing from single polymerase molecules. **Science**, New York, v. 323, n. 5910, p. 133–138, 2009.

ELSHIRE, R. J.; GLAUBITZ, J. C.; SUN, Q.; POLAND, J. A.; KAWAMOTO, K.; BUCKLER, E. S.; MITCHELL, S. E. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. **PloS one**, San Francisco, v. 6, n. 5, p. e19379, 2011.

FERREIRA, M. E. & D. GRATTAPAGLIA. Introdução ao uso de marcadores moleculares em análise genética. Brasília: Embrapa. 1998. 220 p.

FRESNEDO-RAMÍREZ, J.; MARTÍNEZ-GARCÍA, P. J.; PARFITT, D. E.; CRISOSTO, C. H.; GRADZIEL, T. M. Heterogeneity in the entire genome for three genotypes of peach [Prunus persica (L.) Batsch] as distinguished from sequence analysis of genomic variants. **BMC Genomics**, v. 14, n. 1, p. 750, 2013.

GARRISON, E.; MARTH, G. Haplotype-based variant detection from short-read sequencing. p. 9, 2012.

GLENN, T. C. Field guide to next-generation DNA sequencers. **Molecular Ecology Resources**, Oxford, v. 11, n. 5, p. 759–769, 2011.

GRATTAPAGLIA, D.; BRADSHAW, H. D. Nuclear DNA content of commercially

important Eucalyptus species and hybrids. **Canadian journal of forest research**, Ottawa, v. 24, p.1074–1078, 1994.

HAMMOND-KOSACK, K. E.; JONES, J. D. Resistance gene-dependent plant defense responses. **The Plant cell**, v. 8, n. 10, p. 1773–1791, 1996.

HARIDASAN, M. Nutrição mineral de plantas nativas do cerrado. **Revista Brasileira de Fisiologia Vegetal**, v. 12, n. 2, p. 54–64, 2000.

HELENTJARIS, T.; WEBER, D. F.; WRIGHT, S. Use of monosomics to map cloned DNA fragments in maize. **Proceedings of the National Academy of Sciences of the United States of America**, v. 83, n. 16, p. 6035–9, 1986.

HIPP, A. L.; EATON, D. A.; CAVENDER-BARES, J.; FITZEK, E.; NIPPER, R.; MANOS, P. S. A framework phylogeny of the American oak clade based on sequenced RAD data. **PloS one**, San Francisco, v. 9, n. 4, p. 1-12, 2014.

HUTTNER, E.; WENZL, P.; AKBARI, M.; CAIG, V.; CARLING, J.; CAYLA, C.; EVERS, M.; JACCOUD, D.; PENG, K.; PATARAPUWADOL, S.; USZYNSKI, G.; XIA, L.; YANG, S.; KILIAN, A. **Diversity Arrays Technology**: A novel tool for harnessing the genetic potential of orphan crops. Conference of The World Biological Forum Discovery to Delivery: BioVisionAlexandria: CABI Publishing, 2004.

ILLUMINA, Agrigenomics Genotyping Decisions Reach a Crossroads, 2013. Disponível em: <http://res.illumina.com/documents/products/appspotlights/app_spotlight_ngg_ag.pdf>. Acesso em: 13 jun. 2014.

JIANG, C.; MITHANI, A.; BELFIELD, E. J.; et al. Environmentally responsive genome-wide accumulation of de novo Arabidopsis thaliana mutations and epimutations. **Genome research**, v. 24, p. 1821–1829, 2014.

KUMAR, P.; GUPTA, V. K.; MISRA, A K.; MODI, D. R.; PANDEY, B. K. Potential of Molecular Markers in Plant Biotechnology. **Plant Biotechnology**, v. 2, n. 4, p. 141–162, 2009.

LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. **Nature methods**, v. 9, n. 4, p. 357–9, 2012.

LI, H.; DURBIN, R. Fast and accurate short read alignment with Burrows-Wheeler Transform. **Bioinformatics**, v.25, p. 1754-1760, 2009.

LITT, M.; LUTY, J. A. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. **American journal of human genetics**, v. 44, n. 3, p. 397–401, 1989.

LIU, X.; HAN, S.; WANG, Z.; GELERNTER, J.; YANG, B. Z. Variant Callers for Next-Generation Sequencing Data: A Comparison Study. **PLoS ONE**, v. 8, n. 9, p. 1–11, 2013.

LU, F.; LIPKA, A. E.; GLAUBITZ, J.; ELSHIRE, R.; CHERNEY, J. H.; CASLER, M. D.; BUCKLER, E. S.; COSTICH, D. E. Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. **PLoS genetics**, San Francisco, v. 9, n. 1, p. 1-14, 2013.

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

MARTINOTTO, C. et al. **Cagaiteira (Eugenia dysenterica DC.)**. BOLETIM TÉCNICO UNIVERSIDADE FEDERAL DE LAVRAS, v. 1, n. 78, p. 1–21, 2008.

MCKENNA, A.; HANNA, M.; BANKS, E.; et al. Measuring absorptive capacity. **Genome Research**, v. 20, p. 254–260, 2009.

METZKER, M. L. M. L. L. Emerging technologies in DNA sequencing. **Genome research**, New York, v. 15, n. 12, p. 1767–76, 2005.

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

MILLER, M. R.; DUNHAM, J. P.; AMORES, A.; CRESKO, W. A.; JOHNSON, E. A. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. **Genome research**, New York, v. 17, n. 2, p. 240-248, 2007.

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

NOWAK, M. D.; RUSSO, G.; SCHLAPBACH, R.; et al. The draft genome of Primula veris yields insights into the molecular basis of heterostyly. **Genome Biology**, v. 16, n. 1, p. 12, 2015.

NURK, Sergey et al. Assembling Genomes and Mini-metagenomes from Highly Chimeric Reads. In: Research in Computational Molecular Biology. **Proceedings of the 17th Annual International Conference**, RECOMB 2013, Beijing, China, April 7-10, 2013.

PARWEEN, S. et al. An advanced draft genome assembly of a desi type chickpea (Cicer arietinum L.). **Scientific reports**, v. 5, 2015.

PENG, Y.; LAI, Z.; LANE, T.; et al. De Novo Genome Assembly of the Economically Important Weed Horseweed Using Integrated Data from Multiple Sequencing Platforms. **Plant Physiology**, v. 166, n. 3, p. 1241–1254, 2014.

PEGADARAJU, V.; NIPPER, R.; HULKE, B.; QI, L.; SCHULTZ, Q. De novo sequencing of sunflower genome for SNP discovery using RAD (Restriction site Associated DNA) approach. **BMC genomics**, London, v. 14, n. 1, p. 556, 2013.

PROCHNIK, S.; MARRI, P. R.; DESANY, B.; RABINOWICZ, P. D.; KODIRA, C.; MOHIUDDIN, M.; RODRIGUEZ, F.; FAUQUET, C.; TOHME,J.; HARKINS, T.; ROKHSAR, D. S.; ROUNSLEY, S. The cassava genome: current progress, future directions. **Tropical plant biology**, New York, v. 5, n. 1, p. 88-94, 2012.

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.

RESENDE JR, M.; NEVES, L. G.; BALMANT, K. M.; DERVINIS, C.; KULLAN, A. R. K.; VANDYK, D.; MYBURG, A.; KIRST, M. RAPiD-Seq: A method for genome complexity reduction and polymorphism detection. **Plant and Animal Genome XX**, San Diego, 2012.

ROTHBERG, J. M.; LEAMON, J. H. The development and impact of 454 sequencing. **Nature Biotechnology**, New York, v. 26, n. 10, p. 1117-1124, 2008.

RODRIGUES, E. B. **SISTEMA REPRODUTIVO E FLUXO GÊNICO VIA PÓLEN EM UMA COLEÇÃO DE GERMOPLASMA DE Eugenia dysenterica DC.** [s.l.] Universidade Federal de Goiás, 2012.

SAFONOVA, Y.; BANKEVICH, A.; PEVZNER, P. A. dipSPAdes: Assembler for Highly Polymorphic Diploid Genomes. **Journal of Computational Biology**, v. 22, n. 6, p. 528–545, 2015.

SANGER, F.; NICKLEN, S.; COULSON, A. R. DNA sequencing with chain-terminating inhibitors. **Proceedings of the National Academy of Sciences of the United States of America**, Washington, v. 74, n. 12, p. 5463–7, 1977.

SANGER, F.; COULSON, A. R.; FRIEDMANN, T.; et al. The nucleotide sequence of bacteriophage φX174. **Journal of Molecular Biology**, London, v. 125, p. 225–246, 1978.

SANO, S. M. et al. Folhação, floração, frutificação e crescimento inicial da cagaiteira em Panaltina, DF. Pesquisa Agropecuária Brasileira, v. 30, n. 1, p. 5-14, 1995.

SANSALONI, C. P. **Desenvolvimento e aplicações de DArT (*Diversity Arrays Technology*) e genotipagem por sequenciamento (*Genotyping-by- Sequencing*) para análise genética em *Eucalyptus*.** 2012. 145 f. Tese (Doutorado em Ciências Biológicas, área de concentração Biologia Molecular) – Universidade de Brasília, Brasília, 2012.

SCAGLIONE, D.; ACQUADRO, A.; PORTIS, E.; TIRONE, M.; KNAPP, S. J.; LANTERI, S. RAD tag sequencing as a source of SNP markers in Cynara cardunculus L. **BMC Genomics**, London, v. 13, n. 1, p. 3, 2012.

SCHLÖTTERER, C. The evolution of molecular markers--just a matter of fashion? **Nature reviews. Genetics**, v. 5, n. 1, p. 63–69, 2004.

SHENDURE, J.; HANLEE, J. Next-generation DNA sequencing. **Nature biotechnology**, New York, v. 26, n. 10, p. 1135-1145, 2008.

SHENDURE, J.; JI, H. Next-generation DNA sequencing. **Nature biotechnology**, New York, v. 26, n. 10, p. 1135-1145, 2008.

SILVA, R. S. M.; CHAVES, L. J.; NAVES, R. V. Caracterização de frutos e árvores de cagaita (Eugenia dysenterica dc.) no sudeste do estado de Goiás, Brasil. **Revista Brasileira de Fruticultura**, v. 23, n. 2, p. 330–334, 2001.

SIMPSON, J. T.; WONG, K.; JACKMAN, S. D.; SCHEIN, J. E.; JONES, S. J. M. ABySS: A parallel assembler for short read sequence data. , p. 1117–1123, 2009.

SLAVOV, G. T.; NIPPER, R.; ROBSON, P.; FARRAR, K.; ALLISON, G. G.; BOSCH, M.; CLIFTON-BROWN, J.C.; DONNISON, I. S.; JENSEN, E. Genome-wide association studies and prediction of 17 traits related to phenology, biomass and cell wall composition in the energy grass Miscanthus sinensis. **New Phytologist**, London, v. 201, n. 4, p. 1227-1239, 2014.

SMIT, A. F. A.; HUBLEY, R. *RepeatModeler Open-1.0*. Disponível em: http://www.repeatmasker.org. 2015.

SOBRAL, M. et al. Myrtaceae in Lista de Espécies da Flora do Brasil. Jardim Botânico do Rio de Janeiro. Disponível em: <http://reflora.jbrj.gov.br/jabot/floradobrasil/FB10388>. Acesso em: 24 Abr. 2015.

SOLTIS, D. E.; SOLTIS, P. S.; BENNETT, M. D.; LEITCH, I. J. Evolution of genome size in the angiosperms. **American Journal of Botany**, v. 90, n. 11, p. 1596–1603, 2003.

SOUZA, E. R. B. DE et al. Fenologia de cagaiteira (Eugenia dysenterica DC.) no Estado de Goiás. **Revista Brasileira de Fruticultura**, v. 30, n. 4, p. 1009–1014, 2008.

TELLES, M. P. C. et al. Development and characterization of new microsatellites for Eugenia dysenterica DC (Myrtaceae). **Genetics and Molecular Research**, v. 12, n. 3, p. 3124–3127, 2013.

TELLES, M. P. D. C. et al. Autocorrelação espacial das freqüências alélicas em subpopulações de cagaiteira (Eugenia dysenterica DC., Myrtaceae) no sudeste de Goiás. **Revista Brasileira de Botânica**, v. 24, n. 2, p. 145–154, 2001a.

TELLES, M. P. D. C. et al. Divergência entre subpopulações de cagaiteira (Eugenia dysenterica) em resposta a padrões edáficos e distribuição espacial. **Pesquisa Agropecuaria Brasileira**, v. 36, n. 11, p. 1387–1394, 2001b.

TELLES, M. P. D. C. et al. Genetic diversity and population structure of Eugenia dysenterica DC. (“cagaiteira” –Myrtaceae) in Central Brazil: Spatial analysis and implications for conservation andmanagement. **Conservation Genetics**, v. 4, n. 1, p. 1156–1170, 2003.

TRINDADE, M. G.; CHAVES, L. J. Genetic structure of natural Eugenia dysenterica DC
(Myrtaceae) populations in northeastern Goiás, Brazil, accessed by morphological traits

and RAPD markers. **Genetics and Molecular Biology**, v. 28, n. 3, p. 407–413, 2005.

UITDEWILLIGEN, J. G.; WOLTERS, A. M. A.; BJORN, B.; BORM, T. J.; VISSER, R. G.; VAN ECK, H. J. A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. **PLoS One**, San Francisco, v. 8, n. 5, p. 1-14, 2013.

VARSHNEY, R. K.; NAYAK, S. N.; GREGORY D. MAY, G. D.; SCOTT A.; JACKSON, S. A. Next-generation sequencing technologies and their implications for crop genetics and breeding. **Cell Press**, Amsterdam, v. 29, n. 9, 2009.

WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-Seq: a revolutionary tool for transcriptomics. **Nature Reviews Genetics**, London, v. 10, n. 1, p. 57-63, 2009.

ZERBINO, D. R.; BIRNEY, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. **Genome research**, v. 18, n. 5, p. 821–9, 2008.

ZIMIN, A. V; MARÇAIS, G.; PUIU, D.; et al. The MaSuRCA genome assembler. **Bioinformatics (Oxford, England)**, v. 29, n. 21, p. 2669–77, 2013.

ZUCCHI, M. I. et al. Genetic structure and gene flow in Eugenia dysenterica DC in the Brazilian Cerrado utilizing SSR markers. **Genetics and Molecular Biology**, v. 26, n. 4, p. 449-457, 2003.

ZUCCHI, M. I. et al. Genetic structure and gene flow of Eugenia dysenterica natural populations. **Pesquisa Agropecuária Brasileira**, Brasília, v. 40, n. 10, p. 975-980, 2005.

7 ANEXO

```
#!/bin/bash
# Identificação de SNPs em um draft-assembly de Eugenia dysenterica (Cagaiteira)
# Feito por: Rhewter Nunes
# Baseado em: https://www.broadinstitute.org/gatk/guide/best-practices?bpm=DNAseq
# As linhas que começam com “#” são linhas de comentário.

# Versão 1.0 (utilizando um draft-assembly obtido com o programa dipSPAdes).


# ETAPA 1 - Controle de qualidade do draft-assembly:

# 1.1) Blast para retirar as sequências de cloroplasto do consensus_contigs.fasta :
blastn -query consensus_contigs_l500.fasta -db /usr/local/ncbi/db/chloroplast_RefSeq -
out genome_clean_chloro5.txt -outfmt "6 std qlen slen qcovs" -evalue 1e-16 -
perc_identity 95 -qcov_hsp_perc 85 -num_threads 10

# 1.2) Criando uma lista com os nomes dos contigs a serem retirados do genoma:
cut genome_clean_chloro5.txt -f 1 | uniq > contiglist_cleanchloro5.txt

# 1.3) Após separar um lista dos hits, usar essa ferramenta para gerar um arquivo fasta
só com as sequências desejadas:
pyfasta extract --fasta=consensus_contigs.fasta --header --exclude --file
contiglist_cleanchloro.txt > genome_clean_chloro.fasta

# 1.4) Blast para retirar as sequências de mitocôndria do genome_clean_clroro.fasta:
blastn -query genome_clean_chloro.fasta -db
/usr/local/ncbi/db/plant_mitochondria_RefSeq -out genome_clean_chloromito.txt -
outfmt "6 std qlen slen qcovs" -perc_identity 95 -qcov_hsp_perc 85 -evalue 1e-16 -
num_threads 15

# 1.5) Criando uma lista com os nomes dos contigs a serem retirados do genoma:
cut genome_clean_chloromito.txt -f 1 | uniq > contiglist_cleanchloromito.txt

# 1.6) Após separar um lista dos hits, usar essa ferramenta para gerar um arquivo fasta
só com as sequências desejadas:
pyfasta extract --fasta=genome_clean_chloro.fasta --header --exclude --file
contiglist_cleanchloromito.txt > genome_clean_chloromito.fasta

# 1.7) Mascarando as repetições usando o banco de dados de eudicotyledoneas:
RepeatMasker -e crossmatch -pa 20 -species eudicotyledons -a -inv -html
genome_clean_chloromito.fasta
```

1.8) Mascarando as demais repetições usando o banco de dados de repetições de cagaiteira (consensi.fa.lib) obtido de Ribeiro (2015):
RepeatMasker -e crossmatch -a -inv -x -lib /home/rhewter/Edysenterica/gatk/T7_SPA_con500/consensi.fa.lib -html -pa 16 /home/rhewter/Edysenterica/gatk/T7_SPA_con500/spades_clean.fasta

1.9) Alterando o nome do fasta com o genoma para ficar mais fácil:
mv genome_clean_chloromito.fasta.masked.masked Edy_clean.fasta

ETAPA 2 - Processamento dos BAMs:

Criando arquivos de índice:
bwa index Edy_clean.fasta
samtools faidx Edy_clean.fasta
java -jar /usr/local/picard-tools-1.135/picard.jar CreateSequenceDictionary REFERENCE=Edy_clean.fasta OUTPUT=Edy_clean.dict

Realizando os alinhamentos:
bwa mem -t 4 -M -R '@RG\tID:Edy1\tSM:Edy1\tPL:illumina\tLB:paired' Edy_clean.fasta Edy1_PE1_15.fq Edy1_PE2_15.fq > aln_Edy1_PE.sam
bwa mem -t 4 -M -R '@RG\tID:Edy2\tSM:Edy2\tPL:illumina\tLB:paired' Edy_clean.fasta Edy2_PE1_15.fq Edy2_PE2_15.fq > aln_Edy2_PE.sam
bwa mem -t 4 -M -R '@RG\tID:Edy3\tSM:Edy3\tPL:illumina\tLB:paired' Edy_clean.fasta Edy3_PE1_15.fq Edy3_PE2_15.fq > aln_Edy3_PE.sam
bwa mem -t 4 -M -R '@RG\tID:Edy4\tSM:Edy4\tPL:illumina\tLB:paired' Edy_clean.fasta Edy4_PE1_15.fq Edy4_PE2_15.fq > aln_Edy4_PE.sam
bwa mem -t 4 -M -R '@RG\tID:Edy5\tSM:Edy5\tPL:illumina\tLB:paired' Edy_clean.fasta Edy5_PE1_15.fq Edy5_PE2_15.fq > aln_Edy5_PE.sam

Converter SAM em BAM e ordenando:
java -jar /usr/local/picard-tools-1.135/picard.jar SortSam INPUT=aln_Edy1_PE.sam OUTPUT=Edy1_PE_sorted.bam SORT_ORDER=coordinate
java -jar /usr/local/picard-tools-1.135/picard.jar SortSam INPUT=aln_Edy2_PE.sam OUTPUT=Edy2_PE_sorted.bam SORT_ORDER=coordinate
java -jar /usr/local/picard-tools-1.135/picard.jar SortSam INPUT=aln_Edy3_PE.sam OUTPUT=Edy3_PE_sorted.bam SORT_ORDER=coordinate
java -jar /usr/local/picard-tools-1.135/picard.jar SortSam INPUT=aln_Edy4_PE.sam OUTPUT=Edy4_PE_sorted.bam SORT_ORDER=coordinate
java -jar /usr/local/picard-tools-1.135/picard.jar SortSam INPUT=aln_Edy5_PE.sam OUTPUT=Edy5_PE_sorted.bam SORT_ORDER=coordinate

Realizanda a marcação das reads duplicadas;
java -jar /usr/local/picard-tools-1.135/picard.jar MarkDuplicates INPUT=Edy1_PE_sorted.bam OUTPUT=Edy1_PE_sorted_nodup.bam METRICS_FILE=metrics1.txt MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=1000
java -jar /usr/local/picard-tools-1.135/picard.jar MarkDuplicates INPUT=Edy2_PE_sorted.bam OUTPUT=Edy2_PE_sorted_nodup.bam METRICS_FILE=metrics2.txt

```
MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=1000
java -jar /usr/local/picard-tools-1.135/picard.jar MarkDuplicates
INPUT=Edy3_PE_sorted.bam OUTPUT=Edy3_PE_sorted_nodup.bam
METRICS_FILE=metrics3.txt
MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=1000
java -jar /usr/local/picard-tools-1.135/picard.jar MarkDuplicates
INPUT=Edy4_PE_sorted.bam OUTPUT=Edy4_PE_sorted_nodup.bam
METRICS_FILE=metrics4.txt
MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=1000
java -jar /usr/local/picard-tools-1.135/picard.jar MarkDuplicates
INPUT=Edy5_PE_sorted.bam OUTPUT=Edy5_PE_sorted_nodup.bam
METRICS_FILE=metrics5.txt
MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=1000

# Gerando arquivo de índice para os BAMs ordenados e com as duplicatas marcadas:
java -jar /usr/local/picard-tools-1.135/picard.jar BuildBamIndex
INPUT=Edy1_PE_sorted_nodup.bam
java -jar /usr/local/picard-tools-1.135/picard.jar BuildBamIndex
INPUT=Edy2_PE_sorted_nodup.bam
java -jar /usr/local/picard-tools-1.135/picard.jar BuildBamIndex
INPUT=Edy3_PE_sorted_nodup.bam
java -jar /usr/local/picard-tools-1.135/picard.jar BuildBamIndex
INPUT=Edy4_PE_sorted_nodup.bam
java -jar /usr/local/picard-tools-1.135/picard.jar BuildBamIndex
INPUT=Edy5_PE_sorted_nodup.bam

# Contando o número de reads que foram marcadas como duplicatas:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar -T CountReads -R
Edy_clean.fasta -I Edy1_PE_sorted_nodup.bam -rf DuplicateRead
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar -T CountReads -R
Edy_clean.fasta -I Edy2_PE_sorted_nodup.bam -rf DuplicateRead
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar -T CountReads -R
Edy_clean.fasta -I Edy3_PE_sorted_nodup.bam -rf DuplicateRead
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar -T CountReads -R
Edy_clean.fasta -I Edy4_PE_sorted_nodup.bam -rf DuplicateRead
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar -T CountReads -R
Edy_clean.fasta -I Edy5_PE_sorted_nodup.bam -rf DuplicateRead

# Corrigindo BAMs:
java -jar /usr/local/picard-tools-1.135/picard.jar \
        AddOrReplaceReadGroups \
        I=Edy1_PE_sorted_nodup.bam \
        O=Edy1_PE_final.bam \
        SORT_ORDER=coordinate \
        RGID=Edy1 \
        RGLB=Edy1 \
        RGPL=illumina \
        RGPU=run \
        RGSM=Edy1 \
        CREATE_INDEX=True
```

```
java -jar /usr/local/picard-tools-1.135/picard.jar \
        AddOrReplaceReadGroups \
        I=Edy2_PE_sorted_nodup.bam \
        O=Edy2_PE_final.bam \
        SORT_ORDER=coordinate \
        RGID=Edy2 \
        RGLB=Edy2 \
        RGPL=illumina \
        RGPU=run \
        RGSM=Edy2 \
        CREATE_INDEX=True

java -jar /usr/local/picard-tools-1.135/picard.jar \
        AddOrReplaceReadGroups \
        I=Edy3_PE_sorted_nodup.bam \
        O=Edy3_PE_final.bam \
        SORT_ORDER=coordinate \
        RGID=Edy3 \
        RGLB=Edy3 \
        RGPL=illumina \
        RGPU=run \
        RGSM=Edy3 \
        CREATE_INDEX=True

java -jar /usr/local/picard-tools-1.135/picard.jar \
        AddOrReplaceReadGroups \
        I=Edy4_PE_sorted_nodup.bam \
        O=Edy4_PE_final.bam \
        SORT_ORDER=coordinate \
        RGID=Edy4 \
        RGLB=Edy4 \
        RGPL=illumina \
        RGPU=run \
        RGSM=Edy4 \
        CREATE_INDEX=True

java -jar /usr/local/picard-tools-1.135/picard.jar \
        AddOrReplaceReadGroups \
        I=Edy5_PE_sorted_nodup.bam \
        O=Edy5_PE_final.bam \
        SORT_ORDER=coordinate \
        RGID=Edy5 \
        RGLB=Edy5 \
        RGPL=illumina \
        RGPU=run \
        RGSM=Edy5 \
        CREATE_INDEX=True

# Estatisticas descritivas do alinhamento (BWA):
```

```
java -jar /usr/local/picard-tools-1.135/picard.jar \
    CollectAlignmentSummaryMetrics \
    R=Edy_clean.fasta \
    I=Edy1_PE_final.bam \
    O=Edy1_PE_final.bam.txt

# ETAPA3: Identificação de Variantes:
# Identificando as primeiras variantes para serem utilizadas na recalibração da qualidade de bases (BQSR):
# Primeira chamada de variantes:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T HaplotypeCaller \
-R Edy_clean.fasta \
-I Edy1_PE_final.bam -I Edy2_PE_final.bam -I Edy3_PE_final.bam -I Edy4_PE_final.bam -I Edy5_PE_final.bam \
--genotyping_mode DISCOVERY \
-stand_emit_conf 10 \
-stand_call_conf 30 \
--bamOutput raw_variants.bam \
--never_trim_vcf_format_field \
-o raw_variants.vcf

# Selecionando apenas SNPs nas variantes identificadas:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T SelectVariants \
-R Edy_clean.fasta \
-V raw_variants.vcf \
-selectType SNP \
-o raw_snps.vcf

# Aplicando o filtro para SNPs:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T VariantFiltration \
-R Edy_clean.fasta \
-V raw_snps.vcf \
--filterExpression "QD < 2.0 || FS > 20.0 || MQ < 40.0 || MQRankSum < -8.0 || ReadPosRankSum < -8.0" \
--filterName "my_snp_filter" \
-o filtered_snps.vcf

# Selecionando somente variantes do tipo INDEL:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T SelectVariants \
-R Edy_clean.fasta \
-V raw_variants.vcf \
-selectType INDEL \
-o raw_indels.vcf

# Aplicando filtro para INDELs:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
```

```
-T VariantFiltration \
-R Edy_clean.fasta \
-V raw_indels.vcf \
--filterExpression "QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0" \
--filterName "my_indel_filter" \
-o filtered_indels.vcf

# Criando lista dos INDELs:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar -T
RealignerTargetCreator -R Edy_clean.fasta -I Edy1_PE_final.bam --know
filtered_indels.vcf -o target_intervals1.list

java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar -T
RealignerTargetCreator -R Edy_clean.fasta -I Edy2_PE_final.bam --know
filtered_indels.vcf -o target_intervals2.list

java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar -T
RealignerTargetCreator -R Edy_clean.fasta -I Edy3_PE_final.bam --know
filtered_indels.vcf -o target_intervals3.list

java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar -T
RealignerTargetCreator -R Edy_clean.fasta -I Edy4_PE_final.bam --know
filtered_indels.vcf -o target_intervals4.list

java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar -T
RealignerTargetCreator -R Edy_clean.fasta -I Edy5_PE_final.bam --know
filtered_indels.vcf -o target_intervals5.list

# Aplicando realinhamento:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar -T IndelRealigner -R
Edy_clean.fasta -I Edy1_PE_final.bam -targetIntervals target_intervals1.list --know
filtered_indels.vcf -o Edy1_PE_final_.bam

java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar -T IndelRealigner -R
Edy_clean.fasta -I Edy2_PE_final.bam -targetIntervals target_intervals2.list --know
filtered_indels.vcf -o Edy2_PE_final_.bam

java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar -T IndelRealigner -R
Edy_clean.fasta -I Edy3_PE_final.bam -targetIntervals target_intervals3.list --know
filtered_indels.vcf -o Edy3_PE_final_.bam

java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar -T IndelRealigner -R
Edy_clean.fasta -I Edy4_PE_final.bam -targetIntervals target_intervals4.list --know
filtered_indels.vcf -o Edy4_PE_final_.bam

java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar -T IndelRealigner -R
Edy_clean.fasta -I Edy5_PE_final.bam -targetIntervals target_intervals5.list --know
filtered_indels.vcf -o Edy5_PE_final_.bam

# Base Quality Score Recalibration (BQSR) 1:
```

```
# Edy1:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T BaseRecalibrator \
-R Edy_clean.fasta \
-I Edy1_PE_final_.bam \
-knownSites filtered_snps.vcf \
-knownSites filtered_indels.vcf \
-o recal_data1.table

# Edy2:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T BaseRecalibrator \
-R Edy_clean.fasta \
-I Edy2_PE_final_.bam \
-knownSites filtered_snps.vcf \
-knownSites filtered_indels.vcf \
-o recal_data2.table

# Edy3:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T BaseRecalibrator \
-R Edy_clean.fasta \
-I Edy3_PE_final_.bam \
-knownSites filtered_snps.vcf \
-knownSites filtered_indels.vcf \
-o recal_data3.table

# Edy4:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T BaseRecalibrator \
-R Edy_clean.fasta \
-I Edy4_PE_final_.bam \
-knownSites filtered_snps.vcf \
-knownSites filtered_indels.vcf \
-o recal_data4.table

# Edy5:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T BaseRecalibrator \
-R Edy_clean.fasta \
-I Edy5_PE_final_.bam \
-knownSites filtered_snps.vcf \
-knownSites filtered_indels.vcf \
-o recal_data5.table

# Segunda etapa da recalibração:
# Edy1:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T BaseRecalibrator \
-R Edy_clean.fasta \
```

```
-I Edy1_PE_final_.bam \
-knownSites filtered_snps.vcf \
-knownSites filtered_indels.vcf \
-BQSR recal_data1.table \
-o post_recal_data1.table

# Edy2:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T BaseRecalibrator \
-R Edy_clean.fasta \
-I Edy2_PE_final_.bam \
-knownSites filtered_snps.vcf \
-knownSites filtered_indels.vcf \
-BQSR recal_data2.table \
-o post_recal_data2.table

#Edy3:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T BaseRecalibrator \
-R Edy_clean.fasta \
-I Edy3_PE_final_.bam \
-knownSites filtered_snps.vcf \
-knownSites filtered_indels.vcf \
-BQSR recal_data3.table \
-o post_recal_data3.table

#Edy4:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T BaseRecalibrator \
-R Edy_clean.fasta \
-I Edy4_PE_final_.bam \
-knownSites filtered_snps.vcf \
-knownSites filtered_indels.vcf \
-BQSR recal_data4.table \
-o post_recal_data4.table

# Edy5:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T BaseRecalibrator \
-R Edy_clean.fasta \
-I Edy5_PE_final_.bam \
-knownSites filtered_snps.vcf \
-knownSites filtered_indels.vcf \
-BQSR recal_data5.table \
-o post_recal_data5.table

# Gerando graficos antes/depois da recalibracao
# Edy1:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T AnalyzeCovariates \
```

```
-R Edy_clean.fasta \
-before recal_data1.table \
-after post_recal_data1.table \
-plots recalibration_plots1.pdf

# Edy2:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T AnalyzeCovariates \
-R Edy_clean.fasta \
-before recal_data2.table \
-after post_recal_data2.table \
-plots recalibration_plots2.pdf

# Edy3:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T AnalyzeCovariates \
-R Edy_clean.fasta \
-before recal_data3.table \
-after post_recal_data3.table \
-plots recalibration_plots3.pdf

# Edy4:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T AnalyzeCovariates \
-R Edy_clean.fasta \
-before recal_data4.table \
-after post_recal_data4.table \
-plots recalibration_plots4.pdf

# Edy5:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T AnalyzeCovariates \
-R Edy_clean.fasta \
-before recal_data5.table \
-after post_recal_data5.table \
-plots recalibration_plots5.pdf

# Aplicando a tabela de recalibracao:
# Edy1:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T PrintReads \
-R Edy_clean.fasta \
-I Edy1_PE_final_.bam \
-BQSR recal_data1.table \
-o recal_reads1.bam

# Edy2:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T PrintReads \
-R Edy_clean.fasta \
```

```
-I Edy2_PE_final_.bam \
-BQSR recal_data2.table \
-o recal_reads2.bam

# Edy3:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T PrintReads \
-R Edy_clean.fasta \
-I Edy3_PE_final_.bam \
-BQSR recal_data3.table \
-o recal_reads3.bam

# Edy4:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T PrintReads \
-R Edy_clean.fasta \
-I Edy4_PE_final_.bam \
-BQSR recal_data4.table \
-o recal_reads4.bam

# Edy5:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T PrintReads \
-R Edy_clean.fasta \
-I Edy5_PE_final_.bam \
-BQSR recal_data5.table \
-o recal_reads5.bam

# Variant calling com os BAMs recalibrados:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T HaplotypeCaller \
-R Edy_clean.fasta \
-I recal_reads1.bam \
-I recal_reads2.bam \
-I recal_reads3.bam \
-I recal_reads4.bam \
-I recal_reads5.bam \
--genotyping_mode DISCOVERY \
-stand_emit_conf 10 \
-stand_call_conf 30 \
-o raw_variants_recal1.vcf

### SEGUNDA RODADA DE RECALIBRAÇÃO:
# Selecionando apenas SNPs nas variantes identificadas:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T SelectVariants \
-R Edy_clean.fasta \
-V raw_variants_recal1.vcf \
-selectType SNP \
-o raw_snps2.vcf
```

```
# Aplicando o filtro para SNPs:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T VariantFiltration \
-R Edy_clean.fasta \
-V raw_snps2.vcf \
--filterExpression "QD < 2.0 || FS > 20.0 || MQ < 40.0 || MQRankSum < -8.0 ||
ReadPosRankSum < -8.0" \
--filterName "my_snp_filter" \
-o filtered_snps2.vcf

# Selecionando somente variantes do tipo INDEL:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T SelectVariants \
-R Edy_clean.fasta \
-V raw_variants_recal1.vcf \
-selectType INDEL \
-o raw_indels2.vcf

# Aplicando filtro para INDELs:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T VariantFiltration \
-R Edy_clean.fasta \
-V raw_indels2.vcf \
--filterExpression "QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0" \
--filterName "my_indel_filter" \
-o filtered_indels2.vcf

# Base Quality Score Recalibration (BQSR) 1:
# Edy1:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T BaseRecalibrator \
-R Edy_clean.fasta \
-I Edy1_PE_final.bam \
-knownSites filtered_snps2.vcf \
-knownSites filtered_indels2.vcf \
-o recal_data12.table

# Edy2:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T BaseRecalibrator \
-R Edy_clean.fasta \
-I Edy2_PE_final.bam \
-knownSites filtered_snps2.vcf \
-knownSites filtered_indels2.vcf \
-o recal_data22.table

# Edy3:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T BaseRecalibrator \
```

```
-R Edy_clean.fasta \
-I Edy3_PE_final.bam \
-knownSites filtered_snps2.vcf \
-knownSites filtered_indels2.vcf \
-o recal_data32.table

# Edy4:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T BaseRecalibrator \
-R Edy_clean.fasta \
-I Edy4_PE_final.bam \
-knownSites filtered_snps2.vcf \
-knownSites filtered_indels2.vcf \
-o recal_data42.table

# Edy5:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T BaseRecalibrator \
-R Edy_clean.fasta \
-I Edy5_PE_final.bam \
-knownSites filtered_snps2.vcf \
-knownSites filtered_indels2.vcf \
-o recal_data52.table

# Segunda etapa da recalibração:
# Edy1:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T BaseRecalibrator \
-R Edy_clean.fasta \
-I Edy1_PE_final.bam \
-knownSites filtered_snps2.vcf \
-knownSites filtered_indels2.vcf \
-BQSR recal_data12.table \
-o post_recal_data12.table

# Edy2:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T BaseRecalibrator \
-R Edy_clean.fasta \
-I Edy2_PE_final.bam \
-knownSites filtered_snps2.vcf \
-knownSites filtered_indels2.vcf \
-BQSR recal_data22.table \
-o post_recal_data22.table

#Edy3:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T BaseRecalibrator \
-R Edy_clean.fasta \
-I Edy3_PE_final.bam \
```

```
-knownSites filtered_snps2.vcf \
-knownSites filtered_indels2.vcf \
-BQSR recal_data32.table \
-o post_recal_data32.table

#Edy4:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T BaseRecalibrator \
-R Edy_clean.fasta \
-I Edy4_PE_final.bam \
-knownSites filtered_snps2.vcf \
-knownSites filtered_indels2.vcf \
-BQSR recal_data42.table \
-o post_recal_data42.table

# Edy5:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T BaseRecalibrator \
-R Edy_clean.fasta \
-I Edy5_PE_final.bam \
-knownSites filtered_snps2.vcf \
-knownSites filtered_indels2.vcf \
-BQSR recal_data52.table \
-o post_recal_data52.table

# Gerando graficos antes/depois da recalibracao
# Edy1:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T AnalyzeCovariates \
-R Edy_clean.fasta \
-before recal_data12.table \
-after post_recal_data12.table \
-plots recalibration_plots12.pdf

# Edy2:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T AnalyzeCovariates \
-R Edy_clean.fasta \
-before recal_data22.table \
-after post_recal_data22.table \
-plots recalibration_plots22.pdf

# Edy3:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T AnalyzeCovariates \
-R Edy_clean.fasta \
-before recal_data32.table \
-after post_recal_data32.table \
-plots recalibration_plots32.pdf
```

```
# Edy4:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T AnalyzeCovariates \
-R Edy_clean.fasta \
-before recal_data42.table \
-after post_recal_data42.table \
-plots recalibration_plots42.pdf

# Edy5:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T AnalyzeCovariates \
-R Edy_clean.fasta \
-before recal_data52.table \
-after post_recal_data52.table \
-plots recalibration_plots52.pdf

# Aplicando a tabela de recalibracao:
# Edy1:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T PrintReads \
-R Edy_clean.fasta \
-I Edy1_PE_final.bam \
-BQSR recal_data12.table \
-o recal_reads12.bam

# Edy2:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T PrintReads \
-R Edy_clean.fasta \
-I Edy2_PE_final.bam \
-BQSR recal_data22.table \
-o recal_reads22.bam

# Edy3:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T PrintReads \
-R Edy_clean.fasta \
-I Edy3_PE_final.bam \
-BQSR recal_data32.table \
-o recal_reads32.bam

# Edy4:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T PrintReads \
-R Edy_clean.fasta \
-I Edy4_PE_final.bam \
-BQSR recal_data42.table \
-o recal_reads42.bam

# Edy5:
```

```
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T PrintReads \
-R Edy_clean.fasta \
-I Edy5_PE_final.bam \
-BQSR recal_data52.table \
-o recal_reads52.bam

# Variant calling com os BAMs recalibrados:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T HaplotypeCaller \
-R Edy_clean.fasta \
-I recal_reads12.bam \
-I recal_reads22.bam \
-I recal_reads32.bam \
-I recal_reads42.bam \
-I recal_reads52.bam \
--genotyping_mode DISCOVERY \
-stand_emit_conf 10 \
-stand_call_conf 30 \
-o raw_variants_recal2.vcf

# Selecionando apenas SNPs nas variantes identificadas:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T SelectVariants \
-R Edy_clean.fasta \
-V raw_variants_recal2.vcf \
-selectType SNP \
-o raw_snps3.vcf

# Aplicando o filtro para SNPs:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T VariantFiltration \
-R Edy_clean.fasta \
-V raw_snps3.vcf \
--filterExpression "QD < 2.0 || FS > 20.0 || MQ < 40.0 || MQRankSum < -8.0 ||
ReadPosRankSum < -8.0" \
--filterName "my_snp_filter" \
-o filtered_snps3.vcf

# Selecionando somente variantes do tipo INDEL:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T SelectVariants \
-R Edy_clean.fasta \
-V raw_variants_recal2.vcf \
-selectType INDEL \
-o raw_indels3.vcf

# Aplicando filtro para INDELs:
java -jar /usr/local/GenomeAnalysisTK/GenomeAnalysisTK.jar \
-T VariantFiltration \
```

```
-R Edy_clean.fasta \
-V raw_indels3.vcf \
--filterExpression "QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0" \
--filterName "my_indel_filter" \
-o filtered_indels3.vcf
```

Caracterização das variantes utilizando snpEff e VCFTools.