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

AVALIAÇÃO DA TECNOLOGIA OXFORD NANOPORE PARA ANÁLISE DE IDENTIDADE GENÉTICA DE CLONES DE CANA-DE-AÇÚCAR (*Saccharum* spp.)

SÂMELLA DE SOUZA BORGES

Orientador(a):
Prof. Alexandre Siqueira Guedes Coelho

Goiânia - GO
Brasil

Março – 2020

UNIVERSIDADE FEDERAL DE GOIÁS
ESCOLA DE AGRONOMIA

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

E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

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

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

1. Identificação do material bibliográfico

[**x**] Dissertação [ ] Tese

2. Nome completo do autor

Sâmella de Souza Borges

3. Título do trabalho

"AVALIAÇÃO DA TECNOLOGIA OXFORD NANOPORE PARA ANÁLISE DE IDENTIDADE GENÉTICA DE CLONES DE CANA-DE-AÇÚCAR (*Saccharum* spp.)"

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

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

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

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

Documento assinado eletronicamente por **SÂMELLA DE SOUZA BORGES**, **Discente**, em 07/07/2020, às 16:02, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Alexandre Siqueira Guedes Coelho**, **Professor do Magistério Superior**, em 08/07/2020, às 11:34, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1426232** e o código CRC **12223D7D**.

Referência: Processo nº 23070.012771/2020-73 SEI nº 1426232

SÂMELLA DE SOUZA BORGES

AVALIAÇÃO DA TECNOLOGIA OXFORD NANOPORE PARA ANÁLISE DE IDENTIDADE GENÉTICA DE CLONES DE CANA-DE-AÇÚCAR (*Saccharum* spp.)

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

Coorientadora:
Dr.ª Ludmila Ferreira Bandeira

Goiânia, GO – Brasil
2020

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

de Souza Borges, Sâmella
Avaliação da tecnologia Oxford Nanopore para análise de identidade genética de clones de cana-de-açúcar (Saccharum spp.). [manuscrito] / Sâmella de Souza Borges. - 2020.
65 f.

Orientador: Prof. Dr. Alexandre Siqueira Guedes Coelho; co orientadora Dra. Ludmila Ferreira Bandeira.
Dissertação (Mestrado) - Universidade Federal de Goiás, Escola de Agronomia (EA), Programa de Pós-graduação em Genética e Melhoramento de Plantas, Goiânia, 2020.
Bibliografia.
Inclui siglas, abreviaturas, símbolos, gráfico, tabelas.

1. Oxford Nanopore Technologies. 2. Genotipagem por sequenciamento. 3. Identificação clonal. I. Siqueira Guedes Coelho, Alexandre , orient. II. Título.

CDU 575

UNIVERSIDADE FEDERAL DE GOIÁS

ESCOLA DE AGRONOMIA

ATA DE DEFESA DE DISSERTAÇÃO

Ata nº **0062/2020** da sessão de Defesa de Dissertação de **SÂMELLA DE SOUZA BORGES**, que confere o título de Mestre em **Genética e Melhoramento de Plantas**, na área de concentração em **Genética e Melhoramento de Plantas**.

Aos **18/03/2020 - Dezoito dias do mês de março do ano de dois mil e vinte**, a partir das 14:00 **horas**, no **Auditório Roland Vencovsky - Prédio do Setor de Melhoramento de Plantas**, realizou-se a sessão pública de Defesa de Dissertação intitulada **"AVALIAÇÃO DA TECNOLOGIA OXFORD NANOPORE PARA ANÁLISE DE IDENTIDADE GENÉTICA DE CLONES DE CANA-DE-AÇÚCAR (*Saccharum* spp.)"**. Os trabalhos foram instalados pelo Orientador, Professor Doutor **Alexandre Siqueira Guedes Coelho - EA/UFG**, com a participação dos demais membros da Banca Examinadora: Doutora **Tereza Cristina de Oliveira Borba - Embrapa Arroz e Feijão**, membro titular externo; Professor Doutor **Evandro Novaes - UFLA**, membro titular interno. Durante a arguição os membros da banca **não fizeram** sugestão de alteração do título do trabalho. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Dissertação, tendo sido a candidata **aprovada** pelos seus membros. Proclamados os resultados pelo Professor Doutor **Alexandre Siqueira Guedes Coelho - EA/UFG**, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, aos **18/03/2020 - Dezoito dias do mês de março do ano de dois mil e vinte**.

TÍTULO SUGERIDO PELA BANCA

Documento assinado eletronicamente por **Alexandre Siqueira Guedes Coelho**, **Professor do Magistério Superior**, em 18/03/2020, às 16:39, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Evandro Novaes**, **Usuário Externo**, em 18/03/2020, às 16:41, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Tereza Cristina de Oliveira Borba**, **Usuário Externo**, em 18/03/2020, às 16:47, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1225025** e o código CRC **728B610D**.

AGRADECIMENTOS

A presente dissertação de mestrado é prova de uma conclusão, conclusão que cheguei a acreditar que não aconteceria. O modelo da construção da ciência assim como de recursos humanos para esta, já se inicia em uma competição, consigo mesmo, com seus colegas. Ou seja, para chegar neste ponto há um longo caminho a ser percorrido, que nem sempre é trivial. No meu caso especificamente, apesar de cronologicamente ter feito o mestrado em três anos, praticamente toda a execução do projeto foi em seis meses. Antepor aos meus agradecimentos, quero dizer que ciência não se faz individualmente, não se faz sem um time, dito isso, quero agradecer todos que colaboraram para que eu conquistasse este título.

Primeiramente quero agradecer quem participou afinco de todo esse processo, agradeço ao meu orientador Alexandre por todo conhecimento teórico fornecido de diversas áreas, por toda disposição e paciência para ensinar praticamente tudo o que sei de bioinformática, por meio de aulas, ou em particular. À liberdade e confiança de manusear o MinIon sem experiência prévia. Agradeço profundamente, sua atitude no fim, quando os prazos já estavam completamente extrapolados e eu já não tinha nenhuma esperança, o fornecimento financeiro para compra tanto do sequenciador como de todos os reagentes para execução desse projeto, isto sem sombra de dúvidas não era sua obrigação. Quero terminar meus agradecimentos ao meu orientador com uma frase que é bastante dita por mim e por todos orientados dele “Precisa-se de mais Alexandres neste programa”. Contextualizando, esta frase significa que precisamos de mais profissionais dentre outros atributos, com habilidades em bioinformática. Finalizando, agradeço à minha coorientadora Ludmila, por toda disponibilidade, envolvimento na parte inicial experimental do meu trabalho, pelos ensinamentos laboratoriais.

Dentre tantas atividades executadas no mestrado, uma me contentou imensamente, então quero enaltecer meu agradecimento a quem participou ativamente dessa. Ao professor Sibov que foi uma peça de grande importância para meu crescimento, sou extremamente grata por ter me orientado no estágio docência, abrindo meus olhos a uma possibilidade de carreira que ainda não havia pensado. Agradeço a confiança depositada em mim, ao me dar a oportunidade lecionar até mesmo uma aula para pós-graduação, e à prontidão ao me chamar para um segundo estágio docência. Estas oportunidades podem facilmente não ter significado a ele, mas para mim, foi algo indescritível. Como coordenador

da pós-graduação, agradeço a empatia e todo auxílio oferecido para solucionar os problemas ao longo dessa trajetória. Agradeço também todos professores do PGMP que contribuíram durante minha jornada.

Sou grata a todos discentes do PGMP que convivi durante esses anos e se envolveram no meu crescimento pessoal e profissional. Em especial, às meninas que se tornaram grandes amigas, as quais quero levar para a vida inteira: Stelinha, Naíze e Bianquinha. A Stelinha agradeço por me apresentar calmamente o mundo da pós, por me ajudar a coletar minhas amostras, isso foi bastante importante pois eu não tinha nenhuma noção de campo. Além disso, agradeço por toda troca de dizeres sobre a vida. A Naíze agradeço por toda ajuda na execução final do meu projeto, ao desespero compartilhado e aos risos que foram consequências disso, agradeço também pelo suporte e esclarecimento de cada fase da depressão, com certeza aprendi bastante. A Bia agradeço pelo ombro amigo durante os momentos difíceis, o equilíbrio dessa relação em que ora era eu, ora era ela apoiando uma a outra. Enfim, por mais clichê que seja... vocês têm um "lugarzinho no meu coração".

Agradeço os meus pais Miraci, Marcos. Esse momento chega a ser contraditório, como há uma infinidade de motivos para agradece-los é bastante complicado descrever minha gratidão. Resumindo bem, agradeço por todos valores ensinados durante a vida, por todo amor, carinho e apoio. Eu sei que fizeram mais por mim e pelo meu irmão do que para vocês, somos eternamente gratos, fazemos o possível para orgulhar vocês. Aproveitando a citação ao irmão, quero agradecer o meu, Alex você não sabe a admiração que tenho por ti, digo e repito, você foi, é e será meu espelho, principalmente pela sua garra, dedicação e foco para seu crescimento profissionalmente.

Ao meu namorado Sales, por toda paciência, suporte e companheirismo, agradeço pelo incentivo e confiança, sempre me dizendo que sou capaz. Você é uma pessoa fora do comum. Agradeço também meus amigos que estão ao meu lado durante as diferentes fases da vida, de longa data: Paty e Arielly, e os mais recentes, minha família de jiu-jitsu que atuou como minha válvula de escape: mestre Leo, Gabriel, Sales, Henrique, Paulo, Neto, Alê, entre outros.

À UFG, à CAPES pela concessão da bolsa de estudos e à RIDESA pelo material fornecido. À minha banca, Evandro, Tereza, Adriana, Sibov e Alexandre pela disponibilidade e sugestões para melhorias deste trabalho.

SUMÁRIO

RESUMO

BORGES, S.S. **Avaliação da tecnologia Oxford Nanopore para análise de identidade genética de clones de cana-de-açúcar (*Saccharum* spp.).** 2020. 65 f. Dissertação (Mestrado em Genética e Melhoramento de Plantas) – Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2020.[1]

O Brasil é, atualmente, o maior produtor mundial de cana-de-açúcar, que é matéria prima de dois importantes produtos para a economia brasileira: o açúcar e o etanol. Com intuito tanto de se avaliar a tecnologia Oxford Nanopore para aplicações de genotipagem por sequenciamento (*Genotyping By Sequencing* – GBS), tendo como diferencial a obtenção de uma alta cobertura de sequenciamento (>1.000X), como de se desenvolver uma plataforma GBS capaz de identificar clones de cana-de-açúcar, com segurança e praticidade, este trabalho utilizou a plataforma MinION para a realização da genotipagem com base no sequenciamento de 48 indivíduos. Para isto, utilizando-se o genoma de *Sorghum bicolor* como referência e um conjunto de bibliotecas de transcritos de cana-de-açúcar, foram desenhados 20 pares de *primers*, os quais foram utilizados para obtenção de *amplicons* , nos quais foram identificados marcadores moleculares SNPs (*Single Nucleotide Polymorphisms*). Foram construídas seis bibliotecas de sequenciamento, sendo as duas primeiras utilizadas em ensaios-piloto. O alinhamento das sequências obtidas no genoma de referência foi realizado utilizando-se o programa BWA. A identificação e genotipagem dos SNPs foi realizada utilizando-se o software *SAMtools*. A identificação dos clones de cana-de-açúcar foi feita a partir do cálculo da distância genética entre os indivíduos. A análise de agrupamento foi realizada utilizando-se um *script* escrito no *software* R. O sequenciamento resultou em cerca de 841 mil sequências. O tamanho médio dos *amplicons* foi de 1,6 kb. Obteve-se uma alta cobertura de sequenciamento (média de 10.498X/*amplicon*). Nove *amplicons* foram selecionados, nos quais foram identificados 356 sítios SNPs. A porcentagem de *mismatches* entre as sequências obtidas e a de referência variou de 8% a 20% e a porcentagem de *indels* manteve-se homogênea (~6%). As duplicatas de um mesmo indivíduo utilizadas como controle biológico formaram nó com consistência de 94% no dendrograma obtido, entretanto não apresentaram identidade genética perfeita entre elas. Sugere-se que tal fato esteja associado principalmente à alta taxa de erro de sequenciamento da tecnologia de sequenciamento Oxford Nanopore, evidenciando a dificuldade de sua utilização em aplicações que demandem uma identificação genética com alto grau de segurança, como ocorre em problemas envolvendo a identificação clonal em cana-de-açúcar.

Palavras-chave: Oxford Nanopore Technologies, genotipagem por sequenciamento, identificação clonal.

---

[1] Orientador: Prof. Dr. Alexandre Siqueira Guedes Coelho – EA/UFG.
Coorientadora: Dr.ª Ludmila Ferreira Bandeira.

ABSTRACT

BORGES, S.S. **Evaluation of Oxford Nanopore technology to analyze the genetic identity of sugarcane clones (Saccharum spp.).** 2020. 65 f. Dissertation (Master of Science in Genetics and Plant Breeding) – Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2020.[1]

Brazil is currently the world's largest producer of sugarcane, which is the raw material for two important products for the Brazilian economy: sugar and ethanol. In order to evaluate both Oxford Nanopore technology for genotyping applications by sequencing (Genotyping By Sequencing - GBS), with the differential of obtaining a high coverage of sequencing (> 1,000X), as well as to develop a GBS platform capable of to identify sugarcane clones, safely and conveniently, this work used the MinION platform to perform genotyping based on the sequencing of 48 individuals. For this, using the genome of Sorghum bicolor as reference and a set of libraries of sugarcane transcripts, 20 pairs of primers were designed, which were used to obtain amplicons, in which SNPs molecular markers were identified (Single Nucleotide Polymorphisms). Six sequencing libraries were built, the first two being used in pilot trials. The alignment of the sequences obtained in the reference genome was carried out using the BWA program. The identification and genotyping of the SNPs was performed using the SAMtools software. The identification of the sugarcane clones was done by calculating the genetic distance between the individuals. The cluster analysis was performed using a script written in software R. The sequencing resulted in approximately 841 thousand sequences. The average size of the amplicons was 1.6 kb. High sequencing coverage (average of 10,498X / amplicon) was obtained. Nine amplicons were selected, in which 356 SNPs sites were identified. The percentage of mismatches between the obtained and the reference sequences ranged from 8% to 20% and the percentage of indels remained homogeneous (~ 6%). The duplicates of the same individual used as biological control formed a knot with a consistency of 94% in the obtained dendrogram, however they did not present perfect genetic identity between them. It is suggested that this fact is mainly associated with the high rate of sequencing error of the Oxford Nanopore sequencing technology, showing the difficulty of its use in applications that require a genetic identification with a high degree of security, as it occurs in problems involving clonal identification in sugar cane.

Keywords: Oxford Nanopore Technologies, genotyping by sequencing, clonal identification.

[1] Advisor: Prof. Dr. Alexandre Siqueira Guedes Coelho – EA/UFG.
Co-Advisor: Dr.ª Ludmila Ferreira Bandeira.

1 INTRODUÇÃO

As tecnologias de sequenciamento de DNA têm evoluído rapidamente e hoje permitem a análise genômica de qualquer organismo vivo. A partir dos anos 2000, surgiram as plataformas denominadas de sequenciamento de nova geração (*Next Generation Sequencing* – NGS), cuja grande vantagem é a produção de grande quantidade de dados a um custo reduzido por *data point* (Metzker, 2010; Elshire et al., 2011). Desde então, as tecnologias NGS vêm sendo utilizadas para descoberta e avaliação de marcadores genéticos em diferentes populações, em estudos de expressão gênica, de genômica comparativa, na análise genética de doenças, na realização de testes pré-natais não invasivos, entre outras aplicações (Davey et al., 2011).

Com o desenvolvimento das tecnologias NGS, a identificação de polimorfismos de um único nucleotídeo (*Single Nucleotide Polymorphisms* – SNP) passou a ser possível por meio da análise comparativa de alinhamento dos genomas sequenciados (Davey et al., 2011). Os primeiros métodos de genotipagem SNP utilizavam produtos de reação em cadeia da polimerase (*Polymerase Chain Reaction* – PCR) para detecção por fluorescência de hibridização de sondas alelo-específicas, como pelos sistemas *TaqMan* e *Invader*. Foram também desenvolvidas técnicas de genotipagem SNP baseadas na utilização de espectrometria de massa, por exemplo, pelo sistema *iPLEX* da empresa *Sequenom* (Deschamps et al., 2012).

Após o desenvolvimento das tecnologias NGS, surgiu uma abordagem denominada de genotipagem por sequenciamento (*Genotyping By Sequencing* – GBS). Esta abordagem pode ser utilizada no contexto da biologia vegetal para a predição de fenótipos, seleção de cultivares melhoradas, reprodução assistida, entre outras aplicações, contribuindo, por exemplo, para o aumento do abastecimento alimentício global (Poland & Rife, 2012). Como a maioria dos objetivos de estudos genéticos não necessita do conhecimento da sequência completa dos genomas de todos os indivíduos, as técnicas de genotipagem por sequenciamento em uso têm se baseado na abordagem de sequenciamento de uma representação reduzida dos genomas de interesse, e vêm sendo aplicadas em organismos poliploides, como a cana-de-açúcar.

A cana-de-açúcar fornece matéria prima para a fabricação de dois produtos de grande importância para a economia brasileira: o açúcar empregado na indústria alimentícia, e o etanol para produção de biocombustível de uso disseminado no país (Conab, 2017). O Brasil é o maior produtor mundial de cana-de-açúcar. Apesar deste sucesso, a cadeia produtiva da cana-de-açúcar tem forte demanda por cultivares mais produtivas, com maior resistência a pragas, a doenças e a condições de estresse abiótico. Atualmente, o ciclo completo de melhoramento genético de cana-de-açúcar leva cerca de 13 anos (https://www.ridesa.com.br/).

A Rede Interuniversitária para o Desenvolvimento do Setor Sucroenergético (Ridesa) é um programa de melhoramento genético de cana-de-açúcar que tem adotado a estratégia de semeio e plantio denominada *tapetinho*, a qual consiste no plantio agrupado por cova, na fase de seleção de plântulas (Mendes, 2015). Assim, além de simplificar o plantio, o processo inicia-se com uma competição natural entre os indivíduos de um mesmo *tapetinho*. Entretanto, apesar da vantagem de acelerar o programa de melhoramento ainda na primeira fase de seleção, devido ao perfilhamento dos colmos de cana-de-açúcar presentes em uma touceira, é possível que ocorra uma seleção errônea, de modo que clones sejam tratados como genótipos diferentes nas etapas seguintes do programa.

O genoma da cana-de-açúcar é bastante complexo, contendo alta taxa de DNA repetitivo. As cultivares modernas de cana-de-açúcar são tipicamente híbridos interespecíficos de duas espécies poliploides (*S. officinarum* e *S. spontaneum*) e apresentam níveis de ploidia variáveis por grupos de homeologia entre $2n=6x$ a $2n=20x$ (Garcia et al., 2013). Desta forma, o desenvolvimento de novas tecnologias de avaliação genética, notadamente métodos de genotipagem por sequenciamento capazes de identificar clones de cana-de-açúcar, com praticidade e segurança, permitirá a identificação de genótipos propagados e utilizados comercialmente em usinas de cana-de-açúcar com uma maior precisão, podendo auxiliar também, a redução do tempo da avaliação de clones nos programas de melhoramento de cana-de-açúcar. Neste contexto, este trabalho foi realizado com o objetivo de se desenvolver e avaliar um conjunto de marcadores SNPs, obtidos por sequenciamento pela tecnologia Oxford Nanopore, no intuito de se obter uma plataforma de genotipagem rápida e segura para identificação de clones de cana-de-açúcar.

2 REVISÃO BIBLIOGRÁFICA

2.1 A CULTURA DA CANA-DE-AÇÚCAR

2.1.1 Importância econômica da cana-de-açúcar no Brasil

No período colonial do Novo Mundo, a agroindústria da cana-de-açúcar foiuma das principais atividades econômicas, de modo que até nos dias atuais os subprodutos da cana-de-açúcar exercem um papel importante para economia de alguns países, incluindo o Brasil. Desde o início da expansão da produção da cana-de-açúcar até o fim do século XVII, o Brasil foi o maior exportador de açúcar do mundo, tendo ocorrido uma retração no século seguinte e uma nova expansão por volta de 1920, com a crise do café (Brandão, 1985; Santos et al., 2016).

Com a crise do petróleo em 1973, iniciou-se uma busca mundial por opções de fontes alternativas de energia. Apenas dois anos depois, o Brasil foi o único país que lançou um programa de biocombustíveis. Devido ao problema de esgotamento das jazidas petrolíferas e consequentemente o alto preço do petróleo, a agroindústria sucroalcooleira apresentou-se como uma alternativa benéfica para o setor de biocombustíveis, uma vez que o álcool é um combustível ecologicamente correto obtido por meio de fontes renováveis. Por isto, já em 2011, o etanol oriundo do bagaço da cana-de-açúcar correspondia a cerca de 16% da matriz energética brasileira (Nitsch, 1991; Shikida & Perosa, 2012).

É irrefutável a importância econômica da cana-de-açúcar no Brasil, pois o seu plantio em áreas brasileiras tem grande extensão. A produção estimada para a safra 2018/19 é superior a 635 milhões de toneladas. Os principais subprodutos da cana-de-açúcar são o açúcar, com estimativa de produção em torno de 34,25 milhões de toneladas e o etanol, com a produção estimada em 30 bilhões de litros (safra 2018/19). Atualmente as regiões nacionais Sudeste e Centro-Oeste são as que mais contribuem, tanto em termos de produtividade (76,62 t/ha e 74,04 t/ha respectivamente) quanto em termos de fabricação de açúcar e etanol, tendo como destaque os estados de São Paulo e Goiás. Quanto à produção de açúcar, destaca-

se a região Sudeste com estimativa de produção superior a 25 milhões de toneladas, representando 74,1% de todo açúcar produzido no Brasil (safra 2018/17), enquanto na região Centro-Oeste esta estimativa é de 3,8 milhões de toneladas. Já em relação à produção de etanol, a estimativa para a região Sudeste na safra 2018/19 é de 17,5 bilhões de litros, enquanto esta estimativa para região Centro-Oeste é de 9,4 bilhões de litros (Conab, 2018).

A cana-de-açúcar ainda é utilizada nos setores de bebidas e alimentos, como por exemplo, na fabricação da cachaça, que é uma bebida alcoólica, introduzida no Brasil no período colonial, obtida através da fermentação do caldo da cana-de-açúcar seguida pela destilação, que pode conter de 38% a 54% de álcool. Em termos de consumo, a cachaça é a terceira bebida destilada mais ingerida no mundo, e a segunda bebida alcoólica mais consumida no Brasil. Ainda no setor alimentício, mas de modo artesanal, a cana-de-açúcar é matéria prima para a produção de açúcar mascavo, de rapadura e de melado, os quais são subprodutos obtidos principalmente em pequenas propriedades rurais. Dentre tais produtos, o açúcar mascavo é o que tem o nível de açúcar mais elevado, seguido pela rapadura, e por fim o melado (Cib, 2009).

Um dos subprodutos do processamento industrial da cana-de-açúcar para produção de etanol é a vinhaça, que é rica em substâncias orgânicas e minerais. A vinhaça pode ser utilizada para diferentes aplicações como: adubo, produção de proteínas por fermentação anaeróbica, e formulação de ração animal. Além disto, nos últimos anos a cultura da cana-de-açúcar tem sido destinada à obtenção de biomassa. Enquanto outras espécies de plantas convertem pouca quantidade de luz solar em energia química (menos de 1%), no caso da cana-de-açúcar esta conversão é de aproximadamente 2%. Neste sentido, o bagaço da cana-de-açúcar vem sendo utilizado como matéria prima para produção de energia. Estima-se que pela utilização do bagaço de cana-de-açúcar como combustível para usinas termoelétricas no Brasil sejam gerados mais de 11 milhões de kW. Esta forma de biomassa gera 26,9% do total de energia das usinas termoelétricas no país, ficando atrás somente do gás natural que representa 31,2% (Cib, 2009; Aneel, 2018).

Sendo assim, a importância da cana-de-açúcar no mercado sucroalcooleiro no Brasil é enorme, pois além de ser um produto agrícola é também a mais importante fonte de biomassa energética. A cana-de-açúcar em 2016 foi a segunda principal atividade da agroindústria movimentando R$ 53,4 bilhões, ficando atrás apenas da soja (R$ 116,5 bilhões) o que representa cerca de 2% do Produto Interno Bruto (PIB) brasileiro, envolvendo cerca de um milhão de empregos, o que representa 6% dos empregos no setor da agroindústria

nacional (Cib, 2009; Cni, 2017).

2.1.2 Taxonomia e Origem

A cana-de-açúcar é uma gramínea, alógama, semi-perene, cultivada em regiões tropicais e subtropicais. Esta planta pertence à família *Poaceae*, subfamília *Panicoideae*, tribo *Andropogoneae*, subtribo *Saccharinae* e gênero *Saccharum*. As formas selvagens de *Saccharum* podem ser encontradas em diferentes regiões ao norte do continente Africano, algumas regiões do continente Asiático (Afeganistão, Turquestão, Paquistão, Índia, Myanmar, Tailândia, Japão, Indonésia, Filipinas e sul da China) e ainda em regiões do continente da Oceania (Nova Guiné e algumas ilhas da Melanésia e Polinésia) (Bremer, 1961; Jannoo et al., 1999).

Acredita-se que a inserção da cana-de-açúcar nas Américas deu-se no ano de 1493 através da segunda expedição de Cristóvão Colombo. Já a introdução desta cultura no solo brasileiro ocorreu antes da chegada da realeza portuguesa ao Brasil, pela utilização de de mudas provenientes da Ilha de Madeira, provavelmente no início do XVI, por Martim Afonso de Souza (Araújo & Santos, 2014).

2.1.3 As espécies do gênero *Saccharum*

Atualmente, existem seis espécies descritas no gênero *Saccharum*, sendo elas: *S. officinarum* L., *S. spontaneum* L., *S. barberi*, *S. sinense*, *S. robustum* e *S. edule*. De acordo com estudos filogenéticos, apenas três, destas seis espécies, são de fato espécies filogenéticas: *S. officinarum*, *S. spontaneum* e *S. robustum*, as demais espécies são híbridas naturais podendo ser chamadas de "pseudo-espécies" (Bremer, 1961; Oliveira, 2014).

O gênero *Saccharum* compreende plantas com alta complexidade genômica, as quais são todas poliploides constituídas por um grande número de cromossomos e consequentemente um genoma extenso, podendo ter o número básico de cromossomos de 6 a 20 (Bremer, 1961; D'Hont et al., 1996; Garcia et al., 2013).

Entre as espécies de *Saccharum* há diferenças anatômicas (macroscópicas) e histológicas (microscópicas). Plantas da espécie *S. officinarum,* também conhecida como "cana nobre" devido ao grande teor de sacarose e baixa porcentagem de fibra, têm as hastes

grossas que podem ter uma pequena variação na cor e na cobertura de cera. A maioria dos clones de *S. officinarum* tem número gamético de cromossomos igual a 40, ou seja, $2n = 80$, com número básico de cromossomo $x = 10$. Sendo assim, células somáticas de indivíduos desta espécie têm oito cópias de cada cromossomo (Bremer, 1961; Garcia et al., 2013).

Plantas da espécie *S. spontaneum* possuem colmos finos, curtos e fortes, contendo um alto teor de fibras e baixa quantidade de sacarose. O número de cromossomos pode variar de $2n = 40$ a $2n = 128$, sendo o número básico de cromossomo sugerido de $x = 8$, ou seja, é possível a ocorrência de 5 a 16 cromossomos homólogos de um conjunto básico de 8 cromossomos (Bremer, 1961; Irvine, 1999).

As plantas da espécie *S. robustum* possuem caules com colorações que podem variar entre verde, vermelho e roxo. Os colmos de plantas dessa espécie têm diâmetros maiores, rígidos, grandes, fibrosos e pobres em sacarose. Quanto ao número de cromossomos, é possível encontrar uma variação de $2n = 60$ a $2n = 170$, sendo que $2n = 60$ e $2n = 80$ são os valores mais comuns. O número básico de cromossomos sugerido para esta espécie também é de $x = 10$ (Bremer, 1961; Irvine, 1999; Garcia et al., 2013).

A variação do número de cromossomos que a espécie *S. sinense* apresenta é de $2n = 81$ a $2n = 124$. As plantas da espécie *S. barberi* possuem a variação no número de cromossomos de $2n = 111$ a $2n = 120$. Já a espécie *S. edule* contém a variação no número de cromossomos de $2n = 60$ a $2n = 80$, sendo que o número mais comum, assim como em *S. officinarum*, *S. spontaneum* e *S. robustum* é $2n = 80$ (D'Hont et al., 1996; Irvine, 1999; Morais et al., 2015).

Atualmente, as cultivares modernas da cana-de-açúcar são tipicamente híbridos interespecíficos de *S. officinarum* com *S. spontaneum*. As espécies *S. officinarum* e *S. robustum* estão intimamente relacionadas nos aspectos de citologia, morfologia e fisiologia, distinguindo-se principalmente quanto aos teores de fibra e sacarose dispostos em seus caules. Já as espécies *S. barberi* e *S. sinense* são ainda mais semelhantes entre si, de modo que muitas vezes são consideradas suficientemente diferentes para serem individualizadas, mas não o suficiente para serem classificadas como espécies diferentes. Indivíduos da espécie *S. edule* possuem flores inférteis, o que geralmente leva ao aborto de suas inflorescências, que são servidas como um alimento, principalmente na Oceania (Irvine, 1999; Grivet et al., 2004).

2.2 TECNOLOGIAS DE SEQUENCIAMENTO DE NOVA GERAÇÃO

2.2.1 Histórico das tecnologias de sequenciamento de nova geração

Historicamente, a primeira tecnologia de sequenciamento a ser desenvolvida foi o método de degradação química de Maxam & Gilbert, em 1977. No mesmo ano, Frederick Sanger e colaboradores também desenvolveram uma tecnologia de sequenciamento a qual foi posteriormente automatizada, tornando-se amplamente utilizada. Após estes métodos, surgiram vários outros, os quais foram denominados de tecnologias de sequenciamento de nova geração (*Next Generation Sequencing* – NGS) (Goodwin et al., 2016).

A primeira plataforma NGS só foi comercializada após aproximadamente três décadas desde o desenvolvimento do método de Sanger. A plataforma 454 foi lançada em 2005, em 2007 foi adquirida pela Roche, e em 2013 esse equipamento deixou de ser produzido. Essa tecnologia faz uso da abordagem de PCR em emulsão para obtenção de fragmentos clonados, e baseia-se na estratégia de pirossequenciamento, em que há liberação de pirofosfato (PPi) a cada incorporação de um nucleotídeo. A leitura das sequências é feita a partir da liberação de um pirofosfato, o qual é convertido em ATP por meio da enzima ATP sulfurilase. Este ATP é utilizado pela enzima luciferase, para oxidar a luciferina, produzindo então um sinal de luz, captado por uma câmera CCD (*Charge Coupled Device*) (Goodwin et al., 2016) (http://allseq.com).

A segunda técnica de NGS foi introduzida pela Solexa, através do equipamento *Genome Analyzer* (GA), no ano de 2006. No ano seguinte, esta empresa foi adquirida pela Illumina. Esta tecnologia é atualmente responsável por mais de 90% do volume de dados de sequenciamento no mundo, provavelmente devido à elevada acurácia, ao baixo custo, e ao alto volume de dados produzido (Goodwin, et al., 2016).

O método da Illumina ocorre em uma superfície sólida de vidro (*flow cell*), onde ficam dispostos dois tipos de oligonucleotídeos. Os fragmentos de DNA a serem sequenciados possuem adaptadores em suas extremidades que são complementares aos oligonucleotídeos da *flow cell*. Inicialmente é feita uma amplificação com nucleotídeos não marcados para formação de *clusters* (Liu et al., 2012). Em cada *cluster* ocorre a clivagem de todos os fragmentos que estão ligados a um dos tipos de oligonucleotídeos da *flow cell*, para que ocorra o sequenciamento por síntese com terminação cíclica reversível (CTR) com nucleotídeos marcados com uma fluorescência diferente, e com o grupamento ribose 3'-OH

bloqueado, o que impede a extensão da fita de DNA (Glenn, 2011). A visualização se dá por meio de microscopia de reflexão de fluorescência utilizando canais de laser e uma câmera CCD (Glenn, 2011; Goodwin et al., 2016).

A plataforma de sequenciamento por ligação e detecção de oligonucleotídeos (*Sequencing by Oligonucleotide Ligation and Detection* – SOLiD), terceira tecnologia NGS comercial, foi lançada pela Life Technologies também em 2006 e atualmente pertence à Thermo Fisher. Esta tecnologia também utiliza PCR em emulsão para amplificação clonal de fragmentos e fundamenta-se no sequenciamento por ligação (*Sequencing by Ligation* - SBL) (Goodwin et., 2016). As bibliotecas são construídas com fragmentos de DNA de tamanhos específicos, ligados a adaptadores em suas extremidades. As bibliotecas são submetidas a uma PCR, na qual os adaptadores de uma das extremidades dos fragmentos se hibridizam com *primers* de um suporte sólido ocorrendo a amplificação do DNA. Em seguida, ocorre a desnaturação parcial do ácido desoxirribonucleico fita dupla (dsDNA), possibilitando a hibridização com um novo *primer*, resultando em uma nova amplificação. Este ciclo se repete permitindo a clonagem dos fragmentos. As bibliotecas são sequenciadas em uma *flow cell* pelo sequenciamento baseado em ligação de sondas "di-bases" (Goodwin, et., 2016). Após a hibridização da sonda ao adaptador do fragmento de DNA, ocorre a extensão e a identificação das duas primeiras bases conhecidas, e em seguida a clivagem das bases degeneradas e do fluoróforo previamente ligados à sonda. Este processo se repete resultando em uma extensão de até 60 pb. Posteriormente é feita a remoção de todas as sondas e repete-se todo o processo de ligação de sondas, extensão, identificação de bases e clivagem por quatro vezes. Em cada ciclo as sondas utilizadas contêm uma base a mais: n + 1, n + 2, n + 3 e n + 4, de modo que cada base seja sequenciada duas vezes (Goodwin, et al., 2016).

A tecnologia Ion Torrent, comercializada a partir de 2010 e também adquirida pela Thermo Fisher, inovou os métodos de sequenciamento por ser a primeira tecnologia a não utilizar fluorescência e sensor óptico no processo. A detecção de bases incorporadas é feita por meio de íons de hidrogênio, logo, diferentemente das anteriores, não há formação de imagens para identificação das bases, acarretando vantagem na redução de tempo de sequenciamento (Glenn, 2011; Liu et al., 2012; Buermans & Den Dunnen, 2014). O método de sequenciamento Ion Torrent utiliza esferas e PCR em emulsão para formação clonal de fragmentos. A reação de PCR em emulsão contém: esferas com adaptadores fixados a elas (complementares a uma das extremidades dos adaptadores da biblioteca de DNA), reagentes

necessários para a amplificação e outro tipo de adaptador complementar à outra extremidade dos fragmentos das bibliotecas de DNA. A detecção das bases nucleotídicas é feita a partir da mudança de 0,02 unidades no pH que é obtida após da incorporação de um desoxirribonucleosídeo trifosfato (dNTP) na fita de DNA, na qual há liberação do próton (H^+) no meio da reação, de modo que a alteração de pH identificada pelo sensor é imperfeitamente proporcional à quantidade de nucleotídeos detectados. Os dNTPs são adicionados na reação de sequenciamento em múltiplos ciclos ordenados. Isto é necessário pois a liberação do próton dos diferentes dNTPs não é diferenciado (Buermans & Den Dunnen, 2014; Goodwin, et al., 2016).

O primeiro sequenciamento em tempo real de molécula única (*Single-Molecule Real-Time* - SMRT) foi lançado pela Pacific Biosciences também em 2010. Suficientemente sensível para identificar a incorporação de um único nucleotídeo marcado com fluorescência, na tecnologia PacBio é possível se obter sequências longas (de até aproximadamente 175 kb), sem que haja pré-amplificação dos fragmentos a serem sequenciados (Buermans & Den Dunnen, 2014; Van Dijk et al., 2014). Nesse método, os nucleotídeos são marcados por um fluoróforo ligado à sua porção fosfato terminal. Assim, durante a atividade da DNA polimerase ocorre a liberação do fluoróforo do nucleotídeo incorporado, resultando apenas em DNA não modificado. A preparação da biblioteca de DNA utiliza adaptadores em "alça" para tornar o fragmento de dsDNA circular. Estes fragmentos circulares são então sequenciados repetidas vezes. O sequenciamento inicia-se pela interação da DNA polimerase com a região do adaptador. A incorporação dos dNTPs é visualizada, em tempo real, por meio de um sistema de feixes de laser acoplado a um sistema de gravação confocal localizado abaixo das ZMW (*Zero-Mode Waveguides*), que registram tanto a cor quanto a duração do sinal de fluorescência emitido devido à incorporação dos dNTPs. Após a incorporação de um dNTP a DNA polimerase cliva o fluoróforo para que haja a incorporação e identificação de um novo dNTP (Buermans & Den Dunnen, 2014; Goodwin, et al., 2016).

Em 2015 foi lançado o MinION, que foi o primeiro sequenciador da empresa Oxford Nanopore Technologies (ONT). Na maioria dos casos, cada empresa desenvolveu mais de uma plataforma de sequenciamento, sendo que estas são melhoradas rapidamente em relação à quantidade de dados gerados, ao tamanho das sequências produzidas e ao tempo de corrida.

Tabela 1. Características gerais das diferentes plataformas de sequenciamento de nova geração (NGS). Fonte: https://allseq.com, https://www.illumina.com, https://www.thermofisher.com e https://www.pacb.com.

Empresa	Sequenciador	Total de bases por corrida (Gb)	Tamanho das sequências (pb)	Tempo de corrida (h)	Preço do equipamento (US$)
Illumina	iSeq 100	1,2	2 x 150	9,5 – 19	190.900
	MiniSeq	7,5	2 x 150	4 – 24	50.000
	MiSeq	15	2 x 300	4 – 55	99.000
	NextSeq 550	120	2 x 150	12 – 30	285.000
	NextSeq 1000	120	2 x 150	29	215.000
	NextSeq 2000	300	2 x 150	24 – 48	335.000
	NovaSeq	6000	2 x 150	13 – 45	985.000
Thermo Fisher	SOLiD 5500xl	95	2 x 60	144	595.000
	SOLiD 5500xl Wildfire	240	2 x 50	240	70.000
	SOLiD 5500	48	2 x 60	144	349.000
	SOLiD 5500 Wildfire	120	2 x 50	240	70.000
	Ion Gene Studio S5 System	15	600	19	120.000
	Ion Gene Studio S5 Plus System	30	600	10 – 11,5	150.000
	Ion Gene Studio S5 Prime System	50	600	6,5	180.000
Pacific Biosystems	Sequel	20	1.000 – 20.000	20	350.000
	Sequel II	160	15.000 – 20.000	30	495.000

2.2.2 A Tecnologia Oxford Nanopore

A empresa Oxford Nanopore Technologies foi fundada em 2005 com a finalidade de criar um sistema de detecção de molécula única fundamentada na tecnologia de

nanoporos. Após dez anos de desenvolvimento, o primeiro sequenciador *MinION* foi comercializado e, diferentemente da maioria dos sequenciadores, este equipamento é pequeno (10 cm de comprimento x 3 cm de largura x 2 cm de altura, pesando apenas 90 g), podendo ser equiparado a um *pendrive* não só pelo tamanho, mas também pela conexão direta a uma porta USB. Em seguida, foi disponibilizado o *GridION* em 2017 e no ano seguinte o *PromethION*. Além destes, outros equipamentos foram desenvolvidos pela empresa, como o *Flongle* e *SmidgION*, este último com a proposta de ser compatível com celulares (Figura 1). Os diferentes equipamentos possuem diferentes características – os equipamentos maiores possibilitam a utilização de várias *flow cells* ao mesmo tempo, aumentando tanto o rendimento quanto o tempo de sequenciamento (Tabela 2).

Figura 1. Sequenciadores desenvolvidos pela Oxford Nanopore Technologies (Fonte: https://nanoporetech.com/).

Assim como no caso da tecnologia PacBio, essa tecnologia se baseia no sequenciamento de moléculas únicas, sendo possível observar o sequenciamento em tempo real e produzir sequências longas. Até o momento, a maior sequência obtida foi de cerca de 2 Mb (https://nanoporetech.com). Ao contrário do que ocorre com a tecnologia PacBio, na tecnologia Oxford Nanopore o sequenciamento não é limitado pela própria tecnologia, mas sim pelo tamanho e quantidade das moléculas de DNA a serem sequenciadas (Van Dijk et al., 2018).

Tabela 2. Características gerais dos sequenciadores da Oxford Nanopore Technologies (Fonte: https://nanoporetech.com).

Sequenciador	Rendimento por *flow cell* (Gb)	Nº de *flow cells*	Nº de canais por *flow cell*	Tempo de corrida	WGS*
Flongle	2	1	126	1 min - 16 horas	
MinION Mk1B	15 – 30	1	512	1 min - 48 horas	
MinION Mk1C	15 – 30	1	512	1 min - 48 horas	X
GridION	15 – 30	5	512	1 min - 48 horas	X
PromethION 24	100 –180	24	3.000	1 min - 72 horas	X
PromethION 48	100 –180	48	3.000	1 min - 72 horas	X

* *Whole Genome Sequencing.*

A preparação de bibliotecas pode ser feita por diversos kits disponibilizados pela própria empresa. É relevante salientar a existência do kit de sequenciamento rápido, no qual o manejo para o preparo da biblioteca é simples e rápido (cerca de 10 minutos), tornando-o visado para estudos em campo (Van Dijk et al., 2018). De modo geral, o preparo das bibliotecas consiste em: obtenção do DNA purificado, indexação com *barcodes* e ligação dos adaptadores em "Y" (Lu et al., 2016).

A tecnologia Oxford Nanopore se fundamenta na utilização de um nanoporo de proteína que separa dois compartimentos distintos. O sequenciamento inicia-se pela extremidade 5' de fita simples do adaptador em "Y" com o auxílio de uma proteína motora (localizada acima do nanoporo) que desnatura o dsDNA (Figura 2). No momento em que a molécula de DNA atravessa o nanoporo, ocorre uma mudança temporária no potencial entre os compartimentos denominado de *squiggle space*, no qual as mudanças de voltagens correspondem a uma sequência específica de DNA, que são interpretadas como *k-meros* (de 3 a 6 bases). A *flow cell* destes sequenciadores é constituída por um microchip de circuito integrado específico para aplicações (*Application-Specific Integrated Circuit* - ASIC) no qual são depositados os dados para posterior análise pelo *software* MinKNOW (Buermans & Den Dunnen, 2014; Lu et al., 2016; Goodwin et al., 2016).

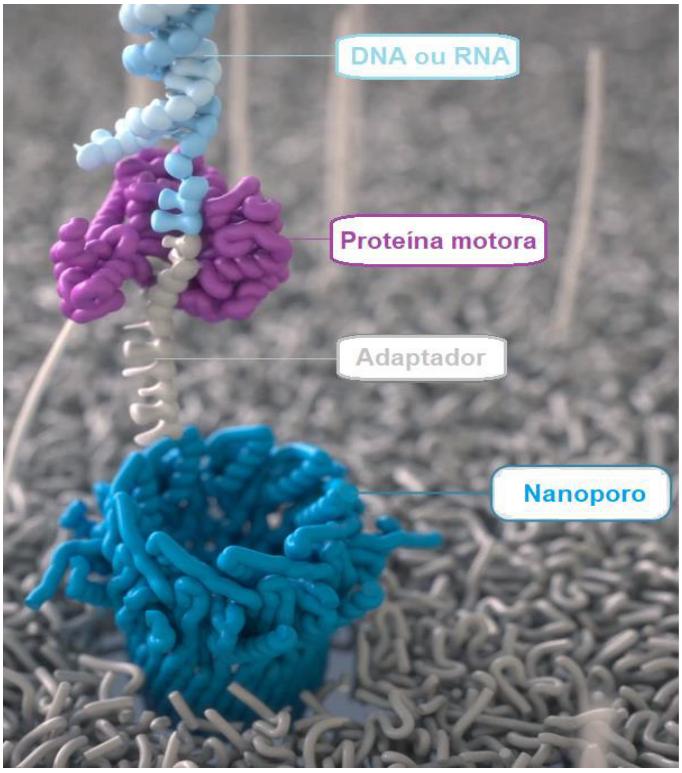

Figura 2. Representação esquemática dos componentes do sequenciamento pela tecnologia Oxford Nanopore (Fonte: https://nanoporetech.com/).

Atualmente o sequenciamento da ONT é apenas unidirecional (*1-Directional* - 1D), ou seja, é feito utilizando apenas uma das fitas do DNA. As vantagens desse método são: obtenção de sequências longas, preparo rápido de bibliotecas, portabilidade devido ao seu tamanho e alto rendimento. Em contrapartida, a desvantagem dessa tecnologia é a taxa de erro que é de aproximadamente 11%. Porém, como a maioria destes erros de sequenciamento são aleatórios, com uma alta cobertura é possível se obter uma sequência consenso com acurácia de aproximadamente 99% (VanDijk et al., 2018).

Ao realizar a compra dos sequenciadores ONT, têm-se direito à licença do *software* desenvolvido pela empresa, que é utilizado na tecnologia (MinKNOW). O tempo da permissão dessa licença depende tanto da plataforma adquirida quanto da escolha do pacote do sistema, que pode ser básico (*System Starter Pack*) ou CapEx (Tabela 2). Este último permite que os usuários acessem a plataforma usando fundos de capital garantindo o título do dispositivo. Os preços referentes ao preparo de biblioteca, ao tipo de *flow cell* e ao sequenciamento propriamente dito também diferem entre as plataformas (Tabela 4).

Tabela 3. Características gerais dos pacotes comercializados pela Oxford Nanopore Technologies (Fonte: https://nanoporetech.com).

Pacote	Característica	Plataforma					
		Flongle	MinION Mk1B	MinION Mk1C	GridION Mk1	PromethION 24	PromethION 48
Básico	Custo do equipamento (US$)	1.860	1.000	4.900	49.995	165.000	285.000
	Nº de *flow cells*	12	2	6	48	60	120
	Tempo de licença dos *softwares*	-	-	12 meses	4 meses	4 meses	4 meses
Estendido	Custo do equipamento (US$)	5.052	4.500	9.995	95.455	297.000	595.000
	Nº de *flow cells*	48	8	8	48	60	120
	Tempo de licença dos *softwares*	-	-	3 anos	3 anos	3 anos	3 anos

Tabela 4. Preços (US$) mínimos referentes ao preparo de bibliotecas, das *flow cells* e do sequenciamento propriamente dito pelas plataformas da Oxford Nanopore Technologies (Fonte: https://nanoporetech.com).

Itens	Flongle	MinION Mk1B	MinION Mk1C	GridION Mk1	PromethION 24	PromethION 48
Preparo da biblioteca	25	99	99	99	99	99
Flow cell	90	475	475	475	625	625
Sequenciamento (por Gb)	45	15	15	15	2	2

A tecnologia de sequenciamento da Oxford Nanopore vem sendo utilizada para genotipagem de *amplicons*. Cornelis et al. (2017) avaliaram a aplicabilidade da plataforma na realização de genotipagem forense. Utilizando o DNA controle (9947A) e o teste SNP-plex, apenas um loco dos 52 avaliados não foi genotipado corretamente quando foram evitados sítios SNPs em regiões homopoliméricas. O sequenciamento durou 24 horas

gerando 367.920 sequências bidirecionais (2D) de alta qualidade com comprimento médio de 625 pb. A taxa média de mapeamento às sequências de referência foi de 60%. Os autores consideraram, no entanto, que quando se comparam os custos, o tempo de análise e a taxa de erro, a tecnologia de sequenciamento da Oxford Nanopore ainda é inferior para genotipagem forense em relação ao sequenciamento Illumina (Cornelis et al., 2017).

2.3 GENOTIPAGEM POR SEQUENCIAMENTO

O surgimento das tecnologias NGS propiciou o advento das tecnologias de genotipagem por sequenciamento (*Genotyping By Sequencing* – GBS). Existem duas abordagens para os métodos GBS: ressequenciamento do genoma completo (*Whole-Genome Resequencing* - WGR) em que se realiza o sequenciamento de todo o genoma, e sequenciamento de representação reduzida (*Reduced-Representation Sequencing* - RRS) em que são sequenciadas apenas regiões específicas do genoma obtidas por redução de complexidade (Tabela 5).

Os métodos de sequenciamento completo do genoma possuem maior resolução. Porém, quando aplicados a genomas eucarióticos, ainda são bastante onerosos. Para a análise eficiente dos dados de sequenciamento de genomas completos se faz necessária a utilização de uma alta cobertura de sequenciamento por indivíduo, devido à existência dos erros de sequenciamento. Isto, em geral, inviabiliza a utilização dos métodos WGR na maior parte das aplicações em escala populacional. Consequentemente, técnicas de genotipagem baseadas no sequenciamento de representações reduzidas de genomas são amplamente utilizadas. Estas técnicas realizam a redução da complexidade do genoma por meio de estratégias que utilizam enzimas de restrição, captura de alvos específicos pela utilização de sondas, amplificação de alvos específicos por PCR, ou ainda o sequenciamento exclusivo da fração transcrita do genoma (RNAseq) (Davey et al., 2011; Ali et al., 2016; Scheben et al., 2017).

Os métodos que adotam a redução da complexidade do genoma utilizando enzimas de restrição têm a vantagem de não exigirem o conhecimento prévio do genoma a ser analisado (Andrews et al., 2016). No entanto, não é possível se determinar com precisão os SNPs que serão efetivamente avaliados pois, com as coberturas tipicamente utilizadas nestes estudos, o efeito da amostragem realizada durante o processo de redução da complexidade do genoma não garante o sequenciamento de todos os sítios de interesse em

todos os indivíduos. Por outro lado, a existência de polimorfismo no sítio de restrição pode fazer com que determinada região alvo seja sequenciada em alguns indivíduos e em outros não. Este aspecto dificulta a realização de estudos populacionais em que a comparação dos genótipos de diferentes indivíduos para os mesmos locos é necessária. As abordagens de genotipagem por sequenciamento mais utilizadas são RADseq, DArTseq e GBS de Cornell e se baseiam no uso de enzimas de restrição para redução da complexidade dos genomas analisados.

Tabela 5. Métodos de genotipagem por sequenciamento descritos na literatura. Adaptado de Scheben et al. (2017).

Abordagem	Métodos	Referências
Reduced Representation Sequencing (*RRS*)	*Target sequencing*	
	PCR amplification	Bevan et al. (1992)
	Selective circularization	Dahl et al. (2005)
	Hybrid Capture	Hodges et al. (2007)
	Targeted Amplicon Sequencing (TAS)	Bybee et al. (2011)
	RAD	
	Restriction site-associated DNA sequencing	Baird et al. (2008)
	(RADseq) Double-digest RAD sequencing (ddRAD)	Peterson et al. (2012)
	2bRAD	Wang et al. (2012)
	ezRAD	Toonen et al. (2013)
	Triple-digest RAD sequencing (3RAD)	Graham et al. (2015)
	RAD Capture (Rapture)	Ali et al. (2016)
	RAD Capture (RADcap)	Hoffberg et al. (2016)
	GBS	
	Genotyping-by-sequencing (Elshire GBS)	Elshire et al. (2011)
	Two-enzyme GBS	Poland et al. (2012)
	Outros	
	Complexity reduction of polymorphic sequences (CRoPS)	Van Orsouw et al. (2007)
	Reduced-representation library (RRL)	Van Tassell et al. (2008)
	Multiplexed shotgun genotyping (MSG)	Andolfatto et al. (2011)
	Sequence-based genotyping (SBG)	Truong et al. (2012)
	Restriction fragment sequencing (RESTseq)	Stolle & Moritz (2013)
	Specific length amplified fragment sequencing (SLAF-Seq)	Sun et al. (2013)
	DArTseq	Ren et al. (2015)
Whole Genome Resequencing (*WGR*)	*Sliding window WGR*	Huang et al. (2009)
	Parental inference WGR	Xie et al. (2010)
	Parental inference WGR with individualized mode	Rowan et al. (2015)
	Skim genotyping-by-sequencing (SkimGBS)	Bayer et al. (2015)

A metodologia 3RADseq consiste nas seguintes etapas: digestão do DNA genômico e a ligação dos adaptadores P1 e P2 (cada um com um *barcode*), seguida da

execução de um único ciclo de amplificação, em que se adiciona um *mix* de *primers* iTru5-8N com oito nucleotídeos degenerados (65.536 diferentes combinações). Posteriormente é realizada uma amplificação que tem como objetivo aumentar a quantidade de fragmentos disponíveis para sequenciamento. Como RADcap utiliza tanto enzimas de restrição, como captura de fragmentos, a segunda fase dessa metodologia é feita por meio de sondas biotiniladas (complementares a fragmentos específicos de RAD *tags*) e *beads* magnéticas revestidas com estreptavidina para captura dos fragmentos para posterior sequenciamento (Miller et al., 2007; Ali et al., 2016; Hoffberg et al., 2016).

O método DArTseq é feito por meio das etapas: digestão do DNA com três enzimas de restrição; ligação dos adaptadores (comum e com *barcode*); amplificação dos fragmentos; purificação e quantificação dos fragmentos; sequenciamento em plataformas NGS; alinhamento a um genoma de referência. Nessa tecnologia, a redução da complexidade do genoma é feita visando a seleção inteligente da porção do genoma que corresponde majoritariamente a genes ativos. Por meio da utilização de uma combinação de enzimas de restrição sensíveis à metilação, separam-se sequências com poucas cópias de sequências repetitivas do genoma (Sansaloni, 2012) (https://www.diversityarrays.com/).

Já na metodologia *two-enzyme* GBS as etapas são: digestão do DNA com duas enzimas de restrição (uma de corte raro e uma de corte frequente); ligação de adaptadores (um complementar à extremidade gerada pela enzima de restrição e que possui o *barcode*, e outro adaptador comum compatível apenas com a extremidade feita pela enzima de restrição); purificação por meio de coluna de exclusão por tamanho; PCR feita com intuito de aumentar o *pool* amostral; e sequenciamento. Esta técnica permite a identificação de dezenas de milhares de marcadores moleculares SNPs (Poland; & Rife, 2012; Poland et al., 2012).

Assim como os métodos que utilizam enzimas de restrição para reduzir a complexidade do genoma, os métodos baseados em RNAseq não exigem o conhecimento prévio do genoma. Essa abordagem permite a obtenção de transcritos ainda não caracterizados, sendo uma alternativa para organismos não-modelo. Contudo, a abordagem baseada em RNAseq, por acessar somente a fração transcrita do genoma, é limitada quanto à abundância dos transcritos, que é dependente dos níveis de expressão de cada gene nas diferentes células conforme o tempo de desenvolvimento e/ou o ambiente/situação (Scheben et al., 2017; Wang et al., 2010).

Os métodos de genotipagem por sequenciamento que realizam a redução da

complexidade do genoma pela utilização de sondas, genericamente denominados de sequenciamento por captura, determinam de maneira mais específica as regiões que serão sequenciadas. As regiões alvo são escolhidas *a priori* e a sequência das sondas ou *primers* (no caso do sequenciamento de *amplicons*) determina as regiões que serão capturadas e sequenciadas. Dessa maneira, para a maioria dos indivíduos, são recuperados genótipos para os mesmos SNPs de interesse. O uso destas estratégias, no entanto, requer o conhecimento prévio do genoma sob estudo, o que implica em um aumento dos custos envolvidos, pois as regiões a serem capturadas devem ser pré-estabelecidas para o desenho das sondas ou *primers*. É importante ressaltar que em estudos de genômica de populações, normalmente, há necessidade de se utilizar um grande número de marcadores.

Onda et al. (2018) avaliaram a capacidade do método *Multiplex PCR targeted amplicon sequencing* (MTA-seq) por meio de um estudo comprobatório em que foi feita a genotipagem de *Brachypodium distachyon*. A análise se deu através da comparação dos resultados obtidos com MTAseq com dados de ressequenciamento do genoma de referência (Bd21), resultando na correta chamada de SNPs em aproximadamente 95% dos casos.

No campo agronômico, os métodos *target sequencing* podem ser aplicados para inferência filogenética, em estudos de estabilidade genética de organismos geneticamente modificados (OGMs), no desenvolvimento de painéis de marcadores SNPs, os quais podem ser utilizados para análises populacionais, mapeamento genético, seleção assistida, melhoramento molecular e diagnósticos moleculares (Azhakanandam et al., 2007; Fonseca & Lohmann, 2018; Boutigny et al., 2019; Onda et al., 2018; Sato et al., 2019).

2.4 APLICAÇÕES DE MARCADORES MOLECULARES EM TESTES EMPREGADOS EM ANÁLISES DE GENÉTICA FORENSE

Os marcadores moleculares são utilizados para identificar polimorfismos em sequências homólogas, são ferramentas que permitem fazer a distinção de indivíduos. Comumente são utilizados dois tipos de marcadores moleculares: os marcadores conhecidos como sequências simples repetidas (*Simple Sequence Repeats* - SSR), baseados no polimorfismo de regiões denominadas de microssatélites, e os polimorfismos de um único nucleotídeo (*Single Nucleotide Polymorphisms* - SNP) (Kayser & De Knijff, 2011). A classificação dos marcadores SSR se dá pelo tamanho dos motivos que se repetem, que podem apresentar de 1 a 6 nucleotídeos (correspondentes às denominações de mono-, di-,

tri-, tetra-, penta- e hexanucleotídeos). Já os marcadores SNPs possuem esse nome devido ao fato de que o polimorfismo é observado na alteração de um único par de bases no DNA.

Os marcadores SSR estão sujeitos a uma taxa de mutação mais elevada do que os SNPs, mas ambos são codominantes, ou seja, conseguem distinguir os indivíduos heterozigotos dos indivíduos homozigotos, são abundantes e tendem a ser distribuídos em todo o genoma. Embora possam ser encontrados tanto em regiões gênicas como em regiões intergênicas, geralmente, estes marcadores são mais prevalentes em regiões não codificantes. Em regiões codantes a ocorrência de polimorfismo pode modificar a expressão gênica e alterar a sequência de aminoácidos codificada. Caso essas modificações sejam deletérias elas tendem a ser eliminadas pela seleção natural.

Como os marcadores SSR são repetições em tandem, evolutivamente, o tamanho do fragmento aumenta ou diminui tipicamente devido ao "escorregamento" da DNA polimerase no momento da replicação, embora esta variação também possa resultar de mutações como retrotransposições ou *crossing over* desigual. Assim, o polimorfismo destes marcadores resulta do número diferente de repetições do motivo do microssatélite que ocorre em cada alelo, fazendo com que estes marcadores sejam tipicamente multialélicos. Essa é a principal vantagem dos marcadores SSR. Assim, mesmo com o aumento de interesse nas últimas décadas pelos marcadores SNP, os marcadores SSR atualmente ainda são os mais utilizados no campo forense (Turchetto-Zolet et al., 2017; Cornelis et al., 2019).

O uso do polimorfismo das regiões microssatélites como marcadores moleculares é possível dado que geralmente as regiões que flanqueiam estes elementos são conservadas entre indivíduos de uma mesma espécie. Por outro lado, para amplificação destas regiões se faz necessário o conhecimento prévio das sequências franqueadoras para o desenho de *primers*, o que acarreta em custos adicionais para o seu desenvolvimento e utilização. Além disso, é possível que ocorram mutações nessas regiões, o que impediria a amplificação do microssatélite, o que acarretaria em um resultado errôneo de ausência de polimorfismo quando na realidade ele estaria presente.

Uma outra característica dos marcadores SSR é que, geralmente, as análises genéticas são baseadas na utilização de múltiplos locos. Assim, é comum o uso de pares de *primers* em reações *multiplex*, possibilitando que mais de um loco seja amplificado em uma mesma PCR. As características de cada par de *primer* e as possíveis interações entre eles, podem tornar a multiplexação uma tarefa complicada – o que representa uma desvantagem destes marcadores. Como os marcadores SNP se baseiam em polimorfismos de um único

nucleotídeo, e como atualmente é possível se realizar o sequenciamento e a genotipagem simultaneamente, a multiplexação não costuma ser problema no caso dos marcadores SNP.

As alterações de um único nucleotídeo dos marcadores SNP podem ser originadas de mutações pontuais, como transições (substituições entre purinas ou entre pirimidinas) e transversões (substituições entre uma purina e uma pirimidina e vice-versa). Em pesquisas genômicas, de modo geral, são também considerados como SNP, os polimorfismos do tipo *indel*, que são regiões onde o polimorfismo se deve à inserção ou deleção de nucleotídeos (Turchetto-Zolet, et al.; 2017).

Os marcadores SNP possuem vantagens em relação aos marcadores SSR quando se trata de amostras de DNA altamente fragmentado (fragmentos com menos de 50 pb), uma vez que o polimorfismo resulta de alterações em apenas uma base. No entanto, a principal desvantagem dos marcadores SNP em relação aos SSR, é que como os SNPs são baseados em polimorfismos de um único nucleotídeo, esses marcadores são menos polimórficos e consequentemente são menos informativos do que os SSR. Embora os marcadores SNP possam ter quatro alelos possíveis, geralmente eles são bialélicos. A ocorrência de uma única modificação em uma base é mais frequente do que a ocorrência de duas ou mais mutações persistentes em um mesmo sítio do genoma.

Por isto, normalmente, os marcadores SNP necessitam de um maior número de locos do que os marcadores SSR para que sejam produzidos resultados com precisões semelhantes. Assim, de modo geral, são necessários cerca de cinco SNPs para cada loco SSR para se atingir um mesmo nível de precisão. Ao se trabalhar com haplótipos, ou pelo aumento do número de marcadores SNP e uso concomitante de tecnologias de genotipagem multiplex, é possível se minimizar esse obstáculo (Kayser & De Knijff, 2011).

Os marcadores moleculares são amplamente utilizados em análises de genética forense, a qual utiliza ferramentas de genética e biologia molecular a fim de que sejam resolvidas questões jurídicas, como por exemplo, em casos em que se demanda a comprovação de paternidade. Em testes de paternidade, é possível se determinar os perfis genotípicos dos indivíduos envolvidos e assim, com o auxílio da estatística, se avaliar as hipóteses de paternidade e de não-paternidade. Todavia, para se certificar uma plataforma de genotipagem altamente eficiente, é preciso se atingir uma probabilidade de exclusão de paternidade *a priori* igual ou maior que 99,9% (Sarmento, 2006). Para realizar o teste de paternidade, normalmente, as hipóteses adotadas são: H_p $(+)$, que geralmente refere-se a alegação de que o suposto genitor é de fato o genitor, e H_d ("hipótese da defesa") $(-)$ que

refere-se a alegação contraditória a H_p (Evett & Weir, 1998; Kayser & De Knijff, 2011).

As análises forenses em seres humanos utilizando marcadores moleculares são tipicamente feitas a fim de se identificar suspeitos de crimes, identificar um possível pai biológico de uma criança e identificar vítimas de desastres (Weir, 1996). Já no contexto de plantas, os testes aplicados em genética forense podem ser utilizados em diferentes estudos agronômicos, como: identificação de genitores e de indivíduos com genótipos superiores; identificação de clones; análise de fluxo gênico por meio da dispersão de pólen; análises de estruturação de subpopulações e determinação do sistema de cruzamento de uma população (Melo, 2000; Jones et al., 2008; Rodrigues et al., 2016).

Em testes de paternidade, geralmente são calculadas duas estatísticas: o índice de paternidade e a probabilidade de paternidade. Tais valores são associados com as frequências alélicas populacionais dos locos sob análise, e estas frequências alélicas devem ser previamente estimadas. O índice de paternidade, PI, e a probabilidade de paternidade, $P(+/G)$, podem ser descritos no contexto da abordagem Bayesiana (Evett & Weir, 1998; Buckleton et al., 2005) como:

$$\frac{P(+/G)}{P(-/G)} = \frac{P(G/+)}{P(G/-)} \cdot \frac{P(+)}{P(-)}$$

em que G é a evidência genética, neste caso representada pelos genótipos dos indivíduos relacionados ao teste, $+$ representa a hipótese de paternidade e $-$ a hipótese de não-paternidade.

O índice de paternidade (PI) representa a razão entre a verossimilhança da hipótese do suposto genitor de fato ser o genitor, e a verossimilhança da hipótese do suposto genitor não ser o genitor (Weir, 1996; Buckleton et al., 2005):

$$PI = \frac{P(G/+)}{P(G/-)}$$

2.4.1 Probabilidade de Exclusão

A probabilidade de exclusão é um parâmetro comumente estimado para um conjunto de marcadores genéticos, e se refere à capacidade de se excluir um indivíduo aleatório de ser genitor, quando de fato este não é o genitor, dadas as frequências alélicas populacionais dos marcadores. Logo, a probabilidade de exclusão depende das frequências alélicas dos locos analisados e depende do número de locos considerados na análise (Weir,

1996).

Suponha um organismo diploide em que o genitor feminino tenha o genótipo A_1A_1 para o loco A, e a prole tenha este mesmo genótipo. Desta forma, logicamente, excluem-se todos os possíveis genitores masculinos que não tenham ao menos um alelo A_1. A proporção de genitores masculinos que não têm esse alelo é de $(1-p)^2$. De modo análogo, outras probabilidades de exclusão podem ser calculadas, conforme apresentado na Tabela 6 (Evett & Weir, 1998).

Tabela 6. Probabilidade de exclusão de paternidade em um determinado loco com *n* alelos. Adaptado de Evett & Weir (1998).

Genitor feminino		Progênie		Genitores masculinos excluídos	
Genótipo	Prob.	Genótipo	Prob.	Genótipos	Prob.
A_iA_i	p_i^2	A_iA_i	p_i	$A_wA_x, w,x \neq i$	$(1-p_i)^2$
		A_iA_j	p_j	$A_wA_x, w,x \neq j$	$(1-p_j)^2$
A_iA_j $j \neq i$	$2p_ip_j$	A_iA_i	$p_i/2$	$A_wA_x, w,x \neq i$	$(1-p_i)^2$
		A_jA_j	$p_j/2$	$A_wA_x, w,x \neq j$	$(1-p_j)^2$
		A_iA_j	$(p_i+p_j)/2$	$A_wA_x, w,x \neq i,j$	$(1-p_i-p_j)^2$
		A_iA_k	$p_k/2$	$A_wA_x, w,x \neq k$	$(1-p_k)^2$
		A_jA_k	$p_k/2$	$A_wA_x, w,x \neq k$	$(1-p_k)^2$

O cálculo da probabilidade de exclusão é dado pela soma das probabilidades conjuntas de todas as possíveis combinações alélicas do genitor feminino-prole-suposto genitor masculino, de modo que a exclusão da paternidade se dá quando a prole tem alelos que o suposto genitor masculino não tem. A probabilidade de exclusão para múltiplos locos pode ser calculada por (Weir, 1996):

$$Q = 1 - \prod_l (1 - Q_l)$$

em que Q é a probabilidade de exclusão multiloco e Q_l é a probabilidade de exclusão do loco l.

Dessa forma a probabilidade de exclusão em múltiplos locos é 1 menos a probabilidade de que nenhum dos locos analisados permita a exclusão do genitor, admitindo

independência entre os locos (Weir, 1996).

Os problemas relacionados à exclusão de paternidade são devido à não conclusão de paternidade quando o verdadeiro genitor masculino não está na análise, ou quando dois ou mais genitores masculinos tenham alelos compatíveis com a prole em análise. No segundo cenário, é necessário se aumentar a quantidade de locos para se aumentar o poder de exclusão.

2.4.2 Análise de Identidade Genética

A análise de identidade genética se refere à comparação dos genótipos de duas amostras supostamente independentes, para se verificar se há evidência de que se tratam na verdade de amostras de um mesmo indivíduo ou não. Assim como ocorre na análise de paternidade, a estatística de teste de interesse neste caso, é dada por uma razão de verossimilhança.

Neste contexto, resultam duas possíveis situações, ou os genótipos sob análise são idênticos, ou não são. A título de exemplo, suponha que se deseja averiguar as hipóteses de que dois indivíduos de uma determinada espécie vegetal arbórea tenham sido obtidos por clonagem ou não. As hipóteses de interesse neste caso são: hipótese $(+)$, os indivíduos (R e C) são réplicas de um mesmo clone; hipótese $(-)$, os indivíduos não são réplicas de um mesmo clone; em que R é o genótipo de referência e C é um possível clone.

Geralmente, assume-se equilíbrio de Hardy-Weinberg e segregação independente para a análise. A evidência estatística neste caso é obtida pela razão entre a verossimilhança da hipótese de identidade genética dos indivíduos por clonagem e a verossimilhança da hipótese de identidade genética dos indivíduos por mero acaso (Weir, 1996):

$$LR = \frac{P[(R = G, C = G)/+]}{P[(R = G, C = G)/-]}$$

Uma vez que a probabilidade de R ter genótipo G é independente das hipóteses $+$ e $-$, esta razão de verossimilhança pode ser apresentada como (Weir, 1996):

$$LR = \frac{P[C = G/(R = G, +)]}{P(C = G)}$$

Quando se ignoram a possibilidade de determinações incorretas dos perfis de R e C e o erro experimental, o numerador de LR será igual a 0 ou a 1. Se o numerador for igual

a 0, o resultado final de *LR* será 0 e conclui-se que os indivíduos são geneticamente diferentes o que anula a possibilidade de que sejam réplicas de um mesmo clone. Caso o numerador de *LR* seja 1, calcula-se a probabilidade de que *R* e *C*, apesar de serem indivíduos diferentes, apresentarem o mesmo genótipo por acaso. Uma vez obtida a estatística *LR*, no contexto Bayesiano:

$$P(+/\mathrm{G}) = \frac{LR \cdot P(+)}{LR \cdot P(+) + [1 - P(+)]}$$

representa a probabilidade de que os indivíduos geneticamente idênticos sejam réplicas de um mesmo clone (Weir, 1996).

3 MATERIAL E MÉTODOS

3.1 EXTRAÇÃO DO DNA GENÔMICO

Foram coletados 47 indivíduos de diferentes clones experimentais de cana-de-açúcar provenientes do campo experimental do Programa de Melhoramento da Ridesa, desenvolvido na Escola de Agronomia/UFG. Como controle, o indivíduo RB006819 foi analisado em duplicata técnica (RB006819a e RB006819b), ou seja, utilizou-se a mesma alíquota de DNA para esses. Dentre os 47 indivíduos foram incluídos ainda dois indivíduos (RB006995x e RB006995y) provenientes de touceiras distintas, mas supostamente réplicas biológicas de um mesmo clone. No total, foram utilizadas 48 amostras de DNA para a realização deste estudo, como apresentado na Tabela 7.

A extração de DNA genômico se fez a partir da maceração de tecidos foliares na presença de nitrogênio líquido. A extração foi conduzida no Laboratório de Genética e Genômica de Plantas, situado na Escola de Agronomia da Universidade Federal de Goiás. O procedimento de extração do DNA genômico teve como base o protocolo Doyle & Doyle (1990). Os *pellets* foram submetidos à secagem à temperatura ambiente e ressuspendidos em água ultrapura.

A integridade do DNA genômico extraído foi avaliada por eletroforese em gel de agarose 1%, comparando-se o peso molecular e o tamanho dos fragmentos de DNA extraídos com um marcador-controle (DNA de fago λ). Os géis foram visualizados em transiluminador e registrados em fotodocumentador L-Pix Ex (*Loccus*). A avaliação do grau de pureza e da concentração do DNA extraído foram realizadas nos equipamentos Nanodrop® Spectrophotometer (*Invitrogen*) e Qubit® Fluorometer Quantitation (*Life Techonologies*), respectivamente.

Tabela 7. Identificação das 48 amostras de DNA de cana-de-açúcar, do programa de melhoramento da Ridesa, utilizadas neste trabalho.

Indivíduos	Indivíduos
RB991536	RB006819b
RB971739	RB037170
RB011549	RB036152
RB041443	RB985476
RB955977	RB987917
RB021754	RB093070
RB058046	RB008310
RB006819a	UFG06109
RB046299	UFG06125
RB106822	RB961552
RB006995x	RB036091
RB056380	RB006655
RB034045	RB061276
RB036088	RB005014
RB975201	RB975375
RB073034	RB036066
RB988082	RB975952
RB867515	RB987935
RB006995y	RB975242
RB961003	RB015935
RB006629	UFG06292
RB056351	RB037194
RB056396	RB083088
RB991536	RB969017

3.2 OBTENÇÃO DOS PRIMERS

Os *primers* utilizados neste trabalho foram desenvolvidos utilizando-se como referência a sequência consenso entre o genoma de referência de *Sorghum bicolor* v3.1.1 (McCormick et al., 2017) e o transcritoma de cana-de-açúcar caracterizado por Melo (2015) (Tabela 8). O sequenciamento das bibliotecas de cDNA foi realizado na plataforma HiSeq2000 (*Illumina*). O transcritoma foi obtido a partir do sequenciamento de amostras de RNA obtidas de 30 indivíduos selecionados aleatoriamente do programa de melhoramento genético em cana-de-açúcar da Ridesa/UFG. Foram sequenciadas dez bibliotecas de cDNA, sendo duas de cada órgão vegetal utilizado (colmo, gemas laterais, plântulas, folhas e gemas apicais) (Melo, 2015).

Os dados de sequenciamento do transcritoma foram alinhados ao genoma de referência utilizando-se o *software* STAR (Dobin et al., 2013). O alinhamento foi visualizado através do programa *Integrative Genomics Viewer* (IGV) (Robinson et al., 2011). Foram selecionadas subjetivamente 20 regiões que continham vários sítios SNPs, sendo uma para

cada um dos braços dos dez cromossomos de sorgo. Os *amplicons* foram identificados de acordo com o cromossomo de sorgo utilizado como referência e seus respectivos braços. De modo que os números 1 e 2 após o a identificação do cromossomo, se referem à regiões iniciais e terminais das sequências de referência, respectivamente. A Tabela 8 apresenta as coordenadas iniciais e finais dos fragmentos para os quais foram desenhados os pares de *primers*, com a finalidade de se obter *amplicons* com tamanhos de 1,5 a 3,0 kb. Os *primers* foram desenhados utilizando-se o programa Primer3Plus (Untergasser et al., 2007). A Tabela 9 exibe as sequências dos *primers* e o tamanho esperado de cada *amplicon*.

Tabela 8. Coordenadas iniciais e finais dos fragmentos utilizados para desenho dos *primers*.

Loco	Posição Inicial (nº da base)	Posição Final (nº da base)
chr01_1	5434372	5437143
chr01_2	78725340	78727922
chr02_1	4160078	4162100
chr02_2	73016381	73018083
chr03_1	3976592	3978590
chr03_2	60287700	60290164
chr04_1	2053727	2056222
chr04_2	62210829	62212864
chr05_1	2075259	2077427
chr05_2	63392946	63394902
chr06_1	1253093	1255547
chr06_2	56317706	56319781
chr07_1	2774635	2776932
chr07_2	60270747	60272716
chr08_1	823079	824679
chr08_2	58376339	58378598
chr09_1	4015727	4017950
chr09_2	55649581	55651912
chr10_1	4475112	4476959
chr10_2	59187660	59189430

Tabela 9. Dados relativos aos *primers* utilizados para obtenção dos *amplicons*.

Loco	Sequência *forward* (5'→ 3')	Sequência *reverse* (5'→ 3')	Tamanho esperado (pb) dos *amplicons*
chr01_1	GTTTGGTAGTGGGGAACCTCCAGA	TGCCTAGCTCTCCACCTTGGGG	2.660
chr01_2	ACAAATGAAGCTGCGCCACAGT	CCCACGCTCCGCAAGCTCAA	2.521
chr02_1	ACTGTCACTTGGCGAGGGGT	TGGCAATGGCCTCATAGCTCGC	1.906
chr02_2	TCATCCCAGGAATTCCCTCTATCAAGC	AGCCGGCATTCTGGGAAATGAGC	1.638
chr03_1	ACTGCCGATCTCAAGCATGTGTCA	TGCCTTGCCTAATGGGGGTGGA	1.751
chr03_2	GTCTTCGCCCGCTCACACCC	AGATTGCTTCAATAGCTGCACAAGACA	2.304
chr04_1	GGGCGCTCGACGCGATCTAC	TGGCTGATACCCCTGCCTGC	2.293
chr04_2	GCTTGTGGGACAGGAGTTGTAGTCC	TGCACAGTAATTTGGATTGCTCCGA	1.914
chr05_1	TGGACGGTTTCACCCTGCCG	TCGAGAAACTGGGCTCGGTGC	1.867
chr05_2	AGACGCCTCTGCGGGGAAGA	TCAGCGGAGGCACTTGCAGC	1.784
chr06_1	ACTCCGACCTCGTCCTGGGC	TGTCACCCAGCTGGAGTTGAGCTT	2.197
chr06_2	CGCATCGCATCCAGTGAATTGGC	ATGCTGCTCGGACATCGGGC	1.946
chr07_1	GCTTCACCACCACCAGGCCC	TCTCTTCTCATCCATCCATGTTGACCA	2.119
chr07_2	AGGTCGCCTCGTCGTCGTCA	CCTGGGCGAAGGGTGTTGGC	1.937
chr08_1	CAATGCGAGACGCTGCCCGA	TGCAGCTTCACTGTGGAGGGGA	1.546
chr08_2	AGGCAAAGCGTTGGCCCTTAGT	GCTTTGTTTGGCCTCTGAAACCGA	2.117
chr09_1	ACAAAGTGTGTTGCATCGGCCA	TCACAGGAGCCTGGTGATCTGTCA	2.011
chr09_2	TGCCTACGGCTTGATCCCAGA	TGTGAGAGTGGTGTTAGCGCGG	2.176
chr10_1	TGCACACCAACCCGACAGCC	CAAGCCGCCGCTCCTGTCTC	1.788
chr10_2	TCGCCGCCAACTCGCTCATC	TGTGCCTAAACCATCAAGCAGAGGA	1.548

3.3 GENOTIPAGEM POR SEQUENCIAMENTO

As bibliotecas de sequenciamento foram construídas a partir dos produtos de PCR, obtidos por amplificação das regiões-alvo utilizando o kit PCR Multiplex (*Qiagen*). Este kit, apesar de não ser indicado para amplificação de fragmentos maiores que 1,5 kb, não demonstrou problemas para amplificação dos fragmentos-alvo, com base nos testes realizados nas duas primeiras bibliotecas de sequenciamento. Em uma primeira etapa, duas das 48 amostras de DNA foram arbitrariamente escolhidas para se realizar reações de PCR amplificando as 20 regiões genômicas individualmente. Em uma segunda etapa, as reações de PCR foram realizadas em *multiplex*, utilizando os 20 pares de *primers* em uma mesma

reação. A integridade dos produtos de amplificação obtidos nas duas etapas foi avaliada por eletroforese em gel de agarose 1%. Os produtos de PCR obtidos na segunda etapa foram também utilizados para construção de uma biblioteca e submetidos ao sequenciamento na plataforma MinION.

Os resultados deste sequenciamento foram utilizados para otimização das combinações dos locos em cinco reações *multiplex* (de acordo com os padrões obtidos de cobertura). Os produtos das cinco reações *multiplex* foram também submetidos ao sequenciamento na plataforma MinION, utilizando-se uma mistura de volumes iguais de cada uma das reações.

A seleção dos pares de *primers* para execução das análises posteriores foi realizada de acordo com esse segundo sequenciamento, eliminando-se os *amplicons* que resultaram em coberturas menores que 1.000X e aqueles em que os fragmentos amplificados apresentaram tamanhos menores do que 400 pb, verificados por meio de alinhamento das sequências obtidas às sequências de referência dos fragmentos utilizados para desenho dos *primers*. Este alinhamento foi realizado pelo programa BLAST (Altschul et al., 1990).

Após a otimização e seleção, o sequenciamento dos *amplicons* obtidos com as 48 amostras de DNA de cana-de-açúcar, foi realizado em quatro bibliotecas com 12 indivíduos cada. Para o sequenciamento desses 48 indivíduos foi utilizada uma nova *flow cell*.

Para a construção das bibliotecas de sequenciamento, os fragmentos amplificados foram inicialmente purificados com o kit *PCR Purification Agencourt AMPure XP* (*Beckman Coulter*). As quatro bibliotecas foram feitas utilizando-se o kit *PCR Barcoding Expansion* 1-12 *(EXP-PBC001)* (*Oxford Nanopore Technologies*) de acordo com o protocolo estabelecido pelo fabricante com uma modificação em uma etapa. Uma vez que o sequenciamento deste trabalho era de fragmentos pequenos, e o protocolo é originalmente descrito para sequenciamento de genomas inteiros, e tendo em vista a redução dos custos de sequenciamento, os volumes dos reagentes da etapa de ligação dos *barcodes* foram reduzidos em 10 vezes. Todos os processos de preparo das bibliotecas e sequenciamento ocorreram no Laboratório de Genética e Genômica de Plantas, localizado na Escola de Agronomia da Universidade Federal de Goiás.

Para todas as bibliotecas de sequenciamento foram utilizadas apenas as sequências que passaram no filtro (fastq_pass) do MinKNOW™ (Oxford Nanopore Technologies). Uma vez que o sequenciamento executado pelo MinION é unidirecional e os

arquivos em formato *fastq* de um mesmo indivíduo são fragmentados em pastas distintas, os arquivos *fastq* de um mesmo indivíduo foram concatenados e submetidos ao diagnóstico de controle de qualidade pelo *software* FastQC v. 0.11.5 (Andrews, 2010). Os módulos *per base sequence quality* e *per sequence quality scores* foram avaliados a fim de se verificar a qualidade das bases e das sequências. Os módulos *Basic Statistics*, *Sequence Length Distribution* e *Per sequence GC content* foram avaliados para apurar as caraterísticas gerais das bibliotecas, tais como o número total de sequências, números de sequências consideradas de baixa qualidade, tamanho das sequências e conteúdo GC.

O *base-calling*, a demultiplexação e a remoção dos adaptadores das sequências foram realizados utilizando o programa MinKNOW™ (Oxford Nanopore Technologies), que por meio do sinal bruto advindo do MinION, detecta e avalia a mudança de sinal com a passagem do DNA pelo nanoporo, fornecendo um arquivo *fastq* com informações das bases, as quais podem ser observadas em tempo real pelo MinKNOW GUI.

Como as tecnologias de sequenciamento de fragmentos longos são mais recentes, ainda existem poucos *softwares* desenvolvidos para lidar com essa abordagem. O algoritmo "*mem*" do alinhador *Burrows-Wheeler Aligner* (BWA) (Li & Durbin, 2009) foi criado para possibilitar alinhamentos de sequências longas. Com isso, neste trabalho, o alinhamento das sequências obtidas à referência (sequências consenso de cada *amplicon*) foi realizado pelo programa BWA. A inspeção da qualidade de sequenciamento foi feita pelo o *software* FastQC (Andrews, 2010), que não foi desenvolvido para lidar com dados de sequenciamentos de sequências longas. A qualidade do alinhamento foi averiguada através de métricas obtidas pela ferramenta "*CollectAlignmentSummaryMetrics*" do programa Picard Tools (Broad Institute - http://broadinstitute.github.io/picard/). A identificação e genotipagem dos SNPs foi realizada pelo *software SAMtools* utilizando-se a ferramenta "*mpileup*" (Li et al.,2009).

3.4 IDENTIFICAÇÃO DOS CLONES DE CANA-DE-AÇÚCAR

As análises genético-estatísticas para fins de identificação dos clones de cana-de-açúcar foram feitas utilizando-se um *script* especificamente construído e executado no *software* R v.3.6.1. O arquivo contendo os dados de genotipagem das 48 amostras de DNA para os 356 SNPs identificados foi inicialmente transformado em um arquivo no formato *fstat*. Em seguida, foram calculadas as frequências alélicas para cada um dos locos. Em

seguida foi realizado o cálculo da distância genética entre os indivíduos utilizando-se a distância euclidiana de Rogers (1972) modificada por Wright (1978). A matriz de distâncias foi então utilizada em uma análise de agrupamento pelo método UPGMA (*Unweighted Pair-Group Method with Arithmetic Averages*) que resultou em um dendrograma. A partir do dendrograma foi estimado o coeficiente de correlação cofenética para avaliação da representatividade do dendrograma obtido. A significância deste coeficiente foi avaliada pelo teste de Mantel (1967). A consistência dos nós do dendrograma foi avaliada por reamostragem *bootstrap* com 10.000 repetições.

4 RESULTADOS E DISCUSSÃO

4.1 CONTROLE DE QUALIDADE DO DNA EXTRAÍDO E OTIMIZAÇÃO DAS REAÇÕES DE PCR

O controle de qualidade foi realizado para cada uma das 48 amostras de DNA extraído. A variação inicial da concentração de DNA foi de 46,3 ng/µL a 508 ng/µL. Todas as alíquotas de DNA genômico extraídas foram visualizadas em gel de agarose e apresentaram valores para a razão entre as absorbâncias a 260 nm e 280 nm entre 1,8 e 2,2, sugerindo ausência de contaminantes como proteínas, fenóis e outros, cuja presença é normalmente associada a valores inferiores aos obtidos.

Quanto à integridade do produto de PCR obtido nos testes de amplificação dos locos utilizando pares de *primers* separadamente (*simplex*) e em conjunto (*multiplex*), não foram detectados sinais de degradação dos fragmentos amplificados.

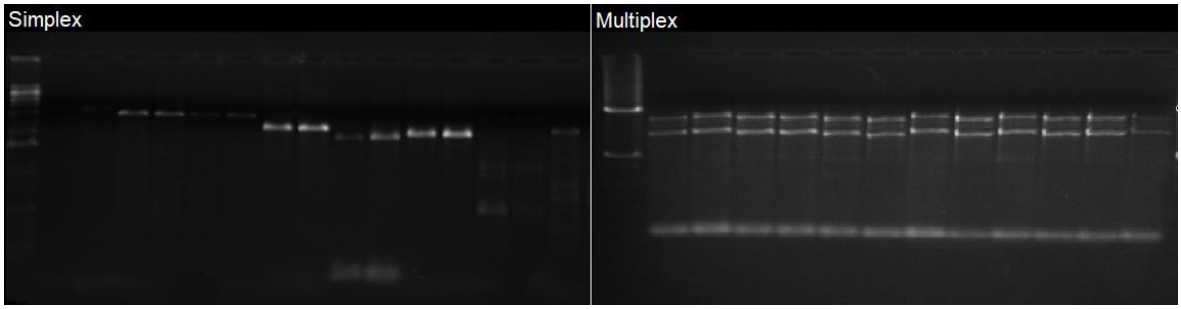

Figura 3. Apresentação dos produtos de PCR obtidos a partir da amplificação dos locos em *simplex* à esquerda e *multiplex* à direita.

O alinhamento das sequências obtidas pelo primeiro sequenciamento (dos produtos de PCR obtidos em *multiplex* com os 20 pares de *primers*) nas sequências de referência (as mesmas utilizadas para o desenho dos *primers*), demonstrou que a cobertura de sequenciamento não foi homogênea para os diferentes *amplicons* (Tabela 10). Como, anteriormente, cada par de *primer* foi testado individualmente e houve amplificação, essa heterogeneidade provavelmente deve-se à competição dos reagentes durante a PCR. As combinações dos pares de *primers* foram então otimizadas com intuito de que todos os locos

apresentassem coberturas semelhantes. Uma segunda etapa de sequenciamento foi realizada para se verificar eficiência desta otimização.

Tabela 10. Comparação das coberturas obtidas antes e após a otimização das reações *multiplex.*

Loco	Antes da otimização (X)	Após a otimização (X)
chr01_1	2	1
chr01_2	0	0
chr02_1	94	1.808
chr02_2	204	100
chr03_1	910	7.999
chr03_2	9	5
chr04_1	29	71
chr04_2	177	1.227
chr05_1	280	2.243
chr05_2	3	7
chr06_1	4	0
chr06_2	44	0
chr07_1	142	814
chr07_2	717	1.493
chr08_1	77	7.903
chr08_2	20	133
chr09_1	26	1.495
chr09_2	0	1
chr10_1	3685	5.153
chr10_2	5983	7.373

Na etapa de otimização foram avaliados cinco grupos de *amplicons* em *multiplex*. O primeiro com os seis locos que tiveram coberturas abaixo de 10X (*chr01_1*, *chr01_2*, *chr03_2*, *chr05_2*, *chr06_1* e *chr09_2*). No segundo foram agrupados os seis locos com coberturas acima de 10X e abaixo de 100X (*chr02_1*, *chr04_1*, *chr06_2*, *chr08_1*, *chr08_2* e *chr09_1*). O terceiro grupo foi construído com os quatros locos que tiveram coberturas acima de 100X e abaixo de 300X (*chr02_2*, *chr04_2*, *chr05_1* e *chr07_1*). Como não houve locos com coberturas entre 300X a 700X, no quarto grupo foram agrupados os locos *chr03_1* e *chr07_2*, os quais tinham coberturas de 910X e 717X, respectivamente. O quinto grupo corresponde aos locos *chr10_1* e *chr10_1*, que ficaram super-representados com coberturas de 3.685X e 5.983X, respectivamente.

O resultado do segundo sequenciamento, após a otimização das reações *multiplex*, demonstrou uma melhora substancial nas coberturas de diversos locos, embora

ainda tenha se detectado uma heterogeneidade que deve ser reduzida em trabalhos futuros de otimização.

Para o sequenciamento das 48 amostras de DNA de cana-de-açúcar, foram selecionados os nove locos que tiveram cobertura acima de 1.000X no segundo sequenciamento. Esses locos foram separados em quatro conjuntos *multiplex* com as seguintes configurações: *chr03_1* e *chr07_2*; *chr08_1* e *chr09_1*; *chr10_1* e *chr10_2*; *chr02_1*, *chr04_2* e *chr05_1*.

4.2 SEQUENCIAMENTO DOS AMPLICONS

No total foram obtidas 841 mil sequências, compreendendo cerca de 1,3 Gb de sequenciamento (Tabela 11). A fim de se reutilizar a *flow cell*, foi calculado o número de bases esperado para que, em média, os 12 indivíduos sob análise tivessem seus *amplicons* sequenciados com cobertura acima de 1.000X. Os sequenciamentos foram então interrompidos após o número total de bases atingir 250 Mb, resultando em um tempo entre três e quatro horas (Tabela 11) para cada rodada de sequenciamento. A produção de novas sequências após 20 horas de sequenciamento com a plataforma MinION é insignificante (Cornelis et al., 2019).

Tabela 11. Estatísticas descritivas dos resultados obtidos no sequenciamento das quatro bibliotecas.

Biblioteca	Número de sequências (x1000)	N50 (kb)	Número total de bases (Mb)	Tempo de sequenciamento (hh:mm)
1	156,16	2,30	248,18	03:55
2	219,55	1,60	353,73	03:43
3	232,68	1,54	363,90	03:35
4	232,81	1,43	358,80	03:21
Total	841,20	-	1.324,61	14:34

Nota-se que a biblioteca 1 foi a que teve maior variação de tamanhos de fragmentos, com 21,2 Mb correspondentes às sequências mais abundantes, as quais tinham aproximadamente 2,0 kb (Figura 4). As bibliotecas 2 e 4 tiveram as distribuições de fragmentos de tamanhos diferentes semelhantes, sendo que a quantidade de bases relacionadas aos fragmentos de tamanhos mais abundantes foram de 36,6 Mb e 33,8 Mb respectivamente. O tamanho mais comum das sequências dessas bibliotecas foi em torno de 1,5 kb (Figuras 5 e 6, respectivamente). A biblioteca quatro foi a que teve menor variação em relação aos diferentes tamanhos de fragmentos, com 62,9 Mb referentes aos tamanhos mais comuns de sequências, que também apresentaram aproximadamente 1,5 kb (Figura 7). Esse resultado coincide com o obtido ao realizar as estatísticas descritivas do sequenciamento (Tabela 13, página 47) em que as médias do tamanho dos fragmentos foi cerca de 1,5 kb.

Essas diferenças dos tamanhos das maiores sequências, não era esperada, uma vez que todas as bibliotecas foram feitas utilizando os mesmos pares de *primers* na etapa de amplificação. O esperado era que os picos desses histogramas tivessem tamanhos concordantes com o estimado na etapa de desenho dos *primers* (Tabela 9, página 34). Apesar de pouco provável, talvez essa discordância aconteceu devido a degradações das extremidades das sequências amplificadas multiplexadas e armazenadas a -20 ºC. Todas as bibliotecas foram feitas, teoricamente, da mesma maneira, diferenciando-se apenas no tempo entre a obtenção dos *amplicons* e o preparo das bibliotecas. A biblioteca 1 foi a primeira a ser feita. A diferença do tempo entre o sequenciamento dessa para a segunda foi de duas semanas e um dia. Após o sequenciamento da segunda, levou-se seis dias para o sequenciamento da terceira. Com cinco dias posteriores foi feita a quarta.

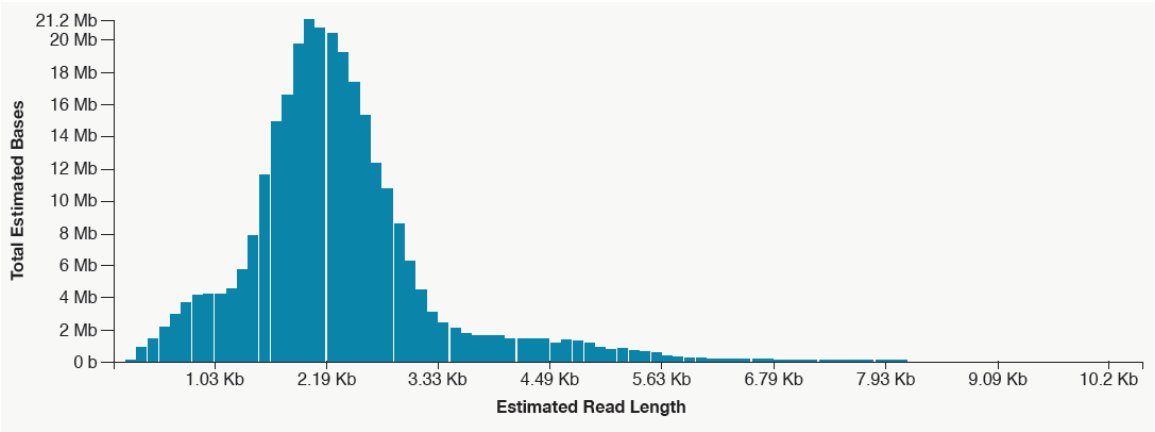

Figura 4. Distribuição dos tamanhos dos fragmentos sequenciados na biblioteca 1.

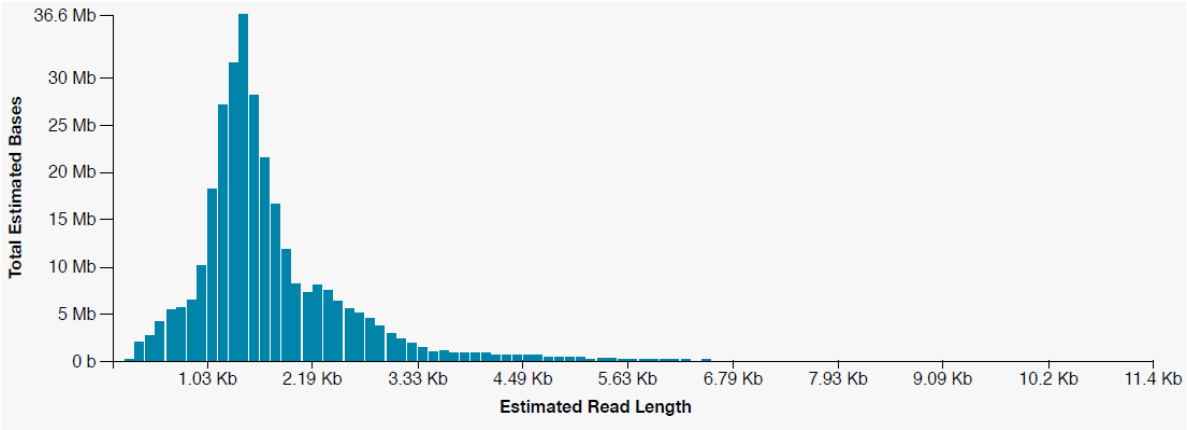

Figura 5. Distribuição dos tamanhos dos fragmentos sequenciados na biblioteca 2.

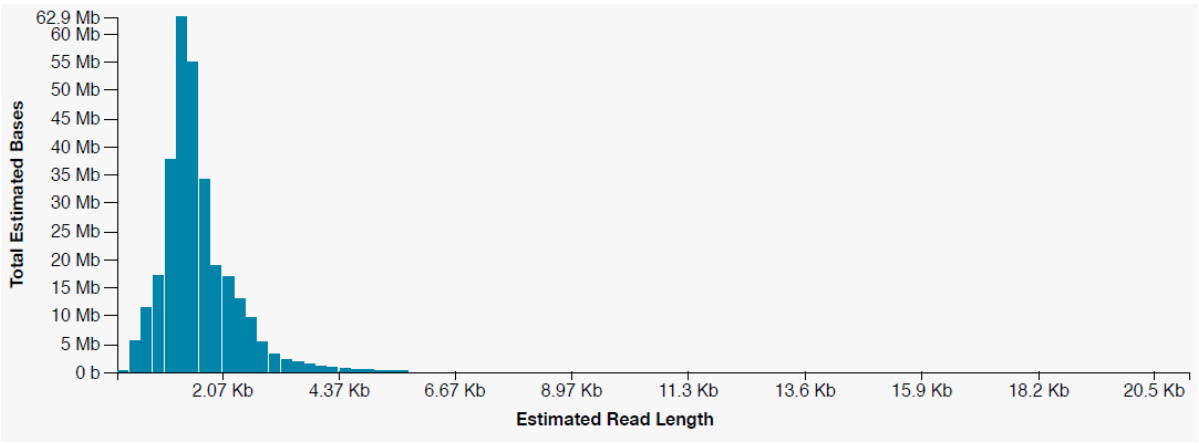

Figura 6. Distribuição dos tamanhos dos fragmentos sequenciados na biblioteca 3.

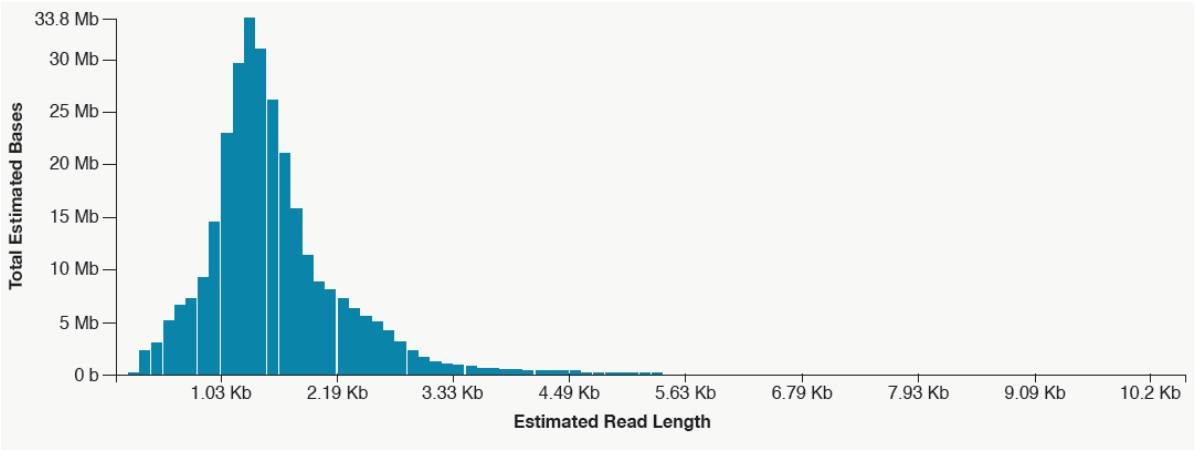

Figura 7. Distribuição dos tamanhos dos fragmentos sequenciados na biblioteca 4.

A inspeção do controle de qualidade dos dados brutos obtidos no sequenciamento, realizado pelo FastQC, revelou resultados semelhantes para todos os indivíduos. É valido lembrar que o FastQC foi desenvolvido para lidar com dados advindos de tecnologias que produzem sequências menores com uma pequena taxa de erro, opostamente aos dados obtidos pela tecnologia Oxford Nanopore.

De acordo com os relatórios do FastQC, todos os 48 indivíduos apresentaram valores para qualidade de sequenciamento por cada base (*per base sequence quality*), bem como por sequência (*per sequence quality score*) entre 14 e 16 (Figuras 8 e 9). As frequências estimadas de cada base (A, G, C e T) (*per base sequence content*) e para o conteúdo GC (*per sequence GC content*) por sequência revelaram-se variáveis (Figuras 10 e 11 respectivamente). Deve-se levar em consideração que neste estudo foram sequenciados apenas fragmentos do genoma da cana-de-açúcar, logo as frequências estimadas das bases assim como o conteúdo GC não representam bem o genoma completo.

Tais resultados se devem provavelmente à alta taxa de erro e a possíveis vieses da tecnologia de sequenciamento utilizada. Apesar dos erros das plataformas ONT serem predominantemente aleatórios, existem pelo menos dois vieses intrínsecos: um relacionado a substituições e outro a deleções. O primeiro é recorrente com as bases guanina e citosina, resultando em um enriquecimento irreal na frequência dessas bases nos dados gerados. O segundo está presente com maior frequência nas bases adenina e timina (Magi et al., 2017).

Ainda sobre os resultados do FastQC, não foram detectados problemas relacionados ao conteúdo de bases indefinidas (*per base N content*) e níveis de sequências duplicadas (*sequence duplication levels*).

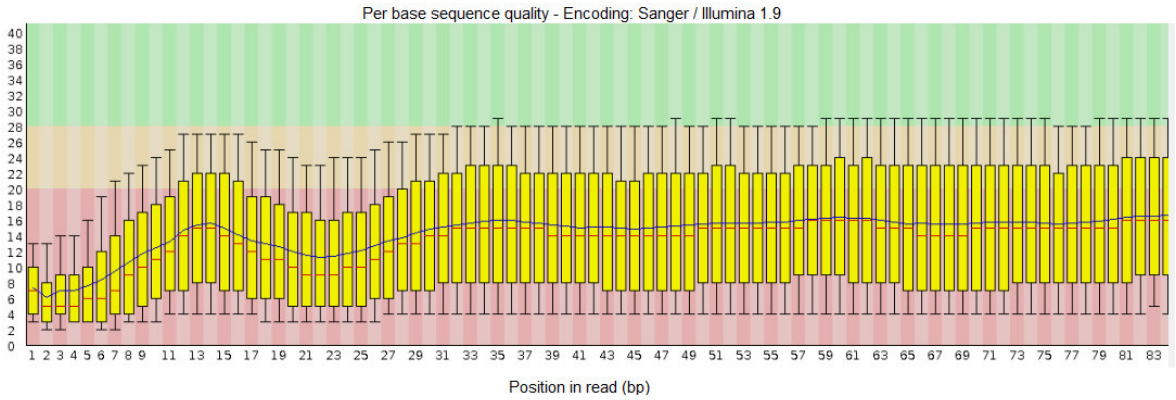

Figura 8. Representação da distribuição de qualidade das sequências obtidas, por base.

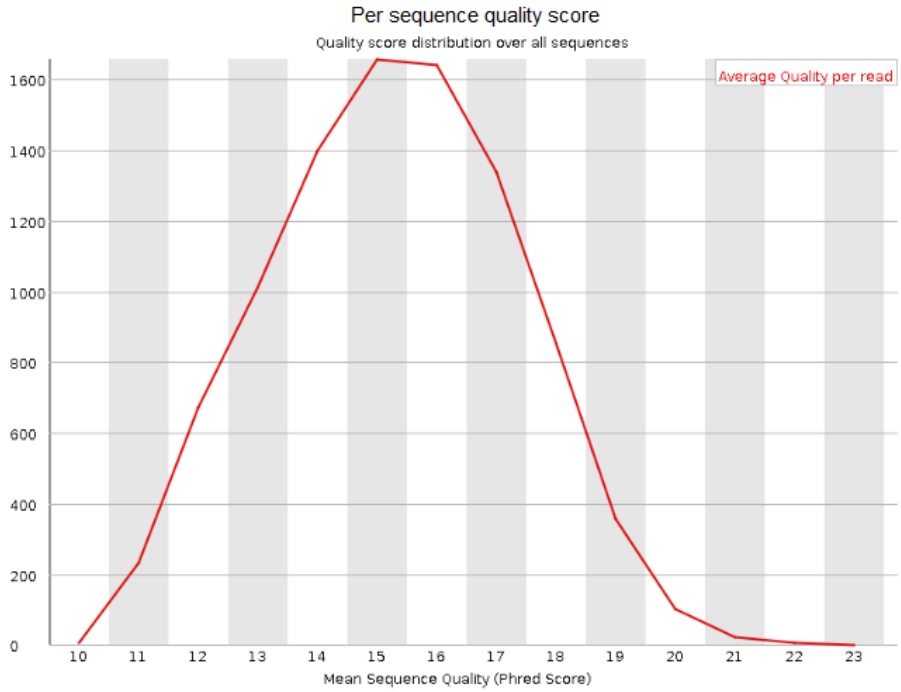

Figura 9. Representação da distribuição de qualidade média das sequências obtidas.

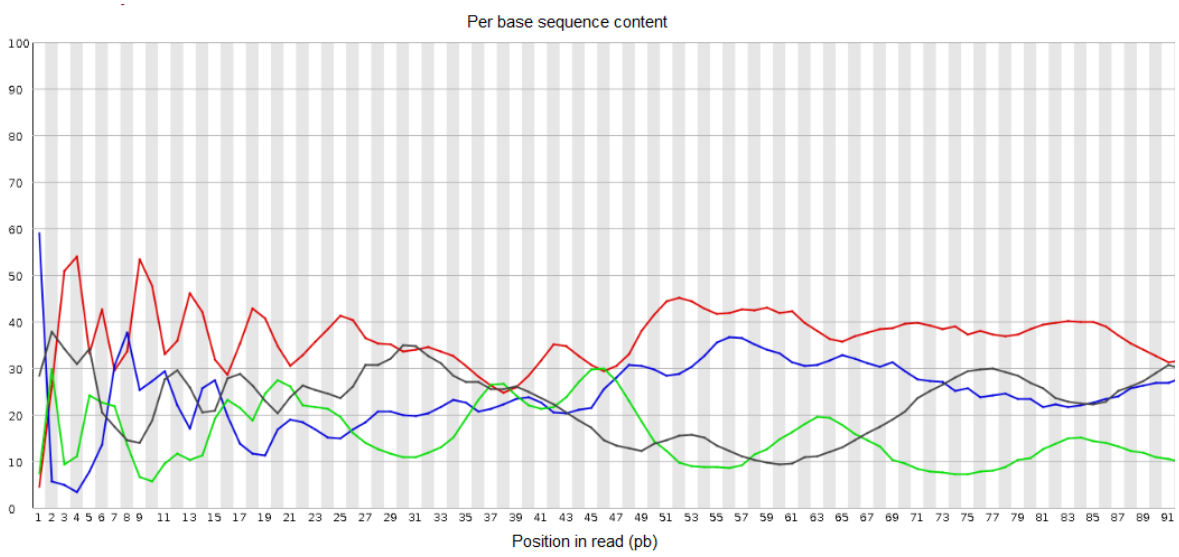

Figura 10. Representação da distribuição do conteúdo de cada base nas sequências obtidas.

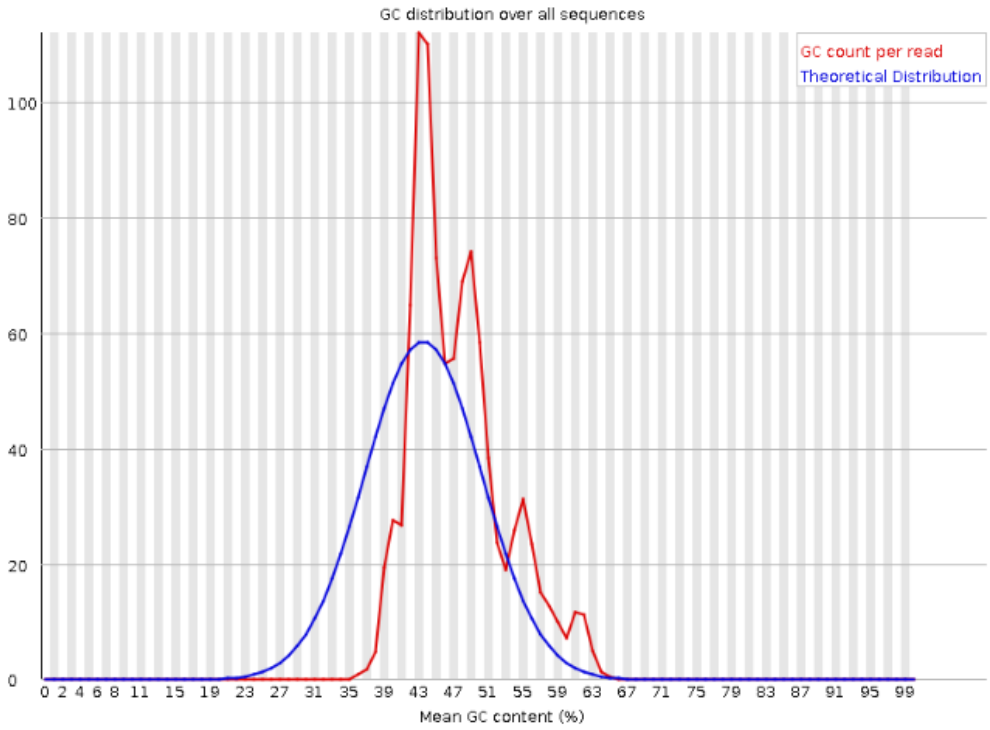

Figura 11. Distribuição do conteúdo GC nas sequências obtidas.

Pode-se observar na Tabela 12 que as sequências obtidas para cada indivíduo apresentaram variações em tamanho em relação ao esperado para cada *amplicon* (Tabela 9, página 34). Os fragmentos menores podem ser resultantes de degradação e os maiores podem resultar de concatenação de fragmentos e contaminantes de adaptadores. Estes fragmentos não são prejudiciais pois são eliminados na etapa de alinhamento. Observa-se ainda na Tabela 12 que o esforço de sequenciamento não foi homogêneo entre indivíduos. Tal fato pode ser explicado por erros na quantificação e normalização dos produtos de PCR utilizados na construção das bibliotecas, já que durante o sequenciamento a seleção de fragmentos para entrada no nanoporo é aleatória.

Após o alinhamento das sequências obtidas às sequências de referência de cada loco (as mesmas que foram utilizadas para o desenho dos *primers*), foi feita uma nova sequência consenso para cada *amplicon* utilizando o programa IGV, de modo que as novas sequências consenso de cada loco foram obtidas apenas dos fragmentos do genoma de cana-de-açúcar que foram efetivamente sequenciados. Essas novas sequências de referência foram então utilizadas nas etapas seguintes do trabalho, para fins de genotipagem. Realizou-se o alinhamento das sequências obtidas às novas referências e foram obtidas as estatísticas descritivas desse alinhamento (Tabela 13).

Tabela 12. Dados referentes ao diagnóstico de qualidade obtido pelo FastQC.

Amostra	Número total de sequências	Tamanho das *s*equências (pb)		%CG
		mínimo	máximo	
RB991536	7.465	102	7.699	45
RB971739	29.168	92	7.293	46
RB011549	4.616	95	5.999	44
RB041443	13.542	103	6.849	44
RB955977	4.375	144	7.428	44
RB021754	19.078	108	6.211	46
RB058046	3.810	134	7.350	47
RB006819a	10.043	103	7.673	45
RB046299	9.847	123	6.308	45
RB106822	9.325	108	9.202	46
RB006995x	126	134	6.199	45
RB056380	3.617	130	7.589	46
RB034045	10.891	155	7.482	44
RB036088	27.171	150	9.583	44
RB975201	10.108	120	9.927	44
RB073034	14.981	139	6.999	44
RB988082	10.529	143	8.894	44
RB867515	14.819	136	7.295	45
RB006995y	2.886	149	9.474	45
RB961003	11.184	131	8.517	44
RB006629	13.061	142	9.193	44
RB056351	10.995	148	10.917	47
RB056396	5.789	132	7.258	47
RB006819b	6.124	142	9.519	44
RB037170	4.707	124	6.893	44
RB036152	27.275	121	9.091	44
RB985476	9.216	157	5.896	44
RB987917	14.826	164	9.073	44
RB093070	7.699	135	6.846	44
RB008310	11.479	131	7.114	44
UFG06109	2.567	144	5.780	44
UFG06125	3.492	179	6.267	44
RB961552	21.491	109	13.153	45
RB036091	11.768	172	6.806	45
RB006655	5.666	178	8.358	44
RB061276	4.841	150	7.096	44
RB005014	7.830	140	7.870	45
RB975375	18.512	138	7.103	45
RB036066	13.231	155	8.851	46
RB975952	14.124	137	7.706	45
RB987935	7.929	144	8.505	45
RB975242	13.293	143	5.667	45
RB015935	6.305	160	5.920	46
UFG06292	8.221	174	6.177	45

Tabela 13. Estatísticas descritivas do alinhamento das sequências nas sequências de referência dos *amplicons*.

Indivíduo	Cobertura (X)	Sequências alinhadas (%)	Tamanho médio das sequências	*Mismatches* (%)	*Indels* (%)
RB991536	7.465	75,97	1.683	11,78	5,56
RB971739	29.168	50,68	1.404	10,32	5,59
RB011549	4.616	83,67	1.767	11,05	5,38
RB041443	13.542	81,24	1.783	10,04	5,50
RB955977	4.375	84,34	1.696	10,32	5,53
RB021754	19.078	86,05	1.744	18,75	5,56
RB058046	3.810	86,85	1.706	20,36	5,42
RB006819a	10.043	88,17	1.729	14,68	5,52
RB046299	9.847	81,15	1.668	15,20	5,45
RB106822	9.325	84,83	1.711	15,64	5,49
RB006995x	126	80,16	1.655	16,75	5,56
RB056380	3.617	85,18	1.751	15,81	5,41
RB034045	10.891	82,92	1.731	10,88	5,67
RB036088	27.171	71,09	1.629	9,93	5,85
RB975201	10.108	83,33	1.736	8,93	5,56
RB073034	14.981	83,93	1.744	9,80	5,89
RB988082	10.529	82,51	1.722	9,30	5,57
RB867515	14.819	86,95	1.718	16,59	6,15
RB006995y	2.886	86,04	1.664	17,15	6,05
RB961003	11.184	87,10	1.691	11,42	5,85
RB006629	13.061	83,21	1.669	11,38	5,67
RB056351	10.995	81,29	1.651	18,85	5,80
RB056396	5.789	83,50	1.662	17,41	5,31
RB006819b	6.124	93,06	1.686	14,61	5,56
RB037170	4.707	80,18	1.632	10,89	6,37
RB036152	27.275	76,81	1.650	10,54	6,02
RB985476	9.216	82,89	1.640	10,49	6,00
RB987917	14.826	81,16	1.593	9,89	6,03
RB093070	7.699	81,62	1.672	8,91	6,21
RB008310	11.479	82,82	1.675	11,26	6,02
UFG06109	2.567	80,99	1.673	11,31	5,89
UFG06125	3.492	87,83	1.655	11,24	6,53
RB961552	21.491	72,50	1.633	10,37	5,84
RB036091	11.768	80,01	1.645	14,75	5,79
RB006655	5.666	85,67	1.674	13,17	5,89
RB061276	4.841	81,86	1.631	13,06	5,97
RB005014	7.830	74,73	1.584	16,42	5,99
RB975375	18.512	75,36	1.605	13,24	6,18
RB036066	13.231	68,37	1.579	12,84	6,00
RB975952	14.124	74,04	1.538	12,10	6,15
RB987935	7.929	76,37	1.599	14,34	6,10
RB975242	13.293	82,53	1.590	18,08	6,04
RB015935	6.305	72,56	1.697	16,55	5,89
UFG06292	8.221	77,57	1.602	17,50	6,00
RB037194	13.117	73,02	1.563	13,11	6,05
RB083088	13.360	72,37	1.572	14,01	5,92
RB969017	13.768	75,92	1.676	13,62	5,85
RB068027	3.345	73,51	1.537	16,05	5,99

De modo geral, obteve-se uma alta cobertura de sequenciamento por indivíduo, com média de 10.498X, como desejável (Tabela 13). Apenas o indivíduo RB006995x teve uma cobertura abaixo de 2.000X, esse corresponde a um dos dois indivíduos utilizados como controle biológico. Nota-se também que cerca de 79,2% dos indivíduos apresentaram porcentagem de sequência*s* alinhadas na referência acima de 75%, e apenas dois indivíduos apresentaram este valor abaixo de 70%. A porcentagem de bases não coincidentes com a referência (*mismatches*) variou de 8% a 20%, como contribuição tanto do polimorfismo entre indivíduos quanto dos erros de sequenciamento. A porcentagem de *indels* manteve-se ao redor de 6%, tendo sido homogênea entre os indivíduos.

4.3 GENOTIPAGEM POR SEQUENCIAMENTO E ANÁLISE DE IDENTIFICAÇÃO INDIVIDUAL

A chamada de polimorfismos realizada pela ferramenta *mpileup* do *software* SAMtools resultou na identificação de um total de 356 SNPs. As regiões genômicas com menor número de sítios polimórficos foram os *amplicons chr03_1* e *chr04_2*, ambas com menos de 20 SNPs. O *amplicon chr08_1* destacou-se por apresentar 90 SNPs, sendo a região mais polimórfica dentre as analisadas (Tabela 14).

Tabela 14. Número de SNPs identificados em cada *amplicon*.

Amplicon	Número de SNPs identificados
chr02_1	24
chr03_1	16
chr04_2	17
chr05_1	77
chr07_2	45
chr08_1	90
chr09_1	24
chr10_1	39
chr10_2	24

A análise de divergência genética entre os indivíduos, realizada a partir da matriz de distância euclidiana de Rogers-Wright, produziu um dendrograma com coeficiente de correlação cofenética de 0,75 (Figura 12). Nota-se que as duplicatas técnicas utilizadas como controle (RB006819a e RB006819b, destacados em vermelho) formaram um nó com consistência de 94,03%, denotando similaridade consistente entre elas. Entretanto, mesmo apresentando a menor distância genética dentre todos os indivíduos analisados, há

inexistência de identidade genética perfeita entre eles, resultante de erro de genotipagem. Isso evidencia a dificuldade de utilização da tecnologia Oxford Nanopore em estudos de genotipagem SNP.

Os indivíduos RB006995x e RB006995y (destacados em verde) representam amostras obtidas de colmos diferentes, previamente identificados como supostos clones. Os resultados obtidos não corroboram com esta hipótese. Tal fato pode ser explicado ou pela taxa de erro de genotipagem, a qual pode ter sido aumentada devido à baixa cobertura do indivíduo RB006995x (126 X) ou pela hipótese de que esses indivíduos não sejam, de fato, clones. A divergência genética entre eles é superior à de cada um deles a outros genótipos sabidamente distintos.

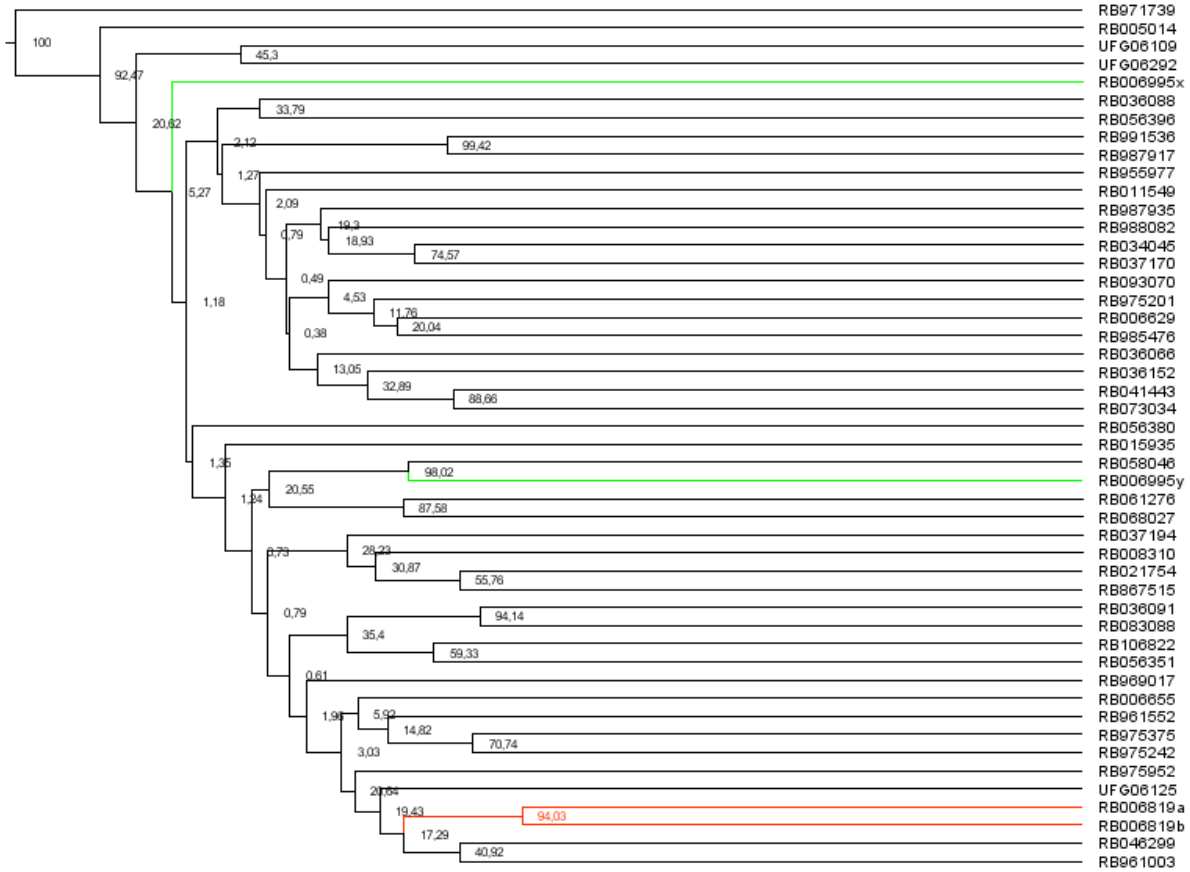

Figura 12. Dendrograma obtido pela análise de agrupamento UPGMA da matriz de distância genética de Rogers-Wright. Em cada nó são apresentados os valores de *bootstrap* obtidos a partir de 10.000 reamostragens.

A falta de confiabilidade para testes de identidade observada neste estudo ocorre principalmente devido à alta taxa de erro de sequenciamento da tecnologia de sequenciamento da Oxford Nanopore. Cumpre salientar que neste trabalho as taxas de erro

foram inferiores às relatadas na literatura. Segundo Cornelis et al. (2019) as taxas de erro da tecnologia Oxford Nanopore se situam entre 10% e 32% com um índice de qualidade *phred* variando de 5 a 10. Mesmo com uma cobertura de sequenciamento de 40X, a plataforma MinION pode inserir ou substituir uma base a cada 10-50 kb, e excluir erroneamente uma base a cada 1 kb (Magi et al., 2017). Tal fato é ainda mais complicado de se resolver quando a genotipagem é realizada em organismos poliploides e altamente heterozigóticos, sobretudo se for realizada sem a identificação prévia dos SNPs, ou em regiões homopoliméricas (Cornelis et al., 2019; Malmberg et al., 2019).

O que se tem feito para realização de genotipagem por sequenciamento com essa tecnologia é utilizar sítios SNPs previamente conhecidos. Evitando locos em regiões homopoliméricas e utilizando um limiar brando, que considera um sítio SNP heterozigótico quando o alelo de menor frequência é apresentado em pelo menos 25% das sequências, a tecnologia ONT foi considerada viável para análises forenses por Cornelis et al. (2019). Outra forma de se aumentar a precisão da genotipagem em sítios SNPs previamente conhecidos se dá pela exclusão de todos os genótipos heterozigotos da análise (Malmberg et al., 2019).

No entanto, para maioria das pesquisas no campo agronômico, o desejável são análises sem a necessidade de informações prévias. Deve se considerar ainda que, majoritariamente, as cultivares de interesse comercial de cana-de-açúcar possuem genoma complexo, inviabilizando essas estratégias que estão sendo utilizadas para contornar os problemas relacionados à elevada taxa de erro das plataformas de sequenciamento Oxford Nanopore.

5 CONSIDERAÇÕES FINAIS

A cana-de-açúcar é um dos principais produtos agrícolas do Brasil. Perduram programas de melhoramento dessa cultura há mais de 100 anos, logo há uma ampla gama de genótipos melhorados. Entretanto, não há técnicas seguras e práticas para a acurada identificação dos genótipos utilizados comercialmente, podendo ocorrer propagação de genótipos distintos de cana-de-açúcar como se fossem clones e vice-versa. O desenvolvimento de métodos de genotipagem por sequenciamento permite a identificação dos clones de cana-de-açúcar por meio de marcadores moleculares.

Os métodos de genotipagem por sequenciamento encontram aplicações em muitas das pesquisas genômicas. Em muitos casos, estas técnicas visam a redução dos custos envolvidos. Isso pode ser feito a partir da redução da complexidade do genoma e por meio de processos de multiplexação tanto de locos quanto de amostras. A redução da complexidade do genoma possibilita o aumento da cobertura de sequenciamento, melhorando a qualidade dos dados obtidos, sem que haja um aumento dos custos envolvidos.

Recentemente surgiu uma nova tecnologia de sequenciamento (Oxford Nanopore Technologies). Esta tecnologia vem sendo utilizada, dentre outras aplicações, para montagem de genomas devido à obtenção de sequências longas. A sua aplicação como método de genotipagem por sequenciamento, no entanto, ainda requer avanços. Ao que tudo indica, os resultados são razoáveis quando se trabalha com organismos de genoma simples, com altos níveis de homozigose, e quando se dispõe da identificação prévia de SNPs. Contudo para maioria das espécies vegetais de interesse agronômico essas premissas ainda não são realidade.

A tecnologia de sequenciamento Oxford Nanopore é recente e tem avançado rapidamente, tanto em termos de kits de reagentes e equipamentos disponíveis, quanto em termos dos programas de processamento dos dados resultantes do sequenciamento. Apesar do esforço relacionado ao desenvolvimento de novos programas/versões capazes de minimizar os problemas dessa tecnologia, a alta taxa de erros associada aos dados de sequenciamento obtidos ainda é um entrave. Estes erros, ainda que sejam de natureza majoritariamente aleatória, são um problema principalmente aplicação desta tecnologia em

estratégias de genotipagem por sequenciamento, sobretudo em casos em que a identificação genética com alto grau de segurança é exigida. Todavia, esses esforços associados às vantagens dessa tecnologia como o baixo custo do equipamento, portabilidade (MinION), rápida obtenção dos resultados, alta cobertura e obtenção de sequências longas, fazem dessa tecnologia uma ferramenta promissora no mercado das plataformas de sequenciamento.

Neste trabalho a aplicabilidade do teste de identidade genética em cana-de-açúcar utilizando a plataforma de sequenciamento MinION da Oxford Nanopore foi investigada. A amplificação de nove locos contendo SNPs foi feita utilizando reações *multiplex* e os indivíduos foram submetidos ao sequenciamento utilizando *barcodes*. Tais estratégias implicam na redução de custos da análise. Este trabalho demostra que mesmo obtendo uma elevada cobertura dos locos sequenciados, problemas com a alta taxa de erro da plataforma impediram a genotipagem precisa e consequentemente a identificação clonal em cana-de-açúcar de uma maneira segura. Com o aprimoramento desta tecnologia de sequenciamento, novos estudos desta natureza devem ser realizados no sentido de se avaliar a aplicabilidade desta tecnologia de genotipagem para fins de identificação individual.

REFERÊNCIAS

ALI, O. A.; O'ROURKE, S. M.; AMISH, A. J.; MEEK, M. H.; LUIKART, G.; JEFFRES, C.; MILLER, M. R. Rad capture (Rapture): Flexible and efficient sequence-based genotyping. **Genetics**, v. 202, n. 2, p. 389–400, 2016.

ALJANABI, S.; FORGET, L.; DOOKUN, A. An improved and rapid protocol for the isolation of polysaccharide-and polyphenol-free sugarcane DNA. **Plant Molecular Biology Reporter**, v. 17, n. 3, p. 281-281, 1999.

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

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

ANEEL. Agência Nacional de Energia Elétrica. **Informações gerencias:** Segundo semestre, setembro 2014.: Itaipu: Agência Nacional de Energia Elétrica. 2014. Disponível em: < http://www.aneel.gov.br/informacoes-gerenciais>. Acesso em: 31 ago. 2018.

ANDOLFATTO, P.; DAVISON, D.; EREZYILMAZ, D.; HU, T. T.; MAST, J.; SUNAYAMA-MORITA, T.; STERN, D. L. Multiplexed shotgun genotyping for rapid and efficient genetic mapping. **Genome Research**, v. 21, n. 4, p. 610–617, 2011.

ARAÚJO, E. DA S.; SANTOS, J. A. P. O desenvolvimento da cultura da cana-de-açúcar no Brasil e sua relevância na economia nacional. **Igarss 2014**, n. 1, p. 1–5, 2014.

AZHAKANANDAM, K.; WEISSINGER, S. M.; NICHOLSON, J. S.; QU R.; WEISSINGER, A. K. Amplicon-plus targeting technology (APTT) for rapid production of a highly unstable vaccine protein in tobacco plants. **Plant Molecular Biology**, v. 63, n. 1, p. 393–404, 2007.

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**, v. 3, n. 10, p. 1–7, 2008.

BAYER, P.E.; RUPERAO, P.; MASON, A. S.; STILLER, J.; CHAN, C. K.; HAYSHI, S. et al. High-resolution skim genotyping by sequencing reveals the distribution of crossovers and gene conversions in Cicer arietinum and Brassica napus. **Theoretical and Applied Genetics**, v.128, n.6, p. 1039–1047, 2015.

BEVAN, I. S.; RAPLEY, R.; WALKER, M. R. Sequencing of PCR-amplified DNA. **Cold Spring Harbor Laboratory Press**, v.1054, n.9803, p. 222–228, 1992.

BOUTIGNY, A. L.; BARRANGER, A.; BOISSÉSON, C.D.; BLANCHARD, Y.; ROLLAND M. Targeted Next Generation Sequencing to study insert stability in genetically modified plants. **Scientific Reports**, v.9, n.2308, p. 1-9, 2019.

BRANDÃO, A. **Cana-de-açúcar:** álcool e açúcar na história e no desenvolvimento social do Brasil. Brasília: Horizonte, 1985. 269 p.

BREMER, G. Problems in breeding and cytology of sugar cane - III. The cytological crossing research of sugar cane. **Euphytica**, v. 10, n. 2, p. 229–243, 1961.

BUCKLETON, J.; TRIGGS, C. M.; WALSH, S. J. Forensic DNA evidence interpretation. In: BUCKLETON, J.; CLAYTON, T.; TRIGGS, C. **Parentage testing**. 1. ed. 2005. cap. 10, p. 349 - 402.

BUERMANS, H. P. J.; DEN DUNNEN, J. T. Next generation sequencing technology: Advances and applications. **Elsevier**, v. 1842, n. 10, p. 1932-1941, 2014.

BYBEE, S. M.; BRACKEN-GRISSOM, H.; HAYNES, B. D.; HERMANSEN, R. A.; BYERS, R. L.; CLEMENT, M. J.; UDALL, J. A.; WILCOX, E. R.; CRANDALL, K. A. Targetes Amplicon Sequencing (TAS): A Scalable Next-Gen Elsevier, Approach to Multilocus, Multitaxa Phylogenetics. **Genome Biology and Evolution**, v. 3, n1. 1312, p. 1312-1323, 2011.

CIB. Conselho de Informações sobre Biotecnologia. **Guia da cana-de-açúcar: avanço científico beneficia o país**: Conselho de Informações sore Biotecnologia, 2009. Disponível

em: < https://cib.org.br>. Acesso em: 30 ago. 2018.

CNI. Confederação Nacional da Industria. **O setor sucroenergético em 2030: dimensões, investimentos e uma agenda estratégica.** Confederação Nacional da Industria, 2017. Disponível em: < http://www.portaldaindustria.com.br/>. Acesso em: 30 ago. 2018.

Conab. Companhia Nacional de Abastecimento. **Acompanhamento de safra brasileira - cana-de-açúcar:** Segundo levantamento, agosto 2017 – safra 2017/2018. Companhia Nacional de Abastecimento. 2017. Disponível em: <https://www.conab.gov.br/info-agro/safras/cana>. Acesso em: 7 nov. 2017.

Conab. Companhia Nacional de Abastecimento. **Acompanhamento de safra brasileira - cana-de-açúcar:** Segundo levantamento, agosto 2018 – safra 2017/2019. Companhia Nacional de Abastecimento. 2017. Disponível em: <https://www.conab.gov.br/info-agro/safras/cana>. Acesso em: 28 ago. 2018.

CORNELIS, S.; GANSEMANS, Y.; DELEYE, L.; DEFORCE, D.; NIEUWERBURGH, F. V. Forensic SNP Genotyping using Nanopore MinION Sequencing**. Scientific Reports**, v. 7, n. 41759, p. 1-5, 2017.

CORNELIS, S.; GANSEMANS, Y.; PLAETSEN, A. S. V.; WEYMAERE, J.; WILLEMS, S.; DEFORCE, D.; NIEUWERBURGH, F. V. Forensic tri-allelic SNP genotyping using nanopore sequencing. **Forensic Science International: Genetics**, v. 38, n. 1, p. 204-210, 2019.

DAVEY, J.; 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**, v. 12, n. 7, p. 499–510, 2011.

DESCHAMPS, S.; LLACA, V.; MAY, G. D. Genotyping-by-Sequencing in Plants. **Biology**, v. 1, n. 3, p. 460–483, 2012.

DAHL, F.; GULLBERG, M.; STENBERG, J.; LANDEGREN, U.; NILSSON, M. Multiplex amplification enabled by selective circularization of large sets of genomicDNAfragments. **Nucleic Acids Research**, v. 33, n. 8, p. 1-7, 2005.

D'HONT, A.; GRIVET, L.; FELDMANN, P.; RAO, S.; BERDING, N.; GLASZMANN, J.

C. Characterisation of the double genome structure of modern sugarcane cultivars (Saccharum spp.) by molecular cytogenetics. **Molecular & General Genetics**, v. 250, n. 4, p. 405–413, 1996.

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

ELSHIRE, R. J.; GLAUBITZ, J. C.; SUN, Q.; POLAND, J. A.; KAWAMOT, K.; BUCKLER, E. S.; MITCHELL, S. E. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. **Plos One**, v. 6, n. 5, p. 1–10, 2011.

EVETT, I. W.; WEIR, B. S. Interpreting DNA Evidence: Statiscical Genetics for Forensic Scientists. In: EVETT, I. W.; WEIR, B. S. **Parentag Testing**. 1998. cap.6, p. 282-285.

FONSECA, L. H. M.; LOHMANN, L. G. Combining high-throughput sequencing and targeted loci data to infer the phylogeny of the "*Adenocalymma-Neojobertia*" clade (Bignonieae, Bignoniaceae). **Molecular Phylogenetics and Evolution**, v. 123, n.1, p. 1-15, 2018.

GARCIA, A. A. F; MOLLINARI, M.; MARCONI, T. G.; SERANG, O. R.; SILVA, R. R.; VIEIREA, M. L. C. et al. SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. **Scientific Reports**. v. 3, n. 3399 p. 1–10, 2013.

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

GOODSTEIN, D. M.; SHU, S.; HOWSON, R.; NEUPANE, R.; HAYES, R. D.; FAZO, J.;

MITROS, T.; DIRKS, W.; HELLSTEN, U.; PUTNAM, N.; ROKHSAR, D. S. Phytozome:

85 a comparative platform for green plant genomics. **Nucleic Acids Research**,

Oxford, v. 40, p.1178-1186, 2011.

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

351, 2016.

GRABHERR, M. G.; HAAS, B. J.; YASSOUR, M.; LEVIN, J. Z.; THOMPSON, D. A.; AMIT, I. et al. A. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. **Nature Biotechnology**, New York, v. 29, n. 7, p. 644-652, 2011.

GRAHAM, C. F.; GLENN, T. C.; McARTHUR, A, G.; BOREHAM, D. R.; KIERAN, T.; LANCE, S.; MANZON, R. G.; MARTINO, J. A.; PIERSON, T.; ROGERS, S. M.; WILSON, J. Y.; SOMERS, C. M. Impacts of degraded DNA on restriction enzyme associated DNA sequencing (RADSeq). **Molecular Ecology Resources**, v. 15, n. 6, p. 1304–1315, 2015.

GRIVET, L.; DANIELS, C.; GLASZMANN, J. C.; D'HONT, A. A review of recente molecular genetics evidence for sugarcane evolution and domestication. **Ethnobotany Research & Applications**, v. 2, n. 0, p. 9–17, 2004.

HODGES, E.; XUAN, Z.; BALIJA, V.; KRAMER, M.; MOLLA, M. N.; SMITH, S. W.; MIDDLE, C. M.; RODESCH, M. J.; ALBERT, T. J.; HANNON, G. J.; MCCOMBIE, W. R. Genome-wide in situ exon capture for selective Resequencing**. Nature Genetics**, v. 29, n. 12, p. 1522–1527, 2007.

HOFFBERG, S. L; KIERAN, T.J .; CATCHEN, J .M.; DEVAULT, A.; FAIRCLOTH, B. C.; MAURICIO, R.; GLEEN, T. C. RADcap: sequence capture of dual-digest RADseq libraries with identifiable duplicates and reduced missing data. **Molecular Ecology Resources**, v. 16, n. 5, p. 1264–1278, 2016.

HUANG, X.; FENG, Q.; QIAN, Q.; ZHAO, Q.; WANG, L. et al. High-throughput genotyping by whole-genome resequencing. **Genome Research**, v. 19, n. 9, p. 1068–1076, 2009.

IRVINE, J. E. Saccharum species as horticultural classes. **Theoretical and Applied Genetics**, v. 98, n. 2, p. 186–194, 1999.

JANNOO, N.; GRIVET, L.; SEGUIN, M.; PAULET, F.; DOMAINGUE, R.; RAO, P. S. et al. Molecular investigation of the genetic base of sugarcane cultivars. **Theoretical and**

Applied Genetics, v. 99, n. 1–2, p. 171–184, 1999.

JONES, M. E.; SHEPHERD, M.; HENRY, R. Pollen flow in Eucalyptus grandis determined by paternity analysis using microsatellite markers. **Tree Genetics & Genomes**, v. 4, n. 1, p. 37–47, 2008.

KAYSER, M.; DE KNIJFF, P. Improving human forensics through advances in genetics, genomics and molecular biology. **Nature Reviews Genetics**, v. 12, n. 3, p. 179–192, 2011.

LANGMEAD, B.; TRAPNELL, C.; POP, M.; SALZBERG, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. **Genome biology**,

London, v. 10, n. 3, p. R25, 2009.

LI, H. Minimap2: pairwise alignment for nucleotide sequence. **Bioinformatics**, v. 34, n. 18, p. 3094-3100, 2018.

LIU, L.; LI, Y.; LI, S.; HU, N.; HE, Y.; PONG, R.; LIN, D.; LU, L.; LAW, M. Comparison of next-generation sequencing systems. **Journal of Biomedicine and Biotechnology**, v. 2012, p. 1-12, 2012.

LU, H.; GIORDANO, F.; NING, Z. Oxford nanopore minion sequencing and genome assembly. **Genomics Proteomics Bioinfomartics**, v. 14, p. 265–279, 2016.

MAGI, A.; GIUSTI, B.; TATTINI, L. Characterization of MinIon nanopore data for resequencing analyses. **Briefings**, v. 18, n. 6, p. 940-653, 2017.

MALMBERG, M. M.; SPANGENBERG, G. C. S.; DAETWYLER, H. D.; COGAN, N. O. I. Assessment of low-coverage nanopore long sequência sequencing for SNP genotyping I doubled haploid canola (*Brassica napus* L.). **Scientific Report**, v. 9, n. 8688, p. 1-12, 2019.

MANTEL, N. The detection of diase clustering and generalized regression approach. **Cancer Research**, v. 27, n. 1, p. 209-220, 1967.

McCORMICK RF, TRUONG SK, SREEDASYAM A, JENKINS J, SHU S, SIMS D, et al.

The Sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. **The Plant journal: for cell and molecular biology**. v. 93, n. 2, p. 338 – 354, 2017.

MELO, V. J. R. **Determinação de Paternidade de Indivíduos Superiores de Eucalyptus com Base em Marcadores Microssatélites**. 2000. 178 f. Dissertação (Mestrado em Agronomia) - Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2000.

MELO, A. T. O. **Montagem e caracterização do transcriptoma de cana-de-açúcar (Saccharum spp.) utilizando dados de sequenciamento de nova geração**. 2015. 105 f. Tese (Doutorado em Genética e Melhoramento de Plantas) – Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2015.

MENDES, C. A. **Construção de um modelo de seleção genômica ampla para cana-de-açúcar (*Saccharum spp*.) no contexto do programa de melhoramento da RIDESA - Goiás.** 2015. 80 f. Teese (Doutorado em Genética e Melhoramento de Plantas) - Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2015.

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

MIKHEYEV, A. S.; TIN, M. M. A first look at the Oxford Nanopore MinION sequencer. **Molecular Ecology**, v. 14, n. 6, p. 1097-1102, 2014.

MILLER, M. R.; DUNHAM, J. P.; AMORES, A.; CRESKO, W. A.; JHONSON, E. A. Rapid and cost-effective polumorphism identification and genotyping using restriction site associated DNA (RAD) markers. **Genome Research**, v.1, n. 17, p.240-248, 2007.

MORAIS, L. K.; CURSI, D. E.; SANTOS, J. M.; SAMPAIO, M.; CÂMARA, T. M. M.; SILVA, P. A.; BARBOSA, G. V.; HOFFMANN, H. P.; CHAPOLA, R. G.; FERNANDES JÚNIOR, A. R.; GAZAFFI, R. **Melhoramento Genético da Cana-de-Açúcar**. 1. ed. Embrapa Tabuleiros Costeiros, 2015. 38 p.

NITSCH, M. O programa de biocombustíveis Proalcool no contexto da estratégia energética brasileira. **Revista de Economia Política**, v. 11, n. 2, p. 123-138, 1991.

OLIVEIRA, I. B. **Desequilíbrio de ligação e análise de seleção desequilíbrio de ligação e**

análise de seleção genômica em cana-de-açúcar. 2014. 85 f. Tese (Mestrado em Genética e Melhoramento de Plantas) - Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2014.

ONDA, Y.; TAKAHAGI, K.; SHIMIZU, M.; INOUE, K.; MOCHIDA, K. Multiplex PCR Targeted Amplicon Sequencing (MTA-Seq): Simple, Flexible, and Versatile SNP Genotyping by Highly Multiplexed PCR Amplicon Sequencing. **Frontiers in Plant Science**, v. 9, n.201, p.1-10, 2018.

REN, R.; RAY, R.; LI, P.; XU, J.;ZHANG, M.; LIU, G.; YAO, X.; KILIAN, A.; YANG, X. Construction of a high-density DArTseq SNP-based genetic map and identification of genomic regions with segregation distortion in a genetic population derived from a cross between feral and cultivated-type watermelon. **Molecular Genetics Genomics**, v. 290, n.1, p. 1457-1470, 2015.

PETERSON, B. K.; WEBER, J. N.; KAY, E. H.; FISHER, H. S.; HOEKSTRA, H. E. Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. **Plos One**, v. 7, n. 5, p. 1-11, 2012.

ROBINSON, M. D.; McCARTHY, D. J.; SMYTH, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. **Bioinformatics**, Oxford, v. 26, p. 139-140, 2010.

RODRIGUES, E. B.; COLLEVATI, R. G.; CHAVES, L. J.; MOREIRA, L. R.; TELLES, M. P. C. Mating system and pollen dispersal in Eugenia dysenterica (Myrtaceae) germplasm collection : tools for conservation and domestication. **Genetica**, v. 144, n. 2, p. 139–146, 2016.

POLAND, J. A. BROWN, P. J.; SORRELLS, M. E.; JANNINK, J. L. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. **Plos One**, v. 7, n. 2, p. 1-8, 2012.

POLAND, J.; RIFE, T. Genotyping-by-Sequencing for Plant Breeding and Genetics. **The Plant Genome**, v. 5, n. 3, p. 92–102, 2012.

QUACKENBUSH, J.; LIANG, F.; HOLT, I.; PERTEA, G.; UPTON, J. The TIGR Gene Index: reconstruction and representation of expressed gene sequences. **Nucleic Acid**

Research, Oxford, v. 28, n. 1, p. 141-145, 2000.

QUAST, C.; PRUESSE, E.; YILMAZ, P.; GERKEN, J.; SCHWEER, T.; YARZA, P.; PEPLIES, J.; GLÖCKNER, F. O. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. **Nucleic Acids Research**, v. 41, p. D590-

D596, 2013.

ROGERS, J. S. **Measures of genetic similarity and genetic distance**. Austin: University of Texas, 1972. p. 145-154.

ROWAN, B. A.; PATEL, V.; WEIGEL, D.; SCHNEEBERGER, K. Rapid and Inexpensive Whole-Genome Genotyping-by-Sequencing for Crossover Localization and Fine-Scale Genetic Mapping. **G3**, v. 13, n. 5, p. 385–398, 2015.

SANTOS, F.; BORÉM, A.; CALDAS, C. Cana-de-açúcar: bioenergia, açúcar e etanol - tecnologias e perpectivas. In: MATSUOKA, S.; BRESSIANI, J.; MACCHERONI, W.; FOUTO, I. **Bioenergia de Cana**. 1. ed. 2016, cap. 20, p. 547 - 571.

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

SARMENTO, F. J. Q. **Modelagem de um Ambiente para Análise de DNA em Genética Forense Modelagem de um Ambiente para Análise de DNA em Genética Forense**. 2006. 46 f. Tese (Mestrado em Ciência) - Instituto de Computação de Conhecimento, Universidade

Federal de Alogoas, Maceió, 2006.

SATO, M.; HOSOYA, S.; YOSHIKAWA, S.; OHKI, S.; KOBAYASHI, Y.; ITOU, T.; KIKUCHI, K. A highly flexible and repeatable genotyping method for aquaculture studies based on target amplicon sequencing using next-generation sequencing technology. **Scientific Reports**, v.9, n.6904, p.1-9, 2019.

SCHEBEN, A.; BATLEY, J.; EDWARDS, D. Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. **Plant Biotechnology Journal**, v. 15, n. 1, p. 149–161, 2017.

SHIKIDA, P. F. A.; PEROSA, B. B. Álcool combustível no Brasil e path dependence. **Revista de Economia e Sociologia Rural**, v. 50, n. 2, p. 243–262, 2012.

STOLLE, E.; MORITZ, R. F. A. RESTseq – Efficient Benchtop Population Genomics with RESTriction Fragment SEQuencing. **Plos One**, v. 8, n. 5, p. 4–8, 2013.

SUN, X.; LIU, D.; ZHANG, X.; LI, W.; LIU, H HONG, W. et al. SLAF-seq : An Efficient Method of Large-Scale De Novo SNP Discovery and Genotyping Using High-Throughput Sequencing. **Plos One**, v. 8, n. 3, p. 1-15, 2013.

TOONEN, R. J.; PURITZ, L. B.; FORSMAN, Z. H.; WHITNEY, J. L.; FERNANDEZ-SILVA, I.; ANDREWS, K. R.; BIRD, C. E. zRAD: a simplified method for genomic genotyping in non-model organisms. **PeerJ**, v. 1, n. 1, p. 1–15, 2013.

TRUONG, H. T.; RAMOS, M.; YALCIN, F.; RUITER, M.; POEL, H. J. A.; HUVENAARS, K. H. J. et al. Sequence-Based Genotyping for Marker Discovery and Co-Dominant Scoring in Germplasm and Populations. **Plos One**, v. 7, n. 5, p. 1-9, 2012.

TURCHETTO-ZOLET, A. C.; TURCHETTO, C.; ZANELLA, C. M.; PASSAIA, G. **Marcadores Moleculares na Era genômica: Metodologias e Aplicações**. 1. ed. 2017. 180 p.

UNTERGASSER, A.; NIJVEEN, H.; RAO, X.; BISSELING, T.; GEURTS, R.; LEUNISSEN, J. A. M. Primer3Plus, an enhanced web interface to Primer3. **Nucleic Acids Research**, v. 35, n. 1, p. 71-74, 2007.

VAN DIJK, E. L.; AUGER, H.; JASZCZYSZYN, Y.; THERMES, C. Ten years of next-generation sequencing technology. **Trends in genetics**, v. 30, n. 9, p. 418–426, 2014.

VAN DIJK, E. L.; JASZCZYSZYN, Y.; NAQUIN, D.; THERMES, C. The Third Revolution in Sequencing Technology. **Trends in Genetics**, v. 34, n. 9, p. 666–681, 2018.

VAN ORSOUW, N. J.; HOGERS, R. C. J.; JANSSEN, A.; YALCIN, F.; SNOEIJERS, S., VERSTEGE, E. et al. Complexity Reduction of Polymorphic Sequences (CRoPS^{TM}): A Novel Approach for Large-Scale Polymorphism Discovery in Complex Genomes. **Plos One**, v. 11, n. 1, p. 1-10, 2007.

VAN TASSELL, C. P., SMITH, T. P., MATUKUMALLI, L. K., TAYLOR, J. F., SCHNABEL, R. D., LAWLEY, C. T., HAUDENSCHILD, C. D.; MOORE, S.S.; WARREN, W.C.; SONSTEGARD, T. S. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. **Nature Methods**, v. 5, n. 3, p. 247–252, 2008.

WRIGHT, S. **Evolution and genetics of population**. Chicago: University of Chicago Press, 1978, 520 p.

XIE, W.; FENG, Q.; YU, H.; HUANG, X.; ZHAO, Q.; XING, Y.; YU, S.; HAN, B.; ZHANG, Q. Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. **PNAS**, v. 107, n. 23, p. 10.578- 10.583, 2010.

WANG, S.; MEYER, E.; MCKAY, J. K.; MATZ, M. V. 2b-RAD: a simple and flexible method for genome-wide genotyping. **Nature Methods**. v. 8, n. 9, p. 808–810, 2012.

WEIR, B. S. Genetic Data Analysis II. In: WEIR, B. S. **Individual Identification**. 2. ed. 1996. cap. 6, p. 202 – 228.