UFG

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

GENÔMICA DE ORGANELAS DE CANA-DE-AÇÚCAR *Saccharum* spp. - CULTIVAR RB867515

MAYARA STEFANY DA S. M. FEITOSA

Orientador:
Prof. Alexandre Siqueira Guedes Coelho

Goiânia - GO
Brasil

Setembro – 2017

sistema de bibliotecas ufg

PRPG
PRÓ-REITORIA DE PÓS-GRADUAÇÃO

UFG

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO 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 nº 9610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou *download*, a título de divulgação da produção científica brasileira, a partir desta data.

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

2. Identificação da Tese ou Dissertação:

Nome completo do autor: Mayara Stefany da Silva Mariano Feitosa

Título do trabalho: GENÔMICA DE ORGANELAS DE CANA-DE-AÇÚCAR (*Saccharum* spp. - CULTIVAR RB867515)

3. Informações de acesso ao documento:

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

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

Mayara Stefany da S. M. Feitosa
Assinatura do(a) autor(a)[2]

Ciente e de acordo:

Alexandre Siqueira Guedes Coelho
Assinatura do(a) orientador(a)[2]

Data: 10 / 08 / 17

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

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

[2]A assinatura deve ser escaneada.

MAYARA STEFANY DA SILVA MARIANO FEITOSA

GENÔMICA DE ORGANELAS DE CANA-DE-AÇÚCAR (*Saccharum* spp. - CULTIVAR RB867515)

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

Goiânia, GO – Brasil
2017

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

Feitosa, Mayara Stefany da Silva Mariano
GENÔMICA DE ORGANELAS DE CANA-DE-AÇÚCAR (Saccharum spp. - CULTIVAR RB867515) [manuscrito] / Mayara Stefany da Silva Mariano Feitosa. - 2017.
LX, 65 f.

Orientador: Prof. Dr. Alexandre Siqueira Guedes Coelho.
Dissertação (Mestrado) - Universidade Federal de Goiás, Escola de Agronomia (EA), Programa de Pós-Graduação em Genética & Melhoramentos de Plantas , Goiânia, 2017.
Bibliografia.

1. Genômica organelar. 2. Cana-de-açúcar. 3. Cloroplastos. 4. Mitocôndrias. 5. Heteroplasmia. I. Coelho, Alexandre Siqueira Guedes, orient. II. Título.

CDU 575

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

UFG

ATA DE DEFESA PÚBLICA DE DISSERTAÇÃO DE MAYARA STEFANY DA SILVA MARIANO FEITOSA. Aos vinte e nove dias do mês de setembro do ano de dois mil e dezessete (29/09/2017), às 13:30, no Auditório do Melhoramento, da Escola de Agronomia da Universidade Federal de Goiás, reuniram-se os membros da Banca Examinadora: Dr. Alexandre Siqueira Guedes Coelho (Presidente/Orientador), Drª. Tereza Cristina de Oliveira Borba e Dr. Evandro Novaes. Sob a presidência do Orientador, em sessão pública, procedeu-se à avaliação da defesa da dissertação intitulada: **GENÔMICA DE ORGANELAS DE CANA-DE-AÇÚCAR (*Saccharum* spp. – CULTIVAR RB867515)**, de autoria de **Mayara Stefany da Silva Mariano Feitosa**, discente do Programa de Pós-Graduação em Genética e Melhoramento de Plantas, no nível de Mestrado, da Universidade Federal de Goiás. A sessão foi aberta pelo presidente da Banca Examinadora, Dr. Alexandre Siqueira Guedes Coelho, que fez a apresentação formal dos membros da Banca. Em seguida, a palavra foi concedida a autora da dissertação que, em 40 minutos, apresentou o seu trabalho. Terminada a apresentação, cada membro da Banca arguiu a mestranda, tendo-se adotado o sistema de diálogo sequencial. Ao final, a Banca reunida em separado procedeu à avaliação da defesa. A Dissertação foi considerada APROVADA pela Banca Examinadora, cumprindo integralmente este requisito para fins de obtenção do título de MESTRE EM GENÉTICA E MELHORAMENTO DE PLANTAS, pela Universidade Federal de Goiás, em conformidade com o estabelecido pela Resolução nº 1403/2016, do Conselho de Ensino, Pesquisa, Extensão e Cultura da UFG (CEPEC/UFG), que regulamenta o Programa de Pós-Graduação em Genética e Melhoramento de Plantas. Para fins de publicação eletrônica, a mestranda poderá efetuar as modificações eventualmente sugeridas pela Banca Examinadora e encaminhar à Secretaria do PGMP, respeitando-se o prazo máximo de 30 dias após a data da Defesa. A conclusão do curso e a emissão do diploma dar-se-ão em conformidade com o estabelecido pela Resolução CEPEC nº 1403/2016. Cumpridas as formalidades de pauta, às 18:00 a presidência da mesa encerrou esta sessão de defesa de Dissertação e, para constar eu, Jéssica Almeida Silva, Secretária do PGMP, lavrei a presente Ata que depois de lida e aprovada, segue assinada pelos membros da Banca Examinadora, em duas vias de igual teor.

Alexandre Siqueira Guedes Coelho
Presidente/Orientador

Drª. Tereza Cristina de Oliveira Borba
Membro Externo

Dr Evandro Novaes
Membro Interno

Prof. Sérgio Tadeu Sibov
Coordenador do Programa de Pós-Graduação em Genética e Melhoramento de Plantas
Escola de Agronomia/UFG

AGRADECIMENTOS

Agradeço à toda minha família por me apoiarem incondicionalmente. Em especial à minha mãe, a guerreira em quem me espelho, por que desistir não existe no dicionário dela. À minha avozinha que com todo carinho sempre acalentou minha cabecinha ansiosa "calma minha filha, tem tempo pra tudo na vida". E à minha tia Dairan, que de um modo muito especial abriu os caminhos dos estudos para que nós pudéssemos seguir mais facilmente. Muito obrigado!

Agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela bolsa de estudos de Mestrado. Agradeço ao meu orientador professor Alexandre Siqueira Guedes Coelho por todos os ensinamentos e por ter confiado em mim para realização desse trabalho e por sempre ter me estimulado pelas minhas potencialidades, de modo que nunca me senti incapaz de realizar algo. Pelo contrário, cada orientação eu entrava achando que seria recriminada por não ter conseguido fazer ou entender algo e saía achando que iria dominar o mundo e que nada era tão difícil que eu não pudesse alcançar, e isso faz muita diferença, pois o foco não eram minhas limitações e sim o meu aprendizado! Obrigada prof. Alexandre!

Agradeço a todos os meus professores, desde a pré-escola que sempre me incentivaram a buscar meus sonhos e que com sua crença em mim acabaram por me fazer acreditar que era possível. Professores nem sempre sabem o quanto marcam nossas trajetórias de vida, mas queria deixar registrado aqui, aos que não estão presentes nesse momento que foi tentando me ver sob seus olhos que segui em frente, mesmo quando tudo parecia mais difícil demais, eu me espelhei em vocês, na coragem, na força e na esperança de que é sim pela educação que conseguimos grandes transformações.

Me sinto grata por ter tido tantos bons mestres e por ter aprendido mais que conteúdos com eles, e eu não poderia deixar de agradecer aqui a professora Carolina Novaes, que me mostrou um outro modo de enxergar os problemas e de seguir em frente. Muito obrigada, jamais vou esquecer suas palavras e o que elas representaram.

Agradeço a toda à equipe do Laboratório de Genética e Genômica de Plantas que direta ou indiretamente contribuíram para a realização desse trabalho. Agradeço ao Osvaldo Reis Júnior por ter me recebido tão bem no LaCTAD (Laboratório Central de

Tecnologias de Alto Desempenho em Ciências da Vida - Unicamp) para um estágio em bioinformática.

Agradeço aos amigos e amigas de longa data que mesmo longe sempre me deram apoio. Agradeço aos amigos que fiz aqui, vocês tornaram essa caminhada mais feliz! Obrigada GDC!!! Obrigada Letycia Basso, parceira e ouvinte das minhas lamentações! Em especial agradeço a minha amiga Karla Carneiro, amiga de longa data e por isso já estamos habituadas em dividir aflições e risadas tantas e tantas vezes, fizemos ensino fundamental e médio juntas e parte da graduação e nos reencontramos aqui, foi bom poder desabafar com quem me conhecia há tanto tempo! Obrigada amiga!

Agradeço à mais nova melhor amiga, Stefany, que conheci aqui na pós-graduação e sem dúvida alguma será amiga pra vida toda. Amiga, obrigada por tornar meus dias mais alegres e por viver comigo histórias que só acontecem com a gente mesmo. E obrigada por fazer terapia comigo, a velha e infalível terapia de amiga, um bom papo, um cafezinho, um passeio e umas gordices pra aliviar a ansiedade. Agradeço à um grande amigo da vida, Rafael, por ter me apoiado e incentivado sempre e por sempre ter acreditado e investido junto comigo nos meus sonhos. Obrigado!

SUMÁRIO

RESUMO

FEITOSA, M.S.S.M. **Genômica de organelas de cana-de-açúcar (*Saccharum* spp. - cultivar RB867515)**. 2017. 65 f. Dissertação (Mestrado em Genética e Melhoramento de Plantas) – Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2017[1].

A cana-de-açúcar (*Saccharum* spp.) é uma das mais importantes culturas das regiões tropicais e subtropicais do mundo. A cana é cultivada em mais de 100 países, fornecendo matéria-prima para a obtenção de produtos como açúcar e bioetanol. Dada sua importância, diversos esforços vêm sendo realizados com o objetivo de se realizar a caracterização genômica de cultivares de cana-de-açúcar. Os genomas eucarióticos são distribuídos em diferentes compartimentos celulares que apresentam padrões distintos de herança. Plastídeos e mitocôndrias possuem sistema genético próprio, contendo DNA, RNA e todos os componentes necessários para os processos de replicação, transcrição e síntese de proteínas, que ocorrem nestas organelas. Cloroplastos e mitocôndrias são organelas que têm como função principal a transdução de energia. Os cloroplastos são responsáveis pela conversão de energia luminosa em energia química, durante a fotossíntese. As mitocôndrias fornecem energia em forma de ATP, por meio da respiração celular. O presente trabalho foi desenvolvido com o objetivo de se realizar a montagem e a caracterização dos genomas cloroplastidial e mitocondrial da cultivar de cana-de-açúcar RB867515, utilizando dados de sequenciamento de nova geração Illumina e PacBio. Em cloroplastos, buscou-se identificar, pela utilização de *reads* longos obtidos pela tecnologia PacBio no processo de montagem, evidências de ocorrência de heteroplasmia cloroplastidial em cultivares modernas de cana-de-açúcar. No genoma mitocondrial investigou-se a ocorrência de variações genéticas e genômicas estruturais. Os *assemblies* foram obtidos pela utilização de *reads* organelares, selecionados através do mapeamento a sequências de referência de cloroplastos e mitocôndrias, disponíveis publicamente. Os *assemblies* obtidos foram realizados com os softwares SPAdes e Organelle_PBA. A anotação gênica foi realizada utilizando as ferramentas DOGMA, GeSeq e Mitofy. Foram identificados dois haplótipos (isoformas) de cloroplastos na cultivar RB867515. Estas isoformas diferem entre si pela ocorrência de orientações distintas da região SSC (*small single copy*), confirmando a hipótese de heteroplasmia cloroplastidial em cana-de-açúcar. Cada haplótipo é constituído por 141.181 pb e exibe uma estrutura quadripartida típica, que inclui uma região longa de cópia única (LSC) de 83.047 pb flanqueada por duas regiões de repetições invertidas (IRs) de 22.795 pb e uma pequena região de cópia única (SSC) entre as IRs de 12.544 pb. O genoma mitocondrial montado foi constituído por dois cromossomos: o cromossomo 1 de comprimento total de 300.765 pb e o cromossomo 2, de 194.383 pb. As estimativas obtidas para os conteúdos GC (~44%) e AT (~56%) foram concordantes com as de outras angiospermas. Foram anotados 39 CDSs, 5 genes hipotéticos conservados, 5 rRNAs, 18 tRNAs e 9 fragmentos de genes transferidos de cloroplastos. A comparação dos cromossomos mitocondriais da cultivar RB867515 com aqueles de *S. officinarum* permitiu a identificação de polimorfismos de bases únicas (SNPs), duplicações gênicas e expansões genômicas.

Palavras-chave: cana-de-açúcar, cloroplasto, heteroplasmia, SSC, mitocôndria.

1 Orientador: Prof. Dr. Alexandre Siqueira Guedes Coelho. EA – UFG.

ABSTRACT

FEITOSA, M.S.S.M. **Organellar genomics in sugarcane (*Saccharum* spp. - cultivar RB867515**). 2017. 60 f. Dissertation (Master in Genetics and Plant Breeding) – Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2017[1].

Sugarcane (*Saccharum* spp.) is one of the most important crops in tropical and subtropical regions of the world. It is cultivated in more than 100 countries, where it is used as raw material to obtain sugar and bioethanol. Given its importance, many efforts have been carried out to characterize the genome of sugarcane cultivars. The eukaryotic genomes are confined in different cellular compartments that present different inheritance patterns. Plastids and mitochondria have their own genetic system, comprising DNA, RNA and all the demanded components for replication, transcription and protein synthesis, that occur inside these organelles. The primary function of chloroplasts and mitochondria is energy transduction. Chloroplasts are responsible to convert light into chemical energy during photosynthesis, while mitochondria provides energy to the cell in form of ATP molecules during respiration. In this work, the chloroplast and mitochondrial genomes of sugarcane cultivar RB867515 are assembled and characterized, using data from two next generation sequencing technologies – Illumina and PacBio. In chloroplasts, we sought to identify evidences of heteroplasmy, by using long reads from PacBio technology in the assembly process. In mitochondria, we investigated the occurrence of genetic and structural genomic variations. The assemblies were carried out using screened reads for the organellar genomes. These reads were selected by mapping whole genome shotgun reads to reference genome sequences of chloroplast and mitochondria, that are publicly available. The organellar reads were assembled using SPAdes and Organelle_PBA. Gene annotation was obtained using DOGMA, GeSeq and Mitofy tools. Two chloroplast haplotypes (isoforms) were identified in the cultivar RB867515. These isoforms are different from each other because they present a distinct orientation of the SSC (*small single copy*) region, confirming the hypothesis of chloroplast heteroplasmy in sugarcane. The genome of each chloroplast isoform comprises 141,181 bp, and shows a typic quadripartite structure, that includes a long single copy region (LSC) of 83,047 bp, which is flanked by two inverted repeat regions (IRs) of 22,795 bp and a small-single copy region (SSC), between IRs, of 12,544 bp. The assembled mitochondrial genome comprised two chromosomes of 300,765 bp and 194,383 bp. The estimates of GC (~44%) and AT (~56%) contents were similar to those obtained for other angiosperms. A total of 39 CDSs, 5 hypothetical conserved genes, 5 rRNAs, 18 tRNAs and 9 gene fragments transferred from chloroplast were annotated. The RB867515 mitochondrial chromosomes showed differences when compared to those from *S. officinarum*, including single nucleotide polymorphisms (SNPs), genetic duplications and genomic expansions.

Keywords: sugarcane, chloroplast, heteroplasmy, SSC, mitochondria.

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

1 INTRODUÇÃO GERAL

1.1 CONSIDERAÇÕES INICIAIS

O presente trabalho foi desenvolvido com o objetivo de se realizar a montagem e caracterização dos genomas organelares da cultivar de cana-de-açúcar RB867515, utilizando dados provenientes de uma iniciativa voltada para o sequenciamento do seu genoma nuclear, utilizando as tecnologias de sequenciamento de nova geração Illumina e PacBio. Especificamente, buscou-se identificar, pela utilização de *reads* longos obtidos pela tecnologia PacBio no processo de montagem, evidências de ocorrência de heteroplasmia em cultivares modernas de cana-de-açúcar.

Cumpre destacar que, embora existam outras iniciativas de sequenciamento de genoma cloroplastidial de cana-de-açúcar, este trabalho é o primeiro a investigar especificamente a ocorrência de heteroplasmia cloroplastidial empregando dados de sequenciamento de nova geração produzidos pela tecnologia PacBio.

1.2 GENOMAS ORGANELARES

Os genomas eucarióticos são distribuídos em diferentes compartimentos genéticos que apresentam padrões distintos de herança. Algumas organelas, como plastídeos e mitocôndrias, possuem sistema genético próprio, contendo DNA, RNA e os componentes necessários para a replicação, transcrição e síntese de proteínas, que ocorrem em seu interior (Allen, 2003). Essas organelas têm como função primária a conversão de energia, sendo a mitocôndria responsável pela respiração e o cloroplasto pela fotossíntese. Ambos não seguem o padrão de herança mendeliano, a herança é predominantemente uniparental, sendo frequentemente materna em angiospermas (Greiner et al., 2015).

A presença de plastídeos é umas das principais características que distinguem as células vegetais de outras células eucarióticas, sendo os cloroplastos um grupo de

organelas vital para estes organismos. Existem diferentes tipos de plastídeos, como proplastídeos (plastídeos indiferenciados), cromoplastos (armazenam carotenoides), amiloplastos (armazenam amido), elaioplastos (armazenam lipídeos), leucoplastos (plastídeos não pigmentados presentes principalmente nas células da raiz), etioplastos (presentes nos cotilédones das mudas de angiospermas, nunca foram expostos à luz) e cloroplastos (especializados em fotossíntese) (Rogalski et al., 2015).

Os genomas de organelas evoluíram a partir de procariotos de vida livre através de eventos de endossimbiose. Segundo esta teoria, atualmente aceita, os cloroplastos se originaram a partir de cianobactérias, enquanto as mitocôndrias se originaram a partir de uma α-proteobactéria (O'Malley, 2015). O tamanho dos genomas de organelas é bastante reduzido se comparado a seu antecessor de vida livre devido à maciça translocação de genes do genoma organelar para o genoma nuclear, seguida de sua deleção no genoma da organela (Lloyd & Timmis, 2011).

Em células foliares de angiospermas existem cerca de 700 a 2000 cópias do genoma cloroplastidial (Golczyk et al., 2014). Cada cópia possui tamanho entre 107 e 218 kb e contém de 120 a 130 genes. Cloroplastos de plantas terrestres possuem topologia circular e estrutura geral quadripartida representada por uma longa região de cópia única (*large single copy*/LSC), flanqueada por duas regiões de repetições invertidas (*inverted repeats*/IRs: IRa e IRb) que são separadas por uma pequena região de cópia única (*small single copy*/SSC) (Figura 1) (Ravi et al., 2008; Wang et al., 2012; Golczyk et al., 2014).

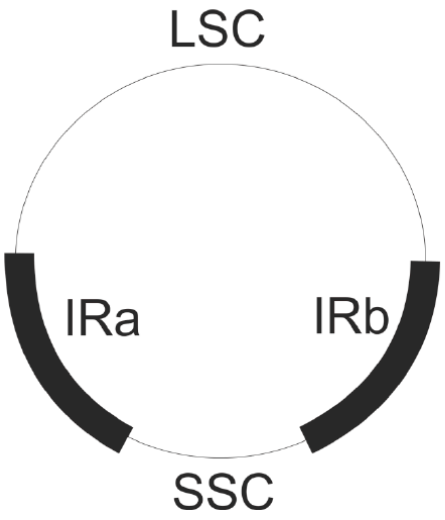

Figura 1. Representação esquemática da organização estrutural dos cloroplastos. Fonte: Mota (2012).

O genoma mitocondrial também está presente em múltiplas cópias, sendo possível encontrar até 100 cópias em uma célula foliar (Wang et al., 2012). O tamanho dos genomas mitocondriais de plantas varia de cerca de 200 kb, como em *Brassica hirta*, a pouco mais do que 10 Mb, como em *Silene conica* (11,3 Mb) (Shearman et al., 2016) e contém de 50 a 60 genes (não se considerando o número de cópias de cada um deles). Estes genomas possuem estrutura multipartida e são compostos por regiões de repetições longas e regiões de repetições curtas (Kubo & Newton, 2008). Essa estrutura diversa ocorre devido ao acúmulo de sequências repetitivas, que podem levar a eventos de recombinação e rearranjos genômicos constantes dentro da mesma espécie. Estes eventos por sua vez promovem o surgimento de múltiplas regiões no genoma organelar com sequências idênticas e diferentes números de cópias (Richardson & Palmer, 2007; Alverson et al., 2010; Goremykin et al., 2012).

As topologias das moléculas de mtDNA podem ser lineares ou circulares, únicas ou concatenadas (Oldenburg & Bendich, 2015). As múltiplas moléculas circulares de DNA mitocondrial presentes em angiospermas são descritas como uma única molécula de DNA circular que aloja o conjunto completo de genes chamado de "cromossomo mestre". Cada cromossomo mestre é composto por várias sequências repetidas onde ocorrem os eventos de recombinação (Figura 2) (Sloan, 2013).

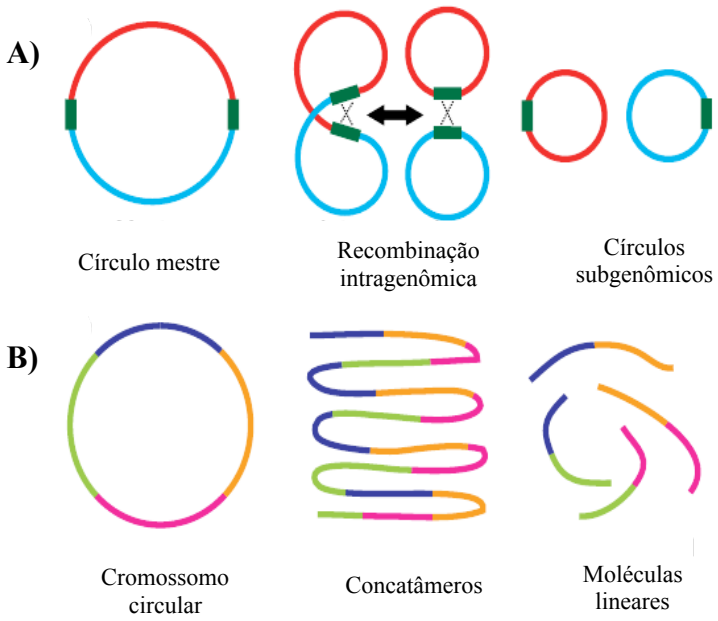

Figura 2. Estrutura de DNA mitocondrial de plantas. (A) Modelo multipartido convencional. (B) Estruturas alternativas que podem produzir mapas de genomas circulares. Os segmentos verdes representam grandes repetições. Adaptado de Sloan (2013).

A expansão dos genomas mitocondriais em plantas comparada à dos genomas cloroplastidiais ocorre principalmente devido ao acúmulo de sequências repetidas, expansão de íntrons e incorporação de DNA exógeno de plastídeos ou do núcleo (Richardson & Palmer, 2007; Alverson et al., 2010; Goremykin et al., 2012). Apesar de possuirem genomas maiores, a densidade de genes nos genomas mitocondriais é menor do que a dos genomas cloroplastidiais. Isso ocorre porque durante a evolução houveram sucessivos eventos de transferência de genes e posterior perda dos genes transferidos. Muitos desses genes perdidos foram transferidos para genomas nucleares, e raramente para genomas de cloroplastos (Adams & Palmer, 2003; Daniell et al., 2016).

O DNA organelar transferido para núcleo pode não ser funcional e ser rapidamente eliminado ou a translocação pode ser acompanhada da aquisição de um promotor nuclear e um sinal de poliadenilação levando à ativação desses genes no núcleo (Sheppard & Timmis, 2009; Lloyd & Timmis, 2011; Rogalski et al., 2015). Do ponto de vista evolutivo, as altas taxas de transferência de genes organelares para o núcleo são consideradas vantajosas por reduzirem o risco de mutações associadas à ação de radicais livres derivados das reações de óxido-redução (redox) e por adicionarem complexidade ao genoma nuclear, fornecendo novas sequências de *exons* ou *introns* aos genes nucleares existentes (Woodson & Chory, 2008; Barbrook et al., 2010).

A manutenção de sistemas genéticos separados aumenta o custo energético para a célula devido à produção de proteínas necessárias à coordenação da comunicação entre organelas e núcleo (Woodson & Chory, 2008). Apesar desta elevação do custo energético, é possível que esse sistema complexo tenha sido mantido devido ao fato de que as proteínas envolvidas na fotossíntese e na respiração celular não são eficientemente transportadas do citoplasma para as organelas em que estes processos ocorrem. Estas proteínas são altamente hidrofóbicas e muitas vezes tóxicas, se acumuladas no citoplasma. Em certos casos, a síntese dessas proteínas deve ocorrer no próprio local em que se dá a montagem de complexos enzimáticos com múltiplas subunidades (Bogorad, 1975). Além disso, a síntese organelar de proteínas envolvidas em reações redox permite a alteração rápida das concentrações destas proteínas sem a necessidade de envio de sinal ao núcleo, permitindo uma rápida alteração do estado redox da organela (Allen, 2003).

Por sua vez, as proteínas codificadas pelos genes que foram transferidos para o genoma nuclear podem reter suas funções originais e serem exportadas de volta para a organela, se elas vierem a adquirir peptídeos sinais apropriados, ou podem assumir novos papeis, não relacionados com suas funções originais (Barbrook et al., 2010).

A troca de DNA entre organelas e o genoma nuclear é conhecida como transferência intracelular de genes (TIG) e a transferência de DNA entre espécies não relacionadas é conhecido como transferência horizontal de genes (THG). As plantas têm altas taxas de transferência horizontal de genes envolvendo genomas mitocondriais. Genomas mitocondriais de plantas incorporam sequências nucleares e plastidiais, bem como DNA mitocondrial derivado de outras espécies (Smith & Keeling, 2015; Smith, 2016). Os segmentos de DNA derivados de plastídeos encontrados em mitocôndrias (MTPTs) de angiospermas representam de 0,1 a 0,3% do total de mtDNA e podem representar uma transferência de até 87% do conteúdo genômico plastidial. Apesar de a maioria dos segmentos derivados de plastídeos resultar em segmentos não-funcionais, uma vez integrados ao genoma mitocondrial, os MTPTs podem afetar a função mitocondrial, tanto pela criação de novas isoformas e de promotores de genes, quanto pela introdução de novos tRNAs funcionais.

Os genomas de plastídeos são mais conservados e exibem taxas menores de recebimento de sequências gênicas de outros compartimentos genéticos. Apenas quatro segmentos derivados de mitocôndrias localizados em cpDNAs (PTMTs) foram descritos até o momento (Gandini & Sanchez-Puerta, 2017).

O conteúdo genético dessas organelas tem sido bastante utilizado para avaliar relações filogenéticas e compreender a evolução das mais variadas espécies de plantas. O conhecimento da dinâmica evolutiva dessas organelas e da função dos genes presentes nos seus genomas pode representar uma ferramenta importante para o melhoramento genético, por subsidiar estratégias baseadas na sua transformação genética (Daniell et al., 2005). Um breve resumo dos trabalhos que investigaram genomas de cloroplastos e mitocôndrias e suas principais características nos últimos anos estão reunidos nas Tabelas 1 e 2.

Tabela 1. Estudos de genomas cloroplastidiais e suas principais características.

Espécie	Tamanho e estrutura do genoma	Principais características	Referência
Astragalus membranaceus	123.582 pb apenas uma IR	5 *loci* de hipermutação intra-específicos, sendo 3 heteroplásmicos	Lei et al. (2016)
Citrus sinensis	160.129 pb IR: 2x26.996 pb LSC: 87.744 pb SSC: 18.393 pb	genes correspondem a 57,3% do genoma - 89 genes codificantes de peptídeos - 4 rRNAs - 30 tRNAs	Bausher et al. (2006)
Cultivar LA3023 tomato	155.461 pb IR: 2x25.611 pb LSC: 85.876 pb SSC: 18.363 pb	genes correspondem a 58,3% do genoma - 133 genes: 113 únicos - 20 duplicados dentro de IRs - 30 tRNAs	Daniell et al. (2006)
Erianthus arundinaceus	141.210 pb IR: 2x22.762 pb LSC: 83.170 pb SSC: 12.516 pb	*S. officinarum* está mais relacionada com *M. sinensis* do que com *E. arundinaceus*	Tsuruta et al. (2017)
Miscanthus sinensis	141.416 pb IR: 2x22.775 pb LSC: 83.141 pb SSC: 12.681 pb	*S. officinarum* está mais relacionada com *M. sinensis* do que com *E. arundinaceus*	Tsuruta et al. (2017)
Oryza minuta	135.094 pb IR: 2x20.840 pb LSC: 81.074 pb SSC: 12.470 pb	Alto grau de similaridade de sequência entre espécies de *Oryza* e divergência relativamente alta nas regiões de espaçadores intergênicos	Asaf et al. (2017)
Saccharum cv. Q155	141.181 pb IR: 2x22.795 pb LSC: 83.047 pb SSC: 12.544 pb	Sem evidência de heteroplasmia	Hoang et al. (2015)
Saccharum cv. RB72454	141.156 pb	-	Evans & Joshi (2016)
Saccharum cv. RB867515	141.181 pb IR: 2x22.795 pb LSC: 83.047 pb SSC: 12.544 pb	Sequência idêntica à da cultivar Q155	Vidigal et al. (2016)
Saccharum cv. SP803280	141.182 pb IR: 2x22.794 pb LSC: 83.048 pb SSC: 12.546 pb	116 genes: 82 genes codificantes de peptídeos - 4 rRNAs - 30 tRNAs	Calsa Júnior et al. (2004)

Espécie	Tamanho e estrutura do genoma	Principais características	Referência
Saccharum cv. NCo310	141.182 pb IR: 2x22.795 pb LSC: 83.048 pb SSC: 12.544 pb	número, conteúdo gênico e ordem de genes funcionais idênticos aos descritos em arroz, milho e trigo	Asano et al., 2004
Solanum tuberosum	155.312 pb IR: 2x25.595 pb LSC: 85.749 pb SSC: 18.373 pb	130 genes - 18 genes com 1 ou 2 *introns* - 4 *introns* em IRs - 1 *intron* em SSC - poucos tRNAs	Chung et al., 2006
Sorghum bicolor	140.754 pb IR: 2x22.782 pb LSC: 82.688 pb SSC: 12.502 pb	genes correspondem a 52,1% do genoma - 131 genes - 113 genes únicos - 4 rRNAs - 30 tRNAs	Saski et al. (2007)

Adaptado de Lichtenstein et al. (2016).

Tabela 2. Estudos de genomas mitocondriais de plantas e suas principais características.

Espécie	Tamanho e estrutura do genoma	Principais características	Referência
Armadillidium vulgare	monômero linear de ~14 kb - dímero circular de ~28 kb	heteroplasmia mitocondrial - genes de tRNA em *loci* espelhados	Peccoud et al. (2017)
Brassica napus	223.412 pb 1 círculo mestre - 2 segmentos subgenômicos	repetições diretas ativas - recombinação intramolecular - presença de cpDNA e de fragmentos nucleares	Chen et al. (2011)
Citrullus lanatus	379.236 pb molécula circular	genes representam 45,9% do genoma - evidência de transferência horizontal	Alverson et al. (2010)
Cucumis sativus	555.935 pb, 83.817 pb , 44.840 pb – 3 círculos autônomos	sequências derivadas do núcleo representam 1/3 do genoma	Alverson et al. (2011)
Cucurbita pepo	982.833 pb molécula circular	genes representam 16,6% do genoma - elementos de transposição derivados do núcleo	Alverson et al. (2010)

Espécie	Tamanho e estrutura do genoma	Principais características	Referência
Cycas taitungensis	414.903 pb 1 círculo mestre e moléculas circulares alternativas	genes representam 10,1% do genoma - elementos transponíveis em regiões não codantes	Chaw et al. (2008)
Daucus carota	281.132 pb 2 possíveis círculos mestres	perda genética por transferência para o genoma nuclear	Iorizzo et al. (2012)
Ferrocalamus rimosivaginus	432.839 pb molécula circular	genes representam 8,9% do genoma - grandes repetições	Ma et al. (2012)
Oryza sativa	~490 pb molécula circular	4.717 variações4.507 SNPs - 210 *indels*	Tong et al. (2017)
Phaeoceros laevis	209.482 pb molécula linear	*introns*: 36,5% - *exons*: 10,9% - sequências intergênicas: 52,6% - edição de RNA em 54 genes	Xue et al. (2010)
Saccharum officinarum	300.778 pb – 144.698 pb	2 círculos independentes que compartilham poucas sequências - retenção do gene *trnL-CAA* perdido em *S. bicolor*	Shearman et al. (2016)
Sorghum bicolor	468.628 pb molécula circular	*assembly* parcial	Zheng et al. (2011)
Tripsacum dactyloides	704.100 pb molécula circular	grande quantidade de segmentos de cpDNA transferidos	Wang et al. (2012)
Vitis vinífera	773.279 pb círculo mestre - segmentos subgenômicos	genes representam 5,0% do genoma – genoma grande devido à expansão de sequências espaçadoras	Goremykin et al. (2007)
Zea luxurians	539.368 pb molécula circular	genes representam 8,6% do genoma - presença de sequências nucleares	Darracq et al. (2010)
Zea perennis	570.354 pb molécula circular	genes representam 8,5% do genoma - presença de sequências nucleares	Darracq et al. (2010)

Adaptado de Lichtenstein et al. (2016).

1.3 ESTRATÉGIAS DE SEQUENCIAMENTO DE GENOMAS ORGANELARES

Dados de sequenciamento de genomas organelares podem ser produzidos a partir da análise de amostras de DNA organelar, obtidas diretamente através de isolamento

por centrifugação, limitando a contaminação com sequências de DNA nuclear, ou a partir do *screening in silico* de dados de sequenciamento de DNA genômico total (Nock et al., 2011; Wicke & Schneeweiss, 2015a). De modo geral, a estratégia de análise baseada no isolamento de DNA organelar requer um esforço adicional significativo, que implica em aumento de custos para obtenção das amostras, e muitas vezes pode não evitar a contaminação (Zhang et al., 2011). Por outro lado, como as amostras de DNA genômico total submetidas às técnicas de sequenciamento incluem também DNA de cloroplastos e de mitocôndrias, além do próprio material genético nuclear, os dados obtidos no âmbito destas análises incluem sequências dos genomas organelares. Assim as sequências correspondentes aos genomas de cloroplastos e de mitocôndrias podem ser identificadas por meio de análises de bioinformática, que podem ser relativamente rápidas e eficientes, preservando-se qualidade dos dados (Zhang et al., 2011).

Os primeiros estudos genéticos de genomas organelares foram feitos por meio do sequenciamento de Sanger de bibliotecas *shotgun* enriquecidas de DNA organelar, garantindo uma cobertura genômica de 6X a 10X, utilizando fosmídeos (*fosmids*) ou BACs. Com o advento das técnicas de sequenciamento de nova geração foi possível substituir o baixo rendimento destas técnicas e o intensivo trabalho de clonagem (Wicke & Schneeweiss, 2015a).

As tecnologias de sequenciamento de nova geração tornaram possível a utilização de DNA diretamente no preparo das bibliotecas, sem a necessidade da etapa de enriquecimento, e a realização do sequenciamento em menor tempo e com custos bem menores, tanto pelo método de sequenciamento por síntese (Illumina), quanto pelo método de sequenciamento em tempo real de moléculas únicas (PacBio) (Nock et al., 2011; Zhang et al., 2012; Wicke & Schneeweiss, 2015a).

Todas as tecnologias de sequenciamento são suscetíveis a erros. Os tipos de erro mais comuns e as frequências com que eles ocorrem (FE) variam de acordo com a tecnologia utilizada. O sequenciamento capilar de Sanger ($FE=10^{-2}$) e o baseado na tecnologia Illumina ($FE=10^{-3}$) são mais suscetíveis a erros envolvendo substituições. O sequenciamento utilizando a plataforma 454 FLX, baseada em pirossequenciamento, é mais suscetível a *indels* ($FE=10^{-2}$), enquanto a tecnologia PacBio apresenta problemas com deleções GC ($FE=10^{-2}$), a tecnologia Ion Torrent com pequenas deleções ($FE=10^{-2}$) e a tecnologia SOLiD apresenta viés A/T ($FE=10^{-2}$) (Wicke & Schneeweiss, 2015a).

Considerando o diminuto tamanho dos genomas de organelas, esperava-se que as plataformas de sequenciamento de nova geração fossem capazes de fornecer dados suficientes para a obtenção de montagens (*assemblies*) extremamente robustos. A realidade evidencia porém que, mesmo com as elevadas coberturas de sequenciamento utilizadas, raramente as sequências destes genomas são montadas completamente a partir dessas tecnologias. A utilização de dados obtidos com a tecnologia PacBio no *assembly* de genomas plastidiais tem sido eficiente em resolver os problemas associados às grandes regiões de repetições invertidas que são características dos genomas de cloroplastos e são dificilmente resolvidas com dados obtidos de outras plataformas de sequenciamento (Ferrarini et al., 2013).

1.4 HETEROPLASMIA

Heteroplasmia é definida como a heterogeneidade de DNA organelar no interior de células individuais (Bock, 2007). A heteroplasmia tem sido relatada em genomas mitocondriais e cloroplastidiais dos mais diversos organismos. Anteriormente, acreditava-se que a heteroplasmia de cloroplastos fosse rara, porém, com o avanço das tecnologias de sequenciamento, estas variações passaram a ser descritas com maior frequência (Sabir et al., 2014) . Não estão completamente claros os mecanismos de manutenção da heteroplasmia, que pode ocorrer em nucleoides únicos ou em nucleoides separados, gerando duas populações distintas de plastídeos dentro de uma célula (Maliga, 2014).

As fontes de diversidade genética para genomas de cloroplastos são mutações, herança biparental dos cloroplastos, recombinação entre genomas de cloroplastos e transferência de genes do cloroplasto ao DNA nuclear e mitocondrial (Wolfe & Randle, 2004) . A detecção de SNPs (*Single Nucleotide Polymorphisms*) intra-cultivares em genomas mitocondriais e de plastídeos configura heteroplasmia de SNPs. Esse tipo de heteroplasmia foi verificado tanto em cloroplastos quanto em mitocôndrias de cultivares de palmeiras (Sabir et al., 2014). A heteroplasmia de SNPs é considerada rara, porém sua natureza pode dificultar a detecção desses polimorfismos, visto que os compartimentos genéticos da célula compartilham segmentos de DNA. Esses segmentos após serem

incorporados sofrem taxas de mutação maiores do que a do segmento original e, caso ambos sejam recuperados em um *assembly*, haverá a identificação equivocada de SNPs (Sabir et al., 2014; Garaycochea et al., 2015; Hoang et al., 2015).

Em 1983, Palmer demonstrou que o DNA de cloroplasto de *Phaseolus vulgaris* exibe uma forma de heteroplasmia em que o plastoma existe em dois estados equimolares, diferindo apenas na orientação relativa da SSC (Palmer, 1983; Walker et al., 2015). Essa variação ocorre devido a presença das IRs, que são duas regiões grandes de repetição invertidas que facilitam recombinação intramolecular homóloga entre os dois IRs, produzindo duas isoformas de plastídeos (Wicke & Schneeweiss, 2015b).

A recombinação intramolecular entre IRs foi sugerida como um mecanismo para evitar a divergência das duas cópias entre si e tem o potencial de reverter a polaridade do segmento entre elas, fenômeno conhecido como recombinação *flip-flop* (Palmer, 1983). Caso ocorra uma mutação que altera a estabilidade termodinâmica da estrutura de modo que uma orientação seja favorecida sobre outra, pode ocorrer a fixação deste haplótipo ao longo do tempo (Maliga, 2014).

1.5 A CANA DE AÇÚCAR

A cana-de-açúcar (*Saccharum* spp.) é reconhecida como uma das mais importantes culturas das regiões tropicais e subtropicais do mundo. Cultivada em mais de 100 países, a planta é responsável por 80% da produção mundial de açúcar, além de ser matéria-prima para produção de álcool, ácido acético, butanol, enzimas industriais, alimentos para animais e biofertilizantes. O Brasil é responsável por cerca de 40% da cana-de-açúcar produzida no mundo, o que o torna o maior produtor mundial (Rao et al., 2016).

A cana-de-açúcar pertence à divisão *Magnoliophyta*, classe *Liliopsida*, ordem *Cyperales*, família *Poaceae*, tribo *Andropogoneae*, subtribo *Saccharininae* e gênero *Saccharum* (Daniels & Roach, 1987). A família *Poaceae* é uma das maiores famílias de monocotiledôneas, com cerca de 10 mil espécies distribuídas em 650 gêneros que incluem cereais extremamente importantes para alimentação humana, como trigo, arroz, milho e sorgo (Glaszmann et al., 1997; Calsa Júnior et al., 2004; Matsuoka et al., 2005).

O gênero *Saccharum* consiste em seis espécies: *S. spontaneum* L., *S. robustum* Brandes e Jeswiet ex Grassl, *S. barberi* Jeswiet, *S. edule* Hassk., *S. officinarum* L. e *S. sinense* Roxborough (Hodkinson et al., 2002). As modernas cultivares de cana-de-açúcar são híbridos interespecíficos envolvendo *S. officinarum*, que contribui com genes para alto teor de açúcar, e *S. spontaneum*, que proporciona ao híbrido a característica de resistência a doenças (Piperidis et al., 2010).

De modo geral, as modernas cultivares de cana-de-açúcar são bem adaptadas a regiões tropicais e subtropicais, possuem metabolismo C4, são alógamas, semi-perenes, de sistema subterrâneo composto por raízes fasciculadas e rizomas. Seu desenvolvimento se dá tipicamente sob a forma de touceiras, com a parte aérea constituída por colmos, folhas e inflorescência. Os toletes, utilizados na propagação vegetativa, são obtidos através dos colmos e é nessa região que estão presentes nós, entrenós e gemas axilares (Cesnik, 2007).

A propagação da cana-de-açúcar é prioritariamente vegetativa, entretanto são possíveis eventos de floração estimulados por fatores abióticos como índice de precipitação, luminosidade, variação de temperatura e altitude. A floração reduz as reservas energéticas da planta e diminui sua produtividade, porém é fundamental como parte das estratégias de melhoramento genético por permitir a obtenção de sementes de cruzamentos entre genótipos de interesse (Matsuoka et al., 2005).

A realização de cruzamentos interespecíficos no gênero *Saccharum* e o processo de nobilização resultaram em um genoma complexo, de origem multiespecífica, caracterizado pela poliploidia, com notável variação no número de cromossomos (Grivet & Arruda, 2002). Nas modernas cultivares de cana-de-açúcar o núcleo celular possui de 100 a 130 cromossomos. Cerca de 70% a 80% do conjunto cromossômico é originário de *S. officinarum*, 10% a 23% de *S. spontaneum* e, de 8% a 13%, de recombinações interespecíficas (Piperidis et al., 2010).

Todo o conteúdo citoplasmático das cultivares comerciais de cana-de-açúcar é proveniente da espécie *S. officinarum*. A literatura tem destacado a importância da ampliação da diversidade genética citoplasmática nas cultivares modernas pela alteração dos genitores femininos utilizados no melhoramento. *S. spontaneum* e espécies do gênero *Erianthus* têm sido sugeridos como potenciais genitores femininos para ampliação desta diversidade citoplasmática (Raj et al., 2015).

2 EVIDÊNCIAS DE HETEROPLASMIA EM *Saccharum* spp. (cultivar RB867515)

RESUMO

A cana-de-açúcar (*Saccharum* spp.) é uma das mais importantes culturas das regiões tropicais e subtropicais do mundo. O gênero é cultivado em mais de 100 países, fornecendo matéria-prima para a obtenção de produtos como açúcar e etanol, de grande importância econômica, além de diversos subprodutos. Dada sua importância, diversos esforços vêm sendo realizados com o objetivo de se realizar a caracterização genômica de cultivares de cana-de-açúcar. Os genomas eucarióticos são distribuídos em diferentes compartimentos genéticos que apresentam padrões distintos de herança. Algumas organelas, como plastídeos e mitocôndrias, possuem sistema genético próprio, contendo DNA, RNA e os componentes necessários para os processos de replicação, transcrição e síntese de proteínas que ocorrem em seu interior. Dentre estas organelas se destacam os cloroplastos que têm como função primária a conversão de energia luminosa em energia química, através da fotossíntese. O presente trabalho foi desenvolvido com o objetivo de se realizar a montagem e caracterização do genoma cloroplastidial da cultivar de cana-de-açúcar RB867515, utilizando dados de sequenciamento de nova geração Illumina e PacBio. Especificamente, buscou-se identificar, pela utilização de *reads* longos obtidos pela tecnologia PacBio no processo de montagem, evidências de ocorrência de heteroplasmia em cultivares modernas de cana-de-açúcar. Amostras de DNA genômico total foram submetidas ao sequenciamento utilizando as plataformas Illumina e PacBio. Após a etapa de controle de qualidade, as sequências foram submetidas ao alinhamento com sequências de referência de genomas cloroplastidiais para fins de *screening*. As sequências filtradas foram então submetidas ao *assembly* utilizando os softwares SPAdes e Organelle_PBA. A anotação gênica foi realizada utilizando-se as ferramentas DOGMA e GeSeq. Foram identificados dois haplótipos (isoformas) de cloroplastos na cultivar RB867515. Estas isoformas diferem entre si pela ocorrência de orientações distintas da região SSC (*small single copy*), confirmando a hipótese de ocorrência de heteroplasmia cloroplastidial em cana-de-açúcar. Cada haplótipo obtido foi constituído por 141.181 pb e exibiu uma estrutura quadripartida típica, que inclui uma região longa de cópia única (LSC) de 83.047 pb flanqueada por duas regiões de repetições invertidas (IRs) de 22.795 pb e uma pequena região de cópia única (SSC) entre as IRs de 12.544 pb.

Palavras-chave: cana-de-açúcar, cloroplasto, heteroplasmia, SSC.

2.1 INTRODUÇÃO

Os cloroplastos originaram-se a partir de um ancestral de cianobactéria engolfado por uma célula hospedeira heterotrófica há cerca de 1,5 bilhões de anos (Nakayama & Ishida, 2009). Posteriormente, eventos secundários de endossimbiose distribuíram os genomas de cloroplastos horizontalmente em todo o domínio vegetal (Gray, 1999). Após a endossimbiose, os genomas de plastídeos sofreram redução de tamanho em sucessivos eventos de perda de genes, eliminação de informação genética redundante entre o nucleoma e o plastoma, e a translocação de genes cloroplastidiais para o núcleo (Rogalski et al., 2015).

Alguns genes sabidamente de cloroplastos como *infA*, *rps16*, *ycf1*, *ycf2* e *ycf4* foram perdidos através da transferência de genes para o núcleo ou foram perdidos completamente da célula (Millen et al., 2001). Atualmente, os plastomas de plantas terrestres contêm de 120 a 130 genes, em contraste com genoma de uma cianobactéria, seu ancestral, que contém cerca de 3200 genes (Rogalski et al., 2015).

O plastoma da maioria das plantas terrestres apresenta uma topologia circular e estrutura geral quadripartida representada por uma longa região de cópia única (*large single copy*/LSC), flanqueada por duas regiões de repetições invertidas (*inverted repeats*/ IRs: IRa e IRb) que são separadas por uma pequena região de cópia única (*small single copy*/SSC) (Ravi et al., 2008). A região SSC abriga dois importantes grupos de genes: o primeiro é constituído por genes que codificam componentes para os fotossistemas I e II (PSI e PSII), para o complexo citocromo *b6f* e para ATP sintase; o segundo inclui os genes que codificam os elementos necessários para expressão de genes dos plastídeos, como subunidades da RNA polimerase codificada por plastídeos (PEP), rRNAs, tRNAs e proteínas ribossomais (Bock, 2007).

Em células foliares de angiospermas existem cerca de 700 a 2000 moléculas de DNA cloroplastidial idênticas (Golczyk et al., 2014). O conteúdo GC representa cerca de 30% do total de pares de bases nestes genomas, em consonância com algumas propriedades específicas da DNA polimerase de cloroplastos e do seu sistema de reparo (Nielsen et al., 2010; Olejniczak et al., 2016). Os cloroplastos da maioria das angiospermas são transmitidos uniparentalmente pelo genitor feminino (Zhu et al., 2014).

O conteúdo genético dos plastomas é bastante informativo para avaliar relações filogenéticas e compreender a evolução das mais variadas espécies de plantas. O conhecimento do genoma de organelas pode auxiliar no esclarecimento de sua dinâmica evolutiva, no aprimoramento do conhecimento da função dos seus genes e na caracterização dos eventos de transferência de segmentos gênicos entre os genomas organelares e nucleares. O desenvolvimento deste tipo de estudo pode ainda resultar na obtenção de conhecimentos importantes para o melhoramento genético, sobretudo para subsidiar estratégias baseadas na transformação genética de cloroplastos (Clarke & Daniell, 2011).

A aplicação das modernas técnicas de análise genômica pode ajudar a compreender as lacunas no conhecimento da função de genes importantes que estão no plastoma. Muitos genes de cloroplastos já foram descritos e têm sua função conhecida. No entanto, muitas ORFs hipotéticas em genomas cloroplastidiais (*ycfs: hypothetical chloroplast frames*) só agora têm sido confirmadas e suas funções ainda são desconhecidas ou não estão completamente descritas (Olejniczak et al., 2016). Um exemplo neste sentido é o do gene *ycf1*, que supostamente é um gene imprescindível para a sobrevivência de alguns organismos como *Nicotiana tabaccum* e *Chlamydomonas reinhardti.* Este gene não está presente no plastoma de *Poaceae* (de Vries et al., 2015).

Apesar de estarem presentes em várias cópias, os genomas de cloroplastos são descritos como haploides. Os eventos genético-evolutivos mais comuns são (1) recombinação de genomas dentro e entre organelas; (2) heteroplasmia, em que coexistem populações de organelas dentro de uma célula com múltiplos haplótipos, levando à distinção de haplótipos entre diferentes células dentro de um indivíduo ou entre indivíduos de uma espécie; e (3) duplicação e transferência de genes para outros compartimentos genéticos dentro de uma célula (Wolfe & Randle, 2004).

O avanço nas tecnologias de sequenciamento tem aumentado o número de sequências de cloroplastos disponíveis (Walker et al., 2015; Daniell et al., 2016). As metodologias podem envolver a utilização de DNA, enriquecido ou não, no preparo de bibliotecas e no sequenciamento que se dá em menor tempo e com custos relativamente acessíveis por diferentes tecnologias, incluindo o sequenciamento por síntese (Illumina), o sequenciamento em tempo real de moléculas únicas (PacBio) entre outras (Nock et al., 2011; Zhang et al., 2012; Wicke & Schneeweiss, 2015b; Santos-Cauz et al., 2017).

As técnicas podem variar desde o momento da obtenção do DNA cloroplastidial (cpDNA), que pode ser ou não isolado antes do sequenciamento. A necessidade de isolamento prévio do cpDNA pode tornar o processo ineficiente, pois representa um esforço adicional e não elimina a possibilidade de contaminação com segmentos de DNA nuclear e mitocondrial. Abordagens recentes, utilizando o sequenciamento do DNA total, têm sido capazes de separar *in silico* e montar com sucesso genomas de cloroplastos completos (Nock et al., 2011; Zhang et al., 2011; Evans & Joshi, 2016).

A cana-de-açúcar é uma das mais importantes culturas da região tropical e subtropical do mundo. É cultivada em mais de 100 países sendo matéria-prima de importantes produtos, como açúcar e o etanol, além de seus subprodutos como o melaço, bagaço, palha e a cachaça (Tomes et al., 2011; Rao et al., 2016). Atualmente, o Brasil produz cerca de 768 milhões de toneladas de cana-de-açúcar/ano, fato que o torna o seu maior produtor mundial, sendo o setor sucroalcooleiro responsável por 2,4% do PIB brasileiro (Rao et al., 2016).

As cultivares modernas de cana de açúcar são derivadas do cruzamento de duas espécies de *Saccharum* e representam uma combinação de *S. spontaneum*, utilizado como genitor masculino, e *S. officinarum*, utilizado como genitor feminino. Como a herança é uniparental e materna, os genomas organelares destas cultivares são provenientes de *S. officinarum* (Raj et al., 2015). Atualmente existem cinco sequências completas de genoma cloroplastidial de cana-de-açúcar disponíveis no GenBank, obtidas das cultivares SP803280, NCo310, Q155, RB867515 (utilizada novamente neste trabalho) e RB72454 (números de acesso: NC_005878.2, NC_006084.1, NC_029221.1, KX507245.1 e LN849914.1, respectivamente) (Asano et al., 2004; Calsa Júnior et al., 2004; Hoang et al., 2015; Evans & Joshi, 2016; Vidigal et al., 2016).

As sequências de cloroplastos já publicadas foram obtidos por meio de diferentes estratégias. A sequência do genoma cloroplastidial da cultivar NCo310 foi obtida por sequenciamento Sanger de produtos de PCR. No caso da cultivar SP803280, a estratégia utilizada foi o sequenciamento *shotgun* (Sanger) de DNA previamente isolado do cloroplasto. Os genomas cloroplastidiais publicados para as cultivares Q155, RB867516 e RB72454 foram montados a partir de sequências pequenas obtidas por sequenciamento Illumina (*paired ends*). É importante destacar que a utilização de tecnologias que

produzem *reads* curtos dificultam sobremaneira a identificação de variações estruturais que podem estar presentes nas sequências de genomas cloroplastidiais (Ferrarini et al., 2013).

O presente trabalho é o primeiro a investigar o genoma cloroplastidial de *Saccharum* utilizando dados de sequenciamento PacBio, que fornecem *reads* longos com maior potencial de detecção de variações estruturais. Pela utilização de *reads* obtidos por esta tecnologia, buscou-se avaliar a hipótese de que ocorre heteroplasmia em cultivares modernas de cana-de-açúcar.

2.2 MATERIAL E MÉTODOS

2.2.1 Extração e quantificação de DNA genômico

O preparo das amostras de DNA genômico foi realizado no Laboratório de Genética e Genômica de Plantas, no Centro de Excelência em Melhoramento Genético da Cana-de-açúcar no Cerrado, localizado na Escola de Agronomia da Universidade Federal de Goiás. O DNA foi obtido a partir de gemas laterais frescas de colmos da cultivar RB867515. As gemas laterais selecionadas foram retiradas com o auxílio de uma espátula, higienizadas e mantidas em freezer a -80ºC até a etapa de extração do DNA conforme protocolo descrito em Aljanabi et al. (1999).

A integridade do DNA extraído foi verificada pela comparação do perfil eletroforético das amostras com diferentes quantidades de DNA padrão (50, 100 e 200ng) e com DNA de fago λ digerido com *HindIII*. Somente amostras com fragmentos superiores a 20kb foram selecionadas. Os perfis eletroforéticos foram obtidos por eletroforese horizontal em gel de agarose 1%, corado em brometo de etídeo (0,5µg/µL). A quantificação de DNA nas amostras foi realizada com fluorímetro Qubit® e como medida de pureza foi utilizada a relação de absorbâncias a 260 nm e 280 nm, obtida pelo espectrofotômetro Nanodrop®. Somente amostras com valores de pureza entre 1,8 e 2,0 foram selecionadas para as etapas seguintes de análise.

2.2.2 Sequenciamento de DNA

2.2.2.1 Sequenciamento pela plataforma Illumina

As amostras foram dispostas em microtubos de 1,5µL, contendo 200µg de DNA genômico, e enviadas à empresa *BGI* (BGI, China) onde foram construídas as bibliotecas (Tabela 2.1) e foi realizado o sequenciamento utilizando a plataforma de nova geração HiSeq2000 (Illumina®), pelas estratégias *paired ends* e *mate pairs* (2 x 100pb).

Tabela 2.1. Bibliotecas de DNA genômico construídas para o sequenciamento pela plataforma Illumina.

Tamanho de insertos	Tipo de Biblioteca	Quantidade de bibliotecas
170 pb	PE	2
500 pb	PE	4
800 pb	PE	2
5 kb	MP	6
Total	-	14

PE: *paired ends*; MP: *mate pairs.*

2.2.2.2 Sequenciamento pela plataforma PacBio

O DNA genômico de alto peso molecular foi enviado para o *Génome Québec Innovation Centre,* na *McGill University* (Canadá). Foram construídas quatro bibliotecas, a partir de quatro amostras, cada uma contendo 10µg de DNA genômico, com tamanho médio de inserto de 20 kb. Cada uma das bibliotecas construídas foi sequenciada na plataforma RSII (PacBio) em 16 SMRT cells utilizando-se o kit P6-C4 *chemistry,* totalizando 64 SMRT cells.

Os dados de sequenciamento PacBio (*reads*) foram disponibilizados em três tipos de arquivos: *subreads.fastq*, *filtered_subreads.fastq* (FS) e *CCS_reads.fastq* (CCS). Os arquivos *subreads.fastq* contêm *reads* flanqueados por adaptadores que são utilizados no sequenciamento e correspondem ao resultado de uma única leitura realizada por uma única passagem da DNA polimerase pelo segmento. Os arquivos FS (*filtered subreads*) contêm *reads* sem as sequências dos adaptadores nas extremidades. Os arquivos CCS (*circular consensus sequence*) contêm as sequências consenso obtidas a partir do alinhamento de pelo menos dois *filtered_subreads* de um mesmo fragmento (Lee, 2017).

2.2.3 Controle de qualidade dos dados de sequenciamento

A avaliação da qualidade dos dados de sequenciamento obtidos pelas tecnologias Illumina e PacBio foi realizada utilizando-se o software FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc). Os *reads* Illumina foram submetidos ao controle de qualidade utilizando-se o software Trimmomatic (Bolger et al., 2014). Foram eliminadas as bases de baixa qualidade das extremidades 5' e 3' (funções LEADING e TRAILING) e dentro de um intervalo de quatro bases (função SLIDINGWINDOW) com valores de *phred* inferiores a 30. Foram selecionadas apenas sequências com comprimento superiores a 80pb (função MINLEN). Após a aplicação destes filtros de qualidade pelo software Trimmomatic, os dados foram reavaliados no software FastQC.

O s *reads* PacBio foram corrigidos utilizando-se os *reads* filtrados de alta qualidade das bibliotecas Illumina (*paired ends*), pelo uso do software Lordec (Salmela & Rivals, 2014), com os seguintes parâmetros -b 200, -e 0.4, -s 3 e -k 21, conforme sugerido por Hoang et al. (2017). As quantidades de *reads* disponíveis em cada biblioteca, antes e após a correção, estão na Tabela 2.2.

Tabela 2.2. Quantidade de *reads* PacBio disponíveis, antes e após a correção com *reads* Illumina de alta qualidade.

Reads PacBio	Total de *reads* não corrigidos	Total de *reads* corrigidos
CCS	258.362	255.147
FS	9.945.534	927.510

CCS: *circular consensus sequence*; FS: *filtered subreads.*

2.2.4 *Screening* de *reads* de cpDNA

O s *reads* obtidos pelas duas tecnologias de sequenciamento (Illumina e PacBio) foram alinhados contra sequências de referência de genomas de cloroplastos. Somente os *reads* mapeados foram utilizados nas etapas posteriores de análise. O alinhamento foi realizado utilizando-se o software BWA (*Burrows-Wheeler Alignment*) (Li & Durbin, 2009), com parâmetros *default* para ambos os tipos de *reads*, exceção feita à flag *-x* que foi

utilizada no alinhamento dos *reads* PacBio. Os arquivos contendo os subconjuntos de *reads* mapeados foram gerados pelo software Samtools (Li et al., 2009).

Nesta etapa de *screening* foram utilizadas como referência as sequências genômicas de cloroplastos de *Saccharum* spp., disponíveis no *GenBank,* das cultivares SP803280, NCo310 e Q155 (números de acesso: NC_005878.2, NC_006084.1 e NC_029221.1, respectivamente). As quantidades de *reads* disponíveis em cada biblioteca, após o *screening*, estão na Tabela 2.3.

Tabela 2.3. Quantidade de *reads* Illumina e PacBio disponíveis, antes e após o *screening* de sequências cloroplastidiais.

Bibliotecas	Total de *reads*	Total de *cp_reads*
PE170A	189.358.279	343.474
PE170B	188.953.086	340.860
PE500A	183.054.232	1.516.448
PE500B	187.302.075	1.532.606
PE500C	188.230.989	1.542.458
PE500D	176.094.426	1.406.830
PE800A	119.037.154	1.123.386
PE800B	142.746.285	1.344.328
CCS	255.147	4.473
FS	927.510	11.889

CCS: *circular consensus sequence*; FS: *filtered subreads;* cp_reads: *reads* mapeados em genomas cloroplastidiais.

2.2.5 *Assembly* do genoma cloroplastidial da cultivar RB867515

Os *reads* selecionados na etapa de *screening* foram utilizados no *assembly* do genoma cloroplastidial da cultivar RB867515. As montagens foram realizadas com os seguintes softwares: SPAdes (Bankevich et al., 2012), NOVOPlasty (Dierckxsens et al., 2016) e Organelle_PBA (Soorni et al., 2017). *Assemblers* distintos e diferentes combinações de *reads* foram testados a fim de se obter a sequência completa do cloroplasto da RB867515.

A inspeção de cada *assembly* foi realizada pelo realinhamento dos *reads* aos *contigs* obtidos. Os resultados destes alinhamentos foram visualizados no software Tablet (Milne et al., 2013).

2.2.6 Anotação dos genes no genoma cloroplastidial da cultivar RB867515

A sequência do genoma cloroplastidial correspondente ao melhor *assembly* obtido na etapa anterior foi submetida à anotação gênica utilizando duas bases de dados: *Dual Organellar Genome Annotator* (DOGMA) (Wyman et al., 2004) e GeSeq (Tillich et al., 2017). O resultado desta anotação, obtido sob a forma de um arquivo Genbank, foi utilizado para a visualização da estrutura circular do cloroplasto com seus respectivos genes no OrganellarGenomeDRAW (Lohse et al., 2013).

2.3 RESULTADOS E DISCUSSÃO

2.3.1 *Assembly* do genoma cloroplastidial da cultivar RB867515

Um conjunto de diferentes estratégias de *assembly* de genomas cloroplastidiais foi testado a fim de se validar a sequência obtida, captar eventuais variações entre elas e permitir comparações entre as estratégias. Todos os *assemblies* foram realizados com *reads* previamente filtrados para genomas de cloroplastos. Essa estratégia, apesar de ser baseada em um filtro, é considerada *assembly de novo* e tem sido utilizada com sucesso na obtenção de sequências completas de organelas (Nock et al., 2011; Evans & Joshi, 2016; Twyford & Ness, 2016). Um resumo dos resultados obtidos por cada uma das estratégias está apresentado na Tabela 2.4.

Tabela 2.4. Estatísticas descritivas relativas aos resultados dos *assemblies* obtidos para o genoma cloroplastidial da cultivar RB867515.

Estratégia	*Assembler*	Nº de *contigs*	Tamanho total de *contigs* (pb)	Maior *contig* (pb)
cp_MP_PE	SPAdes	147	174.221	23.208
PE_CCS	SPAdes.plasmid	1.128	5.742.844	68.929
cp_PE_FS	SPAdes	85	155.495	75.311
cp_PE	SPAdes	82	155.227	76.579
cp_PE_CCS	SPAdes	27	448.197	105.931
cp_MP_PE_CCS	SPAdes	27	402.008	108.625
CCS	Sprai e WGS-Assembler[1]	1	141.129	141.129
FS	Sprai e WGS-Assembler[1]	1	141.160	141.162

1 : *Assemblers* utilizados no *pipeline* Organelle_PBA; cp_: indica que os *reads* utilizados no *assembly* são provenientes do filtro realizado para genomas cloroplastidiais; PE: *paired ends;* MP: *mate pairs;* CCS: *circular consensus sequence;* FS: *filtered subreads.*

2.3.2 *Assemblies* Illumina

O s *reads* Illumina alinhados às sequências de referência de genomas de cloroplasto apresentaram uma alta cobertura, superior a 2000X. Apesar desta cobertura elevada, os *assemblies* realizados somente com *reads* Illumina não permitiram a montagem do genoma em um único *contig,* produzindo *assemblies* fragmentados. Dificuldades semelhantes foram reportadas no *assembly* do cloroplasto de *Potentilla micrantha* que, apesar da alta cobertura utilizada, de modo similar, não permitiu a obtenção da sequência do genoma cloroplastidial sob a forma de um único *contig* (Ferrarini et al., 2013).

Uma das possíveis explicações para a dificuldade de obtenção de um único *contig* representando a sequência completa do genoma cloroplastidial, apesar da elevada cobertura, pode estar relacionada ao filtro realizado antes do assembly, que pode ter resgatado *reads* de cloroplastos inseridos no núcleo e na mitocôndria (Sabir et al., 2014). Quando um segmento de cpDNA é incorporado em outro compartimento genômico as taxas de mutação nas extremidades do segmento são aumentadas, o que faz com que somente uma comparação das extremidades dos *reads* originalmente de cloroplastos os

diferenciem dos *reads* incorporados no núcleo ou na mitocôndria (Garaycochea et al., 2015).

O algoritmo NOVOPlasty utiliza apenas *reads* Illumina e oferece uma abordagem baseada na cobertura de *reads,* sem a necessidade de realizar nenhuma etapa anterior de filtro. Para esse assembler foram utilizados os *reads* não filtrados para sequências de cloroplastos, correspondentes às bibliotecas *paired ends* PE170A e PE170B (unidas), pois este assembler não suporta a montagem com múltiplas bibliotecas com tamanhos de fragmentos diferentes (Dierckxsens et al., 2016).

A montagem dos genomas organelares no NOVOPlasty começa com a sequência *seed* que atua como uma âncora na extensão bidirecional do *assembly*. O *seed* pode ser uma sequência de referência de outra organela filogeneticamente próxima ou apenas um *read* que pertença à organela que se deseja montar. O montador então é capaz de circularizar a sequência se houver uma sobreposição de 200 pb entre as extremidades da sequência final obtida (Dierckxsens et al., 2016).

O *assembly* com o NOVOPlasty foi realizado utilizando-se como *seed input* a sequência completa do genoma cloroplastidial da cultivar Q155. O assembler produziu 14 diferentes versões do genoma cloroplastidial. A análise comparativa destas 14 sequências indicou se tratarem de pares de sequências que diferem entre si somente pela orientação da última região de repetição invertida (IRb). Eliminadas as redundâncias, foram então consideradas oito versões do genoma de cloroplasto. Essas oito sequências quando alinhadas à sequência de referência apresentaram cobertura média de 80%, com uma região de ~50kbp sem nenhuma similaridade com a referência. Os genes dessas sequências foram então anotados e nessa região de ~50kbp só havia um gene, *cox1* (gene mitocondrial). Estes resultados sugerem que os *contigs* obtidos neste caso não podem ser considerados como confiáveis para fins de análise da estrutura do genoma cloroplastidial e foram descartados.

A estratégia de *assembly* utilizada pelo NOVOPlasty é baseada em cobertura, pois sabe-se que as coberturas de sequenciamento são bastante distintas para fragmentos nucleares, cloroplastidiais e mitocondriais. A análise do conjunto de *reads* filtrados para aplicação de outras estratégias de montagem dos genomas organelares (cloroplastidial e mitocondrial) permitiu a observação de que a quantidade de *reads* mitocondriais filtrados (dados não mostrados) foi maior do que a de *reads* cloroplastidiais. Isso provavelmente se

deve ao fato de que as amostras de DNA foram obtidas a partir de gemas laterais frescas, e não de folhas, que é onde são encontradas cópias de cloroplastos em maior abundância.

2.3.3 *Assemblies* híbridos: *reads* Illumina e PacBio

Os melhores resultados de *assemblies* envolvendo *reads* Illumina foram obtidos com estratégias híbridas, executadas no SPAdes utilizando *reads* PacBio CCS com o parâmetro *single read* (flag –s), como recomendado no manual do software. A utilização das *flags* –pacbio para CCS *reads* ou FS *reads* não produziu bons resultados.

Apesar dos *assemblies* híbridos terem melhorado consideravelmente o tamanho dos *contigs* obtidos, estas estratégias não foram capazes de produzir um único *contig* com a sequência completa do genoma cloroplastidial. As estratégias cp_PE_CCS e cp_MP_PE_CCS (Tabela 2.5) recuperaram a sequência completa do genoma cloroplastidial em dois *contigs*: um contendo as regiões LSC e IRa, e o outro contendo a região SSC (Figura 2.1).

A região IRb não foi recuperada por nenhuma destas estratégias. Isso pode ser explicado pela elevada identidade de sequência do par de IRs, IRa e IRb, que são tratadas pelos montadores como uma única sequência. Como resultado, os montadores acabam por, equivocadamente, colapsar estas regiões em uma única, que se apresenta com o dobro da cobertura de sequenciamento (Twyford & Ness, 2016). Em muitos trabalhos a sequência final do *assembly* é obtida pela edição manual das sequências dos *contigs* obtidos, considerando-se as regiões IRa e IRb como regiões de repetição idênticas e invertidas. A utilização desta estratégia, porém, implica na omissão de quaisquer variações entre as IRs que podem ocorrer em maior ou menor escala, como a duplicação de um IR inteiro, ou de seus fragmentos, e até mesmo a perda de um IR, resultando em um cloroplasto com estrutura tripartida (Wu & Chaw, 2014; Lei et al., 2016; Ruhlman et al., 2017).

Tabela 2.5. Estatísticas descritivas relativas aos resultados dos *assemblies* híbridos obtidos para o genoma cloroplastidial da cultivar RB867515.

contig	Estratégia	Tamanho (pb)	Cobertura (%)	*Mismatches*	*Gaps*
contig1	cp_PE_CCS	105.931	75,0	0	0
	cp_MP_PE_CCS	108.625	60,8	3	2
contig2	cp_PE_CCS	12.544	8,9	0	0
	cp_MP_PE_CCS	12.544	8,9	0	0

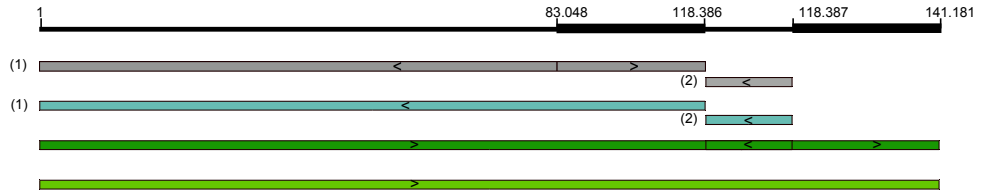

Figura 2.1. Representação esquemática do alinhamento dos *contigs* obtidos por diferentes estratégias de *assembly* da sequência do genoma cloroplastidial da cultivar RB867515. <: *reverse*; >: *forward*. Barras de mesma cor representam *contigs* obtidos pela mesma estratégia. A linha preta representa a sequência de referência da cultivar Q155. As barras mais espessas representam as regiões IRa e IRb. O segmento entre elas representa a região SSC. Barra cinza: estratégia cp_MP_PE_CCS (*contig1* e *contig2*). Barra *cian*: estratégia cp_PE_CCS (*contig1* e *contig2*). Barra verde escuro: estratégia Organelle_CCS. Barra verde claro: estratégia Organelle_FS.

Diferentes orientações para a região SSC são relatadas na literatura e umas das hipóteses para essa inversão é a de que trata-se de um *hotspot* para eventos de inversão. Análises com enzimas de restrição demonstraram que o cpDNA de plantas individuais exibem uma forma de heteroplasmia em que o cloroplasto existe em dois estados equimolares que diferem na orientação relativa da região SSC (Palmer, 1983; Walker et al., 2015).

Os resultados obtidos nos dois *assemblies* híbridos, em que a sequência do genoma cloroplastidial foi completamente recuperada em dois *contigs*, sendo um dos *contigs* correspondente exato à região SSC, sugerem que essa região possa estar presente

em dois estados, o que justificaria o impedimento da sua montagem pelos *assemblers* como parte integrante do maior *contig*.

O algoritmo plasmidSPAdes, segundo informações de seus desenvolvedores, é capaz de detectar e montar plasmídeos em sequências circulares (Antipov et al., 2016). Para a aplicação deste algoritmo foi utilizado todo o conjunto de *reads* Illumina não filtrados, em conjunto com os *reads* PacBio. Os *contigs* obtidos não apresentaram uma boa cobertura do genoma cloroplastidial (dados não mostrados).

2.3.4 *Assembly* utilizando *reads* PacBio

Os *assemblies* obtidos somente com *reads* PacBio foram realizados através do *pipeline* Organelle_PBA, que foi desenvolvido para *assembly* de genomas organelares a partir de *reads* PacBio, utilizando um genoma de referência. Trata-se de um *pipeline* escrito em Perl, sob a forma de um um único *script*, que usa diversos outros programas utilizados em diferentes contextos de análises genômicas como: BlasR (mapeamento de *reads*), Samtools (manipulação de arquivos sam), Blast, Sprai e WGS-Assembler (*assembly*), SSPACE-Long (*rescaffolding*), BEDtools (cálculo de cobertura) e Seqtk (conversão de arquivos fastq para fasta e seleção de *reads* fastq para o Sprai) (Soorni et al., 2017).

Foram realizados dois *assemblies* com esse pipeline, um com os *reads* CCS e o outro com os *reads* FS, ambos corrigidos. *A priori* não eram esperadas diferenças entre os resultados obtidos por estas duas estratégias, haja visto que os *reads* CCS representam as sequências consenso dos alinhamentos dos *reads* FS de mesmos fragmentos.

O *assembly* com *reads* CCS apresentou uma cobertura de 153X e produziu um único *contig* de 141.129 pb, tamanho correspondente ao genoma cloroplastidial inteiro. A comparação da sequência obtida neste caso com a sequência de referência da cultivar Q155 apresentou oito *mismatches* e 43 deleções, com a região SSC invertida (Figura 2.1).

O *assembly* obtido com os *reads* FS produziu um único *contig* de 141.162 pb, tamanho também correspondente ao genoma cloroplastidial inteiro. Os dados de sequenciamento neste caso foram provenientes de 927.510 *reads* (5.061.924.684 pb), dos quais 11.889 *reads* (67.974.778 pb) foram selecionados pelo alinhamento com a sequência

de referência, com cobertura estimada em 481X. A sequência do *contig* obtido tem 19 deleções quando comparada à sequência de referência da cultivar Q155.

Os *assemblies* obtidos utilizando somente *reads* PacBio recuperaram todas as regiões do cloroplasto (LSC, IRa, SSC e IRb) em um único *contig*, porém com elevadas taxas de deleção. É possível que estas deleções sejam consequência de viés presente nos *reads* produzidos pela plataforma PacBio, mesmo se tratando de *reads* que foram previamente corrigidos. A plataforma PacBio produz comprimentos de *reads* bem maiores do que as tecnologias de segunda geração (Illumina), porém as sequências brutas produzidas a partir dessa plataforma são propensas a taxas de erro de até 17,9%, sendo a maioria dos eventos de erros correspondentes a alterações do tipo *indel* (Ferrarini et al., 2013).

Apesar deste viés, os *reads* PacBio tem sido vantajosos na resolução de regiões repetitivas, exatamente como ocorre com as duas regiões de repetição invertida de cloroplastos (Stadermann et al., 2015). Neste sentido, esta tecnologia tem sido recentemente utilizada para obtenção de sequências completas de genomas de cloroplastos, como de *Rhazya stricta* (Park et al., 2014) , *Beta vulgaris* (Stadermann et al., 2015), *Eucommia ulmoides* (Wang et al., 2016), *Amaranthus* spp. (Chaney et al., 2016) e *Vernicia fordii* (Li et al., 2017).

2.3.5 O genoma cloroplastidial da cultivar RB867515

Para integração das informações obtidas pelos diferentes *assemblies* produzidos, utilizou-se como ponto de partida os resultados obtidos pela estratégia cp_PE_CCS, por esta estratégia ter produzido uma sequência quase idêntica à de referência utilizada. Este *assembly* foi constituído por dois *contigs* (*contig1*: LSC+IRa, *contig2*: SSC) (Figura 2.1). A IRb não foi montada, porém sua presença foi confirmada nos outros *assemblies* PacBio. O *contig2*, correspondente à região SSC, foi então integrado ao *contig1* e a sequência da região IRb editada manualmente.

Em todos os *assemblies* obtidos a região SSC foi montada sozinha em um *contig* isolado ou, quando integrada ao *contig* maior, estava disposta de maneira invertida em relação à referência. O realinhamento dos *reads* (Illumina e PacBio) às regiões correspondentes às junções entre IRa-SSC e SSC-IRb foi então realizado para se

determinar a orientação correta deste segmento que, conforme Palmer (1983), pode existir em diferentes polaridades configurando um tipo de heteroplasmia de cloroplastos dentro de indivíduos (Figuras 2.2 e 2.3) (Walker et al., 2015).

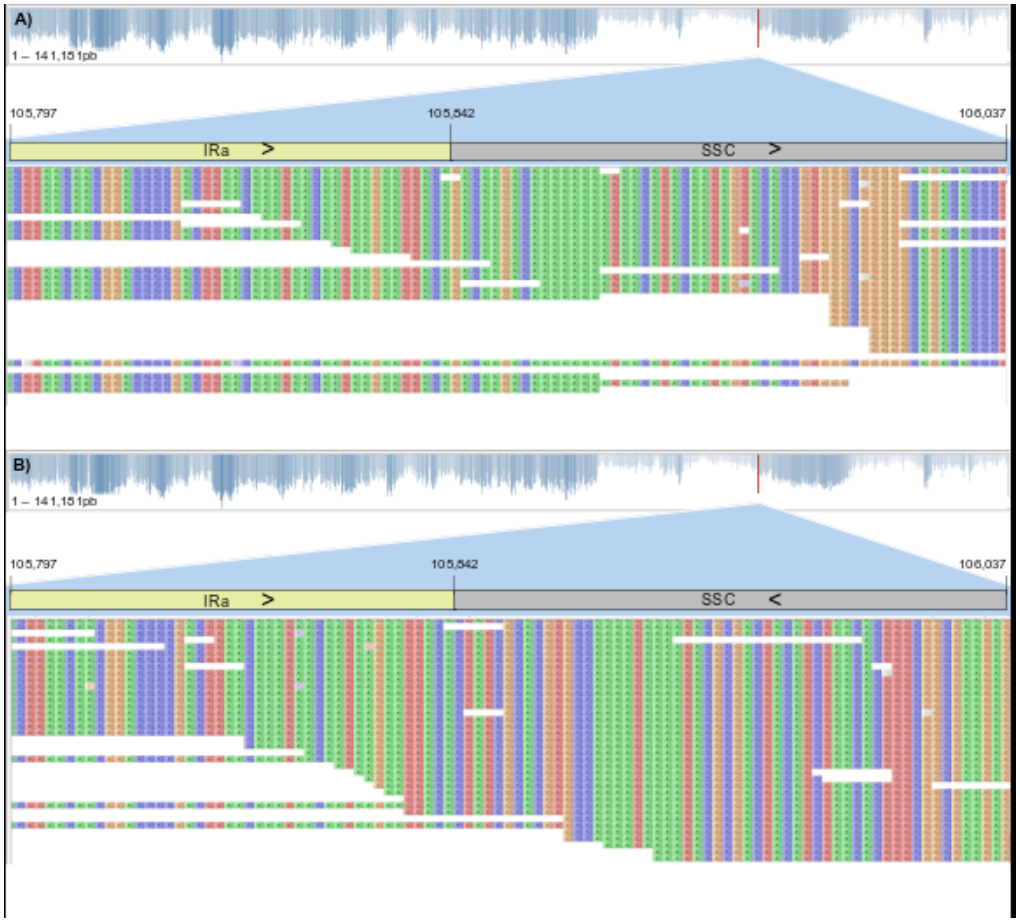

Figura 2.2. Alinhamento de reads PacBio na junção IRa-SSC. A) Isômero de cloroplasto com SSC na orientação *forward* e B) Isômero de cloroplasto com SSC na orientação *reverse*.

Com base nessas alinhamentos foi possível se verificar que existem dois tipos de cloroplastos nas células da cultivar RB867515, que diferem entre si pela orientação da *small single copy* (SSC), resultando em duas populações distintas de cloroplastos no mesmo indivíduo. Esse tipo de evento genético foi descrito primeiramente em *Phaseolus vulgaris* (Fabaceae) (Palmer, 1983) e posteriormente em *Zea mays* (Oldenburg & Bendich, 2004) e *Musa acuminata* (Martin et al., 2013).

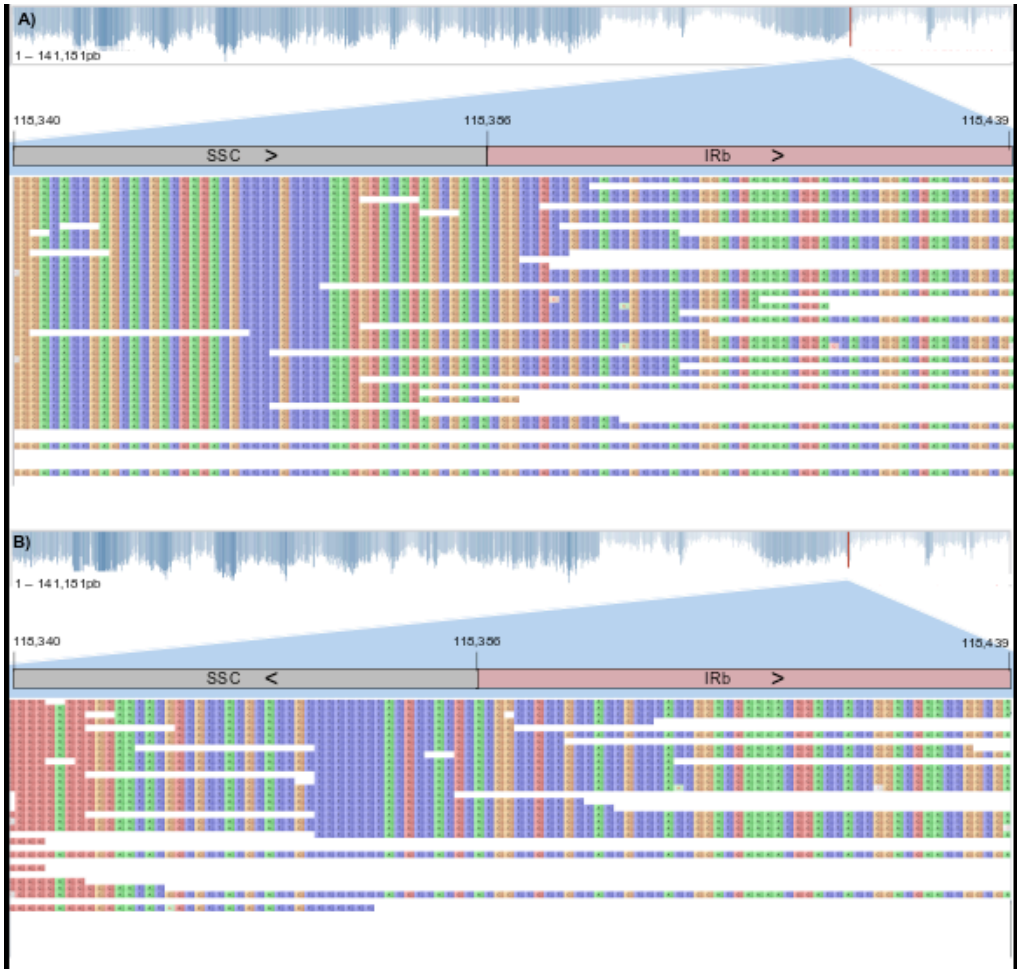

Figura 2.3. Alinhamento de reads PacBio na junção SSC-IRb. A) Isômero de cloroplasto com SSC na orientação *forward* e B) Isômero de cloroplasto com SSC na orientação *reverse*.

2.3.6 Caracterização das isoformas do cloroplasto da cultivar RB867515

Os cloroplastos da cultivar RB867515 são compostos por dois isômeros, que diferem entre si quanto à orientação da região SSC. Cada molécula possui 141.181 pb e exibe uma estrutura quadripartida típica de cloroplastos de angiospermas (Bock, 2007). Essa estrutura inclui uma região longa de cópia única (LSC) de 83.047 pb, flanqueada por duas regiões de repetições invertidas (IRs) de 22.795 pb, e uma pequena região de cópia única (SSC) entre as IRs de 12.544 pb (Figura 2.4).

O tamanho estimado para o genoma cloroplastidial sequenciado é igual ao da cultivar Q155 (141.181 pb) (Hoang et al., 2015) e um nucleotídeo menor que aquele das cultivares NCo310 e SP803280 (141.182 pb) (Asano et al., 2004; Calsa Júnior et al., 2004).

As cultivares modernas de cana-de-açúcar são híbridos interespecíficos entre *S. officinarum* (genitor feminino) e *S. spontaneum* (genitor masculino) e é esperado que o genoma cloroplastidial seja originado de *S. officinarum*.

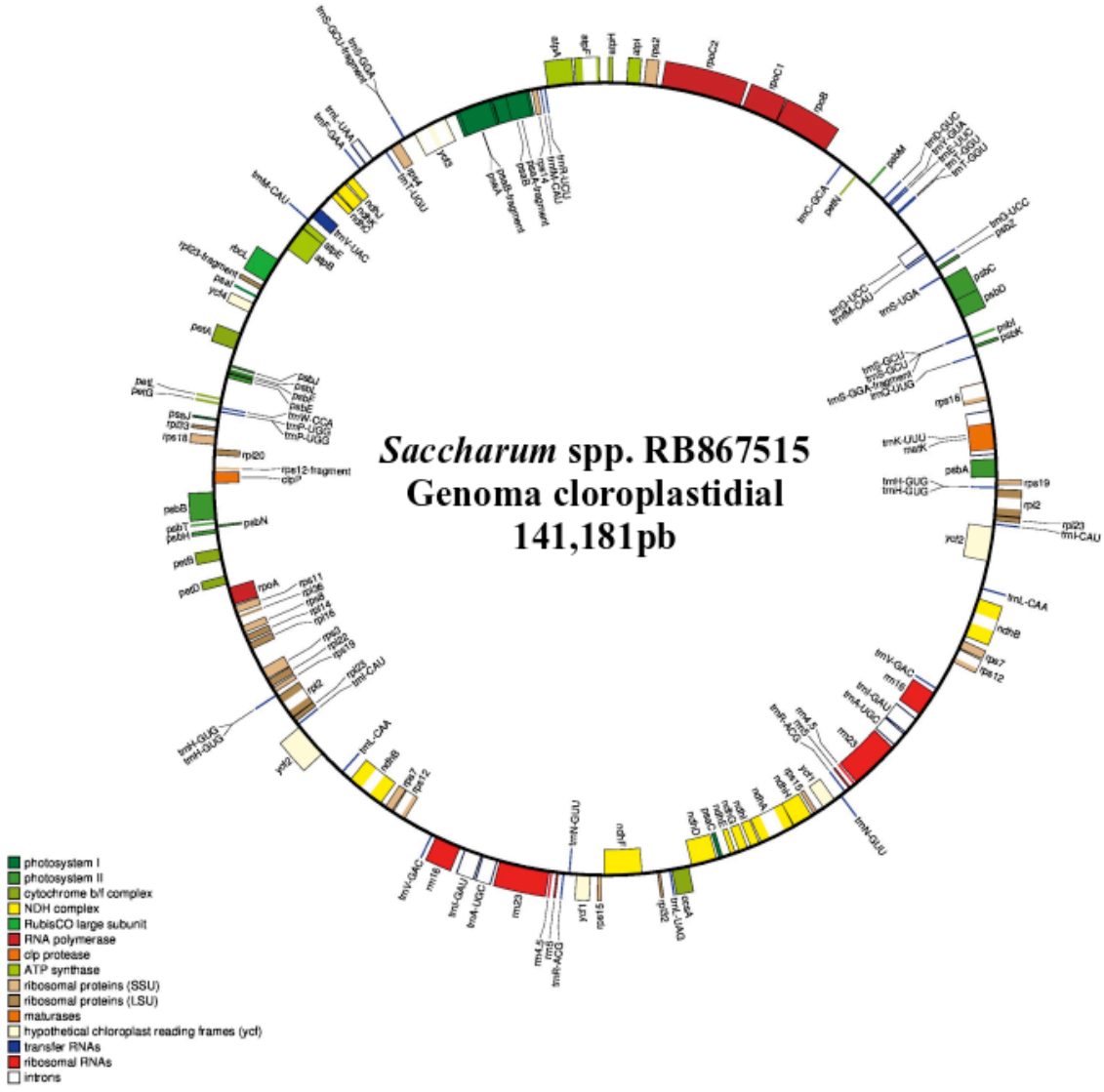

Figura 2.4. Representação esquemática de uma das isoformas do cloroplasto da cultivar de cana-de-açúcar RB867515. Cores diferentes indicam diferentes grupos funcionais de genes.

Os resultados obtidos estão de acordo com os encontrados para a cultivar Q155, que é exatamente equivalente, em cada base, à sequência consenso do cpDNA dos outros dois genótipos com sequências disponíveis (NCo310 e SP803280) (Hoang et al., 2015). Entretanto, as análises realizadas neste trabalho permitiram, de maneira inédita com

dados de sequenciamento de nova geração, se identificar a presença de duas isoformas do genoma cloroplastidial da cana-de-açúcar.

Evans & Joshi (2016) analisaram em conjunto cinco genomas de cloroplastos do Complexo *Saccharum* provenientes de acessos de *S. officinarum* (IJ76-514) (141.176 pb), *S. spontaneum* (SES234B) (141.185 pb), *M. floridulus* (141.356 pb) e duas cultivares híbridas (RB8672454 e Q165). Estes tamanhos reportados para os genomas cloroplastidiais diferem dos obtidos para as cultivares RB867515 e Q155 (141.181 pb) e para as cultivares NCo310 e SP803280 (141.182 pb).

Houve variações mínimas de comprimento dos genomas cloroplastidiais de representantes do gênero *Saccharum,* variando de 141.151 a 141.182 pb. O plastídeo mais longo foi de *M. Floridulus* (141.356 pb), que se comparado a *S. spontaneum* difere em 172 pb. O gênero *Saccharum* é monofilético e irmão de *Miscanthus,* com *M. floridulus* e linhagens de *Saccharum* divergindo há cerca de 3,8 milhões de anos (Evans & Joshi, 2016).

O genoma cloroplastidial de *S. officinarum*, cana selvagem, comparado ao das modernas cultivares híbridas possui um fragmento duplicado no início da região LSC de 1.031 pb, além de uma inserção de 10 pb no gene *rpl23-F* e duas cópias dos genes *orf137*, *trnT*, *orf74* e *rps19* (Paes et al., 2017).

O conteúdo de A+T (61,7%) obtido para o genoma cloroplastidial da cultivar RB867515 é concordante com os valores encontrados em NCo310 e SP803280 (Asano et al., 2004; Calsa Júnior et al., 2004) e é similar ao encontrado em arroz (61,0%), milho (61,5%) e trigo (61,7%).

De modo geral, o número, o conteúdo e a ordem dos genes funcionais dos cloroplastos de cana-de-açúcar estudados até o momento são idênticos aos de arroz, trigo e milho, apresentando maior similaridade com milho, pois as IRs em ambos são expandidas devido a sequências adicionais (*junk* DNA) (Hiratsuka et al., 1989; Maier et al., 1995; Ogihara et al., 2001; Asano et al., 2004). A anotação gênica do genoma cloroplastidial da cultivar RB867515 resultou em 78 CDSs, 4 rRNAs e 29 tRNAs únicos (Tabela 2.6).

Tabela 2.6. Características gerais do genoma cloroplastidial da cultivar RB867515.

Comprimento (pb)	141.181
Conteúdo A (%)	30,83
Conteúdo C (%)	19,18
Conteúdo G (%)	19,26
Conteúdo T (%)	30,73
CDS LSC	*petG, atpA, atpB, atpE, atpF, atpH, atpI, cemA, clpP, infA, matK, ndhC, ndhJ, ndhK, pbf1, petA, petB, petD, petL, petN, psaA, psaB, psaI, psaJ, psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ, rbcL, rpl14, rpl16, rpl20, rpl22 rpl33, rpl36, rpoA, rpoB, rpoC1, rpoC2, rps11, rps14, rps16, rps18, rps2, rps3, rps4, rps8, ycf3, ycf4*
CDS IRs	(2)*rps19,* (2)*rpl2,* (2)*rpl23,* (2)*ycf2,* (2)*ndhB,* (2)*rps7,* (2)*rps12,* (2)*rps15*
CDS SSC	*ndhF, rpl32, ccsA, ndhD, psaC, ndhE, ndhG, ndhI, ndhA, ndhH*
rRNA	(2)*rrn16,* (2)*rrn23,* (2)*rrn4.5,* (2)*rrn5*
tRNA	(2)*trnA-UGC, trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA,* (2)*trnfM-CAU, trnG-GCC,* (2)*trnH-GUG trnI-CAU,* (2)*trnI-GAU, trnK-UUU,* (2)*trnL-CAA, trnL-CAA, trnL-UAG,* (3)*trnM-CAU,* (2)*trnN-GUU, trnP-UGG, trnQ-UUG,* (2)*trnR-ACG, trnR-UCU, trnS-GCU, trnS-GGA, trnS-UGA, trnT-GGU, trnT-UGU,* (2)*trnV-GAC, trnV-UAC, trnW-CCA, trnY-GUA*

* Os números entre parênteses representam a quantidade de cópias de cada gene.

2.4 CONCLUSÕES

As dificuldades encontradas na obtenção da sequência completa de genomas de cloroplastos podem estar relacionadas à presença de polimorfismos estruturais nessas sequências. De maneira inédita, neste trabalho se relatam evidências baseadas na análise de de dados de sequenciamento de nova geração de que os cloroplastos de cana-de-açúcar

apresentam duas isoformas distintas, com diferentes orientações da região SSC, o que configura heteroplasmia individual.

A cultivar RB867515 possui dois cloroplastos, isômeros entre si, que diferem quanto à orientação da SSC. Cada molécula possui 141.181 pb e exibe uma estrutura quadripartida típica, que inclui uma região longa de cópia única (LSC) de 83.047 pb, flanqueada por duas regiões de repetições invertidas (IRs) de 22.795 pb, e uma pequena região de cópia única (SSC) entre as IRs de 12.544 pb. A sequência obtida é idêntica à da cultivar australiana Q155.

3 MONTAGEM E CARACTERIZAÇÃO DO GENOMA MITOCONDRIAL DE *Saccharum* spp. (cultivar RB867515) UTILIZANDO DADOS DE SEQUENCIAMENTO DE NOVA GERAÇÃO

RESUMO

O Brasil é o maior produtor mundial de cana-de-açúcar, com produção anual de 768 milhões de toneladas. Cultivada em mais de 100 países, a cana-de-açúcar é uma das culturas mais importantes da região tropical e subtropical do mundo por ser matéria-prima de produtos de grande importância econômica como o açúcar e o etanol. Dada sua importância, diversos esforços vêm sendo realizados com o objetivo de se realizar a caracterização genômica de cultivares de cana-de-açúcar. Os genomas eucarióticos são distribuídos em diferentes compartimentos genéticos que apresentam padrões distintos de herança. Algumas organelas, como plastídeos e mitocôndrias, possuem sistema genético próprio, contendo DNA, RNA e os componentes necessários para os processos de replicação, transcrição e síntese de proteínas que ocorrem em seu interior. As informações sobre a estrutura dos genomas mitocondriais são úteis em estudos evolutivos e de identificação de origem materna. O sequenciamento e análise desses genomas pode fornecer uma visão da história evolutiva da cana-de-açúcar além de ser uma ferramenta útil na solução de problemas práticos de melhoramento genético. O genoma mitocondrial da cultivar RB867515 foi obtido utilizando dados de sequenciamento das tecnologias Illumina e PacBio. As estratégias de montagem foram baseadas na utilização de dados obtidos a partir do *screening* de sequências mitocondriais realizado pelo alinhamento de *reads* obtidos do sequenciamento do DNA genômico total a sequências de referência de genomas mitocondriais de outras espécies vegetais, depositadas em banco de dados públicos. Os *assemblies* foram obtidos utilizando-se os softwares SPAdes e Organelle_PBA. A anotação gênica foi obtida utilizando-se as ferramentas DOGMA e GeSeq. O genoma mitocondrial da cultivar RB867515 é composto por dois cromossomos: o cromossomo 1 com comprimento de 300.765 pb e o cromossomo 2 com comprimento de 194.383 pb. Os conteúdos GC (~44%) e AT (~56%) são concordantes com aqueles de outras angiospermas. Foram anotados 39 CDS, 5 genes hipotéticos conservados, 5 rRNAs, 18 tRNAs e 9 fragmentos de genes prováveis resultantes de transferência horizontal de cloroplastos. Os cromossomos mitocondriais da cultivar RB867515 não são idênticos ao da sua genitora materna. A comparação com outros genomas mitrocondriais de *S. officinarum* permitiu a identificação de SNPs, duplicações gênicas e de eventos de expansão genômica.

Palavras-chave: cana-de-açúcar, genoma mitocondrial, *assembly*.

3.1 INTRODUÇÃO

As mitocôndrias são oriundas de uma α-proteobactéria endossimbionte adquirida por um organismo incapaz de realizar respiração aeróbia (Embley & Martin, 2006). O genoma mitocondrial de plantas é maior e mais complexo comparado a outros de eucariotos unicelulares ou multicelulares pois acumulam sequências repetitivas e segmentos de DNA exógenos por transferência horizontal (Alverson et al., 2010; Smith, 2016).

A topologia do genoma mitocondrial de plantas é bastante diversa, sendo possível encontrar moléculas lineares ou circulares, únicas ou concatenadas (Oldenburg & Bendich, 2015). O genoma mitocondrial de angiospermas é historicamente descrito como uma única molécula de DNA circular que aloja o conjunto completo de genes chamado de "cromossomo mestre". Cada cromossomo mestre é composto por várias sequências repetidas onde ocorrem os eventos de recombinação (Sloan, 2013).

O genoma mitocondrial possui uma organização multipartida e é composto por regiões de repetições longas e regiões de repetições curtas. As regiões de repetições longas são sujeitas a altas taxas de recombinação. Além disso, o genoma mitocondrial recebe constantemente DNA de plastídeos, do núcleo e até mesmo de outras espécies, o que o torna ainda mais diverso e heterogêneo (Kubo & Newton, 2008). Múltiplas cópias do genoma mitocondrial estão presentes nas células. Em plantas, uma célula foliar pode conter até 100 cópias destes genomas (Wang et al., 2012).

O tamanho dos genomas mitocondriais de plantas varia de cerca de 200 kb, como em *Brassica hirta*, a pouco mais de 11 Mb, como em *Silene conica* (Shearman et al., 2016). A expansão dos genomas mitocondriais ocorre principalmente devido ao acúmulo de sequências repetidas, expansão de íntrons e incorporação de DNA exógeno de plastídeos ou do núcleo. O acúmulo de sequências repetitivas pode levar a eventos de recombinação e rearranjos genômicos constantes dentro da mesma espécie, promovendo o surgimento de múltiplos filamentos de DNA circular com sobreposição de sequências e variação no números de cópias de cada segmento (Richardson & Palmer, 2007; Alverson et al., 2010; Goremykin et al., 2012).

Nos casos de múltiplos filamentos circulares de DNA mitocondrial, o genoma completo é referido como círculo principal ou círculo mestre e os demais filamentos de

DNA derivados de recombinação são designados minicírculos subgenômicos (Sloan, 2013). Podem ocorrer situações em que o círculo principal não existe e o genoma é constituído por múltiplos filamentos circulares de DNA sem nenhuma sequência compartilhada que pudesse facilitar a recombinação (Alverson et al., 2010; Sloan, 2013; Shearman et al., 2016).

A análise comparativa dos genomas mitocondriais de plantas melhora a nossa compreensão sobre mecanismos de rearranjos genômicos e de transferência horizontal de DNA (Wei et al., 2016). O número de genes mitocondriais em angiospermas, sem se considerar o número de cópias, varia de 50 a 60. A variação no número de genes é devida ao conteúdo gênico das subunidades do complexo II, aos genes codificantes de proteínas ribossomais e aos genes de tRNAs. Quando o conteúdo de genes codificantes de proteínas ribossomais é comparado entre angiospermas, fica evidente a perda de genes mitocondriais durante a evolução. Muitos desses genes perdidos foram transferidos para genomas nucleares, e raramente para genomas de cloroplastos (Adams & Palmer, 2003; Daniell et al., 2016).

A análise de dados genômicos mitocondriais contribui para a melhor compreensão da biologia de eucariotos, suas origens, diversidade e complexidade. Os genes mitocondriais estão entre os marcadores genéticos mais utilizados em estudos evolutivos, tanto para estudos em nível populacional quanto para análises comparativas em escalas mais abrangentes, como aquelas que buscam a reconstrução da árvore filogenética de todos os eucariotos (Smith, 2016).

As mitocôndrias contribuem para o metabolismo energético e desempenham papéis fundamentais no desenvolvimento, aptidão e reprodução da planta, bem como estão associadas à biossíntese de ácidos graxos e várias proteínas ativas (Ryan & Hoogenraad, 2007). Mutações em genes mitocondriais podem ser letais ou prejudicar irreversivelmente o desenvolvimento de plantas, como ocorreu com milho (*Zea mays*), através de uma deleção no gene *nad4*. A maioria das plantas que carregam deleções em genes mitocondriais essenciais para a respiração celular só sobrevivem em heteroplasmia, caso em que ocorrem haplótipos normais e mutantes na mesma planta (Yamato & Newton, 1999).

O primeiro genoma mitocondrial de cana-de-açúcar a ser sequenciado foi o da espécie *Saccharum officinarum,* cultivar *Khon Kaen 3*, desenvolvida na Tailândia. A

sequência obtida revelou um genoma composto por dois cromossomos circulares distintos. O maior cromossomo (cromossomo 1) é constituído de 300.778 pb, com 15 kb de sequência repetitiva. O cromossomo menor (cromossomo 2) possui 144.698 pb (Shearman et al., 2016).

Esse tipo de genoma mitocondrial, com múltiplos cromossomos independentes é conhecido como multicromossômico (Sloan, 2013). Tal estrutura foi inicialmente descrita para *Cucumis sativus* (pepino), que de modo similar à cana, possui um cromossomo grande de 1.6 Mb e dois cromossomos pequenos de 45-84 Kb, que não compartilham sequências com o cromossomo mestre (Alverson et al., 2010).

A cana-de-açúcar é a principal matéria-prima para a produção de açúcar no mundo, sendo o Brasil o maior produtor mundial, com 768 milhões de toneladas anuais (Rao et al., 2016). As variedades cultivadas são híbridos derivados de duas ou mais espécies de *Saccharum*, geralmente uma combinação de *S. spontaneum* como genitor masculino e *S. officinarum* como genitor feminino. Deste modo, tipicamente, as cultivares modernas apresentam seus genomas organelares provenientes desta última espécie (Raj et al., 2015).

Estudos de genomas mitocondriais são úteis para a identificação de origem materna. Em cana-de-açúcar, o sequenciamento e análise desses genomas pode fornecer informações históricas importantes a respeito das espécies que constituem o complexo *Saccharum* (Shearman et al., 2016). Neste contexto, o presente trabalho objetivou realizar a montagem e a anotação do genoma mitocondrial da cultivar de cana-de-açúcar RB867515.

3.2 MATERIAL E MÉTODOS

3.2.1 Extração e quantificação de DNA genômico

O preparo das amostras de DNA genômico foi realizado no Laboratório de Genética e Genômica de Plantas, no Centro de Excelência em Melhoramento Genético da Cana-de-açúcar no Cerrado, localizado na Escola de Agronomia da Universidade Federal de Goiás. O DNA foi obtido a partir de gemas laterais frescas de clones da cultivar RB867515. As gemas laterais selecionadas foram retiradas com o auxílio de uma espátula,

higienizadas e mantidas em freezer -80ºC até a etapa de extração do DNA conforme Aljanabi et al. (1999).

A integridade do DNA extraído foi verificada pela comparação com diferentes quantidades de amostras-padrão (50, 100 e 200 ng) e DNA de fago λ digerido com *HindIII*. Os fragmentos foram visualizados por eletroforese horizontal em gel de agarose 1% corado em brometo de etídeo (0,5 μg/μL). Somente amostras com fragmentos de tamanho superior a 20 kb foram selecionadas para as etapas seguintes. A quantificação de DNA obtido em cada extração foi realizada com fluorímetro Qubit®. Como medida de pureza foi utilizada a relação de absorbâncias a 260 nm e 280 nm, obtidas pelo espectrofotômetro Nanodrop®. Somente amostras com valores de pureza entre 1,8 e 2,0 foram selecionadas para as etapas seguintes de sequenciamento.

3.2.2 Sequenciamento de DNA

3.2.2.1 Sequenciamento pela plataforma Illumina

As amostras foram dispostas em microtubos de 1,5μL contendo 200μg de DNA genômico e enviadas a empresa BGI (China) onde foram construídas as bibliotecas e foi realizado o sequenciamento de nova geração na plataforma *HiSeq2000* da Illumina®, pelas estratégias *paired ends* e *mate pairs* (2 x 100pb).

Tabela 3.1. Bibliotecas de DNA genômico construídas para fins de sequenciamento pela plataforma Illumina.

Tamanho de fragmentos	Estratégia	Quantidade
170 pb	PE	2
500 pb	PE	4
800 pb	PE	2
~5 kb	MP	6
Total		14

PE: *paired ends,* **MP:** *mate pairs.*

3.2.2.2 Sequenciamento pela plataforma PacBio RS-II

Neste caso o DNA genômico foi enviado para o *Génome Québec Innovation Centre* na *McGill University* (Canadá). Quatro bibliotecas foram construídas a partir de 40µg de DNA genômico de alto peso molecular (>20 kb). Cada uma das bibliotecas construídas foi sequenciada em 16 *SMRT cells* utilizando o kit P6-C4 *chemistry,* totalizando *64 SMRT cells*.

3.2.3 Controle de qualidade dos dados de sequenciamento

A qualidade das sequências obtidas pelas tecnologias Illumina e PacBio foi avaliada utilizando-se o software FastQC[1]. O controle de qualidade dos *reads* Illumina foi realizado pela utilização do software Trimmomatic (Bolger et al., 2014). Inicialmente, foram eliminadas as bases de baixa qualidade (*phred*<30) a partir das extremidades 5' e 3' (LEADING e TRAILING). Além disso, os *reads* foram truncados sempre que o valor de *phred* médio de um intervalo de quatro bases consecutivas (SLIDINGWINDOW) foi inferior a 30. Foram selecionadas apenas sequências com comprimento superiores a 80 pb (MINLEN). Após essas alterações realizadas no Trimmomatic, os *reads* foram reavaliados utilizando-se o FastQC.

Os *reads* PacBio foram corrigidos utilizando os *reads* de todas as bibliotecas lllumina *paired end* pelo software Lordec (Salmela & Rivals, 2014), com os seguintes parâmetros: -b 200, -e 0.4, -s 3 e -k 21 conforme sugerido por Hoang et al., (2017). As quantidades de *reads* disponíveis antes e após a etapa de correção estão na Tabela 3.2.

Tabela 3.2. Resumo da correção de reads PacBio utilizando reads Illumina.

Reads PacBio	número de *reads* sem correção	número de *reads* corrigidos
CCS	258.362	255.147
FS	9.945.534	927.510

CCS: *circular consensus sequences,* **FS:** *filtered subreads.*

1 http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc

3.2.4 *Screening* de *reads* utilizando sequências de referência de mtDNA

Os *reads* obtidos pelo sequenciamento do genoma total pelas diferentes tecnologias (Illumina e PacBio) foram alinhados contra sequências de referência de genomas mitocondriais. Os *reads* mapeados foram então utilizados nas etapas posteriores de *assembly*. A etapa de alinhamento foi realizada utilizando-se o *software* BWA (*Burrows-Wheeler Aligner*) (Li & Durbin, 2009), utilizando-se os parâmetros *default* do programa para ambos os tipos de *reads*, com exceção da flag -x, que foi adicionada para as análises com os *reads* PacBio. A filtragem dos *reads* mapeados foi realizada utilizando-se a ferramenta Samtools (Li et al., 2009).

Foram utilizadas como sequências de referência de genomas mitocondriais as sequências provenientes de *Saccharum officinarum* cromossomo 1 (LC107874.1) e *Saccharum officinarum* cromossomo 2 (LC107875.1) (Shearman et al., 2016). Os números de *reads* disponíveis para o *assembly* após a filtragem estão disponíveis na Tabela 3.3.

Tabela 3.3. Número de *reads* filtrados de cada biblioteca.

Bibliotecas	**Número total de *reads***	**Número de *reads* mitocondriais**
PE170A	189.358.279	570.986
PE170B	188.953.086	565.997
PE500A	183.054.232	2.215.411
PE500B	187.302.075	2.245.237
PE500C	188.230.989	2.248.817
PE500D	176.094.426	2.013.112
PE800A	119.037.154	1.388.663
PE800B	142.746.285	1.669.674
CCS	255.147	4.268
FS	927.510	13.242

CCS: *circular consensus sequences;* **FS:** *filtered subreads.*

3.2.5 Assembly do genoma mitocondrial da cultivar RB867515

Os *reads* provenientes dos filtros realizados foram utilizados no *assembly* do genoma mitocondrial da cultivar RB867515. As montagens foram realizadas com os *assemblers* SPAdes (Bankevich et al., 2012), NovoPlasty (Dierckxsens et al., 2016) e Organelle_PBA (Soorni et al., 2017). A qualidade dos *assemblies* obtidos foi inspecionada pelo realinhamento dos *reads* em cada *assembly*. Os alinhamentos foram visualizados no software Tablet (Milne et al., 2013).

3.2.6 Anotação dos genes do genoma mitocondrial da cultivar RB867515

O genoma obtido como resultado do melhor *assembly* realizado foi anotado utilizando-se duas bases de dados: *Dual Organellar Genome Annotator* (DOGMA) (Wyman et al., 2004) e GeSeq (Tillich et al., 2017) que resultaram em um arquivo no formato Genbank que foi utilizado para se obter a estrutura circular dos dois cromossomos com seus respectivos genes no OrganellarGenomeDRAW (Lohse et al., 2013).

3.3 RESULTADOS E DISCUSSÃO

3.3.1 Estratégias de assembly para o genoma mitocondrial da cultivar RB867515

A sequência completa dos dois cromossomos mitocondriais da cultivar RB867515 foi obtida com *assembly* utilizando somente *reads* PacBio e um pipeline automatizado para *assembly* de organelas, o Organelle_PBA (Soorni et al., 2017). Porém outras estratégias foram também testadas a fim de se validar a sequência obtida. Os resultados dos *assemblies* obtidos com *reads* Illumina, oriundos de bibliotecas PE e MP, e pela utilização conjunta de *reads* Illumina e PacBio estão na Tabela 3.4.

Tabela 3.4. Resumo dos *assemblies* obtidos por diferentes estratégias.

Descrição	Assembler	Nº de contigs	Tamanho total de *contigs*	Maior *contig*
PE	SPAdes	281	561.702	144.806
PE + MP	SPAdes	420	594.333	99.604
PE + MP + CCS	SPAdes	35	542.515	177.794
PE + CCS	SPAdes	31	683.425	239.859
FS	Sprai e WGS-Assembler[1]	1 (chr 1)	300.734	300.734
		1 (chr 2)	194.394	194.394

[1] *Assemblers* utilizados no pipeline Organelle_PBA. PE: *paired ends,* MP: *mate pairs.* CCS: *circular consensus sequences.* FS: *filtered subreads.*

As comparações dos assemblies foram feitas utilizando a refêrencia mais próxima disponível, os dois cromossomos mitocondriais publicados de *Saccharum officinarum*. Os *assemblies* de estratégias híbridas foram significativamente melhores comparados aos *assemblies* usando somente *reads* Illumina.

No *assembly* PE+MP não foram obtidos *contigs* com cobertura significativa da referência. As estratégias PE e PE+CCS montaram em apenas um *contig* de 144.806 pb a sequência completa do cromossomo 2. Em nenhuma dessas três estratégias o cromossomo 1 foi obtido integralmente, porém a utilização da estratégia PE+MP+CCS forneceu *contigs* para os dois cromossomos, cobrindo o cromosso 1 em 6 *contigs* (com alinhamentos em *forward* e *reverse*) e o cromossomo 2 em um único *contig* (Figuras 3.1 e 3.2).

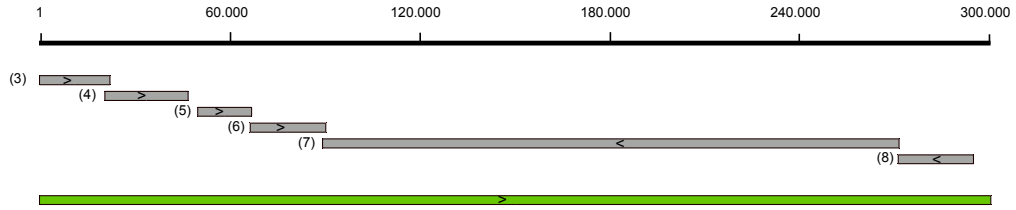

Figura 3.1. Representação esquemática do alinhamento dos *assemblies* obtidos ao cromossomo mitocondrial 1 de *S. officinarum* . **<:** *reverse*. **>:** *forward*. A linha preta representa o comprimento da referência. Barras da mesma cor representam o mesmo *assembly*. Barras em níveis diferentes são *contigs* diferentes. Barra cinza: PE+MP+CCS, *contigs* 3, 4, 5, 6, 7 e 8. Barra verde: *assembly* Organelle_FS.

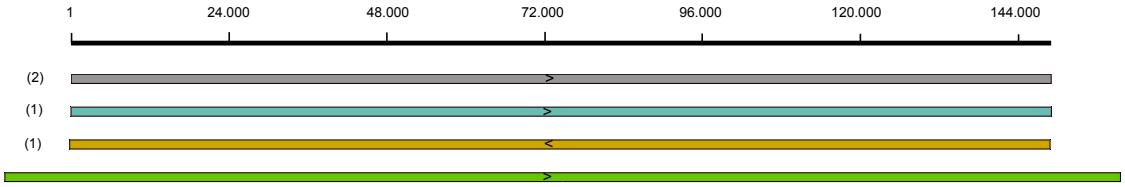

Figura 3.2. Representação esquemática do alinhamento dos *assemblies* obtidos ao cromossomo mitocondrial 2 de *S. officinarum*. **<:** *reverse*. **>:** *forward*. A linha preta representa o comprimento da referência. Barras da mesma cor representam o mesmo *assembly*. Barras em níveis diferentes são *contigs* diferentes. Barra cinza: PE+MP+CCS, *contig* 2. Barra azul: *assembly* PE+CCS, *contig* 1. Barra laranja: PE, *contig* 1. Barra verde: *assembly* Organelle_FS.

A filtragem de *reads* se mostrou eficiente em captar sequências de mtDNA, pois todas as estratégias de *assembly* retornaram a sequência completa do cromossomo 2. Através do *pipeline* Organelle_PBA a sequência completa dos dois cromossomos mitocondrais foi obtida. Cumpre destacar que neste caso, foram realizados dois *assemblies*: um *assembly* utilizando o cromossomo 1 como referência e outro utilizando o cromossomo 2 como referência – pois este *pipeline* só admite uma referência por *assembly*.

A sequência obtida para o cromossomo 1 foi constituída por 300.734 pb, sendo 50 pb menor que a referência (*S. officinarum* chr1, 300.784 pb). O *input* original continha 927.510 *reads* (5.061.924.684 pb) dos quais foram selecionados 9.553 *reads* (59.123.849 bp) pelo mapeamento para o *assembly*, com cobertura estimada de 197X.

A sequência obtida para o cromossomo 2 foi constituída por 194.394 pb, com 49.696 pb a mais do que referência (*S. officinarum* chr2, 144.698 pb). Nesse *assembly* foram selecionadas 4.062 *reads* (23.939.886 pb) na etapa de mapeamento, com cobertura estimada em 165X. As extremidades desse *assembly* (1-36.603 e 181.283-194.394) não alinharam com a referência. No entanto, ao longo destas regiões foram anotados genes mitocondriais, suportando os resultados obtidos.

A partir dos realinhamentos dos *contigs* obtidos pelos *assemblies* Illumina e dos próprios *reads* Illumina nos *assemblies* obtidos pelo Organellle_PBA, realizados para se verificar a ocorrência de *mismatches* e de deleções, foram obtidas sequências consenso para cada um dos cromossomos. A sequência final do cromossomo 1 tem comprimento total de 300.765 pb, com 33 *mismatches* e 19 deleções em relação à referência. A sequência final do cromossomo 2 tem 194.383 pb, com 12 *mismatches*, 11 deleções e duas grandes inserções nas extremidades de 36.603 pb e 13.111 pb, em relação à referência.

3.3.2 O genoma mitocondrial da cultivar RB867515

O genoma mitocondrial da cultivar de cana-de-açúcar RB867515 é composto por dois cromossomos, o cromossomo 1 tem comprimento total de 300.765 pb e o cromossomo 2 com 194.383 pb (Figura 3.3). Os conteúdos GC (~44%) e AT (~56%) são consistentes com aqueles de genomas mitocondriais de outras outras angiospermas (Smith & Keeling, 2015).

Assim como em *S. officinarum*, não foi observado o compartilhamento de grandes regiões pelos dois cromossomos, indicando que o genoma mitocondrial da cana existe sob forma de dois círculos de DNA completamente separados, e não em forma de cromossomo mestre e minicírculos provenientes de recombinação (Shearman et al., 2016). O que indica que esses cromossomos se replicam independentemente um do outro (Alverson et al., 2011).

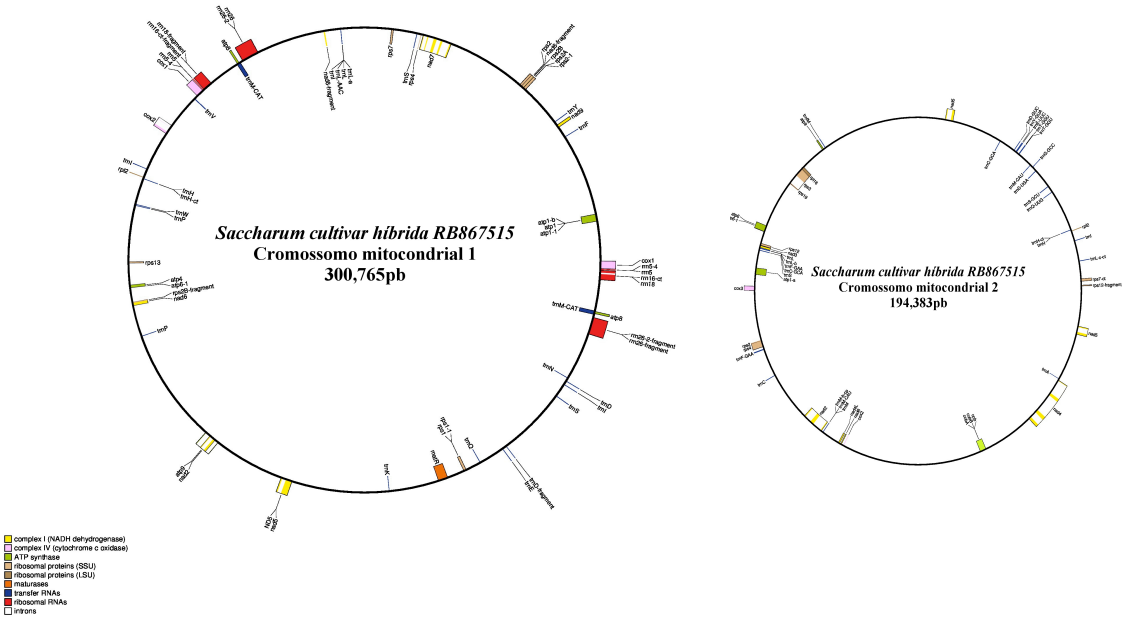

Figura 3.3. Genoma mitocondrial da cultivar de cana-de-açúcar *RB867515*. Representação dos dois cromossomos mitocondriais. As cores dos genes representam seus grupos funcionais conforme a legenda. A escla de representação dos dois cromossomos não é a mesma.

A literatura relata que a presença de minicírculos autônomos não está completamente esclarecida, porém sabe-se que é possível se encontrar mais de dois cromossomos independentes, como verificado em *Cucumis sativus* (Alverson et al., 2011). Os círculos adicionais podem representar "depósitos" de DNA extinto/perdido que podem ser recrutados para recombinação, o que pode representar um mecanismo para conferir variabilidade aos genomas mitocondriais de plantas, que são de herança uniparental (Maréchal & Brisson, 2010).

A anotação do genoma mitocondrial da cultivar RB867515 produziu 39 CDS, 5 genes hipotéticos conservados, 5 rRNAs, 18 tRNAs e 9 fragmentos de genes provavelmente oriundos de cloroplastos por transferência horizontal (Tabela 3.5). O genoma mitocondrial publicado de *Saccharum officinarum*, espécie utilizada no passado como genitor materno de clones que deram origem às cultivares modernas de cana-de-açúcar, existe em dois cromossomos de 300.784 pb e 144.698 pb. A análise comparativa

entre os cromossomos obtidos para a cultivar RB867515 e os cromossomos de *S. officinarum* resultou em 33 *mismatches* e 19 deleções no cromossomo 1 e 12 *mismatches*, 11 deleções, além de uma inserção de 49.714 pb, visualizada como duas inserções nas extremidades da sequência obtida: de 36.603 pb e 13.111 pb, respectivamente, no cromossomo 2.

Os genes a seguir possuem cópias adicionais no genoma da RB867515, quando comparado a *S. officinarum: atp1, atp6, atp9, ccmC, cox1, cox2, cox3, matR, mttB, nad2, nad3, nad4, nad4L, nad5, nad6, nad9, rps2, rps3, rps4, rps12, rps13, rpl14, rps19, rps7, rpl23, trnC-GCA, trnD-GUC, trnE-UUC, trnQ-UUG, trnY-GUA* (todos com uma cópia a mais). Os genes *trnF-GAA* e *trnM-CAU* aparecem com duas e seis cópias adicionais, respectivamente.

Os genes *rrn16* e *rrn23* foram anotados no genoma da RB867515 e não estão presentes na anotação do genoma mitocondrial de *S. officinarum*. O gene *rpl2* foi anotado em *S. officinarum* como fragmento transferido de cpDNA. No genoma mitocondrial da RB867515 esse gene aparece inteiro e está presente em duas cópias, com 4 éxons (Tabela 3.5).

3.4 CONCLUSÕES

O genoma mitocondrial da cultivar de cana-de-açúcar RB867515 é composto por dois cromossomos: o cromossomo 1 tem comprimento total de 300.765 pb e o cromossomo 2 é constituído por 194.383 pb. Os conteúdos GC (~44%) e AT (~56%) são concordantes com aqueles de genomas mitocondriais de outras angiospermas. Foram anotados 39 CDS, 5 genes hipotéticos conservados, 5 rRNAs, 18 tRNAs e 9 fragmentos de genes provavelmente transferidos de cloroplastos por transferência horizontal. Os cromossomos do genoma mitocondrial da cultivar RB867515 apresentam diferenças em relação àqueles da espécie *S. officinarum*, que incluem SNPs, duplicações gênicas, e eventos de expansão genômica.

Tabela 3.5. Resultado da anotação do genoma mitocondrial da cultivar RB867515.

Grupo	Genes
ATP sintase	(2)*atp1*[1,2], *atp4*[1], (2)*atp6*[1,2], (2)*atp8*[1], (2)*atp9*[1,2] [2]
Biogênese citocromo C	*ccmB*[1], (2)*ccmC*[1,2], *ccmFn*[1], *ccmFc*[2] [2]
Ubiquinol citocromo C redutase	*cob*[2]
Citocromo C oxidase	(2)*cox1*[1,2], (2)*cox2*[1,2] [2], (2)*cox3*[1,2]
Maturases	(2)*matR*[1,2]
Proteínas de transporte transmembrana	(2)*mttB*[1,2]
NADH desidrogenase	*nad1*[1][3], (2)*nad2*[1,2] [2], (2)*nad3*[1,2], (2)*nad4*[1,2] [4], (2)*nad4L*[1,2], (2)*nad5*[1,2] [3], (2)*nad6*[1,2], *nad7*[1] [5], (2)*nad9*[1,2]
Proteínas ribossomais	*rps1*[1], (2)*rps12*[1,2], (2)*rps13*[1,2], (2)*rpl14*[1,2], *rpl16*[2], (2)*rps19*[1,2], (2)*rps2*[1,2], (2)*rps4*[1,2], (2)*rps7*[1,2], (2)*rps3*[2] [2]
rRNAs	*rrn16*[1], *rrn18*[1], *rrn23*[1] [6], *rrn26*[1], *rrn5*[1]
tRNAs	(2)*trnC-GCA*, (2)*trnD-GUC*, (2)*trnE-UUC*, (3)*trnF-GAA*, *trnG-GCC*, (2)*trnH-GUG*, *trnK-UUU*, (2)*trnL-CAA*, (8)*trnM-CAU*, *trnN-GUU*, (2)*trnP-UGG*, (2)*trnQ-UUG*, (2)*trnS-GCU*, *trnS-GGA*, (2)*trnS-UGA*, (2)*trnT-GGU*, *trnW-CCA*, (2)*trnY-GUA*
Genes hipotéticos conservados	*orf104*[1], *orf137*[1], *orf179*[1], *orf25*[1], *orf99*[1]
Fragmentos de genes transferidos de cpDNA	*ndhC*[1], *ndhJ*[1], *ndhK*[1], *rbcL*[1], *rpoC1*[1], *rpoC2*[1], *infA*[2], *petB*[2], *petD*[2]
Genes transferidos de cpDNA	(2)*rpl2*[1] [4], *rps8*[2], (2)*rpl23*[1,2], *rpl14*[2], *rpl36*[2]

[1] Anotado no cromossomo 1. [2] Anotado no cromossomo 2. Números entre parênteses representam o número de cópias do gene. Números entre colchetes representam o número de éxons.

4 CONSIDERAÇÕES FINAIS

Estudos de genômica organelar fornecem informações essenciais para a compreensão da dinâmica evolutiva dos genomas de cloroplastos e mitocôndrias, que são organelas fundamentais na conversão de energia. Embora sejam objeto de estudo há mais de dez anos (Tonti-Filippin et al., 2017) muitas características desses genomas ainda não foram completamente esclarecidas, como por exemplo, sua manutenção na célula durante a evolução, tendo em vista que uma grande parcela de genes essenciais ao metabolismo já foram transferidos para o núcleo.

Atualmente sabemos que genes presentes nessas organelas desempenham papeis fundamentais para o crescimento, desenvolvimento e produtividade da planta, e que mutações em genes de vias metabólicas relacionadas à conversão de energia podem ser letais, como já foi amplamente descrito para mutações que ocorreram em genes mitocondriais de milho (Yamato & Newton, 1999).

A sequência do genoma mitocondrial obtida nesse trabalho mostra diferenças significativas em relação àquela de *S. officinarum*. Verificou-se a ocorrência de 9 fragmentos de genes de cpDNA transferidos, 33 SNPs e 19 deleções no cromossomo 1 e 12 SNPs, 11 deleções e uma expansão de 49.714 pb no cromossomo 2. Como descrito para *S. officinarum*, o genoma mitocondrial de cana de açúcar existe em dois cromossomos distintos. A organização em múltiplos círculos autônomos não é tão comum em plantas, pois até o momento foi reportada apenas para três espécies (*Cucumis sativus* (três cromossomos) *Daucus carota* (dois cromossomos) e *Saccharum officinarum* (dois cromossomos)). Genes mitocondriais são fundamentais para a maquinaria energética celular e mutações nesses genes ou são letais ou comprometem o desenvolvimento e produtividade da planta. Sabendo disso, é importante conhecer a organização de genomas mitocondriais e a distribuição dos seus genes entre os múltiplos cromossomos.

No presente trabalho encontramos variações importantes nos genomas de cloroplastos e mitocondriais da cana-de-açúcar, que agregam novas informações que

podem fomentar discussões ainda em aberto sobre a genômica dessas organelas. No caso dos cloroplastos, verificamos a presença de heteroplasmia individual, em que os cloroplastos estão dispostos em duas populações distintas que diferem pela orientação da região SSC (*small single copy*). Existem poucos estudos sobre o mecanismo responsável por essa variação. Até o momento, aceita-se que a inversão da região SSC ocorre devido recombinação *flip-flop* durante a replicação dos cloroplastos (Palmer, 1983).

A herança dessas organelas é uniparental e também não está completamente esclarecido o modo pelo qual elas segregam durante a divisão celular e se mantém nas células ao longo do curso da evolução. Dito isso, a presença de populações distintas de cloroplastos nas células e sua manutenção ao longo do tempo adicionam complexidade na proposição de modelos baseados em segregação citoplasmática aleatória.

Nesse trabalho foram reportados dois tipos de cloroplastos para a cultivar RB867515, mas não é possível se afirmar que essas sejam as únicas formas possíveis de cloroplastos presentes no organismo estudado. Apesar da literatura sempre reportar que cloroplastos têm estrutura muito conservada, sofrem menores taxas de mutação e estão, predominantemente em estado de homoplasmia, é importante ressaltar que as técnicas utilizadas até o momento para obtenção das sequências de cloroplastos e mitocôndrias são limitadas às tecnologias de análise e sequenciamento empregadas nestes estudos.

Abordagens baseadas em filtros podem omitir variações que não estão presentes nos genomas de referência utilizados. O isolamento prévio do DNA organelar e posterior sequenciamento sofre com as baixas coberturas associadas às sequências, e ainda assim não é capaz de isolar de fato sequências compartilhadas entre cloroplastos e mitocôndrias, o que pode levar à dedução errônea de fragmentos com cobertura similar. Nesse sentido, a utilização de *reads* de comprimento longo tem grande potencial de revelar as variações existentes, e por isso conseguiu-se comprovar a ocorrência de heteroplasmia cloroplastidial em cana-de-açúcar.

5 REFERÊNCIAS

ADAMS, K. L.; PALMER, J. D. Evolution of mitochondrial gene content: Gene loss and transfer to the nucleus. **Molecular Phylogenetics and Evolution**, v. 29, n. 3, p. 380-395, 2003.

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.

ALLEN, J. F. The function of genomes in bioenergetic organelles. **Philosophical transactions of the Royal Society of London. Series B, Biological sciences**, v. 358, n. 1429, p. 19-38, 2003.

ALVERSON, A. J.; RICE, D. W.; DICKINSON, S.; BARRY, K.; PALMER, J. D. Origins and recombination of the bacterial-sized multichromosomal mitochondrial genome of cucumber. **The Plant Cell**, v. 23, n. 7, p. 2499-2513, 2011.

ALVERSON, A. J.; WEI, X.; RICE, D. W.; STERN, D. B. BARRY, K.; PALMER, J. D. Insights into the evolution of mitochondrial genome size from complete sequences of citrullus lanatus and cucurbita pepo (*Cucurbitaceae*). **Molecular Biology and Evolution**, v. 27, n. 6, p. 1436-1448, 2010.

ANTIPOV, D.; HARTWICK, N.; SHEN, M.; RAIKO, M.; LAPIDUS, A.; PEVZNER, P. plasmidSPAdes : Assembling Plasmids from Whole. **BioRxiv**, p. 048942, 2016.

ASAF, S.; WAGAS, M.; KHAN, A. The Complete Chloroplast Genome of Wild Rice (*Oryza minuta*) and Its Comparison to Related Species. **Frontiers in Plant Science**, v. 8, 2017.

ASANO, T.; TSUDZUKI, T.; TAKAHASHI, S.; SHIMADA, H.; KADOWAKI, K. I. Complete nucleotide sequence of the sugarcane (*Saccharum officinarum*) chloroplast genome: a comparative analysis of four monocot chloroplast genomes. **DNA Research**, v. 11, n. 2, p. 93-99, 2004.

BANKEVICH, A.; et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. **Journal of Computational Biology**, v. 19, n. 5, p. 455-477, 2012.

BARBROOK, A. C.; HOWE, C. J.; KURNIAWAN, D. P.; TARR, S. J. Organization and expression of organellar genomes. **Philosophical Transactions of the Royal Society of London B: Biological Sciences**, v. 365, n. 1541, p. 785-797, 2010.

BOCK, R. Structure, function, and inheritance of plastid genomes. In: **Cell and Molecular Biology of Plastids**. Springer Berlin Heidelberg, 2007. p. 29-63.

BOGORAD, L. Evolution of organelles and eukaryotic genomes. **Science**, v. 188, n. 4191, p. 891-898, 1975.

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

CAUZ-SANTOS, L. A.; et al. The Chloroplast Genome of *Passiflora edulis* (Passifloraceae) Assembled from Long Sequence Reads: Structural Organization and Phylogenomic Studies in Malpighiales. **Frontiers in Plant Science**, v. 8, 2017.

CESNIK, R. Melhoramento da cana-de-açúcar: marco sucro-alcooleiro no Brasil. **Embrapa Meio Ambiente-Artigo em periódico indexado (ALICE)**, 2004.

CHANEY, L.; MANGELSON, R.; RAMARAJ, T.; JELLEN, E. N.; MAUGHAN, P. J. The complete chloroplast genome sequences for four Amaranthus species (Amaranthaceae). **Applications in Plant Sciences**, v. 4, n. 9, p. 1600063, 2016.

CHAW, S. M.; CHUN-CHIEH SHIH, A.; WANG, D.; WU, Y. W.; LIU, S. M.; CHOU, T. Y. The mitochondrial genome of the gymnosperm Cycas taitungensis contains a novel family of short interspersed elements, Bpu sequences, and abundant RNA editing sites. **Molecular Biology and Evolution**, v. 25, n. 3, p. 603-615, 2008.

CHEN, J.; GUAN, R.; CHANG, S.; DU, T.; ZHANG, H.; XING, H. Substoichiometrically Different Mitotypes Coexist in Mitochondrial Genomes of *Brassica napus* L. **PLoS ONE**, v. 6, n. 3, p. 8, 2011.

CLARKE, J. L.; DANIELL, H. Plastid biotechnology for crop production: Present status and future perspectives. **Plant Molecular Biology**, v. 76, n. 3-5, p. 211-220, 2011.

DANIELL, H.; LIN, C. S.; YU, M.; CHANG, W. J. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. **Genome Biology**, v. 17, n. 1, p. 134, 2016.

DANIELS, J. Taxonomy and evolution. **Sugarcane improvement through breeding**, p. 7-84, 1987.

DE VRIES, J.; et al. *YCF1*: a green TIC?. **The Plant Cell**, v. 27, n. 7, p. 1827-1833, 2015.

DIERCKXSENS, N.; MARDULYN, P.; SMITS, G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. **Nucleic Acids Research**, v. 45, n. 4, p. e18-e18, 2017.

EMBLEY, T. M.; MARTIN, W. Eukaryotic evolution, changes and challenges. **Nature**, v. 440, n. 7084, p. 623-630, 2006.

EVANS, D. L.; JOSHI, S. V. Complete chloroplast genomes of *Saccharum spontaneum*, *Saccharum officinarum* and *Miscanthus floridulus* (Panicoideae: Andropogoneae) reveal the plastid view on sugarcane origins. **Systematics and Biodiversity**, v. 14, n. 6, p. 548-571, 2016.

FERRARINI, M.; et al. An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. **BMC genomics**, v. 14, n. 1, p. 670, 2013.

GANDINI, C. L.; SANCHEZ-PUERTA, M. V. Foreign Plastid Sequences in Plant Mitochondria are Frequently Acquired Via Mitochondrion-to-Mitochondrion Horizontal Transfer. **Scientific Reports**, v. 7, 2017.

GARAYCOCHEA, S.; SPERANZA, P.; ALVAREZ-VALIN, F. A Strategy to Recover a High-Quality, Complete Plastid Sequence from Low-Coverage Whole-Genome Sequencing. **Applications in Plant Sciences**, v. 3, n. 10, p. 1500022, 2015.

GLASZMANN, J. C.; DUFOUR, P.; GRIVET, L.; D'HONT, A.; DEU, M.; PAULET, F.; HAMON, P. Comparative genome analysis between several tropical grasses. **Euphytica**, v. 96, n. 1, p. 13-21, 1997.

GOLCZYK, H.; GREINER, S.; WANNER, G.; WEIHE, A.; BOCK, R.; BÖRNER, T.; HERRMANN, R. G. Chloroplast DNA in mature and senescing leaves: a reappraisal. **The Plant Cell**, v. 26, n. 3, p. 847-854, 2014.

GOREMYKIN, V. V.; LOCKHART, P. J.; VIOLA, R.; VELASCO, R. The mitochondrial genome of Malus domestica and the import−driven hypothesis of mitochondrial genome expansion in seed plants. **The Plant Journal**, v. 71, n. 4, p. 615-626, 2012.

GOREMYKIN, V. V.; SALAMINI, F.; VELASCO, R.; VIOLA, R. Mitochondrial DNA of *Vitis vinifera* and the issue of rampant horizontal gene transfer. **Molecular Biology and Evolution**, v. 26, n. 1, p. 99-110, 2008.

GRAY, M. W. Evolution of organellar genomes. **Current Opinion in Genetics Development**, v. 9, n. 6, p. 678-687, 1999.

GREINER, S.; SOBANSKI, J.; BOCK, R. Why are most organelle genomes transmitted maternally? **BioEssays**, v. 37, n. 1, p. 80-94, 2015.

GRIVET, L.; ARRUDA, P. Sugarcane genomics: Depicting the complex genome of an important tropical crop. **Current Opinion in Plant Biology**, v. 5, n. 2, p. 122-127, 2002.

HIRATSUKA, J.; et al. The complete sequence of the rice (*Oryza sativa*) chloroplast genome: intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. **Molecular and General Genetics MGG**, v. 217, n. 2-3, p. 185-194, 1989.

HOANG, N. V.; et al. A survey of the complex transcriptome from the highly polyploid sugarcane genome using full-length isoform sequencing and de novo assembly from short read sequencing. **BMC Genomics**, v. 18, n. 1, p. 395, 2017.

HOANG, N. V.; Furtado, A.; McQualter, R. B.; Henry, R. J. Next generation sequencing of total DNA from sugarcane provides no evidence for chloroplast heteroplasmy. **New Negatives in Plant Science**, v. 1, p. 33-45, 2015.

HODKINSON, T. R.; CHASE, M. W.; LLEDÓ, D. M.; SALAMIN, N.; RENVOIZE, S. A. Phylogenetics of *Miscanthus*, *Saccharum* and related genera (Saccharinae, Andropogoneae, Poaceae) based on DNA sequences from ITS nuclear ribosomal DNA and plastid *trnL* intron and *trnL*-F intergenic spacers. **Journal of Plant Research**, v. 115, n. 5, p. 381-392, 2002.

IORIZZO, M.; SENALIK, D.; SZKLARCZYK, M.; GRZEBELUS, D.; SPOONER, D.; SIMON, P. De novo assembly of the carrot mitochondrial genome using next generation sequencing of whole genomic DNA provides first evidence of DNA transfer into an angiosperm plastid genome. **BMC Plant Biology**, v. 12, n. 1, p. 61, 2012.

JUNIOR, T. C.; CARRARO, D. M.; BENATTI, M. R.; BARBOSA, A. C.; KITAJIMA, J. P.; CARRER, H. Structural features and transcript-editing analysis of sugarcane (*Saccharum officinarum* L.) chloroplast genome. **Current Genetics**, v. 46, n. 6, p. 366-373, 2004.

KUBO, T.; NEWTON, K. J. Angiosperm mitochondrial genomes and mutations. **Mitochondrion**, v. 8, n. 1, p. 5-14, 2008.

LEE, Y. S. Understanding PacBio transcriptome data. Disponível em: <https://github.com/PacificBiosciences/cDNA_primer/wiki/Understanding-PacBio-transcriptome-data#readexplained>. Acesso em: 3/5/2017.

LEI, W.; et al. Intraspecific and heteroplasmic variations, gene losses and inversions in the chloroplast genome of *Astragalus membranaceus*. **Scientific Reports**, v. 6, p. 21669, 2016.

LI, H.; DURBIN, R. Fast and accurate short read alignment with Burrows–Wheeler transform. **Bioinformatics**, v. 25, n. 14, p. 1754-1760, 2009.

LI, H.; et al. The sequence alignment/map format and SAMtools. **Bioinformatics**, v. 25, n. 16, p. 2078-2079, 2009.

LI, Z.; LONG, H.; ZHANG, L.; LIU, Z.; CAO, H.; SHI, M.; TAN, X. The complete chloroplast genome sequence of tung tree (*Vernicia fordii*): Organization and phylogenetic relationships with other angiosperms. **Scientific Reports**, v. 7, 2017.

LICHTENSTEIN, G.; CONTE, M.; ASIS, R.; CARRARI, F. Chloroplast and Mitochondrial Genomes of Tomato. In: **The Tomato Genome**. Springer Berlin Heidelberg, 2016. p. 111-137.

LLOYD, A. H.; TIMMIS, J. N. The origin and characterization of new nuclear genes originating from a cytoplasmic organellar genome. **Molecular Biology and Evolution**, v. 28, n. 7, p. 2019-2028, 2011.

LOHSE, M.; DRECHSEL, O.; KAHLAU, S.; BOCK, R. OrganellarGenomeDRAW—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. **Nucleic Acids Research**, v. 41, n. W1, p. W575-W581, 2013.

MAIER, R. M.; NECKERMANN, K.; IGLOI, G. L.; KÖSSEL, H. Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. **Journal of Molecular Biology**, v. 251, n. 5, p. 614-628, 1995.

MALIGA, P. (Ed.). **Chloroplast biotechnology: methods and protocols**. Humana Press, 2014.

MARÉCHAL, A.; BRISSON, N. Recombination and the maintenance of plant organelle genome stability. **New Phytologist**, v. 186, n. 2, p. 299-317, 2010.

MARTIN, G.; BAURENS, F. C.; CARDI, C.; AURY, J. M.; D'HONT, A. The complete chloroplast genome of banana (*Musa acuminata*, Zingiberales): insight into plastid monocotyledon evolution. **PloS One**, v. 8, n. 6, p. e67350, 2013.

MATSUOKA, S.; GARCIA, A.A.F.; ARIZONO, H. Melhoramento da cana-de-açúcar. In: BORÉM, A. (Ed.). **Melhoramento de espécies cultivadas**. Viçosa: Ed. da UFV, 2005. p.225-274.

MILLEN, R. S.; et al. Many parallel losses of infA from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. **The Plant Cell**, v. 13, n. 3, p. 645-658, 2001.

MILNE, I.; et al. Using Tablet for visual exploration of second-generation sequencing data. **Briefings in Bioinformatics**, v. 14, n. 2, p. 193-202, 2012.

MOTA, A. P. Z. Sequencia completa do genoma cloroplasmático do feijão-caupi [*Vigna unguiculata* (L.) Walp] e diversidade genética de variedades tradicionais brasileiras e africanas. Sequência completa do genoma cloroplasmático do feijão-caupi (*Vigna unguiculata* (L.) Walp) e diversidade genética de variedades tradicionais brasileiras e africanas. 2013.

NAKAYAMA, T.; ISHIDA, K. ICHIRO. Another acquisition of a primary photosynthetic organelle is underway in *Paulinella chromatophora*. **Current Biology**, v. 19, n. 7, p. 284-285, 2009.

NIELSEN, B. L.; CUPP, J. D.; BRAMMER, J. Mechanisms for maintenance, replication, and repair of the chloroplast genome in plants. **Journal of Experimental Botany**, v. 61, n. 10, p. 2535-2537, 2010.

NOCK, C. J.; et al. Chloroplast genome sequences from total DNA for plant identification. **Plant Biotechnology Journal**, v. 9, n. 3, p. 328-333, 2011.

O'MALLEY, M. A. Endosymbiosis and its implications for evolutionary theory. **Proceedings of the National Academy of Sciences**, v. 112, n. 33, p. 10270-10277, 2015.

OGIHARA, Y.; et al. Structural features of a wheat plastome as revealed by complete sequencing of chloroplast DNA. **Molecular Genetics and Genomics**, v. 266, n. 5, p. 740-746, 2002.

OLDENBURG, D. J.; BENDICH, A. J. DNA maintenance in plastids and mitochondria of plants. **Frontiers in Plant Science**, v. 6, 2015.

OLDENBURG, D. J.; BENDICH, A. J. Most Chloroplast DNA of Maize Seedlings in Linear Molecules with Defined Ends and Branched Forms. **Journal of Molecular Biology**, v. 335, n. 4, p. 953-970, 2004.

OLEJNICZAK, S. A.; et al. Chloroplasts: state of research and practical applications of plastome sequencing. **Planta**, v. 244, n. 3, p. 517-527, 2016.

PAES, D.; et al. A Duplication Lost In Sugarcane Hybrids Revealed By Chloroplast Genome Assembly Of Wild Species *Saccharum officinarum*. **bioRxiv**, p. 141002, 2017.

PALMER, J. D. Chloroplast DNA exists in two orientations. **Nature**, v. 301, n. 5895, p. 92-93, 1983.

PARK, S.; et al. Complete sequences of organelle genomes from the medicinal plant *Rhazya stricta* (Apocynaceae) and contrasting patterns of mitochondrial genome evolution across asterids. **BMC Genomics**, v. 15, n. 1, p. 405, 2014.

PECCOUD, J.; et al. Untangling Heteroplasmy, Structure, and Evolution of an Atypical Mitochondrial Genome by PacBio Sequencing. **Genetics**, p. genetics. 117.203380, 2017.

PIPERIDIS, G.; PIPERIDIS, N.; D'HONT, A. Molecular cytogenetic investigation of chromosome composition and transmission in sugarcane. **Molecular Genetics and Genomics**, v. 284, n. 1, p. 65-73, 2010.

RAJ, P.; SELVI, A.; PRATHIMA, P. T.; NAIR, N. V. Analysis of Genetic Diversity of *Saccharum* Complex Using Chloroplast Microsatellite Markers. **Sugar Tech**, v. 18, n. 2, p. 141-148, 2016.

RAO, V. P.; et al. Genetic variability in sugarcane (*Saccharum* spp. hybrid) genotypes through inter simple sequence repeats (ISSR) markers. **Journal of Applied and Natural Science**, v. 8, n. 3, p. 1404-1409, 2016.

RAVI, V.; KHURANA, J. P.; TYAGI, A. K.; KHURANA, P. An update on chloroplast genomes. **Plant Systematics and Evolution**, v. 271, n. 1-2, p. 101-122, 2008.

RICHARDSON, A. O.; PALMER, J. D. Horizontal gene transfer in plants. **Journal of Experimental Botany**, v. 58, n. 1, p. 1-9, 2007.

ROGALSKI, M.; DO NASCIMENTO VIEIRA, L.; FRAGA, H. P.; GUERRA, M. P. Plastid genomics in horticultural species: importance and applications for plant population genetics, evolution, and biotechnology. **Frontiers in Plant Science**, v. 6, 2015.

RUHLMAN, T. A.; et al. Recombination-dependent replication and gene conversion homogenize repeat sequences and diversify plastid genome structure. **American Journal of Botany**, v. 104, n. 4, p. 559-572, 2017.

RYAN, M. T.; HOOGENRAAD, N. J. Mitochondrial-Nuclear Communications. **Annual Review of Biochemistry**, v. 76, n. 1, p. 701-722, 2007.

SABIR, J. S. M.; et al. Whole Mitochondrial and Plastid Genome SNP Analysis of Nine Date Palm Cultivars Reveals Plastid Heteroplasmy and Close Phylogenetic Relationships among Cultivars. **PLoS One**, v. 9, n. 4, p. e94158, 2014.

SALMELA, L.; RIVALS, E. LoRDEC: Accurate and efficient long read error correction. **Bioinformatics**, v. 30, n. 24, p. 3506-3514, 2014.

SASKI, C.; et al. Complete chloroplast genome sequences of *Hordeum vulgare*, *Sorghum bicolor* and *Agrostis stolonifera*, and comparative analyses with other grass genomes. **Theoretical and Applied Genetics**, v. 115, n. 4, p. 571-590, 2007.

SHEARMAN, J. R.; et al. The two chromosomes of the mitochondrial genome of a sugarcane cultivar: assembly and recombination analysis using long PacBio reads. **Scientific Reports**, v. 6, p. 31533, 2016.

SHEPPARD, A. E.; TIMMIS, J. N. Instability of plastid DNA in the nuclear genome. **PLoS Genetics**, v. 5, n. 1, p. 1-8, 2009.

SLOAN, D. B. One ring to rule them all? Genome sequencing provides new insights into the "master circle" model of plant mitochondrial DNA structure. **New Phytologist**, v. 200, n. 4, p. 978-985, 2013.

SMITH, D. R. The past, present and future of mitochondrial genomics: Have we sequenced enough mtDNAs? **Briefings in Functional Genomics**, v. 15, n. 1, p. 47-54, 2016.

SMITH, D. R.; KEELING, P. J. Mitochondrial and plastid genome architecture: reoccurring themes, but significant differences at the extremes. **Proceedings of the National Academy of Sciences**, v. 112, n. 33, p. 10177-10184, 2015.

SOORNI, A.; HAAK, D.; ZAITLIN, D.; BOMBARELY, A. Organelle_PBA, a pipeline for assembling chloroplast and mitochondrial genomes from PacBio DNA sequencing data. **BMC Genomics**, v. 18, n. 1, p. 49, 2017.

STADERMANN, K. B.; WEISSHAAR, B.; HOLTGRÄWE, D. SMRT sequencing only de novo assembly of the sugar beet (*Beta vulgaris*) chloroplast genome. **BMC Bioinformatics**, v. 16, n. 1, p. 295, 2015.

TILLICH, M.; et al. GeSeq - versatile and accurate annotation of organelle genomes. **Nucleic Acids Research**, v. 22, n. 5, p. 97-104, 2017.

TOMES, D.; LAKSHMANAN, P.; SONGSTAD, D. (Ed.). **Biofuels**. Global Impact on Renewable Energy, Production in Agriculture, and Technology Advancements. Nova York: Springer, 2011. 367 p.

TONG, W.; HE, Q.; PARK, Y. J. Genetic variation architecture of mitochondrial genome reveals the differentiation in Korean landrace and weedy rice. **Scientific Reports**, v. 7, p. 43327, 2017.

TONTI−FILIPPINI, J.; NEVILL, P. G.; DIXON, K.; SMALL, I. What can we do with 1000 plastid genomes?. **The Plant Journal**, v. 90, n. 4, p. 808-818, 2017.

TSURUTA, S.; et al. Complete Chloroplast Genomes of *Erianthus arundinaceus* and *Miscanthus sinensis*: Comparative Genomics and Evolution of the *Saccharum* Complex. **PloS One**, v. 12, n. 1, p. e0169992, 2017.

TWYFORD, A. D.; NESS, R. W. Strategies for complete plastid genome sequencing. **Molecular Ecology Resources**, 2016.

VIDIGAL, P. M.; COELHO, A. S.; NOVAES, E.; BARBOSA, M. H.; PETERNELLI, L. A. Complete Chloroplast Genome Sequence and Annotation of the *Saccharum* Hybrid Cultivar RB867515. **Genome Announcements**, v. 4, n. 5, 2016.

WALKER, J. F.; et al. Sources of inversion variation in the small single copy (SSC) region of chloroplast genomes. **American Journal of Botany**, v. 102, n. 11, p. 1-2, 2015.

WANG, L.; et al. Complete chloroplast genome sequences of *Eucommia ulmoides*: genome structure and evolution. **Tree Genetics Genomes**, v. 12, n. 1, p. 12, 2016.

WANG, W.; WU, Y.; MESSING, J. The Mitochondrial Genome of an Aquatic Plant, *Spirodela polyrhiza*. **PLoS One**, v. 7, n. 10, p. 1-10, 2012.

WEI, S.; et al. Assembly and analysis of the complete *Salix purpurea* L.(Salicaceae) mitochondrial genome sequence. **SpringerPlus**, v. 5, n. 1, p. 1894, 2016.

WICKE, S.; SCHNEEWEISS, G. M. Next-generation organellar genomics: Potentials and pitfalls of high-throughput technologies for molecular evolutionary studies and plant systematics. **Next-Generation Sequencing in Plant Systematics**, v. 158, 2015.

WOLFE, A. D.; RANDLE, C. P. Recombination, Heteroplasmy, Haplotype Polymorphism, and Paralogy in Plastid Genes: Implications for Plant Molecular Systematics. **Systematic Botany**, v. 29, n. 4, p. 1011-1020, 2004.

WOODSON, J. D.; CHORY, J. Coordination of gene expression between organellar and nuclear genomes. **Nat. Rev. Genet.**, v. 9, n. 5, p. 383-395, 2008.

WU, C.; CHAW, S. Highly rearranged and size-variable chloroplast genomes in conifers II clade (cupressophytes): evolution towards shorter intergenic spacers. **Plant Biotechnology Journal**, v. 12, p. 344-353, 2014.

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

YAMATO, K. T.; NEWTON, K. J. Heteroplasmy and homoplasmy for maize mitochondrial mutants: a rare homoplasmic nad4 deletion mutant plant. **Journal of Heredity**, v. 90, n. 3, p. 369-373, 1999.

ZHANG, T.; et al. An efficient procedure for plant organellar genome assembly, based on whole genome data from the 454 GS FLX sequencing platform. **Plant Methods**, v. 7, n. 1, p. 38-45, 2011.

ZHANG, T.; et al. The complete chloroplast and mitochondrial genome sequences of *boea hygrometrica*: Insights into the evolution of plant organellar genomes. **PLoS One**, v. 7, n. 1, p. 101-117, 2012.

ZHENG, L. Y.; et al. Genome-wide patterns of genetic variation in sweet and grain sorghum (*Sorghum bicolor*). **Genome Biology**, v. 12, n. 11, p. R114, 2011.

ZHU, J. R.; et al. Genetic variability among the chloroplast genomes of sugarcane (*Saccharum* spp) and its wild progenitor species *Saccharum spontaneum* L. **Genetics and Molecular Research**, v. 13, n. 2, p. 3037-3047, 2014.