UFG

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

CARACTERIZAÇÃO DA REGIÃO Bru1 NO GENOMA DA CULTIVAR RB867515 (*Saccharum* spp.) UTILIZANDO SEQUENCIAMENTO DE NOVA GERAÇÃO

ISABELA PAVANELLI DE SOUZA

Orientador:
Prof. Dr. Alexandre Siqueira Guedes Coelho

Goiânia - GO
Brasil

Setembro – 2014

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR AS TESES E DISSERTAÇÕES ELETRÔNICAS 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: Isabela Pavanelli de Souza

Título do trabalho: CARACTERIZAÇÃO DA REGIÃO Bru1 NO GENOMA DA CULTIVAR RB867515 (*Saccharum* spp.) UTILIZANDO SEQUENCIAMENTO DE NOVA GERAÇÃO

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.

Isabela Pavanelli de Souza

Assinatura do (a) autor (a) [2]

Data: 15 / 02 / 2017

[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.

[2] A assinatura deve ser escaneada.

ISABELA PAVANELLI DE SOUZA

CARACTERIZAÇÃO DA REGIÃO Bru1 NO GENOMA DA CULTIVAR RB867515 (*Saccharum* spp.) UTILIZANDO SEQUENCIAMENTO DE NOVA GERAÇÃO

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
2014

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

Souza, Isabela Pavanelli de
Caracterização da região Bru1 no genoma da cultivar RB867515 (Saccharum spp.) utilizando sequenciamento de nova geração [manuscrito] / Isabela Pavanelli de Souza. - 2014.
96 f.: il.

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, 2014.
Bibliografia. Apêndice.
Inclui siglas, abreviaturas, gráfico, tabelas.

1. Cana-de-açúcar. 2. Genômica. 3. Bioinformática. I. Coelho, Alexandre Siqueira Guedes, orient. II. Título.

CDU 575

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

UFG

ATA DA DEFESA PÚBLICA DA DISSERTAÇÃO DE ISABELA PAVANELLI DE SOUZA.

Aos vinte e cinco dias do mês de Setembro do ano de dois mil e catorze (25.09.2014), às 14h00min, no Auditório PPGA Escola de Agronomia, reuniram-se os componentes da Banca Examinadora, Prof. Dr. Alexandre Siqueira Guedes Coelho – Presidente/Orientador; Profª. Drª. Monalisa Sampaio Carneiro; Prof. Dr. Evandro Novaes. Sob a presidência do orientador, e em sessão pública, procedeu-se à avaliação da defesa de Dissertação intitulada: **"Caracterização de 1Mb do genoma de *Saccharum* spp. Utilizando sequenciamento de nova geração"**, de autoria de **Isabela Pavanelli de Souza**, 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, Prof. Dr. Alexandre Siqueira Guedes Coelho, que fez a apresentação formal dos membros da Banca. A palavra, a seguir, foi concedida ao autor da Dissertação que, em 40 minutos, apresentou o seu trabalho. Terminada a apresentação, cada membro da Banca argüiu o mestrando, tendo-se adotado o sistema de diálogo seqüencial. Ao final, a banca reunida em separado procedeu à avaliação da defesa. O título da dissertação foi alterado para "CARACTERIZAÇÃO DA REGIÃO Bmb NO GENOMA DA CULTIVAR RB867515 (SACCHARUM SPP.) UTILIZANDO SEQUENCIAMENTO DE NOVA GERAÇÃO."

"De acordo com a Resolução nº 1053/2011, do CEPEC - Conselho de Ensino, Pesquisa, Extensão e Cultura, que regulamenta o Programa de Pós-Graduação em Genética e Melhoramento de Plantas, e desde que procedidas às correções recomendadas, a Dissertação será considerada APROVADA pela Banca Examinadora, estando integralmente cumprido este requisito para fins de obtenção do título de MESTRE EM GENÉTICA E MELHORAMENTO DE PLANTAS, pela Universidade Federal de Goiás. O mestrando deverá efetuar as modificações eventualmente sugeridas pela Banca Examinadora e encaminhar a versão definitiva da Dissertação à Secretaria do PGMP, no prazo máximo de trinta dias após a data da Defesa. A conclusão do Curso e a emissão do Diploma dar-se-ão após o cumprimento do Artigo 52 da Resolução CEPEC nº 1053/2011. A Banca Examinadora recomenda a publicação de artigo(s) científico(s), oriundo(s) dessa Dissertação, em periódicos de circulação nacional e, ou, internacional, depois de procedidas as modificações sugeridas. 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, secretária 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
Prof. Dr. Alexandre Siqueira Guedes Coelho
Presidente/Orientador

Monalisa Sampaio Carneiro
Drª. Monalisa Sampaio Carneiro
Membro Externo

Evandro Novaes
Prof. Dr. Evandro Novaes
Membro Interno

“O degrau de uma escada não serve
simplesmente para que alguém
permaneça em cima dele, destina-se a
sustentar o pé de um homem pelo
tempo suficiente para que ele coloque
o outro um pouco mais alto”

Thomas Huxley

Aos meus pais, Claudinê e Cleide, e ao meu noivo Maurício, pelo apoio incondicional em todos os momentos da minha vida, dedico.

AGRADECIMENTOS

Em primeiro lugar gostaria de agradecer a Deus, pelo dom da vida. Agradeço pela minha saúde, por me dar forças para vencer as barreiras que são impostas, por minha família e meus amigos.

Gostaria de agradecer às instituições que financiaram a execução desse trabalho. À Petrobras pelo custeamento do projeto e à CAPES (Coordenação de Aperfeiçoamento Pessoal de Nível Superior) pela bolsa de mestrado concedida.

Ao Programa de Pós Graduação em Genética e Melhoramento de Plantas da UFG, em especial à coordenadora Dra. Mariana Telles e Dra. Patrícia dos Santos Melo. Agradeço a todos os professores do programa, pela sua contribuição na minha formação acadêmica e como pesquisadora, e que se esforçam em desempenhar seu trabalho com excelência.

Aos membros da banca examinadora, Dr. Evandro Novaes e Dra. Monalisa Sampaio Carneiro, pela disponibilidade e contribuições para o trabalho.

Um agradecimento especial ao meu orientador Dr. Alexandre Siqueira Guedes Coelho pelos valiosos ensinamentos ao longo do mestrado. Posso dizer que existe uma Isabela antes e depois do mestrado, e que a versão "plus" tem grande contribuição das lições que tirei dos anos que convivemos e dos conhecimentos adquiridos com o nosso trabalho, e por isso eu digo ao senhor: Muito obrigada e um "Forte abraço".

À minha família, e em especial aos meus Pais, Claudinê e Cleide, minha fonte de força e inspiração, meus maiores ídolos! Obrigada pelo apoio incondicional em todos os momentos da minha vida, por acreditarem no meu potencial, e por sempre me incentivar a lutar pelos meus sonhos. Aprendi com vocês o valor da família, e pelos exemplos do dia-a-dia aprendi sobre amor, honestidade, paciência, benevolência e educação, e por isso, serei eternamente grata!

Ao meu noivo Maurício, obrigada pelo amor e carinho sempre presentes. Obrigada por todo apoio que você tem me dado, e por sempre me incentivar, comemorando comigo as vitórias e me confortando nas derrotas.

Aos amigos queridos que estiveram comigo durante essa jornada na Pós-graduação. Obrigada por todos os momentos compartilhados, pelas experiências que vivenciamos juntos. Afinal, o que seria da vida se não tivermos amigos para dividir os momentos?

SUMÁRIO

RESUMO

SOUZA, I. P. **Caracterização da região Bru1 no genoma da cultivar RB867515 (*Saccharum* spp.) utilizando sequenciamento de nova geração**. 2014. 100 f. Dissertação (Mestrado em Genética e Melhoramento de Plantas) – Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2014.[1]

A cana-de-açúcar é reconhecida como uma das mais importante culturas do mundo, pela utilização dos seus subprodutos. O genoma da cana-de-açúcar é um dos mais complexos entre as plantas cultivadas, com aproximadamente 10 Gb. Seu genoma completo ainda não foi sequenciado, mas o surgimento e a popularização de novas ferramentas de análise genômica possibilitaram estudos refinados sobre essa cultura. Com o grande volume de informações que é possível gerar, a demanda atual é a produção ferramentas eficientes para o processamento dos dados. Foi realizado um *assembly* e anotação de uma região do genoma da cultivar RB867515 correspondente às sequências de 8 BACs da cultivar R570. As regiões correspondentes foram obtidas por alinhamento usando Bowtie2 com *reads* de bibliotecas *paired-ends* produzidos por sequenciador automático de nova geração e montados *de novo* utilizando MaSuRCA. Os *scaffolds* foram alinhados às sequência de referência usando BWA-SW, e as sequências consenso foram obtidas pela opção mplieup do SAMtools. Reads de cDNA de cinco tecidos vegetais, provenientes de 30 genótipos de cana-de-açúcar obtidos pela estratégia RNA-seq, foram mapeados nas sequências consenso a fim de identificar as regiões gênicas, que foram anotadas utilizando Blastx contra o banco de proteínas não redundante no GenBank. As regiões repetitivas foram determinadas pelo RepeatMasker e os microssatélites pelo IMEX. Para a comparação entre as sequências das duas cultivares, foi realizado uma alinhamento das sequências correspondentes nos dois genomas utilizando ClustalW no software Mega. O *assembly* das oito regiões, gerou de 607 à 2884 *scaffolds* maiores que 1 kb, com o maior *scaffold* chegando a 21 kb. As sequências consenso variam de 81 a 142 kb de tamanho, representando uma taxa de recuperação em relação à referência de 82% a 97%. O tamanho total das sequências montadas somou quase 1 Mb do genoma da cultivar de cana-de-açúcar. Em relação à anotação, foram identificados 5145 elementos genéticos repetitivos, em que 4662 são microssatélites e 460 são transposons, totalizando 225 kb em sequências repetidas ao longo dos BACs. Dentro do grupo dos elementos genéticos móveis os retrotransposons são maioria, com 15% da composição nucleotídica, variando de 8% a 29% entre os BACs. Foram identificados 134 genes nas oito sequências de cana-de-açúcar analisadas, totalizando 243 kb. O número de genes por BAC variou de 11 a 26, com uma média de 16 genes por BAC. Os genes encontrados apresentaram tamanho médio de 1841 pb, variando de 443 (BAC1) à 6316 pb (BAC3). A densidade de genes média foi de 1 gene por 7,2 kb. A porcentagem de mismatches entres as sequências dos BACs de RB867515 variou de 0,27% a 1,32%. Os BACs de cana-de-açúcar correspondem a regiões genômicas homeólogas, com o alinhamento realizado com as duas cultivares pode-se sugerir que existe alta divergência dentro do grupo de homeologia.

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

ABSTRACT

SOUZA, I. P. **Characterization of RB867515 cultivar (*Saccharum* spp.) Bru1 region using next-generation sequencing**. 2014. 100 l. Dissertation (Master in Genetics and Plant Breeding) – Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2014.[2]

Sugarcane is known as one of the most important crops of the word for its sub products utilization. Four countries, led by Brazil, supply the sugar international trade. Ethanol is other important sugarcane sub product, recognized as an alternative product to sugar, and had great demand in Brazilian trade, for its utilization as non-fossil fuel. The sugarcane genome is one of the most complex among crops, with 10 Gb. Its complete genome is not available, but the recent innovations in genomics tools open up new possibilities for the investigations about the sugarcane's genome. We did a genome assembly and annotation of a Brazilian sugarcane cultivar (RB867515) genome region, correspondent to eight R570 homologous sequences already published. We use high qualities paired-ends libraries produced by Illumina HiSeq 2000 sequencing platform. The reads were aligned against eight R570 BACs (Bacterial Artificial Chromosome) sequences stored in NCBI using Bowtie2. We used MaSuRCA to assemble the aligned reads de novo, and the consensus sequences were obtained with SAMtools mpileup option. The transposable elements were identified using RepeatMasker and the gene regions were annotated with Blastx against the GenBank non-redundant protein database. After that, the consensus sequences were aligned with the matching reference (R570) using ClustalW in Mega software, to look for the percentage of mismatches and conserved sites between them. We obtained the number of scaffolds bigger than 1 kb ranging from 607 to 2,884, and the longest scaffold had near 21 kb. The consensus sequence length ranged from 81 to 142 kb, and the recovery rate relative to the reference ranged from 82% to 97%. The sequences amounted 1 Mb of RB867515 cultivar genome. We identified 5,145 repeated elements, which 4,662 were microsatellite and 460 were transposable elements, amounted 225 kb of repeated sequences. Among the mobile elements, the retrotransposons comprises 15% of nucleotide composition, ranging from 8% to 29% among BACs. The 134 genes identified on the eight BAC consensus sequences comprised a total of 243 kb, resulting in a density of one gene per 7.2 kb. The average number of genes per BAC was 16, with an average gene length of 1,841 bp. The percentage of mismatches between the RB867515 and R570 BACs ranged from 0.27% to 1.32%. The sugarcane BACs correspond to homeologous genomic regions, with this alignment we can suggest high divergence inside an homeologous group.

[2] Adviser: Prof. Dr. Alexandre Siqueira Guedes Coelho. EA-UFG

1 INTRODUÇÃO

A exploração da cana-de-açúcar precede a história escrita (Berding & Roach, 1987). A cana-de-açúcar foi introduzida no Brasil pelos portugueses no século XVI e, desde então, destaca-se como uma das principais culturas agrícolas do país. A partir dos primeiros programas de investimento no setor sucroenergético (como o Proálcool em 1975), iniciou-se um processo de intensa transformação do agronegócio da cana, resultante de grandes investimentos em melhorias tecnológicas. Em função da demanda por fontes de energia renováveis, o setor passa por uma nova fase de intensos investimentos diante da perspectiva de aumento no consumo mundial de etanol como fonte de combustível alternativo ao petróleo.

A designação de cana-de-açúcar abrange seis espécies, sendo apenas duas encontradas em seu estado selvagem: *Saccharum spontaneum* e *S. robustum*. As demais são consideradas domesticadas ou híbridas interespecíficas: *S. officinarum*, *S. barberi*, *S. sinense* e *S. edule*. A espécie *S. officinarum* foi e continua sendo altamente explorada como a cana "nobre", com altos teores de sacarose. As variedades modernas são resultantes do intercruzamento dessa espécie com, principalmente, *S. spontaneum*, que confere características de resistência às variedades (Daniels & Roach, 1987).

Devido à sua natureza híbrida, heterozigosidade elevada e poliploidia, o estudo do genoma de cana-de-açúcar tem sido um desafio para geneticistas e melhoristas. D'Hont et al. (1994) foi pioneira em estabelecer os primeiros mapas genéticos de cana-de-açúcar. Desde então, vários grupos de ligação foram identificados no genoma de cana-de-açúcar e sorgo, a partir de dados de polimorfismo de fragmentos de restrição (RFLP), revelando alto grau de sintenia entre os genomas destas espécies (Grivet et al., 1994; Dufour et al., 1997; Guimarães et al., 1997; Ming et al., 1998; Ming et al., 2002).

Apesar dos avanços, a genética de cana-de-açúcar ainda possui muitos aspectos não esclarecidos. Neste contexto, o surgimento, a popularização e o aprimoramento das novas ferramentas de análise genômica, como as plataformas de sequenciamento de DNA

de nova geração, têm grande potencial para o entendimento da complexidade genética e genômica de espécies deste gênero.

O genoma da cana-de-açúcar é conhecido por ser um dos mais complexos entre as plantas cultivadas. As cultivares modernas apresentam elevada poliploidia, com ocorrência de aneuploidias, e têm cromossomos de origem interespecífica. Devido à sua complexidade, o genoma completo de cana-de-açúcar ainda não foi sequenciado e mesmo alguns parâmetros descritivos do genoma, como a frequência de ocorrência de sequências repetitivas, o número e a distribuição dos genes, ainda são pouco conhecidos (D'Hont, 2005).

A importância econômica da cana-de-açúcar e de seus principais produtos, para muitos países de regiões tropicais e subtropicais do mundo, nem sempre tem sido acompanhada por investimentos significativos para pesquisa e desenvolvimento de novas tecnologias de apoio aos programas de melhoramento genético. Uma das razões para isso é, provavelmente, a natureza complexa do genoma da cana e as dificuldades enfrentadas na seleção de novas cultivares em programas de melhoramento, que podem levar até 15 anos. Com o advento das novas ferramentas para análise, novas oportunidades foram criadas para o estudo do genoma de cana-de-açúcar.

Algumas plataformas de sequenciamento de DNA de nova geração (NGS – *Next Generation Sequencing)* são capazes de sequenciar 1 Mb as custos inferiores a U$ 0,10, viabilizando o estudo de genomas complexos em maior escala. O sequenciamento de DNA genômico utilizando plataformas de alta eficiência permite obter um grande volume de informações, necessário para estudo de genomas tão complexos quanto o da cana-de-açúcar. O rendimento do sequenciamento é dependente número e tamanho dos *reads* gerado em cada biblioteca. A título de exemplo, pelo uso da tecnologia Illumina (www.illumina.com), cerca de 1 Tb de sequência podem ser produzidos por corrida (HiSeq2500), e leituras de até 600 bases podem ser obtidas de cada fragmento (MiSeq) (Glenn, 2011).

Definitivamente, estamos em uma era de avanço tecnológico nas áreas ligadas à genômica. A demanda atual é pelo desenvolvimento de metodologias eficientes de análise computacional de dados se quisermos ser capazes de sair à frente nos esforços para conhecer a natureza molecular dos organismos vivos e explorar biotecnologicamente todas as suas possibilidades.

O objetivo desse trabalho foi caracterizar regiões genômicas da cultivar RB867515 hom(e)ólogas com a cultivar de cana-de-açúcar R570 previamente estudadas por Garsmeur et al. (2011). A caracterização foi realizada quanto à presença de elementos genéticos móveis, identificação de microssatélites e identificação de genes. Foi também realizada uma análise comparativa entre as sequências destas regiões nos genomas das duas cultivares.

2 REVISÃO DE LITERATURA

2.1 A CULTURA DA CANA-DE-AÇÚCAR

A cana-de-açúcar é uma importante cultura da região tropical e subtropical do mundo. É cultivada em mais de 20 milhões de hectares principalmente devido a sua grande produção de açúcar. A maioria da produção é processada em usinas especializadas para a extração de açúcar. O seu uso primário é destinado ao consumo humano, mas no Brasil é utilizada na produção de etanol, combustível de fontes renováveis que veio para substituir o uso de combustíveis fósseis (Grivet et al., 2006).

O cultivo da cana-de-açúcar no Brasil teve início em 1532, com as primeiras mudas da cultura trazidas por Martin Afonso. A boa adaptação dessa cultura em solo brasileiro estimulou a instalação das primeiras usinas de cana-de-açúcar, concentradas no Recôncavo Baiano, Pernambuco e Alagoas até o século XX. A produção de cana-de-açúcar passou por um declínio em meados de 1930, influenciado por dois fatores principais: a concentração do cultivo do café e a abolição da escravidão. Após esse período de crise, ocorreu nova expansão da agroindústria canavieira, aliada à criação do IAA (Instituto do Açúcar e do Álcool), órgão que regularia a produção de açúcar e álcool no Brasil, e a decadência da lavoura cafeeira na região Sudeste (BNDES & CGCE, 2008; Barbosa & Silveira, 2011).

Devido a uma crescente demanda em substituir a gasolina derivada do petróleo, vários testes começaram a ser feitos em veículos movidos a etanol. Para diminuir os impactos da total dependência de combustíveis fósseis e utilizar os excedentes de produção da indústria açucareira, o presidente Getúlio Vargas determinou o uso do etanol anidro em mistura com a gasolina em 1931. Ao longo dos anos o percentual mínimo de 5% sofreu alterações, chegando a 7,5% em 1975, quando a crise do petróleo impulsionou o uso desse biocombustível. Nessa época, um conjunto de incentivos foi criado pelo Proálcool (Programa Nacional do Álcool): níveis mais altos de etanol anidro na gasolina, chegando a

25%; preços para o etanol hidratado mais baixos do que os da gasolina; incentivos a usineiros; e redução de impostos na venda de carros novos movidos a etanol hidratado (BNDES & CGCE, 2008).

Com a queda do preço do petróleo e a recuperação do preço do açúcar, em 1985, a produção do etanol deixou de ser tão lucrativa. A mudança na balança dos preços deu início a uma crise no abastecimento desse produto, que refletiu na queda nas vendas de automóveis movidos a etanol. As vendas que chegaram a 85,0% do total de veículos novos em 1985, cinco anos depois caíram para 11,4%. Somente em 2003, com a introdução dos veículos bicombustíveis, a utilização do etanol hidratado voltou a ser expressiva no mercado brasileiro (BNDES & CGCE, 2008).

Segundo a Companhia Nacional de Abastecimento (Conab, 2013), a lavoura de cana-de-açúcar continua em expansão no Brasil, com previsão de aumento de 4,8% em área cultivada em relação à safra anterior (2012/13). Os estados da região Centro-Sul são os responsáveis por esse acréscimo, reflexo da expansão de novas áreas de plantio das usinas já em funcionamento. A área cultivada com cana-de-açúcar na safra 2013/14 está estimada em 8.893,0 mil hectares (mil ha). O Estado de São Paulo é o maior produtor com 51,3% (4.560,9 mil ha), Minas Gerais ficou em segundo lugar com 9,3% (828,0 mil ha), ultrapassando o estado de Goiás com 9,3% (827,0 mil ha). Os estados do Paraná, Mato Grosso do Sul, Alagoas e Pernambuco têm produção representativa, e nos demais estados produtores as áreas são menores, com representações abaixo de 3,0%. A produtividade média brasileira está estimada em 73.520 kg/ha, com perspectiva de crescimento de 4,3% em relação à safra 2012/13.

A produção total de açúcar para a safra 2013/14 no Brasil é estimada em 43,56 milhões de toneladas, 13,6% maior que a produção da safra anterior. Desta produção, 70,9% concentram-se nas usinas da região Sudeste. A produção de etanol, por sua vez, é estimada em 25,77 bilhões de litros, representando um incremento de 2,13 bilhões de litros em relação ao período anterior. Deste total, 11,37 bilhões de litros serão de etanol anidro e 14,40 bilhões de litros serão de etanol hidratado (Conab, 2013).

2.2 ORIGEM DAS CULTIVARES MODERNAS DE CANA-DE-AÇÚCAR

A cana-de-açúcar tem sido a fonte principal de açúcar para os seres humanos por vários milênios (D'Hont et al., 2008). A partir do século XVI, a produção de açúcar para o comércio mundial mudou progressivamente de indústrias caseiras baseadas em *S. sinense* e *S. barberi*, para indústrias baseadas em canas nobres, como *S. officinarum*. Mas alguns problemas começaram a surgir. Apesar das canas nobres, exploradas pelas indústrias, serem mais produtivas que as Crioulas (pela alta concentração de açúcar e baixo teor de fibras), estas eram suscetíveis a doenças. A intensa substituição de variedades pelas indústrias, sem um devido controle fitossanitário, levou à disseminação de diversas doenças nas plantações de cana-de-açúcar (Roach, 1995).

A solução para contornar o problema da vulnerabilidade das canas nobres, mantendo suas propriedades de interesse, veio com o melhoramento genético. Dois eventos podem ser considerados como marcos no processo de melhoramento genético de cana-de-açúcar. O primeiro deles, em meados de 1858, foi a descoberta de sua reprodução sexual por pesquisadores em Barbados, posteriormente também realizada por pesquisadores de Java, em 1885. O segundo, mais tarde, foi a descoberta dos benefícios da hibridização interespecífica (Roach, 1995).

Os primeiros intercruzamentos bem sucedidos de cana-de-açúcar foram realizados em 1893, por pesquisadores em Java, entre *S. officinarum* e a cana Kassoer, considerada como selvagem na época, mas que anos depois foi identificada como uma variedade híbrida entre *S. officinarum* e *S. spontaneum.* Desde então, diversos cruzamentos interespecíficos foram realizados no intuito de se obter uma variedade superior, que combinasse alto teor de açúcar e resistência a doenças (Bremer, 1961).

Em 1921 foi relatado o surgimento da "cana maravilha", a POJ2878 (POJ – *Proefstation Oost-Java*). Esta variedade foi assim chamada pela presença de muitas características desejadas em uma cultivar de cana-de-açúcar: planta alta, entrenós longos, bom perfilhamento, alto teor de açúcar e resistência a doenças como "Sereh" e mosaico. O sucesso na obtenção dessa cultivar, produzida em Java, foi possível graças ao processo de nobilização, em que as características indesejáveis das canas selvagens, utilizadas no intercruzamento com canas nobres, são diluídas por retrocruzamentos sucessivos com as canas nobres (Berding & Roach, 1987; Roach, 1995).

De modo geral, os cruzamento entre a cana nobre e a selvagem foram realizado uma, duas ou três vezes, caracterizando a primeira, a segunda e a terceira nobilização, respectivamente. Assim, a primeira nobilização refere-se ao cruzamento direto da cana nobre (*S. officinarum*) e a selvagem. O retrocruzamento entre o F_1 da primeira nobilização e a cana nobre é chamado segunda mobilização. Quando as variedades de cana pertencentes à segunda nobilização foram cruzadas novamente com a espécie nobre caracteriza-se a chamada terceira nobilização. A variedade POJ2878 é resultado da terceira nobilização. Há relatos de canas de quarta nobilização, mas estas variedades foram, em geral, um pouco mais sensíveis a doenças do que as canas de terceira nobilização (Bremer, 1961).

Durante os processos de nobilização, os melhoristas de Java relataram um padrão diferenciado de transmissão dos cromossomos dos genitores. O cruzamento entre *S. officinarum* (genitor feminino, 2n=80) com *S. spontaneum* (genitor masculino) resultou em progênies com o número somático de cromossomos do genitor feminino mais o número gamético de cromossomos do genitor masculino (*2n*+*n*). Entretanto, no cruzamento recíproco o esperado número de cromossomos (*n*+*n*) foi obtido. Esse padrão aconteceu somente no cruzamento entre essas espécies, tendo *S. officinarum* como genitor feminino (Bremer, 1961; Berding & Roach, 1987; Roach, 1995).

Diversos mecanismos citogenéticos levam à transmissão do número somático do genitor para a progênie (Figura 1). Gametas *2n* podem originar-se por endoduplicação ou por fusão de dois núcleos após a segunda divisão meiótica (os produtos serão qualitativamente diferentes). No primeiro caso, haveria recombinação normal dos genes durante a meiose I e a duplicação na meiose II resultaria em 100% de gametas homozigóticos. No segundo caso, os gametas *2n* seriam geneticamente equivalentes aos produtos da restituição da segunda divisão (SDR), portanto seriam apenas parcialmente homozigóticos. Este fato explica a rápida restituição do genoma do parental recorrente (cana nobre) nos retrocruzamentos (Bremer, 1961; Bhat & Gill, 1985; Roach, 1995).

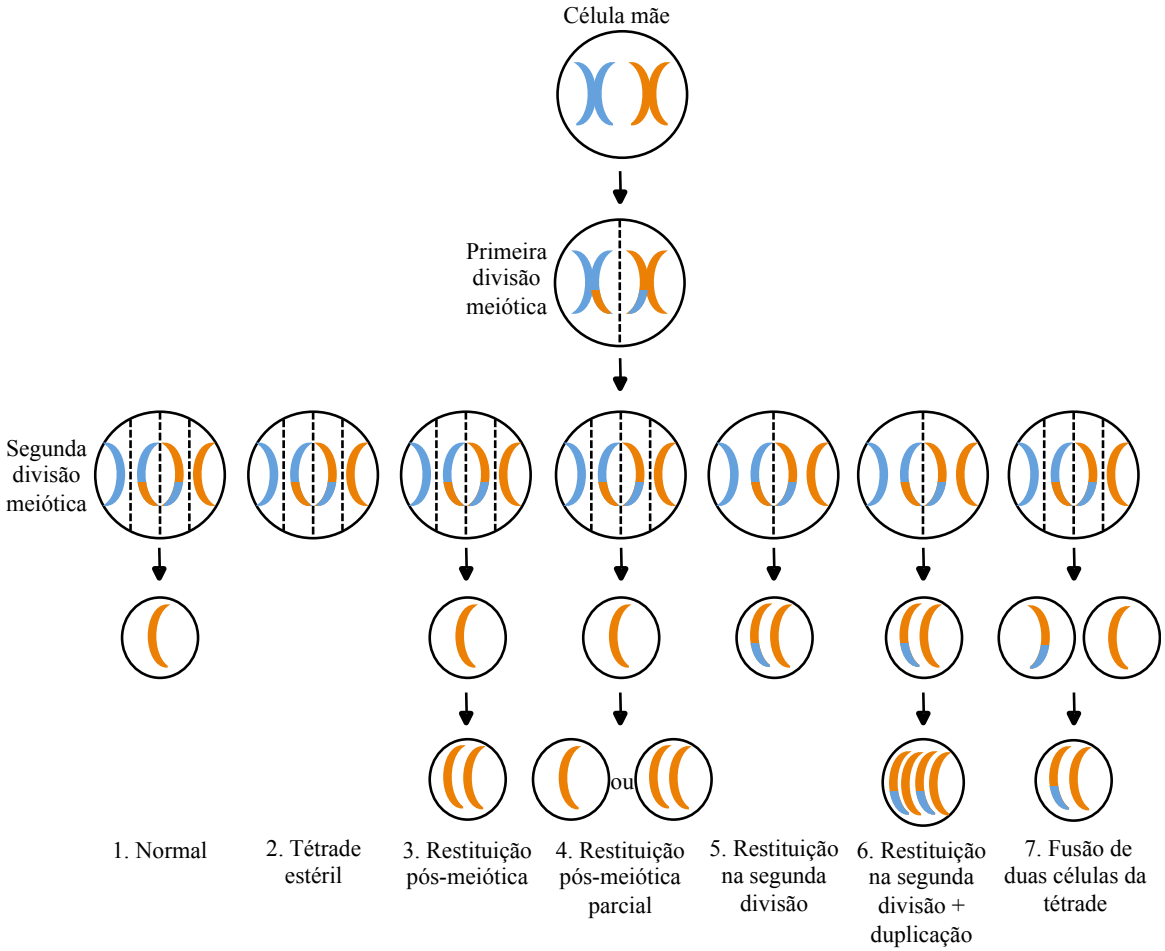


Figura 1. Esquema de sete cenários identificados em megasporogênese de cana-de-açúcar. Os seis primeiros são descritos por Bremer (1959)[3] e o sétimo é descrito por Narayanaswami (1940)[4]. (1) Normal, meiose padrão com gametas reduzidos após duas divisões meióticas; (2) Tétrade estéril, ocorre quando todas as quatro células são degeneradas após a segunda divisão meiótica; (3) Restituição pós-meiótica e (4) Restituição pós-meiótica parcial, ocorrem quando existe uma completa ou incompleta duplicação dos cromossomos após a segunda divisão meiótica, formando gametas *2n* (3), ou um gameta com apenas alguns cromossomos duplicados (4); (5) Restituição na segunda divisão, em que a segunda divisão meiótica pode não ocorrer, resultando em um gameta *2n*; (6) Restituição na segunda divisão seguida por duplicação; (7) Fusão de duas células da tétrade (Adaptado de Hermann et al., 2012).

A perda de vigor que aparece nos produtos da segunda nobilização é explicada parte pelo aumento do grau de homozigose e parte pela diminuição do número de cromossomos de *S. spontaneum*. A partir da terceira nobilização é evidente o aumento de vigor em híbridos provenientes de transmissão *n*+*n* de ambos os genitores, enquanto nova

[3] BREMER, G. Increase of chromosome number in species hybrids of *Saccharum* in relation to the embryosac development. Dordrecht: M. Nijhoff, 1959. 99 p.

[4]NARAYANASWAMI, S. Megasporogenesis and the origin of triploids in *Saccharum*. **Indian J. Agric. Sci.**, New Delhi, v. 10, p. 534, sep. 1940.

queda é observada em híbridos em que a transmissão *2n+n* prevalece. Assim, a expressão fenotípica de vigor e heterose em híbridos interespecíficos de *Saccharum* spp. é função não só da proporção relativa de cromossomos nobres e selvagens, mas também é dependente do grau de heterozigozidade presente nos locos nobres (Bhat & Gill, 1985).

2.2.1 O melhoramento genético e as cultivares de cana-de-açúcar no Brasil

O primeiro programa de melhoramento genético para a cultura da cana-de-açúcar no Brasil foi criado pelo IAC em 1933. Já em 1971 foram criados outros programas, na mesma época da criação do Planalsucar (órgão vinculado ao Instituto do Açúcar e do Álcool – IAA), que tinha como objetivo desenvolver tecnologias para o setor sucroalcooleiro. Em 1989 os órgãos Planalsucar e IAA foram extintos. Um ano depois seis universidades federais brasileiras (UFPR, UFSCar, UFV, UFRJ, UFAL e UFRPE[5]) se uniram e assumiram os trabalhos desses órgãos instituindo a Rede Interuniversitária para o Desenvolvimento do Setor Sucroenergético (RIDESA). Em 2004, a Universidade Federal de Goiás (UFG) foi agregada à rede, e em 2008 foi a vez das Universidades Federais do Mato Grosso (UFMT), Sergipe (UFSE) e Piauí (UFPI) (Barbosa & Silveira, 2011).

O Brasil é um dos países pioneiros na produção de variedades de cana-de-açúcar com qualidade comercial. No país existem, principalmente, quatro programas de melhoramento genético para a espécie: o da RIDESA, o do Centro de Tecnologia Canavieira (CTC, extinta Copersucar) e o do Instituto Agronômico de Campinas (IAC). A Canavialis, criada em 2004, foi comprada pela empresa Monsanto em 2008 (Barbosa & Silveira, 2011).

De acordo com o censo varietal da cana-de-açúcar no Brasil (safra 2011) a cultivar RB867515 (*Saccharum* spp.) é a mais cultivada no país, com 22,1% e uma área de cultivo de 1.331.017 ha (http://www.ridesa.com.br). Essa cultivar foi obtida através do cruzamento entre a RB72454 e pólen de genitor desconhecido, na estação de cruzamento da Serra do Ouro, na Universidade Federal de Alagoas (9º13’ latitude , 35º50’ longitude e 450 m altitude). A cultivar foi lançada pela Universidade Federal de Viçosa em 1997, no

[5] UFPR: Universidade Federal do Paraná; UFSCar: Universidade Federal de São Carlos; UFV: Universidade Federal de Viçosa; UFRJ: Universidade Federal do Rio de Janeiro; UFAL: Universidade Federal de Alagoas; UFRPE: Universidade Federal Rural de Pernambuco.

âmbito da RIDESA, e é protegida pelo Serviço Nacional de Proteção de Cultivares (SNPC) do Ministério da Agricultura (Barbosa et al., 2001).

A cultivar RB867515 apresenta alta produtividade agroindustrial, ótima adaptabilidade e estabilidade de produção em solos de baixa fertilidade natural e menor capacidade de retenção de água. Dada a sua baixa intensidade de perfilhamento, recomenda-se o seu plantio com densidade de 18 a 20 gemas por metro linear. Se mostra muito responsiva à aplicação de maturadores, possibilitando a antecipação do corte para início de safra em solos de menor retenção de água. Em relação a doenças, apresenta tolerância à ferrugem, mosaico e complexo broca-podridões, e tolerância intermediária ao carvão e à escaldadura (Barbosa et al., 2001).

2.3 ASPECTOS EVOLUTIVOS DO GÊNERO *Saccharum*

2.3.1 As espécies do gênero *Saccharum*

A cana-de-açúcar pertence à família Poaceae (Gramineae), subfamília Panicoideae, tribo Andropogoneae, subtribo Saccharinae Benth. e gênero *Saccharum* (Daniels & Roach, 1987). O gênero pertence a um complexo de espécies relacionadas que se intercruzam e têm sua origem envolvida com a da cana-de-açúcar, o complexo *Saccharum*, formado por espécies do gênero *Erianthus*, *Miscanthus*, *Sclerostachya*, *Narenga* e *Saccharum* (Mukherjee, 1957).

O gênero *Saccharum* é constituído por seis espécies: em que *S. spontaneum* e *S. robustum* são selvagens, enquanto *S. officinarum*, *S. barberi*, *S. sinense* e *S. edule* são cultivadas (Screenivasan et al., 1987). As cultivares modernas de cana-de-açúcar (*Saccharum* spp.) são consideradas híbridas entre as espécies *S. officinarum*, a "cana nobre", e a selvagem *S. spontaneum* (D'Hont et al., 1996; Jannoo et al., 1999; Piperidis et al., 2010).

Dados provenientes de isoenzimas, de polimorfismos de comprimento de fragmentos de restrição (RFLP), de polimorfismos de comprimento de fragmentos amplificados (AFLP), de microssatélites (SSR) e de sequenciamento de DNA nuclear e citoplasmático, possibilitaram esclarecer as relações entre as espécies cultivadas de cana-de-açúcar e as espécies selvagens relacionadas. Esses dados apoiam a visão existente de

que a cana-de-açúcar tenha evoluído de uma linhagem específica restrita ao atual gênero *Saccharum*, que inclui canas cultivadas, mais duas espécies selvagens, *S. spontaneum* e *S. robustum* (Figura 2). A implicação desse fato é que canas cultivadas muito provavelmente surgiram a partir espécies selvagens de *Saccharum* e não sofreram introgressões secundárias com outros gêneros do complexo *Saccharum*, como *Miscanthus* e *Erianthus* (Grivet et al., 2004; Grivet et al., 2006; D'Hont et al., 2008).

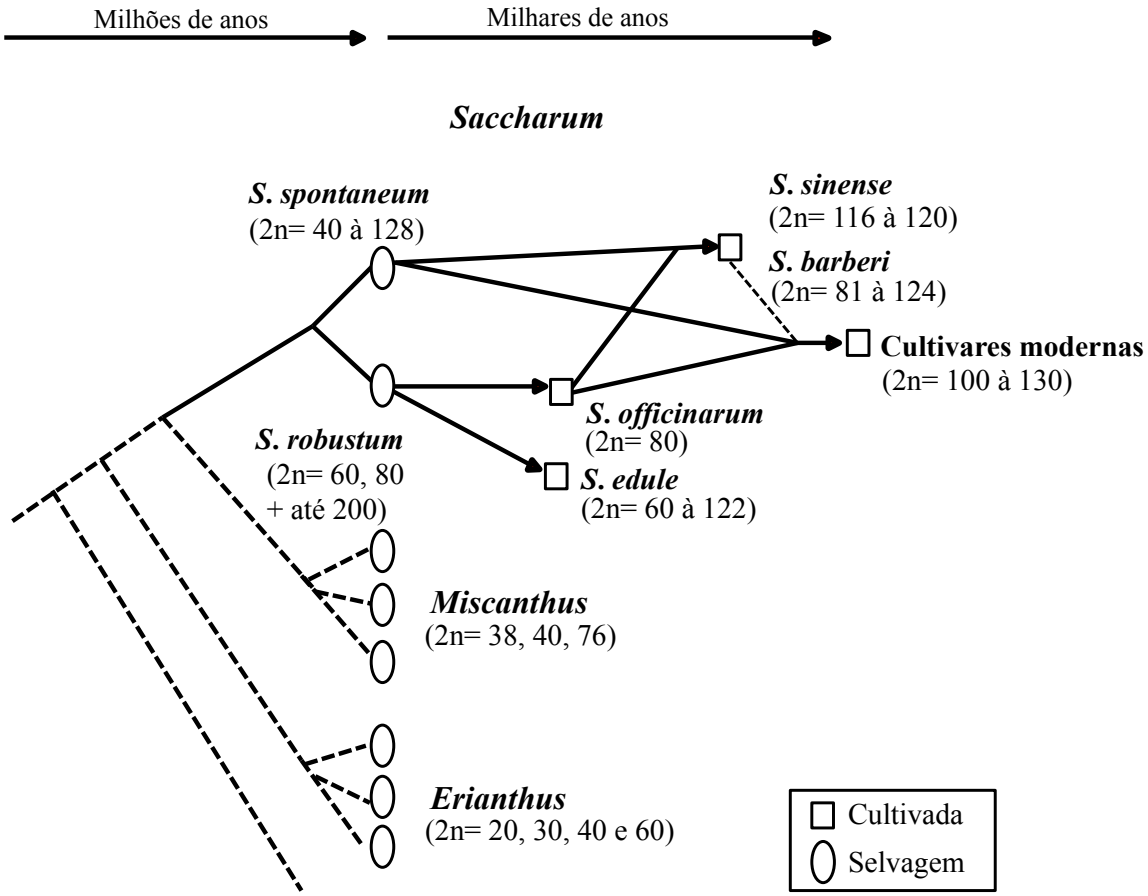


Figura 2. Cenário compatível com os dados moleculares disponíveis para a evolução e domesticação da cana-de-açúcar (adaptado de D'Hont et al., 2008).

As espécies de cana-de-açúcar apresentam grande diversidade de tamanho, forma, composição, origem e centros de distribuição. Embora as espécies possuam fenótipos diferenciados, análises filogenéticas sugerem que as espécies desse gênero têm divergência recente. O tempo de divergência entre as espécies *S. spontaneum* e as outras cinco espécies do gênero *Saccharum* é estimado em 580–780 mil anos atrás, e entre essas espécies do gênero *Saccharum* (exceto *S. spontaneum*) por volta de 0–220 mil anos atrás. O complexo genoma nuclear do gênero pode ser a razão pela qual as espécies alcançaram grande diversidade em um curto período de tempo (Takahashi et al., 2005).

S. spontaneum apresenta desde plantas com aspecto arbustivo e sem colmos a clones de colmos grandes com mais de 5 m de altura. São geralmente finos e têm baixa concentração de açúcar. Possui alta adaptabilidade a diferentes condições ecológicas, crescendo geralmente nos arredores de cursos de água, regiões desérticas ou condições salinas. A espécie tem uma distribuição ampla, que cobre algumas ilhas do Pacífico, Melanésia, Ásia tropical, Oriente Médio e parte da África. A Índia é o seu centro de origem e diversidade (Daniels & Roach, 1987; Roach, 1995; Grivet et al., 2004; Grivet et al., 2006).

S. robustum tem sido reportada em ocorrência natural na Nova Guiné e nas ilhas adjacentes à Melanésia. São plantas extremamente vigorosas, formando touceiras compactas. O caule é duro e lenhoso, as vezes oco no centro, com pouco ou nenhum açúcar. Hipóteses sugerem que *S. officinarum* poderia ter evoluído a partir de *S. robustum* (Jannoo et al., 1999). Essa espécie é considerada como a cana "nobre" pela presença de colmos grandes, espessos, suculentos e ricos em açúcar. Seu centro de origem e diversidade é atribuído à região da Nova Guiné (Bremer, 1961; Daniels & Roach, 1987).

S. barberi e *S. sinense* eram cultivadas na Índia e China, respectivamente, antes da utilização em massa dos híbridos. Têm alto teor de açúcar e muitos clones são tolerantes a doenças e estresses ambientais. Entretanto, o uso dessas espécies no melhoramento é restrito devido à sua baixa floração e fertilidade (Screenivasan et al., 1987). Dados moleculares apontam para uma origem interespecífica dessas espécies, resultantes da migração de *S. officinarum* juntamente com povos Austronésios por volta de 1000 a.C. da Nova Guiné para a Índia e China, onde hibridizou com formas locais de *S. spontaneum* produzindo as espécies *S. barberi* e *S. sinense*, respectivamente (D'Hont et al., 2002).

S. edule é caracterizada pela sua inflorescência abortiva, colmos grandes e espessos, mas com ausência de açúcar. São cultivadas tradicionalmente na Melanésia e em jardins da Nova Guiné até Fiji. Sua distribuição restrita indica que foi reconhecida depois da dispersão de *S. officinarum* pelos viajantes na área da Nova Guiné (Daniels & Roach, 1987; Roach, 1995; Grivet et al., 2006).

2.3.2 Estrutura genômica das cultivares modernas de cana-de-açúcar

O genoma da cana-de-açúcar é altamente complexo: poliploide, aneuploide e com número variável de cromossomos entre as espécies. Trata-se de um genoma grande, o

que dificulta a compreensão de sua arquitetura genética. Em relação ao número de cromossomos, as espécies de cana-de-açúcar apresentam variação, tanto entre elas, quanto em clones da mesma espécie. Em decorrência da variação desses valores, o tamanho do genoma também é variável (Tabela 1).

Tabela 1. Número de cromossomos, número básico e tamanho do genoma de espécies de cana-de-açúcar

Espécie	Número diploide de cromossomos	Número básico (*x*)	Tamanho genoma (Gb)	Tamanho monoploide (Mb)
S. spontaneum	40 a 128	8	3,36 a 12,64	843,1
S. robustum	60 e 80	8	7,65 e 11,78	984,9
S. officinarum	80	10	7,50 a 8,55	984,9
S. barberi	81 a 124	-	-	-
S. sinense	116 a 120	-	-	-
S. edule	60, 70 e 80	10*	-	-

* Geralmente aneuploides.

Fonte: Price (1965); Screenivasan et al. (1987); Roach (1995); D'Hont et al. (1998); Grivet et al. (2006); Zhang et al. (2012a).

Variedades modernas de cana-de-açúcar representam híbridos complexos entre *S. officinarum* e *S. spontaneum*, principalmente, com mínima contribuição de *S. sinense*, *S. barberi* e *S. robustum*. D'Hont et al. (1996) estudaram a composição cromossômica da cultivar R570 por hibridização genômica *in situ* (GISH). Os autores afirmaram que 80% da constituição cromossômica dessa cultivar é proveniente de *S. officinarum*, por volta de 10% proveniente de *S. spontaneum* e os 10% restantes são resultado de recombinação interespecífica. Dados de hibridização *in situ* mais recentes confirmam esses resultados. Os cromossomos das cultivares atuais derivam 70-80% de *S. officinarum*, 10-23% de *S. spontaneum* e no restante dos cromossomos (8-13%) são observadas recombinações interespecíficas (Piperidis et al., 2010).

Híbridos entre *S. officinarum* e *S. spontaneum* do Instituto Cubano INICA (*Instituto Nacional de Investigaciones de la Caña de Azúcar*) foram analisados quanto à composição cromossômica por hibridização *in situ* com sondas fluorescentes por Cuadrado et al. (2004). Todas as cultivares híbridas se mostraram aneuploides para ambos os genomas parentais, com número de cromossomos variando de $2n$ = 102-127. Desses cromossomos, 16% foram herdados de *S. spontaneum* e menos de 5% são cromossomos recombinantes ou com translocações, contendo sequências das duas espécies parentais.

Jannoo et al. (1999) realizaram um estudo para investigar a base genética dos cultivares modernos de cana-de-açúcar. Foram analisados 162 clones de cana-de-açúcar, em que 109 representam cultivares modernas de origem interespecífica e 53 representam a espécie *S. officinarum*, em relação a marcadores RFLP (12 sondas de DNA combinadas com duas enzimas de restrição). Foram obtidos 386 fragmentos polimórficos. Cada clone mostrou uma grande quantidade de fragmentos por sonda/enzima, o que é esperado pela característica poliploide da espécie. Mais de 80% das marcas obtidas para *S. officinarum* foram encontradas também nos cultivares modernos, o que ilustra o efeito da nobilização.

A partir de informações obtidas por estudos de hibridização *in situ* em cana-de-açúcar foi proposta uma representação esquemática da organização cromossômica dos cultivares modernos (Figura 3). Esse esquema ilustra as principais características da constituição cromossômica da cana-de-açúcar: poliploidia elevada, aneuploidia, origem interespecífica dos cromossomos com ocorrência de cromossomos recombinantes interespecíficos e existência de diferenças estruturais entre cromossomos das espécies que a originaram (D'Hont, 2005).

Um total de 23.914 sequências de DNA e RNA de cana-de-açúcar (*Saccharum* spp.), incluindo variedades comerciais híbridas, foram listados no GenBank (http://www.ncbi.nlm.nih.gov/gquery/?term=sugarcane) em julho de 2014. Em relação às GSS (*Genome Survey Sequences*), que são sequências de regiões genômicas específicas, foram encontradas 83,176 sequências depositadas no GenBank (http://www.ncbi.nlm.nih.gov/nucgss/?term=sugarcane), na mesma data. Informações sobre genes também são bastante limitadas para o gênero. No GenBank (http://www.ncbi.nlm.nih.gov/gene/?term=sugarcane) sequências de apenas 396 genes estão depositadas, menos de 1,2% dos genes previstos para cana-de-açúcar por Vettore et al. (2003).

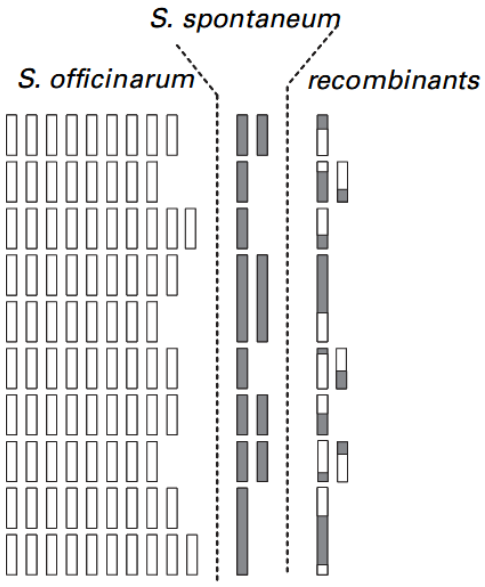


Figura 3. Representação esquemática do genoma das cultivares modernas de cana-de-açúcar a partir de dados FISH e GISH. Cada barra representa um cromossomo. As barras brancas e cinza correspondem a cromossomos, ou segmentos cromossômicos, de S. *officinarum* e *S. spontaneum*, respectivamente. Cromossomos da mesma linha são homólogos (ou homeólogos) (D'Hont, 2005).

Mesmo na ausência de uma sequência de referência para o genoma completo de cana-de-açúcar, algumas informações sobre a composição do genoma foram publicadas a partir do sequenciamento de ESTs (*Expressed Sequence Tags*). A partir de 2008, uma coleção de sequências expressas de cana-de-açúcar foram publicadas, abrindo novas oportunidades para explorar o seu genoma (Menossi et al., 2008). Uma das iniciativas de maior impacto foi o projeto brasileiro de sequenciamento de ESTs de cana-de-açúcar (SUCEST, http://sucest-fun.org). O projeto gerou 237,954 ESTs, organizados em 43,141 transcritos putativos únicos (26,803 *contigs* e 16,338 *singletons*). Com uma taxa de 22% de redundância dos transcritos, foram identificados 33,620 genes únicos (Vettore et al., 2003).

Após o sequenciamento dos ESTs de cana-de-açúcar pelo SUCEST, o passo seguinte foi a realização do estudo de identificação de funções dos genes putativos identificados, pelo projeto SUCEST-FUN (Menossi et al., 2008). Foram identificados 179 genes que são diferencialmente expressos em resposta à seca, à deficiência de fósforo, à herbivoria e à presença de bactérias endofíticas (pelo menos em uma condição) (Rocha et al., 2007). Em relação à resposta ao frio, foram identificados 64 genes diferencialmente expressos. Dentre eles, 20 genes ainda não tinham sido associados a essa condição (Nogueira et al., 2003). O mesmo conjunto de dados, analisado por outra abordagem, permitiu a identificação de 30 novos genes de reposta ao frio (Vicentini & Menossi, 2007).

Genes que regulam o teor de sacarose foram identificados como diferencialmente expressos por Papini-Terzi et al. (2009), comparando plantas de populações de baixo e alto teor de sacarose. Foram descritos 238 genes com nível expressão associado ao conteúdo de sacarose, que podem ser potencialmente úteis no desenvolvimento de plantas com maiores teores desse polissacarídeo. A análise de expressão gênica diferencial em populações contrastantes para teor de sacarose sugeriu uma possível sobreposição dos processos de biossíntese de sacarose com processos de resistência à seca e metabolismo de parede celular.

Depois da descoberta de genes desencadeada pelo sequenciamento de ESTs pelo SUCEST, outros projetos envolvendo a identificação de novos genes de cana-de-açúcar foram publicados. Iskandar et al. (2011) investigaram genes associados ao estresse abiótico para determinar sua expressão associada ao acúmulo de sacarose em células de cana-de-açúcar. No subconjunto de genes identificados como potencialmente associados ao acúmulo de sacarose foram encontrados genes que codificam enzimas envolvidas no metabolismo de aminoácidos, transporte de açúcar e fatores de transcrição. Estes autores observaram uma correlação significativa entre a expressão de genes relacionados ao estresse com aqueles relacionados ao conteúdo de sacarose em cana-de-açúcar. Correlações significativas foram encontradas não apenas em vários tecidos do colmo, mas também em colmos maduros de diferentes genótipos.

Genes relacionados ao estresse hídrico foram estudados também pela abordagem SuperSAGE (*Super Serial Analysis of Gene Expression*). Genótipos contrastantes de cana-de-açúcar (resistentes e não-resistentes à seca) foram analisados sob estresse hídrico, utilizando *bulks* de tecidos retirados da raiz. Dos 213 genes candidatos identificados, 145 foram considerados completamente novos (sem anotação no BlastN) e demandam esforços para reconhecer sua função (Kido et al., 2012).

Zhang et al. (2013) investigaram genes que sintetizam sacarose em três espécies do gênero *Saccharum* que contribuíram para os cultivares modernos de cana-de-açúcar: *S. officinarum*, *S. robustum* e *S. spontaneum*. Cinco genes foram identificados e caracterizados, constituindo a família gênica ScSuSy (*Sugarcane Sucrose Synthase*). Pela análise da frequência de SNPs nos transcritos, os autores concluíram que estes genes se diferenciaram antes da divergência do gênero na tribo Andropogoneae, há pelo menos 12 milhões de anos. Foi também descoberto um novo gene que codifica uma família de proteínas denominadas de Metalotioneínas (ScMT2-1-3). Essas proteínas desempenham

um papel importante no mecanismo de tolerância e acumulação de metais pesados em plantas (Guo et al., 2013).

Com o intuito de reunir informações de sequências do transcriptoma, anotação e dados de expressão gênica de organismos simbiontes de cana-de-açúcar foi desenvolvido um banco de dados, denominado SymGRASS (www.symgrass.org). O banco contém dados de genes ortólogos que fazem parte de vias comuns de organismos simbiontes de cana-de-açúcar, com a intensão de expandir as informações para outras gramíneas (Belarmino et al., 2013).

As bibliotecas obtidas pelo SUCEST também permitiram a investigação da composição de elementos transponíveis no genoma de cana-de-açúcar e seu padrão de expressão. Um total de 276 fragmentos foram completamente sequenciados e classificados em famílias, de acordo com a sua similaridade a um elemento já caracterizado (considerado como referência). Dos 276 fragmentos homólogos identificados, 128 eram elementos representantes da Classe I (retrotransposons) e 148 da Classe II (transposons), agrupados em 21 famílias de elementos diferentes. Os retrotransposons foram classificados nas superfamílias *Copia-Ty1* e *Gypsy-Ty3* e aqueles não classificados foram descritos apenas como LTR. A superfamília *Copia-Ty1* foi a mais abundante, com 64% dos elementos Classe I (Rossi et al., 2001).

Araujo et al. (2005) caracterizaram 68 fragmentos de cDNA, previamente identificados como de elementos transponíveis de cana-de-açúcar, representando 11 famílias. *Mutator* (38%) e *Hopscotch* (18%) foram as mais representadas entre as famílias de transposons e retrotransposons, respectivamente, no transcriptoma de cana-de-açúcar. *Callus* foi identificado como o tecido com maior número de elementos transponíveis expressos. Diferentes representantes dentro de uma mesma família exibiram padrões de expressão diferenciais. Cada família apresentou expressão em quase todos os tecidos investigados, não tendo sido identificadas famílias com expressão tecido-específica.

É de conhecimento geral que a cana-de-açúcar é uma espécie poliploide, mas o comportamento dos genes e dos elementos genéticos repetitivos nesse contexto ainda é pouco caracterizado. Garsmeur et al. (2011) analisaram sete haplótipos hom(e)ólogos provenientes do sequenciamento de oito BACs (*Bacterial Artificial Chromosomes*) da cultivar R570, que apresentam o gene de resistência à seca (Bru1), além de três BACs de sorgo (ortólogos aos BACs de cana-de-açúcar). Em quase 1 Mb do genoma dessa cultivar sequenciado, foram anotados 15 genes, com suas versões alélicas distribuídas em seis ou

sete haplótipos. Foi observada alta colinearidade geral dos genes entre os haplótipos de cana-de-açúcar, além de uma alta conservação da estrutura e sequência dos alelos homólogos e homeólogos. A colinearidade geral foi perturbada por apenas algumas duplicações segmentares. Foram identificados também 66 elementos genéticos transponíveis, que representaram, em média, 35% de todas as sequências (entre 15% e 54% de cada um dos BACs). A maioria deles apresentou similaridade com famílias já descritas, mas 21% foram descritos pela primeira vez. Os retrotransposons (LTR) foram os mais frequentes, representando 65% de todos os elementos, divididos nas duas superfamílias: *Gypsy/Ty3* (58%) e *Copia/Ty1* (42%). Os retrotransposons não-LTR somaram 17% e transposons de DNA representaram 18% do total. A maioria dos elementos foram encontrados em vários haplótipos hom(e)ólogos. Em contraste com a colinearidade e conservação dos genes, não foi encontrada colinearidade generalizada para os elementos móveis entre haplótipos da cultivar R570.

Em concordância com os achados de Jannoo et al. (2007), Garsmeur et al. (2011) sugeriram que a alta poliploidia em cana-de-açúcar não induziu um rearranjo generalizado do seu genoma, desafiando, assim, a ideia de que a poliploidia induz rapidamente um rearranjo generalizado de genomas. A conservação entre alelos hom(e)ólogos, apesar da extrema redundância observada nesse caso, levanta a questão dos mecanismos genéticos envolvidos. Conforme salientam estes autores, o modo de pareamento cromossômico pode ter um forte impacto sobre a evolução do genoma. Em particular, o recorrente rearranjo aleatório dos cromossomos na meiose em autopoliploides poderia atuar contra a estabilidade funcional de todos os alelos, porque poderia dar origem a indivíduos (e gametas) sem um conjunto completo de genes funcionais.

2.3.3 Relações filogenéticas do gênero *Saccharum* com outros gêneros de gramíneas

Muitos estudos já foram realizados no intuito de situar as espécies do gênero *Saccharum* dentro do grupo das gramíneas e avaliar a existência de sintenia com outras espécies dessa família. A partir dos anos 80, com o surgimento de marcadores moleculares, tornou-se possível a construção de mapas genéticos densos e a utilização de sondas de DNA para o desenvolvimento de trabalhos de genética comparativa entre espécies relacionadas.

No trabalho de Moore et al. (1995), realizado com marcadores RFLP (*Restriction Fragment Length Polymorphism*), os genomas de algumas gramíneas (arroz, milho, sorgo, milheto, cana-de-açúcar e espécies da tribo Triticeae) apresentaram similaridade com grupos de ligação identificados em arroz (Figura 4). O "*cropcircle*" que representa o alinhamento dos genomas destas culturas foi atualizado dez anos mais tarde, ao nível de sequências de DNA. As novas marcas descobertas possibilitaram identificar genes por sintenia nestas espécies e selecionar genes candidatos para características de interesse (Devos, 2005).

O mapa integrado do genoma de gramíneas revelou a existência de padrões coerentes com as relações taxonômicas entre as espécies analisadas (Figura 5). O primeiro clado é representado pelo arroz (único representante da subfamília Bambusoideae); o segundo pela aveia e demais gêneros da tribo Triticeae (subfamília Pooideae); e o terceiro é representado pelos vários membros da subfamília Panicoideae, em que a cana-de-açúcar foi alocada. Cumpre destacar a proximidade identificada entre os genomas de *Sorghum* e *Saccharum*, gêneros pertencentes à tribo Andropogoneae (Devos & Gale, 1997).

D'Hont et al. (1994) foram pioneiros em estabelecer os primeiros estudos de genômica comparativa em cana-de-açúcar, utilizando informações da literatura sobre mapas genéticos de milho como referência. Um total de 32 indivíduos provenientes da variedade SP701006 (desenvolvida pela Copersucar, São Paulo, Brasil) foram analisados, utilizando quatro isoenzimas e 53 sondas de milho. Foi possível identificar alto grau de sintenia entre milho e cana-de-açúcar. Dados de sequenciamento do genoma cloroplastidial de cana-de-açúcar (*S. officinarum*) também apontaram alta similaridade com milho, mas não com arroz ou trigo (Asano et al., 2004).

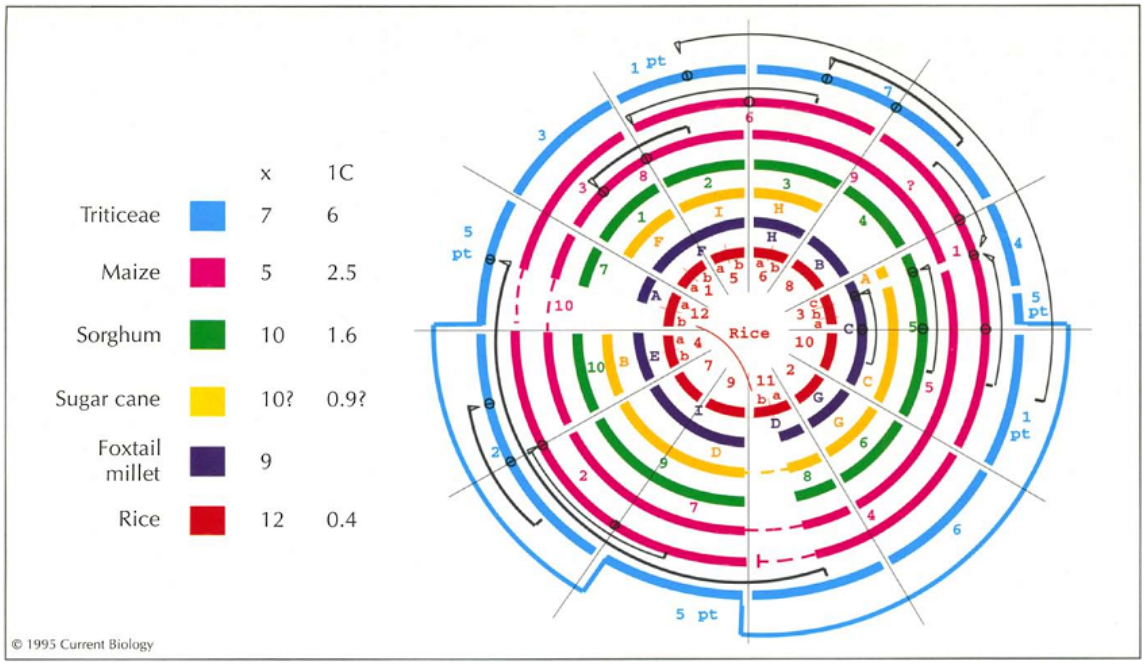


Figura 4. "*Cropcircle*". Alinhamento do genoma de seis gramíneas com 19 grupos de ligação de arroz, cuja ordem reflete o genoma circularizado do ancestral das gramíneas. Os dados foram redesenhados pelos segmentos de ligação de arroz (definido por linhas radiais) formados nos cromossomos (codificados por cores e numeração de linhas). As linhas finas tracejadas correspondem aos segmentos duplicados. Inversões de conjuntos de sequências dentro de um segmento de ligação não são mostradas. Cromossomos formados pela inserção de um segmento no outro são mostrados pelas linhas pretas com as setas indicando a direção e o ponto de inserção. Os pontos de quebra nos cromossomos (envolvidos nos eventos de inserção) são indicados pelos círculos divididos (Moore et al., 1995).

Com o desenvolvimento das sondas genômicas de milho, e sua capacidade de hibridização interespecífica, estudos de mapeamento comparativo de representantes da tribo Andropogoneae começaram a se difundir. Vários grupos de ligação foram localizados nos genomas de cana-de-açúcar e sorgo, a partir de dados de polimorfismo de fragmentos de restrição (RFLP), revelando alto grau de sintenia entre esses grupos (Grivet et al., 1994; Dufour et al., 1997; Guimarães et al., 1997; Ming et al., 1998; Ming et al., 2002).

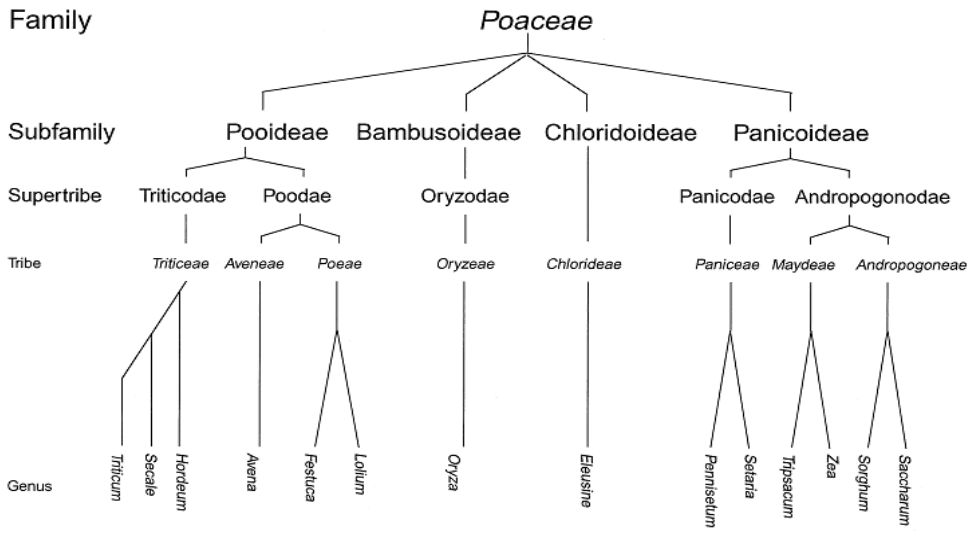


Figura 5. Relações taxonômicas entre alguns gêneros da família Poaceae (Devos & Gale, 1997).

A filogenia entre cana-de-açúcar e grupos relacionados também foi estudada pela análise de regiões de espaçadores intergênicos. Em geral, os acessos de *Erianthus* spp. (incluindo *E. giganteus*, também denominada de *Saccharum giganteum*), *Miscanthus sinensis*, *Sorghum bicolor* e *Zea Mays* tenderam a agrupar-se para formar grupos distintos. Verificou-se que as espécies de cana-de-açúcar estão mais estreitamente relacionadas com *M. sinensis* e *E. giganteus* (Complexo *Saccharum*) do que com o sorgo e milho (Pan et al., 2000).

Um estudo recente tentou avaliar a microssintenia entre 20 cromossomos artificiais de bactéria (BAC) de cana-de-açúcar e sequências de sorgo, através de sequenciamento de nova geração (pirossequenciamento 454). As regiões gênicas dos BACs de cana-de-açúcar compartilharam, em média, 95,2% de identidade com sorgo. Aproximadamente 53% das sequências dos BACs de cana alinharam com o genoma de sorgo. As regiões não alinhadas consistiram, basicamente, de sequências não codificadoras e repetitivas (Wang et al., 2010).

2.4 EVOLUÇÃO DO GENOMA DAS GRAMÍNEAS

A poliploidia é considerada como uma força importante e significativa na evolução das plantas, em diversas escalas temporais, e com amplos efeitos na configuração do genoma destas espécies (Adams & Wendel, 2005; Doyle et al., 2008).

Independentemente do tipo de evento de poliploidização – se fusão de dois genomas completamente diferenciados (alopoliploides), duplicação de um único genoma (autopoliploides), ou algo entre os dois – todos, ou quase todos os genes e outras sequências genômicas se duplicam no processo (Wendel, 2000). Os eventos de poliploidia implicam muito mais do que a simples fusão de dois genomas, pois envolvem vários processos de ajustes fisiológicos e moleculares (Adams & Wendel, 2005).

Duplicações genômicas têm ocorrido repetidamente durante a evolução das plantas, mas inferências sobre o número e o tempo desses eventos ainda variam na literatura (Adams & Wendel, 2005). O genoma das gramíneas foi submetido a diferentes eventos de duplicações completas e segmentares durante a sua evolução a partir de um ancestral comum há 50-70 milhões de anos, com número de cromossomos *n*=5 (Salse et al., 2008). Um evento antigo de duplicação do genoma é relatado para o ancestral comum das gramíneas há 50-70 milhões de anos atrás e um outro evento mais recente na linhagem que origina o gênero *Saccharum* (Figura 6) (Adams & Wendel, 2005).

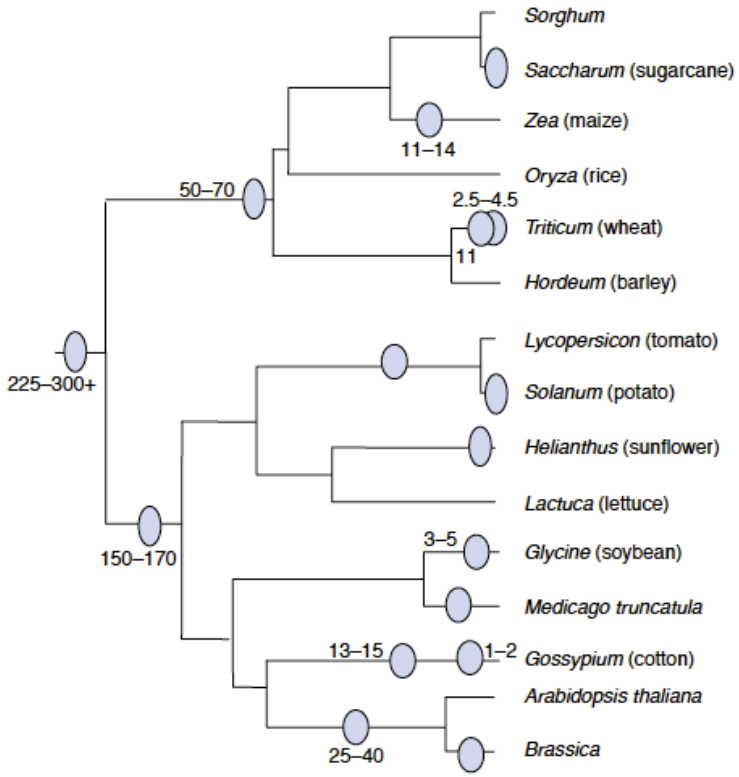


Figura 6. Eventos de poliploidização durante a evolução das angiospermas. Círculos azuis indicam os eventos de duplicação em larga escala. O comprimento dos ramos não estão em escala. Os números indicam datas aproximadamente estimadas (em milhões de anos) desde os eventos de duplicação (Adams & Wendel, 2005).

Apesar da relação evolutiva próxima entre as gramíneas, os cromossomos sofreram extensos rearranjos. Uma reconstrução cromossômica ancestral das gramíneas revelou que o ancestral tinha 12 cromossomos (depois de um evento de duplicação completa e dois eventos de fissões cromossômicas). O ramo Panicoideae passou por mais dois eventos de fissões cromossômicas há aproximadamente 48 milhões de anos (Figura 7), dando origem às gramíneas C_4 como milheto, sorgo e milho (e cana-de-açúcar, não incluída na figura) (Zhang et al., 2012b).

A poliploidia duplica todos os genes do genoma, fornecendo material bruto para divergência ou a divisão de funções nas cópias homólogas. A retenção preferencial e a perda de genes ocorrem em taxas variáveis, o que sugere que há uma série de princípios que governam o destino dos genes duplicados (Doyle et al., 2008).

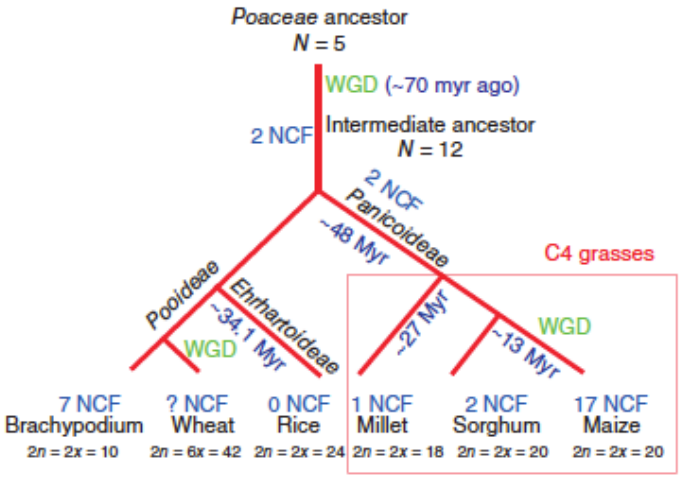


Figura 7. Relações evolutivas de três subfamílias de gramíneas, incluindo tempos de divergência, eventos de duplicação total do genoma (WGD) e fissão cromossômica (NCF) (Adaptado de Zhang et al., 2012b).

Um resultado evolutivo frequente da duplicação genética é a retenção da função para ambas as cópias do gene. O aumento do número de cópias do genoma pode alterar os níveis de expressão dos genes, que podem sofrer efeitos de dosagem. A expressão do gene pode ser diminuída quando a sua dosagem aumenta (*down-regulation*), ou pode ocorrer a compensação desses efeitos, quando os níveis de expressão são mantidos independentemente do número de cópias do gene (*dosage compensation*). Muitos aspectos sobre a evolução dos poliploides ainda estão obscuros, como o papel dos elementos transponíveis na evolução estrutural e regulatória dos genes; os processos de silenciamento epigenético e sua significância; os controles subjacentes ao emparelhamento de

cromossomos; os mecanismos de mudanças rápidas no genoma e o seu significado funcional (Wendel, 2000).

O sucesso evolutivo dos poliploides pode estar associado ao relaxamento da pressão seletiva sob uma cópia do gene, permitindo a divergência entre os genes duplicados e a aquisição de novas funções. Em contrapartida, pode haver silenciamento de genes por meio de mutação ou por processos epigenéticos, como a ação de elementos transponíveis que podem ser inseridos dentro de genes alterando sua expressão (Wendel, 2000; Adams & Wendel, 2005).

2.4.1 Elementos genéticos móveis nos genomas das plantas

Os padrões estruturais do genoma nuclear de plantas com sementes têm sido conservados durante os 100 milhões de anos da evolução das angiospermas. Todas as angiospermas apresentam genomas relativamente complexos, em que as sequências gênicas e regulatórias somam uma pequena fração do conteúdo do DNA nuclear e a maior porção é ocupada por DNA repetitivo. Nas gramíneas, por exemplo, eles respondem por 50 a 80% do tamanho do genoma (Kellog & Bennetzen, 2004).

O conteúdo de DNA repetitivo pode ser dividido em dois grupos principais, que se diferenciam pela organização e localização nos cromossomos. O primeiro grupo compreende as sequências com repetições em tandem, encontradas preferencialmente em posições específicas nos cromossomos, como pericentroméricas, subteloméricas, teloméricas ou regiões intercalares. Estão incluídos nesse grupo os DNAs satélites, repetições teloméricas e o DNA ribossômico. O segundo grupo é composto pelos elementos com distribuição dispersa, espalhados pelo genoma e intercalados com outras sequências ao longo dos cromossomos, os transposons (Kubis et al., 1998; Kellog & Bennetzen, 2004).

Os elementos genéticos móveis foram descobertos na década de 50 por Barbara McClintock, durante seus estudos sobre os diferentes padrões de coloração dos grãos de milho. Esses elementos têm dirigido a evolução do genoma das plantas pela sua capacidade de se integrar ao genoma, em um novo local, dentro de sua célula de origem (Kazazian Jr., 2004). Esses elementos são ainda reconhecidos como ferramentas moleculares naturais que têm moldado a organização, a estrutura e a função dos genes e genomas ao longo da sua evolução (Miller & Capy, 2006).

Os elementos genéticos móveis podem atuar em diferentes vias, classificadas como destrutivas, quando envolvem inserções e rearranjos (devido à recombinação homóloga desigual), e construtivas, quando agem no reparo da dupla hélice do DNA, na regulação da expressão gênica e em outros processos. Embora existam mecanismos pelos quais o genoma hospedeiro pode controlar o seu número, expansões maciças de retrotransposons foram aparentemente toleradas durante a evolução (Kazazian Jr., 2004).

Um estudo baseado nas relações evolutivas de transcriptases reversas permitiu a construção de uma árvore filogenética para 82 retroelementos de animais, plantas, protozoários e bactérias. A partir da comparação da organização genética dessas transcriptases pode-se inferir que o mais provável ancestral dos retroelementos atuais foi um elemento retrotransponível com genes gag-like e pol-like (Xiong & Eickbush, 1990).

O primeiro elemento genético que foi reconhecido como sendo transponível foi um sítio de quebra em um cromossomo de milho, nomeado de *Dissociation* (Ds). Esse elemento podia se transpor, ou quebrar cromossomos, somente na presença de outro loco, chamado de *Activator* (Ac), que também promove sua própria transposição. Juntos esses elementos constituem uma família que inclui elementos autônomos (Ac) e não-autônomos (Ds) (Feschotte et al., 2002).

Foi proposta uma classificação unificada para os transposons de eucariotos (Figura 8). Nesse sistema hierárquico, os elementos são classificados de acordo com seus mecanismos de transposição, similaridades de sequências e relações estruturais. O primeiro nível hierárquico continua sendo as classes, em seguida subclasses (opcional), ordem, superfamília, família e subfamília (Wicker et al., 2007).

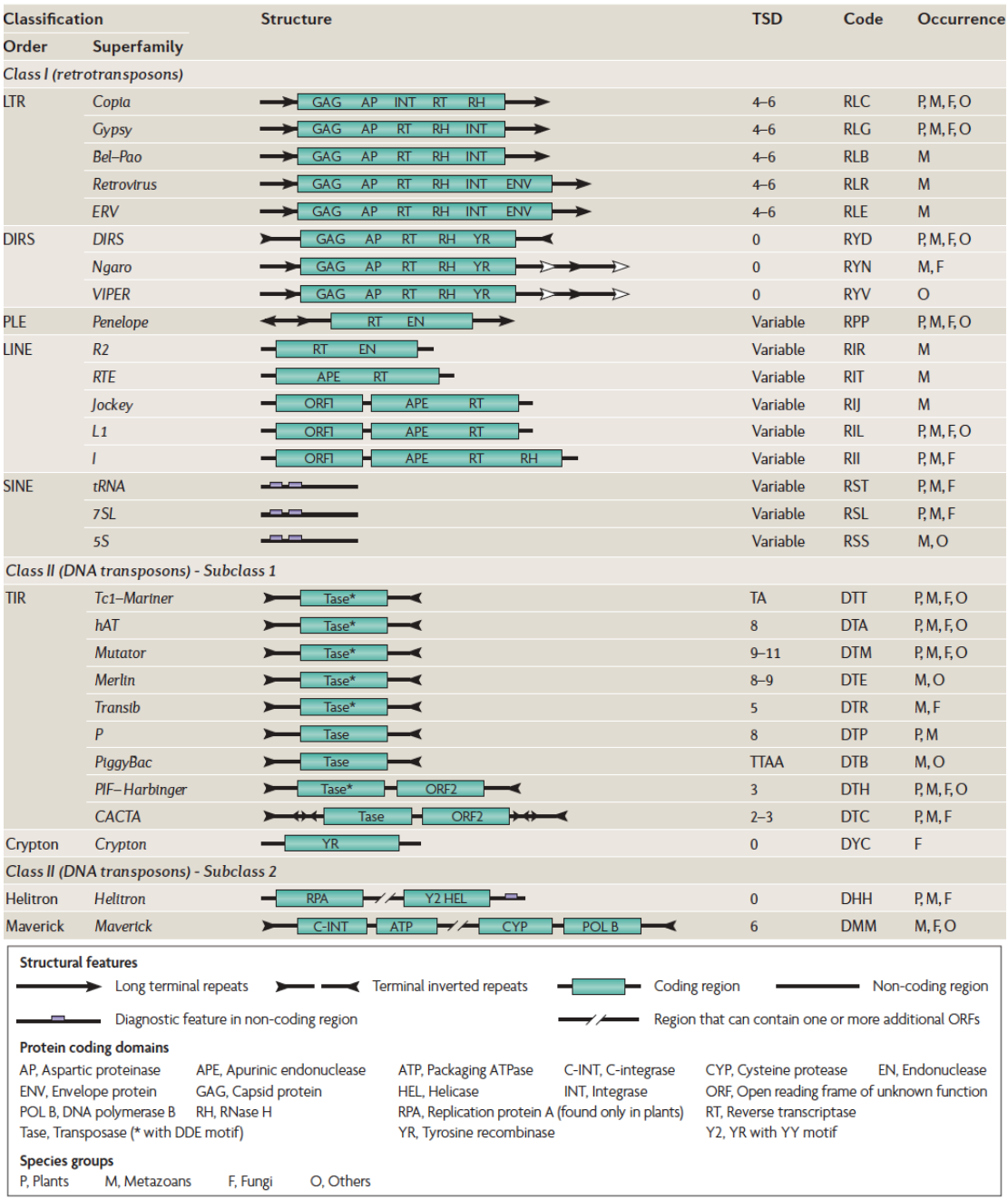


Figura 8. Sistema de classificação para transposons. A classificação é hierárquica e divide os elementos em duas classes principais, com base na ausência ou presença de RNA como transposição intermediária. Essas são divididas, por sua vez, em subclasses, ordens e superfamílias. O tamanho do sítio alvo de duplicação (TSD), característico de muitas superfamílias é usado para diferencia-las. A tabela apresenta um sistema de três letras que descreve os três maiores grupos: DIRS, *Dictyostelium intermediate repeat sequence*; LINE, *long interspersed element*; LTR, *long terminal repeat*; PLE, *Penelope-like elements*; SINE, *short interspersed element*; TIR, *terminal inverted repeat* (adaptado de Wicker et al., 2007).

Os transposons estão divididos em duas classes (I e II), classificados pela presença ou ausência da transposição de um RNA intermediário. Subclasses são distinguidas por elementos que se autocopiam para inserção, daqueles que deixam o local

original para se integrar em outro lugar (Figura 9). As superfamílias são caracterizadas pelo sistema de replicação, mas distinguidas pelas estruturas de proteínas e regiões não codificadoras. Uma das principais diferenças entre superfamílias é a presença e tamanho do sítio alvo de duplicação (TSD, *Target Site Duplication*), que é uma repetição direta pequena que é gerada nas extremidades de inserção de um transposon. As superfamílias, por sua vez, são divididas em famílias, definidas pela conservação da sequência de DNA (Wicker et al., 2007).

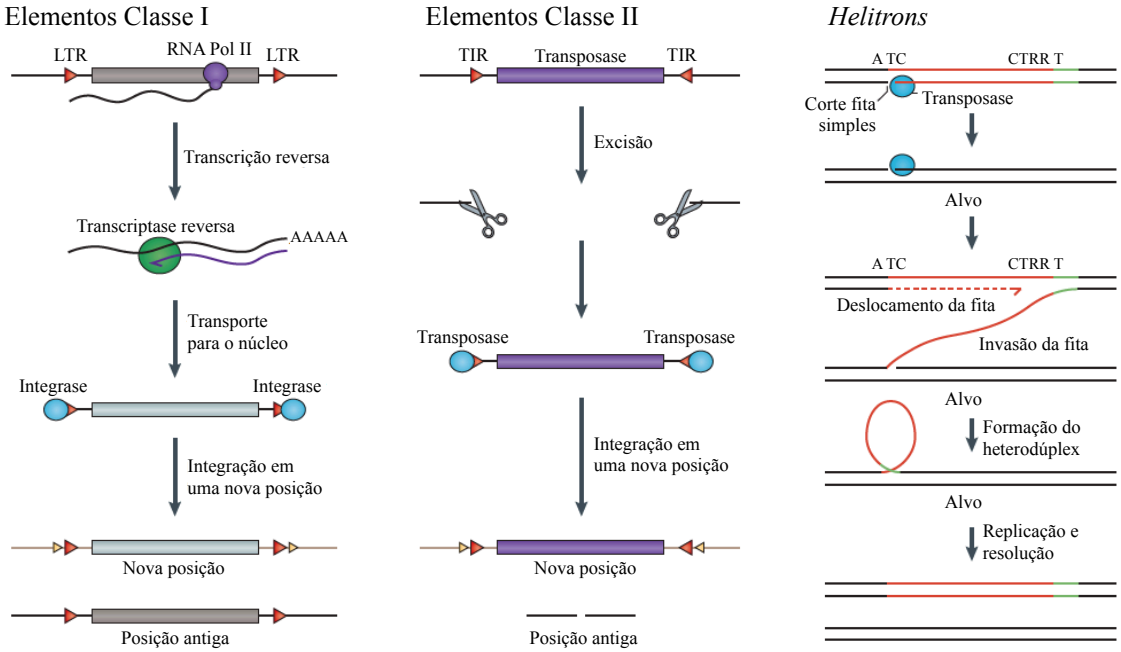


Figura 9. Mecanismos de transposição dos elementos móveis classe I, II e Helitrons (Adaptado de Lisch, 2013).

Elementos da classe I, também conhecidos como retrotransposons, transpõem-se por um RNA intermediário. Uma cópia do elemento é produzida na forma de transcrito de RNA, e revertido (via transcrição reversa) em DNA para se reintegrar no genoma (Figura 9). Esses elementos não possuem subclasses e são divididos em cinco ordens: retrotransposons LTR (*Long Terminal Repeats*), *DIRS*-like, PLEs (*Penelope*), LINEs (*Long Interespersed Elements*) e SINEs (*Short Interespersed Elements*). As três ordens mais relevantes para o genoma de plantas são os retrotransposons LTR, LINEs e SINEs (Wicker et al., 2007), que serão descritos com mais detalhes.

Os retrotransposons LTR variam amplamente de tamanho nos cerais, de 100 pb até mais de 5 kb. São compostos por várias ORFs (do inglês, *Open Reading Frames*), que

codificam proteínas específicas. As repetições terminais invertidas (TIR) dão a esses elementos uma estrutura universal 5' TG....CA 3'. Internamente às regiões 5' e 3' estão os sítios de ligação da transcriptase reversa (PBS – do inglês, *Priming Binding Sites*) e entre esses dois locais estão as ORFs (Figura 10) que codificam as proteínas dos retrotransposons (Todorovska, 2007).

Dois tipos principais de retrotransposons são abundantes no genoma de plantas, *Gypsy-Ty3* e *Copia-Ty1*, que diferem na posição do gene da integrase, na segunda ORF. Em elementos *Copia-Ty1* os domínios internos codificam: proteína do capsídeo (GAG), proteinase aspártica (AP) – que cliva a poliproteína em componentes funcionais –, integrase (IN) – insere o fragmento de cDNA no genoma –, transcriptase reversa (RT) – responsável pela criação das cópias de cDNA – e RNase H (RH) que é importante para a replicação (Figura 10). Nos elementos *Gypsy-Ty3* o gene da integrase é localizado a jusante ao gene da transcriptase reversa e da RNaseH (Kubis et al., 1998; Feschotte et al., 2002; Kazazian Jr., 2004; Todorovska, 2007).

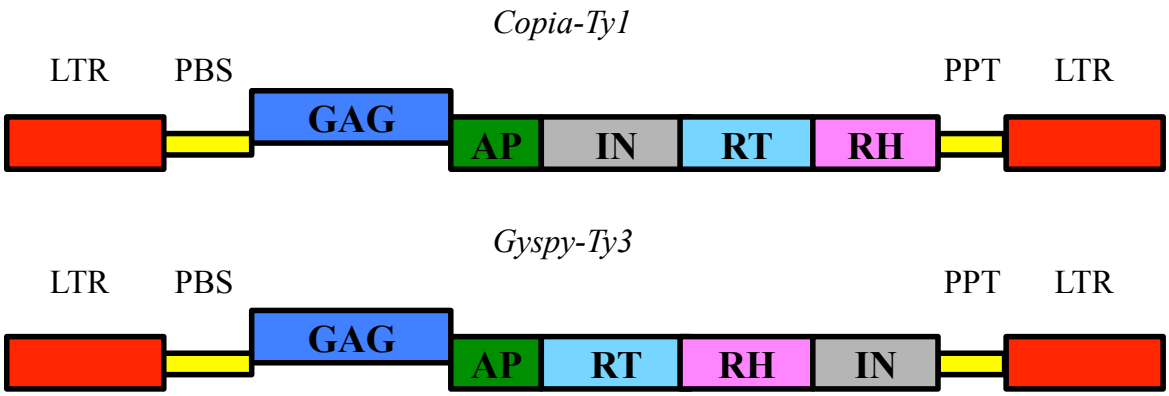


Figura 10. Esquema da composição dos elementos *Copia-Ty1* (superior) e *Gypsy-Ty3* (inferior). LTR – terminações longas repetidas (vermelho); PBS – sítios de ligação da transcriptase reversa (amarelo); GAG – proteína do capsídeo (azul escuro); AP – proteinase aspártica (verde); IN – integrasse (cinza); RT – transcriptase reversa (azul claro); RH – RNase H (rosa) (Adaptado de Todorovska, 2007).

Os elementos da ordem LINE não possuem as repetições terminais como nos LTR. São divididos em cinco superfamílias: *R2*, *L1*, *RTE, I* e *Jockey*. Membros autônomos codificam pelo menos uma transcriptase reversa (RT) e uma nuclease, que diferencia os membros das diferentes famílias. Geralmente esses elementos formam TSDs (sítios alvo de duplicação) após a inserção, mas as extremidades 5' truncadas que são formadas (provavelmente por uma terminação prematura da transcrição reversa) dificultam a localização. Na extremidade 3' podem apresentar uma cauda poli-A, repetições em tandem

ou região rica em "A". São elementos grandes, podendo chegar a kilobases de comprimento. Em plantas, as superfamílias mais comuns são *L1* e *RTE* (Wicker et al., 2007).

A ordem SINE é composta por elementos não-autônomos, portanto, necessitam de enzimas codificadas por outros elementos (LINEs) para sua transposição. São originados de retrotransposição acidental de vários transcritos de polimerase III (Pol III). Possuem um promotor para Pol III, o que permite sua expressão. Como o próprio nome sugere, *Short Interespersed Elements*, são elementos pequenos, variando de 80 a 500 pb, e geram TSDs, de 5-15 pb. A região promotora para Pol III define as superfamílias dessa ordem: *tRNA*, *7SL RNA* e *5S RNA* (Wicker et al., 2007).

Elementos da classe II geralmente estão presentes em pequeno número de cópias. São divididos em duas subclasses, distinguidas pelo número de fitas de DNA que são cortadas durante a transposição. Na subclasse 1 estão os elementos que se movem pelo tradicional mecanismo "corta-e-cola", da ordem TIR (*Terminal Inverted Repeat*), que possuem repetições terminais invertidas de tamanho variado. O reparo do sítio doador do transposon geralmente acontece pelo uso da cromátide irmã ou de um cromossomo homólogo como molde, assim, nenhuma excisão é detectada (Bennetzen & Wang, 2014). São descritas nove superfamílias que se diferenciam pela sequência TIR e o tamanho dos TSDs (Wicker et al., 2007).

Uma das superfamílias mais conhecidas é a *hAT*, bastante difundida em plantas e animais (Arensburger et al., 2011). Seus elementos caracterizam-se pela geração sítios de duplicação (TSD) de 8 pb, como consequência do mecanismo de transposição, e a presença de repetições terminais invertidas curtas (TIR) de 5-27 pb. A maioria dos elementos são menores que 4 kb de tamanho, mas podem ser encontrados alguns com 12 kb. São elementos autônomos, ou seja, possuem o gene para a transposase, que codifica o polipeptídeo catalisador da reação de transposição (Kempken & Windhofer, 2001; Wicker et al., 2007).

Umas das famílias que compõem a superfamília *hAT* é *Tag1*, que foi descoberta como uma inserção no quarto íntron de um gene transportador de nitrato (CHL1) e mapeado no cromossomo 1 de *Arabidopsis thaliana*. Tem comprimento de 3,3 kb, com TIR de 22 pb e produzem TSDs de 8 pb (Tsay et al., 1993; Shankar et al., 2001). Outro elemento dessa superfamília é *Tip100*, descoberto pela primeira vez em *Ipomoea*

purpurea (glória-da-manhã). Possui TIRs com 11 pb e produz TSDs de 8 pb (Christoff et al., 2012).

Membros da superfamília *Stowaway* estão amplamente associados à genes de plantas. Esses elementos são caracterizados pela presença de repetições terminais invertidas conservadas de 11 pb, especificidade por sítio alvo (TA), e potencial para formar estruturas de DNA secundárias estáveis. São elementos pequenos, variando de 80 a 323 pb, ricos em AT (Bureau & Wessler, 1994).

A família *En/Spm* pertence à superfamília *CMC* (*CACTA*, *Mirage*, *Chapaev*) e estão amplamente distribuídos em certas plantas. Elementos dessa família possuem uma sequência conservada na região terminal TIR (5'-CACTA-3') e genes que codificam transposase (TnpA e TnpD, em milho) com diversos domínios conservados. Não estão distribuídos igualmente nos cromossomos, mas em aglomerados, como, por exemplo, em sítios 5S rDNA (Altinkut et al., 2006; Krishnan et al., 2009).

A superfamília *Mutator* (Mule-MUDR) tem distribuição ampla em todos os reinos de eucariotos. Suas repetições terminais invertidas podem variar de tamanho, de centenas de pares de bases até pequenas demais, ao ponto de não serem detectadas. Elementos dessa superfamília produzem TSDs de 9-11 pb (Wicker et al., 2007).

Algumas superfamílias de transposons da classe II apresentam preferência por um sítio-alvo, como é o caso da superfamília *PIF-Harbinger*, que tem preferência por "TAA". Elementos dessa superfamília apresentam duas ORFs, uma para proteína que se liga ao DNA e outra para transposase (Wicker et al., 2007). Elementos autônomos *PIF/Harbinger* carregam repetições terminais invertidas de 14-25 pb, flanqueadas por 3pb (TTA/TAA) de duplicações de sítio alvo (TSD) (Grzebelus et al., 2007). Foi primeiramente reconhecida em milho, tem uma origem antiga, explicada pela sua relação com transposases bacterianas, e identificação em uma vasta gama de eucariotos, incluindo plantas, fungos e nematóides (Feschotte et al., 2002).

A subclasse 2 engloba os elementos cujo mecanismo de transposição não cliva a dupla hélice do DNA, diferente da subclasse 1. Nesta subclasse os elementos mais comuns são os da ordem Helitron e Maverick (ou Politrons) (Wicker et al., 2007). *Helitrons* são elementos presentes em muitas plantas e são caracterizados pelo seu mecanismo de transposição por "círculo rolante" (Figura 9). No processo de transposição, a região terminal do elemento é inicialmente clivada, ocorre uma invasão no sítio alvo que também foi quebrado, síntese do DNA, deslocamento das fitas e resolução do heterodúplex

(Lisch, 2013).

Os transposons não são distribuídos uniformemente no genoma. DNA transposons são preferencialmente encontrados em regiões ricas em genes e os retrotransposons são concentrados em regiões de heterocromatina. A teoria proposta é que os DNA transposons em regiões de heterocromatina seriam menos prováveis de serem transpostos devido à topologia do DNA, que o deixa inacessível às transposases. Os elementos que são excisados tendem a se reinserir em regiões próximas à sua origem, também de eucromatina. Como as transposições para regiões de éxons são mais provavelmente eliminadas pela seleção natural, a maioria desses elementos são encontrados em íntrons ou regiões adjacentes 5' ou 3' (Civán et al., 2011).

Por outro lado, os LTR-retrotransposons são encontrados preferencialmente em regiões de heterocromatina. Como esses elementos são gerados via RNA, a probabilidade da nova cópia ser introduzida próxima da origem é extremamente baixa. Assim, os retroelementos são reintegrados aleatoriamente no genoma (Civán et al., 2011).

O estudo de sequências de DNA repetitivo é essencial para a compreensão da natureza e das consequências da variação no tamanho do genoma de espécies diferentes, e para o melhor entendimento da organização e evolução dos genomas vegetais em grande escala (Kubis et al., 1998). Os elementos genéticos transponíveis são responsáveis por extensivas variações no genoma de gramíneas, tanto em regiões intergênicas quanto em relação ao conteúdo gênico. As variações podem ser observadas não somente entre espécies intimamente relacionadas, mas também entre indivíduos da mesma espécie (Morgante et al., 2007).

As grandes variações existentes nos genomas de gramíneas e a ocorrência de características não compartilhadas podem ser atribuídas à idade muito jovem do conteúdo de elementos repetitivos. Os retrotransposons LTR foram submetidos a amplificações independentes em linhagens distintas de plantas, dentro do mesmo gênero, em um curto período de tempo. Em milho, por exemplo, a maioria dos eventos de inserção de retrotransposons ocorreram nos últimos 400 mil anos. Nesse curto período é provável que tenha havido tempo suficiente para que muitas das novas inserções tenham sido eliminadas ou fixadas no conjunto de genes, por deriva genética ou seleção, de modo que aparecem hoje como polimorfismos intraespecíficos (Morgante et al., 2007).

Certas sequências de retrotransposons são conservadas em centrômeros de cereais. Elementos autônomos e não-autônomos colonizaram os centrômeros de

representantes da família Poaceae por ocasião, ou antes, do último ancestral comum, pelo menos há 60 milhões de anos. Após esse evento divergiram independentemente em cada espécie, com a seleção mantendo o potencial de retrotransposição (Langdon et al., 2000).

As sequências desses elementos, disponíveis na literatura, são hoje consideradas como originárias de uma única família ancestral, denominada *crwydryn*. Esse elemento ancestral é um típico *Gypsy-Ty3*, com uma região UTR 5' maior que 1 kb e uma ORF sobreposta à região de terminação repetida à jusante. Um dos efeitos dessa colonização foi o aumento progressivo do tamanho dos genomas, reflexo da amplificação desses elementos (Langdon et al., 2000).

Dos elementos repetitivos encontrados em plantas, os LTR-retroelementos têm desempenhado um papel significativo no processo evolutivo e contribuído com o aumento do tamanho dos genomas. A presença de retrotransposons foi detectada em todas as linhagens de plantas superiores e algas, mostrando que estes componentes são onipresentes nos genomas das plantas (Kubis et al., 1998).

3 MATERIAL E MÉTODOS

3.1 EXTRAÇÃO E QUANTIFICAÇÃO DO DNA GENÔMICO

O DNA genômico foi extraído a partir de gemas laterais frescas de clones da cultivar RB867515. A extração do DNA foi conduzida no Laboratório de Genética e Genômica de Plantas, na Escola de Agronomia da Universidade Federal de Goiás, utilizando o protocolo descrito por Aljanabi et al. (1999), com modificações. Diversos cuidados adicionais envolvendo o manuseio das amostras foram tomados, como o corte da ponta das ponteiras utilizadas e a não utilização de agitadores do tipo vórtex, com o intuito de preservar a integridade do material genético. O DNA foi ressuspendido em tampão com TE (Tris-EDTA pH=8,0) contendo RNAse.

Para a verificação da qualidade das extrações, as amostras foram submetidas à eletroforese em gel de agarose 1%. Para avaliação visual da concentração do DNA extraído foram utilizados dois padrões de peso molecular correspondentes a 50 e 100 ng/μL de DNA de fago λ. O gel foi corado com brometo de etídeo a 0,5 μg/mL por 15 minutos. A imagem das bandas no gel foi obtida em um transiluminador de luz UV e uma câmera escura acoplada a um sistema de captura de imagem digital.

Para sequenciamento em larga escala foram selecionadas somente as amostras de DNA de alto peso molecular, uma vez que a qualidade inicial do material genético reflete diretamente na qualidade das sequências finais. As amostras selecionadas foram quantificadas utilizando o fluorímetro Qubit®. Como medida de pureza foi utilizada a relação de absorbâncias a 260 nm e 280 nm, obtidas pelo espectrofotômetro Nanodrop®. Foram utilizadas somente amostras com valores para essa relação entre 1,8 e 2,0. Valores inferiores indicam a presença de proteínas, fenóis ou outros contaminantes que absorvem luz em comprimentos de onda próximos a 280 nm.

Para o sequenciamento em larga escala foram preparados aproximadamente 200 µg de DNA genômico de alto peso molecular, dispostos em microtubos de 1,5 µL, de acordo com as exigências da empresa de sequenciamento. As amostras foram enviadas para a empresa BGI Americas, onde foram construídas as bibliotecas genômicas e foi realizado o sequenciamento propriamente dito. O pipeline da obtenção dos *reads* até o *assembly* das sequências consenso da cultivar RB87515 estão ilustrados na Figura 11.

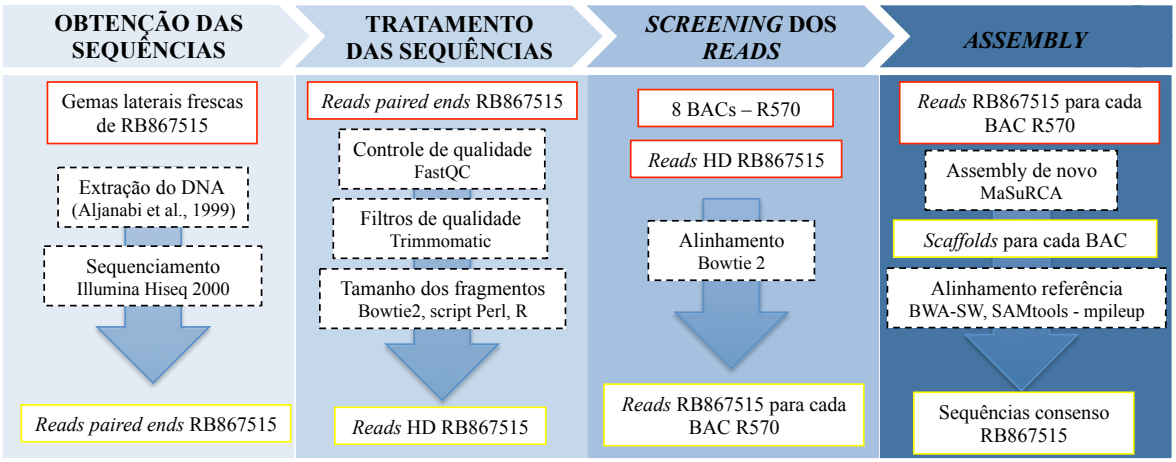


Figura 11. *Pipeline* das etapas da obtenção dos *reads paired ends*, tratamento das sequências, *screening* dos *reads* e *assembly* das sequências consenso. As caixas em vermelho representam os arquivos de entrada e em amarelo os arquivos de saída de cada etapa. As caixas com linhas tracejadas representam as etapas/análises realizadas juntamente com os softwares ou protocolos utilizados.

3.2 ESTRATÉGIA DE SEQUENCIAMENTO

A estratégia adotada foi a construção de quatro bibliotecas de DNA genômico da cultivar RB867515 com tamanhos de fragmentos distintos. Cada uma delas foi sequenciada duas vezes utilizando a plataforma de sequenciamento de nova geração HiSeq2000 da Illumina®, pelas estratégias de *paired ends*, em que foram sequenciados 100 pb de cada extremidade do fragmento. No total foram utilizados oito *lanes* de sequenciamento, com rendimento médio previsto de 37,5 Gb em cada *lane* (Tabela 2).

Tabela 2. Resumo das bibliotecas de DNA genômico construídas para o sequenciamento

Tamanho médio dos fragmentos	**Quantidade de bibliotecas**	***Lanes* de sequenciamento/biblioteca**	**Total**
170 pb	1	2	2
500 pb	2	2	4
800 pb	1	2	2
		Total Geral	8

3.3 TRATAMENTO DAS SEQUÊNCIAS

Para proceder às análises das sequências genômicas, os arquivos brutos recebidos pela empresa de sequenciamento foram depositados no servidor local, onde foram realizado todos os procedimentos de análise dos dados. Os *reads* (*forward* e *reverse*) obtidos de cada biblioteca, e de *lanes* diferentes da mesma biblioteca, encontram-se em arquivos compactados separados. Primeiramente as bibliotecas foram descompactadas e renomeadas de acordo com o tipo de biblioteca (*paired end*), tamanho do fragmento utilizado para a construção da biblioteca*, lane* de sequenciamento e tipo do *read* (*forward* ou *reverse*). Por exemplo: PE170_A_1.fq e PE170_A_2.fq representam, respectivamente, os *reads forward* e *reverse*, obtidos no *lane* A da biblioteca *paired end* de fragmentos de 170 pb.

3.3.1 Controle de qualidade

Inicialmente as sequências obtidas foram avaliadas em termos de qualidade pelo *software* FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc). Os resultados das análises de qualidade foram obtidos em formato HTML, com o resumo das análises em formato de gráficos e tabelas de modo a facilitar a tomada de decisões sobre os filtros de qualidade que foram aplicados a cada caso.

Dentre os resultados apresentados, alguns merecem uma atenção mais criteriosa, como a qualidade da sequência por base (*per base sequence quality*), conteúdo da sequência por base (*per base sequence content*) e sequências super-representadas (*overrepresented sequences*). O primeiro mostra uma visão geral da distribuição de valores de qualidade phred (Ewing & Green, 1998) em todas as bases em cada posição no arquivo formato “fastq”. O segundo mostra o conteúdo de cada base (A, T, C ou G) em cada posição dos *reads*. Regiões com proporções de cada base muito discrepantes (diferenças

entre A e T ou C e G maiores que 10%) foram removidas da análise.

As bibliotecas foram filtradas utilizando o *software* Trimmomatic (Bolger et al., 2014), que executa uma variedade de tarefas de controle de qualidade de *reads*. Inicialmente foram eliminadas as bases de baixa qualidade a partir das extremidades 5' e 3' de cada sequência, utilizando como critério de corte o valor de phred < 30. Depois as sequências foram cortadas cada vez que a qualidade média dentro de uma janela de quatro bases fosse inferior ao valor de phred 30. Pares de *reads* que passaram pelos filtros foram considerados de alta qualidade se tivessem no mínimo 50 pb (HQ). *Reads* "órfãos", em que somente um dos elementos do par permaneceu após os filtros de qualidade, foram combinados em um único arquivo e tratados como *single reads* (SR).

Uma vez realizados os tratamentos dos dados pelo Trimmomatic, a ferramenta FastQC foi utilizada novamente para fins de novo diagnóstico. Essa análise permitiu a identificação de regiões com distorções nos conteúdos de cada base. Foi necessário então realizar um novo corte no início de cada sequência de 10 a 11 bases, utilizando o comando HEADCROP do Trimmomatic. Para o corte do início das sequências, todos os comandos anteriores foram executados novamente nos arquivos brutos, adicionando essa nova etapa no final, mantendo-se a exigência de um tamanho mínimo de 50 bases por *read*.

3.3.2 Estimação do tamanho dos fragmentos

O tamanho dos fragmentos das bibliotecas genômicas da RB867515 foi estimado com base no alinhamento dos *reads* no genoma de sorgo (Paterson et al., 2009), utilizando o *software* Bowtie2 (Langmead & Salzberg, 2012). O arquivo com extensão "sam" criado pelo Bowtie2 foi utilizado como entrada para a execução de um script em Pearl, originalmente criado por Matthew Conte, com adaptações, de modo a se estimar o tamanho dos fragmentos correspondentes aos *reads pareado*s mapeados. A saída, sob a forma de um arquivo texto, foi posteriormente analisada através de um *script* em R para obtenção dos gráficos e estatísticas descritivas dos tamanhos de fragmentos de cada biblioteca.

3.4 *ASSEMBLY* DAS REGIÕES DE INTERESSE DO GENOMA DA CULTIVAR RB867515

3.4.1 Sequências dos BACs da cultivar R570

Para a obtenção das sequências de interesse do genoma da cultivar RB867515 foram utilizadas como referência oito sequências provenientes de BACs (*Bacterial Artificial Chromossomes*) da cultivar R570 publicadas por Garsmeur et al. (2011). Essas sequências pertencem a sete haplótipos do mesmo grupo de homeologia, e todas carregam o gene de resistência à ferrugem (*Bru1*) (Figura 12).

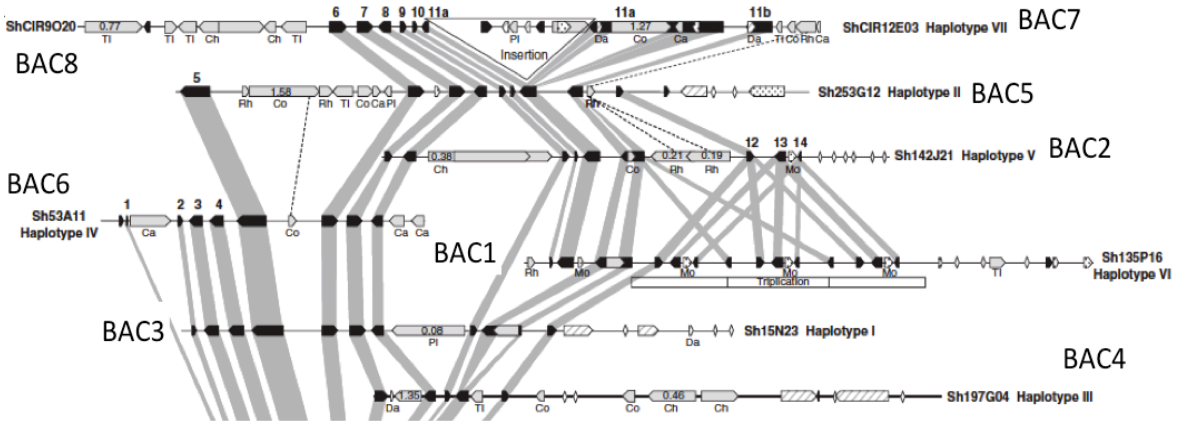


Figura 12. Representação esquemática dos BACs de R570 (Adaptado de Garsmeur et al., 2011)

As sequências foram renomeadas conforme apresentado na Tabela 3 e foram designadas como BACs RB867515, para as novas sequências consenso obtidas, e BACs R570, para as sequências de referência. Os tamanhos dos BACs de R570 variaram de ~82 (Sh53A11) a 161 kb (Sh253G12), e totalizaram quase 1 Mb do genoma da cultivar R570.

Tabela 3. Renomeações dos BACs R570 e tamanho das sequências de referência

RB867515	R570	Tamanho (pb)
1	Sh135P16	142.236
2	Sh142J21	128.444
3	Sh15N23	139.911
4	Sh197G04	143.745
5	Sh253G12	160.839
6	Sh53A11	82.414
7	ShCIR12E03	86.245
8	ShCIR9O20	88.987
	Total	**972.831**

3.4.2 *Screening* dos *reads* provenientes da cultivar RB867515

Para obtenção das sequências correspondentes aos BACs da cultivar R570 no genoma da cultivar RB867515, os *reads* das bibliotecas de DNA genômico da cultivar RB867515 foram alinhados às oito sequências de referência da cultivar R570 (Garsmeur et al., 2011), utilizando-se o *software* Bowtie2 (Langmead & Salzberg, 2012). Os pares de *reads* que alinharam em cada uma das sequências de referência foram agrupados em três bibliotecas (com tamanhos de fragmentos em média iguais a 170, 500 e 800 pb) para cada um dos oito BACs da cultivar R570.

3.4.3 Assembly das regiões correspondentes aos BACs

As bibliotecas correspondentes a cada um dos oito BACs foram submetidas ao *software* MaSuRCA 2.0.3.1 para realização do *assembly* de cada BAC individualmente. A ideia principal do MaSuRCA é reduzir a complexidade dos dados, transformando a alta cobertura obtida pela grande quantidade de *reads* curtos, gerados pela tecnologia Illumina, em menor cobertura de *super-reads*, com tamanho maior. Essa estratégia alia a eficiência da utilização de memória dos grafos *de Bruijn*, usados na construção dos *super-reads*, e a flexibilidade e robustez da estratégia de *assembly* por *Overlap-layout-consensus* (OLC). O *software* utiliza uma versão modificada do Celera Assembler, contida no MaSuRCA, que é responsável pela construção dos *contigs* e *scaffolds*. As duas consequências mais relevantes decorrentes da criação dos *super-reads* é que cada um dos *reads* originais está

contido nos *super-reads*, ou seja, nenhuma informação é perdida no processo; e muitos dos *reads* originais produzem o mesmo *super-read*, reduzindo a complexidade do conjunto de dados (Zimin et al., 2013).

3.4.4 Alinhamento com sequências de referência e obtenção da sequência consenso

Os *scaffolds* obtidos no *assembly* de cada BAC (RB867515) foram alinhados às sequências de referência correspondentes (R570) utilizando o *software* BWA/SW (Li & Durbin, 2010), que é uma versão do programa BWA para alinhamento de sequências longas. O alinhamento gera um arquivo com extensão "sam" que contém as coordenadas e informações do alinhamento das sequências. Esse arquivo foi convertido para o formato binário "bam", e dele foi criado um arquivo ordenado "sorted" e um arquivo de índices com formato "bai" pelo SAMtools. As sequências-consenso das regiões genômicas da cultivar RB867515 correspondentes a cada BAC da cultivar R570 foram obtidas pela função *mpileup* do SAMtools com o respectivo arquivo de alinhamentos ordenados (Li et al., 2009).

3.5 ANOTAÇÃO E ANÁLISE GENÔMICA

3.5.1 Identificação das regiões repetitivas

Para quantificação da riqueza e distribuição de elementos repetitivos ao longo das sequências, foi utilizada a versão online do programa *RepeatMasker* (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker). O programa faz uma procura por elementos repetitivos em sequências de DNA no formato "fasta" contra o banco de dados Repbase e fornece as sequências com esses elementos mascarados além de uma tabela com a anotação e classificação desses elementos. O banco de dados escolhido para a identificação dos elementos repetitivos em cana-de-açúcar foi o da subfamília Panicoideae, que inclui elementos de milho, sorgo, cana-de-açúcar e milheto. Os elementos foram classificados em DNA transposons, retrotransposons e *helitrons*. Para os dois primeiros, a classificação foi até o nível de famílias. Foram considerados tanto o número quanto o

tamanho dos elementos identificados. A Figura 13 ilustra o pipeline da anotação das sequências da cultivar RB867515 e a comparação em relação a cultivar R570.

Os microssatélites foram identificados pela versão online do *software* IMEX (http://imex.cdfd.org.in/IMEX/imex_basic.html). Optou-se pela procura de microssatélites perfeitos com motivos de mono, di, tri, tetra, penta e hexanucletídeos, com, no mínimo 5 repetições. Os resultados foram obtidos para cada BAC separadamente.

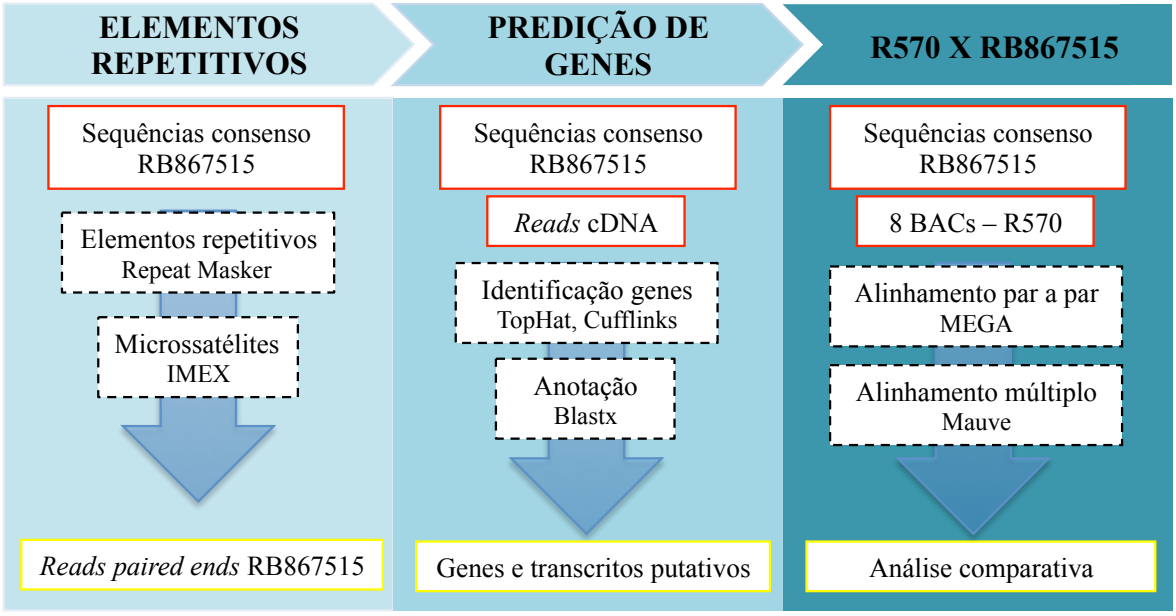


Figura 13. *Pipeline* das etapas da anotação das sequências consenso e análise comparativa das duas cultivares (RB867515 X R570). As caixas em vermelho representam os arquivos de entrada e em amarelo os arquivos de saída de cada etapa. As caixas com linhas tracejadas representam as etapas/análises realizadas juntamente com os *softwares* utilizados.

3.5.2 Predição de genes

As regiões gênicas foram identificadas utilizando sequências de cDNA obtidas pela extração do mRNA total de diversos tecidos vegetais de genótipos de cana-de-açúcar. Foram coletados 30 indivíduos no total, provenientes de uma população de melhoramento constituída por 48 genótipos selecionados e em fase final de avaliação do programa de melhoramento genético para a cultura da RIDESA, e representam a mesma base genética da cultivar RB867515. De cada um dos trinta indivíduos foram coletados cinco tecidos vegetais: colmo, gemas laterais, plântulas, folhas e gemas apicais.

O RNA total foi extraído em *bulk*, formado pelos 30 genótipos, utilizando kits de extração de RNA (TruSeq RNA Library Prep da Illumina). O sequenciamento dos cDNAs foi realizado pela empresa BGI Americas utilizando a plataforma HiSeq2000 da Illumina. Os *reads* das bibliotecas de cDNA de cada tecido vegetal foram submetidos a filtros de qualidade, com os mesmos critérios aplicados às bibliotecas de DNA genômico.

Os *reads* de RNA foram mapeados em cada uma das sequências-consenso dos BACs de RB867515 para a identificação dos genes. Para isso foi utilizado um protocolo descrito por Trapnell et al. (2012), em que se utiliza o TopHat (http://tophat.cbcb.umd.edu/), para o alinhamento dos *reads* e identificação dos sítios de *splicing*, e o Cufflinks (http://cufflinks.cbcb.umd.edu/) para montagem dos *reads* em transcritos.

O arquivo com extensão "gtf" criado, que contém as coordenadas dos éxons identificados, em relação a sequência de referência fornecida, foi utilizado para anotar os genes de cada BAC de RB867515. Para a identificação das funções putativas das proteínas, as sequências dos CDS (*Coding DNA Sequences*) foram comparadas com o banco de sequências de proteínas não redundantes no NCBI através de Blastx (Altschul et al., 1990).

3.6 ANÁLISE COMPARATIVA DOS BACs DA CULTIVAR R570 E RB867515

Para realizar a análise comparativa das regiões genômicas correspondentes aos BACs da cultivar R570 e aquelas da cultivar RB867515 foi utilizado o *software* MEGA 5.2 (Tamura et al., 2011), que é uma ferramenta integrada de alinhamento de sequências, estimação de tempos de divergência, taxas de evolução molecular, entre outras atribuições. As 16 sequências correspondentes aos BACs, oito da cultivar R570 e oito da cultivar RB867515, foram alinhadas par a par, utilizando-se o algoritmo ClustalW (Larkin et al., 2007). O programa Mauve (Darling et al., 2004) foi utilizado para construção do alinhamento múltiplo das sequências correspondentes aos 16 segmentos e visualização das regiões conservadas entre elas.

4 RESULTADOS E DISCUSSÃO

O sequenciamento das bibliotecas de DNA genômico da cultivar RB867515 produziu um total de 278,68 Gb com *reads* pa*read*os de 100 pb cada (Tabela 4). O rendimento médio foi menor do que previsto, considerando um rendimento esperado para cada *lane* de 37,5 Gb (37,5 Gb * 8 lanes de sequenciamento = 300 Gb). Essa diferença foi consequência dos rendimentos mais baixos das bibliotecas com fragmentos de 800pb.

Tabela 4. Rendimento das bibliotecas de DNA genômico da cultivar RB867515 antes e depois da aplicação dos filtros de qualidade pelo *software* Trimmomatic

	Bibliotecas[1]		**Número de *reads* (milhões)**	**Total de bases (Gb)**	**Tamanho médio dos *reads***	**Conteúdo GC (%)**	**Cobertura (/10 Gb)**
Sequências brutas	170		764,40	76,44	100	44,00	7,60
	500		1491,80	149,18	100	43,00	14,90
	800		530,60	53,06	100	42,00	5,30
		Total	**2786,80**	**278,68**	**100**	**43,00**	**27,80**
Sequências filtradas	170	HQ[2]	473,40	39,90	84	42,00	4,00
		SR[3]	72,70	5,50	76	46,00	0,60
	500	HQ	900,60	75,10	83	41,00	7,50
		SR	150,80	11,40	76	42,00	1,10
	800	HQ	336,80	27,80	83	41,00	2,80
		SR	56,50	4,30	77	42,00	0,40
		Total	**1990,80**	**164,00**	**80**	**42,33**	**16,40**

[1] Referente ao tamanho médio dos fragmentos de sequenciamento contratados; [2] Bibliotecas consideradas *High quality* (HQ), que passaram pelos filtros de qualidade e mantiveram de *reads* pareados; [3] Bibliotecas consideradas *Single reads* (SR) que passaram pelos filtros de qualidade, mas não mantiveram os *reads* pareados.

4.1 ANÁLISE DE CONTROLE DE QUALIDADE

Foram aplicados filtros de controle de qualidade nas sequências brutas, que eliminaram as bases com qualidade phred inferior a 30 no início e no final do *reads*. Além

de um corte realizado no início de cada *read* que variou de 10 a 12 pb, com a finalidade de eliminar fragmentos com conteúdo GC discrepante.

Após a análise dos *reads* filtrados no FastQC, pode-se notar que todas as bases restantes possuem alta qualidade (Figura 14). Em relação ao conteúdo das bases em cada posição no *read*, depois do corte realizado, o conteúdo GC ficou mais homogêneo ao longo dos *reads* (Figura 15). Essa distorção no conteúdo GC no início das sequências pode ser um viés resultante da presença de fragmentos de adaptadores usados nas etapas de sequenciamento.

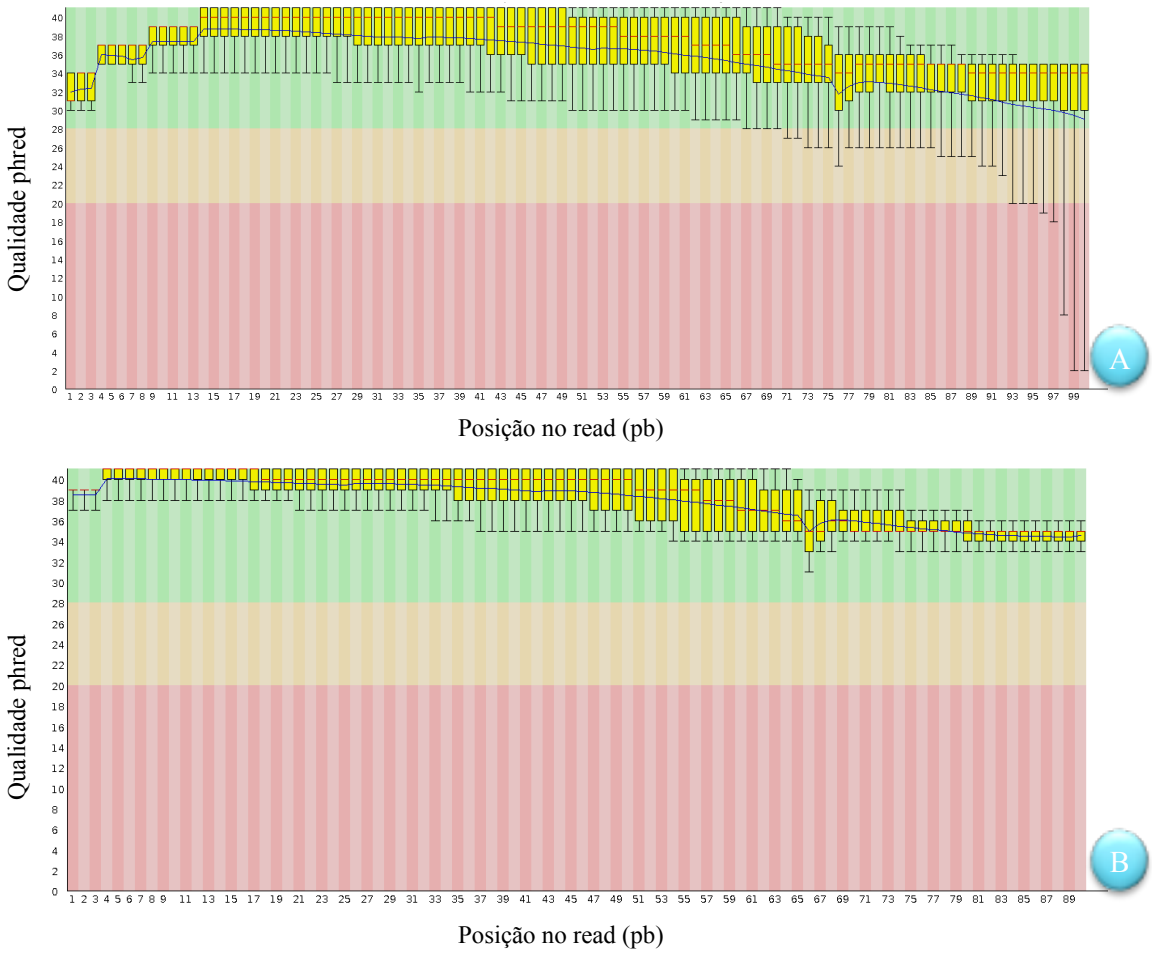


Figura 14. Qualidade de cada base dos *reads* da biblioteca 170_A de DNA genômico da cultivar RB867515, atribuída por valores phred antes (A) e depois dos filtros de qualidade (B). O eixo x representa a posição no read e o eixo y representa os valores de qualidade phred.

Houve uma redução de aproximadamente 29% no número de *reads*, e 41% do total de bases em relação aos arquivos originais. Além disso, houve também uma diminuição de 20 pb, em média, no tamanho dos *reads*. As bibliotecas ficaram divididas

em HQ (*High quality*), que possuem os *reads* de alta qualidade e que permaneceram pareados após os filtros, e SR (*Single reads*), que possuem *reads* de alta qualidade mas que ficaram despareados (o seu par não passou nos critérios do filtro) (Tabela 4).

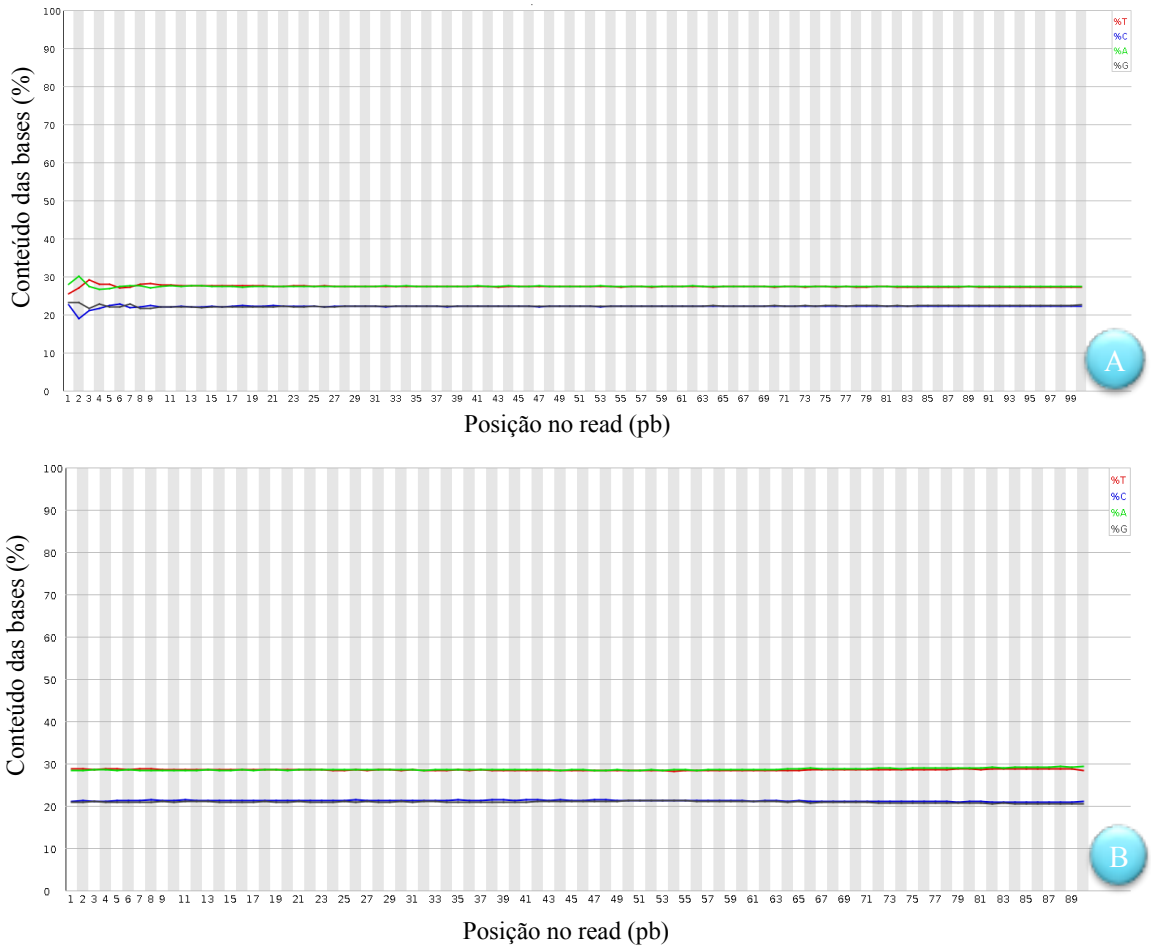


Figura 15. Conteúdo de cada base nos *reads* da biblioteca 170_A de DNA genômico da cultivar RB867515 antes (A) e depois dos filtros de qualidade (B). As linhas na cor vermelho, azul, verde e preto representam, respectivamente, as bases nitrogenadas timina, citosina, adenina e guanina. O eixo x representa a posição no read e o eixo y representa o conteúdo das bases em porcentagem.

Um outro critério analisado pelo software FastQC são as sequências super-representadas no conjunto de dados, que também podem ser originadas de contaminações por fragmentos de adaptadores. Nenhuma das bibliotecas apresentou sequências super-representadas, apontando estarem livres de contaminação remanescente dessa natureza.

4.2 ESTIMATIVA DO TAMANHO DOS FRAGMENTOS

Para a realização do *assembly* é necessário estimar o tamanho dos fragmentos das bibliotecas *paired ends*, e os respectivos desvios padrão. Essa estimativa foi obtida pelo alinhamento das sequências de cana-de-açúcar, de cada biblioteca separadamente, com o genoma de sorgo. As distâncias calculadas entre os *reads* pareados que alinharam concordantemente representam o tamanho dos fragmentos. De posse desses valores foi possível obter estimativas para as médias e desvios padrões dos tamanhos dos fragmentos em cada biblioteca (Tabela 5) e construir histogramas da distribuição desses valores (Figura 16).

Tabela 5. Estimativas de médias e desvios padrões do tamanho dos fragmentos das bibliotecas *paired ends* (PE) de DNA genômico da cultivar de cana-de-açúcar RB867515

Bibliotecas	Média	Desvio padrão
PE170_A	141	10
PE170_B	139	10
PE500_A	448	20
PE500_B	448	20
PE500_C	449	20
PE500_D	449	20
PE800_A	712	30
PE800_B	712	29

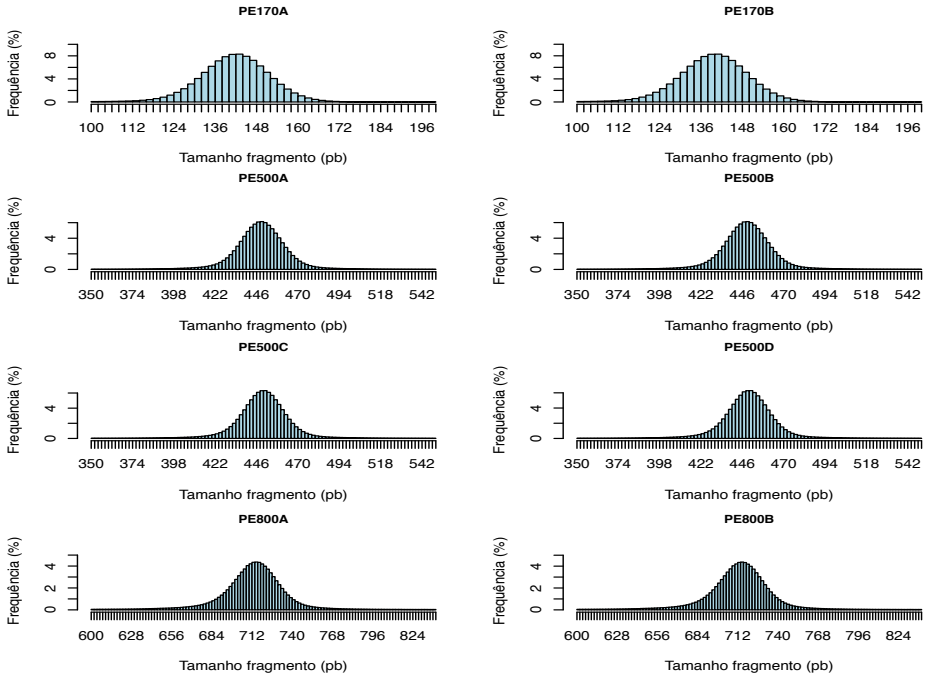


Figura 16. Distribuição do tamanho dos fragmentos de cada biblioteca de DNA genômico da cultivar de cana-de-açúcar RB867515.

Pode-se observar uma distribuição aproximadamente normal dos tamanhos dos fragmentos em cada biblioteca (Figura 16).

4.3 OBTENÇÃO DAS SEQUÊNCIAS-CONSENSO DE INTERESSE NO GENOMA DA CULTIVAR RB867515

4.3.1 Alinhamento dos *reads* de DNA genômico da cultivar RB867515 nas sequências de referência dos BACs da cultivar R570

As oito sequências dos BACs da cultivar R570 foram obtidas no banco de dados do NCBI (http://www.ncbi.nlm.nih.gov). Foi realizado o alinhamento dos *reads* gerados a partir do DNA genômico da cultivar RB867515 em cada um dos BACs da R570. Os *reads* positivos (alinhados) para cada uma das sequências de referência foram utilizados para a montagem das sequências-consenso.

Como a estratégia de *assembly* leva em consideração o tamanho dos fragmentos que gerou os *reads* no processo de sequenciamento, as bibliotecas continuaram divididas pelo tamanho de seus fragmentos. Os *reads* de diferentes bibliotecas com mesmo tamanho médio de fragmentos foram agrupados em uma única biblioteca. Para cada um dos BACs foram utilizadas três bibliotecas de *reads* pareados (PE – *paired ends*), com tamanhos médios de fragmentos de 140, 448 e 712 pb (Tabela 5).

A existência de múltiplas bibliotecas *paired ends* com diferentes tamanhos de fragmentos é muito importante em projetos de sequenciamento de genomas. A criação de bibliotecas de longo alcance, ou seja, que a distância entre os dois *reads* (*foward* e *reverse*) seja grande, é muito útil na montagem dos genomas, pois permite fazer a conexão entre dois *contigs* não ligados anteriormente (Salzberg et al., 2011).

A Tabela 6 apresenta o número de *reads* utilizados no *assembly* da região correspondente a cada BAC. A região correspondente ao BAC 3 foi a que apresentou o menor número de *reads* pareados alinhados, pouco mais de 4 milhões de *reads*. A região correspondente ao BAC 5 apresentou o maior número, com quase 44 milhões de *reads*.

4.3.2 Montagem dos BACs usando MaSuRCA

As bibliotecas de DNA genômico com os *reads* positivos da cultivar RB867515, para cada um dos BACs, foram montadas separadamente, utilizando o software de montagem de genomas *de novo* MaSuRCA 2.0.3 (Zimin et al., 2013). O arquivo de *scaffolds* resultante foi submetido ao script assemblathon_stats.pl, que calcula um conjunto de estatísticas para genomas montados (Bradnam, 2011).

Tabela 6. Número de *reads* das bibliotecas de DNA genômico da cultivar RB867515 que alinharam nos BACs da cultivar R570 e foram utilizados para a montagem das sequências consenso

	BACs	**Tamanho (pb)**	**Bibliotecas**	**Número de *reads* PE**
1	Sh135P16	142.236	PE170	4.014.958
			PE500	7.864.571
			PE800	2.548.197
			Total	14.427.726
2	Sh142J21	128.444	PE170	8.223.712
			PE500	17.166.288
			PE800	5.542.298
			Total	30.932.298
3	Sh15N23	139.911	PE170	1.003.752
			PE500	2.310.609
			PE800	843.327
			Total	4.157.688
4	Sh197G04	143.745	PE170	3.677.973
			PE500	7.734.263
			PE800	2.599.428
			Total	14.011.664
5	Sh253G12	160.839	PE170	12.087.515
			PE500	24.175.902
			PE800	7.728.022
			Total	43.991.439
6	Sh53A11	82.414	PE170	6.833.167
			PE500	13.790.136
			PE800	4.323.154
			Total	24.946.457
7	ShCIR12E03	86.245	PE170	8.026.052
			PE500	16.395.751
			PE800	5.315.143
			Total	29.736.946
8	ShCIR9O20	88.987	PE170	7.362.153
			PE500	7.864.571
			PE800	5.606.887
			Total	20.833.611

Em relação aos resultados dos *assemblies* das regiões correspondentes aos oito BACs no genoma da cultivar RB867515, foram obtidos 607 *scaffolds* maiores que 1 kb

para o BAC 1 e 2.884 para o BAC 7, o que corresponde a, respectivamente, 35,5% e 51,1% do total de *scaffolds* gerados. O grande número de *scaffolds* reflete a dificuldade em montar genomas complexos como o da cana-de-açúcar, mesmo em uma análise concentrada em uma microrregião como um BAC. Ainda assim, foram obtidas sequências contíguas de até 21 kb de tamanho (maior *scaffold*), que correspondentes a 25% do tamanho total do BAC 6 (Tabela 7).

Os valores de N50 e L50 são métricas comuns utilizadas para avaliar a qualidade de *assemblies*. O N50 é a mediana ponderada dos comprimentos dos *scaffolds* (ou *contigs*), ou seja, é o tamanho do menor *scaffold* que faz com que a soma de comprimentos de sequências do mesmo tamanho ou maiores que ele resulte na metade do tamanho total do *assembly*. O L50 corresponde à contagem dessas sequências maiores do que o N50 (Salzberg et al., 2011). O maior N50 foi observado para o BAC 7 (1764), e o menor foi para o BAC 8 (1062) (Tabela 7).

Tabela 7. Estatísticas descritivas da montagem de oito regiões genômicas da cultivar RB867515 correspondentes a oito BACs da cultivar R570

BACs	Número de *scaffolds**	% Total *scaffolds***	Maior *scaffold* (pb)	N50	L50
1	607	35,5	15.240	1.228	413
2	1.885	46,4	12.786	1.740	872
3	817	42,6	10.982	1.537	422
4	894	37,0	10.072	1.506	516
5	2.119	40,6	9.713	1.377	1.263
6	962	41,2	21.026	1.454	533
7	2.884	51,1	11.648	1.764	1.249
8	684	22,0	11.329	1.062	617

* *Scaffolds* > 1 kb; ** Porcentagem dos scaffolds > 1 kb em relação ao número total.

O valor de N50 fornece uma noção de escala e potencial contiguidade do *assembly*. Teoricamente, quanto maior o seu valor e menor o L50 melhor a qualidade do *assembly* gerado. Entretanto, estes valores isoladamente não fornecem informações sobre a cobertura ou a verdadeira acurácia dos *assemblies* (Salzberg et al., 2011).

4.3.3 Obtenção das sequências-consenso

Devido ao grande número de *scaffolds* obtidos para cada BAC no *assembly*, optou-se pelo alinhamento desses com suas respectivas sequências de referência para fins de obtenção de uma sequência consenso correspondente a cada BAC. A sequência

consenso foi obtida pela opção *mpileup* do SAMtools utilizando os arquivos gerados no alinhamento produzido pelo BWA-SW.

Sobre as sequências consenso foi possível estimar algumas estatísticas (Tabela 8), como a porcentagem da sequência de referência recuperada no *assembly*; a porcentagem de N (%N), regiões não montadas na sequência consenso e que foram preenchidas com a letra N; porcentagem de "letras" que não estão incluídas em ATCGN (% não-ATCGN), refere-se a polimorfismos entre os *scaffolds* que o script não conseguiu solucionar na sequência consenso, e que são representados por um alfabeto especial (Tabela 9); a porcentagem do BAC montado equivale à taxa efetiva de recuperação, ou seja, a porcentagem de recuperação da referência subtraída da porcentagem de N's.

Tabela 8. Estatísticas descritivas das sequências consenso dos oito BACs da cultivar RB867515 obtido pelo alinhamento dos scaffolds de cada um dos BACs nas respectivas sequências da cultivar R570 usada como referência

	BACs	Tamanho (pb)	% referência	% N	% não-ATCGN*	% BAC montado	Tamanho consenso (pb)
1	Sh135P16	142.236	100,0	18,3	2,2	81,7	142.236
2	Sh142J21	128.444	98,5	3,4	2,4	95,1	126.547
3	Sh15N23	139.911	98,5	6,4	3,1	92,2	137.851
4	Sh197G04	143.745	98,5	7,1	5,4	91,4	141.630
5	Sh253G12	160.839	98,2	2,8	5,3	95,4	157.913
6	Sh53A11	82.414	98,5	1,2	4,3	97,3	81.164
7	ShCIR12E03	86.245	98,5	4,4	8,4	94,1	84.968
8	ShCIR9O20	88.987	97,9	5,2	5,0	92,7	87.153
	Total	**972.831**			**Média**	**92,5**	**959.462**

* Refere-se a polimorfismos entre as sequências das duas cultivares

O tamanho dos BACs montados em relação a referência variou de 97,9% para o BAC 8 até 100% para o BAC 1. Essas taxas não revelam o tamanho real da montagem dos BACs, pois a sequência com o maior valor também foi a que apresentou a maior proporção de bases não identificadas, preenchidas com N's. O BAC 6 é o menor BAC entre os oito e apresentou a menor proporção de N's (Tabela 8), além de apresentar o maior *scaffold* contíguo (Tabela 7).

Tabela 9. Códigos IUPAC (*International Union of Pure and Applied Chemistry*) para nucleotídeos ambíguos

Símbolo	Significado
M	A ou C
R	A ou G
W	A ou T
S	C ou G
Y	C ou T
K	G ou T
V	A ou C ou G
H	A ou C ou T
D	A ou G ou T
B	C ou G ou T
N	qualquer base

Fonte: http://www.bioinformatics.org/sms/iupac.html

Para a taxa de recuperação dos BACs em relação às sequências de referência da cultivar R570, obteve-se valores variando de 82% para o BAC 1 à 97% para o BAC 6 (Tabela 8). Esses valores são reflexo da proporção de bases não identificadas.

As sequências apresentaram um tamanho final que variou de 81 kb para o BAC 6 à 158 kb para o BAC 5. O somatório dos tamanhos totais das oito sequências consenso obtidas para a cultivar RB867515 representou um amostragem de quase 1 Mb do genoma dessa cultivar. Considerando o tamanho do genoma total de cana-de-açúcar como sendo 10 Gb, tem-se uma representação de 1% do genoma, contida em oito sequências com tamanho médio de 120 kb.

4.4 ANOTAÇÃO

4.4.1 Regiões repetitivas

Foram identificadas 5.145 sequências repetitivas nos BACs da cultivar RB867515, dos quais 4.662 são sequências microssatélites, 477 são elementos genéticos móveis (326 retrotransposons + 134 transposons de DNA + 17 *Helitrons*) e seis regiões de baixa complexidade (Tabela 10).

Tabela 10. Número de elementos genéticos repetitivos encontrados nas oito sequências de RB867515 correspondentes aos BACs da cultivar R570

			BAC1	BAC2	BAC3	BAC4	BAC5	BAC6	BAC7	BAC8	TOTAL
Retrotransposons	LTR	*Gypsy-Ty3*	9	24	19	21	18	3	14	46	154
		Copia-Ty1	4	25	8	9	16	32	19	21	134
	LINE	*L1*	3	1	9	10	5	0	3	0	31
		RTE-BovB	0	0	1	2	0	0	2	0	5
	SINE/*tRNA*		0	0	1	0	1	0	0	0	2
	TOTAL		16	50	38	42	40	35	38	67	326
DNA transposons	*TcMar-Stowaway*		3	3	2	2	5	2	1	1	19
	CMC-EnSpm		33	10	8	1	7	0	3	0	62
	En-Spm		5	0	0	0	3	0	0	0	8
	hAT-Ac		4	2	1	0	2	0	1	0	10
	hAT-Tag1		0	0	1	2	0	0	0	0	3
	hAT-Tip100		0	0	0	0	1	0	1	0	2
	PIF-Harbinger		1	2	1	0	8	7	3	5	27
	MULE-MuDR		2	1	0	0	0	0	0	0	3
	TOTAL		48	18	13	5	26	9	9	6	134
RC/Helitron			5	1	2	0	3	2	3	1	17
Baixa complexidade			3	0	1	0	1	1	0	0	6
Microssatélites			636	581	720	678	865	402	420	360	4,662
TOTAL GERAL			708	650	774	725	935	449	470	434	5.145

Do total de nucleotídeos sequenciados, 225 kb compreendem sequências repetitivas, o que equivale a 24% da amostra do genoma da cultivar RB867515. Esse valor está distribuído em 147 kb de retrotransposons, 50 kb de transposons de DNA (incluindo os *Helitrons*), 27 kb de microssatélites e 336 pb de regiões de baixa complexidade (Tabela 11).

Os 477 elementos genéticos móveis encontrados correspondem a 197 kb do genoma da cultivar que foi montado. Em termos de proporção, tem-se que pouco mais de 20% das sequências são constituídas por elementos genéticos móveis. Levando em consideração as sequências repetitivas como um todo, os retrotransposons representam quase 66% dessas sequências, enquanto os transposons de DNA (exceto *Helitrons*) compõe aproximadamente 19%, totalizando 85% destas regiões.

O número de retrotransposons (326) encontrado foi maior do que o de transposons de DNA (134) no genoma da cultivar RB867515 (Tabela 10). Em relação ao tamanho total dos elementos, os retrotransposons somaram mais de 147 kb (15% da composição do genoma avaliado), enquanto os transposons de DNA contribuem com aproximadamente 43 kb (~4%).

Tabela 11. Tamanho total dos elementos genéticos repetitivos em pares de bases encontrados nas oito sequências de RB867515 correspondentes aos BACs de R570

			BAC1	BAC2	BAC3	BAC4	BAC5	BAC6	BAC7	BAC8	TOTAL
Retrotransposons	LTR	Gypsy-Ty3	4.610	14.272	3.955	8.804	9.131	313	9.100	14.215	64.400
		Copia-Ty1	5.193	10.692	2.489	6.507	6.931	13.460	10.006	5.946	61.224
	LINE	L1	1.263	1.206	3.192	5.774	3.887	0	1.774	0	17.096
		RTE-BovB	0	0	774	628	0	0	3.618	0	5.020
	SINE/tRNA		0	0	45	0	45	0	0	0	90
	TOTAL		11.066	26.170	10.455	21.713	19.994	13.773	24.298	20.161	147.830
DNA transposons	TcMar-Stowaway		585	608	422	353	1.013	472	236	236	3.925
	CMC-EnSpm		16.331	5.487	1.784	274	2.015	0	1.291	0	27.182
	En-Spm		1.353	0	0	0	2.471	0	0	0	3.824
	hAT-Ac		489	328	115	0	335	0	77	0	1.344
	hAT-Tag1		0	0	48	325	0	0	0	0	373
	hAT-Tip100		0	0	0	0	60	0	2.193	0	2.253
	PIF-Harbinger		130	251	242	0	1.069	876	354	784	3.706
	MULE-MuDR		155	104	0	0	0	0	0	0	259
	TOTAL		19.043	6.778	2.611	952	6.963	1.348	4.151	1.020	42.866
RC/Helitron			907	82	271	0	414	290	4.523	265	6.752
Baixa complexidade			215	0	51	0	26	44	0	0	336
Microssatélites			4.035	3.483	4.383	3.904	5.109	2.420	2.460	2.133	27.927
TOTAL GERAL			35.266	36.513	17.771	26.569	32.506	17.875	35.432	23.579	225.511

Uma procura por elementos transponíveis das bibliotecas de ESTs de cana-de-açúcar do programa SUCEST revelou uma diferença pequena entre a proporção de transposons de DNA e retrotransposons, 54% e 46%, respectivamente (Rossi et al, 2001). O que pode ser decorrente do fato das regiões analisadas no SUCEST serem transcricionalmente ativas, de eucromatina, e os retrotransposons são preferencialmente encontrados em regiões de heterocromatina, por isso os números encontrados por estes autores para as duas classes de transposons não terem sido tão discrepantes.

Os retrotransposons podem ser classificados em LTR e não-LTR. Os representantes do primeiro grupo possuem repetições terminais longas e o segundo é caracterizado pela ausência dessas estruturas em suas regiões terminais. Entre as sequências analisadas, a maioria foi composta por retrotransposons LTR, com 288 elementos que correspondem à 87 % dos retrotransposons. Os retrotransposons não-LTR totalizaram 38 elementos, e contribuíram com 13 % do total.

Retrotransposons da ordem LTR estão representados por duas superfamílias na amostra do genoma de cana-de-açúcar, *Gypsy-Ty3* e *Copia-Ty1*, que diferem entre si pela posição do gene que codifica a enzima integrase. Foram observados uma proporção de 1,15 elementos *Gypsy-Ty3* para *Copia-Ty1*. Em relação ao tamanho dos elementos eles representam, respectivamente, 53,5% e 46,5% do total de retrotransposons LTR. Os

retrotransposons LTR de BACs da cultivar R570 foram analisados por Domingues et al. (2012). Foram identificados 60 sequências completas desses elementos, na proporção de 0,88 elementos *Gypsy-Ty3* (45,6%) para *Copia-Ty1* (53,4%).

Um cenário diferente foi encontrado para os retrotransposons do SUCEST descritos por Rossi et al. (2001), a proporção de elementos da superfamília *Copia-Ty1* (64,1%) foi maior que *Gypsy-Ty3* (14,8%), e ainda uma grande proporção de elementos LTR não foram classificados (21,10%). A superfamília *Copia-Ty1* mostrou-se predominante nas análises de retrotransposons para cana-de-açúcar, diferente do encontrado no presente trabalho, em que a superfamília *Gypsy-Ty3* foi mais expressiva. Novamente, o fato de que os dados do SUCEST se referem a sequências expressas, e não a sequências genômicas como os do presente trabalho, pode explicar esta diferença.

Os elementos não-LTR são menos numerosos no conjunto de retrotransposons. As ordens observadas foram LINE (*Long Interspersed Nuclear Elements*) com 36 elementos (94,7%) e SINE (*Short Interspersed Nuclear Elements*) com 2 representantes (5,3%) (Tabela 9). Nenhum retrotransposon não-LTR foi identificado nas bibliotecas de EST do SUCEST. A ausência de retrotransposons SINE pode ser atribuída ao fato de que esses elementos são não-autônomos, caracterizados por não produzirem enzimas para sua transposição.

Em relação ao tamanho dos elementos, a superfamília *Gypsy-Ty3* foi predominante entre os retroelementos, com 64 kb, representando 43,6% do total de retrotransposons. A superfamília *Copia-Ty1* foi a segunda mais abundante, contribuindo com 62 kb da região analisada do genoma de cana-de-açúcar e 41,4% da totalidade dos retrotransposons. Os elementos não-LTR somaram 15% dos retrotransposons. Das ordens representadas, LINE totalizou 22 kb (14,9%) e SINE foi representado apenas por dois elementos que somaram 90 pb (0,1%) (Tabela 11).

O BAC 7 apresentou a maior proporção de retrotransposons, 29%, que representa 24 kb dos seus 85 kb de comprimento, em que 22% são retrotransposons do tipo LTR e 6% são elementos não-LTR. O BAC com menor proporção de retroelementos foi o 3, com 10 kb (8%). Destes valores, 5% são representados por retrotransposons LTR e 3% por não-LTR (Tabela 12).

Tabela 12. Proporção de elementos genéticos repetitivos em relação ao tamanho dos BACs montados

		Retrotransposons						DNA transposons		Outros*		TOTAL GERAL	
BAC	Tamanho	LTR		Não-LTR		Total							
1	142.236	9.803	7%	1.263	1%	11.066	8%	19.043	13%	5.157	4%	35.266	25%
2	126.547	24.964	20%	1.206	1%	26.170	21%	6.778	5%	3.565	3%	36.513	29%
3	137.851	6.444	5%	4.011	3%	10.455	8%	2.611	2%	4.705	3%	17.771	13%
4	141.630	15.311	11%	6.402	5%	21.713	15%	952	1%	3.904	3%	26.569	19%
5	157.913	16.062	10%	3.932	2%	19.994	13%	6.963	4%	5.549	4%	32.506	21%
6	81.164	13.773	17%	0	0%	13.773	17%	1.348	2%	2.754	3%	17.875	22%
7	84.968	18.906	22%	5.392	6%	24.298	29%	4.151	5%	6.983	8%	35.432	42%
8	87.153	20.161	23%	0	0%	20.161	23%	1.020	1%	2.398	3%	23.579	27%
TOTAL	959.462	125.424	13%	22.206	2%	147.630	15%	42.866	4%	35.015	4%	225.511	24%

* Microssatélites, regiões de baixa complexidade e *helitrons*.

Retrotransposons LTR apresentaram um tamanho médio de 436 pb, sendo que representantes da superfamília *Gypsy-Ty3* tiveram tamanho médio de 458 pb e os representantes da família *Copia-Ty1*, 418 pb. Já os elementos não-LTR apresentaram uma maior discrepância em relação aos seus tamanhos médios, LINEs com 618 pb e SINEs com 45 pb. Esses valores são decorrentes das características intrínsecas de cada elemento. Elementos SINE são não-autônomos e possuem tamanho que varia de 80 a 500 pb, bastante reduzidos em relação aos LINE que possuem ORFs para codificar as enzimas responsáveis pela sua transposição, e que podem chegar até 6 kb de tamanho (Wicker et al., 2007).

Os 134 transposons de DNA encontrados na amostra do genoma de cana-de-açúcar estão distribuídos em 5 superfamílias: *TcMar-Stowaway*, *CMC (En-Spm)*, *hAT* (*Ac*, *Tag1* e *Tip100*), *PIF-Harbinger* e *Mule-MUDR*. A superfamília mais numerosa foi *CMC* com 70 elementos e a menos numerosa foi *Mule-MUDR* (Mutator) com apenas 3 elementos. O BAC 1 foi o mais numeroso em termos de transposons de DNA, e o BAC 4 o menos numeroso, com, respectivamente, 48 (42,3%) e 5 elementos (3,7%). A superfamília *hAT* está representada por elementos de três famílias, *Ac* (10), *Tag1* (3) e *Tip100* (2) (Tabela 10).

Estes elementos estão distribuídos por todos os BACs, mas algumas superfamílias não estão representadas em alguns BACs, como é o caso do elemento mais numeroso, *CMC-EnSpm*, que não foi identificado nos BACs 6 e 8. Em contrapartida, o elemento *TcMar-Stowaway* está distribuído em todos os BACs, variando de 1 a 5 elementos por sequência (Tabela 10).

No total, os transposons de DNA somaram aproximadamente 43 kb, que equivale a 4% do tamanho total das sequências montadas para os BACs de cana-de-açúcar. A superfamília mais numerosa (*CMC-EnSpm*) também apresentou o maior tamanho total, com pouco mais de 31 kb, representando 72% dos elementos dessa mesma classe. Já a superfamília que apresentou o menor tamanho total foi *MULE-MuDR*, com 259 pb (0,6%) (Tabela 11). Os elementos apresentaram tamanho médio de 320 pb, com os representante da família *hAT-Tip100* sendo os maiores, com 1127 pb em média, e os menores foram os elementos *MULE-MuDR* com tamanho médio de 86 pb.

Os BACs 1 e 4, foram os que apresentaram o maior e o menor número de elementos transponíveis de DNA, respectivamente. O primeiro apresentou 19 kb de transposons de DNA, 13% do seu conteúdo de 142 kb. Já para o BAC 4 foi observado apenas 952 pb desses elementos, menos de 1% do seu tamanho total de 141 kb (Tabela 12).

Com as iniciativas de sequenciamento de genomas de várias espécies de plantas, principalmente as pertencentes à família das gramíneas, alguns estudos em larga escala sobre a composição do genoma repetitivo dessas espécies já foram realizados. Uma das primeiras espécies de gramíneas a ter o seu genoma publicado foi o arroz (International Rice Genome Sequencing Project, 2005), seguido do genoma de milho (Schnable et al., 2009) e sorgo (Paterson et al., 2009).

Cerca de 35% do genoma de *Oryza sativa* ssp. *japonica* é composto por transposons. Os elementos transponíveis da classe II (163.800) estão presentes em maior número que os elementos da classe I (61.900). Entretanto, quando o tamanho dos elementos é levado em consideração, os retrotransposons contribuem com 19,4% do genoma enquanto os transposons de DNA somam 13,0%. Em relação ao conteúdo de retrotransposons, a ordem LTR foi predominante, representando 76,2% dos elementos da classe I, enquanto os não-LTR somaram 23,8%. Das superfamílias de retrotransposons LTR, *Gypsy-Ty3* (10,9%) é mais abundante no genoma de arroz do que *Copia-Ty1* (3,9%). Elementos não-LTR contribuíram com uma pequena fração do genoma, LINEs com 1,1%, SINEs com 0,1% e outros elementos não classificados somaram 3,4% (International Rice Genome Sequencing Project, 2005).

A distribuição dos transposons de DNA foi mais uniforme entre as superfamílias identificadas. Das cinco superfamílias presentes no genoma de arroz, *Tc1-Mariner* apresentou o maior números de cópias (67000), mas, quando o tamanho dos elementos é levado em consideração, a superfamília *Mutator* contribuiu com a maior

fração do genoma (3,6%) e superou a superfamília mais numerosa (2,3%). A superfamília que foi menos representada, tanto em número quanto tamanho foi *hAT*, com 1100 elementos e 0,4% do genoma (International Rice Genome Sequencing Project, 2005).

O genoma de sorgo é composto por 55% de retrotransposons, maior do que a fração encontrada em milho e para os BACs da cultivar RB867515. Desses apenas 0,1% são elementos não-LTR e 0,1% ainda não foram classificados, o restante são compostos por LTRs (54,5%). A proporção de retroelementos LTR *Gypsy-Ty3*/*Copia-Ty1* foi de 3,7 para 1, menor que a de arroz (4,9 para 1) e maior que a de cana-de-açúcar (1,14 para 1). Não foram detectados elementos SINE no genoma de sorgo, ou a porcentagem para o genoma não foi significativa (Paterson et al., 2009). Os transposons de DNA contribuem com 7,5% do genoma dessa gramínea. Foram identificadas quatro superfamílias, em que *CACTA* (ou CMC) foi a mais abundante, representando 4,8% do genoma, enquanto a menos representativa foram as *hAT* e *PIF/Harbinger* com 0,02% cada. Foram identificados também elementos *MITE* (*Stowaway* e *Tourist*), que somaram 1,7% do genoma, e *Helitrons* com 0,8% (Paterson et al., 2009).

Um das culturas mais representativas em termos de composição de elementos repetitivos no genoma é o milho, que apresenta quase 85% da sua constituição de elementos genéticos moveis, dispersos não uniformemente pelo genoma. Muitos dos elementos transponíveis, em especial os primeiros representantes das superfamílias *CACTA* (*En/Spm*), *hAT* (*Ac*), *PIF/Harbinger* e *Mutator*, foram inicialmente descobertos em milho (Schnable et al., 2009).

Elementos da classe I também predominam o genoma de milho, com 75,6% contra 8,6% de transposons de DNA. Retrotransposons LTR representam 70,1% do genoma, com predomínio da superfamília *Gypsy-Ty3* (46,4%) em relação à *Copia-Ty1* (23,7%), e uma fração não classificada (4,5%). Já os retrotransposons não-LTR somaram uma pequena fração do genoma (1,0%), com os elementos LINE predominando sobre SINE (17,58 para 1). O genoma de milho contém 855 famílias de transposons de DNA, classificadas em seis superfamílias (incluindo *Helitrons*), que somam 8,6% do genoma dessa gramínea. A superfamília mais representativa foi *CACTA* (3,2%) e a com menor contribuição foi *MLE/Stowaway* (0,1%) (Schnable et al., 2009).

Na fração do genoma de cana-de-açúcar amostrada nesse trabalho foram identificados 4662 microssatélites, que, em termos relativos, correspondem a 90,6% das sequências genéticas repetitivas. Quando agrupados pela tamanho dos motivos repetidos,

foram encontrados microssatélites perfeitos de mono (1 base), di (2 bases), tri (3 bases) e tetranucleotídeos (4 bases). Os microssatélites de mononucletídeos representam a maior proporção de repetições, o motivo A/T com 74,6 % e C/G com 17,6 %, totalizaram 92,2 % (Figura 17).

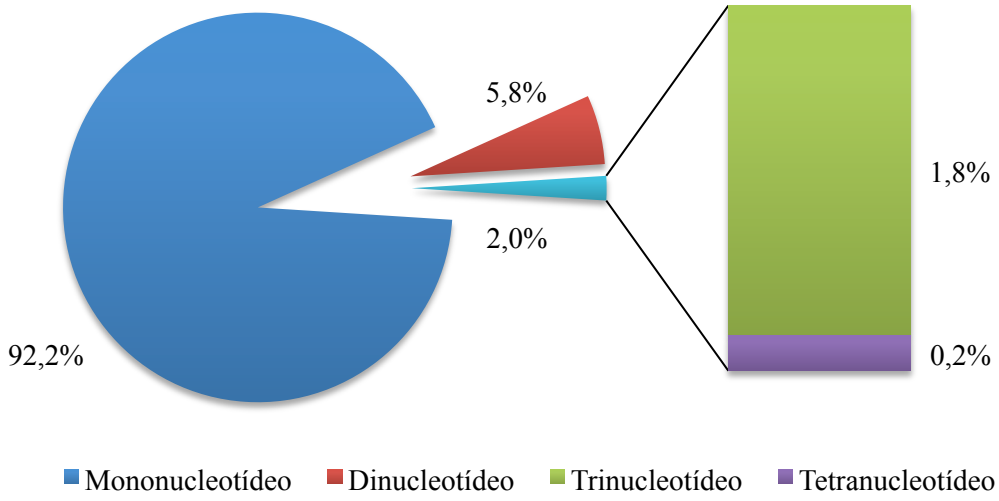


Figura 17. Proporções de microssatélites agrupados pela tamanho dos motivos repetidos.

Entre os dinucleotídeos, que totalizaram 5,8%, os motivos AT/TA foram os mais representativos, aparecendo 42 vezes nas sequências genômicas, enquanto o motivo menos representativo, AG/TC, apareceu 5 vezes. Microssatélites de trinucleotídeos e de tetranucleotídeos representaram uma pequena porção do total de microssatélites, 2,0%. O motivo tri CCG/GGC foi o mais representativo, com 9 repetições. Apenas um motivo de tetranucleotídeos foi identificado nas sequências: ATGT (Tabela 13). Em geral, a quantidade de repetições diminuiu com o aumento do tamanho dos motivos, mostrando que locos microssatélites com motivos longos são mais raros.

Após a anotação dos elementos repetitivos em todos os BACs, foi construído um arquivo no formato "gff3" para cada uma das sequências. Esse tipo de arquivo possui informações sobre a anotação das sequências, como as posições de início e término de cada elemento repetitivo (designadas como *features*), e a classificação dos elementos. Os arquivos de cada BAC foram utilizados no programa DNAplotter (Carver et al., 2009) para a construção de uma representação esquemática dos elementos. Para efeito de visualização, os retrotransposons foram divididos em LTR/*Gypsy-Ty3*, LTR/*Copia-Ty1* e outros

retrotransposons (LINEs e SINEs), os DNA transposons foram agrupados em única categoria, assim como os microssatélites. Outros elementos genéticos como os *Helitrons*, regiões de baixa complexidade, genes e sequências não anotadas não foram representadas separadamente no esquema (Figura 18).

Tabela 13. Motivos, frequência e tamanho dos microssatélites perfeitos identificados nas regiões genômicas caracterizadas da cultivar RB867515, correspondentes aos oito BACs da cultivar R570

motivo	frequência (n)	tamanho (pb)
(A/T)n	3606	20837
(C/G)n	920	4919
(AC/TG)n	17	190
(AG/TC)n	5	62
(AT/TA)n	42	794
(CA/GT)n	7	72
(CG/GC)n	7	70
(CT/GA)n	26	418
(ACA/TGT)n	4	63
(ACG/TGC)n	1	15
(ACT/TGA)n	1	18
(AGA/TCT)n	1	15
(AGC/TCG)n	1	18
(CAG/GTC)n	2	30
(CCA/GGT)n	3	45
(CCG/GGC)n	9	153
(CGA/GCT)n	1	15
(CGC/GCG)n	5	87
(CGG/GCC)n	3	51
(ATGT/TACG)n	1	55
TOTAL	4662	27927

Pela representação esquemática dos elementos repetitivos pode-se notar que eles estão presentes em todas as oito sequências da cultivar RB867515. Os elementos retrotransponíveis sobressaem sobre os transposons de DNA em número e tamanho das sequências, exceto para o BAC 1, que possui mais transposons de DNA em relação aos retrotransposons que os demais (Figura 18).

É possível visualizar grandes regiões colonizadas por retrotransposons, principalmente os da ordem LTR (*Gypsy-Ty3* e *Copia-Ty1*). Pelo seu mecanismo de transposição, a probabilidade reinserção da sua cópia na mesma região, ou em região próxima da origem é baixa (Civán et al., 2011).

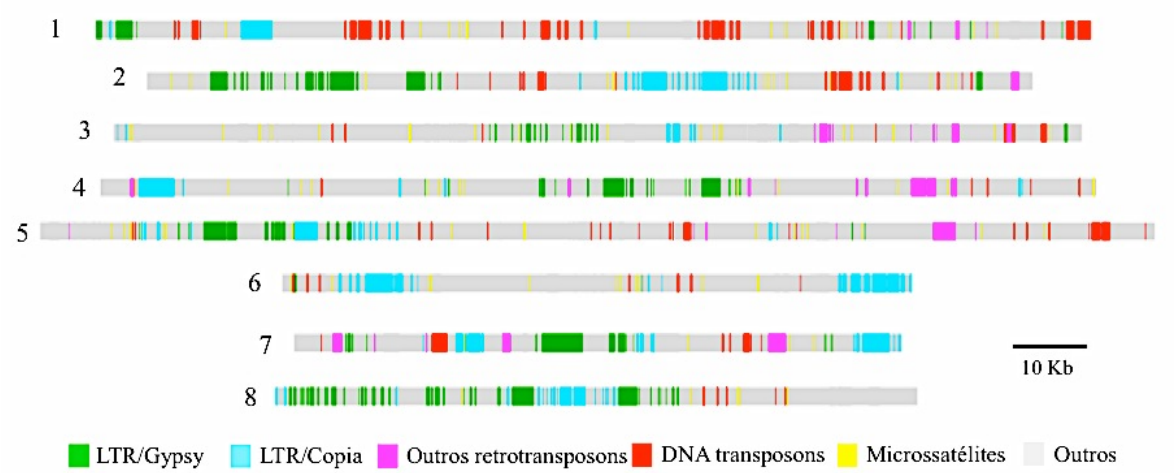


Figura 18. Representação esquemática da anotação dos BACs de RB867515 em relação aos elementos genéticos transponíveis.

Retrotransposons exibiram um padrão família-específico de distribuição não uniforme ao longo dos cromossomos no genoma de milho. Elementos *Copia-Ty1* foram super-representados em regiões de eucromatinas, ricas em genes, enquanto os elementos Gypsy-Ty3 apresentaram maior afinidade por regiões de heterocromatina, pobres em genes (Schnable et al., 2009). Em sorgo, foi possível notar uma grande composição de retrotransposons LTR pericentroméricos. A inserção de LTRs jovens (10 mil anos atrás) apareceram distribuídas aleatoriamente nos cromossomos, o que sugere que esses elementos foram eliminados preferencialmente de regiões ricas em genes e acumulados em regiões pobres em constituição gênica. Foram datadas duas ondas de retrotransposição no genoma dessa gramínea, uma pequena há 1-2 milhões de anos e a maior há 1 milhão de anos atrás (Paterson et al., 2009).

Os transposons de DNA estão mais agrupados em comparação com a disposição dos retrotransposons. Esses elementos se transpõem pelo mecanismo de "corta-e-cola", em que o elemento é excisado da sua região de origem para a inserção em um novo local. Com base nos dados apresentados não há evidências suficientes para confirmar a teoria de que esses elementos tendem a se reinserir em locais próximos ao sítio de origem após a sua excisão (Civán et al., 2011).

4.4.2 Genes e sequências relacionadas

Foram identificados 134 genes nas oito sequências de cultivar de cana-de-açúcar, que totalizaram 243 kb da composição genômica dos BACs. O número de elementos variou de 11 para o BAC 8 a 26 para o BAC 7, com uma média de 16 genes por BAC. Os genes encontrados apresentaram tamanho médio de 1841 pb, variando de 443 (BAC 1) à 6316 pb (BAC 3). A densidade de genes média foi de 1 gene por 7,2 kb, variando de 3,3 para o BAC7 a 10,6 para o BAC3 (Tabela 14).

Tabela 14. Composição em termos de genes preditos nas regiões correspondentes a cada BAC, na cultivar RB867515

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**		
Genes									**Média**	**Total**
Número	17	17	13	21	17	12	26	11	16	134
Tamanho Médio (pb)	443	1.967	6.316	1.018	2.165	1.508	518	790	1.841	
Tamanho Total (kb)	7,5	37,4	75,8	25,4	51,9	18,1	16,6	10,3		243,0
Densidade de genes (kb/gene)	8,4	7,4	10,6	6,7	9,3	6,8	3,3	7,9	7,2	
Isoformas/Gene										
Número médio	1,00	1,12	1,77	1,19	1,44	1,00	1,23	1,27	1,25	
Máximo	1	2	7	3	4	1	4	2		
Número total	17	21	49	30	34	12	42	15		220
Éxons/Gene										
Número médio	1,18	1,32	2,58	1,53	1,75	1,50	1,04	1,09	1,50	
Máximo*	3	4	8	7	9	6	2	2		
Tamanho Médio (pb)	321	658	365	393	431	436	383	728	464	
Tamanho Total (kb)	6,4	17,8	24,4	17,3	19,8	7,8	13,8	10,2		117,5
Íntrons/Gene										
Número médio	0,18	0,32	1,58	0,53	0,75	0,50	0,04	0,09	0,50	
Máximo	2	3	7	6	8	5	1	1		
Tamanho Médio (pb)	376	2,450	1,167	429	1,461	570	697	80	904	
Tamanho Total (kb)	1,1	19,6	51,4	8,1	32,1	10,3	2,8	0,08		125,5

*Número máximo em relação à média do número de éxons dos transcritos por gene

Em relação às isoformas, foi observada uma média de 1,25 transcritos por gene. O maior número de isoformas foi obtido para o gene 8 do BAC 3 (Figura 19), que apresentou sete isoformas (Tabela 14). Pode-se notar que o primeiro éxon é comum a todos as isoformas, mas apresenta tamanho reduzido na isoforma 7. Os demais éxons variaram entre as isoformas, com exceção do ultimo éxon que é comum às isoformas 1, 2 e 3.

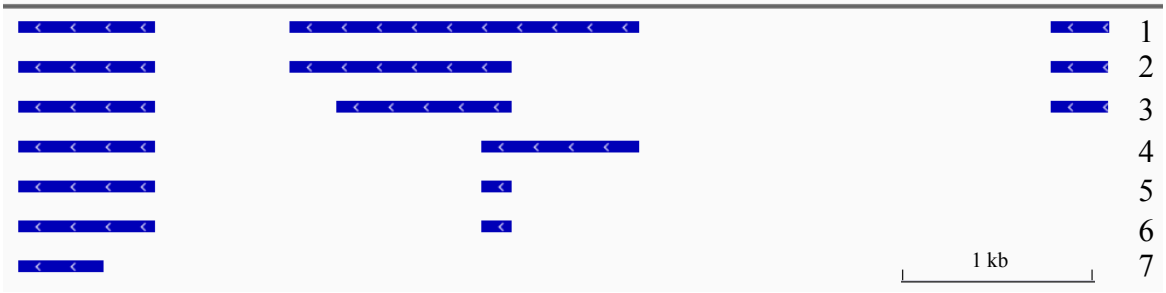


Figura 19. Esquema da composição dos éxons em cada uma das sete isoformas presentes no gene 8 do BAC 3. Figura obtida no visualizador de genomas IGV, a partir da sequência consenso e do arquivo de anotação construído.

Foi observada uma média de 1,50 éxons por gene, com uma flutuação de 1,04 (BAC 7) a 2,58 (BAC 3). O número máximo de éxons observados por gene foi 11, para a isoforma 1 do gene 4 (BAC 5) (Figura 20). O tamanho médio dos éxons foi de 464 pb, variando de 321 para o BAC 1 a 728 para o BAC 8. Em relação aos íntrons, observou-se um tamanho médio para esses elementos de 904 pb. A variação deste tamanho médio para os diferentes BACs foi bastante ampla, com 2450 pb para o BAC 2 e 80 pb para o BAC 8 (Tabela 14).

Figura 20. Esquema do gene 4 do BAC 5, evidenciando as duas isoformas encontradas para esse gene e os éxons representados pelos blocos azuis interligados pela linha azul com as setas indicando o sentido dos genes. Figura obtida no visualizador de genomas IGV, a partir da sequência consenso e do arquivo de anotação construído.

Para as sequências da cultivar R570 que foram utilizadas como referência para a montagem das sequências consenso da cultivar RB867515 foram encontrados 15 genes com suas versões alélicas distribuídas em seis ou sete haplótipos. Com exceção de um gene que foi determinado como fragmento de gene, os outros 14 genes apresentam uma estrutura éxons/íntron conservada em todos os hom(e)ólogos. A colinearidade geral entre os haplótipos foi rompida por poucas duplicações segmentares. A densidade de genes encontrada em quase 1 Mb de sequências foi de 1 gene cada 9 kb (Garsmeur et al., 2011)

No genoma de arroz foram preditos 37.544 genes, com uma densidade de 1 gene cada 9,9 kb, e tamanho médio de 2.699 pb. Foram identificados 175.203 éxons, totalizando 44 Mb. O número médio de éxons por gene foi de 4,7 com tamanho médio de 254 pb. Em relação aos íntrons, os 137.659 elementos somaram 56 Mb do genoma de

arroz, com uma média de 3,7 íntrons por gene e tamanho médio de 413 pb (International Rice Genome Sequencing Project, 2005).

Dos 34.496 modelos gênicos identificados no genoma de sorgo, 27.640 foram considerados genes codificadores de proteínas *bona fide*, por acumularem evidências de homologia e predição de genes *ab initio* com sequências expressas de sorgo, milho e cana-de-açúcar. Os genes têm tamanho médio de 2.873 pb (sem regiões UTR), e densidade de 1 gene por 24 kb. Foram identificados 129.000 éxons, com uma média de 4,7 éxons por gene. Foi previsto um tamanho médio dos íntrons de 436 pb (Paterson et al., 2009).

Foram preditos 32.540 genes para o genoma de milho, apresentando um tamanho médio de 3.733 pb e densidade de 1 gene por 70,7 kb. Em relação aos éxons, foi observado um tamanho médio de 304 pb e uma densidade média de 5,3 éxons por gene. Já o tamanho médio para os íntrons foi de 516 pb. Em comparação com as outras gramíneas, o elevado tamanho médio dos íntrons é atribuído às inserções por elementos transponíveis no seu genoma (Schnable et al., 2009).

Os genes identificados pelo mapeamento dos *reads* de RNAseq foram inspecionados visualmente, quanto a sua composição de éxons e as diferentes isoformas antes da identificação de suas funções. A partir dessa inspeção observou-se alguns problemas na identificação dos genes pelo TopHat e Cufflinks. Alguns genes foram identificados como distintos, quando na verdade tratam-se de um único gene com diferentes números/estruturas de éxons. Outro caso que exigiu intervenção manual na predição de genes foi a identificação de um mesmo gene, na mesma posição, mas ao longo das duas fitas de DNA. Como essa estrutura gênica é, no mínimo, improvável, a identificação dos genes envolvidos foi ignorada. Alguns genes, por apresentaram tamanhos de íntrons muito grandes, maiores de 30 kb, não haviam sido identificados ou haviam sido separados em genes diferentes. Estes erros foram corrigidos manualmente.

Os genes preditos foram submetidos ao Blastx contra o banco de proteínas não redundantes do GenBank para a anotação das proteínas presentes nas regiões genômicas caracterizadas (Tabela 15). Dos 134 genes identificados pelo TopHat e Cufflinks, 62 (46,3%) não tiveram correspondência no banco de proteínas e foram consideradas como "No hit". Os genes indicados como "Conserved protein" e "Hypothetical protein" foram agrupados em uma categoria Conserved/Hypothetical protein" e compreendem 34 genes, representando 24,4% do conjunto. Alguns elementos genéticos móveis foram identificados dentro do conjunto gênico, como algumas proteínas produzidas pelos retrotransposons.

Esses elementos foram agrupados na categoria “Transposon”, com 16 elementos (11,9%) e, em uma categoria separada, foram identificadas duas “Transcriptases” previamente identificadas na cultivar R570.

Tabela 15. Descrições das proteínas identificadas pelo Blastx contra o banco de proteínas não redundantes do GenBank, ordenadas pelo número de genes identificados com a mesma descrição

	Descrição	Nº Genes
1	No hit	62
2	Conserved/Hypothetical protein	34
3	Transposon	16
4	NADP-dependent D-sorbitol-6-phosphate dehydrogenase [*Saccharum* hybrid cultivar]	9
5	Endoglucanase 4 precursor [*Saccharum* hybrid cultivar R570]	3
6	Putative ulp1 protease [*Saccharum* hybrid cultivar R570]	3
7	Transcriptase [*Saccharum* hybrid cultivar R570]	2
8	Transporter [*Saccharum* hybrid cultivar R570]	2
9	Putative shrunken seed protein [*Saccharum* hybrid cultivar R570]	1
10	Putative trehalose-phosphatase (C-terminal fragment) [*Saccharum* hybrid cultivar R570]	1
11	Subtilisin-like protease [*Saccharum* hybrid cultivar R570]	1
	TOTAL	134

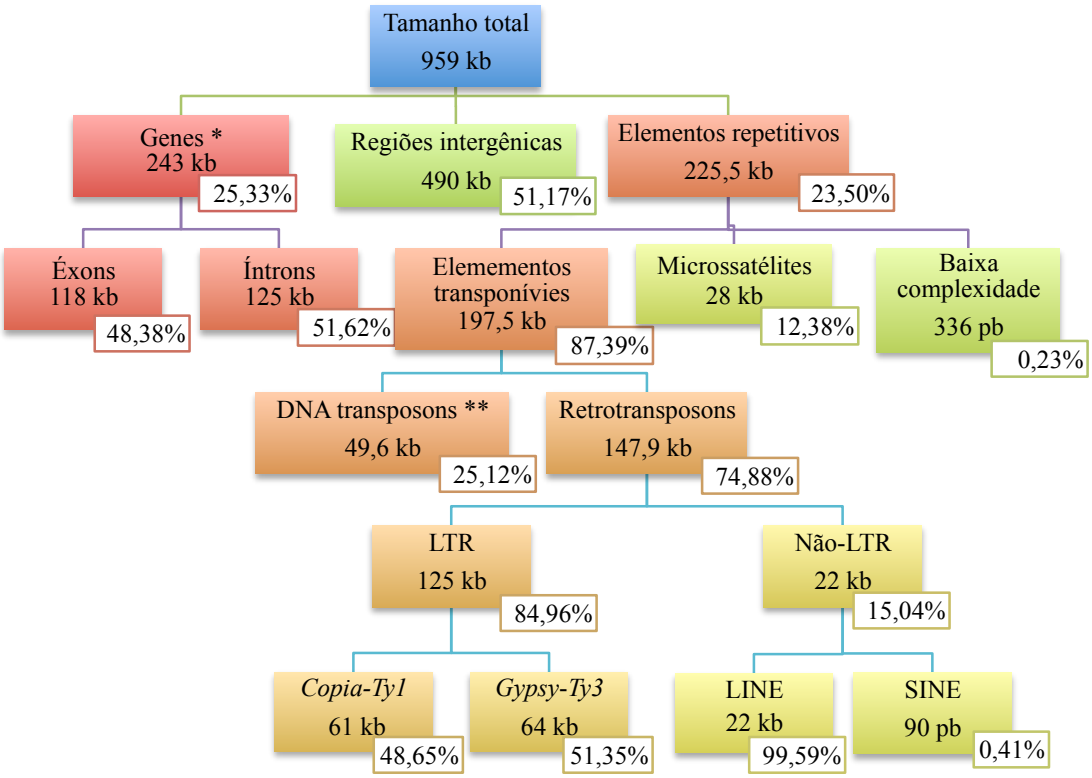
As 22 proteínas restantes foram idenficadas pelo Blastx. Os critérios utilizados na análise pelo Blastx foram: evalue $< 10^{-20}$, cobertura de alinhamento > 50 pb e identidade > 70%. Desse total, nove proteínas foram associadas à “NADP-dependent D-sorbitol-6-phosphate dehydrogenase”. O gene codificante para essa proteína apresentou a maior distribuição entre os BACs e está presente em quase todos eles, exceto no BAC 6, além de estar duplicado nos BACs 2 e 7 (Apêndice A).

A primeira descrição da proteína NADP-dependent D-sorbitol-6-phosphate dehydrogenase (NADP-S6PDH) foi feita para uma espécie de maçã (*Malus domestica*). A proteína NADP-S6PDH sintetiza sorbitol-6-phosphate, que é um intermediário chave na síntese de sorbitol, o maior produto fotossintético da família Rosaceae (Kanayama et al., 1992). Na cultivar comercial de cana-de-açúcar R570 essa proteína também foi identificada, e se apresentou duplicada no mesmo haplótipo (Garsmeur et al., 2011).

A proteína “Endoglucanase 4 precursor” esta presente em três haplótipos do genoma amostrado da RB867515. Essa proteína possui uma atividade catalítica de endohidrólise de ligações glicosídicas da celulose em plantas (Juturu & Wu, 2014). A proteína “putative ulp1 protease” também foi identificada em três dos oito haplótipos de

cana-de-açúcar anotados. A protease ulp1 (Ubiquitin-like-specific protease 1) foi descrita em levedura, como atuante no ciclo celular (Fase G2/Mitose) (Li & Hochstrasser, 1999).

A partir da anotação das sequências obtidas para a cultivar de cana-de-açúcar foi possível construir um esquema da composição da amostra do seu genoma, em termos de sequências codificadoras e elementos genéticos móveis (Figura 21). Levando em consideração o tamanho total dos genes, incluindo as regiões de íntrons, tem-se uma composição quase equivalente entre regiões repetitivas (226 kb) e gênicas (243 kb). Mas é preciso atentar-se para o fato que há inserções de elementos repetitivos dentro dos genes, em regiões de íntrons, contribuindo para o aumento do tamanho destas regiões. Esse fato já foi relatado no genoma de milho, por exemplo, que possui tamanho médio de íntrons superior às demais gramíneas, por ter sofrido diversas invasões de elementos transponíveis em regiões gênicas (Schanable et al., 2009).



*Não estão representadas as regiões UTR 5'e 3'
** Incluíndo os *Helitrons* (Transposons de DNA subclasse II)
Figura 21. Composição estimada do genoma da cultivar de cana-de-açúcar RB867515.

Quando o conteúdo de éxons (118 kb) é comparado com o dos elementos repetitivos (226 kb), nota-se que as sequências repetitivas compõem a maior parte da composição do genoma. Dentro dos elementos genéticos transponíveis, a grande maioria

das sequências pertencem à classe dos retrotransposons (74,9%). Um padrão que já foi observado em outras gramíneas como milho (Schanable et al., 2009) e sorgo (Paterson et al., 2009).

4.5 ANÁLISE COMPARATIVA DA R570 X RB867515

Organismos poliploides são caracterizados pela presença de mais de um par de cromossomos homólogos na sua constituição. Nessas espécies, o pareamento dos cromossomos na meiose será diferenciado dependendo da natureza da sua poliploidia. Os alopoliploides, no qual o genoma é derivado de uma combinação de espécies, apresentam um pareamento cromossômico preferencial na meiose, com a formação de bivalentes. No entanto, nas espécies que são consideradas autopoliploides, derivadas da duplicação do seu próprio genoma, ou derivadas de uma combinação de dois ou mais genomas geneticamente muito similares, o pareamento dos cromossomos na meiose não ocorrerá de forma preferencial, e pode ser observado a formação de multivalentes (Albino et al., 2006).

A construção de mapas genéticos permite a definição dos grupos de hom(e)ologia formados pelos cromossomos. Um grupo de ligação é designado com base nos marcadores em associação (segregam juntos na meiose) que correspondem ao mesmo cromossomo. Um conjunto de cromossomos homólogos são agrupados pela presença de marcadores comuns, derivados de um mesmo loco gênico (em diferentes grupos de ligação), independente do tipo do marcador genético utilizado (Microssatélite, AFLP, DArT, entre outros) (Aitken et al., 2005).

Em poliploides, além da presença de cromossomos homólogos, há uma diferenciação para os homeólogos, que são conjuntos cromossômicos geneticamente distintos entre si (podem ser decorrentes de origem distinta) e são agrupados em grupos de homeologia (Albino et al., 2006). Em cana-de-açúcar o número de grupos de ligação e homologia pode variar, em consequência da presença de conjuntos de cromossomos aneuploides no genoma, além da origem biespecífica e de recombinações interespecíficas dos cromossomos.

Para realizar uma análise comparativa dos elementos genéticos presentes nas duas cultivares de cana-de-açúcar, os BACs de R570 foram reanotados, utilizando os mesmos critérios da anotação das sequências de RB867515. Os elementos genéticos

móveis foram identificados pelo *Repeat Masker*. Os genes e sequências relacionadas a genes foram anotados pela ferramenta Blastx contra o banco de proteínas não redundante. O esquema representativo da anotação das sequências das duas cultivares foi produzido pelo DNAplotter (Figura 22).

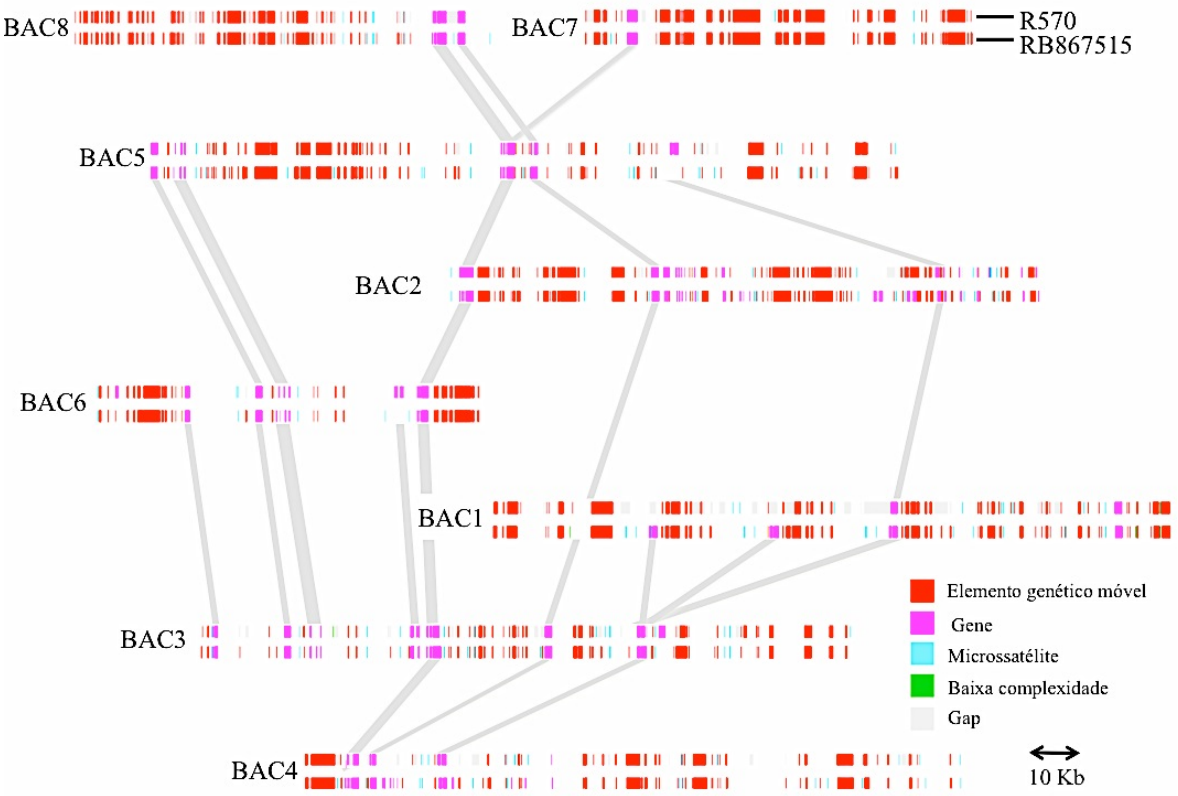


Figura 22. Representação esquemática da anotação dos BACs de R570 (superior) e RB867515 (inferior). As linhas cinza conectam genes homólogos entre as sequências.

Pela anotação realizada, notou-se a presença dos mesmos genes e demais elementos genéticos em posições análogas nas duas cultivares. Além disso, genes homólogos foram identificados entre as sequências, o que as incorpora em um mesmo grupo de homeologia (Figura 22).

Essa semelhança entre as sequências também foi analisada em relação ao conteúdo nucleotídico. Através do alinhamento das sequências de R570 e RB867515 foi possível estimar a porcentagem de *mismatches* entre elas, que representa a fração dos nucleotídeos que não são correspondentes nas duas sequências. A porcentagem de *mismatches* entre as sequências homólogas foi baixa e variou de 388 pb (0,3%) para o BAC1, a 1869 pb (1,3%) para o BAC4 (Figura 23), mais uma vez confirmando a semelhança entre os pares de homólogos.

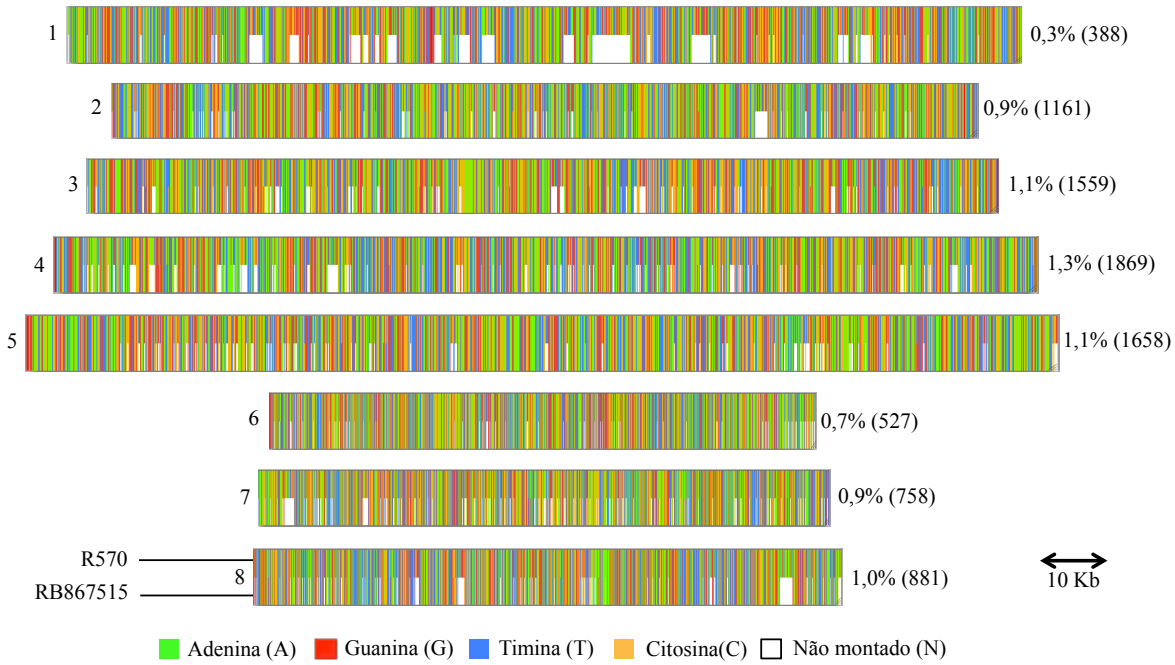


Figura 23. Representação da porcentagem dos *mismatches* entre sequências homólogas de R570 e RB867515, entre parênteses estão representados os números absolutos em pares de bases. As quatro cores representam as quatro bases nitrogenadas e espaços em branco simbolizam regiões que não foram montadas. Barras verticais da mesma cor significam um *match* (mesma base) e barras verticais com duas cores diferentes representam um *mismatch* (bases diferentes).

Afim de verificar a organização de blocos conservados entre e dentro dos grupos de homologia, as sequências foram alinhadas no software Mauve (Darling et al., 2004). A partir do esquema obtido pelo alinhamento das 16 sequências (8 BACs de cada cultivar) pode-se observar que existem grandes blocos conservados de sequências entre as sequências homólogas, e nenhum rearranjo estrutural foi observado entre elas. Além disso, foi detectado uma grande divergência entre os diferentes membros do grupo de hom(e)ologia (Figura 24).

Figura 24. Esquema do alinhamento enfatizando a divergência entre os membros de um mesmo grupo de hom(e)ologia. As cores diferentes representam blocos de sequências homólogas conservadas entre os grupos. As sequências estão em ordem crescente, de cima para baixo, sendo uma sequência de BAC R570 (superior) e a correspondente RB867515 (inferior).

.

5 CONCLUSÕES

O método de sequenciamento utilizado foi eficiente na montagem nas sequências da cultivar de cana-de-açúcar RB867515. As sequências obtidas se mostraram bastante fragmentadas, mas com a inclusão de bibliotecas de DNA com tamanho de fragmentos maiores do que 1 kb espera-se uma melhora na montagem das sequências, ligando *scaffolds* antes desconectados. A semelhança entre as estruturas genômicas das duas cultivares permitiu uma taxa de recuperação das sequências de até 97%.

Como sugerido pela literatura, os resultados obtidos nestes trabalho evidenciam que o genoma de cana-de-açúcar é abundante em elementos genéticos repetitivos. Os retrotransposons compõem a maior porção entre os elementos genéticos móveis, com predomínio dos LTR. Em virtude do mecanismo de transposição desses elementos, eles podem ter contribuído para o aumento de tamanho do genoma destas plantas.

Foram identificadas 4662 regiões microssatélites, em que os motivos mononucleotídeos foram os mais abundantes, seguidos pelos motivos de di, tri e tetranucleotídeos. Essas sequências identificadas, bem como suas regiões flanqueadoras podem ser utilizadas para a construção de iniciadores para estudos posteriores de análise de diversidade genética dessa cultura.

O presente trabalho foi pioneiro na identificação de genes na cultivar RB867515. A estrutura de éxons e íntrons destes genes foi elucidada, incluindo aspectos relacionados ao número dessas estruturas por gene e seu tamanho médio. Além disso, foi possível identificar diferentes isoformas para alguns desses genes.

As sequências dos BACs de R570 e as das regiões correspondentes no genoma da cultivar RB867515 mostraram-se bastante semelhantes, tanto em termos de composição nucleotídica, quanto pela constituição em elementos genéticos móveis, genes e blocos conservados. A presença de genes homólogos nas diferentes regiões genômicas estudadas permitiu posicionar as 16 sequências (oito de cada cultivar) em um mesmo grupo de homeologia.

A grande similaridade entre os cromossomos de cultivares de base genética tão distinta quanto as cultivares R570 e RB867515, tem implicações importantes no que diz respeito ao uso das sequências obtidas com uma eventual montagem do genoma completo de cana-de-açúcar. Esta similaridade deverá permitir o uso mais abrangente das informações produzidas em uma futura sequência de referência para o genoma de cana-de-açúcar, do que antecipado pela complexidade citogenética da espécie reportada na literatura.

6 REFERÊNCIAS

ADAMS, K. L.; WENDEL, J. F. Polyploidy and genome evolution in plants. **Current Opinion in Plant Biology**, Philadelphia, v. 8, n. 1, p. 135-141, jan. 2005.

AITKEN, K. S.; JACKSON, P. A.; MCINTYRE, C. L. A combination of AFLP and SSR markers provides extensive map coverage and identification of homo(eo)logous linkage groups in a sugarcane cultivar. **Theoretical and Applied Genetics**, Heidelberg, v. 110, n. 4, p. 789-801, feb. 2005.

ALBINO, J.C.; CRESTE, S.; FIGUEIRA, A. Mapeamento genético da cana-de-açúcar. **Biotecnologia Ciência & Desenvolvimento**, v. 9, n. 36, p. 82-91, jan. 2006.

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**, Amsterdam, v. 17, n. 3, p. 281-281, june 1999.

ALTINKUT, A.; RASKINA, O.; NEVO, E.; BELYAYEV, A. En/Spm-like transposons in Poaceae species: transposase sequence variability and chromosomal distribution. **Cellular & Molecular Biology Letters**, Heidelberg, v. 11, n. 2, p. 214-229, june 2006.

ALTSCHUL, S. F.; GISH, W.; MILLER, W.; MYERS, E. W.; LIPMAN, D. J. Basic local alignment search tool. **J Mol Biol**., New York, v. 5, n. 215, p. 403-410, oct. 1990.

ARAUJO, P. G.; ROSSI, M.; JESUS, E. M.; SACCARO JR, N. L.; KAJIHARA, D.; MASSA, R.; FELIX, J. M.; DRUMMOND, R. D.; FALCO, M. C.; CHABREGAS, S. M.; ULIAN, E. C.; MENOSSI, M.; VAN SLUYS, M. Transcriptionally active transposable elements in recent hybrid sugarcane. **The Plant Journal**, Malden, v. 44, p. 707-717, aug. 2005.

ARENSBURGER, P.; HICE, R. H.; ZHOU, L.; SMITH, R. C.; TOM, A. C.; WRIGHT, J. A.; KNAPP, J.; O'BROCHTA, D. A.; CRAIG, N. L.; ATKINSON, P. W. Phylogenetic and functional characterization of the hAT transposon superfamily. **Genetics**, Pittsburg, v. 188, n. 1, p. 45-57, may 2011

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

BANCO NACIONAL DO DESENVOLVIMENTO; CENTRO DE GESTÃO E ESTUDOS ESTRATÉGICOS. Bioetanol de cana-de-açúcar no Brasil. In: BANCO NACIONAL DO DESENVOLVIMENTO; CENTRO DE GESTÃO E ESTUDOS ESTRATÉGICOS (Org.). **Bioetanol de cana-de-açúcar:** energia para o desenvolvimento sustentável. Rio de Janeiro. BNDES, 2008. p. 153-176.

BARBOSA, M. H. P.; SILVEIRA, L. C. I. Melhoramento genético e recomendação de cultivares. In: SANTOS, F.; BORÉM, A.; CALDAS, C. (Ed.). **Cana-de-açúcar, Bioenergia, Açúcar e Etanol**: Tecnologia e Perspectivas. 2 ed. Viçosa: Editora UFV, 2011. p. 313-332.

BARBOSA, M. H. P.; SILVEIRA, L. C. I.; OLIVEIRA, M. W.; SOUZA, V. F. M.; RIBEIRO, S. N. N. RB867515 Sugarcane cultivar. **Crop Breeding and Applied Biotechnology**, Viçosa, v. 1, n. 4, p. 437-438, sep. 2001.

BELARMINO, L. C.; SILVA, R. L. O.; CAVALCANTI, N. M. S.; KREZDORN, N.; KIDO, E. A.; HORRES, R.; WINTER, P.; KAHL, G.; BENKO-ISEPPON, A. M. SymGRASS: a database of sugarcane orthologous genes involved in arbuscular mycorrhiza and root nodule symbiosis, **BMC Bioinformatics**, London, v. 14, n. S2, p. 1-8, sep. 2012.

BENNETZEN, J. L.; WANG, H. The contributions of transposable elements to the structure, function, and evolution of plant genomes. **Annual Review of Plant Biology**, Palo Alto, v. 65, n. 1, p. 505-530, jan. 2014.

BERDING, N.; ROACH, B. T. Germplasm collection, maintence and use. In: HEINZ, D. (Ed.). **Sugarcane improvement through breeding**. Amsterdam: Elsevier, 1987. p. 143-210.

BHAT, S. R.; GILL, S. The implications of 2n egg gametes in nobilization and breeding of sugarcane. **Euphytica**, Wageningen, v. 34, n. 2, p. 377-384, may 1985.

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

BRADNAM, K. Assemblathon_stats.pl: A script to calculate a basic set of metrics from a genome *assembly*. 2011.

BREMER, G. Problems in breeding and cytology of sugar cane. **Euphytica**, Wageningen, v. 10, n. 1, p. 59-78, feb. 1961.

BUREAU, T. E.; WESSLER, S. R. Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. **The Plant Cell**, Waterbury, v. 6, n. 6, p. 907-916. june 1994.

CARVER, T.; THOMSON, N.; BLEASBY, A.; BERRIMAN, M.; PARKHILL, J. DNAPlotter: circular and linear interactive genome visualization. **Bioinformatics**, Oxford, v. 25, n. 1, p. 119-120, 2009.

CHRISTOFF, A.; LORETO, E. L. S.; SEPEL, L. M. N. Evolutionary history of the Tip100 transposon in the genus *Ipomoea*. **Genetics and Molecular Biology**, São Paulo, v. 35, n. 2, p. 460-465, apr. 2012.

CIVÁÑ, P.; SVEC, M.; HAUPTVOGEL, P. On the coevolution of transposable elements and plant genomes. **Journal of Botany**, New York, v. 2011, p. 1-9, sep. 2011.

COMPANHIA NACIONAL DE ABASTECIMENTO. **Acompanhamento de safra brasileira**: Cana-de-açúcar. 2° levantamento. Brasília: CONAB, 2013, 18 p.

CUADRADO, A.; ACEVEDO, R.; DE LA ESPINA, S. M. D.; JOUVE, N.; DE LA TORRE, C. Genome remodelling in three modern *S. officinarum* × *S. spontaneum* sugarcane cultivars. **Journal of experimental botany**, Lancaster, v. 55, n. 398, p. 847-854, apr. 2004.

D'HONT, A.; GRIVET, L.; FELDMANN, P.; GLASZMANN, J.; RAO, S.; BERDING, N. Characterisation of the double genome structure of modern sugarcane cultivars (*Saccharum* spp.) by molecular cytogenetics. **Molecular and General Genetics**, Heidelberg, v. 250, n. 4, p. 405-413, oct. 1996.

D'HONT, A.; ISON, D.; ALIX, K.; ROUX, C.; GLASZMANN, J. C. Determination of basic chromosome numbers in the genus *Saccharum* by physical mapping of ribosomal RNA genes. **Genome**, Birmingham, v. 41, n. 2, p. 221-225, jan. 1998.

D'HONT, A.; LU, Y. H.; DE LEÓN, D. G.; GRIVET, L.; FELDMANN, P.; LANAUD, C.; GLASZMANN, J. C. A molecular approach to unraveling the genetics of sugarcane, a complex polyploid of the Andropogoneae tribe. **Genome**, Birmingham, v. 37, n. 2, p. 222-230, oct. 1994.

D'HONT, A.; PAULET, F.; GLASZMANN, J. C. Oligoclonal interspecific origin of 'North Indian' and 'Chinese' sugarcanes. **Chromosome Research**, Oxford, v. 10, n. 3, p. 253-262, jan. 2002.

D'HONT, A. Unraveling the genome structure of polyploids using FISH and GISH; examples of sugarcane and banana. **Cytogenetic and genome research**, Basel, v. 109, n. 1-3, p. 27-33, july 2005.

D'HONT, A.; SOUZA, G. M.; MENOSSI, M.; VINCENTZ, M.; VAN-SLUYS, M. A.; GLASZMANN, J. C.; ULIAN, E. Sugarcane: a major source of sweetness, alcohol, and bio-energy. In: MOORE, P.; MING, R. (Ed.). **Genomics of tropical crop plants**. New York: Springer, 2008. p. 483-513.

DANIELS, J.; ROACH, B. T. Taxonomy and evolution. In: HEINZ, D. (Ed.). **Sugarcane improvement through breeding**. Amsterdam: Elsevier, 1987. p. 7-84.

DARLING, A. C. E.; MAU, B.; BLATTER, F. R.; PERNA, N. T. Mauve: multiple alignment of conserved genomic sequence with rearrangements. **Genome Research**, New York, v. 14, n. 7, p. 1394-1403, 2004.

DEVOS, K. M. Updating the ‘crop circle’. **Current opinion in plant biology**, Amsterdam, v. 8, n. 2, p. 155-162, jan. 2005.

DEVOS, K. M.; GALE, M. D. Comparative genetics in the grasses. **Plant Molecular Biology**, Dordrecht, v. 35, n. 1, p. 3-15, feb. 1997.

DOMINGUES, D. S.; CRUZ, G. M. Q.; METCALFE, C. J.; NOGUEIRA, F. T. S.; VICENTINI, R. V.; ALVES, C. S.; VAN SLUYS, M. Analysis of plant LTR-retrotransposons at the finescale family level reveals individual molecular patterns. **BMC Genomics**, London, v. 13, n. 137, p. 1-13, apr. 2012.

DOYLE, J. J.; FLAGEL, A. E.; PATERSON, A. H.; RAPP, R. A.; SOLTIS, D. E.; SOLTIS, P. S.; WENDEL, J. F. Evolutionary genetics of genome merger and doubling in plants. **Annu. Rev. Genet**, Palo Alto, v. 42, n. 1, p. 443-461, aug. 2008.

DUFOUR, P.; DEU, M.; GRIVET, L.; D’HONT, A.; PAULET, F.; BOUET, A.; LANAUD, C.; GLASZMANN, J.; HAMON, P. Construction of a composite sorghum genome map and comparison with sugarcane, a related complex polyploid. **Theoretical and Applied Genetics**, Heidelberg, v. 94, n. 3, p. 409-418, aug. 1997.

EWING, B.; GREEN, P. Base-calling of automated sequencer traces using phred. **Genome Research**, New York, v. 8, p. 186-194, 1998.

FESCHOTTE, C.; JIANG, N.; WESSLER, S.R. Plant transposable elements: where genetics meets genomics. **Nature Reviews Genetics,** London, v. 3, n. 5, p. 329-341, may 2002.

GARSMEUR, O.; CHARRON, C.; BOCS, S.; JOUFFE, V.; SAMAIN, S.; COULOUX, A.; DROC, G.; ZINI, C.; GLASZMANN, J.; VAN SLUYS, M.; D’HONT, A. High homologous gene conservation despite extreme autopolyploid redundancy in sugarcane. **New Phytologist**, Lancaster, v. 189, p. 629-642, sep. 2011.

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

GRIVET, L.; D'HONT, A.; DUFOUR, P.; HAMON, P.; ROQUES, D.; GLASZMANN, J. C. Comparative genome mapping of sugar cane with other species within the Andropogoneae tribe. **Heredity**, New York, v. 73, n. 5, p. 500-508, feb. 1994.

GRIVET, L.; DANIELS, C.; GLASZMANN, J. C.; D'HONT, A. A review of recent molecular genetics evidence for sugarcane evolution and domestication. **Ethnobotany Research & Applications**, Manoa, v. 2, n. 1, p. 9-17, mar. 2004.

GRIVET, L.; GLASZMANN, J.; D’HONT, A. Molecular Evidence of Sugarcane Evolution and Domestication. In: MOTLEY, T. J.; ZEREGA, N.; CROSS, H. B. (Ed.). **Darwin's harvest**: new approaches to the origins, evolution, and conservation of crops. New York: Columbia Univ Pr, 2006. p. 49-66.

GRZEBELUS, D.; LASOTA, S.; GAMBIN, T.; KUCHEROV, G.; GAMBIN, A. Diversity and structure of PIF/Harbinger-like elements in the genome of *Medicago truncatula*. **BMC Genomics**, London, v. 8, n. 409, p. 1-14, nov. 2007.

GUIMARÃES, C. T.; SILLS, G. R.; SOBRAL, B. W. S. Comparative mapping of Andropogoneae: *Saccharum* L.(sugarcane) and its relation to sorghum and maize. **Proceedings of the National Academy of Sciences**, Washington, v. 94, n. 26, p. 14261-14266, dec. 1997.

GUO, J.; XU, L.; SU, Y.; WANG, H.; GAO, S.; XU, J.; QUE, Y. ScMT2-1-3, a metallothionein gene of sugarcane, plays an important role in the regulation of heavy metal tolerance/accumulation. **BioMed Research International**, New York, v. 2013, p. 1-12, may 2013.

HERMANN, S.; AITKEN, K.; JACKSON, P.; GEORGE, A.; PIPERIDIS, N.; WEI, X.; KILIAN, A.; DETERING, F. Evidence for second division restitution as the basis for 2n + n maternal chromosome transmission in a sugarcane cross. **Euphytica**, Wageningen, v. 10, p. 1-10, apr. 2012.

INTERNATIONAL RICE GENOME SEQUENCING PROJECT. The map-based sequence of the rice genome. **Nature**, New York, v. 436, n. 7052, p. 793-800, aug. 2005.

ISKANDAR, H. M.; CASU, R.; FLETCHER, A. T.; SCHMIDT, S.; XU, J.; MACLEAN, D. J.; MANNERS, J. M. M.; BONNETT, G. D. Identification of drought-response genes and a study of their expression during sucrose accumulation and water deficit in sugarcane culms. **BMC Plant Biology**, London, v. 11, n. 12, dec. 2011.

JANNOO, N.; GRIVET, L.; CHANTRET, N.; GARSMEUR, O.; GLASZMANN, J. C.; ARRUDA, P.; D'HONT, A. Orthologous comparison in a gene-rich region among grasses reveals stability in the sugarcane polyploid genome. **Plant Journal**, Malden, v. 50, p. 574–585, mar. 2007.

JANNOO, N.; GRIVET, L.; SEGUIN, M.; PAULET, F.; DOMAINGUE, R.; RAO, P.; DOOKUN, A.; D'HONT, A.; GLASZMANN, J. Molecular investigation of the genetic base of sugarcane cultivars. **Theoretical and Applied Genetics**, Heidelberg, v. 99, n. 1, p. 171-184, nov. 1999.

JUTURU, V.; WU, J. C. Microbial cellulases: Engineering, production and applications. **Renewable and Sustainable Energy Reviews**, New York, 33 (2014) 188–203

KANAYAMA, Y.; MORI, H.; IMASEKI, H.; YAMAKI, S. Nucleotide sequence of a cDNA enconding NADP-Sorbitol-6-Phosfatase Dehydrogenase from apple. **Plant Physiology**, Waterbury, v. 100, n. 3, p. 1607-1608, nov. 1992.

KAZAZIAN JR, H. H. Mobile elements: drivers of genome evolution. **Science**, Washington, v. 303, n. 5664, p. 1626-1632, june 2004.

KELLOGG, E. A.; BENNETZEN, J. L. The evolution of nuclear genome structure in seed plants. **American Journal of Botany**, St. Louis, v. 91, n. 10, p. 1709-1725, oct. 2004.

KEMPKEN, F.; WINDHOFER, F. The hAT family: a versatile transposon group common to plants, fungi, animals, and man. **Chromosoma**, Heidelberg, v. 110, n. 1, p. 1-9, apr. 2001.

KIDO, E. A.; FERREIRA NETO, J. R. C.; SILVA, R. L. O.; PANDOLFI, V.; GUIMARÃES, DA. C. R.; VEIGA, A. T.; CHABREGAS, S. M.; CROVELLA, S.; BENKO-ISEPPON, A. M. New insights in the sugarcane transcriptome responding to drought stress as revealed by Supersage. **The Scientific World Journal**, New York, v. 2012, p. 1-14, dec. 2012.

KRISHNAN, A.; GRECO, R.; PEREIRA, A. Diversity of En/Spm transposons in maize and rice. **Maydica**, Bergamo, v. 53, n. 3, p. 181-187, oct. 2008.

KUBIS, S.; SCHMIDTJ, T.; HESLOP-HARRISON, J. S. Repetitive DNA elements as a major component of plant genomes. **Annals of Botany**, Exeter, v. 82, n. 9, p. 45-55, sep. 1998.

LANGDON, T.; SEAGO, C.; MENDE, M.; LEGGETT, M.; THOMAS, H.; FORSTER, J. W.; THOMAS, H.; JONES, R. N.; JENKINS, G. Retrotransposon evolution in diverse plant genomes. **Genetics**, Pittsburg, v. 156, n. 3, p. 313-325, sep. 2000.

LANGMEAD, B; SALZBERG, S. Fast gapped-read alignment with Bowtie 2. **Nature Methods**, New York, v. 9, p. 357-359, 2012.

LARKIN, M. A.; BLACKSHIELDS, G.; BROWN, N. P.; CHENNA, R.; MCGETTIGAN, P. A.; MCWILLIAM, H.; VALENTIN, F.; WALLACE. I. M.; WILM, A.; LOPEZ, R.; THOMPSON, J. D.; GIBSON, T. J.; HIGGINS, D. G. Clustal W and Clustal X version 2.0. **Bioinformatics**, Oxford, v. 23, n. 21, p. 2947-2948, sep. 2007.

LI, H.; DURBIN, R. Fast and accurate long-read alignment with Burrows-Wheeler Transform. **Bioinformatics**, Oxford, v. 25, n. 5, p. 589-595, mar. 2010.

LI, H.; HANDSAKER, B.; WYSOKER, A.; FENNELL, T.; RUAN, J.; HOMER, N.; MARTH, G.; ABECASIS, G.; DURBIN, R. The sequence alignment/map format and SAMtools. **Bioinformatics**, Oxford, v. 25, n. 16, p. 2078-2079, aug. 2009.

LI, S.; HOCHSTRASSER, M. A new protease required for cell-cycle progression in yeast. **Nature**, New York, v. 318, n. 6724, p. 246-251, mar. 1999.

LISCH, D. How important are transposons for plant evolution? **Nature reviews Genetics**, New York, v. 14, n. 1, p. 50-63, jan. 2013.

MENOSSI, M.; SILVA-FILHO, M. C.; VINCENTZ, M.; VAN-SLUYS, M.-A.; SOUZA, G.M. Sugarcane functional genomics: gene discovery for agronomic trait development. **International Journal of Plant Genomics**, New York, v. 2008, p. 1-11, nov. 2008.

MILLER, W. J.; CAPY, P. Applying mobile genetic elements for genome analysis and evolution. **Molecular Biotechnology**, Heidelberg, v. 33, n. 2, p. 161-174, june 2006.

MING, R.; IRVINE, J. E.; LIU, S. C.; MOORE, P. H.; PATERSON, A. H.; BOWERS, J. E. Construction of a consensus genetic map from two interspecific crosses. **Crop science**, Madison, v. 42, n. 2, p. 570-583, apr. 2002.

MING, R.; LIU, S. C.; LIN, Y. R.; DA SILVA, J.; WILSON, W.; BRAGA, D.; VAN DEYNZE, A.; WENSLAFF, T.; WU, K.; MOORE, P. Detailed alignment of *Saccharum* and *Sorghum* chromosomes: comparative organization of closely related diploid and polyploid genomes. **Genetics**, Pittsburg, v. 150, n. 4, p. 1663-1682, dec. 1998.

MOORE, G.; DEVOS, K. M.; WANG, Z.; GALE, M. D. Cereal Genome Evolution: Grasses, line up and form a circle. **Current biology**, Maryland Heights, v. 5, n. 7, p. 737-739, july 1995.

MORGANTE, M.; PAOLI, E.; RADOVIC, S. Transposable elements and the plant pan-genomes. **Current Opinion in Plant Biology**, Philadelphia, v. 10, n. 2, p. 149-155, feb. 2007.

MUKHERJEE, S. K. Origin and distribution of *Saccharum*. **Botanical Gazette**, Chigago, v. 119, n. 1, p. 55-61, feb. 1957.

NOGUEIRA, F. T. S.; DE ROSA JR., V. E.; MENOSSI, M.; ULIAN, E. C.; ARRUDA, P. RNA expression profiles and data mining of sugarcane response to low temperature. **Plant Physiology**, Waterbury, v. 132, n. 4, p. 1811–1824, oct. 2003.

PAN, Y. B.; BURNER, D. M.; LEGENDRE, B. L. An Assessment of the Phylogenetic Relationship Among Sugarcane and Related Taxa Based on the Nucleotide Sequence of 5S rRNA Intergenic Spacers. **Genetica**, Amsterdam, v. 108, n. 3, p. 285-295, oct. 2000.

PAPINI-TERZI, F. S.; FELIX, J. M.; ROCHA, F. R.; WACLAWOVSKY, A. J.; ULIAN, E. C.; CHABREGAS, S. M.; FALCO, M. C.; NISHIYAMA JR. M. Y.; VÊNCIO, R. Z. N.; VICENTINI, R.; MENOSSI, M.; SOUZA, G. M. The SUCEST-FUN project: identifying genes that regulate sucrose content in sugarcane plants. **Proc. Int. Soc. Sugar Cane Technol**., v. 26, p. 1-10, 2007.

PAPINI-TERZI, F. S.; ROCHA, F. R.; VÊNCIO, R. Z.; FELIX, J. M.; BRANCO, D. S.; WACLAWOVSKY, A. J.; DEL BEM, L. E.; LEMBKE, C. G.; COSTA, M. D.; NISHIYAMA JR. M. Y.; VICENTINI, R.; VINCENTZ, M. G.; ULIAN, E. C.; MENOSSI, M.; SOUZA, G. M. Sugarcane genes associated with sucrose content. **BMC Genomics**, London, v. 10, n. 120, p. 1-21, mar. 2009.

PATERSON, A. H.; BOWERS, J. E.; BRUGGMANN, R.; DUBCHAK, I.; GRIMWOOD, J.; GUNDLACH, H.; HABERER, G.; HELLSTEN, U.; MITROS, T.; POLIAKOV, A. SCHMUTZ, J.; SPANNAGL, M.; TANG, H.; WANG, X.; WICKER, T.; BHARTI, A. K.; CHAPMAN, J.; FELTUS, F. A.; GOWIK, U.; GRIGORIEV, I. V.; LYONS, E.; MAHER, C. A.; MARTIS, M. A.; NARECHANIA, A.; OTILLAR, R. P.; PENNING, B. W.; SALAMOV, A. S.; WANG, Y.; ZHANG, L.; CARPITA, N. C.; FREELING, M.; GINGLE, A. R.; HASH, T. A.; KELLER, B.; KLEIN, P.; KRESOVICH, S.; CCANN, M. C.; MING, R.; PETERSON, D. G.; MEHBOOB-UR-RAHMAN, WARE, D.;

WESTHOFF, P.; MAYER, K. F. X.; MESSING, J.; ROKHSAR, D. S. The *Sorghum bicolor* genome and the diversification of grasses. **Nature**, New York, v. 457, n. 7229, p. 551-556, jan. 2009.

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

PRICE, S. **Cytology of Saccharum robustum and related symapatric species and natural hybrids**. Washington: US Dept. of Agriculture, 1965. 57p.

ROACH, B. T. Sugar canes. In: SMARTT, J.; SIMMONDS, N. W. (Ed.). **Evolution of crop plants**. 2. ed. Singapore: Longman Scientific & Technical, 1995. p. 160-166.

ROCHA, F. R.; PAPINI-TERZI, F. S.; NISHIYAMA JR. M. Y. Signal transduction-related responses to phytohormones and environmental challenges in sugarcane. **BMC Genomics**, London, v. 8, p. 71-80, 2007.

ROSSI, M.; ARAUJO, P. G.; VAN SLUYS, M. Survey of transposable elements in sugarcane expressed sequence tags (ESTs). **Genetics and Molecular Biology**, São Paulo, v. 24, n. 1-4, p. 147-154, dec. 2001

SALSE, J.; BOLOT, S.; THROUDE, M. L.; JOUFFE, V.; PIEGU, V. T.; QURAISHI, U. M.; CALCAGNO, T.; COOKE, R.; DELSENY, M.; FEUILLET, C. Identification and characterization of shared duplications between rice and wheat provide new insight into grass. **The Plant Cell**, Waterbury, v. 20, n.1, p.11-24, jan. 2008.

SALZBERG, S. L.; PHILLIPPY, A. M.; ZIMIN, A.; PUIU, D.; MAGOC, T.; KOREN, S.; TREANGEN, T. J.; SCHATZ, M. C.; DELCHER, A. L.; ROBERTS, M.; MARCXAIS, G.; POP, M.; YORKE, J. A. GAGE: A critical evaluation of genome assemblies and assembly algorithms. **Genome Research**, New York, v. 22, p. 557–567, 2011.

SCHNABLE, P. S. et al. The B73 maize genome: complexity, diversity, and dynamics. **Science**, Washington, v. 326, n. 1112, p. 1112-1115, nov. 2009.

SCREENIVASAN, T. V.; AHLOOWALIA, B. S.; HEIZ, D. J. Cytogenetics. In: HEINZ, D. (Ed.). **Sugarcane improvement through breeding**. Amsterdam: Elsevier, 1987. p. 211-253.

SHANKAR, P. C.; ITO, S.; KATO, M.; MATSUI, M.; KODAIRA, R.; HAYASHIDA, N.; OKAZAKI, M. Analysis of Tag1-Like elements in *Arabidopsis thaliana* and their distribution in other plants. **DNA Research**, Oxford, v. 8, n. 3, p. 107-113, may 2001.

TAKAHASHI, S.; FURUKAWA, T.; ASANO, T.; TERAJIMA, Y.; SHIMADA, H.; SUGIMOTO, A.; KADOWAKI, K. Very close relationship of the chloroplast genomes among Saccharum species. **Theoretical and Applied Genetics**, Heidelberg, v. 110, n. 8, p. 1523-1529, apr. 2005.

TAMURA, K.; PETERSON, D.; PETERSON, N.; STECHER, G.; MASATOSHI, N.; KUMAR, S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. **Mol. Biol. Evol**., Oxford, v. 28, n. 10, p. 2731–2739, may 2011.

TODOROVSKA, E. Retrotransposons and their role in plant – Genome evolution. **Biotechnol. & Biotechnol**., Sofia, v. 21, n. 3, p. 294-305, mar. 2007.

TRAPNELL, C.; ROBERTS, A.; GOFF, L.; PERTEA, G.; KIM, D.; KELLEY, D. R.; PIMENTEL, H.; SALZBERG, S. L.; RINN, J. L.; PACHTER, L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. **Nature Protocols**, New York, v. 7, n. 3, p. 562-578, 2012.

TSAY, Y. F.; FRANK, M. J.; PAGE, T.; DEAN, C.; CRAWFORD, N. M. Identification of a mobile endogenous transposon in *Arabidopsis thaliana*. **Science**, Whashington, v. 260, n. 5106, p.342-344, apr. 1993.

VETTORE, A. L.; SILVA, F. R.; KEMPER, E. L.; SOUZA, G. M.; SILVA, A. M. ; FERRO, M. I. T.; HENRIQUE-SILVA, F.; GIGLIOTI, E. A.; LEMOS, M. V. F.; COUTINHO, L. L.; NOBREGA, M. P.; CARRER, H.; FRANÇA, S. C.; BACCI JR., M.; GOLDMAN, M. H. S.; GOMES, S. L.; NUNES, L. R.; CAMARGO, L. E. A.; SIQUEIRA, W. J.; VAN SLUYS, M.; THIEMANN, O. H.; KURAMAE, E. E.; SANTELLI, R. V.; MARINO, C. L.; TARGON, M. L. P. N.; FERRO, J. A.; SILVEIRA, H. C. S.; MARINI, D. C.; LEMOS, E. G. M.; MONTEIRO-VITORELLO, C. B.; TAMBOR, J. H. M.; CARRARO, D. M.; ROBERTO, P. G.; MARTINS, V. G.; GOLDMAN, G. H.; OLIVEIRA, R. C.; TRUFFI, D.; COLOMBO, C. A.; ROSSI, M.; ARAUJO, P. G.; SCULACCIO, S. A.; ANGELLA, A.; LIMA, M. M. A.; ROSA JR., V. E.; SIVIERO, F.; COSCRATO, V. E.; MACHADO, M. A.; GRIVET, L.; DI MAURO, S. M. Z.; NOBREGA, F. G.; MENCK, C. F. M.; BRAGA, M. D. V.; TELLES, G. P.; CARA, F. A. A.; PEDROSA, G.; MEIDANIS, J.; ARRUDA, P. Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. **Genome Research**, New York, v. 13, p. 2725–2735, nov. 2003.

VICENTINI, R.; MENOSSI, M. Pipeline for macro and microarray analyses. **Brazilian Journal of Medical and Biological Research**, v. 40, n. 5, p. 615–619, 2007.

WANG, J.; ROE, B.; MACMIL, S.; YU, Q.; MURRAY, J. E.; TANG, H.; CHEN, C.; NAJAR, F.; WILEY, G.; BOWERS, J. Microcollinearity between autopolyploid sugarcane and diploid sorghum genomes. **BMC genomics**, London, v. 11, n. 1, p. 261, apr. 2010.

WENDEL, J. F. Genome evolution in polyploids. **Plant Molecular Biology**, Netherlands, v. 42, n. 1, p. 225-249, jan. 2000.

WICKER, T.; SABOT, F.; HUA-VAN, A.; BENNETZEN, J. L.; CAPY, P.; CHALHOUB, B.; FLAVELL, A.; LEROY, P.; MORGANTE, M.; PANAUD, O.; PAUX, E.; SANMIGUEL, P.; SCHULMAN, A. H. A unified classification system for eukaryotic transposable elements. **Nature reviews Genetics**, New York, v. 8, n. 12, p. 973-982, dec. 2007.

XIONG, Y.; EICKBUSH, T. H. Origin and evolution of retroelements based upon their reverse transcriptase sequences. **The EMBO Journal**, Malden, v. 9, n. 10, p. 3353-3362, oct. 1990.

ZHANG, J.; ARRO, J.; CHEN, Y.; MING, R. Haplotype analysis of sucrose synthase gene family in three *Saccharum* species. **BMC Genomics**, London, v. 14, n. 314, p. 1-14, may 2013.

ZHANG, J.; NAGAI, C.; YU, Q.; PAN, Y.-B.; AYALA-SILVA, T.; SCHNELL, R.; COMSTOCK, J.; ARUMUGANATHAN, A.; MING, R. Genome size variation in three *Saccharum* species. **Euphytica**, Wageningen, v. 185, n. 3, p. 511-519, mar. 2012a.

ZHANG, G.; LIU, X.; QUAN, Z.; CHENG, S.; XU, X.; PAN, S.; XIE, M.; ZENG, P.; YUE, Z.; WANG, W.; TAO, Y.; BIAN, C.; HAN, C.; XIA, Q.; PENG, X.; CAO, R.; YANG, X.; ZHAN, D.; HU, J.; HANG, Y.; LI, H.; LI, H.; LI, N.; WANG, J.; WANG, C.; WANG, R.; GUO, T.; CAI, Y.; LIU, C.; XIANG, H.; SHI, Q.; HUANG, P.; CHEN, Q.; LI, Y.; WANG, J.; ZHAO, Z.; WANG, J. Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. **Nature biotechnology**, New York, v. 30, n. 6, p. 549-554, jun. 2012b.

ZIMIN, A. V.; MARC¸ G.; PUIU, D.; ROBERTS, M.; SALZBERG, S. L.; YORKE, J. A. The MaSuRCA genome assembler. **Bioinformatics**, Oxford, v. 6, n. 17, p. 1-9, sep. 2013.

APÊNDICE

Apêndice A. Anotação dos genes presente nos oito BACs de RB867515, de acordo com a ordem dos genes nas sequências, e os diferentes transcritos (quando presentes) para o mesmo gene

	Genes	Transcritos/ Isoformas	Descrição
BAC 1	3	1	No hit
	4	1	Hyphotetical protein
	5	1	No hit
	1	1	NADP-dependent D-sorbitol-6-phosphate dehydrogenase [Saccharum hybrid cultivar]
	6	1	No hit
	7	1	No hit
	8	1	Hyphotetical protein
	2	1	Hyphotetical protein
	9	1	No hit
	10	1	No hit
	11	1	No hit
	12	1	No hit
	13	1	No hit
	14	1	No hit
	15	1	No hit
	16	1	No hit
	17	1	No hit

	Genes	Transcritos/ Isoformas	Descrição
BAC 2	3	1	No hit
	4	1	No hit
	5	1	Hyphotetical protein
	6	1	Hyphotetical protein
	7	1	No hit
	8	1	No hit
	9	1	No hit
	10	1	Retrotransposon/LTR-Gypsy-Ty3
	11	1	No hit
	1	1	NADP-dependent D-sorbitol-6-phosphate deshydrogenase [Saccharum hybrid cultivar R570]
		2	NADP-dependent D-sorbitol-6-phosphate deshydrogenase [Saccharum hybrid cultivar R570]
	12	1	No hit

continua...

Apendice A. Continuação

BAC 2	13	1	NADP-dependent D-sorbitol-6-phosphate deshydrogenase [Saccharum hybrid cultivar R570]
	14	1	Gag-pol polyproteina
	15	1	Hyphotetical protein
	16	1	Hyphotetical protein
	2	1	Retrotransposons/ LTR-Copia-Ty1
		2	Retrotransposons/ LTR-Copia-Ty1
	17	1	Endoglucanase 4 precursor [Saccharum hybrid cultivar R570]

	Genes	**Transcritos/ Isoformas**	**Descrição**
BAC 3	9	1	No hit
	3	1	Putative ulp1 protease [Saccharum hybrid cultivar R570]
		2	Putative ulp1 protease [Saccharum hybrid cultivar R570]
		3	Putative ulp1 protease [Saccharum hybrid cultivar R570]
	4	1	Conserved hypothetical protein [Saccharum hybrid cultivar R570]
		2	Conserved hypothetical protein [Saccharum hybrid cultivar R570]
	5	1	Hyphotetical protein
	6	1	NADP-dependent D-sorbitol-6-phosphate deshydrogenase [Saccharum hybrid cultivar R570]
	15	1	Endoglucanase 4 precursor (Sh142J21g280-g330 modules) [Saccharum hybrid cultivar R570]
	17	1	Hyphotetical protein
	18	1	Hyphotetical protein
	19	1	Transcriptase [Saccharum hybrid cultivar R570]
	20	1	No hit
	7	1	No hit
	2	1	Hyphotetical protein
		2	Hyphotetical protein
	8	1	Hyphotetical protein
		2	Hyphotetical protein
		3	Hyphotetical protein
		4	Hyphotetical protein
		5	Hyphotetical protein
		6	Hyphotetical protein
		7	Hyphotetical protein

Continua...

Apendice A. Continuação

	Genes	Transcritos/ Isoformas	Descrição
BAC 4	1	1	Putative shrunken seed protein [Saccharum hybrid cultivar R570]
		2	Putative shrunken seed protein [Saccharum hybrid cultivar R570]
	6	1	transcriptase [Saccharum hybrid cultivar R570]
	7	1	Hyphotetical protein
	8	1	Retrotrasnposon/LTR-Copia-Ty1
	3	1	Conserved protein
	4	1	NADP-dependent D-sorbitol-6-phosphate deshydrogenase [Saccharum hybrid cultivar R570]
	9	1	Endoglucanase 4 precursor
	11	1	No hit
	12	1	No hit
	13	1	No hit
	14	1	No hit
	15	1	Hyphotetical protein
	16	1	No hit
	2	1	No hit
		2	No hit
	17	1	Putative trehalose-phosphatase (C-terminal fragment) [Saccharum hybrid cultivar R570]
	18	1	No hit
	19	1	Subtilisin-like protease [Saccharum hybrid cultivar R570]
	20	1	Putative transposon element
	5	1	Hyphotetical protein
		2	Hyphotetical protein
		3	Hyphotetical protein
	21	1	No hit
	22	1	No hit

	Genes	Transcritos/ Isoformas	Descrição
BAC 5	4	1	Putative ulp1 protease [Saccharum hybrid cultivar R570]
		2	Putative ulp1 protease [Saccharum hybrid cultivar R570]
	9	1	Hyphotetical protein
	10	1	Hyphotetical protein
	5	1	No hit

Continua...

Apendice A. Continuação

BAC 5	2	1	No hit
		2	No hit
		3	No hit
		4	Hyphotetical protein
	6	1	Hyphotetical protein
		2	Hyphotetical protein
	11	1	No hit
	3	1	No hit
		2	No hit
	12	1	No hit
	13	1	Hyphotetical protein
	14	1	No hit
	7a	1	No hit
	7b	1	No hit
	8	1	NADP-dependent D-sorbitol-6-phosphate deshydrogenase [Saccharum hybrid cultivar R570
	16	1	Hyphotetical protein
	17	1	No hit
	18	1	No hit
		2	No hit

	Genes	**Transcritos/ Isoformas**	**Descrição**
BAC 6	3	1	Retrotransposon putative Copia-Ty1
	4	1	No hit
	5	1	No hit
	6	1	No hit
	1	1	putative ulp1 protease
	7	1	No hit
	8	1	No hit
	9	1	No hit
	10	1	No hit
	11	1	Hyphotetical protein
	12	1	No hit
	2	1	Hyphotetical protein

	Genes	**Transcritos/ Isoformas**	**Descrição**
BAC 7	3	1	No hit
	4	1	Hyphotetical protein
	5	1	Putative retrotransposon
	6	1	Retrotranposon protein putative LTR/Copia-Ty1
	7	1	Retrotranposon protein putative LTR/Copia-Ty1

Continua...

Apêndice A. Continuação

	8	1	Retrotranposon protein putative LTR/Copia-Ty1
	9	1	Hyphotetical protein
	10	1	NADP-dependent D-sorbitol-6-phosphate dehydrogenase [Saccharum hybrid cultivar]
	1	1	Hyphotetical protein transpcriptase reversa
	11	1	No hit
	2a	1	Retrovirus-related Pol polyprotein LINE-1 [Triticum urartu]
		2	Retrovirus-related Pol polyprotein LINE-1 [Triticum urartu]
	2b	1	Retrovirus-related Pol polyprotein LINE-1 [Triticum urartu]
		2	Retrovirus-related Pol polyprotein LINE-1 [Triticum urartu]
		3	Retrovirus-related Pol polyprotein LINE-1 [Triticum urartu]
BAC 7	2c	1	NADP-dependent D-sorbitol-6-phosphate dehydrogenase [Saccharum hybrid cultivar]
		2	NADP-dependent D-sorbitol-6-phosphate dehydrogenase [Saccharum hybrid cultivar]
		3	NADP-dependent D-sorbitol-6-phosphate dehydrogenase [Saccharum hybrid cultivar]
		4	NADP-dependent D-sorbitol-6-phosphate dehydrogenase [Saccharum hybrid cultivar]
	13	1	No hit
	14	1	Hyphotetical protein
	15	1	Hyphotetical protein
	16	1	No hit
	17	1	No hit
	18	1	No hit
	19	1	No hit
	20	1	No hit
	21	1	Hyphotetical protein
	22	1	Retrotransposon LTR/Copia-Ty1
	23	1	Retrotransposon LTR/Copia-Ty1
	24	1	Hyphotetical protein
	25	1	No hit

	Genes	**Transcritos/ Isoformas**	**Descrição**
BAC 8	3	1	hypotetical protein
	1a	1	No hit
		2	No hit
	1b	1	No hit

Continua...

Apêndice A. Continuação

BAC 8	2a	1	Transporter [Saccharum hybrid cultivar R570]
		2	Transporter [Saccharum hybrid cultivar R570]
	2b	1	Hypothetical protein ShCIR9O20g_040 [Saccharum hybrid cultivar]
	4	1	Hyphotetical protein
	5	1	Hyphotetical protein
	6	1	Retrotransposon LTR/Gypsy-Ty3
	7	1	Transporter [Saccharum hybrid cultivar R570]
	8	1	No hit
	9	1	NADP-dependent D-sorbitol-6-phosphate dehydrogenase [Saccharum hybrid cultivar R570]