UFG

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

MONTAGEM E CARACTERIZAÇÃO DO TRANSCRITOMA DE CANA-DE-AÇÚCAR (*Saccharum* spp.) UTILIZANDO DADOS DE SEQUENCIAMENTO DE NOVA GERAÇÃO

ARTHUR TAVARES DE OLIVEIRA MELO

Orientador:
Prof. Dr. Alexandre Siqueira Guedes Coelho

Goiânia - GO
Brasil

Janeiro - 2015

ARTHUR TAVARES DE OLIVEIRA MELO

MONTAGEM E CARACTERIZAÇÃO DO TRANSCRITOMA DE CANA-DE-AÇÚCAR (*Saccharum* spp.) UTILIZANDO DADOS DE SEQUENCIAMENTO DE NOVA GERAÇÃO

Tese 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 Doutor em Genética e Melhoramento de Plantas.

Orientador:
Prof. Dr. Alexandre Siqueira Guedes Coelho

Goiânia, GO - Brasil
2015

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR AS TESES E DISSERTAÇÕES ELETRÔNICAS (TEDE) NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), sem ressarcimento dos direitos autorais, de acordo com a Lei nº 9610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou *download*, a título de divulgação da produção científica brasileira, a partir desta data.

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

2. Identificação da Tese ou Dissertação

Autor (a):	Arthur Tavares de Oliveira Melo					
E-mail:	arthurmelobio@gmail.com					
Seu e-mail pode ser disponibilizado na página? [X]Sim [ ] Não						
Vínculo empregatício do autor						
Agência de fomento:		CAPES			Sigla:	
País:	Brasil	UF:		CNPJ:		
Título:	Montagem e caracterização do transcritoma de cana-de-açúcar (*Saccharum* spp.) utilizando dados de sequenciamento de nova geração					
Palavras-chave:	*Saccharum* spp.; transcritoma, RNA-seq; Trinity					
Título em outra língua:	Assembly and characterization of sugarcane (*Saccharum* spp.) transcriptome using next generation sequencing data					
Palavras-chave em outra língua:	*Saccharum* spp.; transcriptome; RNA-seq; Trinity					
Área de concentração:	Genética e Melhoramento de Plantas					
Data defesa: (dd/mm/aaaa)	22/01/2015					
Programa de Pós-Graduação:	Genética e Melhoramento de Plantas					
Orientador (a):	Dr. Alexandre Siqueira Guedes Coelho					
E-mail:	asgcoelho@gmail.com					
Co-orientador (a):*						
E-mail:						

*Necessita do CPF quando não constar no SisPG

3. Informações de acesso ao documento:

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

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

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

Arthur T. de O. Melo
Assinatura do (a) autor (a)

Data: 15 / 05 / 2015

[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.

ARTHUR TAVARES DE OLIVEIRA MELO

TÍTULO: **"Montagem e caracterização do transcritoma de cana-de-açúcar (*Saccharum* spp.) utilizando dados de sequenciamento de nova geração".**

Tese DEFENDIDA em 22 de Janeiro de 2015, e APROVADA pela Banca

Examinadora constituída pelos membros:

Prof. Dr. Alexandre Siqueira Guedes Coelho
Orientador – EA/UFG

Drª. Georgios Joannis Pappas Júnior
ICB/UNB

Dr. Tereza Cristina de Oliveira Borba.
Embrapa Arroz e Feijão

Prof. Dr. Claudio Brondani
Embrapa Arroz e Feijão

Prof. Dr. Evandro Novaes
EA/UFG

“A coisa mais bela que podemos experimentar é o mistério. É a fonte de toda arte verdadeira e ciência.”

Albert Einstein

“Acho que é muito difícil lidar com os fatos, Holmes, sem nos perdermos atrás de teorias e fantasias.”

Inspetor Lestrade para Sherlock Holmes
O mistério do Vale Boscombe

AGRADECIMENTOS

Primeiramente, gostaria de agradecer às instituições que financiaram a execução deste trabalho. À Petrobras Biocombustíveis pela disponibilidade de recursos financeiros, à Capes, pela bolsa de doutorado concedida e à Universidade Federal de Goiás (UFG), pela infraestrutura e apoio no desenvolvimento da pesquisa.

Ao Programa de Pós-Graduação em Genética e Melhoramento de Plantas da UFG, em especial à coordenadora do Programa Dra. Mariana Pires de Campos Telles. A todos os professores do programa que estão envolvidos no crescimento e na excelente qualidade das atividades científico-acadêmicas do Programa, meu muito obrigado.

Aos membros da banca examinadora: Dra. Tereza Borba, Prof. Dr. Georgios Pappas, Dr. Claudio Brondani, Prof. Dr. Evandro Novaes, e em especial meu orientador Dr. Alexandre Siqueira Guedes Coelho, pelas valiosas contribuições para a finalização deste trabalho.

Um agradecimento especial à minha família e mais especial ainda aos meus pais, Marinei Jane de Melo e Newton Tavares de Oliveira, por acreditarem no meu potencial, pela confiança, pela condição e pelo exemplo de dedicação ao trabalho. Não há palavras que descrevem o quanto sou grato a vocês dois!

Um agradecimento também muito especial à Fernanda Ramos Cyríaco, minha eterna companheira.

Ao meu orientador, Dr. Alexandre Siqueira Guedes Coelho eu quero agradecer imensamente pelo exemplo de profissional acadêmico e pelos inúmeros ensinamentos científicos ao longo da graduação e da pós-graduação. Agradeço também por todas as correções feitas a este trabalho.

Ao pessoal que participou direta ou indiretamente dando apoio nas várias etapas de condução do trabalho. Agradeço especialmente à Dra. Ludmila Ferreira Bandeira e à Stela Barros Ribeiro pelas excelentes extrações de RNA. À Dra. Rosana Pereira Vianello e ao Dr. Claudio Brondani da Embrapa Arroz e Feijão, pelo empréstimo do equipamento de análise de qualidade do RNA extraído. Ao Professor Dr. Cirano Ulhoa por ceder cordialmente equipamentos do seu laboratório. Sem vocês este trabalho não poderia ser concluído.

A todos os amigos, professores (em especial ao Dr. Evandro Novaes pelas discussões e ensinamentos de bioinformática) e companheiros do Setor de Melhoramento de Plantas da Escola de Agronomia da UFG, um muito obrigado!

SUMÁRIO

RESUMO GERAL

MELO, A.T.O. **Montagem e caracterização do transcritoma de cana-de-açúcar (*Saccharum* spp.) utilizando dados de sequenciamento de nova geração.** 2015. 102 f. Tese (Doutorado em Genética e Melhoramento de Plantas) – Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2015. [2]

A cana-de-açúcar é uma das principais espécies cultivadas para o fornecimento mundial de açúcar e energia renovável. Devido à elevada quantidade de elementos repetitivos e os vários eventos de poliploidização, o genoma da espécie ainda não foi montado e anotado, diferentemente de outras espécies de interesse agronômico. Assim, as informações do transcritoma da espécie se tornam ainda mais úteis por dar suporte ás iniciativas de análises genômicas. Um *draft assembly* do transcritoma de cana-de-açúcar foi montado a partir do sequenciamento Illumina de bibliotecas *paired-ends* de cinco órgãos distintos da planta, obtidos de uma amostra de trinta clones elite. Os dados de RNA-seq passaram por análises de controle de qualidade e normalização. O software Trinity foi utilizado para montagem *de novo* do transcritoma. Os *scaffolds* obtidos identificados como *ORFs* completas foram anotados conforme os termos do *Gene Ontology*. O *draft assembly* obtido para o transcritoma de cana-de-açúcar foi caracterizado pela identificação de marcadores moleculares do tipo microssatélites e SNPs e pela avaliação da contribuição dos diferentes órgãos vegetais para constituição final do transcritoma. O transcritoma obtido compreendeu 178 Mb, distribuídos em 131.831 *scaffolds*, representando 61.225 genes. O tamanho médio dos transcritos foi de 1.350 pb, com valor de N50 igual a 1.667 pb. Um total de 1.250 transcritos, identificados como *ORFs* completas, não apresentaram similaridade com sequências do banco de dados nr do NCBI, sendo considerados novas regiões transcricionalmente ativas (nTARs). A anotação realizada através do banco de dados do KEGG identificou 234 transcritos codificantes para enzimas integrantes do metabolismo de sacarose e amido, uma importante rota metabólica para compreensão da relação entre taxa fotossintética e o acúmulo de sacarose no colmo. Os cinco órgãos vegetais utilizados contribuíram igualmente para a constituição do *draft* do transcritoma de cana-de-açúcar. Foram identificadas 12.931 regiões genômicas contendo microssatélites perfeitos, com predomínio de di e tri nucleotídeos. Em média, identificou-se um SNP a cada 18 pares de bases, com mais de quatro milhões de SNPs identificados. A diversidade nucleotídica dos trinta clones elites utilizados é elevada. A identificação destes marcadores moleculares, principalmente os marcadores SNPs, fornece a possibilidade de utilização destes polimorfismos em estudos genéticos e genômicos de cana-de-açúcar, incluindo o emprego em abordagens como seleção genômica ampla no melhoramento da espécie. O *draft assembly* do transcritoma de cana-de-açúcar proposto neste estudo possui qualidade de dados e de análise suficiente para ser utilizado na tentativa de abranger um transcritoma de referência para as espécies de *Saccharum* spp.

Palavras chave: *Saccharum* spp.; transcritoma; RNA-seq; Trinity

[2] Orientador: Prof. Dr. Alexandre Siqueira Guedes Coelho

GENERAL ABSTRACT

MELO, A.T.O. **Assembly and characterization of sugarcane (*Saccharum* spp.) transcriptome using Next Generation Sequencing data.** 2015. 102 f. Tese (Doutorado em Genética e Melhoramento de Plantas) – Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2015. [3]

The sugarcane is one of the most important crop species to provide sugar and renewable energy in the world. Due to the high amount of repetitive elements and the various polyploidization events suffer during its evolution, the *Saccharum* spp. genome has not yet been assembled and annotated, unlike other agronomic important species. So, the knowledge about sugarcane transcriptome become even more useful for supporting genomic analyzes studies. A draft assembly of sugarcane transcriptome was obtained from Illumina sequencing paired-ends libraries of five different plant organs, sampled from thirty elite clones. Analyzes of quality control and normalization was done in the RNA-seq data. Trinity package was used for *de novo* assembly. The scaffolds obtained and identified as complete ORFs were annotated according to Gene Ontology terms. The draft assembly was characterized by the identification of microsatellites and SNPs molecular markers and for assessing the contribution of different plant organs for transcriptome final assembly. The draft sugarcane transcriptome comprised 178 Mb, over 131,831 scaffolds, representing 61,225 genes. The transcripts average size was 1,350 bp and N50 value was 1,667 bp. A total of 1,250 transcripts identified as complete ORFs showed no similarity to sequences of the nr NCBI database, are considered new Transcript Active Regions (nTARs). The annotation performed using the KEGG database identified 234 transcripts coding for enzymes members of sucrose and starch metabolism, an important metabolic pathway for understanding the relationship between photosynthetic rate and sucrose accumulation in the stalk. The five plant organs used contributed equally for the draft sugarcane transcriptome. A total of 12,931 genomic regions were identified containing perfect microsatellites, with a predominance of di and tri nucleotide. On average, one SNP every 18 bp was identified, with more than four million SNPs identified with satisfactory values of haplotype and quality scores. The nucleotide diversity of thirty elite clones used in this study was high. The identification of these molecular markers, particularly SNPs markers, provides the possibility of using these polymorphisms in genomic and genetic studies of sugarcane, including the possibility of application of genome wide selection like breeding strategy. The sugarcane transcriptome draft assembly proposed in this study has data and analysis quality sufficient to be used in attempt to encompass a reference transcriptome for the species of *Saccharum* spp.

Key-words: *Saccharum* spp.; transcriptome; RNA-seq; Trinity;

[3] Adviser: Prof. Dr. Alexandre Siqueira Guedes Coelho

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO GERAL

A cana-de-açúcar (*Saccharum* spp.) é a espécie cultivada mais importante para o fornecimento mundial de açúcar e energia (Henry, 2010). Ocorre, nos últimos anos, um elevado crescimento anual de área cultivada nas regiões tropicais e subtropicais em todo o mundo. A produção brasileira de cana-de-açúcar foi, no ano de 2013, bem maior que a soma da produção dos outros quatro países maiores produtores (Índia, China, Tailândia e Paquistão) (FaoStats, 2013). Atualmente, o Brasil é o país de maior produção mundial e lidera o mercado de etanol e açúcar derivados de cana-de-açúcar, em que se estima que mais da metade do açúcar comercializado no mundo seja de produção brasileira (MAPA, 2013). A produção mundial de cana-de-açúcar, na safra de 2013/14, foi de 1,8 bilhões de toneladas, sendo 658,8 milhões de toneladas produzidas somente no Brasil, o que corresponde a aproximadamente 35% da produção mundial, cultivados em mais de nove milhões de hectares em território brasileiro (FaoStats, 2013).

O genoma das cultivares modernas de cana-de-açúcar é grande e complexo, formado pelo cruzamento interespecífico de dois táxons próximos e silvestres (*Saccharum officinarum* x *Saccharum spontaneum*). *Saccharum* spp. é considerada a espécie cultivada que produz a maior quantidade de produto na colheita, devido ao seu mecanismo fotossintético C4 que converte, com muita eficiência, moléculas de carbono em biomassa (Henry, 2010). Estima-se que sejam colhidas, anualmente, cerca de dois bilhões de toneladas de cana-de-açúcar em todo o mundo, enquanto que os valores médios da colheita de alguns grãos como soja, milho e trigo ficam em torno das 600 milhões de toneladas anuais. A cana-de-açúcar é a principal espécie cultivada utilizada para o abastecimento energético (etanol e eletricidade) (Tew & Cobill, 2008), de açúcar (Cordeiro et al., 2007) e para o mercado de fibras (Zandersons et al., 1999; Lavarack et al., 2002).

O uso das ferramentas genético-moleculares no auxílio ao melhoramento de espécies cultivadas tem crescido no decorrer das duas últimas décadas. Atualmente,

estamos inseridos na era genômica, pois o uso destas ferramentas acontece em grande escala no melhoramento genético das mais diversas espécies cultivadas. Um exemplo claro é a crescente utilização dos marcadores SNPs (*Single Nucleotide Polymorphisms*) na construção de mapas genéticos e no emprego das técnicas de seleção assistida por marcadores moleculares (MAS – *Marker Assisted Selection*) e seleção genômica ampla (*WGS - Whole Genome Selection*). Outro exemplo do atual nível de desenvolvimento tecnológico nas áreas da genética e biologia molecular voltadas para o melhoramento vegetal é a possibilidade de sequenciamento e/ou ressequenciamento, por completo, do genoma de uma espécie em tempo reduzido e a preços cada vez mais baixos. Com isso, perguntas acadêmicas para compreensão dos padrões genético-populacionais mudaram de escala. Não se trata mais de inferências paramétricas com base na caracterização genética de poucos locos, mas sim de estimativas populacionais dos parâmetros de interesse com informações de milhares de locos distribuídos no genoma. Tal fato permite um entendimento mais profundo e detalhado a respeito da estrutura e composição dos genomas, da identificação de polimorfismos de interesse agronômico, sobre o comportamento da expressão diferencial dos genes transcritos em diferentes condições ambientais e a respeito das interações das vias metabólicas que controlam mecanismos de resposta aos estresses biótico e abiótico.

Neste contexto, se desenvolveram nos últimos dez anos, as plataformas de sequenciamento de DNA/RNA de nova geração. Tratam-se, na grande maioria, do uso de micro e/ou nano tecnologias com a finalidade de sequenciar em larga escala fragmentos relativamente pequenos de DNA e obter Gigabases de sequência do genoma ou do transcritoma de uma espécie (Schuster, 2008). O aumento da capacidade em sequenciar o DNA e produzir um grande volume de informação genética desencadeou uma mudança de paradigma na área da genômica, permitindo estudos genéticos com resoluções no nível de pares de bases. Entre estes estudos incluem-se: o ressequenciamento completo de genomas ou o sequenciamento *de novo*; a identificação de polimorfismos nucleotídicos ao longo do genoma e/ou transcritoma; o mapeamento das mutações; a compreensão dos padrões de metilação do DNA e das modificações no posicionamento das histonas; o sequenciamento do transcritoma; o descobrimento e a análise da expressão diferencial de genes; a identificação de *splicings* alternativos e a análise dos perfis de expressão de *small* RNAs e das interações DNA – proteínas e proteínas – proteínas (Lyster et al., 2009).

Transcritoma é o conjunto completo de transcritos de uma célula e sua quantificação em um estágio específico de condições fisiológicas (Wang et al., 2009). A técnica de sequenciamento de RNA (RNA-seq) é uma abordagem recente que utiliza o sequenciamento de elevada cobertura dos mRNAs ou cDNAs com o objetivo de compreender o perfil do transcritoma de uma espécie (Lyster et al., 2009, Haas & Zody, 2010; Nagalakshmi et al., 2010). O tratamento dos dados produzidos pode ser iniciado a partir da disponibilidade prévia de um genoma e/ou transcritoma de referência. Caso não exista esta referência, as estratégias de bioinformática utilizadas no tratamento dos dados são outras e a análise passa a ser caracterizada como uma montagem *de novo*, produzindo assim, um genoma e/ou transcritoma de referência. A compreensão do transcritoma de uma espécie, por exemplo, tem auxiliado na interpretação dos elementos funcionais do genoma e revelado os constituintes moleculares de células e tecidos. A mudança de escala para um nível de identificação de polimorfismos nucleotídicos permitiu uma melhor compreensão da complexidade dos transcritos dos eucariotos, de modo que as análises de RNA-Seq estão revolucionando a maneira como os transcritomas de eucariotos são analisados (Wang et al., 2009; Groba & Burgos, 2010; Garber et al., 2011).

Neste contexto, o presente trabalho tem como objetivo utilizar sequências genômicas obtidas pelo sequenciamento de nova geração de moléculas de mRNA provenientes de diferentes órgãos vegetais amostrados de 30 clones elites, para montar, através da estratégia *de novo*, um *draft assembly* do transcritoma de cana-de-açúcar (*Saccharum* spp.). Além disso, objetivou-se a anotação funcional deste transcritoma e sua caracterização.

2 REVISÃO BIBLIOGRÁFICA

2.1 A CULTURA DA CANA-DE-AÇÚCAR

Trata-se de uma cultura perene e subtropical. A cana-de-açúcar é uma gramínea pertencente à família Poaceae. A família das gramíneas (Poaceae), pertencente ao grupo das Monocotiledôneas, é dividida em três subfamílias. O grupo das Panicoidae, formado por sorgo (*Sorghum bicolor*), milho (*Zea mays*) e cana-de-açúcar (*Saccharum* spp.), a subfamília Ehrhartoideae formada pelo arroz (*Oryza sativa*) e a subfamília Pooideae formada pela espécie *Brachypodium distachyon*. O gênero Saccharum, do qual a cana-de-açúcar faz parte, pertence à tribo Andropogoneae e a subtribo Sacchacarinae. Nesta subtribo inclui as espécies com maior eficiência de acúmulo de biomassa, através da assimilação eficiente de carbono em elevadas temperaturas, o que é típico de plantas que possuem o mecanismo fotossintético C4 (Paterson et al., 2009).

Acredita-se que a cana-de-açúcar foi inicialmente cultivada na Nova Guiné por volta de 6000 anos a.c. No entanto, o desenvolvimento do cultivo aconteceu na Índia, anos depois. Existem evidências de que a cana-de-açúcar possui seu centro de origem na região da Indonésia e Nova Guiné e tem sido cultivada na Ásia, desde épocas pré-históricas (Burr et al., 1956). A cana-de-açúcar chegou ao Brasil no século XVI, junto com os portugueses. As primeiras mudas vieram em 1532, na expedição marítima de Martim Afonso de Souza. A cana-de-açúcar possui uma domesticação antiga e complexa, relacionada à existência de vários cruzamentos interespecíficos entre cultivares tradicionais e parentes silvestres (Grivet et al., 2004).

A cana-de-açúcar apresenta uma importância histórica para a economia brasileira. No Período Colonial, durante o sistema de capitanias hereditárias, a Capitania de Pernambuco no nordeste brasileiro se tornou um centro de crescimento populacional e econômico devido à exploração da cana-de-açúcar. Os elevados preços que o açúcar era

cotado na Europa e a pequena oferta do produto fez com que no final do século XV, o Brasil Colônia fosse o maior produtor de açúcar do mundo, representando um dos maiores momentos de crescimento econômico do Brasil Colônia (Schwartz, 2005). Outro momento de desenvolvimento econômico brasileiro relacionado com a cultura da cana-de-açúcar se deu em meados da década de 70 com a implementação do Programa Nacional do Álcool (Proálcool). Este Programa objetivava a substituição em larga escala dos combustíveis veiculares derivados de petróleo por álcool, devido à crise do petróleo em 1973. Assim, a produção do álcool oriundo da cana-de-açúcar foi altamente financiada em todo o território nacional, representando um passo importante no financiamento dos mais diversos estudos sobre a biologia da espécie, permitindo a criação de programas de melhoramento genético e o desenvolvimento de cultivares nacionais de cana-de-açúcar (Giacomazzi, 2012).

As espécies do complexo *Saccharum* são plantas que utilizam vias metabólicas C4, permitindo uma fotossíntese mais eficiente, sobretudo em regiões de elevada temperatura. Em alguns países, produz-se 40 toneladas de matéria seca por hectare, em outros, a produtividade pode chegar a 70 toneladas por hectare. No entanto, em condições experimentais ideais a produção pode chegar a 100 toneladas por hectare, fazendo da cana-de-açúcar a espécie com maior rendimento de cultivo (matéria seca/biomassa) no mundo (Henry, 2010).

Abrangendo cerca de 35% da produção mundial de cana-de-açúcar na safra 2013/2014, o Brasil se destaca como o maior exportador mundial de açúcar e etanol derivados da cana-de-açúcar. O crescimento de produtividade no Brasil tem se mostrado contínuo ao longo dos anos, aumentando de algo em torno de 271 milhões de toneladas, na safra de 1992 para os 658 milhões de toneladas colhidas na safra de 2013/14. Dentre os estados brasileiros produtores, deve-se destacar o estado de São Paulo como sendo o maior produtor, com 52% (4,6 milhões ha) da área plantada, seguido pelos estados de Goiás, Minas Gerais e Mato Grosso do Sul com cerca de 9,5% (852 mil ha), 8,9% (800 mil ha) e 7,4% (668 mil ha) da área plantada no Brasil, respectivamente[4]. O Ministério de Agricultura Pecuária e Abastecimento (MAPA)[5] estima que o país deve alcançar taxa média de aumento anual da produção de açúcar de 3,25% até 2018/19, e produzir 47,34 milhões de toneladas do produto, o que corresponde a um acréscimo de 14,6 milhões de

[4] Informação disponível em: www.conab.gov.br
[5] Informação disponível em: www.agricultura.gov.br

toneladas em relação ao período 2007/2008. Para as exportações de açúcar, o volume previsto para 2019 é de 32,6 milhões de toneladas.

Apesar de o Brasil ser o maior produtor mundial de cana-de-açúcar, a produtividade de açúcar tem se estabilizado sem apresentar ganhos significativos nos últimos dez anos (Figura 1), o que justifica a importância de melhorar o entendimento genético/genômico que se tem sobre a espécie. Os dados mostram que o aumento anual de produtividade (biomassa) é de 1,2% enquanto que o acúmulo de açúcar cresce num ritmo insignificante de 0,2% ao ano (Dal-Bianco et al., 2011).

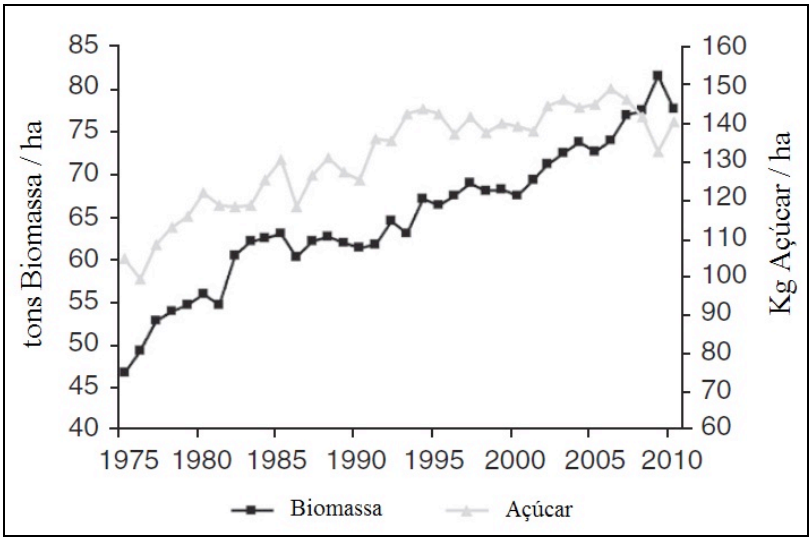


Figura 1. Evolução da produtividade de biomassa e de açúcar de *Saccharum* spp., evidenciando o crescimento de 1,2% ao ano da produtividade de biomassa e 0,2% ao ano da produtividade de açúcar.

2.2 EVOLUÇÃO DO GENOMA DAS ESPÉCIES DO COMPLEXO *Saccharum*

Ao fazer uma revisão do gênero *Saccharum* e de outros gêneros próximos, Mukherjee (1957) demonstrou que os gêneros *Saccharum*, *Ripidium*, *Sclerostachya* e *Narenga* constituíam um grande e bastante relacionado grupo de intercruzamentos, os quais deram origem à cana-de-açúcar. Foi este autor que cunhou o termo "Complexo *Saccharum*", com o objetivo de descrever este enorme pool gênico de cruzamentos.

A cana-de-açúcar compreende as várias espécies do gênero *Saccharum*. Estas espécies já foram caracterizadas e podem ser identificadas com base na taxonomia tradicional em: *S. spontaneum*, *S. robustum*, *S. officinarum*, *S. barberi*, *S. sinense*, e *S. edule* (Daniels & Roach, 1987). *S. spontaneum* e *S. robustum* são espécies silvestres com número básico de cromossomo x = 8 e x = 10, respectivamente (D'Hont et al., 1998). *S. officinarum* é a espécie domesticada de cana-de-açúcar com nível de ploidia igual a oito (autooctaploide – 2n = 8x = 80) com provável origem a partir da espécie silvestre autopoliploide *S. robustum* (2n = 60 ou 80). As outras três espécies são híbridas. *S. barberi* e *S. sinense* são híbridos interespecíficos entre *S. officinarum* e *S. spontaneum*. *S. edule* pode ser um híbrido interespecífico ou intragenérico entre *S. officinarum* ou *S. robustum* com outra espécie do complexo *Saccharum* (D`Hont et al., 2004).

Eventos de poliploidização são forças evolutivas importantes, existentes principalmente no grupo taxonômico das angiospermas (Adams & Wendel, 2005; Doyle et al., 2008; Soltis & Soltis, 1999). Adams & Wendel (2005) e Masterson (1994) ainda afirmam que a poliploidização é o principal evento de duplicação gênica, que ocorre em aproximadamente 70% das angiospermas. Paterson (2005) chama a atenção para o fato de que eventos de duplicação do genoma como um todo (*genome wide chromatin duplication events*) podem ser responsáveis pela origem das angiospermas, moldando toda a biologia das espécies florais. Acredita-se que estes eventos são responsáveis pelos mecanismos de adaptação de algumas espécies, principalmente as gramíneas, às pressões de domesticação impostas pelo ser humano. Devido à elevada frequência com que os eventos de poliploidização ocorrem em plantas, pode-se afirmar que as espécies provavelmente formam um grupo polifilético (Soltis & Soltis, 1999; Soltis et al., 2009). A poliploidização pode representar um período de transição, durante o qual, alterações genômicas ocorrem, com o potencial de produzir novos complexos gênicos, facilitando uma rápida evolução molecular (Wendel, 2000).

Indivíduos poliploides possuem algumas características que contribuem para uma melhor adaptação às variações ambientais, permitindo uma sobrevivência diferenciada em relação a indivíduos diploides, por exemplo (Hancock, 2004). As principais vantagens adaptativas são: aumento da quantidade de DNA, do tamanho celular e alteração nas taxas de desenvolvimento (efeito nucleotípico); aumento do nível de produção enzimática (efeito de dosagem) e aumento da heterozigosidade. Este último fator, determinado pela

duplicação gênica, consegue explicar o aumento da plasticidade fenotípica e a elevada capacidade de adaptação às diversas condições ambientais apresentadas por espécies poliploides. Esta plasticidade fenotípica se caracteriza, pois várias enzimas serão produzidas pelas diversas cópias gênicas existentes e cada uma destas enzimas pode estar relacionada a condições ambientais diferentes e específicas (Hancock, 2004).

Após os eventos de duplicação gênica, os genes duplicados têm três destinos: (1) continuar ativos com a mesma função, (2) continuar ativos, mas com funções diferentes e (3) serem silenciados. Mas, há evidências de que a grande maioria destes genes permanece ativos. Comai et al. (2000) simularam populações alotetraploides artificiais de *Arabidopisis thaliana* e *Cardaninopsis arenosa* e compararam os níveis de expressão gênica em populações diploides e poliploides. Concluíram que somente 0,4% dos genes foram realmente silenciados nas populações tetraploides.

As variedades modernas de cana-de-açúcar (*Saccharum* spp.) são formadas pelo cruzamento interespecífico entre *S. officinarum* (2n = 80) x *S. spontaneum* (2n = 40 a 128) que aconteceram no decorrer do último século, com início na Índia, na década de 1920 (Hermann et al., 2012). Esses híbridos apresentam eventos de poliploidização e aneuploidia com um número de cromossomos variando de 100 a 130, em que 85-90% do genoma é proveniente da espécie *S. officinarum* e 15-10% proveniente do parente silvestre *S. spontaneum* (Paterson, et al. 2010; Piperidis et al., 2010). Durante estes processos de hibridização através sucessivos retrocruzamento utilizando como parental recorrente *S. officinarum*, ocorreu um fenômeno chamado de nobilização nos primeiros ciclos de retrocruzamentos. Trata-se de uma peculiaridade citológica em que, com alta frequência de ocorrência, 2n dos gametas de *S. officinarum* foram transmitidos durante o cruzamento com *S. spontaneum*, quando *S. officinarum* foi tratado como parental feminino (Bhat & Gill, 1984; Roach, 1987; Paterson et al., 2010). Este processo acelerou a recuperação de alelos responsáveis pela produção de açúcar de *S. officinarum* (Tabela 1), além de ter introduzido alelos de tolerância e resistência existentes em *S. spontaneum*, explicando o enorme vigor híbrido apresentado pela progênie deste cruzamento (Paterson et al., 2010). Daniels & Roach (1987) fizeram uma ótima e detalhada revisão sobre a taxonomia do gênero *Saccharum*, esclarecendo sobre as principais hipóteses a respeito da evolução do gênero.

Tabela 1. Número de cromossomos em três estágios de nobilização em cruzamentos entre *S. officinarum* (2n = 80) e *S. spontaneum* (2n = 64), assumindo a participação de 2n gametas nos três estágios.

Estágio de nobilização	Geração	Número de cromossomos	Proporção (%) de *S. off.* : *S. spont.*
I	F_1 : *S. off.* x *S. spont.*	2n = 80 + 32 = 112	71,4 : 28,6
II	RC_1 : *S. off.* x F_1	2n = 80 + 56 = 136	88,2 : 11,8
III	RC_2 : *S. off.* x RC_1	2n = 80 + 68 = 148	92,6 : 7,4

S. off. = Saccharum officinarum
S. spont. = Saccharum spontaneum

Utilizando técnicas citogenéticas de hibridização *in situ* (GISH), Piperidis et al. (2010) demonstraram que algo em torno de 25 a 27,5% do genoma das cultivares modernas de cana-de-açúcar são derivadas de *S. spontaneum*, enquanto que 8 a 13% do genoma têm origem nas recombinações interespecíficas. Estes autores também confirmaram a ocorrência de transmissão de 2n + n gametas em cruzamentos de *S. officinarum* x *S. spontaneum*, porém, relataram a possibilidade de existência desse fenômeno também entre cruzamentos de cultivares modernas (*Saccharum* spp.) e *S. officinarum*. Alguns autores sugerem que esse fenômeno não é bem definido e de fácil compreensão como apresentado acima, sugerindo o acontecimento tanto em gametas masculinos quanto femininos (Bielig et al., 2003).

Existe uma diferença estrutural entre os genomas de *S. officinarum* e *S. spontaneum*, havendo certa independência entre os grupos de ligação das duas espécies (Ming et al., 2008). Estes mesmos autores encontraram onze rearranjos cromossômicos distintos entre *S. officinarum* e *S. spontaneum* e treze rearranjos cromossômicos diferentes entre *Saccharum* spp. e *Sorgum bicolor*. Isto pode ser um indício de que a divergência entre *S. officinarum* e *S. spontaneum* pode ter sido tão antiga quanto a divergência entre cana-de-açúcar e sorgo a qual, pela comparação entre *Miscanthus* e *Saccharum*, é datada em aproximadamente 7-9 milhões de anos atrás (Paterson et al., 2009).

A relação evolutiva do complexo *Saccharum*, em relação às espécies da família Poaceae, apresenta uma sintenia interessante com a espécie *Sorghum bicolor*, uma vez que ambas fazem parte da subtribo Sacchacarinae, o que indica a existência de um ancestral comum entre elas há aproximadamente 7-9 milhões de anos atrás (Jannoo et al., 2007). Existem muitos genes parálogos entre as duas espécies, mostrando que neste curto período de evolução divergente, o complexo *Saccharum* passou por pelo menos dois eventos de

duplicação gênica completa (Paterson et al., 2009). Estes eventos de duplicação gênica possuem uma importância central na evolução e adaptação da cana-de-açúcar.

2.2.1 Os desafios dos estudos genômicos em cana-de-açúcar

Até o final da década de 90, o conhecimento sobre a genética/genômica da cana-de-açúcar era relativamente limitado, pois a enorme complexidade do genoma, o pouco desenvolvimento tecnológico das ferramentas de sequenciamento e o elevado custo de projetos desta natureza impediam grandes avanços nesta área. Foram nos últimos 20 anos, principalmente com a redução do custo de obtenção de informações genéticas, que houve um crescente número de trabalhos com os mais diversos objetivos de compreensão genômica e do transcritoma da espécie.

Lakshmanan et al. (2005) sugerem a cana-de-açúcar como uma espécie em que o melhoramento genético apoiado pela utilização de ferramentas molecular, teria grandes vantagens em ser aplicado. Com isso, o emprego dos métodos biotecnológicos existentes atualmente possui uma capacidade de produzir ótimas mudanças na cultura da cana-de-açúcar, principalmente devido à complexidade do genoma (poliploide e aneuploide), a baixa fertilidade, a susceptibilidade a doenças, e a longa duração para produção de cultivares elites. Neste contexto, podem ser destacadas as principais áreas de atuação das pesquisas genético-moleculares com a espécie: (1) técnicas de cultura de tecidos e células para o melhoramento molecular e a propagação vegetativa; (2) engenharia genética de novos genes de interesse agrícola; (3) diagnóstico molecular de patógenos para aperfeiçoamento do uso de germoplasma exótico de gêneros próximos (*Miscanthus* e *Erianthus*); (4) desenvolvimento de mapas genéticos usando marcadores moleculares atuais como os SNPs e (5) compreensão das vias metabólicas de acúmulo de sacarose no colmo de cana-de-açúcar (Suprasanna et al., 2011).

Butterfield et al. (2001) relataram que o tamanho básico do genoma de *Saccharum* spp. é cerca de duas vezes maior se comparado com o genoma de arroz (*Oryza sativa*). O genoma monoploide de *S. officinarum* (x = 10) apresenta um tamanho de

aproximadamente 926 Mpb, enquanto que em *S. spontaneum* (x = 8), aproximadamente 760 Mpb. Portanto, o tamanho aproximado do genoma de cana-de-açúcar, tratada como uma espécie octaploide a dodecaploide, pode chegar aos 10 Gb (Setta et al., 2014). O genoma de sorgo (*Sorghum bicolor*) com aproximadamente 700 Mpb é o genoma mais próximo da cana-de açúcar em termos de tamanho (Paterson et al., 2009). Dentre as gramíneas, o milho (*Zea mays*) é a espécie que apresenta o maior genoma completamente sequenciado, com cerca de 2,3 Gb (Schnable et al., 2009).

Com o atual desenvolvimento das plataformas de sequenciamento de nova geração, o acesso aos dados genéticos se tornou mais rápido e mais barato. Existe uma expectativa muito grande quanto ao uso destas ferramentas para produção de informações genômicas de cana-de-açúcar. Com isso, uma compreensão mais detalhada da organização e estrutura do genoma, da existência de genes parálogos que esclarecem sobre os eventos de duplicação do genoma, da existência de genes ortólogos que revelam as relações filogenéticas, da existência de SNPs, da expressão diferencial de genes e futuramente das vias metabólicas associadas a características fenotípicas de interesse, poderão ser mais bem aproveitadas e utilizadas com maiores expectativas no melhoramento genético da espécie.

2.3 AS PLATAFORMAS DE SEQUENCIAMENTO DE NOVA GERAÇÃO (NGS – *NEXT GENERATION SEQUENCING*)

Durante o projeto de sequenciamento do genoma humano (HGP – *Human Genome Project*), realizado através do sequenciamento Sanger, iniciou-se o desenvolvimento das plataformas de sequenciamento que atualmente são conhecidas como sequenciadores de nova geração. Atualmente presenciamos a produção de dados genéticos (sequenciamento de genomas e transcritomas) em larga escala com custos cada vez mais baixos. Esta redução do custo de sequenciamento permitiu um aumento do volume de projetos de genômica estrutural e funcional em todo o mundo, viabilizando o sequenciamento de genomas de espécies modelos e não modelos (Metzker, 2010; Green, 2001).

Shendure et al. (2004) e Shendure & Ji (2008) classificam os métodos de sequenciamento de DNA em quatro abordagens diferentes. A primeira abordagem são os métodos de eletroforese. A segunda abordagem compreende o sequenciamento por hibridização (SBH – *Sequencing By Hybridization*). A terceira abordagem se refere ao sequenciamento de moléculas individuais de DNA e/ou RNA em tempo real. A quarta abordagem são as metodologias de sequenciamento cíclico de matrizes. Esta abordagem utiliza inúmeros ciclos de reações enzimáticas para a manipulação de matrizes de fragmentos de DNA. Cada ciclo de sequenciamento é capaz de decodificar poucos pares de base da sequência alvo, porém o procedimento é feito simultaneamente para bilhões de fragmentos de DNA, com uma capacidade de decodificação de milhares de nucleotídeos em pouco tempo de sequenciamento. Trata-se de um método que não utiliza a eletroforese capilar e está presente nas plataformas de sequenciamento de nova geração (NGS), principalmente nos sequenciadores de segunda geração, também conhecido como tecnologias de sequenciamento de alta cobertura (*High Throughput Sequencing* - HTS).

As descobertas científicas que resultaram na aplicação das tecnologias de sequenciamento de nova geração tiveram um impacto muito grande em diversas áreas da biologia, principalmente na genética, além de permitirem uma análise ampla dos genomas com precisão ao nível de nucleotídeos/pares de base (Mardis, 2008). Com isso, estudos que vão desde a construção de mapas genéticos em humanos com a intenção de associar doenças hereditárias a polimorfismos de uma única base (SNPs) (Baird et al., 2008), passando pelo melhoramento/seleção genômica ampla das espécies cultivadas (Kruglyak, 1999; Jannink et al., 2010), pela metagenômica (Mardis, 2008) até a genômica de populações (Davey & Blaxter, 2010; Hohenlohe et al., 2011) tiveram um avanço enorme na quantidade e qualidade de informações disponíveis e na precisão das análises genético-estatísticas.

As plataformas de NGS começaram a ser comercializadas em 2005 (Liu et al., 2012) e estão evoluindo rapidamente. Todas essas tecnologias promovem o sequenciamento de DNA em plataformas capazes de gerar informação sobre milhões ou até mesmo bilhões de pares de bases em uma única corrida. Dentre estas, destacaram-se: a 454 FLX (Roche), que foi a primeira plataforma de NGS desenvolvida, a Solexa (Illumina), a SOLiD (*Applied Biosystems)*, a Ion Torrent da *Life Tecnhologies*, que detecta os nucleotídeos com base nas variações de pH do meio bioquímico, a *Heliscope* (tSMS)

(Helicos), a PacBio (Pacific Bioscience) e a *Nanopore* (*Oxford Nanopore Technologies)*. As duas últimas plataformas são conhecidas como sequenciamento de terceira geração (Aluru, 2012). A plataforma de sequenciamento da Illumina se destacou entre as concorrentes, sendo, atualmente, a mais utilizada.

Essas novas plataformas possuem como características comuns um poder de gerar informação numa quantidade milhares de vezes maior que o sequenciamento de Sanger, com uma grande economia de tempo e dinheiro, revolucionando as técnicas de sequenciamento de moléculas (Glenn et al., 2011). Essa capacidade extraordinária de produção de elevada quantidade de dados advém do uso de reações químicas complexas e de um enorme desenvolvimento tecnológico, na área da genética molecular, que fornece sistemas sólidos como unidades de sequenciamento e diferentes métodos de detecção de *base calling*. Estas plataformas de sequenciamento de genomas aliviam o intensivo trabalho laboratorial de preparação de amostras, reações de PCR e de sequenciamento. As reações moleculares realizadas *in vitro* em suportes sólidos dentro destas plataformas de sequenciamento permitem que as leituras da sequência de milhares de fragmentos de DNA possam produzir Gigabases ou até mesmo Terabases de sequências em tempos curtos e de forma relativamente barata (Mardis, 2008; Shendure & Ji, 2008; Ansorge, 2009; Carvalho & Silva, 2010). Estas tecnologias abriram a oportunidade para o sequenciamento amplo do genoma de qualquer organismo (modelos e não modelos), além de acelerar o ritmo com que a exploração do genoma é feita, proporcionando até o ressequenciamento genômico e análises robustas sobre o transcritoma de qualquer espécie (Lyster et al., 2009). No entanto, todas estas plataformas possuem pontos negativos, principalmente quanto ao tamanho pequeno dos reads sequenciados e relacionados aos erros de sequenciamento.

Os erros de sequenciamento existem em ambas as plataformas e em sua maioria podem ser classificados em inserções/deleções – conhecidos como *indels* – e substituições (Tabela 2). Sabe-se que quanto maior o tamanho dos reads sequenciados maior será a taxa de erro, isto é, o tamanho máximo dos reads está relacionado ao quanto é aceitável de erros de sequenciamento (Glenn, 2011). Estes erros devem ser levados em consideração durante o desenvolvimento de algoritmos matemáticos de análise da sequência, principalmente nos algoritmos de *base calling*. Glenn (2011) e Ross et al. (2013) discutem a dificuldade de se comparar os erros existentes por detrás de cada plataforma de NGS, pois a média da taxa de erro por pares de bases pode variar de 0,01% à

16% entre as plataformas de NGS. A plataforma SOLiD apresenta a menor taxa de erro dos dados acessíveis aos usuários, enquanto a PacBio apresenta a maior taxa de erro. Esta baixa taxa de erro por nucleotídeo sequenciado no sistema SOLiD é explicada pelo fato de que cada nucleotídeo é sequenciado duas e/ou três vezes (Glenn, 2011). Além dos erros de sequenciamento, cada plataforma apresenta um viés quanto à distribuição e cobertura dos reads sequenciados. Este viés pode ser produzido durante a construção das bibliotecas, amplificação dos fragmentos de sequenciamento e durante o próprio sequenciamento e possui implicações diretas nos dados obtidos e consequentemente nas análises de bioinformática. Métodos computacionais capazes de identificar e quantificar este viés já foram desenvolvidos (Ross et al., 2013).

Tabela 2. Taxas de erro das principais plataformas de sequenciamento de DNA. Todas as taxas de erro estão em porcentagem, que significa a porcentagem de erro por base dentro de um único read com comprimento máximo.

Plataforma de sequenciamento	Tipos de erros	Taxa de erro inicial (%)	Taxa de erro final (%)
3730xl (Sanger/Capilar)	Substituição	0,1-1	0,1-1
454 (Pirossequenciamento)	*Indel*	1	1
Illumina (Todos os modelos)	Substituição	~0,1	~0,1
Ion Torrent (Todos os chips)	*Indel*	~1	~1
SOLiD – 5500xl	A-T viés	~5	≤0,1
Oxford Nanopore	Deleção	≥4*	4*
PacBio RS	*Indel*	~13	≤1

Fonte: Gleen (2011)

*Informações com base em fontes da empresa. Não é claro se os 4% são referentes ao sequenciamento de ambas às fitas ou de uma sequência consenso.

Shendure & Ji (2008) discutem que a criação destas plataformas de sequenciamento de alta cobertura surgiu com o desenvolvimento de quatro áreas. A primeira foi o projeto de sequenciamento do genoma humano, em que disputas entre instituições públicas e privadas sobre quem sequenciaria o genoma com menor custo, permitiu um primeiro desenvolvimento de técnicas mais elaboradas de sequenciamento. A segunda foi na adoção de fragmentos curtos (20-50 pb) de DNA para serem sequenciados (Tecnologia de sequenciamento de *reads* curtos – SRS – *Short Reads Sequencing*), em comparação com os 450 a 900 pb que eram gerados no sequenciamento de Sanger. A terceira foi o crescente desenvolvimento das técnicas moleculares, que forneceu uma enorme variedade de alternativas às trabalhosas reações necessárias para o sequenciamento. Em quarto, está o progresso tecnológico por detrás de alguns campos

importantes como a microscopia ótica, a bioquímica de nucleotídeos, a engenharia da polimerase, a computação de softwares e hardwares, o armazenamento de dados e outros.

Atualmente, já são comercializadas máquinas capazes de gerar uma enorme quantidade de dados, porém ocupando um espaço bem menor no laboratório. Estas máquinas são chamadas de sequenciadores de alto desempenho de bancada (*Benchtop high-throughput sequencing platforms*). Existem três principais equipamentos de sequenciamento de bancada. O 454 Junior (Roche), o MiSeq (Illumina) e o *Ion Torrent* PGM (Life Technologies). As metodologias de sequenciamento inseridas nas plataformas 454 Junior e no MiSeq são idênticas às apresentadas no pirossequenciamento e no equipamento HiSeq (Illumina), respectivamente. Já a plataforma *Ion Torrent* PGM foi proposta no começo de 2011, usando PCR em emulsão e o sequenciamento por síntese. Parte-se do princípio que cada um dos quatro nucleotídeos incorporados a fita molde de DNA, pela ação da DNA polimerase, altera o pH do meio de modo diferente, liberando íons H^+ (Loman et al., 2012; Liu et al., 2012). É o primeiro método de sequenciamento que não utiliza a detecção de fluorescência como determinação da posição dos nucleotídeos na sequência de DNA. A comparação entre estas plataformas de sequenciamento é algo inevitável devido à competição existente entre as empresas detentoras destas tecnologias. Glenn (2011), Loman et al. (2012) e Liu et al. (2012) fizeram uma ótima revisão entre as diversas abordagens moleculares implementadas nas plataformas de sequenciamento de nova geração, dando ênfase aos pontos positivos e negativos de cada tecnologia.

2.3.1 A plataforma de sequenciamento da Illumina

Inicialmente conhecida como plataforma *Solexa*, esta metodologia de sequenciamento foi proposta por Turcatti et al. (2008) como uma nova metodologia de sequenciamento de nova geração caracterizada pelo uso de nucleotídeos modificados. Características como a proteção do grupamento hidroxila na posição 3’ permitem que o nucleotídeo fluorescente e reversível seja incorporado na fita de DNA e/ou RNA e, posteriormente identificado. Este processo de sequenciamento por adição de nucleotídeos é chamado de sequenciamento por síntese do DNA/RNA (SBS - *Sequencing By Synthesis*). Este método SBS permite que os quatro nucleotídeos sejam incorporados simultaneamente

durante o sequenciamento que ocorre em células sólidas fixas chamadas de *flow cells* (Mardis, 2008).

Atualmente, a empresa Illumina já desenvolveu diversas máquinas de sequenciamento, incluindo Genome Analyzer IIx, HiSeq, MiSeq e o NextSeq, além de máquinas de *Arrays* como o HiScanSQ e o iScan[6]. O sequenciador HiSeq é plataforma mais utilizada na produção de dados genômicos com elevada densidade de cobertura.

O sequenciamento na plataforma Illumina é realizado por síntese usando a DNA polimerase e nucleotídeos terminadores marcados com diferentes fluoróforos. A inovação dessa plataforma consiste na clonagem *in vitro* dos fragmentos em uma plataforma sólida de vidro, processo também conhecido como PCR de fase sólida (Carvalho & Silva, 2010). Bibliotecas genômicas são construídas por qualquer método que garanta a adição de adaptadores nas extremidades 3' e 5' nos fragmentos de aproximadamente 100-800 pb de comprimento. Estes adaptadores fazem a fixação por hibridação destes fragmentos em uma célula de sequenciamento sólida altamente preenchida de oligonucleotídeos que servirão como primers durante a PCR (Shendure & Ji, 2008). Vários nucleotídeos não marcados são fornecidos, no primeiro ciclo de amplificação, para que haja a síntese complementar do fragmento ancorado na célula. O anelamento com os primers (oligonucleotídeos) existentes na célula fazem com que o fragmento assuma um formato de "ponte" (*bridge PCR*). A extensão é feita pela DNA polimerase e a fita complementar formada também assume o formato de "ponte", o que caracteriza a PCR. No ciclo de desnaturação, as fitas são separadas e linearizadas. Esses ciclos são repetidos cerca de 40 vezes e pelo menos mil cópias são geradas de cada fragmento, aos quais permanecem próximas umas das outras, formando uma espécie de *cluster* de sequenciamento (Ansorge, 2009).

Alguns milhões de *clusters* são amplificados em até oito linhas independentes existentes em cada célula de sequenciamento, de modo que oito bibliotecas genômicas podem ser sequenciadas em conjunto utilizando uma única corrida da plataforma (Shendure & Ji, 2008). Posteriormente, alguns iniciadores universais de sequenciamento, formados por nucleotídeos modificados são usados durante a reação de sequenciamento, que realiza a determinação dos quatro nucleotídeos simultaneamente (Shendure & Ji, 2008;

[6] Informação disponível em: http://systems.illumina.com/systems.html

Ansorge, 2009). Após a incorporação dos nucleotídeos modificados nos fragmentos sendo sequenciados em síntese (SBS - *Sequencing By Synthesis*), a leitura do sinal de fluorescência é feita simultaneamente para os quatro pares de bases sequenciados nos milhões de grupos de fragmentos amplificados. Em seguida, ocorre uma etapa de lavagem para remoção dos reagentes excedentes e remoção do terminal 3' bloqueado e do fluoróforo do nucleotídeo incorporado no ciclo anterior para que a reação de sequenciamento prossiga. A leitura das bases é feita pela análise sequencial das imagens capturadas em cada ciclo de sequenciamento (Shendure & Ji, 2008).

No início, a plataforma Solexa GA conseguia produzir uma quantidade de 1 Gb/corrida. Posteriormente, conseguiu-se um rendimento de sequenciamento de 20 Gb/corrida em bibliotecas *Paired-Ends* (PE) (reads sequenciados nas duas extremidades, 3'e 5' do fragmento de sequenciamento) com reads de 75 pares de base. Com o desenvolvimento tecnológico da plataforma estes valores foram aumentando para 30, 50 e 85 Gb/corrida com reads PE de 100 pb. Atualmente, o sequenciador HiSeq 2000 consegue produzir cerca de 600 Gb/corrida. Estima-se que com a queda dos custos de sequenciamento, será possível obter 1 Tb/corrida em um tempo de cerca de oito dias (Liu et al., 2012). A taxa de erro de um reads de 100 pares de bases de tamanho é, em média 2%, após a etapa de filtragem. Comparado com as plataformas 454 e SOLiD, o sequenciamento Illumina é muito mais barato, com um custo de 0,02 dólares por *datapoint*. Com a possibilidade de realização de sistemas de multiplex, através dos adaptadores P5/P7, cerca de cem amostras podem ser sequenciadas simultaneamente. Existem dois softwares principais, embutidos na plataforma HiSeq 2000, responsáveis pelo controle de qualidade do sequenciamento (HCS – *HiSeq Control System*) e do processo de *base calling* (RTA – *Real-Time Analyzer*). Outro importante algoritmo existente nesta plataforma é conhecido como CASAVA, responsável pelas análises subsequentes de processamento dos reads. O HiSeq 2000 utiliza dois lasers e quatro filtros para detectar os quatro tipos de nucleotídeos com uma emissão de fluorescência simultânea para os quatro tipos de nucleotídeos, de maneira que a imagem dos quatro nucleotídeos não é independente. Assim, a distribuição dos nucleotídeos sequenciados pode afetar a qualidade do sequenciamento (Liu et al., 2012).

Algumas limitações da metodologia são expostas por Mardis (2008) e Shendure & Ji (2008). Estes autores discutem que a leitura, através de algoritmos de *base*

calling de fragmentos muito grandes pode gerar sequências de baixa qualidade nas extremidades de leitura do fragmento. O tipo de erro dominante nesta plataforma é a substituição de nucleotídeos, ao contrário da plataforma 454 em que predominam os erros do tipo *indels* em homopolímeros. O algoritmo de *base calling* existente dentro da plataforma Illumina possibilita a eliminação das bases de má qualidade usando valores de phred (Ewing et al., 1998) como referência (Mardis, 2008).

2.4 ESTUDOS GENÔMICOS EM CANA-DE-AÇÚCAR

O melhoramento genético é considerado uma das principais estratégias para aumentar a produtividade das espécies cultivadas. A compreensão da composição e da estrutura de um genoma tem sido cada vez mais importante para a eficiência dos programas de melhoramento, permitindo que a seleção de genótipos superiores possa ser realizada com base nas características genômicas e não somente em observações fenotípicas, o que de certa forma aumenta os ganhos genéticos com a seleção (Resende et al., 2008).

D'Hont & Glasman (2001) fizeram uma revisão da literatura sobre o progresso das pesquisas genéticas em cana-de-açúcar e discutiram que as informações levantadas até aquele momento eram de grande valia para auxiliar os programas de melhoramento da espécie, mas muito ainda deveria ser feito para possibilitar a implementação real de técnicas de melhoramento como a seleção assistida por marcadores. Arruda (2001) também acredita que a coleção de trabalhos sobre a caracterização do genoma de cana-de-açúcar publicada até 2001 seria de extrema importância para direcionar os estudos futuros, além de permitir uma compreensão da relação de sintenia existente entre o genoma de espécies filogeneticamente próximas à cana-de-açúcar.

Nos últimos quarenta anos houve um avanço significativo dos estudos genéticos-genômicos das espécies do complexo *Saccharum*, principalmente devido ao desenvolvimento biotecnológico e bioquímico existente nas atuais plataformas de sequenciamento de moléculas de DNA ou RNA que permitiram que genomas complexos

como o genoma da cana-de-açúcar fossem mais bem compreendidos quanto a sua composição, estrutura e evolução. Mesmo assim, a compreensão detalhada do genoma de cana-de-açúcar ainda é muito limitada quando comparado com as informações existentes para outras espécies agronomicamente importantes da família das gramíneas. Um exemplo claro é a dificuldade de montagem e a não existência de um genoma de referência para a espécie, diferentemente do encontrado em sorgo (Paterson et al., 2009), arroz (Kawahara et al., 2013) e milho (Hirsch et al., 2014), por exemplo, que já possuem seus genomas sequenciados e anotados.

2.4.1 Caracterização da diversidade genética e construção de mapas genéticos

Os primeiros estudos sobre a genética da cana-de-açúcar se iniciaram com a utilização de locos isoenzimáticos para caracterização da diversidade genética no início dos anos 70 (Thom & Maretzki, 1970; Waldron & Glasziou, 1971). Mesmo sendo considerada uma cultura extremamente importante, os primeiros mapas genéticos para cana-de-açúcar, construídos com base em locos AFLP *single doses*, por exemplo, são do início dos anos 90 (Da Silva et al., 1993; D'Hont et al., 1994). A identificação de genes candidatos através de análises de QTL e estudos de associação eram escassos até início dos anos 2000, quando o primeiro gene associado à resistência em cana-de-açúcar foi mapeado (Asnaghi et al., 2000). Mesmo assim, a grande maioria dos mapas genéticos para cana-de-açúcar não são saturados (Garcia et al., 2013), pois as marcas *single doses* não são suficientes para amostrar a enorme variação de ploidia do genoma da cana-de-açúcar.

2.4.2 Sequenciamento de bibliotecas de ESTs e identificação de genes de interesse

A identificação de ESTs (*Expressed Sequence Tags*), considerados regiões gênicas que fazem parte do transcritoma das espécies, é importante para identificação, caracterização e validação de genes de interesse agronômico. Tomkins et al. (1999) construíram a primeira biblioteca de BACs para cana-de-açúcar. Carson & Botha (2000) construíram o primeiro banco de dados de sequências de ESTs para cana-de-açúcar com o objetivo de dar suporte aos programas de melhoramento genético através da identificação e caracterização gênica. No entanto, o trabalho caracterizado como SUCEST (*SUgarCane ESTs*) pode ser considerado um dos trabalhos pioneiros e mais completos sobre a disponibilidade de bibliotecas de ESTs para cana-de-açúcar (Vettore et al., 2001). Aproximadamente 238 mil sequências de ESTs de alta qualidade foram sequenciadas em tecnologia Sanger a partir de nove diferentes tecidos vegetais derivados de treze diferentes variedades de cana-de-açúcar. Vettore et al. (2003) realizaram a montagem e a anotação funcional destas 237954 sequências de ESTs e evidenciaram a existência de pouco mais de 43 mil transcritos, dos quais 35% não estavam presentes em bancos de dados públicos. A anotação funcional dos genes foi realizada para 33% do total de transcritos que apresentaram pelo menos um clone com ORF completa e revelou que 50% destes insertos em *full-length* estavam relacionados com o metabolismo de proteína, a comunicação celular, a funções bioenergéticas e a resposta a estresses bióticos e abióticos. O banco de dados de EST de cana-de-açúcar permitiu que estudos de associação fossem feitos entre estas regiões genômicas e variações de caracteres quantitativos, revelando uma grande quantidade de genes associados às características de interesse agronômico. Genes envolvidos na desintoxicação de espécies reativas de oxigênio (Kurama et al., 2002), envolvidos em mecanismos de tolerância à baixas temperaturas e resistência ao ataque de patógenos (Nogueira et al., 2003) e a identificação de genes da enzima álcool hidrogenase (Adh) (Grivet el al., 2003) são alguns dos exemplos da utilização do banco de dados do SUCEST em estudos genômicos para cana-de-açúcar. Vicentini et al. (2012) analisaram o conteúdo genético do banco de dados de ESTs de cana-de-açúcar e revelaram a existência de aproximadamente dez mil genes ainda não identificados e anotados para a espécie, além de inferirem que 58% dos ESTs são considerados regiões ortólogas ao proteoma de *Sorghum bicolor*. Estes autores ainda revelaram a existência de mais de dois mil RNAs não codificantes de proteínas conservados entre *S. bicolor* e *Saccharum* spp.

Casu et al. (2004), Casu et al. (2005) e Casu et al. (2007) também realizaram estudos referente a caracterização do genoma de cana-de-açúcar através de bibliotecas de ESTs. O primeiro estudo retrata a identificação, através de técnicas de hibridização de *microarrays*, de transcritos diferencialmente expressos durante a maturação do colmo em cana-de-açúcar. É considerado um dos primeiros estudos sobre a compreensão do metabolismo de acúmulo de açúcar em espécies do gênero *Saccharum*. O segundo estudo também adotou a união de técnicas de sequenciamento de ESTs e *microarrays* para identificar genes relacionados com características quantitativas de interesse em populações segregantes de cana-de-açúcar. Os autores discutem que devido à complexidade genômica da cana-de-açúcar, esta estratégia pode ser eficiente na identificação de genes candidatos que controlam o metabolismo de acúmulo de sacarose, por exemplo. O terceiro estudo destes autores discute que o acesso à coleção de ESTs de cana-de-açúcar foi fundamental para o desenvolvimento da primeira ferramenta comercial de estudos do perfil de expressão gênica em cana-de-açúcar. Foi desenvolvido um *chip* de genotipagem array da *Affymetrix* chamado *GeneChip® Sugarcane Genome Array* (Casu et al., 2007), que foi utilizado em estudos de associação entre os transcritos e o metabolismo de parede celular e a maturação do colmo em cana-de-açúcar. Manners & Casu (2011) ao analisarem as regiões funcionais do genoma de cana-de-açúcar, discutiram que o seu transcritoma é complexo e inclui transcritos de diferentes grupos de homo(eo)logia. Esta complexidade do transcritoma é reflexo dos elevados índices de ploidia apresentados pelas cultivares comerciais de cana-de-açúcar.

Houve um esforço para unificar os bancos de dados públicos de sequências de ESTs para cana-de-açúcar, unindo principalmente os dados genômicos produzidos no Brasil pelo projeto SUCEST (Vettore et al., 2001), na África do Sul (Carson & Botha, 2000) e na Austrália (Casu et al., 2001). A união destes bancos de dados produziu um banco de dados mais completo chamado de SoGI (*Saccharum officinarum Gene Index*), onde houve a tentativa de montar dos ESTs em sequências maiores chamadas de *Tentative Consensus* (TCs) (Quackenbush et al., 2000). A última atualização deste banco de dados revelou a existência de 116588 *contigs*, divididos em 40016 TCs e 76572 *singletons* de ESTs. Este banco de dados de ESTs e TCs representa uma ferramenta poderosa para obtenção e anotação de sequências gênicas para cana-de-açúcar (Souza et al., 2011).

2.4.3 Estudos de genômica comparativa

Sabe-se que um dos casos mais relatados de sintenia e colinearidade genômica acontecem entre as espécies da família Poaceae (Gale & Devos, 1998; Freeling, 2001; Paterson et al., 2009), principalmente quando se compara espécies de subfamílias específicas. Por exemplo, o sorgo e o milho apresentam o mesmo número de cromossomos (n = 10), embora se saiba que o milho sofreu um evento de duplicação genômica após a sua divergência (Swigonova et al., 2004). Paterson et al. (2004) e Paterson et al. (2009) evidenciaram que muitos dos eventos recentes de duplicação genômica sofridos por *S. bicolor* são compartilhados com outras espécies de cereais. A ocorrência de muitos eventos de condensação de regiões genômicas pode ser a explicação para a evidência de que um simples braço dos cromossomos 10 e 5 em milho corresponderem inteiramente aos cromossomos 6 e 4, em sorgo, respectivamente (Bowers et al., 2003). Devos & Gale (2000) realizaram um estudo de genômica comparativa entre quatro subfamílias de gramíneas e concluíram haver uma conservação da ordem de disposição dos genes nas diferentes espécies, além de que é possível identificar e caracterizar um genoma ancestral existindo entre as subfamílias, principalmente dentro do grupo Panicoidae, do qual cana-de-açúcar faz parte. Houve uma espécie ancestral, carregando combinações alélicas específicas das gramíneas, a partir da qual a dispersão adaptativa deste grupo taxonômico aconteceu.

Glaszmann et al. (1997), utilizando o mapeamento genético através de sondas de locos RFLP, já evidenciaram a existência de sintenia entre as espécies da família Poaceae. Estes autores, ao analisar as sondas na cultivar de cana-de-açúcar R570, mostraram a correlação genética entre grupos de ligação em cana-de-açúcar e sorgo. Grivet et al. (1996), mostraram existir um elevado nível de sintenia e colinearidade entre os dois parentais (*S. officinarum* e *S. spontaneum*) formadores das variedades comerciais de cana-de-açúcar. Estudos mais completos como o de Jannoo et al. (2007), conseguiram identificar, através da comparação de genes ortólogos, regiões homólogas entre cana-de-açúcar e outras espécies de gramíneas, mostrando que o genoma de *Saccharum* spp. possui uma estabilidade genômica, mesmo com elevados índices de poliploidia. Assim, a identificação precisa de genes ortólogos entre espécies filogeneticamente próximas e a

distinção destes genes dos genes parálogos, faz-se necessário uma compreensão de regiões que apresentam sintenia e colinearidade genômica entre as espécies.

Em um recente e impactante estudo de caracterização e anotação de BACs (*Bacterial Artificial Chromossome*), Setta et al. (2014) caracterizaram mais de três mil BACs de eucromatinas referentes a cultivar australiana R570. Um conjunto de1.400 proteínas foram anotadas, além da caracterização das regiões repetitivas destas eucromatinas. Análises de RNA-seq foram utilizadas para explorar os padrões de expressão gênica e as vias metabólicas relacionadas ao metabolismo da sacarose. Este trabalho pode ser considerado um dos maiores estudos genômicos em cana-de-açúcar, pois fornece uma quantidade de dados importantes para a compreensão da estrutura do genoma de uma das espécies com maior nível de complexidade genômica relatado. Os autores mostraram elevada semelhança genômica entre cana-de-açúcar e *Sorghum bicolor*. A elevada quantidade de genes ortólogos entre as duas espécies e a existência dessemelhança genômica quando aos elementos transponíveis e regiões genômicas não caracterizadas corroboram os estudos de Paterson et al. (2004), Paterson (2005) e Paterson et al. (2009). Foi identificada também, uma variação genômica expressiva em regiões gênicas e não gênicas entre os cromossomos hom(e)ologos da espécie, mostrando evidências aos eventos de duplicação genômica sofrida pela espécie e o comportamento de elementos transponíveis no genoma da espécie, aumentando o número de genes parálogos e a diversidade alélica para um mesmo loco gênico em cana-de-açúcar.

2.4.4 Identificação e caracterização de marcadores moleculares

Juntamente com regiões de substituição de um único nucleotídeo (SNPs) e inserções e/ou deleções em segmentos de DNA, as regiões de microssatélites são muito utilizadas em estudos que objetivam identificar e explorar o polimorfismo genético existente dentro e/ou entre populações e são classificados entre os principais tipos de polimorfismo genético existente (Mammadov et al., 2012). Os marcadores microssatélites são utilizados principalmente em estudos de caracterização da diversidade genética e como marcadores na construção de mapas genéticos densos que auxiliam na tomada de decisões

em programas de melhoramento de plantas ou animais que utilizam procedimentos de análises de QTL, estudos de associação e seleção genômica, por exemplo.

Regiões de microssatélites podem ser identificadas em regiões de DNA repetitivo, regiões intergênicas, regiões de íntrons e em regiões gênicas (Varshney et al., 2005), sendo este último caracterizado como EST (*Expressed Sequence Tag*) ou GBM (*Gene Based Markers*) e possuem importância crucial por estarem relacionados com o transcritoma da espécie em questão (Gao et al., 2003; Varshney et al., 2005), podendo ser utilizados na identificação e caracterização de genes candidatos para futura utilização em estudos e/ou tecnologias de engenharia genética. Primers que permitam a amplificação destas regiões de microssatélites e consequentemente o estudo de polimorfismos de interesse podem ser desenhados para locos específicos. A transferibilidade entre primers de amplificação de regiões microssatélites entre espécies filogeneticamente próximas é uma alternativa viável e aplicável a espécies como cana-de-açúcar e sorgo, por exemplo (Cordeiro et al., 2001, Decroocq et al., 2003, Gupta et al., 2003, Sasha et al., 2004, Yadav et al., 2008).

Marcadores microssatélites genômicos ou derivados de regiões gênicas foram identificados e descritos para cana-de-açúcar através do enriquecimento de bibliotecas (Cordeiro et al., 2001; DaSilva, 2001; Parrida et al., 2006; Parrida et al., 2009). Muitos destes locos, utilizados em análises genético-genômica da espécie foram obtidos de projetos como o *UniGene derived Sugarcane Microsatellites* (UGSM) e o *Sugarcane Enriched Genomic Microsatellites* (SEGM). Singh et al. (2010) utilizaram microssatélites obtidos nestes projetos para avaliar a diversidade genética em 84 genótipos de *S. barberi*, *S. spontaneum* e *S. officinarum* através de estimativas do conteúdo de informação polimórfica (PIC – *Polymorphism Information Content*). Os padrões de agrupamentos identificados pelos autores sugerem alguns genótipos como interessantes genitores em programas de melhoramento da espécie. Outros 387 locos SSR também derivados dos projetos UGSM e SEGM foram utilizados em estimativas de diversidade genética para genes relacionados ao conteúdo de açúcar em seis cultivares comerciais de um programa de melhoramento indiano. Foram encontrados 158 microssatélites robustos e polimórficos para uma importante das mais importantes características fenotípicas da espécie. Cardoso-Silva et al. (2014), ao encontrar 5.106 sequências simples repetidas em regiões transcritas

do tecido vegetal de folhas em seis variedades comerciais de cana-de-açúcar, avaliando 72.269 unigenes.

Os marcadores SNPs vem ganhando destaque nos estudos de genética vegetal nos últimos quinze anos, devido ao baixo custo de obtenção, a grande quantidade de marcadores espalhados no genoma e capacidade destes marcadores explorarem o tipo de polimorfismo genético mais basal que se possa existir: a substituição nucleotídica através de mutações pontuais. Em cana-de-açúcar, Bundock et al. (2009), resequenciaram, usando a plataforma 454 (pirossequenciamento), regiões genômicas de uma população de mapeamento de duas variedades comerciais australianas com o objetivo de identificar SNPs ligados a uma característica quantitativa. Como resultado verificaram a presença de SNPs a cada 35 pb, sendo a transição o tipo de mutação mais frequente. A cobertura de sequenciamento ficou próxima a 300X e a média de tamanho de reads produzidos foi de 220 bases. Foram encontrados 1.632 SNPs para genótipo Q165 enquanto 1.013 SNPs foram encontrados para o parental feminino IJ76-514 (*S. officinarum*). Foram testados 225 SNPs candidatos e 93% foram validados como polimórficos. Com o uso de *Sequenom MassArray* 209 dos 225 candidatos a SNPs para as duas espécies foram validados. Cardoso-Silva et al. (2014) analisaram o transcritoma foliar de cana-de-açúcar através da metodologia de RNA-seq e identificaram pouco mais de 708 mil SNPs espalhados em cerca de 72 mil unigenes.

3 MONTAGEM DO TRANSCRITOMA DE CANA-DE-AÇÚCAR (*Saccharum* spp.) UTILIZANDO DADOS DE SEQUENCIAMENTO DE NOVA GERAÇÃO

RESUMO

A cana-de-açúcar é uma das principais espécies cultivadas no mundo devido à sua eficiência de conversão de carbono atmosférico em biomassa. Devido à elevada quantidade de elementos repetitivos e os vários eventos de poliploidização, o genoma da espécie ainda não foi montado e anotado, diferentemente de outras gramíneas de interesse agronômico. Assim, as informações do transcritoma da espécie se tornam ainda mais úteis por dar suporte as iniciativas de análises genômicas. O transcritoma de cana-de-açúcar foi montado a partir do sequenciamento Illumina de bibliotecas *paired-ends* de cinco órgãos distintos da planta, obtidos de uma amostra de trinta clones elite. Os dados de RNA-seq passaram por análises de controle de qualidade e normalização. O software Trinity foi utilizado para montagem e a qualidade do *assembly* foi avaliada por estimativas de treze parâmetros diferentes. Os *scaffolds* obtidos identificados como *ORFs* completas foram anotados conforme os termos do *Gene Ontology*. O transcritoma obtido compreendeu 178 Mb, distribuídos em 131.831 *scaffolds*, representando 61.225 genes. O tamanho médio dos transcritos foi de 1.350 pb, com valor de N50 igual a 1.667 pb. A distribuição do tamanho dos scaffolds mostrou que grande maioria (99,3%) teve tamanho entre 500 e 5000 pares de bases. Cerca de 32 mil transcritos são exclusivos de cana-de-açúcar e há um indício de que a grande maioria deles, por não apresentar *ORFs* completas, pode ser caracterizado como RNAs longos e não codificantes (lncRNAs). Um total de 1.250 transcritos, identificados como *ORFs* completas, não apresentaram similaridade com sequências do banco de dados do NCBI, sendo considerados novas regiões transcricionalmente ativas (nTARs). O transcritoma de cana-de-açúcar obtido neste estudo possui qualidade de dados e de análise suficiente para ser considerado um transcritoma de referência para as espécies de *Saccharum* spp.

Palavras-chave: *Saccharum*; transcritoma; RNA-seq; *de novo assembly*

ABSTRACT

Sugarcane (*Saccharum* spp.) is one of the most important crops for global agribusiness due to its high energy conversion efficiency from photosynthesis into biomass. Due to the high amount of repetitive DNA elements added to several polyploidization events in its evolutionary history, the sugarcane genome has not yet been sequenced, making the information about its transcriptome the most useful tool for supporting sugarcane genomic analysis. A *de novo* draft assembly of sugarcane transcriptome was generated using paired end libraries from Illumina sequencing from five different plant organs collected from a sample of thirty elite clones. The sequencing data was submitted to quality control and normalization analyses. The draft transcriptome was assembled using Trinity package. The assembly quality was accessed through thirteen different estimated parameters. The scaffolds identified as complete ORFs were annotated according to GO terms. The draft sugarcane transcriptome was assembled with a total size of 178 Mb comprising 131,831 scaffolds related to 61,225 genes. The average size of the transcripts was 1,350 bp, whereas the value of N50 was 1,667 bp. The distribuition of scaffolds length showed that the most scaffolds (99.3%) are between 500 and 5000 base pair. Near of 32 hundred transcripts are exclusive of sugarcane and there are some evidence a huge amount of this transcripts are characterized long non-coding RNAs (lncRNAs), because its do not have complete ORFs. A total of 1,250 transcripts identified as complete ORFs, showed no similarity to sequences in NCBI database and can be considered new Transcripted Active Regions (nTARs). Annotation using the KEGG database identified 234 transcripts participating in the metabolism of sucrose and starch, an important metabolic pathway for understanding the relationship between photosynthesis rates and sucrose accumulation in the stalks. The identification of genomic regions that control agronomic traits is important step that enables the use of genome or transcriptome information in plant breeding procedures. The sugarcane transcriptome draft assembly proposed in this study has a quality data and consistent bioinformatic analysis that allow its transcriptome could be considered a reference transcriptome to *Saccharum* spp.

Key-words: *Saccharum*; transcriptome; RNA-seq; *de novo* assembly

3.1 INTRODUÇÃO

A cana-de-açúcar (*Saccharum* spp.) é uma das espécies cultivadas mais importantes para o agronegócio mundial. O Brasil é o maior produtor de cana-de-açúcar e o crescimento de produtividade nas safras brasileiras tem se mostrado contínuo ao longo dos últimos vinte anos, com um aumento de cerca de 400 milhões de toneladas neste período (FaoStats, 2013). A cana-de-açúcar se caracteriza por possuir uma enorme complexidade genômica, devido, principalmente, aos elevados níveis de poliploidia e aneuploidia sofrido pelos genitores (*S. officinarum* e *S. spontaneum*) que deram origem às cultivares modernas híbridas de cana-de-açúcar (Hermann et al., 2012). A complexidade genômica pode dificultar a compreensão de aditividade alélica e o seu emprego no melhoramento genético de características agronômicas de interesse, fazendo com que os ganhos genéticos para características quantitativas ao longo dos ciclos de melhoramento sejam pouco expressivos. Dal-Bianco et al. (2011) mostraram que os ganhos genéticos para acúmulo de açúcar aumentam de maneira pouco expressivacrescem em ritmos insignificantes ao ano. As técnicas atuais de sequenciamento de DNA permitem obter informações relevantes sobre os constituintes genéticos que controlam a expressão de caracteres agronômicos de interesse, aumentando a compreensão sobre a genética e genômica de espécies importantes (Seeb et al., 2011). Considerando a elevada quantidade de elementos genéticos repetitivos e a dificuldade de compreender a associação destas regiões com as características fenotípicas de interesse, o estudo do transcritoma se torna uma alternativa bastante atraente neste contexto (Wang et al., 2009; Garber et al. 2011).

A compreensão sobre as regiões funcionais do genoma de uma espécie são fundamentais para a correta interpretação dos elementos genéticos responsáveis pela produção de proteínas, além de revelar os constituintes moleculares presentes em células e tecidos. O transcritoma representa o conjunto completo de transcritos de uma célula e sua quantificação em um estágio específico de condições fisiológicas (Wang et al., 2009). A metodologia de RNA-seq é caracterizada pela sua elevada qualidade nas análises que objetivam o entendimento do perfil do transcritoma de uma espécie a partir de dados genéticos fornecidos pelo sequenciamento de alto desempenho do mRNA ou cDNA. Esta metodologia é capaz de fornecer medidas mais seguras sobre a quantidade, o perfil e a

orientação de transcritos produzidos por tecidos fisiológicos em condições ambientais específicas quando comparada com outros métodos de análise de transcritoma, como as técnicas baseadas em *microarrays*, por exemplo (Nagalakshmi et al., 2010; Dillies et al., 2012).

A metodologia de RNA-seq aumentou significativamente a qualidade das análises de transcritomas, permitindo uma aplicação desta metodologia a diversas espécies de procariotos e eucariotos (Wang et al., 2009). Além disso, as análises de RNA-seq são realizadas com resolução de SNPs, isto é, a fronteira entre dois transcritos pode ser discriminada em nível de nucleotídeos (pares de bases) e genes diferencialmente expressos podem ser caracterizados por variações alélicas específicas (Haas et al., 2013).

A montagem de um transcritoma pode ser realizada por duas alternativas mutuamente excludentes e a escolha de uma delas dependerá da existência prévia ou não de informações genômicas de referência para a espécie. A disponibilidade de genomas de referência significa que os reads (sequências curtas de DNA e/ou RNA sequenciada através das plataformas de sequenciamento de nova geração) de mRNA/cDNA poderão ser mapeados contra um genoma de referência, de modo que o seu posicionamento e orientação serão definidas com base nessa referência. Esta metodologia é conhecida como abordagem de mapeamento (*mapping-first approach*). No entanto, caso esta informação não esteja disponível, será necessário adotar uma metodologia alternativa conhecida como montagem do transcritoma *de novo* (*de novo transcriptome assembly* ou *assembly-first approach*) (Grabherr et al., 2011).

As ferramentas de bioinformática são específicas para cada uma destas duas metodologias de análise de transcritomas. Garber et al. (2011) e Trapnell et al. (2012) defendem o uso de pacotes computacionais como o TopHat e o Cufflinks para montagem *de novo* de transcritomas. Trata-se de softwares baseados na metodologia de mapeamento de reads em genomas de referência (*mapping-first approach*), assim como a ferramenta Scripture (Guttman et al., 2010). Os softwares TopHat-Cufflinks e Scripture são capazes de analisar reads provenientes do sequenciamento de mRNA e montar o transcritoma da espécie, além de inferir o número de unigenes, o número e a estrutura de *splicings* alternativos e a quantificação de transcritos diferencialmente expressos. No entanto, existem metodologias que não utilizam um genoma de referência (*assembly-first approach* ou montagem *de novo*) para a montagem do transcritoma e estão implementadas em

softwares como, por exemplo o Trans-ABySS (Birol et al., 2009), o SOAPdenovo-Trans (Li et al., 2009), o Velvet-Oases (Schulz et al., 2012) e o Trinity (Grabherr et al., 2011). Estes algoritmos de *assembly de novo* de transcritomas foram comparados e suas performances avaliadas em diversos trabalhos (Groba & Burgos 2010; Hass et al., 2013; O'Neil & Emrich, 2013). O algoritmo implementado na plataforma do Trinity vem se destacando como a principal ferramenta de bioinformática utilizada na montagem *de novo* de transcritoma baseada na construção de grafos de Bruijn.

Neste contexto, o presente trabalho tem como objetivo a utilização da metodologia de RNA-seq juntamente com o pacote computacional Trinity (Grabherr et al., 2011) para montagem *de novo* de um *draft assembly* para o transcritoma de cana-de-açúcar, utilizando cinco diferentes órgãos vegetais coletados de uma amostra de trinta clones elites do programa de melhoramento genético da Ridesa/UFG.

3.2 MATERIAL E MÉTODOS

3.2.1 Material vegetal e sequenciamento do mRNA

O material vegetal foi obtido a partir de 30 clones elites selecionado aleatoriamente de uma população de melhoramento formada por 48 genótipos em fase final de avaliação pelo programa de melhoramento genético da cana-de-açúcar da Ridesa/UFG, mantida em campo experimental da Escola de Agronomia da Universidade Federal de Goiás. Esta população era formada por indivíduos adultos que foram coletados em aproximadamente dez meses após o transplantio. Cinco tipos diferentes de órgãos vegetais foram coletados de cada um dos 30 clones elites. Os órgãos amostrados foram: colmo, gemas laterais, plântulas, folhas e gemas apicais.

Foi amostrada a mesma quantidade de material vegetal para cada órgão coletado a partir dos 30 clones elites. Imediatamente após a coleta, o material vegetal foi armazenado em freezer a -80ºC. Para cada órgão, todo o material coletado foi macerado juntamente utilizando nitrogênio liquido, formando cindo tipos de bibliotecas distintas referente a cada órgão vegetal. O RNA total de cada órgão foi extraído em *bulk* (o bulk foi

formado antes da extração, na etapa de maceração) constituído por todos os 30 genótipos utilizando o kit *RNeasy® Plant Mini Kit* (Qiagen)[7] . O RNA extraído foi tratado com a enzima DNAse para evitar contaminação por DNA. A integridade, a qualidade e a quantidade de RNA extraído foram inferidas utilizando-se o aparelho Bioanalyzer 2100 (Agilent). As amostras de RNA de alta qualidade e integridade física foram enviadas em colunas (RNAstable™ Biomatrica)[8] para a empresa BGI Co. Ltd. para o sequenciamento. O preparo das bibliotecas de sequenciamento foi realizado através do *TruSeq Stranded mRNA* que isolam os mRNA com base na sua cauda poli-A através de beads de oligonucleotídeos dT. O sequenciamento de bibliotecas *paired ends* foi feito a partir de moléculas de cDNA e realizado utilizando a plataforma de NGS da Illumina HiSeq2000, utilizando ½ *lane* para cada biblioteca, com exceção da biblioteca de gema apical que foi sequenciada em um único *lane*. As moléculas de cDNA foram normalizadas a partir da técnica DSN (*Duplex-Specific Thermostable Nuclease*) com o objetivo de amostrar transcritos pouco abundantes.

3.2.2 Controle de qualidade das sequências

O procedimento de determinação da qualidade dos reads produzidos pela plataforma Illumina foi realizado em duas etapas. Na primeira, realizada pela própria empresa contratada para o sequenciamento, foram eliminados os adaptadores e os reads que continham mais de 50% de suas bases de baixa qualidade (valor de qualidade ≤ 5). A segunda etapa da análise de controle de qualidade, suportada pelas estimativas do fastQC (Andrews, 2010), foi realizada pela ferramenta Trimmomatic (Bolger et al., 2014), através dos seguintes parâmetros: LEADING:39 TRAILING:30 SLIDINGWINDOW:4:30 MINLEN:36 HEADCROP:15. Quinze pares de bases iniciais de todos os reads, identificados como fragmentos com contaminantes (pelo seu conteúdo GC), foram inicialmente eliminados. Posteriormente, nucleotídeos com qualidade de sequenciamento baixa foram eliminados, de modo a permitir um erro de sequenciamento a cada mil pares

[7] Informação disponível em: http://www.qiagen.com/products/rnastabilizationpurification/rneasysystem
[8] Informação disponível em: http://www.biomatrica.com

de bases. Após esse filtro, reads que tiveram tamanhos menores que 50 pares de base também foram eliminados. Os dados finais foram divididos em arquivos de reads pareados e arquivos de reads "órfãos".

3.2.3 Normalização dos reads sequenciados

A normalização dos dados é uma etapa importante no tratamento das sequências produzidas pelas plataformas de nova geração, pois estas plataformas introduzem erros de sequenciamento e de amostragem. Os algoritmos normalizadores sistematizam a cobertura de sequenciamento, eliminam reads redundantes e os erros de sequenciamento, o que melhora a eficiência computacional dedicada à construção e resolução dos grafos de Bruijn sem afetar o conteúdo dos *contigs* e *scaffolds* produzidos (Brown et al., 2014).

Os reads de alta qualidade pareados (PE) e órfãos (SR) das cinco bibliotecas de sequenciamento foram normalizados separadamente por duas metodologias diferentes antes da montagem do transcritoma de cana-de-açúcar. Foram utilizados o normalizador Khmer (Crusoe et al., 2014), que faz uso de uma normalização pela mediana da abundancia dos k-meros (Brown et al., 2014), e o normalizador disponibilizado pela própria plataforma de análise do Trinity (Grabherr et al., 2011), caracterizada como um normalizador *in silico* que também utiliza da abundância dos k-meros além da profundidade de cobertura do sequenciamento.

3.2.4 Montagem *de novo* do transcritoma de cana-de-açúcar

O *draft assembly* do transcritoma de cana-de-açúcar foi obtido a partir de uma abordagem *de novo*, através do pacote computacional Trinity (Grabherr et al., 2011). Todas as cinco bibliotecas dos diferentes tipos de órgãos vegetais foram utilizadas na montagem. Somente os *contigs* com comprimento maior de 500 pb foram mantidos nas análises

subsequentes. Os parâmetros de qualidade da montagem *de novo* foram obtidos utilizando um *script* (*assemblathon_stats.pl*) disponibilizado pelo grupo Assemblathon (Earl et al., 2011) e os arquivos de resultados do Trinity. Para fins de comparação entre as diferentes estratégias de tratamento inicial dos dados foram realizadas análises comparativas entre o transcritoma de cana-de-açúcar montado (sequências *query*) e o transcritoma de referência de *Sorghum bicolor* v2.1 (Paterson et al., 2009) (sequências *subject*) através do algoritmo blastx v2.2.30 (Altschul et al., 1997). Assim, a melhor montagem foi definida com base em treze parâmetros, sendo oito referentes a própria montagem e cinco referentes aos resultados da análise blastx. Os parâmetros estimados com base na montagem propriamente dita foram: I) número de *scaffolds* (transcritos + isoformas); II) número de genes (transcritos); III) número de *scaffolds* acima de 1 Kb; IV) % de *scaffolds* acima de 1 Kb; V) tamanho médio dos *scaffolds*; VI) N50; VII) número de *scaffolds* acima de 10 Kb e VIII) Total de pares de bases no *assembly*. Com base nos resultados da análise com o blastx foram estimados: I) número de proteínas com cobertura de 100%; II) % média de cobertura; III) % de hits no transcritoma de *S. bicolor*; IV) comprimento médio dos hits e V) probabilidade média de identidade.

A qualidade do transcritoma montado também foi avaliada pelo alinhamento contra bancos de dados de sequências de possíveis contaminantes. O alinhador Bowtie2 (Langmead et al., 2009) foi utilizado no alinhamento do transcritoma montado (sequências *query*) contra cinco bancos de dados de possíveis contaminantes criados como subamostras a partir do banco de dados do GeneBank (VexScreen) e do banco de dados SILVA (Quast et al., 2013). Os bancos de dados de possíveis contaminantes forma: 1) cpDNA de plantas, 2) o genoma de *Escherichia coli*, 3) mtDNA plantas, 4) rRNA de angiospermas e 5) banco de dados de possíveis vetores.

Foi realizada ainda uma análise comparativa com o banco de dados SoGI (*Saccharum officinarum* Gene Index) (Quackenbush et al., 2000) (sequência *query*) e o transcritoma obtido de cana-de-açúcar montado (sequência *subject*) com o objetivo de estimar a porcentagem de ESTs e transcritos existentes no banco de dados SoGI coberta pelo *draft assembly* produzido. Esta comparação também permitiu inferir a quantidade de transcritos representados na montagem do transcritoma de cana-de-açúcar obtido que não estão presentes no banco de dados do SoGI.

O draft assembly do transcritoma de cana-de-açúcar também foi comparado contra um banco de dados chamado *"Grass_DB"*, formado pelo transcritoma de seis espécies de gramíneas (*Sorghum bicolor* v2.1 (Paterson et al., 2009), *Setaria itálica* v1.0 (Zhang et al., 2012), *Oryza sativa* v1.0 (Matsumoto et al., 2005), *Zea mays* v1.0 (Hirsch et al., 2014), *Brachypodium distachyon* v1.0 (Lucas et al., 2009) [9] e *Panicum virgatum* v1.0 (JGI, 2014) [10]), contidas no banco de dados do Phytozome (GOODSTEIN, et al., 2011).

Diagramas de Venn, representando os resultados das comparações com os diferentes bancos de dados, foram produzidos usando o pacote do software R (R Core Team, 2013) "VennDiagram" (Chen & Boutros, 2011).

O pipeline de análise contendo os softwares utilizados em cada etapa das análises de bioinformática está disponibilizado no Apêndice A.

3.3 RESULTADOS E DISCUSSÃO

3.3.1 Estatísticas descritivas e normalização dos dados

O sequenciamento das bibliotecas obtidas a partir dos cinco órgãos vegetais amostrados produziu um total de 809.858.868 *reads Paired-Ends* (PE), de cem pares de bases cada, totalizando cerca de 80 Gb de sequências de mRNA de cana-de-açúcar (Tabela 3). Após a eliminação dos *reads* de baixa qualidade, 743.503.344 *reads* foram mantidos (91,80%), dos quais 94,79% eram *reads* pareados (PE). Tanto os reads PE quanto os SR possuíam um tamanho médio de 75 pb. O tamanho médio dos insertos de sequenciamento foi de 141 pb.

[9] Sequências submetidas diretamente no banco de dados do Phytozome.
[10] Sequências submetidas diretamente no banco de dados do Phytozome.

Tabela 3. Resumo dos resultados de sequenciamento Illumina do mRNA de cinco órgãos vegetais de cana-de-açúcar utilizados para obtenção do *draft assembly* do transcritoma. Os dados eliminados referem-se a quantidade de dados eliminados pelas análises de controle de qualidade. A biblioteca de gema apical foi sequenciado em um *lane* de sequenciamento enquanto as outras em ½ *lane*.

Órgãos Vegetais	Tamanho dos *reads* (bp)	Dados do sequenciamento (Gb)	Dados eliminados (Gb)	GC (%)	Q30 (%)
Colmo	100	13,14	5,02	50,75	88,76
Gema lateral	100	13,80	5,19	49,50	89,70
Plântulas	100	14,00	5,41	50,37	88,79
Folhas	100	12,88	5,12	53,00	88,26
Gema apical	100	27,08	11,77	52,87	83,32

Um primeiro *assembly* foi conduzido após a eliminação de reads de baixa qualidade e normalizado pelo Khmer, totalizando 257.284.911 reads (31,77% do total) utilizados no *assembly*. Um segundo *assembly* foi realizado também após a eliminação dos *reads* de baixa qualidade e normalizados pelo algoritmo *in silico* disponibilizado na ferramenta Trinity, totalizando 384.690.774 reads (51,25% do total) utilizados no *assembly*.

3.3.2 O *de novo draft assembly* do transcritoma de *Saccharum* spp.

A comparação entre as duas montagens conduzidas pelo Trinity, utilizando duas estratégias diferentes de normalização dos dados, mostra que ambos os normalizadores produzem resultados muito semelhantes (Tabela 4). No entanto, o normalizador *in silico* do Trinity apresentou uma superioridade quanto à avaliação do número de genes estimado e a probabilidade média de identidade em relação ao transcritoma de *S. bicolor*. Além disso, há discussões que mostram que o normalizador do Trinity já possui o algoritmo do normalizador do Khmer conhecido como "*diginorm*" (http://ivory.idyll.org/blog/trinity-in-silico-normalize.html).

O *draft* do transcritoma de cana-de-açúcar foi montado com um tamanho de 178 Mb distribuídos em 131.831 *scaffolds* relativos a 61.225 genes. Paterson et al. (2009)

estimaram o tamanho médio das regiões de eucromatina em *S. bicolor* em cerca de 252 Mb, as quais eram formadas por 34.496 genes. Em *O. sativa*, estas regiões foram estimadas em 309 Mb contemplando 37.554 genes (Matsumoto et al., 2005). Hirsch et al. (2014) amostraram plântulas de 503 linhas endogâmicas de milho (*Z. mays*) e usando a técnica de RNA-seq identificaram 31.398 transcritos. Estima-se que o tamanho do transcritoma de milho seja de 177 Mb (Messing et al., 2004). Estes resultados sugerem que existem mais genes em cana-de-açúcar do que em outras espécies filogeneticamente próximas da família das gramíneas.

O tamanho médio dos *scafoolds* obtidos foi de 1.350 pb, variando entre 501 pb e 15.506 pb. O valor de N50 obtido foi de 1.667 pb e a média do conteúdo GC foi de 51,29%.

Tabela 4. Estimativas dos parâmetros dos *draft assemblies* do transcritoma de cana-de-açúcar.

Parâmetros do *assembly*	Normalizadores	
	Khmer (Crusoe et al. 2014)	Trinity (Grabherr et al. 2011)
Parâmetros do *assembly*		
1. Número de *scaffolds* (transcritos + isoformas)	151.806	131.831
2. Número de genes (transcritos)	60.073	61.225
3. Número de *scaffolds* maiores que 1 Kb	83.542	69.131
4. % de *scaffolds* maiores que 1 Kb	55,03	52,44
5. Número de *scaffolds* maiores que 10 Kb	8	18
6. Tamanho médio dos *scaffolds*	1.390	1.350
7. N50	1.727	1.667
8. Total de pares de bases no assembly (Mb)	211,03	177,98
Parâmetros da análise blastx		
9. Número de proteínas com cobertura 100%	10.287	10.624
10. % media de cobertura	17,12	17,35
11. % de hits no transcritoma de *S. bicolor*	62,46	64,65
12. Comprimento médio dos hits (pb)	294,04	309,65
13. Probabilidade média de identidade	62,46	83,74

O número de genes de cana-de-açúcar permanece em aberto considerando que cinco diferentes órgãos vegetais foram amostrados e que o *assembly* apresentou cerca de 61 mil genes, em contraste com a montagem do transcritoma foliar de seis variedades de cana-de-açúcar com mais de 72 mil genes e 119.768 transcritos obtida por Cardoso-Silva et al. (2014). Houve uma diferença entre as duas montagens referente ao tamanho dos *scaffolds* amostrados. A montagem aqui proposta foi feita utilizando somente *scaffolds*

acima de 500 pb, enquanto na montagem do transcritoma foliar foi realizada utilizando *scaffolds* acima de 300 pb (Cardoso-Silva et al., 2014). Esses autores também estimaram o tamanho médio dos genes em 921 pb, com uma medida de N50 igual a 1.367 pb. Cerca de 44% dos unigenes revelados no transcritoma foliar de cana-de-açúcar apresentaram tamanho entre 300 e 500 pb (Cardoso-Silva et al., 2014). Geralmente, transcritos menores que 500 pb são referentes a RNAs não codificantes (ncRNAs) que podem ter tamanhos variados com média ao redor de 200 pb. RNAs menores que 200 pb são caracterizados em outras classes de RNA (miRNAs, siRNAs, piRNAs). Estes RNAs pequenos em geral atuam no controle da expressão gênica e não na expressão direta de um fenótipo molecular de interesse (Perkel, 2013), entretanto, são consideradas moléculas importantes para caracterização molecular de um genoma. Neste trabalho, o *draft* do transcritoma obtido foi montado com mais de 52% dos transcritos com tamanhos acima de 1 Kb e cerca de 47% dos *scaffolds* com tamanhos entre 500 pb e 1 Kb. A Tabela 5 mostra a distribuição do tamanho dos *scaffolds*.

Tabela 5. Distribuição dos tamanhos e a porcentagem dos *scaffolds* produzidos pelo Trinity.

Comprimento dos *scaffolds* (pb)	**Total de *scaffolds***	**Porcentagem (%)**
500 – 1000	62.676	47,54
1000 – 5000	68.238	51,76
5000 – 10000	899	0,681
> 10000	18	0,013

O número detectado de sequências contaminantes no *draft* do transcritoma de referência para cana-de-açúcar foi muito baixo. Em média, 0,12% dos *reads* alinharam em bancos de dados de contaminantes, com uma variação de 0,01% de sequências de possíveis vetores, a 0,34% de sequências de mtDNA de plantas (Apêndice B). Estes resultados confirmam a eficiência das análises de qualidade inicialmente realizadas nos dados originais do sequenciamento Illumina.

O número de sequências gênicas no banco de dados SoGI é de 121.342. Desse total, 112.988 alinharam em 39.888 transcritos obtidos para cana-de-açúcar. Cerca de 30,25% do transcritoma obtido foi suficiente para representar 93,11% da totalidade das sequências gênicas do banco de dados do SoGI. Além disso, o *draft* do transcritoma de cana-de-açúcar produzido apresentou mais de 90 mil transcritos que não estão incluídos no

maior banco de dados de sequências gênicas para cana-de-açúcar (SoGI) (Figura 2). Destaca-se o fato do banco de dados SoGI conter quase a totalidade das sequências do banco de dados do SUCEST (Vettore et al., 2001). Estes resultados sugerem que o banco de dados do SUCEST não contempla a totalidade de transcritos existentes no transcritoma de cana-de-açúcar.

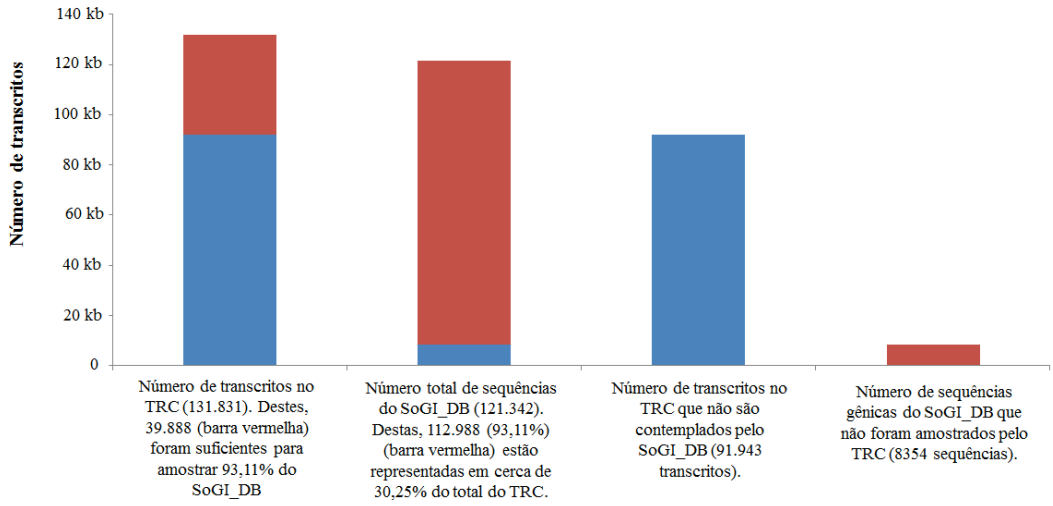


Figura 2. Representatividade do banco de dados SoGI (SoGI_DB) no *draft* do transcritoma de cana-de-açúcar (TRC), mostrando a relação entre o transcritoma proposto e o maior banco de dados público de sequências gênicas de cana-de-açúcar.

A busca por similaridade de sequências, utilizando o programa blastx, entre o transcritoma montado de cana-de-açúcar contendo 131.831 transcritos relativos a 61.225 genes contra o banco de dados SoGI, revelou que 111.527 transcritos apresentaram similaridade significativa, com pouco mais de 18 mil transcritos apresentando um alinhamento com 100% de cobertura. O alinhamento contra o transcritoma de *S. bicolor* apresentou uma similaridade significativa para cerca de 67.247 sequências, das quais, 10.624 apresentaram alinhamento com cobertura de 100%, evidenciando a similaridade entre o transcritoma das duas espécies, o que já havia sido sugerido em outros estudos de genômica comparativa (Devos & Gale, 2000; Jannoo et al., 2007; Paterson et al., 2009). O alinhamento do transcritoma foliar de cana-de-açúcar contra o genoma de *S. bicolor* feito por Cardoso-Silva et al. (2014), revelou similaridade significativa com somente cerca de 28 mil proteínas. Os resultados da busca por similaridade de sequências entre o

transcritoma obtido e o banco de dados "Grass_DB" resultou na identificação de 88 mil sequências entre os dois bancos de dados, das quais 11.923 apresentaram 100% de cobertura de alinhamento (Figura 3). Um número semelhante de transcritos foram identificados nos bancos de dados "Grass_DB" e SoGI, mas com um número um pouco maior de transcritos identificados no banco de dados "Grass_DB". No entanto, foi identificado um maior número de sequências com 100% de similaridade quando o transcritoma de cana-de-açúcar montado foi comparado com o banco de dados do SoGI. Este resultado corrobora com o elevado número de sequências genica do banco de dados do SoGI amostradas pelo transcritoma proposto.

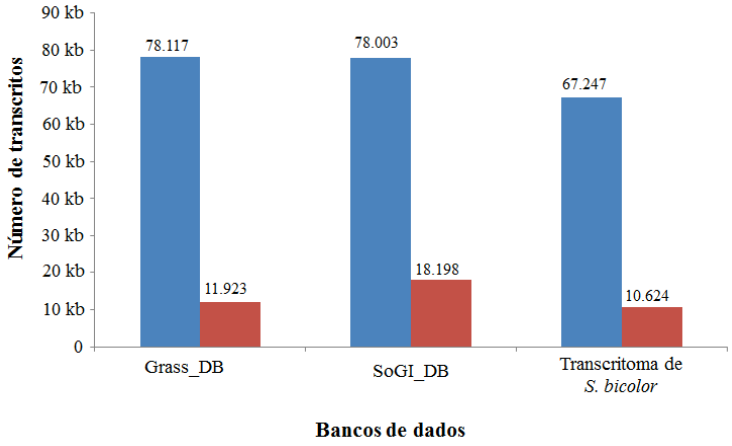


Figura 3. Resultado da análise de busca por similaridade de sequências do *draft* do transcritoma de cana-de-açúcar (sequência query) contra o banco de dados SoGI, *GrassDB* e o transcritoma de *S. bicolor*, utilizados como sequências *subject*. As barras azuis representam o total de transcritos com hits significativos (evalue $\leq 10^{-6}$), enquanto as barras vermelhas representam o número de transcritos com 100% de similaridade.

Cerca de 50 mil transcritos foram identificados quando os três bancos de dados foram analisados simultaneamente (Figura 4). A comparação do número de transcritos identificados simultaneamente entre os bancos de dados revela que 32.507 transcritos são exclusivos de cana-de-açúcar, não sendo identificado em nenhum dos três bancos de dados (Apêndice C).

A busca pelos quadros abertos de leitura (ORFs) se deu em todos os transcritos e revelou a existência de 84.180 ORFs, o que corresponde a 63,85% dos transcritos identificados. Destas, 42.932 eram ORFs completas. Entre os 32.507 transcritos exclusivos

de cana-de-açúcar, 1.381 apresentaram ORFs completas, indicando que os restantes destes transcritos exclusivos (31.126 transcritos) talvez possam ser considerados como outras classes de RNAs não codificantes, como os RNAs longos e não codantes (lncRNAs), por exemplo.

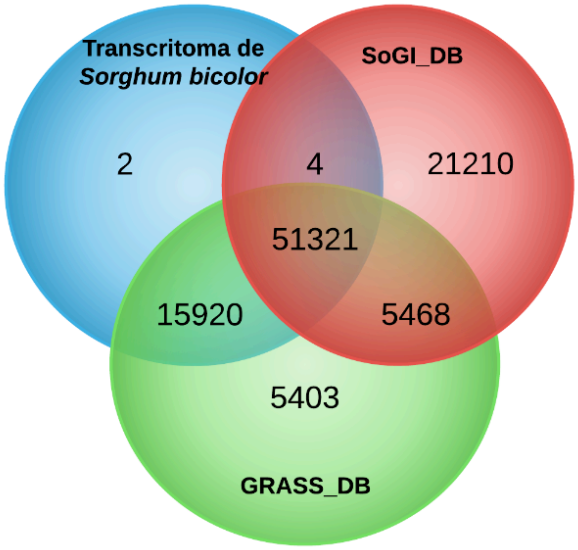


Figura 4. Diagrama de Venn representando o número de transcritos montados pelo Trinity e identificados em cada um dos três bancos de dados.

Os transcritos exclusivos de cana-de-açúcar e que apresentaram *ORFs* completas foram anotados utilizando o banco de dados nr do NCBI na sua versão 2.2.30, com o objetivo de identificar transcritos novos (nTARs – *novel Transcripts Active Regions*) ou *splicings* alternativos não descritos nos bancos de dados anteriormente utilizados (Figura 5). Entre os transcritos com *ORFs* completas, 1.250 não apresentaram *hits* significativos contra o banco de dados nr do NCBI, sugerindo que estes transcritos podem ser considerados genes novos, ainda não identificados (Apêndice E). Venturinni et al. (2013), avaliando o transcritoma de uva comercial (*Vitis vinifera*), identificaram através da metodologia de RNA-seq, 2.321 genes novos, codificadores de proteínas, em regiões não anotadas ou não montadas do genoma de referência da espécie. Xu et al. (2012) ressequenciaram cinquenta acessos de arroz (*O. sativa*) cultivado e silvestre e identificaram 1.415 genes novos de importância agronômica, além de identificarem marcadores

moleculares relacionados à estes genes. Em milho (*Z. mays*), Thiebaut et al. (2014) utilizaram a técnica de RNA-seq para explorar e identificar pequenos RNAs relacionados ao controle da expressão gênica. Um total de 25 famílias de miRNAs e 15 novos miRNAs foram identificados como resposta à relação endofítica benéfica provocada por bactérias diazotróficas. Chen et al. (2014), estudando uma importante planta medicinal (*Stevia rebaudiana*), identificaram, também através da metodologia de RNA-seq, novos genes importantes e participantes da via metabólica de produção de diterpênicos de esteviol, um importante composto bioquímico amplamente utilizado na indústria farmacêutica e alimentícia. Lu et al. (2010), ao estudar o transcritoma das duas principais subespécies de arroz (*Oryza sativa índica* e *japônica*), conseguiram identificar uma quantidade elevada de regiões transcricionais ativas novas. Entre os mais de 15 mil transcritos, declarados como novos por Lu et al. (2010), cerca de 51% (8.011) não apresentaram similaridade a sequências de nenhum banco de dados público. Estes resultados sugerem que a metodologia de RNA-seq vem se destacando no cenário atual de estudos de transcritomas como uma ferramenta útil na descoberta de novos transcritos e/ou isoformas que compõem o transcritoma de referência de espécies modelo e não modelo.

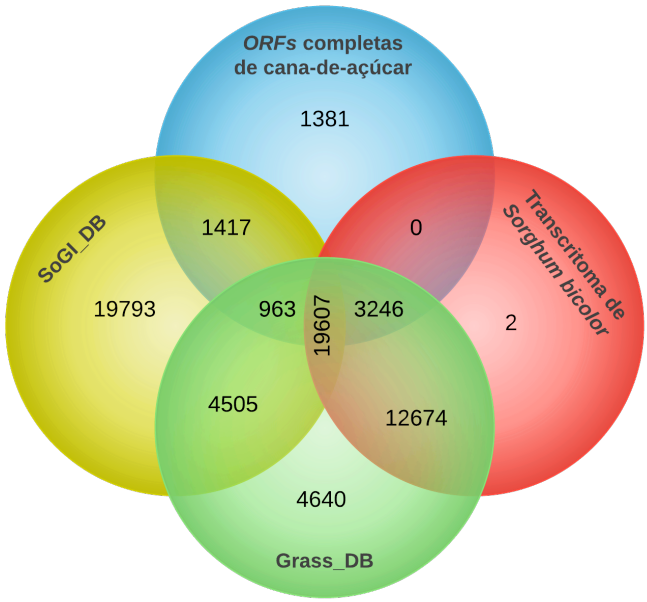


Figura 5. Diagrama de Venn mostrando a existência de 1.381 transcritos com ORFs completas, identificados no *draft* do transcritoma de cana-de-açúcar, que não apresentam similaridade às sequências depositadas nos três bancos de dados utilizados.

Paterson et al. (2009), em análises de genômica ampla pelo sequenciamento completo do genoma de *S. bicolor*, mostraram que o tamanho da sequência de nucleotídeos que formam famílias gênicas em sorgo é muito semelhante ao tamanho das mesmas famílias gênicas em *O. sativa*, *A. thaliana* e *Populus trichocarpa*. O sequenciamento de vinte BACs de cana-de-açúcar utilizando a tecnologia de NGS 454 (pirossequenciamento) revelou que cerca de 95% das regiões gênicas destes BACs eram correspondentes e idênticas a regiões genômicas de *S. bicolor* (Wang et al., 2010). Os fragmentos do genoma de sorgo cobriram cerca de 78,2% das sequências de DNA obtidas a partir dos BACs, mostrando uma elevada sintenia e colinearidade entre as duas espécies, o que segundo os autores, pode ser caracterizado como uma microcolinearidade genômica entre cana-de-açúcar e sorgo. Estes autores concluíram que o genoma de sorgo, por ser muito menos complexo, pode ser utilizado como um genoma de referência para identificação gênica e para inferências sobre caracteres de interesse agronômico em cana-de-açúcar (Wang et al., 2010).

3.4 CONCLUSÕES

Os normalizadores utilizados na análise que antecede a montagem do transcritoma apresentaram resultados bastante semelhantes com uma ligeira superioridade do normalizador *in silico* disponibilizado pela plataforma Trinity. Esta ferramenta de análise se mostrou eficiente na montagem do transcritoma de cana-de-açúcar.

Um *draft assembly* para o transcritoma de cana-de-açúcar foi gerado, a partir amostras de amostras de RNAs obtidos de cinco órgãos vegetais de um *pool* de trinta clones elites, com tamanho médio de 178 Mb distribuídos em 131.831 *scaffolds* relacionados a 61.225 genes.

O *assembly* construído foi mais rico em número de genes e mais consistente quando comparado aos demais obtidos em trabalhos que objetivaram montar um transcritoma para cana-de-açúcar e representa um passo fundamental na construção de um transcritoma de referência para *Saccharum* spp.

Existem 32.507 transcritos exclusivos de cana-de-açúcar, não sendo identificado em nenhum dos três bancos de dados. A grande maioria destes transcritos não apresentam ORFs completas e por isso há evidências de que podem ser considerados RNAs longos e não codificantes.

Um total de 1.250 transcritos não apresentaram *hits* significativos quando comparados contra o banco de dados nr do NCBI, sendo considerados transcritos novos (nTARs – *novel Transcripts Active Regions*), ainda não identificados e anotados.

A comparação do transcritoma obtido com o banco de dados do SoGI (*Saccharum officinarum* Gene Index), considerado um dos maiores bancos de dados de sequências gênicas de cana-de-açúcar, evidencia que este banco de dados não contempla a totalidade de transcritos da espécie. Aproximadamente 30% do transcritoma montado foi suficiente para cobrir cerca de 93% deste banco de dados.

Existem mais de 90 mil transcritos no transcritoma de cana-de-açúcar proposto que não estão inseridos no banco de dados do SoGI. Portanto este trabalho será útil por fornecer dados que completam os bancos de dados públicos com o objetivo de definir um transcritoma de referência para cana-de-açúcar.

4 ANOTAÇÃO E CARACTERIZAÇÃO PRELIMINAR DO TRANSCRITOMA DE CANA-DE-AÇÚCAR (*Saccharum* spp.)

RESUMO

O Brasil é o país que lidera o mercado econômico mundial de produção de açúcar e etanol derivados da cana-de-açúcar (*Saccharum* spp.), fazendo desta cultura umas das mais importantes no cenário agrícola nacional. A compreensão detalhada do transcritoma de uma espécie é importante e fornece informações básicas para o desenvolvimento de estudos posteriores de caracterização funcional de genes de interesse. O *draft assembly* obtido para o transcritoma de cana-de-açúcar foi anotado e caracterizado. A anotação dos transcritos que possuem ORFs completas foi feita através do BLAST2GO, enquanto a caracterização através da identificação de marcadores moleculares do tipo microssatélites e SNPs e pela avaliação da contribuição dos diferentes órgãos vegetais para constituição do transcritoma final. A anotação realizada através do banco de dados do KEGG identificou 234 transcritos codificantes para enzimas integrantes do metabolismo de sacarose e amido, uma importante rota metabólica para compreensão da relação entre taxa fotossintética e acúmulo de sacarose no colmo. Os cinco órgãos vegetais utilizados contribuíram igualmente para a constituição do *draft* do transcritoma de cana-de-açúcar. Foram identificadas 12.931 regiões genômicas contendo microssatélites perfeitos, com predomínio de di e tri nucleotídeos. Em média, identificou-se um SNP a cada 18 pares de bases, com mais de quatro milhões de SNPs identificados. A profundidade média de sequenciamento para identificação dos SNPs foi de 75X. A estimativa da diversidade nucleotídica para o transcritoma entre os 30 genótipos elite avaliados foi elevada (estimativa de $\pi = 0{,}931$). A identificação destes marcadores moleculares, principalmente os marcadores SNPs, fornece a possibilidade de utilização destes polimorfismos em estudos genéticos e genômicos de cana-de-açúcar, incluindo a possibilidade de desenvolvimento de aplicações, como o desenvolvimento de modelos de seleção genômica ampla.

Palavras-chave: *Saccharum*; RNA-seq; SNPs; microssatélites.

ABSTRACT

Brazil is the country that leads the world economic market of sugar and ethanol production derived from sugarcane (*Saccharum* spp.), making this one of the most important cultures in the national agricultural scenario. The effective understanding of the transcriptome of an important species allows the development of further gene functions studies. The obtained draft assembly of sugarcane transcriptome was annotated and characterized. The annotation of transcripts with complete ORFs was done using BLAST2GO suite, while transcriptome characterization by the identification of microsatellites regions, SNPs and analysis of contribution of the five different plant organs used in the assembly to its constitution. The annotation performed using the KEGG database identified 234 transcripts coding for enzymes members of sucrose and starch metabolism, an important metabolic pathway for understanding the relationship between photosynthetic rate and sucrose accumulation in the sugarcane stalk. The five plant organs used have contributed equally to the assembly. A total of 12,931 perfect microsatellites regions were found, predominantly di and tri nucleotides. On average, one SNP every 18 bp was found, with more than four million SNPs identified. The average depth sequencing to identify the SNPs was 75X. The nucleotide diversity estimate for the sugarcane transcriptome for the 30 evaluated elite clones was high (π estimate = 0.931). The identification of molecular markers, specially the SNP markers, provides the possibility of using these polymorphisms in further genetic/molecular studies in sugarcane. High-throughput genotyping techniques can be derived from this information, including the development of techniques such as genome wide selection.

Key-words: *Saccharum*; RNA-seq; SNPs, microsatellites.

4.1 INTRODUÇÃO

As variedades modernas de cana-de-açúcar (*Saccharum* spp.) foram formadas pelo cruzamento interespecífico entre *S. officinarum* x *S. spontaneum* (Hermann et al., 2012) e, normalmente, exibem mais de oito cópias homólogas de cada cromossomo de *S. officinarum* e várias cópias homólogas de cromossomos de *S. spontaneum* (Ming et al., 2008), ou seja, apresentam elevada complexidade genômica.

A cana-de-açúcar é uma espécie de extrema importância para o cenário agronômico mundial e nacional, principalmente devido a crescente demanda para substituição da matriz energética de combustíveis fósseis para combustíveis renováveis. Neste contexto, as espécies do complexo *Saccharum* se destacam pela eficiente capacidade de conversão de energia bioquímica em biomassa, através do mecanismo fotossintético C4. Atualmente, o melhoramento genético de plantas vem sendo auxiliado pelo desenvolvimento das ferramentas genômicas e o uso destas ferramentas tem-se mostrado crescente no decorrer das duas últimas décadas. Um exemplo claro é o elevado uso dos marcadores SNPs (*Single Nucleotide Polymorphism*) em diversos contextos da genética e melhoramento de plantas visando à identificação de genótipos agronomicamente superiores e mais produtivos (Mammadov et al., 2012). Outro ponto importante é a possibilidade de sequenciar genomas/transcritomas completos para espécies de interesse.

Os marcadores moleculares SNPs, capazes de identificar polimorfismos através da detecção de mutações pontuais em sequências de DNA, são os marcadores mais utilizados na atualidade, juntamente com os microssatélites (Vignal et al., 2002; Brumfield et al., 2003; Morin et al., 2004). No entanto, os marcadores SNPs possuem como uma das suas principais vantagens a possibilidade de serem utilizados nas plataformas de genotipagem de alto desempenho como os chips de genotipagem, além de que o sequenciamento de genomas/transcritomas utilizando as plataformas de NGS permite a detecção em larga escala destes marcadores (Pérez-Castro et al., 2012). A genotipagem de marcadores SNPs através do sequenciamento de alto rendimento e com elevada cobertura tem despertado muito interesse, sendo amplamente utilizada nos estudos com espécies modelo e não modelo (Seeb et al., 2011).

A identificação e validação de marcadores SNPs são etapas iniciais para utilização destes marcadores em aplicações da genômica ao melhoramento vegetal de espécies de interesse agronômico. Uma vez identificados, os SNPs podem ser convertidos em marcadores genéticos e utilizados na caracterização de populações e características fenotípicas de interesse. Devido à sua elevada abundância no genoma, mapas genéticos densos podem ser construídos e utilizados como suporte aos programas de melhoramento genético que utilizam a estratégia de Seleção Assistida por Marcadores (MAS – *Marker Assisted Selection*) ou Seleção Genômica. A identificação de SNPs em regiões genômicas já foi realizada para inúmeras espécies incluindo as principais espécies cultivadas. O sequenciamento de transcritomas através de plataformas de NGS permite uma identificação de SNPs em regiões gênicas, evitando-se assim as regiões de DNA repetitivo. Essa busca por SNPs auxiliada pelo sequenciamento de nova geração tornou-se uma técnica rápida e de baixo custo (Morozova & Marra, 2008). Esta metodologia tem sido aplicada com sucesso em diversas espécies, incluindo milho (Barbazuk et al., 2007), trigo (Parchman et al., 2010), canola (Trick et al., 2009), eucalipto (Novaes et al., 2008), a cana-de-açúcar (Bundock et al., 2009; Cardoso-Silva et al., 2014) e outras.

Marcadores microssatélites genômicos ou derivados de regiões gênicas já foram identificados e descritos para cana-de-açúcar através do enriquecimento de bibliotecas (Cordeiro et al., 2001; DaSilva, 2001; Parrida et al., 2006; Parrida et al., 2009). Muitos destes locos, utilizados em análises genético-genômica da espécie foram obtidos de projetos como o *UniGene derived Sugarcane Microsatellites* (UGSM) e o *Sugarcane Enriched Genomic Microsatellites* (SEGM). Os marcadores microssatélites são considerados excelentes ferramentas de análises genéticas de populações, pois são caracterizados pela capacidade de detectar elevados níveis de polimorfismo intra e inter populacional (Schlötterer, 2004). Juntamente com os marcadores SNPs, os microssatélites se destacam na atualidade pela ampla utilização em diversos tipos de estudos genéticos de diferentes espécies (Mammadov et al., 2012).

Neste contexto, o presente trabalho teve por objetivo caracterizar o *draft assembly* do transcritoma de cana-de-açúcar através da identificação de marcadores SNPs, microssatélites e pela análise da contribuição dos cinco diferentes órgãos amostrados na montagem final do transcritoma de cana-de-açúcar.

4.2 MATERIAL E MÉTODOS

4.2.1 O *draft assembly* do transcritoma de cana-de-açúcar

O *draft assembly* do transcritoma de cana-de-açúcar foi obtido pela análise de 30 clones elites de uma população de melhoramento formada por 48 genótipos selecionados e em fase final de avaliação pelo programa de melhoramento genético de cana-de-açúcar da Ridesa/UFG. Esta população era formada por indivíduos adultos que foram coletados em aproximadamente dez meses após o transplantio Cinco tipos diferentes de órgãos vegetais foram coletados de cada um dos trinta genótipos. Os órgãos amostrados foram: colmo, gemas laterais, plântulas, folhas e gemas apicais.

Foi amostrada a mesma quantidade de material vegetal para cada órgão coletado a partir dos 30 clones elites. Imediatamente após a coleta, o material vegetal foi armazenado em freezer a -80ºC. Para cada órgão, todo o material coletado foi macerado juntamente, utilizando nitrogênio liquido, formando cinco tipos de bibliotecas distintas referente a cada órgão vegetal. O RNA total de cada órgão foi extraído em *bulk* (o bulk foi formado antes da extração, na etapa de maceração) constituído por todos os 30 genótipos utilizando o kit *RNeasy® Plant Mini Kit* (Qiagen). O sequenciamento de bibliotecas *paired ends* foi feito utilizando a plataforma de NGS da Illumina HiSeq2000, utilizando ½ *lane* para cada biblioteca, com exceção da biblioteca de gema apical que foi sequenciada em um único *lane*. As moléculas de cDNA foram normalizadas a partir da técnica DSN (*Duplex-Specific Thermostable Nuclease*) com o objetivo de amostrar transcritos pouco abundantes.

Os dados de RNA-seq passaram por análises de controle de qualidade e normalização antes do assembly *de novo do draft* do transcritoma de cana-de-açúcar, que foi realizado pelo pacote computacional Trinity (Grabherr et al., 2011) (ver Capítulo 1 – "Obtenção de um *de novo draft assembly* do transcritoma de cana-de-açúcar utilizando dados de sequenciamento de nova geração").

4.2.2 Análise funcional dos *scaffolds*

Para cada *scaffold* montado pelo Trinity foi realizada uma análise para a identificação de quadros abertos de leitura (*ORFs - Open Reading Frames*). Esta análise foi conduzida através da ferramenta computacional TransDecoder (http://transdecoder.sourceforge.net/), considerando 300 pb como o tamanho mínimo de *ORFs*. As *ORFs* identificadas como completas, isto é, com códons de início e terminação, foram anotadas conforme os termos do Gene Ontology (The Gene Ontology Consortium, 2000), utilizando a ferramenta BLAST2GO (Conesa & Gotz 2008; Gotz et al., 2011). A análise funcional do transcritoma de cana-de-açúcar foi realizada utilizando-se como referência o banco de dados "*Grass_DB*". A busca por vias metabólicas representadas pelos transcritos com ORFs completas foi conduzida no banco de dados KEGG (*Kyoto Encyclopedia of Genes and Genomes*) (Kanehisa & Goto, 2000).

4.2.3 Contribuição dos diferentes órgãos para a constituição do transcritoma

Os cinco órgãos de cana-de-açúcar utilizados na montagem do *draf*t do transcritoma foram comparados par-a-par quanto à origem dos transcritos de cada um dos genes identificados. Os níveis normalizados de expressão gênica foram quantificados através da abundância do número de transcritos de cada órgão mapeados contra o *draft* do transcritoma, utilizando as estimativas de FPKM, estimadas com base no algoritmo RSEM (Li & Dewey, 2011). Para identificar a contribuição genotípica de cada tecido vegetal para montagem do transcritoma, os reads de cada biblioteca vegetal foram mapeados contra o *draft assembly* do transcritoma, usando o alinhador de reads curtos, BWA (Li & Durbin, 2009). Os arquivos .BAM, gerados pelo BWA foram ordenados e utilizados para estimar o nível de abundância de cada transcrito em cada um dos cinco tecidos vegetais, usando o software RSEM (Li & Dewey, 2011).

4.2.4 Identificação de marcadores SNPs

Os marcadores SNPs (*Single Nucleotide Polymorphisms*) foram identificados através de três plataformas de *SNP calling* diferentes. Foram identificados SNPs somente nas maiores isoformas de cada transcrito. A busca por variantes SNPs e polimorfismos do tipo *indels* foi realizada a partir do arquivo .BAM, após o alinhamento de cada uma das bibliotecas de sequenciamento dos cinco órgãos vegetais no transcritoma de cana-de-açúcar, representado pelas maiores isoformas de cada gene. O alinhamento foi feito utilizando o software BWA (Li & Durbin, 2009). Um limiar de 99.9% de certeza de *base calling* correto escolhido como filtro para definição correta das substituições nucleotídicas. A ferramenta VCFTools (Danecek et al., 2011) foi utilizada para análise descritiva do arquivo VCF contendo as informações de identificação dos SNPs.

A primeira ferramenta utilizada foi o software GATK (McKenna et al., 2010) pelo uso da função "*UnifiedGenotyper*", própria para busca de SNPs em espécies poliploides. Foram realizadas, previamente, uma eliminação dos locos considerados duplicados através da ferramenta de bioinformática Picard (http://picard.sourceforge.net/) e uma recalibração dos erros usando uma identificação prévia dos locos SNPs, através da função "*BaseRecalibrator*". A segunda, foi a opção "mpileup" do SAMTools (Li et al., 2009) que utiliza a ferramenta BCFTools (https://github.com/samtools/bcftools) para gerar os arquivos de *variant calling* (VCF - *Variant Calling Format*). Já a terceira ferramenta utilizada foi o software FreeBayes (Garrison & Marth, 2012), utilizado para identificação e genotipagem de polimorfismos haplotípicos.

4.2.5 Identificação de marcadores microssatélites

A busca por regiões microssatélites foi realizada em dois softwares distintos. O primeiro software chamado GMATo (Wang et al., 2013), destinado a busca de microssatélites em grandes genomas e busca somente por regiões simples repetitivas

perfeitas. Entretanto, o segundo software, chamado MISA (http://pgrc.ipk-gatersleben.de/misa/) busca por regiões microssatélites perfeitas e compostas. Em ambas as análises foram amostradas regiões simples repetitivas de seis tipos diferentes: mono, di, tri, tetra, penta e hexa nucleotídeos. Na busca por microssatélites di-nucleotídicos e tri-nucleotídicos assumiu-se como seis o número mínimo de motivos de repetição, enquanto para os microssatélites tetra-nucleotídeos, penta-nucleotídeos e hexa-nucleotídeos assumiu-se como quatro o número mínimo de motivos de repetição. Definiu-se como microssatélite do tipo mono-nucleotídeo as sequências que se repetiam em tandem por no mínimo 12 vezes. Foram amostradas regiões de microssatélites com uma distância mínima de 100 pares de bases uma da outra. Microssatélites imperfeitos não foram amostrados.

O pipeline de análise contendo os softwares utilizados em cada etapa das análises de bioinformática está disponibilizado no Apêndice D.

4.3 RESULTADOS E DISCUSSÃO

4.3.1 Anotação gênica

Foram realizadas cerca de 80 mil anotações e 95% destas foram feitas através do banco de dados UniProt utilizando principalmente informações funcionais disponibilizadas para *O. sativa*, seguida de *Z. mays*. Os resultados da análise blastx utilizadas durante a anotação funcional mostraram elevada similaridade entre as sequências gênicas de cana-de-açúcar e aquelas do banco de dados do “Grass_DB”.

Em média, 5.761 anotações gênicas foram identificadas para cada um dos três termos (Processo Biológico – BP, Função Molecular – MF e Componente Celular – CC) de classificação do GO. O termo Processo Biológico apresentou o maior número de anotações, com 27.971 transcritos classificados como fazendo parte do algum processo biológico da planta. Em número de classes dentro de cada termo de classificação, 15 diferentes tipos de Processos Biológicos foram identificados, destacando-se processos

metabólicos e celulares. O maior número de transcritos foi identificado como componente de estruturas celulares e organelares dentro do termo de classificação Componente Celular. Mais de dez mil transcritos foram relacionados a organelas, enquanto 1.423 transcritos anotados como componentes de membrana (Figura 6).

A anotação realizada pela utilização do banco de dados KEGG identificou transcritos relacionados a 902 enzimas ativas e participantes de 129 rotas metabólicas em pelo menos um órgão de cana-de-açúcar. Um total de 376 transcritos foram identificados como codificantes para proteínas atuantes na via metabólica de produção de Purina, enquanto outros 234 transcritos foram identificados como relacionados à produção de proteínas integrantes do metabolismo da sacarose e amido (Apêndice F). Esta conversão é realizada através da fotossíntese e o carbono fixado durante este evento é transformado em açúcar ou outras moléculas derivadas de açúcar (Wang et al., 2013). Eficiências fotossintéticas e altas taxas de acumulação de carbono elevam a produtividade de açúcar e o valor econômico e agronômico da cultura da cana-de-açúcar. A relação entre a atividade fotossintética foliar e o acúmulo de sacarose no colmo ainda não é bem compreendida. As observações de que a atividade fotossintética diminui durante a maturação do colmo em cultivares comerciais, aliado ao fato de que os genótipos de *Saccharum officinarum* possuem uma taxa fotossintética e acúmulo de sacarose duas a três vezes maiores que genótipos de *S. spontaneum*, evidenciam que a compreensão detalhada entre fotossíntese e acúmulo de sacarose no colmo pode desempenhar um papel fundamental no aumento do rendimento da sacarose em cultivares elites (McCormick et al., 2009). Jackson (2005) relata que em condições ideais de crescimento, 25% do peso fresco da cana-de-açúcar é devido ao acúmulo de sacarose.

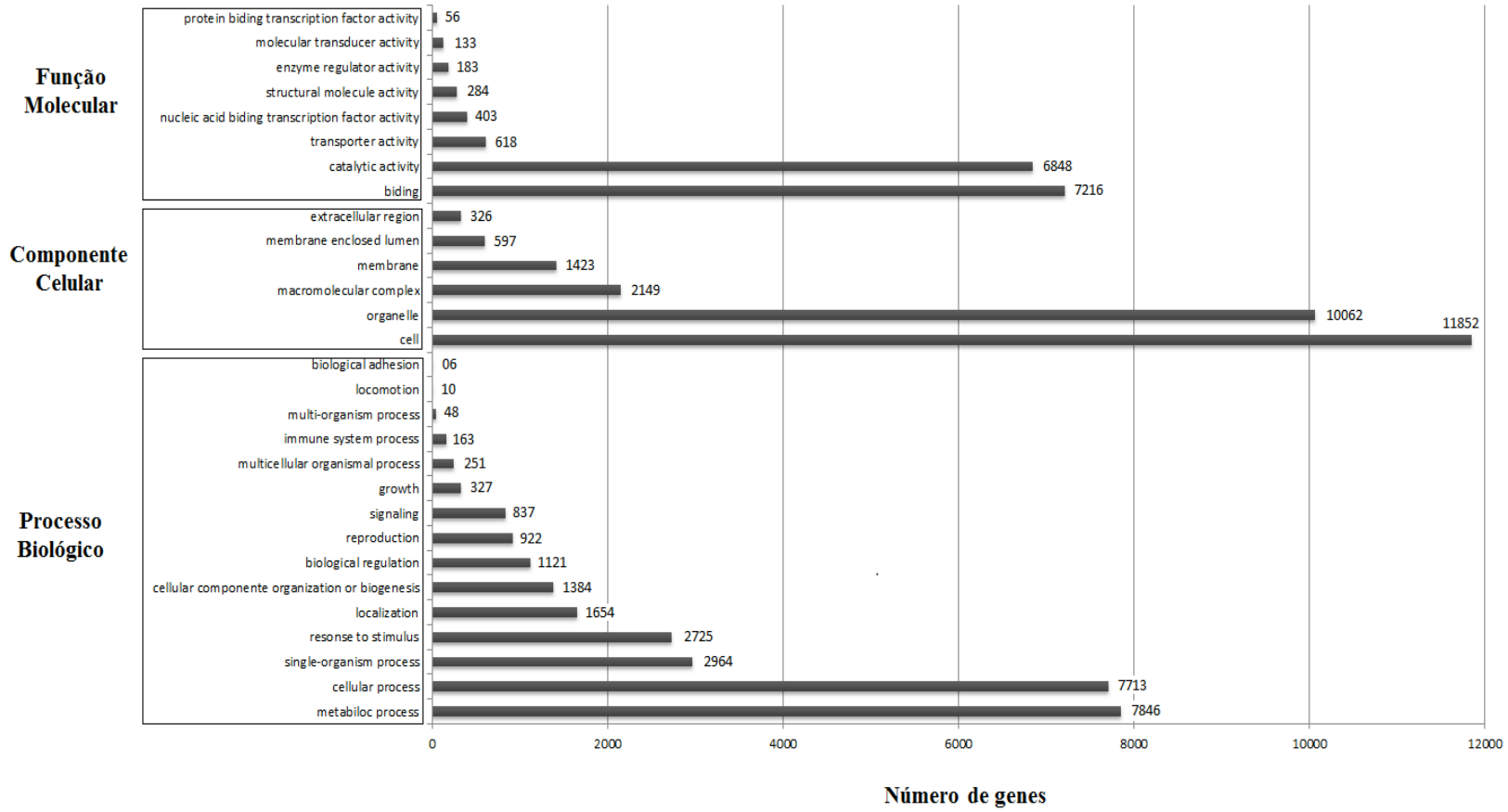


Figura 6. Anotação dos transcritos identificados no *draft assembly* do transcritoma de cana-de-açúcar que apresentam ORFs completas. A anotação foi realizada conforme os três termos do Gene Ontology (Componente Celular, Função Molecular e Processos Biológicos).

Em cana-de-açúcar, a sacarose começa a se acumular nos entrenós quando eles começam a alongar-se e continua até depois deste alongamento (Lingle & Smith, 1991). Durante o amadurecimento, as concentrações de sacarose ao longo de todo o colmo aumentam significativamente, diminuindo as concentrações de glicose e frutose (Fernandes & Brenda, 1985). Este padrão sugere que o metabolismo de sacarose no colmo de cana-de-açúcar se altera durante o desenvolvimento de planta.

O aumento de sacarose é uma característica agronômica altamente explorada pelos programas de melhoramento de cana-de-açúcar (Grof & Campbell, 2001; Moore, 2005). No entanto, recentemente com o desenvolvimento de tecnologias genômicas é possível identificar genes de interesse e manipulá-los, permitindo a produção de genótipos com características agronômicas superiores (Rafalski, 2002; Pérez-de-Castro et al., 2012; Chandra et al., 2012). Agora, o objetivo de aumentar a produtividade de sacarose em cana-de-açúcar pode ser alcançado pela regulação de enzimas específicas envolvidas no metabolismo da sacarose. Chandra et al. (2012) destacam que o metabolismo de sacarose é governado inicialmente por três enzimas: uma invertase (E.C.3.2.1.26), uma enzima sintetizadora de sacarose (*Sucrose Synthase* (SS), E.C.2.4.1.13) e uma enzima sintetizadora de fosfatos de sacarose (*Sucrose Phosphate Synthase* (SPS), E.C.2.4.1.14). Todas estas três enzimas destacadas fazem parte da via metabólica identificada pelo KEGG como metabolismo de amido e sacarose (Apêndice F), mostrando a relação de vários genes envolvidos no metabolismo de açúcar, destacando as espécies do gênero *Saccharum* quanto à eficiência em acúmulo de açúcar.

4.3.1 Contribuição dos diferentes órgãos para a constituição do transcritoma de cana-de-açúcar

Os cinco órgãos de cana-de-açúcar utilizados na montagem do *draft* do transcritoma contribuíram igualmente em termos de número de *reads* para a montagem. A média da taxa de alinhamento das diferentes bibliotecas foi de 78,07%, variando de 75,85% para a biblioteca originada de folhas a 80,50% naquela de gemas apicais (Tabela 6).

Em cana-de-açúcar, a sacarose começa a se acumular nos entrenós quando eles começam a alongar-se e continua até depois deste alongamento (Lingle & Smith, 1991). Durante o amadurecimento, as concentrações de sacarose ao longo de todo o colmo aumentam significativamente, diminuindo as concentrações de glicose e frutose (Fernandes & Brenda, 1985). Este padrão sugere que o metabolismo de sacarose no colmo de cana-de-açúcar se altera durante o desenvolvimento de planta.

O aumento de sacarose é uma característica agronômica altamente explorada pelos programas de melhoramento de cana-de-açúcar (Grof & Campbell, 2001; Moore, 2005). No entanto, recentemente com o desenvolvimento de tecnologias genômicas é possível identificar genes de interesse e manipulá-los, permitindo a produção de genótipos com características agronômicas superiores (Rafalski, 2002; Pérez-de-Castro et al., 2012; Chandra et al., 2012). Agora, o objetivo de aumentar a produtividade de sacarose em cana-de-açúcar pode ser alcançado pela regulação de enzimas específicas envolvidas no metabolismo da sacarose. Chandra et al. (2012) destacam que o metabolismo de sacarose é governado inicialmente por três enzimas: uma invertase (E.C.3.2.1.26), uma enzima sintetizadora de sacarose (*Sucrose Synthase* (SS), E.C.2.4.1.13) e uma enzima sintetizadora de fosfatos de sacarose (*Sucrose Phosphate Synthase* (SPS), E.C.2.4.1.14). Todas estas três enzimas destacadas fazem parte da via metabólica identificada pelo KEGG como metabolismo de amido e sacarose (Apêndice F), mostrando a relação de vários genes envolvidos no metabolismo de açúcar, destacando as espécies do gênero *Saccharum* quanto à eficiência em acúmulo de açúcar.

4.3.1 Contribuição dos diferentes órgãos para a constituição do transcritoma de cana-de-açúcar

Os cinco órgãos de cana-de-açúcar utilizados na montagem do *draft* do transcritoma contribuíram igualmente em termos de número de *reads* para a montagem. A média da taxa de alinhamento das diferentes bibliotecas foi de 78,07%, variando de 75,85% para a biblioteca originada de folhas a 80,50% naquela de gemas apicais (Tabela 6).

Em cana-de-açúcar, a sacarose começa a se acumular nos entrenós quando eles começam a alongar-se e continua até depois deste alongamento (Lingle & Smith, 1991). Durante o amadurecimento, as concentrações de sacarose ao longo de todo o colmo aumentam significativamente, diminuindo as concentrações de glicose e frutose (Fernandes & Brenda, 1985). Este padrão sugere que o metabolismo de sacarose no colmo de cana-de-açúcar se altera durante o desenvolvimento de planta.

O aumento de sacarose é uma característica agronômica altamente explorada pelos programas de melhoramento de cana-de-açúcar (Grof & Campbell, 2001; Moore, 2005). No entanto, recentemente com o desenvolvimento de tecnologias genômicas é possível identificar genes de interesse e manipulá-los, permitindo a produção de genótipos com características agronômicas superiores (Rafalski, 2002; Pérez-de-Castro et al., 2012; Chandra et al., 2012). Agora, o objetivo de aumentar a produtividade de sacarose em cana-de-açúcar pode ser alcançado pela regulação de enzimas específicas envolvidas no metabolismo da sacarose. Chandra et al. (2012) destacam que o metabolismo de sacarose é governado inicialmente por três enzimas: uma invertase (E.C.3.2.1.26), uma enzima sintetizadora de sacarose (*Sucrose Synthase* (SS), E.C.2.4.1.13) e uma enzima sintetizadora de fosfatos de sacarose (*Sucrose Phosphate Synthase* (SPS), E.C.2.4.1.14). Todas estas três enzimas destacadas fazem parte da via metabólica identificada pelo KEGG como metabolismo de amido e sacarose (Apêndice F), mostrando a relação de vários genes envolvidos no metabolismo de açúcar, destacando as espécies do gênero *Saccharum* quanto à eficiência em acúmulo de açúcar.

4.3.1 Contribuição dos diferentes órgãos para a constituição do transcritoma de cana-de-açúcar

Os cinco órgãos de cana-de-açúcar utilizados na montagem do *draft* do transcritoma contribuíram igualmente em termos de número de *reads* para a montagem. A média da taxa de alinhamento das diferentes bibliotecas foi de 78,07%, variando de 75,85% para a biblioteca originada de folhas a 80,50% naquela de gemas apicais (Tabela 6).

Tabela 6. Contribuição dos reads de diferentes órgãos vegetais de cana-de-açúcar para a montagem do transcritoma. FPKM é o número de fragmentos por kilobase por milhões de fragmentos mapeados.

Órgão vegetal de cana-de-açúcar	Número de reads	Taxa de alinhamento (%)	Número de transcritos mapeados	FPKM médio
Gema Apical	112888086	80,50	123829	5,364
Gema Lateral	61949766	79,05	128835	5,143
Plântulas	62294415	76,39	125280	5,477
Folhas	56894266	75,85	121990	5,388
Colmos	58384157	78,60	127477	5,323
Média	70482138	78,07	125482,2	5,339

Os valores de FPKM variaram de 5,143 para o órgão gema lateral a 5,477 em plântulas. Adicionado a isto, as análises revelaram que a média de número de transcritos mapeados no transcritoma originados de cada órgão vegetal separadamente foi de pouco mais de 125 mil transcritos. A contribuição de cada órgão vegetal, estimada através da quantificação do número de reads mapeados (FPKM) pode ser considerada confiável, uma vez que este parâmetro consegue captar realmente a variação que existe nas taxas de mapeamento de reads e na quantificação de suas abundâncias (Mortazavi et al., 2008). O órgão vegetal que contribuiu com mais transcritos foi gema lateral (128835 transcritos mapeados), mesmo não sendo o órgão com maior número de reads. A folha foi o órgão vegetal com menor número de transcritos mapeados (Tabela 6). Estes resultados mostram que o tecido foliar apresentou a baixa diversidade de isoformas em relação aos outros quatro órgãos amostrados, sugerindo que a utilização apenas deste órgão vegetal não abrange a totalidade de transcritos para cana-de-açúcar e que a montagem de um transcritoma para cana-de-açúcar utilizando somente este órgão pode superestimar o número de transcritos da espécie.

4.3.2 A identificação de marcadores moleculares microssatélites

A busca por sequências simples repetitivas utilizando ferramentas distintas revelou, praticamente, o mesmo número de regiões de microssatélites no transcritoma de referência de cana-de-açúcar. Utilizando os mesmos parâmetros em ambas as análises, o software MISA conseguiu identificar 12.931 regiões de microssatélites, enquanto o

software GMATo identificou 12.925 sequências simples repetitivas. A diferença de seis regiões aconteceu pela não identificação de três regiões dinucleotídicas e três regiões de hexanucleotídeos (Figura 7). Houve um predomínio, em ambos os softwares, de regiões microssatélites com polimorfismos di e tri nucleotídicos. Juntas, estas duas classes de marcadores corresponderam a aproximadamente 60% do total de microssatélites identificados.

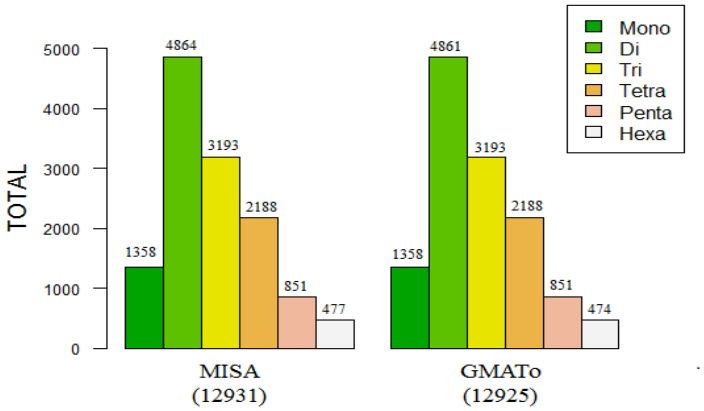


Figura 7. Número total de regiões microssatélites identificados em ambos os softwares utilizados nas análises. Os dois softwares conseguiram identificar praticamente a mesma quantidade de sequências simples repetidas, com um predomínio das repetições di e tri nucleotídicas.

O maior número de regiões de microssatélites identificados por transcrito foi quatro. Nove transcritos apresentaram quatro regiões de microssatélites em cada um. O software MISA conseguiu identificar 74 microssatélites compostos.

O motivo de repetição mais comum nos microssatélites di e tri nucleotídeos são AG/CT e CCG/CGG, respectivamente. 2.597 microssatélites dinucleotídeos apresentam o motivo de repetição AG/CT, enquanto 1.120 microssatélites trinucleotídeos apresentam o motivo de repetição CCG/CGG. O número de regiões microssatélites para as classes de motivos de repetição mais frequente pode ser visualizado na Tabela 7.

Tabela 7. Descrição do número de microssatélites identificados para o motivo de repetição mais frequente em cada um dos seis tipos de microssatélites analisados. Mono = Mono-nucleotídeo; DI = Di-

nucleotídeo; TRI = Tri-nucleotídeo; TETRA = Tetra-nucleotídeo; PENTA = Penta-nucleotídeo; HEXA = Hexa-nucleotídeo.

* Tipo do SSR	Motivo de repetição mais frequente	Número de microssatélites
MONO	A/T	1325
DI	AG/CT	2597
TRI	CCG/CGG	1120
TETRA	AGGC/CCTG	180
PENTA	AAAAG/CTTTT	71
HEXA	AGCAGG/CCTGCT	20

* Classificação das regiões microssatélites quanto tipo de motivos de repetição.

Para cana-de-açúcar, os microssatélites com polimorfismo tri-nucleotídeo, identificados no transcritoma da espécie, apresentaram uma abundância do conteúdo GC (Figura 8), corroborando com os resultados encontrados por Blair et al. (2011). Estes autores, ao caracterizarem regiões microssatélites no genoma de feijão comum (*Phaseolus vulgaris*) com base em sequências de ESTs de tecido foliar e radicular, perceberam que os locos tri-nucleotídeos também apresentavam elevados índices de nucleotídeos GC. Quanto aos microssatélites di-nucleotídeos, houve um predomino de nucleotídeos AG/CT. Em feijão comum, um número elevado de microssatélites dinucleotídeos foi encontrado em tecido radicular, mostrando que alguns polimorfismos podem ser específicos para tecidos vegetais distintos (Blair et al., 2011). Este estudo forneceu uma desenvolveu uma variedade de marcadores moleculares com base em dois genótipos de feijão andino e mesoamericano. A identificação e o desenvolvimento de marcadores microssatélites para ervilha (*Pisum sativum*), derivados de sequências de ESTs, revelu que a grande maioria dos locos possuíam um motivo de três nucleotídeos, no entanto, o motivo GAA foi o mais abundante entre os identificados (Gong et al., 2010). Estes autores conseguiram identificar 503 locos microssatélites em mais de 18 mil sequências de ESTs existentes no banco de dados do NCBI. Moe et al. (2012), desenvolveram marcadores moleculares para uma espécie popular de orquídea (*Cymbidium* spp.) a partir de sequências de cDNA e mostraram que os polimorfismos di-nucleotídeos são mais frequentes em regiões genômicas, enquanto os polimorfismos do tipo tri-nucleotídeo são mais frequente em regiões de cDNA. O motivo mais frequente entre os di-nucleotídeos, tanto em regiões genômicas quanto em regiões de cDNA, foi CT/AG/TC/GA, enquanto entre os tri-nucleotídeos houve dois tipos de motivos muito frequente (CTT/AAG/TCT/AGA/TTC/GAA e GTT/AAC/TGT/ACA/TTG/CAA).

Diferenças entre tipos de marcadores em diferentes regiões do genoma (regiões gênicas e não gênicas) já foi observado para outras espécies agronomicamente importantes, como o feijão comum (De Campos et al., 2007; Hanai et al., 2007). Diferentemente do que foi encontrado para o transcritoma de cana-de-açúcar, onde os motivos di-nucleotídeos foram os mais frequentes (Figura 8), em outras espécies como o trigo tetraploide (*Triticum durum*) (Gadaleta et al., 2010) e o feijão comum (Blair et al., 2011) o motivo tri-nucleotídeo apresentou uma maior frequência.

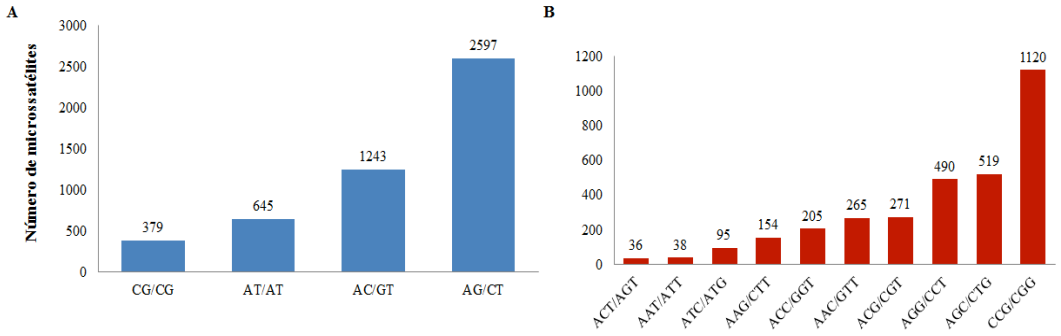


Figura 8. Distribuição dos motivos de repetição nos microssatélites analisados. (A) distribuição dos motivos di-nucleotídeos, mostrando que o motivo AG/TC foi o motivo mais frequente dentre os quatro motivos identificados. (B) Foram identificados dez motivos de repetição do tipo tri-nucleotídeo, com um predomínio do motivo CCG. Os microssatélites do tipo tri-nucleotídeos possuem uma abundância do conteúdo GC.

Blair et al. (2009) identificaram e caracterizaram 248 marcadores microssatélites em regiões gênicas/transcritas de um genótipo de feijão (*Phaseolus vulgaris* L.) andino utilizado como fonte de resistência a fatores bióticos e abióticos. Ding et al. (2011), desenharam marcadores moleculares SSR para regiões gênicas relacionadas à homeostase de fósforo identificados em *A. thaliana* e construíram um mapa genético destes genes em *Brassica napus*. Cardoso-Silva et al. (2014) encontraram 5.106 sequências simples repetidas em regiões transcritas do tecido vegetal de folhas em seis variedades comerciais de cana-de-açúcar ao avaliarem 72.269 unigenes.

A distribuição das regiões de microssatélites quanto ao número de repetição dos motivos, mostra que a grande maioria dos SSRs di e tri nucleotídeos possuem seis ou sete repetições do motivo, podendo ser consideradas regiões estáveis com potencial para serem utilizadas como marcadores moleculares (Figura 9). Ou seja, sequências simples de dois ou

três nucleotídeos repetidas em tandem seis ou sete vezes são regiões para as quais primers podem ser desenhados e utilizados para revelar polimorfismos populacionais de interesse.

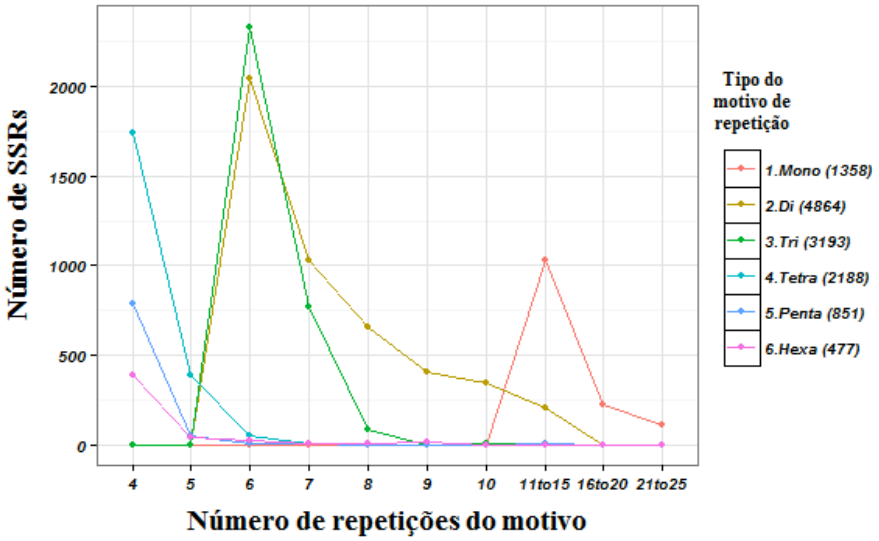


Figura 9. Distribuição das regiões de microssatélites identificadas quanto ao número dos motivos de repetição.

4.3.3 A identificação de marcadores moleculares SNPs

As três principais ferramentas de bioinformática utilizadas na identificação de SNPs se mostraram semelhantes, com destaque para a ferramenta GATK que foi capaz de identificar o maior número de SNPs com cerca de 4,16 milhões, seguida pela ferramenta SAMTools/mpileup, com 4,01 milhões e pela ferramenta FreeBayes com 3,74 milhões de SNPs identificados. No entanto, houve uma diferença expressiva quanto ao número de *indels* identificados. O padrão foi inverso à identificação de SNPs com a ferramenta FreeBayes identificando cerca de 319 mil *indels*, seguida pela ferramenta SAMTools/mpileup com cerca de 210 mil e pela ferramenta GATK que identificou o menor número, somente 4.344 *indels* (Apêndice G). O motivo desta diferença expressiva do número de indels identificados entre os três softwares pode ser explicada pela não

utilização da ferramenta “IndelRealigner” existente na plataforma GATK, que realiza um realinhamento dos reads mal alinhados pela presença de indels.

Nas três ferramentas, o órgão vegetal com maior número de SNPs identificados foi gema apical e o órgão com menor número de SNPs identificados foi folha, que também apresentou em ambas as análises, a menor taxa de diversidade nucleotídica. Considerando que o número de SNPs identificados pelas três abordagens foram semelhantes e que existe pouca diferença de desempenho de análise entre estas ferramentas de *SNP calling* (Yu & Sun, 2013), além de que a ferramenta GATK é considerada a melhor entre elas (Liu et al., 2013), somente as estimativas feitas com esta ferramenta serão discutidas no restante do trabalho.

A busca por polimorfismos do tipo SNPs, revelou, para o transcritoma de cana-de-açúcar obtido, a existência de, em média, uma substituição nucleotídica a cada 18 pb, mostrando uma elevada densidade de SNPs. Existem, em quase 77 Mb de sequências gênicas amostradas (somente as maiores isoformas de cada transcrito), um total de 4.171.246 SNPs. Estes SNPs podem ser utilizados na predição dos valores genéticos-genômicos em abordagens de Seleção Genômica Ampla (Goddard & Hayes, 2007), por exemplo. A profundidade média de sequenciamento para identificação dos SNPs foi de 75X. A estimativa da diversidade nucleotídica para os trinta clones elites amostrados foi muito elevada, com a estimativa de $\pi = 0{,}931$ (Tabela 7). Não foi detectada uma correlação significativa entre o número de SNPs encontrados nas bibliotecas de cada órgão vegetal e o número de transcritos identificados, ou seja, a diversidade de isoformas, representada pelo número de isoformas identificada para cada transcrito descrito pelo Trinity, não está relacionada com a diversidade nucleotídica.

A razão entre a taxa de substituições do tipo Transição (Ts) e a taxa de substituição do tipo Transversão (Tv) foi, em média, de 1,74. Houve cerca de duas vezes mais mutações do tipo Transição (Ts) em relação as mutações do tipo Transversão (Tv) (Figura 12). As mutações do tipo Ts são mais frequentes que as mutações do tipo Tv porque as mutações Ts acontecem entre nucleotídeos da mesma família nucleotídica, isto é, entre Purinas (A/G) ou entre Pirimidinas (C/T), ao contrário das substituições do tipo Tv que acontecem entre nucleotídeos de famílias diferentes.

Morton et al. (2006) estudaram o padrão de mutações pontuais entre linhagens de milho (*Z. mays*), através de um conjunto de dados de mais de 10 mil SNPs e perceberam uma relação direta entre o padrão de mutação e os nucleotídeos que flanqueiam o sítio mutacional. Estes autores ainda discutiram que, geralmente o conteúdo A+T flanqueia sítos de mutação do tipo Transição (Ts). A razão Ts/Tv tem sido estimada para estudos genômicos em algumas espécies de plantas, tais como *Zea mays* (TsTv = 3,9), *Medicago sativa* (Ts/Tv = 3,6), *Triticum monococcum* (Ts/Tv =1,9) e *Hordeum vulgare* (Ts/Tv = 1,6) (Vitte & Bennetzen, 2006). Informações sobre a razão Ts/Tv são escassas em muitas espécies, assim como em cana-de-açúcar. A estimativa da razão Ts/Tv é comumente utilizada em reconstrução filogenética, estimação do tempo de divergência e compreensão dos mecanismos de evolução molecular (Yang & Yoder, 1999).

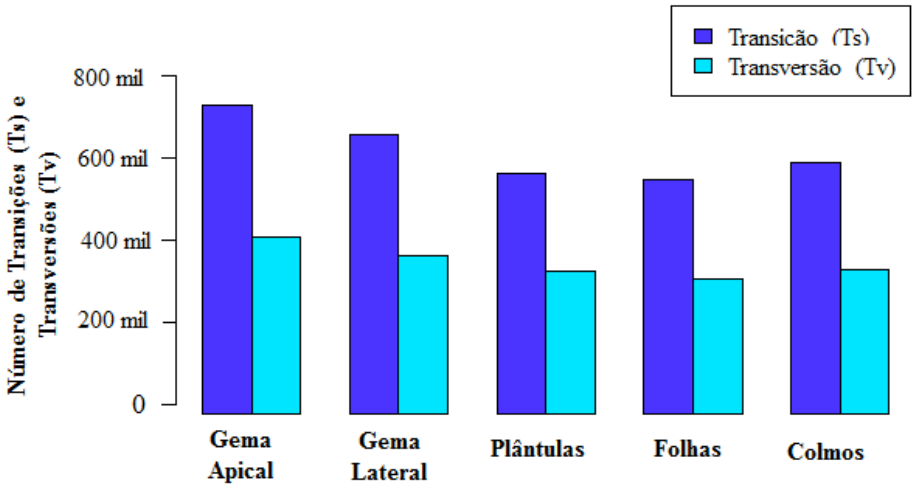


Figura 10. Relação entre o número de substituições nucleotídicas do tipo Transição (Ts) e do tipo Transversão (Ts) para os 4.171.246 SNPs identificados. A razão entre a taxa de Ts/Tv foi de 1,74, mostrando que o número de substituições entre nucleotídeos da mesma família é maior.

Tabela 8. Parâmetros que caracterizam a identificação de SNPs ao longo do transcritoma de cana-de-açúcar. A identificação de SNPs foi realizada separadamente para cada biblioteca oriunda de um tipo específico de órgão vegetal coletado em 30 clones elite.

						Score dos Haplótipos		Scores de Qualidade	
Órgão Vegetal	**Número de SNPs**	**Indels**	**Cobertura**	**Diversidade nucleotídica (π)**	**Razão Ts/Tv**	**Média**	**Desvio Padrão**	**Média**	**Desvio Padrão**
Gema Apical	1.034.475	1.103	93,51	0,947	1,746	15,421	33,192	577,177	841,99
Gema Lateral	877.320	919	75,11	0,940	1,771	14,458	31,574	479,888	746,40
Plântulas	745.286	790	71,69	0,920	1,696	13,925	30,608	474,229	755,22
Folhas	732.788	807	68,37	0,908	1,735	12,569	29,414	457,130	746,44
Colmo	781.377	725	68,64	0,938	1,750	12,826	29,096	444,143	708,45
Total	4.171.246	4344	--	--	--	--	--	--	--
Média	834.249,2	868,8	75,46	0,931	1,740	13,840	30,777	486,513	759,699

Os valores dos scores dos haplótipos, estimados pela ferramenta GATK, estimam a consistência dos sítios polimórficos (que apresentam polimorfismos SNPs) em apresentar somente dois haplótipos, pois se espera que para cada loco somente dois haplótipos sejam possíveis de estarem segregando. Altos valores desta estimativa são indicativos de regiões genômicas com mau alinhamento e possivelmente com identificação errada dos SNPs. Como sugerido pelo time do *Broad Institute*, desenvolvedor do software GATK, um valor dos scores dos haplótipos considerado como limiar para filtragem dos SNPs em regiões de exoma é treze. No entanto, como os valores dos scores do haplótipos são dependentes da cobertura de identificação dos SNPs, um valor limiar deve ser adaptado para cada estudo (http://gatkforums.broadinstitute.org/discussion/2369/calculation-of-haplotypescore). O valor médio dos scores haplotípicos obtidos na identificação de SNPs no transcritoma de cana-de-açúcar foi de 13,84, no entanto, para alguns órgãos vegetais como folhas e colmos os valores foram razoavelmente inferiores ao limiar inferior sugerido pelos desenvolvedores do GATK. A medida scores de qualidade para a identificação dos SNPs pode ser entendida como o limiar a ser assumido como erros de sequenciamento ao invés de um SNP. Este limiar foi de 99,9%, aceitando somente um erro de sequenciamento a cada mil pares de bases. Assim, quanto maior os valores de scores de qualidade mais confiável será a identificação dos SNPs.

A identificação e validação de marcadores SNPs são etapas iniciais para utilização destes marcadores em estudos genômicos voltados para o melhoramento vegetal de espécies de interesse agronômico. Uma vez identificados, os SNPs podem ser convertidos em marcadores genéticos e utilizados nas plataformas de genotipagem de alto desempenho. Devido à sua elevada abundância no genoma, mapas genéticos densos podem ser construídos e utilizados como suporte aos programas de melhoramento genético que utilizam a estratégia de Seleção Assistida por Marcadores (MAS – *Marker Assisted Selection*) ou Seleção Genômica Ampla (Goddard & Hayes, 2007). Inúmeros projetos de identificação extensiva de locos SNPs ao longo do genoma e/ou transcritoma de diversas espécies modelos e não modelos já foram conduzidos. Em espécies de plantas em que não há o genoma de referência sequenciado, a identificação em larga escala de locos SNPs em regiões gênicas pode ser realizado através da caracterização de bibliotecas de ESTs (*Expressed Sequence Tags*) (Bundock et al., 2006) ou com base no desenvolvimento de primers e ressequenciamento (Choi et al., 2007). A cana-de-açúcar não é uma exceção e

sequências de ESTs têm sido utilizadas na busca por locos SNPs (Grivet et al., 2003; Cordeiro et al., 2006).

Em milho, estima-se que ocorra em média, um polimorfismo SNP a cada 28 a 124 pares de bases, dependendo da região genômica e do tipo de população avaliada (Ching et al., 2002). Barbazuk et al. (2007), encontraram 36.000 SNPs em duas populações híbridas de milho após sequenciarem, via pirossequenciamento, o transcritoma de meristemas apicais. Cerca de 85% destes SNPs foram validados utilizando o sequenciamento de Sanger. Choi et al. (2007) construíram o primeiro mapa de transcritoma de soja utilizando três linhas de endocruzamento. Nos 2,44 Mb de sequências alinhadas foram encontrados 5.551 SNPs, além da existência de pelo menos um SNP em cada um dos 1.141 genes identificados. Em cana-de-açúcar, Bundock et al. (2009) sequenciaram, usando a plataforma 454 (pirossequenciamento), regiões genômicas de uma população de mapeamento e duas variedades comerciais australianas com o objetivo de identificar SNPs ligados a uma característica quantitativa Foram encontrados 1.632 SNPs para o genótipo Q165, enquanto 1.013 SNPs foram encontrados para o parental feminino IJ76-514 (*S. officinarum*). Foram testados 225 SNPs candidatos e 93% foram validados como polimórficos. Cardoso-Silva et al. (2014) analisaram o transcritoma foliar de cana-de-açúcar através da metodologia de RNA-seq e identificaram pouco mais de 708 mil SNPs distribuídos em cerca de 72 mil unigenes.

4.4 CONCLUSÕES

A anotação realizada no banco de dados KEGG identificou 234 transcritos participantes do metabolismo da sacarose e amido, uma importante rota metabólica para compreensão da relação entre taxa fotossintética e acúmulo de sacarose no colmo. As três principais enzimas de fundamental importância nesta rota metabólica foram amostradas.

A identificação de genes candidatos que controlam características agronômicas é o primeiro passo que viabiliza a utilização de técnicas de engenharia genética no melhoramento de plantas.

O transcritoma de cana-de-açúcar foi montado abrangendo igualmente os cinco órgãos vegetais amostrados (gema apical, gema lateral, folhas, colmos e plântulas).

Foram identificados mais de quatro milhões de locos SNPs espalhados ao longo do transcritoma de cana-de-açúcar e bem distribuídos nos cinco órgãos vegetais amostrados. Em média, encontrou-se 1 SNP a cada 18 pares de bases, mostrando elevada densidade destes locos ao longo do transcritoma de cana-de-açúcar. Estes SNPs podem ser utilizados no desenvolvimento de tecnologias de genotipagem de alto desempenho, fornecendo suporte a construção de mapas genéticos densos e a identificação precisa de QTLs.

A diversidade nucleotídica encontrada foi elevada para os cinco órgãos vegetais dos trinta clones elites amostrados. Os valores dos scores haplotípicos e scores de qualidade mostram uma robustez das análises para identificação de SNPs, eliminando regiões de mau alinhamento dos reads e aceitando somente um erro de sequenciamento a cada mil pares de bases.

Mais de 12 mil regiões microssatélites foram identificadas com predomínio dos polimorfismos com motivos de di e tri nucleotídeos que apresentaram entre seis e sete repetições, sendo considerados microssatélites estáveis, onde marcadores moleculares podem ser desenvolvidos. As regiões de microssatélites estáveis identificadas permitem a exploração ainda maior deste transcritoma, uma vez que estas regiões podem ser transformadas em marcadores moleculares polimórficos.

Devido à enorme complexidade, o genoma da cana-de-açúcar ainda não foi montado e anotado por completo, reafirmando a importância de um estudo de montagem e caracterização do transcritoma da espécie.

5 CONSIDERAÇÕES FINAIS

A montagem e caracterização de um *draft assembly* para o transcritoma de cana-de-açúcar é de fundamental importância para a utilização destas informações em estudos genéticos e genômicos com a espécie. A cana-de-açúcar é uma das espécies agrícolas de maior complexidade genômica e por isso a montagem da sequência completa do seu genoma ainda não foi possível, o que ressalta a importância de se explorar as informações no contexto do transcritoma da espécie. Neste trabalho, a parte funcional do genoma de cana-de-açúcar foi trabalhada com ênfase, permitindo a montagem e a caracterização de um *draft* do transcritoma, que representa um passo fundamental em direção à obtenção de um transcritoma de referência para uma das espécies mais importantes no cenário agrícola nacional e mundial.

O presente trabalho propõe um *draft assembly* para o transcritoma de cana-de-açúcar com um tamanho de aproximadamente 178 Mb. O transcritoma aqui proposto abrange cerca de 93% do total de sequências do principal banco de dados públicos de sequências gênicas de cana-de-açúcar (SoGI – *Saccharum officinarum Gene Index*). Além disso, foram identificados mais de 90 mil transcritos que não estão representados nos bancos de dados atualmente disponíveis para cana-de-açúcar O pipeline de análise de RNA-seq proposto pela plataforma Trinity mostrou eficiente na detecção de transcritos novos. Foram identificados 1.250 transcritos pela primeira vez (nTAR – *novel Transcripts Active Regions*), não havendo hits no banco de dados nr do NCBI para estes transcritos.

Há evidências de que o transcritoma de cana-de-açúcar possui uma quantidade maior de genes quando comparado com outras espécies da família das Poaceae (*O. sativa*, *Z. mays* e *S. bicolor*), sugerindo um efeito multiplicador de eventos de duplicação gênica ao longo da evolução das espécies do complexo *Saccharum*.

Foi identificada uma quantidade muito grande de SNPs ao longo do transcritoma de cana-de-açúcar, além de uma diversidade nucleotídica elevada. Estimou-se em média, a existência de um SNP a cada 18 pares de bases. A identificação de marcadores

moleculares do tipo SNPs espalhados ao longo do transcritoma de uma espécie, fornece subsídios importantes para construção de chips de genotipagem de alto desempenho e a utilização destas ferramentas em estratégias de melhoramento genético na era da genômica.

A metodologia de sequenciamento de mRNA se mostrou eficiente por permitir uma identificação extensiva de marcadores moleculares em regiões gênicas ao longo do transcritoma de cana-de-açúcar. Neste sentido, pode-se dizer que se trata de uma metodologia eficiente para caracterização do exoma de espécies de plantas poliploides.

A utilização futura dos resultados aqui obtidos deverá permitir a identificação de marcadores SNPs em regiões genômicas de interesse agronômico e deve ser considerada como fundamental para a utilização de ferramentas genômicas com potencialidade de auxiliar efetivamente o melhoramento de plantas na identificação e/ou produção de genótipos com características agronômicas superiores.

6 REFERÊNCIAS BIBLIOGRÁFICAS

ADAMS, K. L. & WENDEL, J. F. Polyploidy and genome evolution in plants. **Current opinion in plant biology**, New Jersey, v. 8, n. 2, p.135-141, 2005.

ALTSCHUL, MADDEN, T. L.; W.; SCHÄFFER, A. A.; ZHANG, J.; ZHANG, Z.; MILLER, W.; LIPMAN, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Research**, Oxford, v. 25, n. 17, p. 3389-3402, 1997.

ALURU, S. Bioinformatics for Next Generation Sequencing. **CSI Journal of Computing**, Mumbai, v. 1, n. 1, p. 1-15, 2012.

ANDREWS, S. FastQC: a quality control tool for high throughput sequence data. Disponível em: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc >, 2010.

ANSORGE, W. J. Next-generation DNA sequencing techniques. **New Biotechnology**, Cambridge, v. 25, n. 4, p. 195-203, 2009.

ARRUDA, P. Sugarcane transcriptome. A landmark in plant genomics in the tropics. **Genetics and Molecular Biology**, Ribeirão Preto, v 24, n.4, pp.1-2, 2001.

ASNAGHI, C.; PAULET, F.; KAYE, C.; GRIVET, L.; DEU, M.; GLASZMANN, J. C.; D'HONT, A. Application of synteny across Poaceae to determine the map location of a sugarcane rust resistance gene. **Theoretical and Applied Genetics**, Stuttgart, v. 101, n. 5-6, p. 962-969, 2000.

BAIRD, N. A.; ETTER, P. D.; ATWOOD, T. S.; CURREY, M. C.; SHIVER, A. L.; LEWIS, Z. A. … JOHNSON, E. A. Rapid SNP discovery and genetic mapping using sequenced RAD markers. **Plos One**, Washington, v. 3, n. 10 p. 1-7, 2008.

BARBAZUK, W. B.; EMRICH, S. J.; CHEN, H. D.; LI, L.; SCHNABLE, P. S. SNP discovery via 454 transcriptome sequencing. **The Plant Journal**, Michigan, v. 51, n. 5, p. 910-918, 2007.

BHAT, S. R. & GILL, S. S. The implication of the 2n egg gametes in nobilization and breedind of sugarcane. **Euphytica**, Wageningen, v. 34, p. 377-384, 1985.

BIROL, I,; JACKMAN, S. D.; NIELSEN, C. B.; QIAN, J. Q.; VARHOL, R.; STAZYK, G. … JONES, S. J. *De novo* transcriptome assembly with ABySS. **Bioinformatics**, London, v. 25, p. 2872-2877, 2009.

BIELIG, L. M.; MARIANI, A.; BERDING, N. Cytological studies of 2n male gamete formation in sugarcane, *Saccharum L*. **Euphytica**, Wageningen, v. 133, p. 117-124, 2003.

BLAIR, M. W.; TORRES, M. M.; GIRALDO, M. C.; PEDRAZA, F. Development and diversity of Andean-derived, gene-based microsatellites for common bean (*Phaseolus vulgaris* L.). **BMC Plant Biology**, London, v. 9, n. 100, p. 1-14, 2009.

BLAIR, M. W.; HURTADO, N.; CHAVARRO, C. M.; MUÑOZ-TORRES, M. C.; GIRALDO, M. C.; PEDRAZA, F.; TOMKINS, J.; WING, R. Gene-based SSR markers for common bean (*Phaseolus vulgaris* L.) derived from root and leaf tissue ESTs: An integration of the BMc series. **BMC Plant Biology**, London, vol. 11, n. 50, p. 1-10, 2011.

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

BOWERS, J. E.; ABBEY, C.; ANDERSON, S.; CHANG, C.; DRAYE, X.; HOPPE, A. H. … PATERSON, A. H. A high-density genetic recombination map of sequence-tagged sites for sorghum, as a framework for comparative structural and evolutionary genomics of tropical grains and grasses. **Genetics**, New York, v. 165, p. 367-386, 2003.

BROWN, C. T.; HOWE, A.; ZHANG Q.; PYRKOSZ, A.; BROM T. H. A reference-free algorithm for computational normalization of shotgun sequencing data. http://arxiv.org/abs/1203.4802, 2014.

BRUMFIELD, R. T.; BEERLI, P.; NICKERSON, D. A.; EDWARDS, S. V. The utility of single nucleotide polymorphisms in inferences of population history. **Trends in Ecology and Evolution**, Cambridge, v. 18, p. 249-256, 2003.

BUNDOCK, P. C.; CROSS, M. J.; SHAPTER, F. M.; HENRY, R. J. Robust allele-specific polymerase chain reaction markers developed for single nucleotide polymorphisms in expressed barley sequences. **Theoretical Applied Genetics**, Stuttgart, v. 112, 358-365, 2006.

BUNDOCK, P. C.; ELIOTT, F. G.; ABLETT, G.; BENSON, A. D.; CASU, R. E.; AITKEN, K. S.; HENRY, R. J. Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploid plant species using 454 sequencing. **Plant Biotechnology Journal**, Atlanta, v. 7, n. 4, p. 347-354, 2009.

BURR, G. O.; HARTT, C. E.; BRODIE, H. W.; TANIMOTO, T.; KORTSCHAK, H. P.; TAKAHASHI, D. … COLEMAN, R. E. The sugarcane, **Annual review of plant physiology**, Los Angeles, v. 1, p. 1-34, 1956.

BUTTERFIELD, M. K.; D'HONT, A.; BERDING, N. The sugarcane genome: a synthesis of current understanding and lessons for breeding and biotechnology. **Proceedings of the South African Sugar Technologists Associations**, Cape Town, v. 75, p. 1-5, 2001.

CARDOSO-SILVA, C. B.; COSTA, E. A.; MANCINI, M. C.; BALSALOBRE, T. W. A.; CANESIN, L. E. C.; PINTO, L. R. ... VICENTINI, R. *De novo* assembly and transcriptome analysis of contrasting sugarcane varieties. **PLOS ONE**, Washington, v. 9, n. 2 p. e88462, 2014.

CARSON, D. L.; BOTHA, F. C. Preliminary analysis of expressed sequence tags for sugarcane. **Crop Science**, Madison, v. 40, n. 6, p. 1769-1779, 2000.

CARVALHO, M. C. G.; SILVA, D. C. G. Sequenciamento de DNA de nova geração e suas aplicações na genômica de plantas. **Ciência Rural**, Santa Maria, v. 40, n. 3, p. 735-744, 2010.

CASU, R.; DIMMOCK, C.; THOMAS, M.; BOWERM, N.; KNIGHT, D.; GROF, C.; MCINTYRE, L.; JACKSON, P.; JORDAN, D.; WHAN, V.; DRENTH, J.; TAO, Y.; MANNERS, J. Genetic and expression profiling in sugarcane. **Proceeding of International Sociaty of Sugar Cane Technologist**, Quatre-Bornes, v. 24, p. 542-546, 2001.

CASU, R. E.; DIMMOCK, C. M.; CHAPMAN, S. C.; GROF, C. P.; MCINTYRE, C. L.; BONNETT, G. D.; MANNERS, J. M. Identification of differentially expressed transcripts from maturing stem of sugarcane by *in silico* analysis of stem expressed sequence tags and gene expression profiling. **Plant Molecular Biology**, Amsterdam, v. 54, n. 4, p. 503-517, 2004.

CASU, R. E.; MANNERS, J. M.; BONNETT, G. D.; JACKSON, P. A.; MCINTYRE, C. L.; DUNNE, R.; CHAPMAN, S. C.; RAE, A. L.; GROF, C. P. Genomics approaches for the identification of genes determining important traits in sugarcane. **Field Crops Research**, Philadelphia, v. 92, n. 2, p. 137-147, 2005.

CASU, R. E.; JARMEY, J. M.; BONNET, G. D.; MANNERS, J. M. Identification of transcripts associated with cell wall metabolism and development in the stem of sugarcane by Affymetrix GeneChip Sugarcane Genome Array expression profiling. **Functional Integrative Genomics**, Perth, v. 7, p. 153-167, 2007.

CHANDRA, A.; JAIN, R.; SOLOMON, S. Complexities of invertases controlling sucrose accumulation and retention in sugarcane. **Current Science**, Bangalore v. 102, n. 6, p. 857-866, 2012.

CHEN, H. & BOUTROS, P. C. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. **BMC Bioinformatics**, London, v. 12, n. 35, p. 1-7, 2011.

CHEN, J.; HOU, K.; QIN, P.; LIU, H.; YI, B.; YANG, W.; WU, W. RNA-Seq for gene identification and transcript profiling of three *Stevia rebaudiana* genotypes. **BMC Genomics**, London, v. 15, n. 571, p. 1-11, 2014.

CHING, A.; CALDWELL, K. S.; JUNG, M.; DOLAN, M.; SMITH, O. S.; TINGEY, S. … RAFALSKI, A. J. SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. **BMC Genetic**, London, v. 3, n. 19, p. 1-14. 2002.

CHINNUSAMY, V. & ZHU, J. K. Epigenetic regulation: chromatin modeling and small RNAs. In: PAREEK. A.; SOPORY, S. K.; BOHNERT, H. J.; GOVINDJEE. (Ed.). **Abiotic Stress Adaptation in Plants: physiological, molecular and genomic foundation**. Amsterdam, 2010, cap. 11, p. 217-236.

CHOI, I. Y.; HYTEN, D. L.; MATUKUMALLI, L. K.; SONG, Q.; CHAKY, J. M.; QUIGLEY, C. V. … CREGAN, P. B. A Soybean Transcript Map: Gene Distribution, Haplotype and Single-Nucleotide Polymorphism Analysis. **Genetics**, New York, v. 176, p. 685-696, 2007.

COLLINS, N. C.; TARDIEU, F.; TUBEROSA, R. Quantitative trait loci and crop performance under abiotic stress: where do we stand? **Plant Physiology**, Los Angeles, v. 147, p. 469-486, 2008.

COMAI, L.; TYAGI, A. P.; WINTER, K.; HOLMES-DAVIS, R.; REYNOLDS, S. H.; STEVENS, Y.; BYERS, B. Phenotypic instability and rapid gene silencing in newly formed Arabidopsis allotetraploids. **The Plant Cell**, Michigan, v. 12, p. 1551-1567, 2000.

CONESA, A.; GOTZ, S.; GARCÍA-GÓMEZ, J. M.; TEROL, J.; TALÓN, M.; ROBLES, M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. **Bioinformatics**, Oxford, v. 21, n. 18, p. 3674-3676, 2005.

CONESA, A. & GOTZ, S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. **International journal of plant genomics**, Tashkent, v. 619832, 2008.

CORDEIRO, G.M.; CASU, R.; McINTYRE, C. L.; MANNERS, J. M.; HENRY, R. J. Microsatellite markers from sugarcane *(Saccharum* spp.) ESTs cross transferable to erianthus and sorghum. **Plant Science**, Davis, v. 160, n. 6, p. 1115-1123, 2001.

CORDEIRO, G. M.; ELIOTT, F.; MCINTYRE, C. L.; CASU, R. E.; HENRY, R. J. Characterisation of single nucleotide polymorphisms in sugarcane ESTs. **Theoretical Applied Genetics**, Sttutgart, v. 113, p. 331-343, 2006.

CORDEIRO, G. M.; AMOUYAL. O.; ELOITT. F.; HENRY, R. J. Sugarcane, In: KOLE, C. (Ed.). **Genome Mapping and Molecular Breeding in Plants: Pulses, Sugar and Tuber Crops**, New York: Springer, 2007, v. 3, pp. 175-204.

CRUSOE, M. R.; EDVENSON, G.; FISH, J.; HOWE, A.; McDONALD, E.; NAHUM, J. … BROW, T. C. The khmer software package: enabling efficient sequence analysis. doi: 10.6084/m9.figshare.979190, 2014.

D'HONT, A.; LU, Y. H.; LEÓN, D. G. D.; 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, 1994.

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**, Toronto v. 41, p. 221-225, 1998.

D'HONT, A. & GLASZMANN, J. C. Sugarcane genome analysis with molecular markers: a first decade of research. **Proceedings of International Society of Sugar Cane Technologists**, Quatre-Bornes, v. 2, p. 556-559, 2001.

D'HONT, A.; LU, Y. H.; FELDMANN, P.; GASZMANN, J. C. Oligoclonal interspecific origin of 'North Indian' and 'Chinese' sugarcanes. **Chromosome Research**, Irvine, v. 10, p. 253-262, 2004.

DA SILVA, J. A. D. **A methodology for genome mapping of auto-polyploids and its application to sugarcane (*Saccharum* spp.)**. Ph.D. dissertation, Cornell University, Ithaca, Nova York, 1993.

DAL-BIANCO, M.; CARNEIRO, M. S.; HOTTA, C. T.; CHAPOLA, R. G.; HOFFMANN, H. P.; GARCIA, A. A. F.; SOUZA, G. M. Sugarcane improvement: how far can we go? **Current Opinion in Biotechnology**, Madri, v. 23, p. 1-6, 2011.

DANECEK, P.; AUTON, A.; ABECASIS, G.; ALBERS, C. A.; BANKS, E.; DEPRISTO, M. A. … 1000 GENOMES PROJECT ANALYSIS GROUP. The Variant Call Format and VCFtools. Bioinformatics, London, v. 1, p. 1-3, 2011.

DANIELS, J. & ROACH, B. T. Taxonomy and Evolution, In: HEINZ, D. J. **Sugarcane Improvement through Breeding**. New York: Elsevier Science Publishing Company, 1987, cap. 3, pp. 7-84.

DAVEY, J. W.; BLAXTER, M. L. RADSeq: next generation population genetics. **Briefings in functional genomics**, Oxford, v. 9, n. 5, p. 416-423, 2011.

DE CAMPOS, T.; BENCHIMOL, L. L.; CARBONELL, S. A. M.; CHIORATTO, A. F.; FORMIGHIERI, E. F.; DE SOUZA, A. P. Microsatellites for genetic studies and breeding programs in common bean. **Pesquisa Agropecuária Brasileira**, Brasília, v. 42, n. 4, p. 589-592, 2007.

DECROOCQ, V. FAVÉ, M. G.; HAGEN, L. BORDENAVE, L.; DECROOCQ, S. Development and transferability of apricot and grape EST microsatellite markers across taxa. **Theoretical and Applied Genetics**, Stuttgart, v.106, n. 5, p. 912-922, 2003.

DEVOS, K. M. & GALE, M. D. Genome relationship: the grass model in current research. **Plant Cell**, Michigan, v. 1, n. 2, p. 636-646, 2000.

DILLIES, M.; RAU, A.; AUBERT, J.; ANTIER, C. H.; JEANMOUGIN, M.; SERVANT, N. … JAFFRE, F.; and on behalf of The French StatOmique. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. **Briefings in bioinformatics**, Oxford, v. 1, p. 1-13, 2012.

DING, G.; LIAO, Y.; YANG, M.; ZUNKANG, Z.; SHI, L.; XU, F. Development of gene-based markers from functional *Arabidopsis thaliana* genes involved in phosphorus homeostasis and mapping in *Brassica napus*. **Euphytica,** Wageningen, v. 181, n. 3, p. 305-322, 2011.

DOYLE, J. J.; FLAGEL, L. 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. **Annual review of genetics**, Madison, v. 42, p. 443-461, 2008.

EARL, D. A.; BRADNAM, K.; JOHN, J. S.; DARLING, A.; LIN, D.; FAAS, J. … PATEN, B. Assemblathon 1: A competitive assessment of de novo short read assembly methods. **Genome Reserch**, Baltimore, v. 24, n. 12, p. 2224-2241, 2011.

EWING, B.; HILLIER, L.; WENDL, M. C.; GREEN, P. Base-Calling of automated sequencer traces using Phred . I . Accuracy assessment. **Genome Research**, Baltimore, v. 8, p.175-185, 1998.

FAOSTATS: Food and Agricultural Organization of the United Nation, 2013. Disponível em <www,faostat3,fao,org>. Acessado em: 28 de novembro de 2013.

FERNANDES, A. C. & BRENDA, G. T. A. Distribution pattern of brix and fibre in the primary stalk of sugarcane. **International Sugarcane Journal**, London, v.5, p. 8-13, 1985.

FREELING, M. Grasses as a single genetic system. Reassessment. **Plant Physiology**, Los Angeles, v. 125, p. 1191-1197, 2001.

GADALETA, A.; MASTRANGELO, A.; RUSSO, M.; GIOVE, S.; D'ONOFRIO, O.; MANGO, T. ... CIFARELLI, R. A. Development and characterization of EST-derived SSRs from a 'totipotent' cDNA library of durum wheat. **Plant Breeding**, Bonn, v. 129, n. 6, p. 715-717, 2010

GALE, M. D. & DEVOS, K. Plant comparative genetics after 10 years. **Science**, New York, v. 282, p. 656-659, 1998.

GAO, L.; TANG, J.; LI, H.; JIA, J. Analysis of microsatellites in major crops assessed by computational and experimental approaches. **Molecular Breeding**, Lleida, v.12, p. 245-261, 2003.

GARBER, M.; GRABHERR, M. G.; GUTTMAN, M.; TRAPNELL, C. Computational methods for transcritoma annotation and quantification using RNA-Seq. **Nature Methods**, Madison, v. 8, n. 6, p. 469-477, 2011.

GARCIA, A. A .; MOLLINARI, M.; MARCONI, T. G.; SERANG, O. R; SILVA, R. R.; VIEIRA, M. L. C. ... SOUZA, A. P. SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. **Nature Scientific Reports**, Nova York, v. 3, n. 3399, 2013.

GARRISON, E. & MARTH, G. Haplotype-based variant detection from short-read sequencing. http://arxiv.org/pdf/1207.3907v2.pdf, p. 1-9, 2012.

GIACOMAZZI, E. **A brief history of brazilian PróÁlcool programme and developments of biofuel and biobased products in Brazil**. FIESP - Industry Federation of Sao Paulo State, Paris, 2012.

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

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

GODDARD, M. E. & HAYES, B. J. Genome Selection. **Journal of animal breeding and genetics**, Berlin, v. 14, p. 323-330, 2007.

GONG, Y. M.; XU, S. C.; MAO, W. H.; HU, Q. Z.; ZHANG, G. W.; DING, J.; LI, Y. D. Developing new SSR markers from ESTs of pea (*Pisum sativum* L.). **Journal of Zhejiang University Science B**, Zhejiang, v. 11, n. 9, p. 702-707, 2010.

GOODSTEIN , D. M.; SHU, S.; HOWSON, R.; NEUPANE, R.; HAYES, R. D.; FAZO, J.; MITROS, T.; DIRKS, W.; HELLSTEN, U.; PUTNAM, N.; ROKHSAR, D. S. Phytozome:

a comparative platform for green plant genomics. **Nucleic Acids Research**, Oxford, v. 40, p.1178-1186, 2011.

GOTZ, S.; ARNOLD, R.; SEBASTIÁN-LEÓN, P.; MARTÍN-RODRÍGUES, S.; TISCHLER, P.; JEHL, M. A. … CONESA, A. B2G-FAR, a species-centered GO annotation repository. **Bioinformatics**, Oxford, v. 27, n. 7, p. 919-924, 2011.

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

GREEN, E. D. Strategies for the systematic sequencing of complex genomes. **Nature Reviews Genetics**, New York, v. 2, p. 573-582, 2001.

GRIVET, L.; D'HONT, A.; ROQUES, D.; FELDMANN, P.; LANAUD, C.; GLASZMANN, J. C. RFLP Mapping in cultivated sugarcane (*Saccharum* spp.): genome organization in a highly polyploid and aneuploid interspecific hybrid. **Genetics**, New York, v. 142, p. 987-1000, 1996.

GRIVET, L.; GLASZMANN, J. C.; VINCENTZ, M.; DA SILVA, F.; ARRUDA, P. ESTs as a source for sequence polymorphism discovery in sugarcane: example of the Adh genes. **Theoretical Applied Genetics**, Sttutgart, v. 106, p. 190-197, 2003.

GRIVET, L.; DANIELS, C.; GLASZMANN, J. C.; D`HON. A Review of Recent Molecular Genetics Evidence for Sugarcane Evolution and Domestication. **Ethobotany Research & Applications**, Manoa, v. 2, n. 1, p. 9-17, 2004.

GROBA, S. Y. & BURGOS, J. I. M. Optimization of de novo transcriptome assembly from next generation sequencing data. **Genome Research**, Baltimore, v. 20, p. 1432-1440, 2010.

GROF, C. P. L. & CAMPBELL, J. A. Sugarcane sucrose metabolism: scope for molecular manipulation. **Australian Journal of Plant Physiology**, Hobart v. 28, p. 1-12, 2001.

GUPTA, P. K.; RUSTGI, S.; SHARMA, R.; SINGH, N.; KUMAR, H.; BALYAN, H. S. Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. **Molecular Genetics and Genomics**, Göteborg, v. 270, n. 4, p. 315-323, 2003.

GUTTMAN, M.; GARBER, M.; LEVIN, J. Z.; DONAGEY, J.; ROBINSON, J.; ADICONIS, X. … REGEV, A. *Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. **Nature Biotechnology**, New York, v. 28, p. 503-510, 2010.

HAAS, B. J. & ZODY, M. C. Advancing RNA-Seq analysis. **Nature Biotechnology**, New York, v. 28, n. 5, p. 421-423, 2010.

HAAS, B. J.; PAPANICOLAOU, A.; YASSOUR, M.; GRABHERR, M.; BLOOD, P. D.; BOWDEN, J. … REVEG, A. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. **Nature Protocols**, Madison, v. 8, n. 8, p. 1494-1512, 2013.

HANAI, L. R.; DE CAMPOS, T.; CAMARGO, L. E. A.; BENCHIMOL, L. L.; DE SOUZA, A. P.; MELOTTO, M. ... VIEIRA, M. L. C. Development, characterization, and comparative analysis of polymorphism at common bean SSR loci isolated from genic and genomic sources. **Genome**, Baltimore, v. 50, n. 3, p. 266-277, 2007.

HANCOCK, J. F. **Plant evolution and the origins of crop science**. 2º ed. London, 2004.

HENRY, R. J. Basic information on the Sugarcane plant. In: HENRY, R. J. & KOLE, C. (Ed.). **Genetics, Genomics and Plant Breeding of Sugarcane**. Enfield: Science Publisher, 2010, cap. 1, pp. 1-7.

HERMANN, S. R.; AITKEN, K. S.; JACKSON, P. A.; GEORGE, A. W.; PIPERIDIS, N.; WEI, X. … DETERING, F. Evidence for second division restitution as the basis for 2n + n maternal chromosome transmission in a sugarcane cross. **Euphytica**, Wageningen v. 187, n. 3, p. 359-368, 2012.

HIRSCH, C. N.; FOERSTER, J. M.; JOHNSON, J. M.; SEKHON, R.S.; MUTTONI, G.; VAILLANCOURT, B.; … BUELL, C. R. Insights into the maize pan-genome and pan-transcriptome. **Plant Cell**, Michigan, v.26, n.1 p. 121-35, 2014.

HOHENLOHE, P. A.; AMISH, S. J.; CATCHEN, J. M.; ALLENDORF, F. W.; LUIKART, G. Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. **Molecular Ecology Resource**, San Antonio, v. 11, p. 117-122, 2011.

HOLFORD, I. C. R. Soil phosphorus: its measurement and, its uptake by plants. **Australian Journal of Soil Research**, Sydney, v. 35, p. 227-239, 1997.

JACKSON, P. A. Breeding for improved sugar content in sugarcane. **Field Crops Research**, Bonn, v. 92, p. 277-290, 2005.

JANNINK, J. L.; LORENZ, A. J.; IWATA, H. Genomic selection in plant breeding: from theory to practice. **Briefings in Functional Genomics**, Oxford, v. 9, n. 2, p. 166-177, 2010.

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. **The Plant Journal**, Michigan, v. 50, p. 574-585, 2007.

KANEHISA, M. & GOTO, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. **Nucleic Acids Research**, Oxford, v. 28, n. 1, p. 27-30, 2000.

KAUR, G.; KUMAR, S.; NAYYAR, H.; UPADHYAYA, H. D. Cold stress injury during the pod-filling phase in chickpea (*Cicer arietinum* L,): effects on quantitative and qualitative components of seeds. **Journal of Agronomy Crop Science**, Pretoria, v. 194, n. 6, p. 457-464, 2008.

KAWAHARA, Y.; BASTIDE M.; HAMILTON, J. P.; KANAMORI, H.; McCOMBIE, W. R. ... MATSUMOTO, T. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. **The Rice Journal**, Tokyo, v. 6, n. 4, pp. 1-10, 2014.

KRUGLYAK, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. **Nature Genetics**, New York, v. 22, p. 134-144, 1999.

KURAMA, E. E.; FENILLE, R. C.; ROSA, V. E.; ROSA, D. D.; ULIAN, E. C. Mining the enzymes involved in the detoxification of reactive oxygen species (ROS) in sugarcane. **Molecular Plant Pathology**, Massachusetts, v. 3, n. 4, p. 251-259, 2002.

LAKSHMANAN, P.; GEIJSKES, J.; AITKEN, K. S.; GROF, C. L. P.; BONNETT, G. D.; SMITH, G. R. Sugarcane biotechnology: the challenges and opportunities. **In Vitro Cellular & Development Biology**, Mobile, v. 41, p. 345-363, 2005.

LANGMEAD, B.; TRAPNELL, C.; POP, M.; SALZBERG, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. **Genome biology**, London, v. 10, n. 3, p. R25, 2009.

LAVARACK, B. P.; GRIFFIN, G. J.; RODMANC, D. The acid hydrolysis of sugarcane bagasse hemicellulose to produce xylose, arabinose, glucose and other products. **Biomass and Bioenergy**, Aberdeen, v. 23, n. 5, p. 367-380, 2002.

LI, B. & DEWEY, C. N. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. **BMC Bioinformatics**, London, v. 12, p. 323, 2011.

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

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

LI, R.; YU, C.; LI, Y.; LAM, T. W.; YIU, S. M.; KRISTIANSEN, K.; WANG, J. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics, London, v. 25, n. 15, p. 1966-1967, 2009.

LINGLE, S. E. & SMITH, R. C. Sucrose metabolism related to growth and ripening in sugarcane internodes. **Crop Science**, Madison. v. 31, p. 172-177, 1991.

LIU, L.; LI, Y.; LI, S.; HU, N.; HE, Y.; PONG, R.; LIN, D.; LU, L.; LAW, M. Comparison of Next-Generation Sequencing Systems. **Journal of Biomedicine and Biotechnology**, Washington, v. 01, p. 1-11, 2012.

LIU, X.; HAN, S.; WANG, Z.; GELERNTER, J.; YANG, B. Z. Variant callers for next-generation sequencing data: a comparison study. **PLOS ONE**, Washington, v. 8, n. 9, p. 1-11, 2013.

LYSTER, R.; GREGORY, B. D.; ECKER, J. R. Next is now: new technologies for sequencing of genomes transcriptomes and beyond. **Plant Biology**, Berlin, v. 12, p. 107-118, 2009.

LOMAN, N. J.; MISRA, R. V.; DALLMAN, T. J.; CONSTANTINIDOU, C.; GHARBIA, S. E.; WAIN, J.; PALLEN, M. Performance comparison of benchtop high-throughput sequencing platforms. **Nature Biotecnology**, New York, v. 30, p. 434-439, 2012.

LU, T.; LU, G.; FAN, D.; ZHU, C.; LI, W.; ZHAO, Q. ... HAN, B. Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. **Genome Research**, Baltimore, v. 20, p. 1238-1249, 2010.

MAMMADOV, J.; AGGARWAL, R.; BUYYARAPU, R.; KUMPATLA, S. SNP markers and their impact on plant breeding. **International Journal of Plant Genomics**, Tashkent, v. 728398, p. 1-11, 2012.

MANNERS, J. M. & CASU, R. E. Transcriptome analysis and functional genomics of sugarcane. **Tropical Plant Biology**, Kunia, v. 4, p. 9-21, 2011.

MANTRI, N.; PATADE, V.; PENNA, S.; FORD, R.; PANG, E. Abiotic stress responses in plants: present and future. In: AHMAD, P. & PRASAD, M. N. V. (Ed.). **Abiotic stress responses in plants: metabolism, productivity and sustainability**. Kashmir, 2012, cap. 1, p. 1-20.

MAPA: Ministério da Agricultura Pecuária e Abastecimento. 2012. Disponível em <http://www,agricultura,gov,br/vegetal/culturas/cana-de-acucar>. Acessado em: 28 de novembro de 2013.

MARDIS, E. R. Next-Generation DNA sequencing methods. **Annual Review of Genomics and Human Genetics**, Baltimore, v. 9, p. 387- 402, 2008.

MASTERSON, J. Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. **Science**, New York, v. 264, p. 421-423, 1994.

MATSUMOTO, T.; WU, J.; KANAMORI, H.; KATAYOSE, Y.; FUJISAWA, M.; NAMIKI, N.; … BURR, B. The map-based sequence of the rice genome. **Nature**, Madison, v. 436, p. 793-800, 2005.

McCORMICK, A. J.; WATT, D. A.; CRAMER, M. D. Supply and demand: sink regulation of sugar accumulation in sugarcane. **Journal of Experimental Botany**, Lancaster v. 60, n. 2, p. 357-364, 2009.

McKENNA, A.; BANKS, H. M.; SIVACHENKO, B. E.; CIBULSKIS, K.; KERNYTSKY, A.; GARIMELLA, K. … DePRISTO, M. A. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. **Genome Research**, Baltome, v. 20, n.9, p. 1297-1303, 2010.

METZKER, M. L. Sequencing technologies – the next generation. **Nature Review Genetics**, New York, v. 11, p. 31-46, 2010.

MESSING, J.; BHARTI, A. K.; KARLOWSKI, W. M.; GUNDLACH, H.; KIM, H. R.; YU, Y.; … WING, R. A. Sequence composition and genome organization of maize. **Proceedings National Academic of Science**, San Diego, v. 101, n. 40, p. 14349-14354, 2004 .

MING, R.; LIU, S. C.; LIN, Y. R.; SILVA, J.; WILSON, W.; BRAGA, D. … PATERSON, A. H. Detailed alignment of *Saccharum* and *Sorghum* chromosomes: Comparative organization of closely related diploid and polyploid genomes. **Genetics**, Baltimore, v. 150, p. 1663-1682, 2008.

MOE, K. T.; HONG, W. J.; KWON, S. W.; PARK, Y. J. Development of cDNA-derived SSR markers and their efficiency in diversity assessment of *Cymbidium* accessions. **Electronic Journal of Biotechnology**, Valparaiso, v. 15, n. 2, p. 1-10, 2012.

MOORE, P. H. Integration of sucrose accumulation processes across hierarchical scales: towards developing an understanding of the gene to crop continuum. **Field Crops Research**, Bonn, v. 92, p. 119-135, 2005.

MORIN, P. A.; LUIKART, G.; WAYNE, R. K. SNPs in ecology, evolution and conservation. **Trends in Ecology & Evolution**,Cambridge, v. 19, p. 208-216, 2004.

MOROZOVA, O.; MARRA, M. A. Applications of next-generation sequencing technologies in functional genomics. **Genomics**, Boston, v. 95, n. 5, p. 255-264, 2008.

MORTAZAVI, A.; WILLIAMS, B. A.; McCUE, K.; SCHAEFFER, L.; WOLD, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. **Nature Mathods**, Madison, v. 5, n. 7, p. 621-628, 2008.

MORTON, B. R.; BI, I. V.; McMULLEN, M. D.; GAUT, B. S. Variation in mutation dynamics across the Maize genome as a function of regional and flanking base composition. **Genetics**, New York, v. 172, n. 1, p. 569-577, 2006.

MUKHERJEE, S. K. Origin and distribution of Saccharum. **Botanical Gazette Journal**, Waterloo, v. 119, p. 55-61, 1957.

NAGALAKSHMI, U.; WAERN, K.; SNYDER, M. RNA-Seq: A Method for Comprehensive Transcriptome Analysis. **Current protocol in molecular biology**, New York, v. 1, p. 1-13, 2010.

NAGARAJAN, S. & NAGARAJAN, S. Abiotic Tolerance and Crop Improvement. In: PAREEK, A.; SOPORY, S. K.; BOHNERT, H. J.; GOVINDJEE. (Ed.). **Abiotic Stress Adaptation in Plants: physiological, molecular and genomic foundation**. Amsterdam, 2010, cap. 1, p. 1-11.

NOGUEIRA, F. T.; DE ROSA, 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, 2003.

NOOKAEW, I.; PAPINI, M.; PORNPUTTAPONG, N.; SCALCINATI, G.; FAGERBER, L.; UHLÉN, M.; NIELSEN, J. A comprehensive comparison of RNA-Seq based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in *Saccharomyces cerevisiae*. **Acid Nucleic Research**, Oxford, v. 40, n. 20, p. 10084-10097, 2012.

NOVAES, E.; DEREK, R. D.; FARMERIE, W. G.; PAPPAS, G. J.; GRATTAPAGLIA, D.; SEDEROFF, R. R.; KIRST, M. High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. **BMC Genomics**, Oxford, v. 9, n. 312, p. 1-12, 2008.

O'NEIL, S. T. & EMRICH, S. J. Assessing de novo transcriptome assembly metrics for consistency and utility. BMC Genomics, Oxford, v. 14, n. 1, p. 1-12, 2013.

PARCHMAN, T. L.; GEIST, K. S.; GRAHNEN, J. A.; BENKMAN, C. W.; BUERKLE, C. A. Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. BMC Genomics, Oxford, v. 11, n. 110, p. 1-16, 2010.

PARIDA, S. K.; RAJKUMAR, K. A.; DALAL, V.; SINGH. N. K.; MOHAPATRA, T. Unigene derived microsatellite markers for the cereal genomes. **Theoretical Applied Genetics**, Sttutgart, v. 112, p. 808-817, 2006.

PARIDA, S. K.; KALIA, S. K.; KAUL, S.; DALAL, V.; HEMAPRABHA, G.; SELVI, A. … MOHAPATRA, T. Informative genomic microsatellite markers for efficient genotyping applications in sugarcane. **Theoretical Applied Genetics**, Sttutgart, v. 118, p. 327-338, 2009.

PATERSON, A. H.; BOWERS, J. E.; CHAPMAN, B. A. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. **Proceedings of the National Academy of Sciences**, San Diego, v. 101, n. 26, p. 9903-9908, 2004.

PATERSON, A. H. Polyploidy, evolutionary opportunity, and crop adaptation. **Genetica**, v. 123, p. 191-196, 2005.

PATERSON, A. H.; BOWERS, J. E.; BRUGGMANN, R.; BUBCHAK, I.; GRIMWOOD, J.; GUNDLACH, H. … ROKHSAR, D. S. The Sorghum bicolor genome and the diversification of grasses. **Nature**, Madison, v. 457, n. 7229, p. 551-556, 2009.

PATERSON, A. H.; SOUZA, G.; SLUYS, M. A. V.; MING, R.; D'HONT, A. Structural genomics and genome sequencing. In: HENRY, R. J. & KOLE, C. (Ed.). **Genetics, Genomics and Plant Breeding of Sugarcane**. Enfield: Science Publisher, 2010, cap. 8, pp. 150-165.

PÉREZ-DE-CASTRO, A. M.; VILANOVA S.; CANIZARES, J.; PASCUAL, L.; BLANCA J.; DÍEZ, M. ... PICÓ B. J. Application of Genomic Tools in Plant Breeding. **Current Genomics**, Paris, v. 13, p. 179-195, 2012.

PERKEL, J. M. Visiting "Noncodarnia". **BioTechniques**, New York, v. 54, p. 301-304, 2013.

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

QUACKENBUSH, J.; LIANG, F.; HOLT, I.; PERTEA, G.; UPTON, J. The TIGR Gene Index: reconstruction and representation of expressed gene sequences. **Nucleic Acid Research**, Oxford, v. 28, n. 1, p. 141-145, 2000.

QUAST, C.; PRUESSE, E.; YILMAZ, P.; GERKEN, J.; SCHWEER, T.; YARZA, P.; PEPLIES, J.; GLÖCKNER, F. O. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. **Nucleic Acids Research**, v. 41, p. D590-D596, 2013.

R DEVELOPMENT CORE TEAM. R: a language and environment for statistical computing. **R Fundation for Statistical Computing**, Viena, 2013.

RAFALSKI, J. A. Novel genetic mapping tools in plants: SNPs and LD-based approaches. **Plant Science**, Davis, v. 162, p. 329-333, 2002.

RAGHOTHAMA, K. G. Phosphate acquisition. **Annual Review of Plant Physiology and Plant Molecular Biology**, Oxford, v. 50, p. 665-686, 1999.

RESENDE, M. D. V.; LOPES, P. S.; SILVA, R. L.; PIRES, I. E. Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. **Pesquisa Florestal Brasileira**, Colombo, n. 56, p. 63, 2008.

ROACH, B. T. Nobilization of sugarcane. **Breeding and Genetics**, Madri, v. 1, p. 206-216, 1987.

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

ROSS, M. G.; RUSS, C.; COSTELLO, M.; HOLLINGER, A.; LENNON, N. J.; HEGARTY, R.; NUSBAUM, C.; JAFFE, D. B. Characterizing and measuring bias in sequence data. **Genome Biology**, London, v. 14, n. R51, p. 1-20, 2013.

SASHA, M. C.; MIAN, M. A. R.; EUJAVL, I.; ZWONITZER, J. C.; WANG, L.; MAY, G. D. Tall fescue EST-SSR markers with transferability across several grass species. **Theoretical and Applied Genetics**, Stuttgart, v. 109, n. 4, p. 783-791, 2004.

SAKANO, K. Proton/phosphate stoichiometry in uptake of inorganic phosphate by cultured cells of *Catharanthus roseus (L,) G.* **Plant Physiology**, Urbana, v. 67, p. 797-801, 1990.

SCHLOTTERER, C. The evolution of molecular markers - just a matter of fashion? **Nature Reviews Genetics**, New York, v. 5, p. 63-69, 2004.

SCHNABLE, P. S.; WARE, D.; FULTON, R. S.; STEIN, J. C.; WEI, F.; PASTERNAK, S. … WILSON, R. K. The B73 maize genome: complexity, diversity, and dynamics. **Science**, New York, v. 326, n. 5956, p. 1112-1115, 2009.

SCHULZ, M. H.; ZERBINO, D. R.; VINGRON, M.; BIRNEY, E. Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. **Bioinformatics**, London, v. 28, p. 1086-1092, 2012.

SCHUSTER, S. C. Next generation sequencing transforms today`s biology. **Nature Methods**, Madison, v. 5, p. 16-18, 2008.

SCHWARTZ, S. B. The early brazilian sugar industry, 1550-1670. **Revista de Indias**, Madrid, v. 65, n. 233, p. 79-116, 2005.

SEEB, J. E.; CARVALHO, G.; HAUSER, K.; NAISH, S.; ROBERTS, S.; SEEB, L. W. Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. **Molecular Ecology Resources**, San Diego, v. 11, p. 1-8, 2011.

SETTA, N.; VITORELLO, C. B. M.; METCALFE, J. C.; CRUZ, G. M. Q.; BEM, L. E. D. ... VAN-SLUYS, M. A. Building the sugarcane genome for biotechnology and identifying evolutionary trends. **BMC Genomics**, London, v. 15, n. 540, p. 1-17, 2014.

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

SHENDURE, J.; MITRA, R. D.; VARMA, C.; CHURCH, G. M. Advanced sequencing technologies: methods and goals. **Nature Reviews Genetics**, Madison, v. 5, p. 335-344, 2004.

SINGH, R. K.; MISHRA, S. K.; SINGH, S. P.; MISHRA, N.; SHARMA, M. L. Evaluation of microsatellite markers for genetic diversity analysis among sugarcane species and commercial hybrids. **Australian Journal of Crop Science**, Nova Scotia, v. 4, n. 2, p. 116-125, 2010.

SINGH, R. K.; SINGH, R. B.; SINGH, S. P.; SHARMA, M. L. Identification of sugarcane microsatellites associated to sugar content in sugarcane and transferability to other cereal genomes. **Euphytica**, Amesterdã, v. 182, n. 3, p. 335-354, 2011.

SOLTIS, D. E. & SOLTIS, P. S. Polyploidy: recurrent formation and genome evolution. **Trends in ecology & evolution**, London, v. 14, n. 9, p. 348-352, 1999.

SOLTIS, D. E.; ALBERT, V. A.; LEEBENS-MACK, J.; BELL, C. D.; ZHENG, C.; SANKOFF, D. … SOLTIS, P. S. Polyploidy and angiosperm diversification. **American Journal of Botany**, St. Louis, v. 96, p. 336-348, 2009.

SUPRASANNA, P.; PATADE, V. Y.; DESAI, N. S.; DEVARUMATH, R. M.; KAWAR, P. G.; PAGARIYA, M. C. ... BABU, K. H. Biotechnological Developments in Sugarcane Improvement: An Overview. **Sugar Tech**, Lucknow, v. 13, n. 4, p. 322-335, 2011.

SWIGONOVA, Z.; LAI, J.; MA, J.; RAMAKRISHNA, W.; LLACA, V.; BENNETZEN, J. L. MESSING, J. On the tetraploid origin of the maize genome. **Comparative and Functional Genome**, London, v. 5, n. 3, p. 281-284, 2004.

TEW, T. L.; COBILL, R. M. Genetic improvement of sugarcane (*Saccharum spp,)* as an energy crop. In: Vermerris, W. (Ed.). **Genetic Improvement of Bioenergy Crops**. New York: Springer, 2008, pp. 249-272.

THAKUR, P.; KUMAR, S.; MALIK, J. A.; BERGER, J. D.; NAYYAR, H. Cold stress effects on reproductive development in grain crops: an overview. **Environmental Experimental Botany**, Paris, v. 67, n. 3, p. 429-443, 2010.

THE GENE ONTOLOGY CONSORTIUM. The Gene Ontology project in 2008. **Nucleic Acids Research**, Oxford, v. 36, p. 440-444, 2008.

THIEBAUT, F.; ROJAS, C. A.; GRATIVOL, C.; MOTTA, M. R.; VIEIRA, T.; REGULSKI, M. ... FERREIRA, P. G. C. Genome-wide identification of microRNA and siRNA responsive to endophytic beneficial diazotrophic bacteria in maize. **BMC Genomics**, London, v. 15, n. 766, p. 1-18, 2014.

THOM, M.; MARETZKI, A. Peroxidase and esterase isozymes in Hawaiian sugar-cane. **Hawaiian Plant Research**, Hawai, v. 58, p. 81-94, 1970.

TOMKINS, J. P.; YU, Y.; SMITH, M. H.; FRISCH, D. A.; WOO, S. S.; WING, R. A. A bacterial artificial chromosome library for sugarcane. Theoretical Applied Genetics, Stuttgart, v. 3, n. 4, p. 419-424, 1999.

TRAPNELL, C.; ROBERTS, A.; GOFF, L.; PERTEA, G.; KIM, D.; KELLEY, D. R. … PACHTER, L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. **Nature Protocols**, Madison, v. 7, n. 3, p. 562-578, 2012b.

TRICK, M.; LONG, Y.; MENG, J.; BANCROFT, I. Single nucleotide polymorphism (SNP) discovery in the polyploid *Brassica napus* using Solexa transcriptome sequencing. Plant Biotechnology Journal, Atlanta, v. 7, p. 334-346, 2009.

TURCATTI, G.; ROMIEU, A.; FEDURCO, M.; TAIRI, A. P. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. **Nucleic Acids Research**, Oxford, v. 36, n. 4, p. 1-13, 2008.

VARSHNEY, R. K.; GRANER, A.; SORRELLS, M. E. Genic microsatellite markers in plants: features and applications. **Trends in Biotechnology**, San Diego, v. 23, p. 48-55, 2005.

VELÁZQUEZ, S. F.; GUERRA, R. R.; CALDERÓN, L. S. Abiotic and biotic stress response crosstalk in plants. In: SHANKER, A. K. & VENKATESWARLU, B. (Ed.). **Abiotic stress response in plants – Physiological, biochemical and genetic perspectives**. Rijeka, 2010, cap. 1, p. 3-26.

VENTURINI, L.; FERRARINI, A.; ZENONI, S.; TORNIELLI, G. B.; FASOLI, M.; SANTO, S. D. ... DELLEDONNE, M. *De novo* transcriptome characterization of *Vitis vinifera* cv. Corvina unveils varietal diversity. **BMC Genomics**, London, v. 14, n. 41, p. 1-13, 2013.

VETTORE, A. L.; SILVA, F. R.; KEMPER, E. L.; ARRUDA, P. The libraries that made SUCEST. **Genetics and Molecular Biology**, Ribeirão Preto, v. 24, n. 4, p. 1-7, 2001.

VETTORE, A. L.; DA SILVA, F. R.; KEMPER, E. L.; SOUZA, G. M.; DA SILVA, A. M.; FERRO, M. I. T. ... ARUUDA, P. Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. **Genome Research**, Nova York, v. 13, n. 12, p. 2725-2735, 2003.

VICENTINI, R.; DEL BEM, L. E. V.; VAN SLUYS, M. A.; NOGUEIRA, F. T. S.; VINCENTZ, M. Gene Content Analysis of Sugarcane Public ESTs Reveals Thousands of Missing Coding-Genes and an Unexpected Pool of Grasses Conserved ncRNAs. **Tropical Plant Biology**, Kunia, v. 5, n. 2, p. 199-205, 2012.

VIGNAL, A.; MILAN, D.; SANCRISTOBAL, M.; EGGEN, A. A review on SNP and other types of molecular markers and their use in animal genetics. **Genetics Selection Evolution**, Ames, v. 34, p. 275-305, 2002.

VITTE, C. & BENNETZEN, J, L. Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. **Proceedings of National Academic of Science**, San Diego, v. 103, p. 1763-17643, 2006.

XU, X.; LIU, X.; GE, S.; JENSEN, J. D.; HU, F.; LI, X. ... WANG, W. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. **Nature Biotechnology**, Madison, v. 30, p. 105-111, 2012.

YADAV, O. P.; MITCHELL, S. E.; FULTON, T. M.; KRESOVICH, S. Transferring molecular markers from sorghum, rice and other cereals to pearl millet and identifying polymorphic markers. **Journal of SAT Agricultural Research**, Nova Deli, v. 6, p. 1-4, 2008.

YANG, Z. & YODER, A. D. Estimation of the transition/transversion rate bias and species sampling. **Journal of Molecular Evolution**, Portland, v. 48, p. 274-283, 1999.

YU, X. & SUN, S. Comparing a few SNP calling algorithms using low-coverage sequencing data. **BMC Bioinformatics**, London, v. 14, n. 274, p. 1-15, 2013.

WALDRON, J. C. & GLASZIOU, K.T. Isozymes as a method of varietal identification in sugarcane. **Proceedings of the International Society for Sugarcane Technologists**, Quatre-Bornes, v. 14, p. 249-256, 1971.

WANG, J.; ROE, B.; MACMIL, S.; YU, Q.; MURRAY, J. E.; TANG, H. … MING, R. Microcollinearity between autopolyploid sugarcane and diploid sorghum genomes. **BMC Genomic**, London, v. 11, n. 261, p. 1-17, 2010.

WANG, J.; NAYAK, S.; KOCH, K.; MING, R. Carbon partitioning in sugarcane (*Saccharum* species). **Frontiers in Plant Science**, Tucson, v. 4, p. 1-6, 2013.

WANG, X.; LU, P.; LUO, Z. GMATo: A novel tool for the identification and analysis of microsatellites in large genomes. **Bioinformation**, Nova Deli, v.9, n.10, p. 541-544, 2013.

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

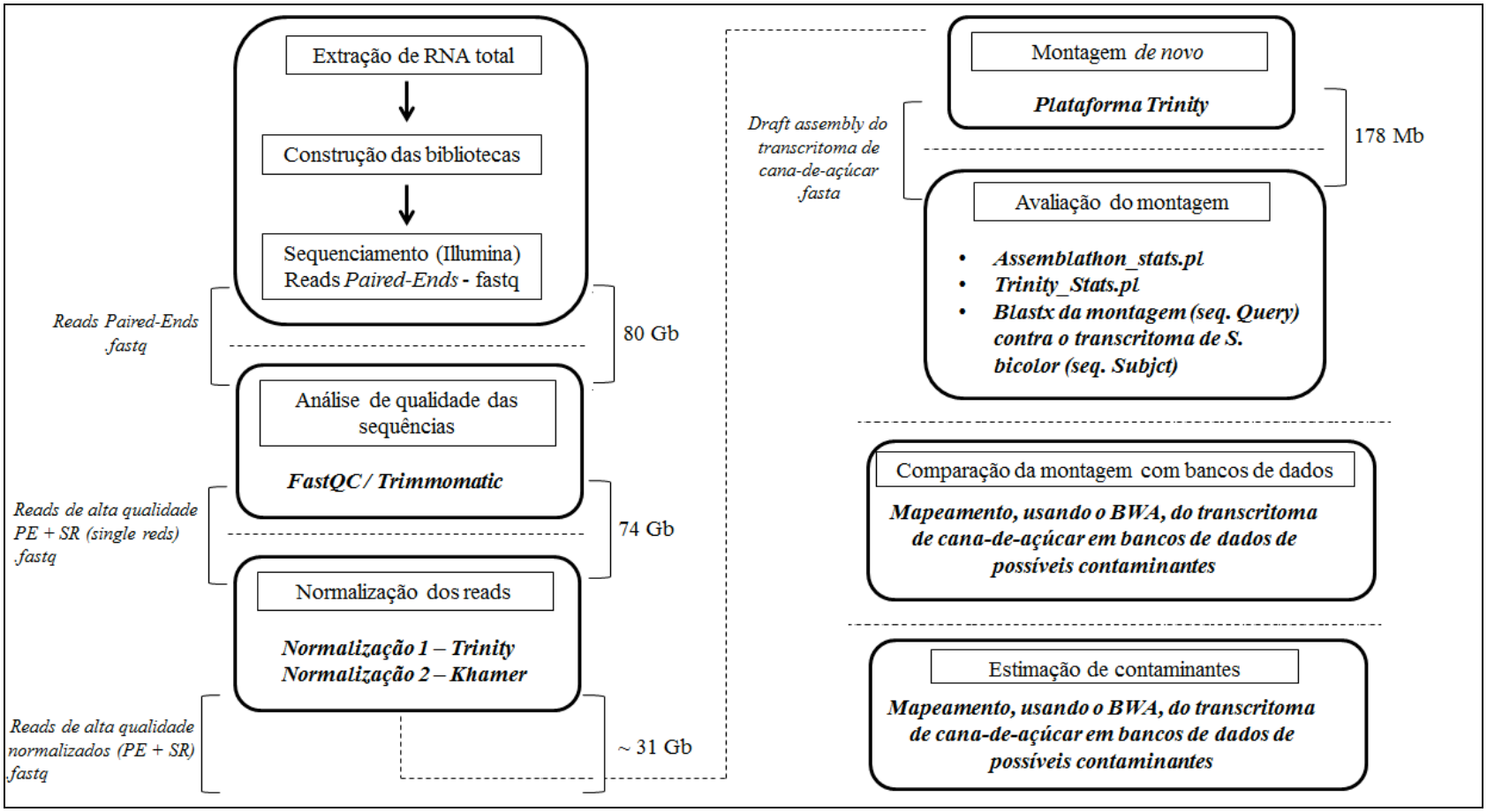
WENDEL, J. F. Genome evolution in polyploids. **Plant molecular biology**, Zurich, v. 42, n. 1, p. 225-249, 2000.

ZANDERSONS, J.; GRAVITIS, J.; KIKIREVICSA, A.; ZHURINSH, A.; BIKOVENS, O.; TARDENAKA, A.; SPINCE, B. Studies of the Brazilian sugarcane bagasse carbonisation process and products properties. **Biomass and Bioenergy**, Aberdeen, v. 17, n. 3, p. 209-219, 1999.

ZHANG, G.; LIU, X.; QUAN, Z.; CHENG, S.; XU, X.; PAN, S. … 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, 2012.

APÊNDICES

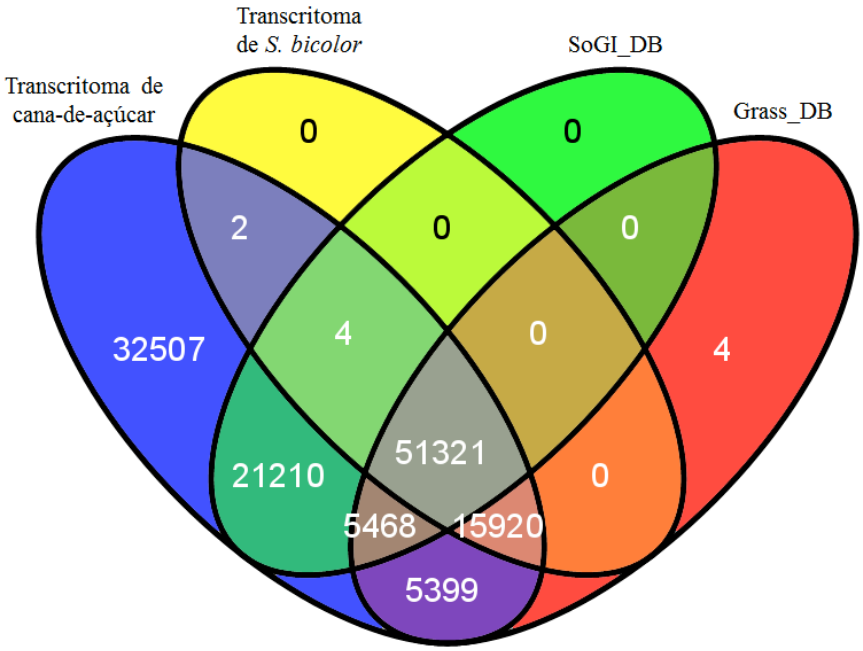
Apêndice A. Pipeline/workflow contendo os softwares utilizados em cada etapa das análises de bioinformática do Capítulo 1 da Tese ("Montagem do transcritoma de cana-de-açúcar (*Saccharum* spp.) utilizando dados de sequenciamento de nova geração").



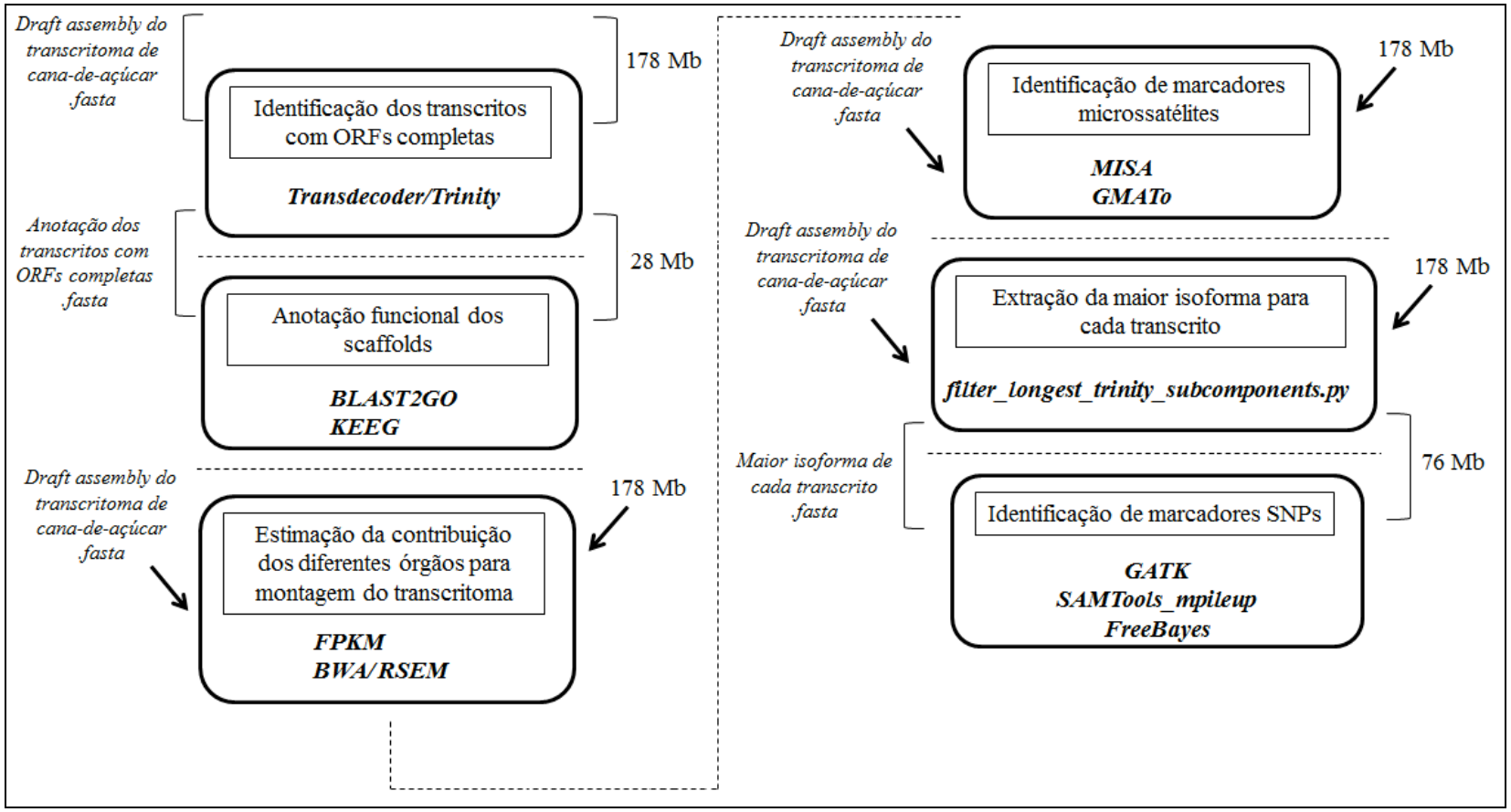
Apêndice B. Porcentagem de possíveis sequências contaminantes detectadas pelo alinhamento do transcritoma obtido para cana-de-açúcar contra bancos de dados de possíveis contaminantes obtidos do NCBI. A porcentagem de alinhamento revela baixa taxa de contaminantes. cpDNA = DNA cloroplastidial, mtDNA = DNA mitocondrial e rRNA = RNA ribossomal.

Banco de dados	**Número de reads alinhados**	**Porcentagem de alinhamento**
cpDNA de plantas	391774888	0,17
mtDNA de plantas	391774888	0,34
rRNA de Angiospermas	391774888	0,05
Genoma de *Echerichia coli*	391774888	0,04
Sequências de vetores	391774888	0,01
Média	--	0,122

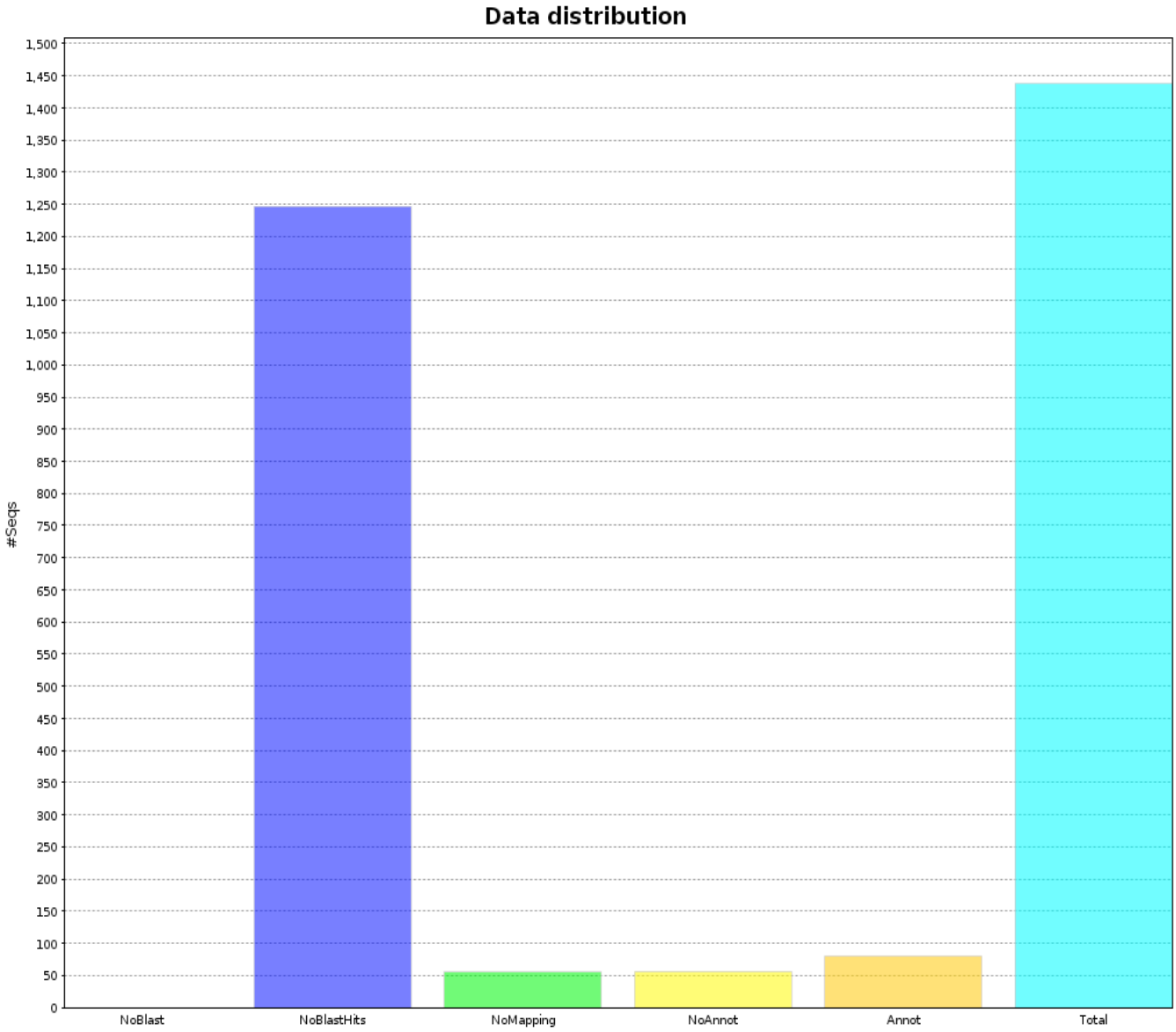
Apêndice C. Diagrama de Venn mostrando a comparação do transcritoma de cana-de-açúcar obtido com outros três bancos de dados. Os bancos de dados são: o transcritoma de *Sorghum bicolor*, o *Sacharum officinarum* Gene Index (SoGI) e o banco de dados formado pelo transcritoma de seis espécies (*Oryza sativa*, *Zea mays*, *Sorghum bicolor*, *Setaria itálica*, *Brachypodium distachyon* e *Panicum virgatum*) de gramíneas (Grass_DB). Existem 32.507 transcritos exclusivos de cana-de-açúcar.



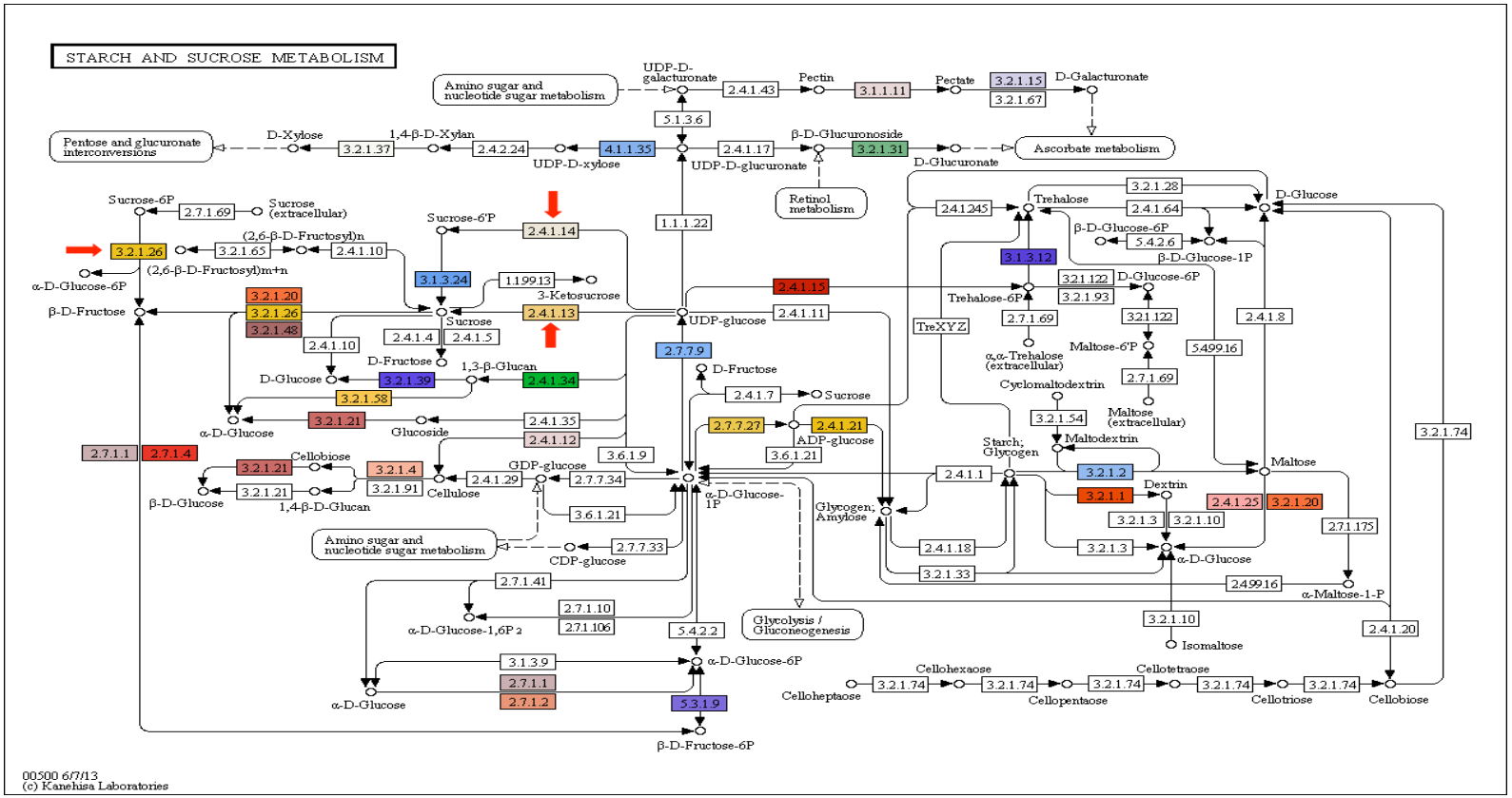
Apêndice D. Pipeline/workflow contendo os softwares utilizados em cada etapa das análises de bioinformática do Capítulo 2 da Tese ("Anotação e caracterização do transcritoma de cana-de-açúcar (*Saccharum* spp.) utilizando dados de sequenciamento de nova geração").



Apêndice E. Distribuição dos 1.380 transcritos que apresentam ORFs completas e não estão representados em nenhum dos três bancos de dados utilizados. Cerca de 1.250 destes transcritos não apresentam hits homólogos no banco de dados nr do NCBI, sendo considerados transcritos novos.



Apêndice F. Metabolismo do amido e da sacarose ativado por 234 transcritos amostrados do *draft assembly* do transcritoma de cana-de-açúcar. As principais enzimas (uma invertase (E.C.3.2.1.26), uma enzima sintetizadora de sacarose (*Sucrose Synthase* (SS), E.C.2.4.1.13) e uma enzima sintetizadora de fosfatos de sacarose (*Sucrose Phosphate Synthase* (SPS), E.C.2.4.1.14)) ativadas nesta via metabólica foram visualizadas nesta rota metabólica. Setas vermelhas indicam estas três enzimas na via metabólica. Estas enzimas foram descritas e caracterizadas por Chandra et al. (2012).



Apêndice G. Comparação entre as três ferramentas (GATK, SAMTools/mpileup e FreeBayes) utilizadas na identificação de variantes do tipo SNPs (a) e *indels* (b) no transcritoma de cana-de-açúcar obtido. Variáveis como a média do *Score* de Qualidade (c), a profundidade de sequenciamento (d), a diversidade nucleotídica (e) e a razão entre as taxas de Transição (Ts) e Transversão (Tv) (f), também foram estimadas.