Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Caracterização de polimorfismos e assinaturas de seleção em genótipos de cana-de-açúcar (*Saccharum* spp.) através de genotipagem-por-sequenciamento

Leonardo Sartori Menegatto

Dissertação apresentada para obtenção do título de Mestre em Ciências. Área de concentração: Genética e Melhoramento de Plantas

Piracicaba
2017

Leonardo Sartori Menegatto
Engenheiro Agrônomo

Caracterização de polimorfismos e assinaturas de seleção entre genótipos de cana-de-açúcar (*Saccharum* spp.) através de genotipagem-por-sequenciamento

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:
Prof. Dr. **GABRIEL RODRIGUES ALVES MARGARIDO**

Dissertação apresentada para obtenção do título de Mestre em Ciências. Área de concentração: Genética e Melhoramento de Plantas

Piracicaba
2017

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Menegatto, Leonardo Sartori

Caracterização de polimorfismos e assinaturas de seleção entre genótipos de cana-de-açúcar (*Saccharum spp.*) através de genotipagem-por-sequenciamento / Leonardo Sartori Menegatto - - versão revisada de acordo com a resolução CoPGr 6018 de 2011 - - Piracicaba, 2017.

96 p.

Dissertação (Mestrado) - - USP / Escola Superior de Agricultura “Luiz de Queiroz”.

1. SNP 2. Poliploide 3. Assinaturas de seleção 4. Genômica populacional 6. Variabilidade genética I. Título

AGRADECIMENTOS

Primeiramente, agradeço a Deus pela vida, sanidade intelectual e suporte em vencer as dificuldades e vilezas do mundo, em especial para a realização desse trabalho acadêmico.

Agradeço ao meu orientador, Prof. Dr. Gabriel Rodrigues Alves Margarido, pela tutela, ajuda, paciência e disponibilidade durante todo o meu período de mestrado, com destaque para a conclusão dessa dissertação. Friso que sua contribuição teve grande propulsão em minha maturidade profissional e compreensão científica.

Remerceio a postura caridosa e gratificante de meus colegas de curso, em especial dos participantes do Laboratório de Bioinformática Aplicada à Bioenergia, além dos docentes e funcionários do Programa de Pós-Graduação em Agronomia (Genética e Melhoramento de Plantas) e do Departamento de Genética. Gratulo, ainda, à Gloriosa Escola Superior de Agricultura "Luiz de Queiroz" e à Universidade de São Paulo pelo suporte acadêmico fornecido desde a minha graduação.

Agradeço ao Conselho Nacional Científico e Tecnológico (CNPq) e à Coordenadoria de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo auxílio financeiro, vital à conclusão dessa pesquisa.

Também agradeço às equipes da Profª. Dra. Monalisa Sampaio Carneiro, do Centro de Ciências Agrárias da Universidade Federal de São Carlos, do Prof. Dr. Antonio Augusto Franco Garcia, do Departamento de Genética da ESALQ-USP, e da Profª. Dra. Anete Pereira de Souza, do Departamento de Genética e Evolução do Instituto de Biologia e do Centro de Biologia Molecular e Engenharia Genética da Universidade Estadual de Campinas, pelo suporte com os dados desse trabalho.

Sou, ainda, grato à minha família, pela ajuda incomensurável em minha formação, e aos meus amigos, destacando-se os de Piracicaba, por serem sempre solícitos na resolução de problemas de cunho pessoal, intelectual e profissional durante a elaboração desse trabalho.

Finalmente, agradeço à sociedade paulista e ao povo brasileiro, que custearam minha formação durante esses importantes anos, reconhecendo minha imponente dívida social pelo privilégio concedido.

Todos os homens, por natureza, desejam saber

Aristóteles

SUMÁRIO

RESUMO

Caracterização de polimorfismos e assinaturas de seleção em genótipos de cana-de-açúcar (*Saccharum* spp.) através de genotipagem-por-sequenciamento

A cana-de-açúcar (*Saccharum* ssp.) é uma cultura valiosa na produção de alimento, fibra e energia para o Brasil e, especialmente, para o estado de São Paulo. Com o advento da biotecnologia, alternativas de melhoramento genético têm despertado a atenção da comunidade científica, sendo etapas cruciais para tais avanços o sequenciamento e a caracterização do genoma das espécies cultivadas. Dada sua natureza poliploide, com frequente aneuploidia, a cana-de-açúcar apresenta dificuldades às práticas convencionais em genômica, de maneira que é vantajoso fazer uso de recursos de sequenciamento de nova geração e de espécies próximas para elucidar de forma mais efetiva o genoma da gramínea. Uma contribuição interessante, nesse sentido, é a caracterização funcional de polimorfismos genéticos existentes entre genótipos do gênero *Saccharum*, auxiliando investigações relacionadas à genômica de poliploides complexos, desenvolvendo um recurso a ser utilizado futuramente por melhoristas. Esse trabalho realizou a caracterização da variabilidade genômica a partir de dados genotípicos de indivíduos do Painel Brasileiro de Genótipos de Cana-de-Açúcar, obtidos via genotipagem-por-sequenciamento, utilizando como referência o genoma já sequenciado do sorgo. Os sítios variantes (sobretudo polimorfismos de nucleotídeo único) foram detectados com o *software* FreeBayes e suas possíveis funções e posições foram anotadas com o programa SnpEff. Utilizaram-se estatísticas de genética de populações, como a frequência alélica para várias classes de polimorfimo, o Teste de McDonald & Kreitman (busca de evidências de evolução adaptativa) e a heterozigosidade combinada (busca de regiões genômicas com assinatura de seleção), de modo a identificar regiões genômicas potencialmente envolvidas em eventos evolutivos. Os resultados demonstraram a perda de variabilidade entre os genótipos melhorados em relação aos ancestrais, com evidências de assinaturas de seleção, envolvendo questões sensíveis ao funcionamento da maquinaria celular (como respiração e fotossíntese) e a características valoradas para a cultura (destacando-se a resistência a patógenos e a biossíntese da sacarose). Tais indícios fornecem subsídios à compreensão do genoma e ao melhoramento genético desse poliploide.

Palavras-chave: SNP; Poliploide; Assinaturas de seleção; Genômica populacional; Variabilidade genética

ABSTRACT

Characterization of polymorphisms and selection signatures in sugarcane genotypes (*Saccharum* spp.) by genotyping-by-sequencing

Sugarcane (*Saccharum* ssp.) is a valuable crop for food, fiber and energy production in Brazil, especially to the São Paulo State. With the advent of biotechnology, alternatives to breeding have enticed attention of the scientific community, with genome sequencing and characterization being crucial steps to these advances. Because sugarcane is polyploid, with frequent aneuploidy, it presents difficulties to the application of standard practices in genomics, such that it is advantageous to make use of next generation sequencing alternatives and resources from related species to more effectively elucidate the genome of this grass. Thus, an interesting contribution is the functional characterization of genetic polymorphisms from the *Saccharum* genus, aiding investigations related to genomics of complex polyploids, developing a resource to be used in the future by breeders. Our goal was to perform this characterization with genotypic data from individuals of the Brazilian Panel of Sugarcane Genotypes, obtained by genotyping-by-sequencing (GBS), using as reference the previously sequenced sorghum genome. We called the variants (mainly single nucleotide polymorphisms) with FreeBayes and annotated their functions and positions with SnpEff. We used population genetics statistics, such as the allele frequency, the McDonald & Kreitman Test and the pooled heterozygosity, to identify genomic regions potentially involved in evolutionary events. The results showed a loss of variability between bred genotypes in relation to the ancestors, with evidences of selective sweeps, involving regions related to the cellular machinery (such as respiration and photosynthesis) and specific crop traits (especially disease resistance and sucrose biosynthesis). These results support understanding of the genome and breeding efforts in this polyploid grass.

Keywords: SNP; Polyploid; Selective sweeps; Population genomics; Genetic variability

1 INTRODUÇÃO

A atividade agropecuária é estimada como existente desde cerca de 10.000 a.C. Separadamente, espécies animais e vegetais tiveram sua reprodução compreendida e formaram-se criações e cultivos, permitindo o desenvolvimento do que se conheceria como Revolução Agrícola ainda no Neolítico, marcando o início de uma nova forma de relação entre homem e ambiente. Consequentemente, o homem desenvolveu métodos empíricos primitivos de seleção de vegetais e animais, que, mais tarde aliados a cruzamentos intencionais, deram origem ao que se chamaria de domesticação (MAZOYER & ROUDART, 2006).

Esse melhoramento genético inicial tornou-se um imponente aliado da sociedade em obter alimentos, fibras e energia. As inovações agrícolas foram, então, exploradas sob o ponto de vista ambiental, com inovações presentes no século XVIII e depois no decorrer dos séculos subsequentes. Sob o ponto de vista genético, foram preponderantes os estudos em evolução de Charles Darwin e inaugurais na ciência genética por Gregor Mendel, cuja unificação na década de 1900 e os subsequentes adventos das genéticas de populações e quantitativa nas três décadas subsequentes impulsionaram o desenvolvimento de metodologias de melhoramento. Uma nova era, porém, adentrou-se com a descoberta do ácido desoxirrobonucleico (DNA) como material genético na década de 1940 e o conhecimento de sua estrutura e funcionamento na década de 1950, casados a outras áreas no desenrolar da Revolução Técnico-Científico-Informacional, e consequentemente da Revolução Verde, nas décadas de 1960 e 1970. Por fim, com a bioinformática, uma nova perspectiva de oportunidades inaugurou-se, com o manejo de dados moleculares complexos, com destaque para a genômica (VEIGA, 1991; CONWAY, 1997).

Considerando as novas tecnologias disponíveis, como a seleção genômica e estudos de mapeamento associativo (*genome-wide association study* - GWAS), incentivou-se abundantemente estudos genômicos em agropecuária, obtendo-se, inclusive, o sequenciamento do genoma de diversas culturas agrícolas de relevância mundial, incluindo muitas gramíneas e leguminosas. Adicionalmente, estudos de genética de populações, importantes no manejo de recursos genéticos para fins de conservação ou melhoramento genético, passaram a ter contornos diferenciados (DEPRISTO *et al.*, 2011). Possibilitou-se a caracterização de populações investigando-se sua variabilidade e compreendendo aspectos de sua evolução, fornecendo resultados sob os pontos de vista histórico, ecológico e tecnológico.

Nesse contexto, a caracterização funcional de locos variantes em genomas de plantas cultivadas passou a ser bastante valorada no meio científico. Genomas poliploides e/ou com eventos recorrentes de aneuploidia e aberrações cromossômicas estruturais (notoriamente elementos de transposição), contudo, apresentam relativa dificuldade de serem sequenciados, sendo custosa e trabalhosa a obtenção de um genoma de referência. Desse fenômeno decorreram dois tipos de estudos. Primeiramente, foram desenvolvidas metodologias de genotipagem simplificadas, sem a necessidade de genoma de referência e focadas em regiões hipometiladas, como a genotipagem-por-sequenciamento (ELSHIRE *et al.*, 2011). Secundariamente, elaboraram-se estudos utilizando genomas de espécies aparentadas.

A cana-de-açúcar é uma cultura de importância histórica no Brasil. Em São Paulo constitui-se a mais importante cultura agrícola, da qual se obtém açúcar, álcool, energia e diversos subprodutos (GOLDEMBERG *et al.*, 2008). Seu melhoramento genético é relativamente recente, começando no final do século XIX em nível mundial e apenas na década de 1930 em nível nacional (FIGUEIREDO, 2008). Quedas de lucratividade nesse setor e o menor desenvolvimento de técnicas de manejo em fitossanidade (BARROS *et al.*, 2014; TOKESHI & RAGO, 2005) valoram aos recursos genéticos uma contribuição expressiva. No entanto, pontua-se que os cultivares usados comercialmente são híbridos interespecíficos, de genoma altamente poliploide e com frequentes eventos de aneuploidia (JANOO, 2007). Assim, é notória a maior restrição em estudos genômicos com essa gramínea, sendo importantes trabalhos genômicos que auxiliem sua elucidação, bem como forneçam subsídios ao melhoramento genético.

Sob esse âmbito, pesquisas que aliem dados genotípicos a *softwares* e metodologias de trabalho em poliploides complexos são relevantes. Tendo em vista o interesse na caracterização funcional de polimorfismos em genótipos de cana-de-açúcar, propõe-se a realização de um trabalho que detecte sítios variantes que possivelmente tenham sido selecionados ao longo da evolução e melhoramento da cultura. Para tal, combina-se um genoma de referência de uma espécie evolutivamente aparentada com um programa de detecção de variações com o uso de metodologias de bioinformática e de estatísticas de genética de populações para, então, atribuí-los função e anotação posicional. Com isso, pretendeu-se observar flutuações nas frequências alélicas dos polimorfismos e a presença de assinaturas de seleção no genoma da cana.

2 REVISÃO BIBLIOGRÁFICA

2.1 A cultura da cana-de-açúcar

2.1.1 Aspectos botânicos

A cana-de-açúcar (*Saccharum* spp.) é uma monocotiledônea da família Poaceae (gramínea), pertencente à tribo Andropogoneae, subtribo Saccharinae. Constitui-se de um complexo de espécies poliploides, notoriamente *S. officinarum* L. e *S. spontaneum* L. na formação de variedades comerciais (BLACKBURN, 1984).

Segundo Van Dillewijin (1952), a planta possui folhas paralelinérveas, de inserção oposta, com bordos serrilhados (corpos silicosos), às vezes com pelos (joçal) e estômatos presentes nas duas faces, com destaque para a abaxial. De porte herbáceo, seu hábito de crescimento é determinado, com raízes superficiais, de sustentação e de cordão, sendo as primeiras mais finas e superficiais, as segundas mais grossas e aprofundadas e as últimas aglomerações profundas. Seu caule é denominado colmo, sendo o tolete o colmo plantado horizontalmente e de onde se replicam os perfilhos.

A sua conformação foliar e sua alta necessidade de insolação relacionam-se à sua elevada eficiência fotossintética. Como outras herbáceas tropicais, seu metabolismo é dito C_4. Após a fase fotoquímica, o dióxido de carbono proveniente da atmosfera não participa imediatamente do ciclo de Calvin, mas antes forma oxalacetato e integra uma rota metabólica alternativa no mesofilo e nas células de Kranz (bainha). Tal diferença culmina no grande aproveitamento luminoso, na não sensibilidade a variações de temperatura na fase química da fotossíntese e na ausência de fotorrespiração (VAN DILLEWIJIN, 1952; UENO, 1996).

Como observam Daniels *et al.* (1975), genótipos de *S. officinarum* possuem boas características produtivas, como alto nível de sacarose no colmo, tamanho avantajado de colmos, baixa concentração de impurezas no caldo e teor de fibras adequado para a moagem, embora possuam altas exigências referentes a solo e clima. Por sua vez, genótipos de *S. spontaneum* são marcantes pela rusticidade, com características favoráveis de vigor, perfilhamento, capacidade de rebrota, expansão do sistema radicular no solo e, principalmente, resistência a pragas e doenças. Seus inconvenientes são o menor porte, colmos curtos, finos e fibrosos e com baixo teor de açúcar.

É importante enfatizar que, dentre as espécies rústicas secundárias, *S. robustum* L. apresenta apenas relevância nos híbridos havaianos, possuindo porte alto, colmos finos e baixo teor de sacarose. *S. sinensis* Roxb., por sua vez, apresenta colmos finos e fibrosos,

ocasionalmente com alto teor de sacarose, e sistema radicular desenvolvido e que suporta estresse hídrico. Por fim, *S. barberi* Jeswiet possui porte médio e colmos finos, fibrosos e pobres em sacarose. *S. edule* Hassk. e *S. bengalensis* Retz. não apresentam importância na história evolutiva dos cultivares comerciais, sendo basicamente utilizadas para alimentações humana na Nova Guiné e animal na Índia, respectivamente (DANIELS *et al.*, 1975; PIPERIDIS, 2000).

A subtribo Saccharinae também inclui os gêneros *Miscanthus*, *Erianthus* e *Sorghum*. O primeiro é uma erva de alto perfilhamento e colmos extremamente finos, que compartilha características de rusticidade com algumas das espécies citadas do gênero *Saccharum*, sendo cultivado para produção de biodiesel e uso ornamental (USDA, 2015). O *Erianthus*, por sua vez, é conhecido como planta daninha em diversas regiões do mundo, embora possa ser utilizado com finalidade ornamental e para prevenção de erosão do solo, possuindo características de rusticidade semelhantes ao *Miscanthus*, mas com colmos com menor teor de açúcar (BERDING *et al.*, 1997).

No que se refere ao gênero *Sorghum*, destaca-se a cultura do sorgo (*S. bicolor* (L.) Moench), com usos relativos às alimentações animal e humana e produções de fibra e biodiesel (PATERSON *et al.*, 2009). Seu genoma diploide, composto de dez pares de cromossomos, foi já sequenciado e anotado. Tal anotação foi facilitada pelo fato de seus números e famílias gênicos serem semelhantes à brassicácea *Arabidopsis thaliana* (L.) Heynh. (conhecidamente uma planta-modelo), como observado por Paterson *et al*. (2000).

Considerando esse arcabouço de informações, incluindo gêneros aparentados, espécies diversas de *Saccharum* e seus cultivares híbridos intra e interespecíficos, foi compilado o Painel Brasileiro de Genótipos de Cana-de-Açúcar (PBGCA), procurando reunir a grande variabilidade de germoplasma existente. Esse painel inclui espécies nativas ancestrais de cana, como genótipos das espécies citadas, além de híbridos não pertencentes a programas de melhoramento recente de *S. officinarum* com *S. spontaneum*, de *S. officinarum* com *S. robustum*, de *S. spontaneum* com *S. bengalensis* e envolvendo mais de duas espécies, bem como espécies aparentadas (*Erianthus spp.* e *Miscanthus* spp.). Também estão inclusos nesse painel cultivares melhorados antigos e recentes do Brasil e do exterior, abarcando 18 programas de melhoramento genético de onze países em quatro continentes. Uma descrição de um subconjunto desses genótipos encontra-se disponível no trabalho de Garcia *et al.* (2013).

2.1.2 Cultivo e importância econômica

Os principais produtos da cana são o açúcar e o etanol, tendo como subprodutos principalmente a vinhaça e a torta de filtro (utilizadas na adubação potássica, nitrogenada e fosforada), além do uso do bagaço na cogeração de energia das usinas com a queima da palhada. Do etanol inclui-se a produção de aguardente, de produtos de uso industrial e doméstico e de combustível, sendo esse de grande relevância com a preocupação global em diminuir a dependência dos combustíveis fósseis. É ainda significativo o uso da planta inteira como forragem para a produção de silagem objetivando a alimentação animal (GOLDEMBERG *et al.*, 2008).

Como observam Marin & Carvalho (2012), a planta possui forte adaptação às condições edafoclimáticas do estado de São Paulo, considerando que as exigências estão intimamente ligadas ao clima tropical e subtropical (com temperatura média adequada entre 25 e 34°C) (EMBRAPA, 2000). A presença de estação seca é vital para produzir estresse para o acúmulo de açúcar. Além disso, a gramínea requer solos profundos, uma vez que é naturalmente perene e tratada como semianual. Nesse contexto, de acordo com o último Censo Agropecuário (IBGE, 2016), os maiores produtores brasileiros são os estados de São Paulo (48,00%), Goiás (15,36%), Mato Grosso do Sul (9,23%), Minas Gerais (9,19%) e Paraná (5,90%). No mundo, depois do Brasil, os maiores produtores são Índia, China, Tailândia e México (FAO, 2016).

Características de manejo são variáveis se em primeiro corte ou nos cortes subsequentes e de acordo com a época de plantio (MARIN & CARVALHO, 2012). As fases fenológicas da cana dividem-se em vegetativas, que incluiem brotação, perfilhamento e elongação, e de maturação, uma vez que a fase reprodutiva (florescimento) é evitada por comprometer o teor de sacarose no colmo, cujo fenômeno é denominado isoporização. Genericamente, as primeiras fases são as mais críticas quanto a danos de pragas, doenças e plantas daninhas. Importante destacar que a patente falta de produtos fitossanitários para uso em canaviais faz dos programas de melhoramento genético a principal metodologia de controle a patologias, pragas e estresses abióticos na cultura (GALLO *et al.*, 2002; TOKESHI & RAGO, 2005; EMBRAPA, 2016).

De acordo com o MAPA (2016), o Brasil corresponde a mais da metade do comércio mundial de açúcar, esperando colher na safra de 2018/2019 47,34 milhões de toneladas de colmo. Barros *et al.* (2014) expõem que no ano de 2013 o setor sucroalcooleiro nacional foi responsável por cerca de 6,2% das exportações de produtos primários para a China, 16,4%

para os EUA e 3,4% para a União Europeia (respectivamente os principais parceiros comerciais do país), totalizando um volume de divisas superior a US$ 3 bilhões.

Em contraposição, o setor canavieiro brasileiro vive uma retração significativa, o que realça a postura dos produtores de traçar estratégias contínuas para garantir a produtividade e a rentabilidade perante o mercado. Além de medidas nas políticas econômicas, de posicionamento comercial e de mudanças ambientais nas lavouras, frisa-se em grande nível a exploração de recursos genéticos. Nesse sentido, tanto programas de melhoramento genético conduzidos através de técnicas clássicas, como iniciativas baseadas no uso de métodos biotecnológicos, são importantes contribuintes para o desenvolvimento competitivo da cana-de-açúcar.

2.1.3 Origem e histórico

O centro de origem da cana é incerto, embora seja apontado no Sudeste Asiático, com reportagens na Polinésia (Fiji e Taiti), Índia, Nova Guiné e China. Nesse sentido, apresenta-se um plural de centros de diversidade, sendo Nova Guiné para *S. officinarum* e *S. robustum*, China para *S. sinensis*, norte da Índia para *S. barberi* e *S. spontaneum*, oeste da Índia para *S. bengalensis* e Melanésia e Polinésia para *S. edule*. Ao fim, importante destacar que variedades nobres de *S. officinarum* já eram conhecidas nas localidades indianas de Assam e Bengala desde 6.000 a.C. (ROACH & DANIELS, 1987; DANIELS *et al.*, 1975).

Segundo Clayton *et al.* (1975), a hipótese mais aceita a respeito da expansão da cana-de-açúcar é que fora cultivada inicialmente na região do Golfo da Bengala, espalhando-se pelo uso dos chineses, persas e árabes. Atribui-se aos últimos a dispersão pelo sul e leste mediterrâneos para uso ornamental a partir da conquista de territórios, sendo a planta cultivada para produção açucareira por Espanha e Sicília. Finalmente, o fim do oligopólio de Gênova e Veneza com o Renascimento tornou o açúcar um produto global, exercendo forte influência sobre as Grandes Navegações ibéricas.

No Brasil, o cultivo da cana é bastante antigo. Como afirma Figueiredo (2008), relatos de sua introdução datam de 1502. No entanto, mudas para amplo cultivo só foram trazidas da Ilha da Madeira para a Capitania de São Vicente em 1532 por Martim Afonso de Souza, quando a região já era objeto de colonização. A exploração da cultura nas antigas colônias portuguesas foi consequência direta das Grandes Navegações, tornando-se base de suas economias, com destaque para Brasil, Agave, Ilha da Madeira, Açores, Canárias, Cabo Verde e São Tomé.

Em meio a tal situação, o capitão-donatário de Pernambuco, Duarte Coelho, iniciou a construção de engenhos de açúcar, permitindo que a cultura prosperasse em função das condições climáticas do local. Em pouco tempo, as capitanias de Alagoas, Bahia e Piauí seguiram o mesmo caminho, conquistando, junto a Pernambuco, o monopólio no comércio mundial de açúcar. Tal desempenho continuou semelhante até que a competição com as antilhas holandesas, em função da União Ibérica, produzissem o declínio da produção nacional, encerrando um ciclo produtivo. Um novo ciclo deu-se com a introdução de cana caiana no século XIX e com a realização de programas de melhoramento genético na primeira metade do século XX. Objetivando a resistência a doenças, tal conjuntura direcionou o plantio ao estado de São Paulo. Por fim, o Proálcool, resposta do governo militar à crise do petróleo de 1975, incentivou a produção de etanol no Brasil, tornando-o o maior produtor sucroalcooleiro mundial (COSTA, 1958; FIGUEIREDO, 2008).

Por séculos não se aplicaram técnicas significativas de melhoramento genético à gramínea, dado o desconhecimento de plantas férteis, o que mudou no fim do século XIX com o aparecimento de programas internacionais (com destaque para Java, na Indonésia) (Jeswit, 1930). Como ressaltam Scarpari & Beauclair (2008), a hibridação interespecífica da cana criou uma situação de grande benefício a seu cultivo, especialmente por aliar as boas características produtivas e industriais de *S. officinarum* e a rusticidade de *S. spontaneum*.

A saber, os principais programas de melhoramento genético de cana-de-açúcar no exterior ocorreram na Argentina, na África do Sul, na Austrália, na Colômbia, em Cuba, nos EUA, na Índia e na Indonésia (SCARPARI & BEAUCLAIR, 2008). São ainda relevantes outros sete países, além do protetorado de Porto Rico e do departamento ultramarino das Ilhas Reunião (Tabela 1).

TABELA 1 – Relação de programas de melhoramento genético de cana-de-açúcar historicamente relevantes no exterior e seus respectivos países ou territórios.

País/Território	Programa	Sigla
África do Sul	Natal	N
	Natal-Coimbatore	Nco
Argentina	Tucman	Tu
	Norte de Argentina	NA
Austrália	Queensland	Q
Barbados	Barbados	B
	Barbados-Trinidade	BT
Bolívia	Santa Cruz	CIMCA
Colômbia	Instituto Colombiano Agropecuario	ICA
Cuba	Cuba	C
EUA	Clewiston/U.S. Sugar Corporation	CL
	Canal Point	CP
	Florida/Belle Grade Experimental Station	F
	Louisiana Experimental Station	L
	Hawaii Sugar Planter Association	H
Fiji	Mali	M
Filipinas	Philippines	P
Ilhas Reunião	Réunion	R
Índia	Coimbatore	Co
Indonésia	Proefstation Oest Java	POJ
Porto Rico	Puerto Rico	PR
República Dominicana	Central Romana	CR
Taiwan	Taiwan Sugar Manufactured Co.	TA
Venezuela	Venezuela	V

No Brasil, cultivou-se durante todo o ciclo canavieiro o cultivar Crioula, também conhecida como Creoula ou Mirim, idêntica ao cultivar Puri, que consistia em um híbrido entre *S. officinarum* e *S. barberi*. Posteriormente, por ação da vinda da família real portuguesa ao Rio de Janeiro (1808), a abertura de portos e consequente introdução de tecnologia ao Brasil trouxe a variedade de *S. officinarum* Caiana, anteriormente conhecida como Caiena. Oriunda da Guiana Francesa, essa se estabeleceu como a mais cultivada entre 1810 e 1880 (COSTA, 1958).

A cana Caiana sofreu, ao final do século XIX, severos danos por conta da gomose (*Xanthomonas axonopodis* Hasse), sendo substituída pelas variedades Roxa, Rosa, Salangor, Louiser, Bois Rouge Cavangirie, Ubá e Cristalina. Na década de 1920, o mosaico da cana disseminou-se pelo Brasil, criando a necessidade do uso de híbridos. Foram introduzidos

cultivares resistentes POJ, Co e CP na Estação Experimental de Piracicaba, embora somente os dois primeiros tenham se estabelecido (FIGUEIREDO, 2008).

O primeiro programa de melhoramento genético de sucesso do país ocorreu em 1934 por ação do Instituto Agronômico de Campinas (IAC), originando o cultivar IAC36-25. Forrageira, essa foi substituída por outras variedades do IAC no fim da primeira metade do século XX, quando cultivares de Campos (CB), no estado do Rio de Janeiro, foram disseminados a partir da Estação Experimental de Piracicaba. Nas décadas de 1970 e 1980, o cultivar NA56-79 tornou-se o mais popular do país, sendo substituído nas décadas posteriores pelos cultivares da Copersucar (SP) e da Planalsucar (RB), essa hoje denominada Rede Universitária para o Desenvolvimento Sucroalcooleiro (RIDESA). São importantes, ainda, os programas de melhoramento do Centro de Tecnologia Canavieira (CTC), de Pedro Ometo - Usina da Barra (PO) e de Pernambuco (PB) (COSTA, 1958; SZMRECSÁNYI & MOREIRA, 1991).

2.2 Melhoramento genético

2.2.1 Aspectos gerais

Um programa de melhoramento genético consiste na adoção, por avaliações mensuradas por parâmetros genéticos adequados, de etapas de seleção e cruzamento de organismos visando à obtenção de variedades superiores para alguma característica (BOS & CALIGARI, 1995). Dentre os caracteres, pode-se elencar resistências a estresses abióticos, injúrias e doenças, aumento de produtividade, aumento da facilidade de manejo no campo e maior teor de características de interesse industrial.

Historicamente, o melhoramento passou por três fases marcantes. A primeira, relativa ao empirismo, em que os cruzamentos e a seleção não obedeciam a critérios científicos, mas experiências corriqueiras de agricultores. A segunda etapa, já em posse de trabalhos de genética e evolução, foi marcada pela integração de fenótipo e genótipo. Por fim, a era da biotecnologia teve como compromisso a rapidez na genotipagem e a exploração de eventos que não ocorreriam naturalmente. Incluíram-se, pois, técnicas como fusão de protoplastos e produção de organismos geneticamente modificados. Nesse contexto, a marcação molecular tem desempenhado um importante papel, encurtando a etapa de seleção e melhorando significativamente a acurácia de estimações importantes, como testes de progênie (SIM *et al.*, 2011; EATHINGTON *et al*., 2007).

2.2.2 Melhoramento da cana-de-açúcar

Os programas de melhoramento genético de cana, de forma geral, foram historicamente focados em aumento do perfilhamento e do teor de sacarose, resistência à seca e, principalmente, resistência a doenças (SCARPARI & BEAUCLAIR, 2008). O programa histórico de Java, por exemplo, tinha como foco o controle de Sereh - doença plástica de impacto muito grande no crescimento, desenvolvimento e perfilhamento das plantas afetadas - que causou perdas significativas nas plantações até o início do século XX (JESWIT, 1930).

Referente às particularidades reprodutivas relacionadas ao melhoramento, a gramínea possui como inflorescência uma panícula com flores monoclinas, cujo grão-de-pólen é muito pequeno e de meia-vida bastante reduzida, criando dificuldades para a fecundação (DILLEWIJIN, 1952). De acordo com Bull *et al.* (1992), outros inconvenientes do melhoramento da cultura são os fatos de não se conseguir formar grupos heteróticos verdadeiros, a recorrente carga genética quando há endogamia, a aparente dificuldade de produzir homozigose pela alta ploidia e os problemas relacionados à sua natureza aneuploide, além das questões do alto custo de propágulos e da vulnerabilidade genética dos clones.

Como descrito, a cana-de-açúcar é uma cultura cujo melhoramento depende da sua forma de reprodução assexuada, diferindo-se de programas convencionais de espécies autógamas – em que há ocorrência de autofecundação em mais de 90% dos eventos reprodutivos – e alógamas – com ocorrência de fecundação cruzada em mais de 90% das fertilizações (BOS & CALIGARI, 1995). Com efeito, o processo concentra-se muito mais intensamente no processo de seleção do que na abrangência de cruzamentos, em que um único evento de recombinação pode sofrer seleção por dez anos para originar um cultivar.

Em linhas gerais, um programa de melhoramento de cana possui as etapas de seleção de genitores superiores, cruzamentos (ampliação de variabilidade), seleção de indivíduos superiores na progênie e clonagem. No que se refere aos cruzamentos, é mais comum a adoção de cruzamentos biparentais na maioria dos programas, embora policruzamentos também sejam presentes. Não se realizam teste de progênie estritamente, mas análises de *score* para que se proceda a seleção nas fases iniciais, intermediárias e finais, chegando-se a um cultivar. Nas primeiras, há o uso de milhares de genótipos, sem repetição, com teste em apenas um local e com seleção de caracteres de alta herdabilidade; nas seguintes adota-se o uso de centenas de genótipos, com repetição e teste em apenas um local e com seleção de caracteres de média herdabilidade; e nas finais há o uso de dezenas de genótipos, com repetição e teste em vários locais - VCU (valor de cultivo e uso) e seleção de caracteres de

baixa herdabilidade. A clonagem se dá através da distribuição de toletes no campo (BULL *et al.*, 1992).

2.3 Aspectos genômicos da cana-de-açúcar

Trabalhos relacionando o gênero *Saccharum* com outras espécies da tribo Andropogoneae, como milho (*Zea mays* L.) e arroz (*Oryza sativa* L.), foram desenvolvidos com o objetivo de mapear regiões conservadas entre as diferentes espécies. Foi relevante a pesquisa de Jannoo *et al*. (2007), que estabeleceu uma comparação de ortólogos em uma região rica em genes entre diferentes espécies da tribo. Mais detalhados, porém, são estudos referentes a plantas da subtribo Saccharinae, a qual inclui o gênero *Sorghum*, apresentando o último elevada sintenia de genes e blocos marcadores com *Saccharum* (GRIVET et al., 1994). Enfatiza-se o fato de estar já disponível um genoma de referência da cultura do sorgo devido à sua importância econômica (PATERSON *et al.* 2009). Nessa perspectiva, a comparação entre dados genômicos da cana-de-açúcar e do sorgo indicou alta microcolinearidade entre seus genomas, o que o sugere como uma espécie proximamente aparentada com a cultura em estudo, cuja divergência de um ancestral comum deu-se há cerca de oito milhões de anos (WANG *et al.*, 2010).

Wang *et al.* (2010) esclarecem que há uma diferença, no que se refere aos números característicos de cromossomos e seus *fingerprints* de DNA distintivos, entre dois principais grupos de estudo – um para *S. spontaneum* e outro para todas as demais espécies, com destaque para *S. officinarum*. Daniels & Roach (1987) admitem, ainda, que *S. officinarum* seja originária de hibridações de *S. spontaneum, S. robustum, Miscanthus spp.* e *Erianthus arundinaceus* (Retz.) Jesw.

Considerando as particularidades genômicas, as variedades comerciais são compostas por hibridações envolvendo as espécies *S. officinarum* (a qual contribuiu com 70 a 80% dos cromossomos das canas estudadas) e *S. spontaneum* (de 10 a 20% dos cromossomos), com contribuição minoritária de genomas como *S. robustum*, *S. sinensis* e *S. barberi*. Como consequência da prática de melhoramento, os cultivares modernos de cana possuem ploidia superior a dez, com aproximadamente 120 cromossomos e genoma de tamanho de cerca de 10 Gpb. É necessário enfatizar o fato de formas aneuploides terem contribuído para o genoma da cana moderna, além de esse ter sofrido pelo menos duas rodadas de eventos de duplicação desde sua divergência de um ancestral comum com o sorgo (JANOO *et al.*, 2007).

Pelas características apontadas anteriormente, observa-se que a planta possui notável complexidade para o desenvolvimento de estudos genômicos. Fatores como o comportamento diferenciado dos cromossomos durante a meiose entre as duas principais espécies, grande ação de transposições no genoma do híbrido e as duplicações citadas terem ocorrido, provavelmente, após a especiação entre *S. robustum* e *S. spontaneum* ampliam a dificuldade de estabelecimento de programas investigativos no genoma da cultura (JANOO *et al*., 2007; WANG *et al*., 2010). Frente a isso, prima-se para a existência de trabalhos que fazem uso de recursos disponíveis para outras espécies proximamente aparentadas, como é o caso já citado do sorgo.

Coletivamente, as complicações decorrentes da complexidade genômica da cana-de-açúcar dificultam a aplicação de técnicas modernas voltadas ao melhoramento da espécie. Felizmente, novas metodologias têm sido desenvolvidas para contornar tal questão, como, por exemplo, o método de genotipagem de polimorfismo de nucleotídeo único (*single nucleotide polymorphism* ou SNP) em poliploides, o qual permite uma caracterização detalhada do genoma da cana e de outros poliploides complexos (Serang *et al.*, 2012; Garcia *et al.*, 2013). A partir das informações obtidas por esta abordagem, novas pesquisas podem passar a ter maior suporte no gênero *Saccharum*, transitando por temas como mapeamento associativo, mapeamento de locos que controlam características quantitativas (*quantitative trait loci* ou QTLs), investigação de assinaturas de seleção, caracterização funcional de variantes, dentre outros.

2.4 Genotipagem-por-sequenciamento

2.4.1. Caracterização e análise da variabilidade

Dentro da perspectiva de alternativas de genotipagem, uma tecnologia em franca ascensão é a genotipagem-por-sequenciamento (*genotyping-by-sequencing* ou GBS). Elshire *et al*. (2011) ressaltam o fato de essa técnica ser viabilizada por recentes avanços na obtenção de dados de sequenciamento de nova geração (*next generation sequencing* ou NGS). Tal metodologia tem aplicações promissoras que incluem estudos de associação genômica (GWAS), seleção em germoplasma de espécies sem a necessidade prévia de desenvolvimento de ferramentas moleculares e determinação de estruturas populacionais sem o conhecimento *a priori* dos genomas ou da diversidade existente na espécie. Dessa forma, a técnica de GBS mostra-se como uma viável alternativa também para estudos envolvendo poliploides

complexos, como a cana-de-açúcar, destacando-se suas notáveis vantagens como a maior facilidade de manuseio das amostras, o custo inferior ao de outras técnicas de genotipagem, o acesso a regiões potencialmente reguladoras da expressão gênica e a não necessidade de haver um genoma de referência para o alinhamento das leituras.

É importante destacar que, de acordo com DePristo *et al.* (2011), tais avanços nas tecnologias de sequenciamento possibilitaram o manejo de informações valiosas a respeito da variabilidade genética em amostras populacionais, permitindo compreensões abrangentes sobre ancestralidade e fenômenos evolutivos. Nesse sentido, o uso de recursos de NGS tem sido bastante relevante para estudos populacionais. Contudo, nota-se que a GBS possui uma cobertura enviesada do genoma, uma vez que há um maior foco nas regiões hipometiladas, que em geral são ricas em genes. Além disso, essa técnica apresenta dificuldades adicionais em poliploides, principalmente no tocante à determinação de dosagens alélicas. Os principais fatores a considerar são a profundidade de cobertura e a dificuldade de distinção entre variações alélicas e locos homeólogos, especialmente para um caso complexo como o da cana.

Como observa Gibson (2006), a análise da variabilidade da sequência de DNA é de importância central em estudos genéticos de diferentes naturezas, de forma que marcadores moleculares foram criados para acessar tal variabilidade, tendo melhorado significativamente a análise genética. No tocante aos SNPs, pode-se dizer que esses potencialmente constituem o tipo mais básico de variação genética, por revelar polimorfismo diretamente em nível de nucleotídeos. Como incentivo de seu uso em relação aos comuns microssatélites, por exemplo, pode-se elencar o fato de apresentarem abundância no genoma, passividade de automação em larga escala, alta adaptabilidade a diversas tecnologias e menor necessidade de manipulação laboratorial (HARA *et al.*, 2010; WELLER *et al.*, 2010). Por fim, Dillon *et al.* (2010) apontam que a elevada cobertura de SNPs aumenta, teoricamente, a possibilidade de se mapear proximidades de regiões reguladoras, facilitando a associação entre genótipo e fenótipo.

Um trabalho interessante aliando NGS com estudo de diversidade populacional foi o de Gheyas *et al.* (2015). Nesse, fez-se uma anotação funcional e posicional de sítios variantes, com base nos modelos gênicos anotados no genoma da galinha doméstica (*Gallus gallus* L.), além de uma análise de aspectos da diversidade em aminoácidos e alterações nas estruturas secundárias de RNAs mensageiros (mRNAs) e RNAs não-codificantes (ncRNAs). Esses autores buscaram também evidências de fenômenos seletivos ao longo do genoma, utilizando parâmetros de genética de populações relacionados ao comportamento de indicadores de

seleção, com destaque para a quantidade relativa de variabilidade, em regiões de interesse e sua predição funcional. A partir de tal trabalho, é latente a aliança entre tecnologias de nova geração e estudos de diversidade genômica, possibilitando-os mesmo em organismos de genoma complexo, como é o caso da cana-de-açúcar.

2.4.2 Separação das leituras por *barcode* e detecção de SNPs

Os *barcodes* podem ser definidos como pequenas sequências de DNA que identificam, singularmente, cada amostra. Dessa forma, o uso desses identificadores em aplicações de NGS permite que um amplo conjunto de amostras seja analisado simultaneamente em uma única plataforma de sequenciamento (DAVEY *et al*., 2011).

Utilizando-se procedimentos criteriosos, como diferença de tamanho, não sobreposição de sequências e ausência de sequências repetitivas nos *barcodes*, essa técnica pode ser aplicada de forma eficiente em GBS. Sua função na genotipagem é relevante pelo uso de bibliotecas reduzidas ser aliado ao potencial de relacionar indivíduos com suas sequências de forma sistemática (ELSHIRE, 2011).

Após o sequencimento, os usos da base de dados são diversos, tendo destaque a identificação de variantes para estudos populacionais. Um *software* de uso comum para identificação de sítios variantes a partir de dados de resequenciamento é o FreeBayes (GARRISON & MARTH, 2012) – um método Bayesiano que encontra polimorfismos menores que o comprimento do alinhamento de uma sequência de leitura curta, como SNPs, *indels* e eventos similares, sendo capaz de modelar locos multialélicos em conjuntos de indivíduos. Seu funcionamento é baseado em haplótipos, detectando os sítios variantes de acordo com sequências de leituras alinhadas a um alvo particular. Esse fato evita o problema comum de haver múltiplos alinhamentos possíveis para sequências idênticas. Por fim, seu uso depende de um genoma de referência em formato FASTA, determinando a combinação de máxima verossimilhança dos genótipos para a população analisada em cada posição da referência. Seu resultado é um conjunto de posições em que se encontram prováveis polimorfismos em formato VCF (*variant call file format*), que, segundo Danecek *et al.* (2011), é um arquivo de texto utilizado em bioinformática para armazenar variações de sequências genômicas.

2.4.3 Genotipagem quantitativa de poliploides

Como dito, um problema na genotipagem de poliploides é a quantificação das dosagens alélicas. Tendo em vista que um autopoliploide possui mais de duas cópias de uma mesma região genômica, pode-se inferir que não há apenas uma dicotomia entre as situações homo e heterozigota como em um diploide, mas existem diversas situações intermediárias com distintos níveis de heterozigosidade. Em um indivíduo tetraploide, por exemplo, entre os genótipos totalmente homozigotos para os alelos *A* (denotado por *AAAA*) e *a* (denotado por *aaaa*), há heterozigotos com uma cópia do alelo *a* (*AAAa*), duas (*AAaa*) e três cópias (*Aaaa*). Raciocínio similar é válido para as ploidias maiores.

Uma metodologia recente para genotipagem nessa situação é o uso do programa computacional SuperMASSA, como apontam Serang *et al.* (2012). Para inferir a ploidia de melhor ajuste aos dados para cada loco consideram-se simultaneamente as intensidades associadas com os diferentes alelos, as frequências esperadas de indivíduos e o erro experimental. Além da estimação da ploidia, busca-se um modelo que ajuste as dosagens alélicas com a complexa estrutura do genoma da cana através da maximização da probabilidade *a posteriori*.

No que se refere ao *software*, trata-se de um programa que realiza estatística Bayesiana, sendo sua principal vantagem o fato de ploidia e dosagem serem determinadas em conjunto para uma população através das intensidades relativas de cada alelo. Assim, é muito útil quando, na obtenção de intensidades alélicas relativas semelhantes, não se tem ideia sobre a ploidia do indivíduo (por exemplo, uma proporção 3:1 não fornece a ideia se esse indivíduo é tetraploide, *AAAa*, ou octaploide, *AAAAAAaa*).

2.5 Anotação funcional

2.5.1 Classes funcionais

Antes de discorrer sobre os papéis evolutivos dos polimorfismos (sejam eles *indels*, SNPs ou das demais categorias), atribui-se alguma tipologia a respeito das suas possíveis consequências moleculares. Nesse sentido, na literatura encontra-se uma série de classificações sobre os efeitos das modificações. As principais são relacionadas às consequências transcricionais ou traducionais, além de serem relevantes, em paralelo, as anotações por região genômica.

Como pontua Ernest (1959a, 1959b), um SNP pode ser classificado, quanto à natureza da substituição, como transição, quando tal substituição conserva a categoria química da base nitrogenada (o nucleotídeo continha uma purina e continuou com uma, apesar da troca, por exemplo), ou como transversão, quando a categoria química é alterada (o nucleotídeo continha uma purina e passa a possuir uma pirimidina, por exemplo). Segundo Maquat (2001), no que se refere às consequências dessa substituição, um SNP é denominado sinônimo ou silencioso quando o códon alterado traduz o mesmo aminoácido e não-sinônimo ou de sentido trocado quando há alteração do aminoácido codificado. Sobre esse último tipo, pode-se denominar o SNP de conservativo, situação em que a alteração é de pouco efeito predito na estrutura da proteína, ou não-conservativo, quando a estrutura proteica muda substancialmente. Ainda se classifica como SNP sem sentido (*nonsense*) aquele que altera um códon tradutor de aminoácido para um códon de fim (terminação), encerrando a tradução dessa sequência a partir desse ponto.

No tocante às regiões de ocorrência de SNPs, são relevantes as separações de regiões gênicas e intergênicas, de éxon (região que permanece no RNA mensageiro maduro após seu processamento) e íntron (porção excisada da sequência após o processamento do transcrito primário) e de regiões regulatórias não transcritas (*untranslated regions* – UTRs) e regiões traduzidas (MAQUAT, 2001). Pode-se, ainda, elencar as classificações de códon de iniciação, códon de terminação, sítio doador ou receptor de *splicing* e região a montante (*upstream*) ou a jusante (*downstream*) de uma região gênica.

Nesse âmbito, um programa disponível para a realização de anotação funcional de locos variantes é o SnpEff. Esse *software* confronta a posição em que cada sítio está situado com anotações pré-existentes de uma sequência de referência (CINGOLANI *et al.*, 2012). Entre as principais classes atribuídas por este programa, destacam-se:

- Variação intergênica, *upstream* (5 Kpb) ou *downstream* (5 Kpb) de uma região gênica;
- Variação na região 5’-UTR ou 3’-UTR;
- Ganho ou perda de códon de iniciação, códon de terminação, sítio doador ou receptor de *splicing*;
- Variação em região de íntron ou éxon;
- Se em região de éxon, se a mutação é sinônima ou não sinônima, se há alteração de quadro de leitura, ganho ou perda de códons.

Dentre as atribuições realizadas por esse programa, é também relevante a predição de efeitos alto, moderado, baixo e modificador. Cingolani *et al.* (2016) definem como de alto

efeito as mutações que provocam a excisão do transcrito, perda ou ganho de códon de terminação, ganho de códon de iniciação, alteração em sítio doador ou receptor de *splicing*, alteração para um códon que traduz um aminoácido raro, perda de éxon e variação no quadro de leitura ou no número de cromossomos. Moderados são os efeitos que produzem o truncamento de UTR com consequente perda de éxon, variação de baixo impacto na sequência codante, deleções ou inserções sem maiores consequências, simples alteração de aminoácido (mutação de sentido trocado), excisão de região regulatória, perda de região de *splicing* e excisão de regiões de ligação de fatores de transcrição.

Alterações de baixo efeito são descritas para os casos de mutação silenciosa, mutação que retém o códon de iniciação ou de terminação, modificações com baixo efeito na região de *splicing* e criação prematura de 5'-UTR. Por fim, as mutações de efeito modificador abarcam um amplo espectro de condições, incluindo variações genéricas em regiões não traduzidas que não alteram propriamente a proteína, como alterações em íntrons e na transcrição de micro-RNA (miRNA) (CINGOLANI *et al.*, 2016). É importante destacar que nem todos os tipos de variações, portanto, referem-se a SNPs e *indels* pontuais, já que algumas das classes correspondem a alterações em maior escala.

Adicionalmente, é relevante expandir as tipologias para um horizonte macrofuncional, que alie as consequências pontuais das alterações à morfologia ou ao metabolismo do indivíduo. Análises de enriquecimento funcional podem ter o papel de entender o ganho ou a perda de determinada função frente ao agrupamento de uma série de mutações de alguma classificação comum. Segundo Subramanian *et al.* (2005), esse tipo de análise é especialmente utilizado objetivando a reunião de grupos com função, localização cromossômica ou regulação semelhantes. Assim, é possível estabelecer relação entre as alterações genômicas e as possíveis consequências morfofisiológicas, ou seja, entre o genótipo e o fenótipo.

2.5.2 Anotação ontológica

Ontologia pode ser definida como um modelo de dados que trata de objetos inseridos em um domínio e sua relação (GUARINO, 1998). Em genética, significa agrupar genes sob alguma ótica funcional comum dentro de uma hierarquia.

Uma base de dados que define e caracteriza anotações gênicas ontológicas é a *Gene Ontology* (GENE ONTOLOGY CONSORTIUM, 2004), reunindo uma série de termos hierarquizados distribuídos em três grandes esferas. Tais instâncias são componente celular

(com genes que participam da constituição anatômica da célula, incluindo a maquinaria de expressão gênica), função molecular (com genes que participam de atividades que ocorrem em nível molecular, como atividades ligante, transportadora e catalítica) e processo biológico (com genes que participam de processos amplos, com mais de uma etapa funcional distinta).

2.6 Domesticação e variabilidade genética

2.6.1 Domesticação

Domesticação é definida como a arte ou prática, consciente ou inconsciente, de transformar uma espécie que não apresenta características de fácil cultivo ou criação (condição selvagem) para uma condição de amplos propagação, domínio de manejo e aproveitamento pelo homem (ZOHARY *et al.*, 2012). Esse fenômeno tem como resultado a presença de síndromes, sendo relevantes em plantas, como demonstra Harlan (1975), o aumento do tamanho e da coloração das sementes, o crescimento determinado do caule, a redução ou a eliminação de dormência das sementes e de compostos tóxicos nas partes comestíveis, a precocidade, a menor sensibilidade ao fotoperíodo e a presença de vagens não fibrosas e indeiscentes.

Segundo Zohary *et al.* (2012), os principais precursores das síndromes de seleção (ou seja, os meios de pressão seletiva durante a domesticação) são a competição entre plântulas na lavoura, a colheita e a escolha deliberada de características favoráveis ao cultivo e/ou ao consumo pelo homem. Na cana-de-açúcar, de acordo com McKey *et al.* (2010), a domesticação provavelmente resultou em características como o aumento do teor de açúcar e do número de perfilhos e crescimento determinado.

Por fim, além do processo de domesticação, as culturas continuam a evoluir a partir das seleções natural e artificial. Nesse último caso, enfatiza-se as diferenças de foco em programas de melhoramento genético, o que é ampliado quando o uso é diverso entre regiões geográficas. Como pontua Harlan (1975), o homem criou uma relação com os animais e as plantas domesticados, cuja sobrevivência passou a depender do primeiro. A domesticação, portanto, modificou as relações das culturas com o ambiente e, mais especificamente, a variação das frequências fenotípica e genotípica, com frequente diminuição da variabilidade alélica.

2.6.2 Variabilidade genômica

A variação populacional foi explicada basicamente por três escolas de pensamento evolutivo. A escola clássica, representada principalmente por Thomas Hunt Morgan e Hermann Müller, foi muito influenciada pelo mendelismo e pelo uso de organismos-modelo em laboratório. Advogava que havia baixa variação na natureza, uma vez que as mutações, ao surgirem, eram rapidamente selecionadas e aumentavam sua frequência ou eram rapidamente eliminadas. Definia-se, com base no ambiente médio, uma variedade selvagem e outra mutante (MAYER, 1998).

Ainda segundo Mayer (1998), a escola do polimorfismo balanceado, por sua vez, representada especialmente por Alfred Sturtevant e Theodosius Dobzhansky, apoiou-se muito na observação de populações naturais. Pontuava que havia alta variação na natureza, o que se devia à não extinção de polimorfismos por questões de adaptabilidade temporal ou espacial e pela ocorrência de heterose.

Testes de sequenciamento de aminoácidos e de eletroforese de proteínas na década de 1960 questionaram, respectivamente, as quantidades de variação inter e intraespecífica. Kimura (1968) propôs o que se conheceria posteriormente como escola neutralista. A partir de tal interpretação, o processo evolutivo em nível molecular era especialmente influenciado pela deriva genética, ou seja, pelo acaso, em oposição ao que posteriormente denominou-se de *selecionismo*, que compreendia que a evolução molecular era conduzida pela seleção de mutações vantajosas, reduzindo a importância da deriva (NEI, 2005).

O *neutralismo* de Kimura (1968) teve como pontos principais o reconhecimento de uma maior frequência de mutações deletérias em relação ao total (como também compreendia o *selecionismo*), mas com um papel significativo das mutações neutras e muito pequeno das mutações positivas, e da existência de um relógio molecular. Também foi entendido que o papel da seleção era extremamente reduzido, sendo pronunciado basicamente na evolução adaptativa.

A compreensão radical da teoria neutralista (teoria totalmente neutra) sofreu críticas inúmeras da comunidade científica. Apontou-se que sob uma seleção branda (*soft selection*), a mortalidade causada pela seleção incorporava-se à mortalidade ecológica, o que enfraquecia o desprezo pela seleção por parte do *neutralismo*. Ainda, medidas do relógio molecular observaram discrepâncias nas taxas de mutação, apontando para uma força inferior da deriva genética na evolução molecular do que a imaginada (NEI, 2005). O descompasso do relógio molecular em relação às mutações sinônimas e não-sinônimas foi então explicado pela teoria

aproximadamente neutra, proposta por Ohta (1973), que considerou que o valor adaptativo (*fitness*) era uma função do tamanho populacional, o que explicaria o fato de populações pequenas e grandes possuírem taxas de variação semelhantes.

Ohta (1973) argumentou que existiam mutações aproximadamente neutras, que se comportavam como neutras em populações pequenas, fixando-se sucessivamente, e como deletérias em populações grandes, diminuindo sua frequência. Tal mensuração foi feita em relação ao coeficiente de seleção (baseado no *fitness*), que, relacionado ao tamanho populacional, determina o comportamento da mutação (neutra, aproximadamente neutra ou não neutra).

2.6.3 Estimativas genômicas populacionais

As consequências da teoria aproximadamente neutra foram diversas, destacando-se a conclusão de que as variações sinônimas são mais frequentes, embora as não-sinônimas possuam maior taxa de variação. Assim, pode-se aferir sobre a influência da seleção em uma população a partir da razão de diferenças não-sinônimas (n) e sinônimas (s), denominada $\omega = {}^{n}/_{S}$. Como pontua Nielsen (2001), trabalha-se com três situações desse quociente:

- $\omega < 1$, que indica a presença de seleção negativa (purificadora);
- $\omega = 1$, que indica seleção relaxada ou a não influência da seleção;
- $\omega > 1$, que indica a ação da seleção positiva (direcional).

Expandindo-se tais consequências, McDonald & Kreitman (1991) demonstraram que a relação era igual entre e dentro de espécies com a ação da deriva genética, mas tal relação diferia com a ação da seleção. Isso significa que é possível, a partir da aplicação de um teste comparativo entre os valores de ω intra e interespecífico, observar em que genes há indício de seleção e qual a sua direção. Como resultado, o pesquisador obtém informações a respeito da história evolutiva que ocorreu desde a especiação em questão e obtém indícios de processos de pressão seletiva. Com efeito, essas evidências podem ser utilizadas pelo geneticista em estudos ecológicos e de biologia evolutiva. Também há aplicações no melhoramento genético através de possíveis ações de seleção assistida. Trabalhos nesse sentido são o de Walker (2002), com seu uso em tamanho populacional efetivo, além de outros exemplos, como de Charlesworth & Walker (2008), que aborda o efeito das mutações deletérias e Messer & Petrov (2013), sobre adaptação.

Sobre o teste, comparando-se as duas espécies, pode-se aferir como diferenças interespecíficas as variações entre elas que foram fixadas na espécie de interesse (denominadas D - portanto, Dn e Ds) e como diferenças intraespecíficas as variações polimórficas em tal espécie (denominadas P - portanto, Pn e Ps). De forma a criar uma relação entre as variações polimórficas e as variações fixadas, tem-se:

$$NI \ = \omega_P/\omega_D = \frac{Ds.Pn}{Ps.Dn}$$

Denomina-se NI o índice de neutralidade. Considerando que mutações silenciosas (ou sinônimas) são neutras, se NI for igual a um, não há diferença entre as taxas inter e intraespecífica, não havendo evidência de seleção. Se NI for maior do que um, há acúmulo de aminoácidos polimórficos, indicando seleção negativa (purificadora). Por fim, se *NI* for inferior a um, há acúmulo de aminoácidos fixados, indicando seleção positiva (direcional). Define-se também o parâmetro $\alpha = 1 - NI$, que indica a proporção de substituições dirigidas pela seleção positiva.

Outros parâmetros podem ser considerados para a obtenção de estimativas relacionadas à variabilidade genômica. É considerada de interesse a diversidade nucleotídica (π) (OLEKSYK et al., 2010), obtida pela seguinte expressão:

$$\pi = \sum_{i=1}^{k} \frac{2j_i(n_i - j_i)}{n_i(n_i - 1)}$$

em que k representa o número de posições variáveis em uma janela de tamanho pré-determinado (por exemplo, 1000 pb); j_i indica o número de ocorrências do alelo alternativo para o i-ésimo SNP ou *indel*; n_i representa o número total de sequências, isto é, a soma das contagens dos alelos de referência e alternativo para o SNP ou *indel*.

Além dois tipos de estudo, é possível observar assinaturas de seleção a partir da comparação entre regiões genômicas. Tais assinaturas podem ser verificadas pela presença de menor variação em algumas regiões em comparação a outras. Isso se deve à seleção de caracteres, que, segregando juntamente a regiões vizinhas de evolução neutra por ação da deriva genética, acabam por apresentar evidência de perda de variabilidade alélica. Esse desequilíbrio de ligação gera uma varredura na região genômica ao longo do tempo, culminando em uma limpeza seletiva (*selective sweep*) (NIELSEN, 2005).

A importância da procura dessas assinaturas relaciona-se a estudos genômicos da evolução de uma espécie ou população, a investigação da função de tais regiões – principalmente quando o genoma não está totalmente descrito – e a marcação molecular para diversos estudos genômicos (PRZEWORSKI *et al.*, 2005). Entre estes, estão inclusos o mapeamento genético e o desenvolvimento de programas de melhoramento. Sim (2011) e Zhang (2012) discutem sua importância em estimativas de estrutura populacional e em processos de seleção, respectivamente. Com efeito, um programa de melhoramento genético de cana-de-açúcar poderia utilizar tais informações para definição de cruzamentos, considerando a dificuldade já citada de se formar grupos heteróticos verdadeiros, e a marcação de regiões de interesse para processos de seleção.

Um parâmetro para aferição de assinaturas de seleção é dado por Rubin *et al.* (2010), denominado heterozigosidade combinada (*pooled heterozigosity*, ou Hp), que utiliza as frequências dos alelos de todos os sítios variantes dentro de uma região genômica. Seu cálculo é obtido por:

$$Hp = \frac{2 \sum nMAJ \sum nMIN}{(\sum nMAJ + \sum nMIN)^2},$$

em que $nMAJ$ e $nMIN$ representam as contagens dos alelos mais frequente e menos frequente presentes na janela genômica em análise, respectivamente.

Dentro dessa perspectiva de análise de variabilidade populacional, pouco se sabe, ainda, a respeito da variação genética existente entre materiais comumente utilizados em programas de melhoramento de cana-de-açúcar em nível de genoma. Adentram a questão a quantificação da variabilidade de SNPs e identificação de suas regiões de incidência, além da caracterização funcional e da avaliação da diversidade alélica existente ao longo do genoma da espécie.

O uso do teste de McDonald & Kreitman mostra-se oportuno primeiramente pela existência de um genoma de referência do sorgo, uma espécie proximamente aparentada à cana. Além disso, ele possibilita análises de evidências seletivas em genes, um tipo de estudo inédito para a cultura canavieira. Em relação à heterozigosidade combinada, trabalhos como o de Gheyas *et al.* (2015) demonstram sua importância em análises de cunho funcional e evolutivo. Sua importância ocorre nesse caso por possibilitar a busca por assinaturas de seleção com base em todos os polimorfismos, localizados em regiões gênicas ou não.

3 OBJETIVOS

3.1. Objetivo geral

Esse trabalho objetivou caracterizar funcionalmente a variabilidade genética entre genótipos de cana-de-açúcar, avaliada através de genotipagem-por-sequenciamento. Mais estritamente, realizou-se uma análise funcional e posicional de SNPs e outros tipos de variações que possam ter ocorrido durante a evolução dessa gramínea, inclusive possivelmente durante sua domesticação, e aferir sobre suas possíveis consequências morfofisiológicas.

3.2. Objetivos específicos

De forma particularizada, esse estudo teve como objetivos (1) testar uma caracterização funcional de polimorfismos identificados em genótipos de um painel populacional de um poliploide complexo genotipado pela metodologia de genotipagem-por-sequenciamento; (2) predizer os efeitos moleculares desses polimorfismos de um poliploide de melhoramento recente como a cana-de-açúcar; (3) avaliar a direção de seleção em todo o genoma dos indivíduos da população ao longo da evolução da gramínea, incluindo a domesticação; (4) encontrar assinaturas de seleção através da identificação de regiões genômicas de reduzida variabilidade com a aplicação da estatística heterozigosidade combinada (Hp); e (5) identificar grupos funcionais sob pressão positiva de seleção pela aplicação do Teste de McDonald & Kreitman. As hipóteses motivadoras deste trabalho foram (1) a perda de variabilidade genética nos genótipos melhorados em relação aos ancestrais; (2) a existência de evidência de seleção purificadora em regiões gênicas; (3) e a evidência de seleção em regiões e genes potencialmente envolvidos com processos importantes à planta. A partir dessas evidências, espera-se ampliar o conhecimento do genoma desse poliploide, fornecendo subsídios a futuros estudos ecológicos e, mais notoriamente, de melhoramento genético da cultura.

4 MATERIAL E MÉTODOS

4.1 Materiais vegetais e sequenciamento

4.1.1 Genótipos

Os indivíduos de *Saccharum spp.* utilizados neste trabalho são integrantes do Painel Brasileiro de Genótipos de Cana-de-Açúcar (PBGCA), composto por 266 genótipos. Esse painel é mantido em campo no Centro de Ciências Agrárias da Universidade Federal de São Carlos, em Araras, estado de São Paulo, que integra a RIDESA. Sua descrição fenotípica e molecular está presente no trabalho de Barreto *et al.* (2016).

Separou-se os genótipos em duas classes contrastantes no quesito de melhoramento genético: genótipos ancestrais e genótipos melhorados. Fazem parte do primeiro grupo os seguintes indivíduos: representantes de *S. robustum* (nove genótipos), *S. sinensis* (quatro), *S. barberi* (cinco), *S. officinarum* (51), *S. spontaneum* (cinco), *S. edule* (um) e *S. bengalensis* (um), híbridos de *S. officinarum* com *S. spontaneum* (três), de *S. officinarum* com *S. robustum* (dois), de *S. officinarum* com *S. barberi* (três) e de *S. spontaneum* com *S. bengalensis* (um), além de híbridos interespecíficos de *Saccharum* spp. (dois), *Erianthus* spp. (seis) e um híbrido de *Erianthus* spp. com *Miscanthus* spp., totalizando 93 genótipos (Tabela 2). Do segundo grupo, por sua vez, são constituintes os programas de melhoramento genético RB (83), SP (33) e IAC (14), além de 43 cultivares originários de outros 15 programas, apresentando 173 genótipos no total (Tabela 3).

TABELA 2 - Relação de genótipos ancestrais presentes no Painel Brasileiro de Genótipos de Cana-de-Açúcar. Cada genótipo está classificado de acordo com a espécie à qual pertence, quando conhecida.

Genótipo	Discriminação	Genótipo	Discriminação
28NG289	*S. robustum*	HJ5741	*S. officinarum*
57NG12	*S. robustum*	H. Kawandang	*Erianthus* spp.
75//09 Erianthus	*Erianthus* spp.	IJ76-293	*S. robustum*
Agau	*S. sinensis*	IJ76-313	*S. officinarum*
Agoule	*S. barberi*	IJ76-314	*S. robustum x S. officinarum*
Ajax	*S. officinarum*	IJ76-317	*S. officinarum*
Akbar	*S. edule*	IJ76-318	*S. robustum*
Ar Chi	*S. sinensis*	IJ76-325	*S. officinarum*
Arundoid B	*Saccharum* spp.	IJ76-326	*S. officinarum*
Badila de Java	*S. officinarum*	IJ76-360	*S. officinarum*
Black Borneo	*S. officinarum*	IJ76-418 Red	*S. officinarum*
Caiana Fita	*S. officinarum*	IJ76-560	*S. officinarum*
Caiana Listrada	*S. officinarum*	IM76-227	*Erianthus* spp.
Caiana Riscada	*S. officinarum*	IM76-228	*S. robustum*
Caiana Verdadeira	*S. officinarum*	IM76-229	*S. robustum*
Cana Alho	*S. officinarum*	IN84-103	*S. officinarum*
Cana Blanca	*S. officinarum*	IN84-104	*S. robustum*
Cana Manteiga	*S. officinarum*	IN84-105	*S. officinarum*
Cayana	*S. officinarum*	IN84-106	*S. officinarum*
Ceran Red	*S. officinarum*	IN84-117	*S. robustum*
Chin	*S. barberi*	IN84-46	*S. officinarum*
China	*S. officinarum*	IN84-58	*S. spontaneum*
Chunnee	*S. barberi*	IN84-73	*Erianthus* spp.
Creoula	*S. officinarum x S. barberi*	IN84-77	*Erianthus* spp.
Criolla Morada	*S. officinarum x S. barberi*	IN84-82	*S. spontaneum*
Criolla Rayada	*S. officinarum x S. barberi*	IN84-83	*Erianthus* spp.
Cristalina	*S. officinarum*	IN84-88	*S. spontaneum*
D11/35	*S. officinarum*	IS76-155	*S. officinarum*
D152	*S. officinarum*	Kavangira	*S. robustum x S. officinarum*
D625	*S. officinarum*	Krakatau	*S. spontaneum*
EK28	*S. officinarum*	Laukona	*S. officinarum*
F76-1762	*Miscanthus* spp. *x Erianthus* spp.	Louser	*S. officinarum*
Flor de Cuba	*S. officinarum*	Mana II	*S. officinarum*
Formosa	*S. officinarum*	Maneria	*S. sinensis*
Ganda Cheni	*S. barberi*	Muntok Java	*S. officinarum x S. spontaneum*
Green German	*S. officinarum*	MZ-151	*S. officinarum*

Genótipo	Discriminação	Genótipo	Discriminação
NG21-17	*S. officinarum*	Sac.Offic. 284	*S. officinarum*
NG21-21	*S. officinarum*	SES 205 A	*S. spontaneum*
NG57-221	*S. officinarum*	Uba Demerara	*S. sinensis*
NG57-50	*S. officinarum x S. spontaneum*	US57-141-5	*S. robustum*
NG57-6	*S. officinarum*	US60-31-3	*S. bengalensis*
NG77-18	*S. officinarum*	US85-1008	*S. spontaneum x S. bengalensis*
Ragnar	*S. officinarum x S. spontaneum*	White Mauritius	*S. officinarum*
Sabura	*S. officinarum*	White Pararia	*S. barberi*
Sac.Offic. 8272	*S. officinarum*	White Transparent	*S. officinarum*
Sac.Offic. 8276	*S. officinarum*	Zwart Manila	*S. officinarum*
Sac.Offic. 8280	*S. officinarum*		

TABELA 3 - Relação de genótipos melhorados presentes no Painel Brasileiro de Genótipos de Cana-de-Açúcar. Cada genótipo está classificado de acordo com o programa ao qual pertence.

Genótipo	Discriminação	Genótipo	Discriminação
CB36-14	Campos Brasil	CP70-1547	Canal Point, EUA
CB36-24	Campos Brasil	CR72/106	Central Romana, República Dominicana
CB36-25	Campos Brasil	F150	Flórida, EUA
CB36-68	Campos Brasil	F31-962	Flórida, EUA
CB40-13	Campos Brasil	F36-819	Flórida, EUA
CB40-77	Campos Brasil	H53-3989	Havaí, EUA
CB41-76	Campos Brasil	H59-1966	Havaí, EUA
CB45-155	Campos Brasil	IAC48-65	Instituto Agronômico de Campinas
CB45-3	Campos Brasil	IAC49-131	Instituto Agronômico de Campinas
CB46-47	Campos Brasil	IAC50-134	Instituto Agronômico de Campinas
CB47-355	Campos Brasil	IAC51-205	Instituto Agronômico de Campinas
CB49-260	Campos Brasil	IAC52-150	Instituto Agronômico de Campinas
CB53-98	Campos Brasil	IAC58-480	Instituto Agronômico de Campinas
CIMCA77-316	Santa Cruz, Bolívia	IAC64-257	Instituto Agronômico de Campinas
CIMCA77-318	Santa Cruz, Bolívia	IAC68-12	Instituto Agronômico de Campinas
Co285	Coimbatore, Índia	IAC82-2045	Instituto Agronômico de Campinas
Co290	Coimbatore, Índia	IAC82-3092	Instituto Agronômico de Campinas
Co331	Coimbatore, Índia	IAC83-4157	Instituto Agronômico de Campinas
Co419	Coimbatore, Índia	IAC86-2210	Instituto Agronômico de Campinas
Co449	Coimbatore, Índia	IAC87-3396	Instituto Agronômico de Campinas
Co740	Coimbatore, Índia	IAC91-1099	Instituto Agronômico de Campinas
Co997	Coimbatore, Índia	L60-14	Louisiana, EUA
CP51-22	Canal Point, EUA	MALI	Fiji
CP52-68	Canal Point, EUA	NA56-79	Norte Argentino
CP53-76	Canal Point, EUA	NCo310	Natal-Coimbatore, África do Sul

Genótipo	Discriminação	Genótipo	Discriminação
POJ161	Proefstation Oest Java, Indonésia	RB845239	Ridesa Brasil
POJ2878	Proefstation Oest Java, Indonésia	RB845257	Ridesa Brasil
Q117	Queensland, Austrália	RB845286	Ridesa Brasil
Q165	Queensland, Austrália	RB855002	Ridesa Brasil
Q70	Queensland, Austrália	RB855035	Ridesa Brasil
R570	Ilhas Reunião	RB855036	Ridesa Brasil
RB721012	Ridesa Brasil	RB855070	Ridesa Brasil
RB72199	Ridesa Brasil	RB855077	Ridesa Brasil
RB72454	Ridesa Brasil	RB855113	Ridesa Brasil
RB725053	Ridesa Brasil	RB855156	Ridesa Brasil
RB725828	Ridesa Brasil	RB855196	Ridesa Brasil
RB732577	Ridesa Brasil	RB855206	Ridesa Brasil
RB735200	Ridesa Brasil	RB855350	Ridesa Brasil
RB735217	Ridesa Brasil	RB855357	Ridesa Brasil
RB735220	Ridesa Brasil	RB855453	Ridesa Brasil
RB735275	Ridesa Brasil	RB855463	Ridesa Brasil
RB736018	Ridesa Brasil	RB855465	Ridesa Brasil
RB739359	Ridesa Brasil	RB855511	Ridesa Brasil
RB739735	Ridesa Brasil	RB855533	Ridesa Brasil
RB75126	Ridesa Brasil	RB855536	Ridesa Brasil
RB765418	Ridesa Brasil	RB855546	Ridesa Brasil
RB765480	Ridesa Brasil	RB855563	Ridesa Brasil
RB785148	Ridesa Brasil	RB855589	Ridesa Brasil
RB785750	Ridesa Brasil	RB855595	Ridesa Brasil
RB805276	Ridesa Brasil	RB855598	Ridesa Brasil
RB806043	Ridesa Brasil	RB865214	Ridesa Brasil
RB815521	Ridesa Brasil	RB867515	Ridesa Brasil
RB815627	Ridesa Brasil	RB925211	Ridesa Brasil
RB815690	Ridesa Brasil	RB925268	Ridesa Brasil
RB825317	Ridesa Brasil	RB925345	Ridesa Brasil
RB825336	Ridesa Brasil	RB92579	Ridesa Brasil
RB825548	Ridesa Brasil	RB935744	Ridesa Brasil
RB83100	Ridesa Brasil	RB965902	Ridesa Brasil
RB83102	Ridesa Brasil	RB965917	Ridesa Brasil
RB83160	Ridesa Brasil	RB966928	Ridesa Brasil
RB835019	Ridesa Brasil	RB975148	Ridesa Brasil
RB835054	Ridesa Brasil	RB975157	Ridesa Brasil
RB835089	Ridesa Brasil	RB975184	Ridesa Brasil
RB835205	Ridesa Brasil	RB975201	Ridesa Brasil
RB835486	Ridesa Brasil	RB975242	Ridesa Brasil
RB835687	Ridesa Brasil	RB975932	Ridesa Brasil
RB845197	Ridesa Brasil	RB975952	Ridesa Brasil
RB845210	Ridesa Brasil	RB975476	Ridesa Brasil

Genótipo	Discriminação	Genótipo	Discriminação
RB002601	Ridesa Brasil	SP79-6134	Copersucar (São Paulo)
RB002700	Ridesa Brasil	SP79-6192	Copersucar (São Paulo)
RB002754	Ridesa Brasil	SP80-1520	Copersucar (São Paulo)
SP70-1005	Copersucar (São Paulo)	SP80-1816	Copersucar (São Paulo)
SP70-1078	Copersucar (São Paulo)	SP80-1836	Copersucar (São Paulo)
SP70-1143	Copersucar (São Paulo)	SP80-1842	Copersucar (São Paulo)
SP70-1284	Copersucar (São Paulo)	SP80-180	Copersucar (São Paulo)
SP70-1423	Copersucar (São Paulo)	SP80-185	Copersucar (São Paulo)
SP70-3370	Copersucar (São Paulo)	SP80-3280	Copersucar (São Paulo)
SP71-1406	Copersucar (São Paulo)	SP80-4966	Copersucar (São Paulo)
SP71-6163	Copersucar (São Paulo)	SP81-1763	Copersucar (São Paulo)
SP71-6949	Copersucar (São Paulo)	SP81-3250	Copersucar (São Paulo)
SP71-799	Copersucar (São Paulo)	SP83-2847	Copersucar (São Paulo)
SP72-4928	Copersucar (São Paulo)	SP83-5073	Copersucar (São Paulo)
SP77-5181	Copersucar (São Paulo)	SP86-155	Copersucar (São Paulo)
SP79-1011	Copersucar (São Paulo)	SP89-1115	Copersucar (São Paulo)
SP79-2233	Copersucar (São Paulo)	SP91-1049	Copersucar (São Paulo)
SP79-2312	Copersucar (São Paulo)	Tu71-7	Tucman, Argentina
SP79-2313	Copersucar (São Paulo)		

4.1.2 Genotipagem-por-sequenciamento

Um subconjunto de genótipos do PBGCA foi previamente genotipado através da plataforma Sequenom iPLEX MassARRAY, como apresentado por Garcia *et al.* (2013). Mais recentemente, foram obtidos dados de genotipagem-por-sequenciamento para todos os indivíduos integrantes do painel, no Institute for Genomic Diversity (IGD), da Universidade de Cornell, nos Estados Unidos da América.

Devido à natureza poliploide da cana-de-açúcar, duas estratégias foram empregadas para aumentar a profundidade da cobertura de sequenciamento. Em primeiro lugar, optou-se por realizar a digestão genômica com a enzima PstI, a qual reconhece a sequência de seis bases CTGCAG e tem corte raro. Assim, um número menor de fragmentos genômicos foi amostrado, de modo que a profundidade de cobertura por fragmento aumentou. Além disso, cada biblioteca de sequenciamento contendo 96 indivíduos foi sequenciada em duas canaletas (*lanes*) de um equipamento HiSeq, o que duplicou o número de leituras por indivíduo. Útil destacar que, com a abordagem de GBS, é importante escolher cuidadosamente a enzima de restrição e delinear apropriadamente o sequenciamento das bibliotecas (ELSHIRE *et al*., 2011).

4.1.3 Genoma de referência

Utilizou-se a montagem de referência do genoma do sorgo, versão 2.1, que se encontra publicada com sequências representando cromossomos inteiros. Esse genoma pôde ser encontrado na plataforma Phytozome (GOODSTEIN *et al.*, 2012), do Joint Genome Institute. Os dados estão disponíveis desde a publicação do trabalho de Paterson *et. al.* (2009), configurando-se como base para o encontro de polimorfismos no genoma da cana.

4.2 Metodologia

4.2.1 Alinhamento e genotipagem

Após a realização da genotipagem com a técnica GBS, foi necessária a identificação dos sítios variantes. Inicialmente, as leituras brutas foram separadas de acordo com as sequências dos *barcodes*, através do script *GBS barcode splitter*, disponível em Source Forge (2016). Em seguida, as leituras de cada genótipo do PBGCA foram alinhadas contra o genoma de referência do sorgo. Utilizou-se para tanto o programa BWA MEM (LI, 2013), mantendo-se o valor padrão para todos os parâmetros.

A chamada inicial de SNPs e demais tipos de variações foi realizada com o software FreeBayes (GARRISON & MARTH, 2012), considerando as seguintes opções: apenas variantes bialélicos (--use-best-n-alleles 2); inclusão de sítios monomórficos nos indivíduos do painel, porém polimórficos em relação à sequência de referência do sorgo (--use-reference-allele); uso de leituras repetidas, visto que múltiplas sequências de GBS têm início em posições comuns, oriundas dos pontos de corte da enzima (--use-duplicate-reads); e ausência de informação *a priori* sobre a população (--no-population-priors). Importante destacar que esse *software* teve a vantagem de detectar locos variantes fixados e polimórficos no painel de cana-de-açúcar e espécies correlatas, possibilitando a aplicação de testes comparativos entre as duas classes.

Por fim, empregou-se uma adaptação do software SuperMASSA, desenvolvido para uso em dados das plataformas Illumina GoldenGate e Sequenom Iplex MassARRAY, à abordagem da genotipagem por GBS. Tal metodologia tem como objetivo estimar com acurácia a abundância relativa de alelos diferentes, mesmo quando a ploidia da população é desconhecida (MOLLINARI & SERANG, 2014). Com tal estratégia foi possível obter dados quantitativos da dosagem alélica de cada SNP. As contagens dos alelos de referência e

alternativo foram assim empregadas para determinar a ploidia mais provável de cada loco, bem como o genótipo quantitativo de cada indivíduo. Foram testados níveis de ploidia variando de dois a 20, sendo que apenas valores estimados entre seis e 14 foram mantidos. Além disso, tolerou-se a presença de até 50% de dados perdidos por sítio. Desta forma foi obtida uma tabela de genótipos adequada ao cálculo de estatísticas genômicas populacionais para essa espécie poliploide.

4.2.2 Anotação dos sítios variantes por posição e função

As posições variantes detectadas através da estratégia de GBS foram inicialmente anotadas quanto à sua posição e aos seus potenciais efeitos moleculares, isto é, caracterizadas de acordo com a região genômica em que se inserem e a classificação funcional descrita por CINGOLANI *et al.* (2016). Para tanto, foi empregado o *software* SnpEff. Tal caracterização teve como objetivo fornecer informações a respeito da natureza dos sítios polimórficos identificados pela técnica de genotipagem-por-sequenciamento em cana-de-açúcar, bem como fornecer subsídios para investigações a respeito de regiões potencialmente enriquecidas com determinado tipo de variação.

Os *outputs* gerados pelo programa apresentaram-se em documentos de natureza exploratória, que incluem descrições sobre as possíveis consequências moleculares dos poliformismos e sua distribuição ao longo do cromossomo, e em tabelas com as anotações por tipo e região e com as trocas de bases e aminoácidos, discriminadas por posição e por gene, quando foi o caso. Foram utilizados os documentos em formato VCF e TXT para sua manipulação por meio de *scripts* originais produzidos na linguagem R. Essa compreende um ambiente de desenvolvimento integrado para análise de dados, ajuste de modelos estatísticos e construção de gráficos (R, 2016).

O banco de dados anotado foi carregado a partir da função *readVcf* e a discriminação de variantes por genótipos e por posição, bem como suas anotações funcionais, foram acessadas pelas funções *geno*, *rowRanges* e *info*, respectivamente. Todas essas operações ocorreram com o uso do pacote *VariantAnnotation* (OBENCHAIN *et al.*, 2014). Além disso, foi calculada a correlação entre o número de locos polimórficos e o número de genes por cromossomo do sorgo.

4.2.3 Distribuição de frequências alélicas por classe funcional e por grupo de genótipos

Após a caracterização posicional e funcional dos sítios variantes detectados, estatísticas de genética de populações foram obtidas para cada polimorfismo a partir dos dados genotípicos dos indivíduos do PBGCA. Tais estatísticas oferecem evidências a respeito de diversos fenômenos evolutivos, com destaque para os eventos seletivos.

Uma dessas estatísticas é a frequência alélica, que descreve a presença relativa de um alelo para um loco em um grupo de indivíduos, podendo evidenciar os mais diferentes processos populacionais e evolutivos (LUIKART *et al.*, 1998). Seu uso nesse trabalho se justifica no fato de, comparando o grupo ancestral com o grupo melhorado ou simplesmente observando a distribuição de classes funcionais dos locos anotados em todo o painel, ser possível aferir como se comportaram as proporções alélicas durante a evolução provocada pelas seleções artificial e natural. Em posse dessa comparação, é possível se obter alguns indícios de seleção positiva, negativa e neutra. Não obstante, é importante ressaltar que essa análise é de natureza exploratória. Ela teve como objetivo comparar as classes funcionais quanto à maior ou menor incidência de frequências alélicas extremas, isto é, variações raras ou próximas à fixação.

Para tal, uma análise gráfica foi produzida empregando-se o pacote *ggplot2* (WICKHAM, 2011) por meio da contagem de locos distribuídos em janelas de frequências alélicas. Foram definidos intervalos equidistantes de frequências, sendo as janelas espaçadas em 2,5%. Em seguida, a distribuição dos locos foi obtida pela simples contagem dos sítios presentes em cada janela de frequência alélica.

Com base nisso, foram discriminadas as frequências alélicas para três grupos de classes: (1) de efeito alto, moderado, baixo e modificador; (2) regiões genômicas UTR, *up* e *dowstream*, intergênica, de *splicing*, de transcrição e íntron; e (3) pelos tipos mutacionais sinônimo, de sentido trocado e sem sentido. Essa análise teve como objetivo identificar variações na distribuição das frequências alélicas, fornecendo as possíveis evidências seletivas citadas.

4.2.4 Teste de evolução adaptativa por classe de ontologia gênica

Além da distribuição das frequências alélicas por classe de efeito e de região de inserção, é interessante investigar genes com evidências de seleção e relacioná-las às suas possíveis consequências morfofisiológicas. Nesse contexto, empregou-se o teste de McDonald & Kreitman (1991) para aferições por gene.

Como esse teste contrasta variações fixadas e polimórficas, procedeu-se com um *script* em R separando a matriz inicial em duas matrizes com os sítios variantes discriminados por transcrito. A primeira matriz englobou todas as variações filtradas com frequência do alelo alternativo superior a 1%, enquanto a segunda matriz reuniu apenas as variações fixadas, com frequência alélica superior a 99%. Ambas foram anotadas funcionalmente pelo uso individual do programa SnpEff, quantificando-se as variações sinônimas e não-sinônimas em cada uma. Em posse de documentos gerados para as anotações por transcrito, foi possível discriminar, por diferença entre as duas matrizes, o número de sítios variantes polimórficos e fixados, sinônimos e não sinônimos, por gene. Foram também obtidos novos relatórios de análises exploratórias desses conjuntos de dados pelo *software* SnpEff.

O conjunto de polimorfismos aqui utilizados resultou em número pequeno de mutações sinônimas ou não sinônimas para muitos dos genes anotados no genoma do sorgo, o que reduziu o poder estatístico dos testes. Assim, procedeu-se agrupando os genes anotados com termos em comum do *Gene Ontolog* e, em seguida, aplicou-se o teste de McDonald & Kreitman por termo ontológico, utilizando todos os polimorfismos presentes nos genes assim agrupados. Primeiro obteve-se o índice de neutralidade (*NI*) e, em caso de seu valor ter sido inferior a um (indicando seleção positiva), calculou-se o parâmetro α, que informa a proporção de substituições que sofreram seleção direcional.

Considerando a ocorrência de falsos positivos em múltiplos testes, utilizou-se a metodologia de correção baseada na taxa de falsas descobertas (*false dicovery rate* – FDR). Esse é um método estatístico utilizado em testes de múltiplas hipóteses que controla a proporção esperada de hipóteses nulas incorretamente rejeitadas (STOREY, 2002). Isso foi feito a partir do pacote *fdrtool* (STRIMMER, 2008), obtendo-se os valores q a partir dos valores *p* para cada termo GO.

Sendo a hipótese de nulidade (H_0) não haver evidência de seleção ($NI = 1$) e a hipótese alternativa (H_a) haver evidência de seleção ($NI \neq 1$), foram considerados significativos aqueles termos que apresentaram valor *q* inferior a 0,05, ou seja, que desviaram significativamente do esperado sob o nível de confiança de 5%. A partir de tal apuração, procedeu-se com a investigação dos termos significativos a partir da plataforma *Gene Ontology* e, então, morfofisiológica com base na literatura.

4.2.5 Identificação de assinaturas de seleção

A investigação por assinaturas de seleção deu-se com o cálculo da estatística *Hp* a partir da criação de uma função própria e sua aplicação em *script* na linguagem R. As estimativas dessa heterozigosidade combinada foram obtidas para janelas genômicas de 40 Kpb distribuídas ao longo dos dez cromossomos do sorgo. Com a finalidade de investigar essas regiões com maior resolução, as janelas foram sobrepostas em passos de 20 Kpb, conforme sugerido por Rubin (2010). Apenas janelas com cinco ou mais polimorfismos foram consideradas.

Tomou-se como hipótese de nulidade a heterozigosidade combinada da região ser igual ou maior que a encontrada em outras regiões do cromossomo, ou seja, de não haver redução da variabilidade alélica, enquanto a hipótese alternativa foi de a heterozigosidade combinada ser inferior à encontrada nas demais janelas do cromossomo. Para avaliar a significância dos valores de *Hp*, procedeu-se com testes de 10.000 permutações por janela, realizados igualmente em *scripts* originais, observando quantas estimativas eram menores ou iguais à obtida a fim de se auferir valores *p*. Esses resultados foram ajustados em valores *q* pelo método do FDR com o uso do pacote *fdrtool*. O índice de significância empregado para a identificação das regiões genômicas foi de 5%.

Para enriquecer a análise das regiões genômicas identificadas, as janelas significativas foram comparadas com uma base de dados de QTLs de cana-de-açúcar mapeados em sorgo, disponível no Comparative Saccharine Genome Resource (ZHANG *et al.*, 2013). Essa base reúne resultados para caracteres como dias para floração, crescimento da planta, massa do colmo, conteúdo de açúcar (aumento, diminuição ou variação em ambas as direções), teor de fibra, teor de cinzas e POL (porcentagem de sacarose aparente no mosto, determinada pelo desvio provocado pela solução no plano da luz polarizada).

Assim, procurou-se contrastar as posições dos QTLs previamente identificados com as regiões de heterozigosidade combinada significativamente reduzida. Tal análise teve como objetivo buscar fontes de evidência adicional que pudessem corroborar os resultados aqui obtidos.

Uma análise objetivando o encontro de genes candidatos foi realizada procurando-se os transcritos inseridos nas regiões detectadas do genoma do sorgo. Para tal, foi utilizado o visualizador IGV (Interative Genomics Viewer), que facilita o estudo de sequências nas regiões de um genoma de interesse compilando-se a sequência do genoma em formato FASTA com a anotação gênica da espécie em formato BED (ROBINSON *et al.*, 2011a). O

primeiro arquivo foi prontamente obtido por meio da plataforma Phytozome, ao passo que o segundo foi inicialmente adquirido nessa em formato GFF3 e, posteriormente, convertido em formato BED pelo programa BEDOPS (NEPH *et al.*, 2012). A partir da sequência dos transcritos anotados nas janelas significativas, buscaram-se suas possíveis funções com base no alinhamento de cada sequência na plataforma BLASTn do NCBI (National Center for Biotechnology Information), dos Estados Unidos da América (NCBI, 2013), baseando-se nos parâmetros de cobertura e identidade (cerca de 70% ao menos para cada). Finalmente, suas possíveis consequências de natureza molecular, morfológica e/ou fisiológica no desenvolvimento da gramínea foram investigadas a partir de pesquisas na literatura.

5 RESULTADOS E DISCUSSÃO

5.1 Anotação funcional e posicional dos sítios variantes

5.1.1 Descrição dos locos e classificação por posição

O *software* SnpEff foi utilizado para anotar 104.861 sítios variantes em todo o conjunto de dados analisado, o que corresponde a um loco variante a cada 6.278 pares de bases, em média. Desses polimorfismos, 99.291 (94,69%) foram SNPs, 2.116 (2,02%) inserções, 1.658 (1,58%) deleções e 1.796 (1,71%) foram anotados em outras classificações, como mistas de classes anteriores. É relevante destacar que, devido à natureza da metodologia de GBS, a distribuição dos polimorfismos no genoma não foi uniforme, uma vez que sua amostragem é enviesada, com grande concentração de sítios variantes em algumas regiões em detrimento de outras. A relação dos dez cromossomos do sorgo, seu comprimento sequenciado, o número de genes anotados e o número de locos variantes encontrados neste trabalho está disponível na Tabela 4.

TABELA 4 – Número locos variantes no Painel Brasileiro de Genótipos de Cana-de-Açúcar, por cromossomo do sorgo. O comprimento dos cromossomos, em pares de bases, e o número de genes correspondem à versão 2.1 da sequência de referência do genoma de *Sorghum bicolor*.

Cromossomo	Comprimento (pb)	Número de genes	Sítios variantes
1	73.727.935	5.447	18.399
2	77.694.824	4.317	13.742
3	74.408.397	4.461	14.680
4	67.966.759	3.599	11.530
5	62.243.505	2.322	6.301
6	62.192.017	2.822	9.450
7	64.263.908	2.277	7.341
8	55.354.556	1.933	6.430
9	59.454.246	2.610	8.052
10	61.085.274	2.801	8.936
TOTAL	658.391.421	32.589	104.861

Foi observada uma alta correlação ($r = 0{,}9938$, $p = 1{,}39 \times 10^{-11}$) entre o número de genes por cromossomo do sorgo e o número de sítios variantes anotados (Tabela 4, Figura 1). Tal resultado provavelmente indica que a técnica GBS amostrou preferencialmente as regiões gênicas à própria sensibilidade a regiões hipometiladas. Além disso, esse fato corrobora evidências de conservação da sequência em regiões não repetitivas entre o sorgo e a cana-de-açúcar (WANG *et al.*, 2010).

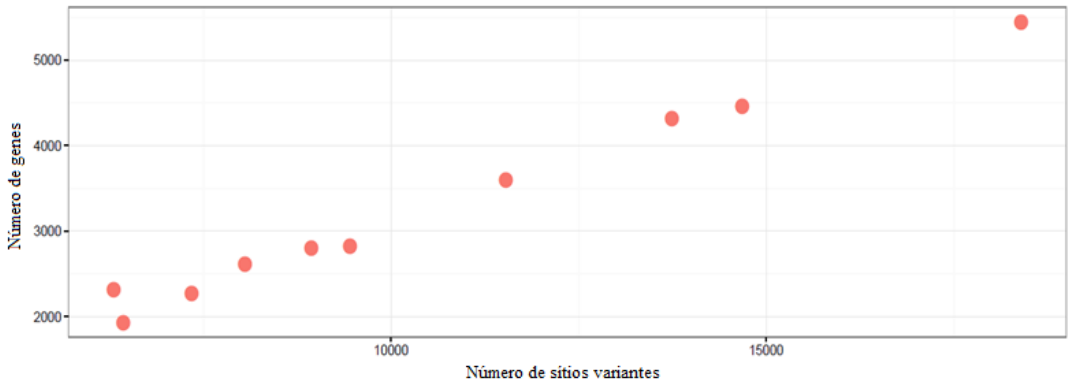


FIGURA 1 – Relação entre o número de sítios variantes detectados em cana-de-açúcar e o número de genes anotados por cromossomo do sorgo.

No que se refere à anotação dos SNPs, foi observada uma taxa de transição por transversão de 1,69. Essa razão é geralmente pouco variável entre populações de uma espécie, podendo também ser discriminada para regiões específicas do genoma. Dessa forma, é útil como indicador de qualidade de dados, além de possibilitar o estabelecimento de estudos filogenéticos entre sequências. Em cana-de-açúcar não há tal descrição disponível na literatura, o que provavelmente se deve à dificuldade de estudos genômicos com esse poliploide. De qualquer maneira, é usual que transições superem transversões na maioria dos casos, havendo algumas exceções bastante específicas, como no caso de comparação de algumas sequências de DNA bacteriano ou mitocondrial (KELLER, 2007; SUTRISNO, 2012).

Bomblies & Doebley (2016) apresentaram estimativas da taxa de transição por transversão a partir de 1,96 para espécies da tribo Andropogoneae. É útil observar que as comparações entre sorgo e cana-de-açúcar e entre indivíduos de cana envolvem genótipos

evolutivamente mais aparentados do que aquelas entre as diversas espécies da tribo, o que pode influenciar o valor da relação.

Anotadas de acordo com a região em que se inseriam, as variações concentraram-se nas regiões exônica (30,01%), na proximidade de genes (48,38%, sendo 28,37% em posições *downstream* e 20,01% *upstream* desses), intrônica (10,16%) e regiões não traduzidas (UTRs) 3’ (2,77%) e 5’ (1,75%) - acumulando 4,52% nessas duas últimas (Figura 2). Como pode ser observado por esse resultado, os locos variantes concentraram-se em regiões gênicas e, mais notoriamente, em regiões transcritas. Esse fato já era esperado, novamente, pela natureza da técnica GBS. Porém, é importante ressaltar que a predominância das mutações nessas posições favorece o estudo da ação da seleção natural e durante o melhoramento genético da cultura.

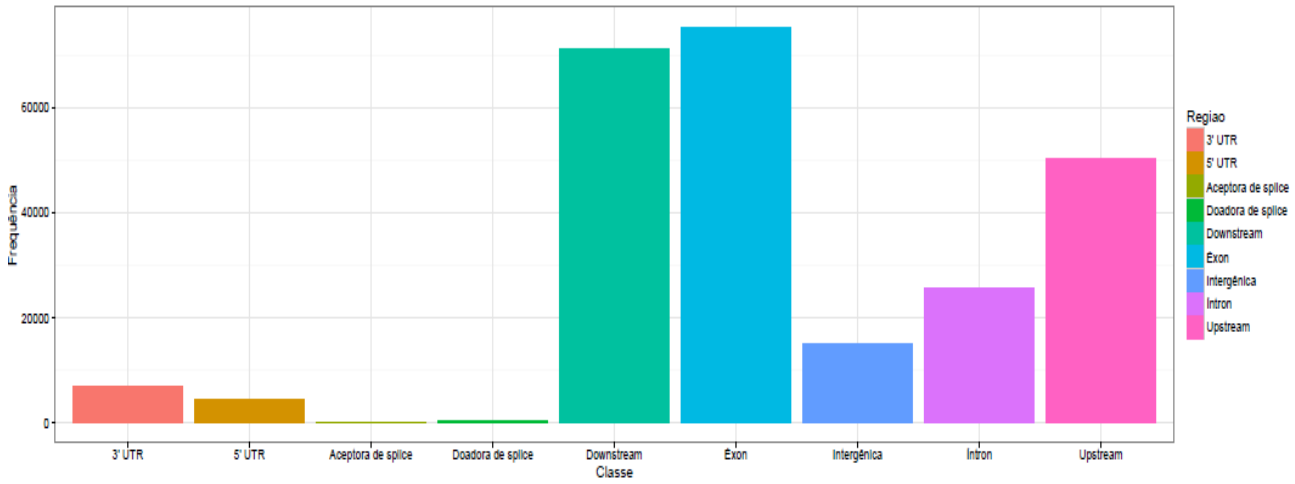


FIGURA 2 - Distribuição dos locos variantes do Painel Brasileiro de Genótipos Cana-de-Açúcar, classificados pelo SnpEff de acordo com sua posição genômica. Foi anotado um total de 104.861 polimorfismos, detectados a partir da genotipagem por sequenciamento utilizando o genoma do sorgo como referência.

Após a filtragem dos dados, com a remoção de locos com frequência do alelo alternativo inferior a 1%, o número de sítios variantes foi reduzido a 34.262, totalizando 32,67% do conjunto original. Esse resultado demonstrou a predominância de variantes de baixa frequência, algo esperado em um genoma altamente heterozigoto, como o de um poliploide (JANOO *et al.*, 2007), embora seja importante frisar que, por ser um painel amplo, foram amostrados alelos raros. Embora a deriva genética tenha uma forte ação sobre a erosão de alelos mutantes, a seleção, positiva ou negativa, pode ter maior dificuldade em fixar um

alelo novo ou eliminá-lo em uma espécie de alta ploidia. Levanta-se a hipótese de que isso seja especialmente verdadeiro em um genoma complexo, com presença de hibridações interespecíficas e eventos de aneuploidia em sua história evolutiva, como é o caso da cana-de-açúcar, além dos programas de melhoramento da cultura serem recentes, o que favorece a ocorrência de variabilidade.

Destaca-se que a natureza das variações (se SNP ou *indel*) não sofreu grande alteração, com 91,43% do primeiro tipo e 5,28% do segundo. Também não se observou diferença na razão entre transições e transversões entre as duas situações, que continuou aproximadamente em 1,69, ou no que se refere à sua distribuição nas regiões genômicas (Figura 3).

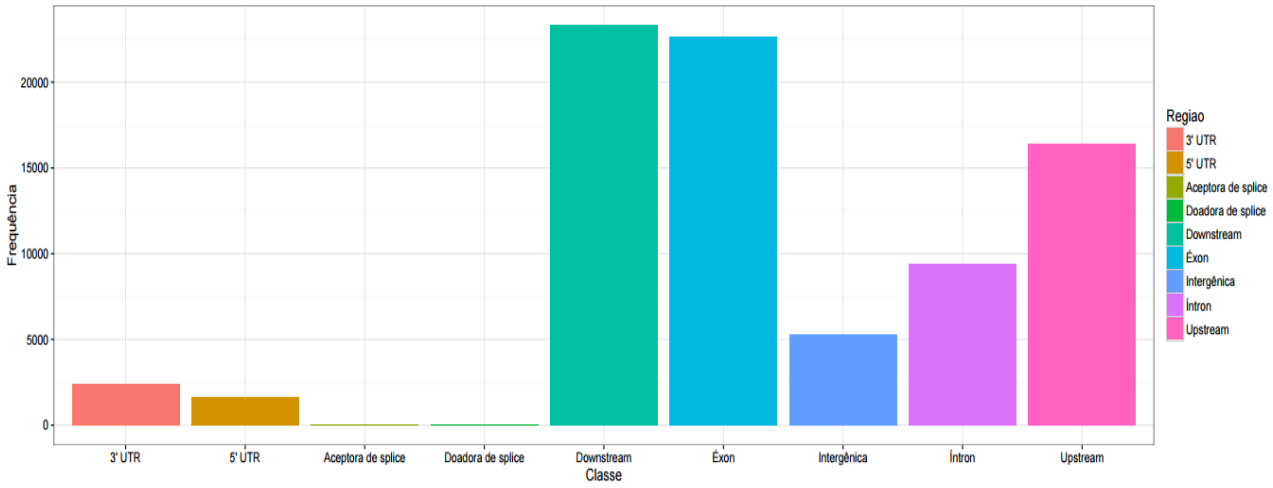


FIGURA 3 - Distribuição por região genômica dos sítios variantes filtrados, realizada pelo SnpEff, no Painel Brasileiro de Genótipos de Cana-de-Açúcar. Os locos filtrados incluem apenas aqueles com frequência do alelo alternativo superior a 1%.

Por fim, foram identificados 4.461 sítios com alelos fixados (frequência alélica superior a 99%), o que corresponde a apenas 13,02% das variações filtradas. Mais uma vez, foram pequenas as diferenças em relação à natureza das variações (89,35% SNP e 5,67% *indel*) e à taxa de transição e transversão (1,66 nesse caso), em comparação com o conjunto de dados antes da filtragem para alelos fixados. Por outro lado, a distribuição de mutações por região apresentou diferenças, embora de maneira pouco distintiva (Figura 4).

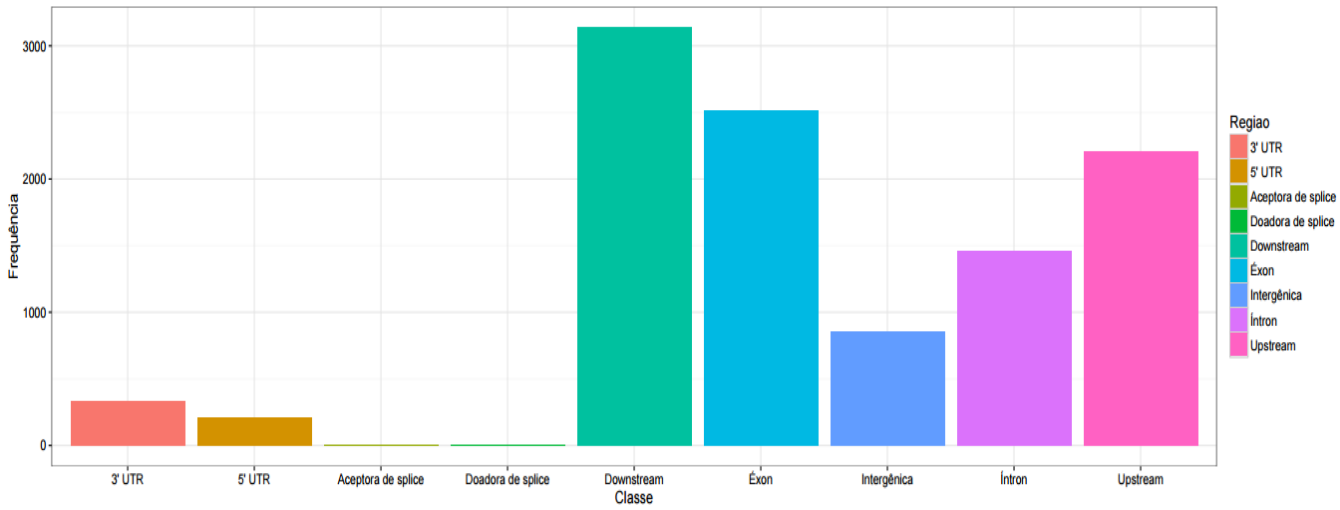


FIGURA 4 - Distribuição dos variantes fixados por região genômica, realizada pelo SnpEff, no Painel Brasileiro de Genótipos de Cana-de-açúcar. Os locos fixados incluem apenas aqueles com frequência alélica superior a 99%.

Comparado aos dados filtrados pela frequência do alelo alternativo, houve a redução da proporção de variantes exônicos (de 27,67% para 23,25%), ao passo que aumentaram os polimorfismos intrônicos (11,45% para 13,54%) e intergênicos (6,48% para 7,93%). Possivelmente, isso se deve ao fato de mutações que ocorrem nas duas últimas regiões, por não serem transcritas (e, portanto, traduzidas), serem mais brandas quanto a seus efeitos. Por esse motivo, durante o melhoramento da cana, essas variações poderiam ter sofrido menor pressão de seleção, como será melhor discutido a seguir.

5.1.2 Classificação de polimorfismos por efeito

Sobre os SNPs no conjunto de dados bruto, foi observada uma taxa de mutações não-sinônimas por sinônimas (ω) de 1,69 (Tabela 5). Considerara-se que os programas de melhoramento genético da cana sejam recentes, com efeito de favorecer a ocorrência de variabilidade. Adicionalmente, é importante ressaltar que o PBGCA reúne principalmente cultivares e genótipos melhorados, de modo que a pressão seletiva possa ter elevado a razão ω propriamente utilizada para locos fixados.

TABELA 5 - Classificação de SNPs em dados brutos, filtrados (frequência do alelo alternativo superior a 1%) e fixados (frequência alélica superior a 99%) nos tipos não-sinônimo, sinônimo e sem sentido, realizada pelo SnpEff, no Painel Brasileiro de Genótipos de Cana-de-Açúcar.

Tipo de polimorfismo	Dados brutos	Dados filtrados	Dados fixados
Não-sinônimo	45.443 (61,95%)	11.259 (52,11%)	945 (39,49%)
Sem sentido	1.115 (1,52%)	178 (0,82%)	3 (0,13%)
Sinônimo	26.798 (36,53%)	10.168 (47,06%)	1.445 (60,38%)
Total	73.356 (100,00%)	21.605 (99,99%)	2.393 (100,00%)

Embora haja evidência de seleção direcional, é importante frisar que possivelmente houve uma seleção purificadora contra mutações de maior impacto na estrutura das proteínas, o que pode ser notado pela baixa presença de variações de efeito alto, segundo a classificação de Cingolani *et al.* (2016) (Tabela 6). Esse fenômeno pode ser explicado pelo fato de a cana-de-açúcar ter sofrido seleção fenotípica durante seu melhoramento, consciente ou inconscientemente, para obtenção de genótipos com características desejáveis. Nesse contexto, mutações favoráveis de pequeno efeito podem ter sido acumuladas gradativamente pela seleção artificial durante diversas gerações. Mutações de efeito mais drástico, porém, teriam sido frequentemente eliminadas, seja pela seleção natural, seja pela seleção artificial, pelo seu provável efeito deletério. Como aponta Ahloowalia *et al.* (2004), é muito pouco provável que a mudança radical da estrutura de uma proteína acarrete em uma maior eficiência metabólica, uma vez que esse fato depende de um ajuste global em uma rota bioquímica, na qual estão envolvidas diversas moléculas e seus genes precursores.

TABELA 6 - Classificação de variações em dados brutos, filtrados (frequência do alelo alternativo superior a 1%) e fixados (frequência alélica superior a 99%) no Painel Brasileiro de Genótipos de Cana-de-Açúcar de acordo com a magnitude predita de seu efeito. Sítios variantes foram classificados pelo SnpEff nas categorias de efeito alto, moderado, baixo e modificador, podendo um mesmo polimorfismo apresentar mais de uma classificação.

Efeito predito	Dados brutos	Dados filtrados	Dados fixados
Alto	2.629 (1,05%)	571 (0,70%)	15 (0,14%)
Moderado	46.691 (18,59%)	12.021 (14,67%)	1.569 (14,52%)
Baixo	29.089 (11,58%)	11.036 (13,47%)	1.047 (9,69%)
Modificador	172.781 (68,79%)	58.295 (71,16%)	8.177 (75,66%)

Comparando-se as anotações oriundas de dados filtrados pela frequência do alelo alternativo com aquelas dos dados brutos, os efeitos preditos por magnitude foram sensivelmente afetados. Houve uma ligeira queda entre os efeitos alto e moderado (que eram 19,64% e passaram a 15,38%) e consequente aumento dos efeitos baixo e modificador (de 80,37% para 84,63%) (Tabela 6). Nota-se uma tendência à eliminação das mutações de efeito mais agressivo, o que seria esperado pelo fato de elas tenderem a ser raras. Ressalta-se que se justifica a eliminação de tais locos variantes das análises subsequentes pela dificuldade de diferenciá-los de erros de sequenciamento, embora diferenças nas dosagens alélicas possam ser percebidas.

Maior alteração ocorreu na classificação dos SNPs quanto aos tipos silencioso, sem sentido e de sentido trocado (Tabela 5). Enquanto os tipos mais agressivos (sem sentido e de sentido trocado) sofreram considerável redução (passando de 1,52% para 0,82% e de 61,95% para 52,11%, respectivamente), as mutações silenciosas subiram de 36,53% para 47,06%. Como consequência, a razão ω reduziu-se a 1,11. Essa queda de evidência de seleção positiva reforça a comum tendência de seleção purificadora em regiões gênicas, majoritárias nas anotações.

Considerando os locos fixados, o número de variações consideradas mais impactantes na estrutura ou função proteicas foi um pouco mais reduzido. As mutações de efeito alto e moderado saíram de um patamar de 15,38% e foram a 14,66%, enquanto as de efeito baixo e modificador deslocaram-se de 84,63% para 85,35% (Tabela 6). Similarmente, as mutações sem sentido reduziram-se de 0,82% para 0,13% e as de sentido trocado de 52,11% para

39,49%, levando ao consequente aumento da proporção de sítios variantes de tipo silencioso de 47,06 para 60,38% (Tabela 5).

Em regiões gênicas, como discutido, é natural que os locos variantes de menor efeito fixem-se em maior proporção. Como observado, comparando-se com o conjunto de dados anterior, houve uma intensa redução de polimorfismos de maior efeito. Sendo assim, espera-se observar uma seleção purificadora (negativa) para a maioria desses locos, o que pôde ser verificado pelo valor de ω igual a 0,65, bastante inferior ao observado para o conjunto de todos os locos.

A relação discriminada das anotações funcional e posicional dos dados brutos, filtrados e fixados está disponível nos Anexo I.

5.2 Efeito das classes de polimorfismos sobre a distribuição das frequências alélicas

5.2.1 Classificação quanto à magnitude dos efeitos

A Figura 5 apresenta as distribuições das frequências do alelo alternativo para os locos classificados de acordo com a magnitude predita de seus efeitos no conjunto completo de genótipos do PBGCA.

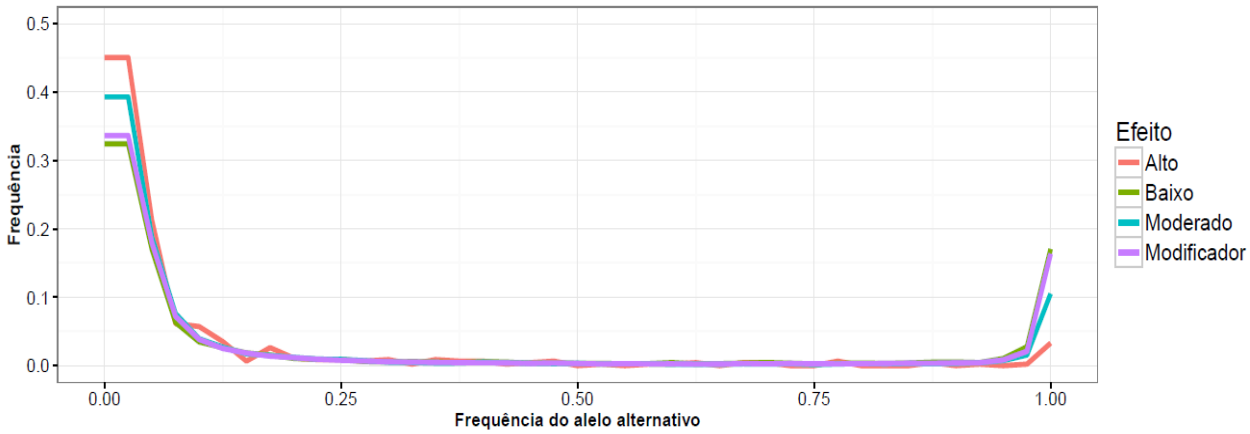


FIGURA 5 - Distribuição da frequência do alelo alternativo para locos variantes classificados de acordo com a magnitude de seus efeitos. Dados referentes aos 266 genótipos do Painel Brasileiro de Genótipos de Cana-de-Açúcar.

Entre a faixa de frequências do alelo alternativo entre 10 e 90%, não houve uma diferença expressiva entre as diversas classes, visto que os polimorfismos tiveram frequência inferior a 10% para todas as classes. Assim, maior atenção é dada aos polimorfismos de frequência mais rara e próximos de serem fixados.

No que se refere à ocorrência de mutações de efeito alto, houve uma predominância das raras, com frequência alélica inferior a 10%. Sua presença diminuiu rapidamente, saindo de um patamar de cerca de 45% para locos em que o alelo alternativo teve frequência entre 0 e 2,5%, para 5% no caso de locos entre 7,5 a 10%. Isso indica que essas mutações são normalmente deletérias, de modo que tendem a ser mantidas em frequências mais baixas.

Em contrapartida, é notável a alteração de posição entre as mutações de efeito moderado com as de efeitos baixo e modificador para polimorfismos situados nos extremos das frequências do alelo alternativo. Tais resultados também eram esperados pelos motivos similares aos expostos para as mutações de tipo alto.

Sabendo-se que esse possível processo purificador ocorreu, provavelmente, pela ação da domesticação e do melhoramento genético da cultura, mostra-se útil analisar uma comparação para os tipos de efeito entre os grupos genotípicos ancestral e melhorado (Figuras 6 e 7).

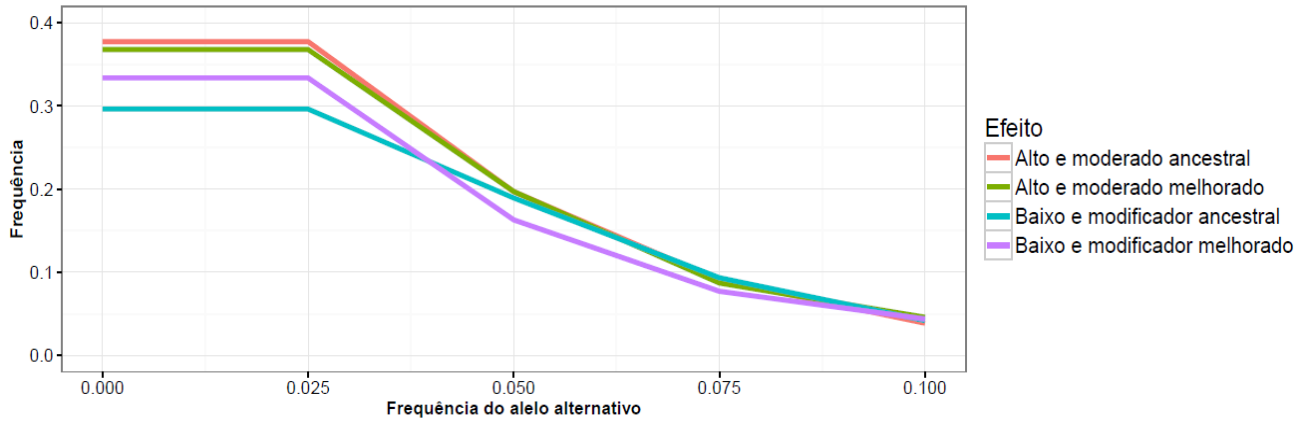


FIGURA 6 – Distribuição da frequência do alelo alternativo de 0 a 10% para locos variantes classificados de acordo com a magnitude de seus efeitos. Dados referentes aos grupos genotípicos ancestral e melhorado do Painel Brasileiro de Genótipos de Cana-de-Açúcar.

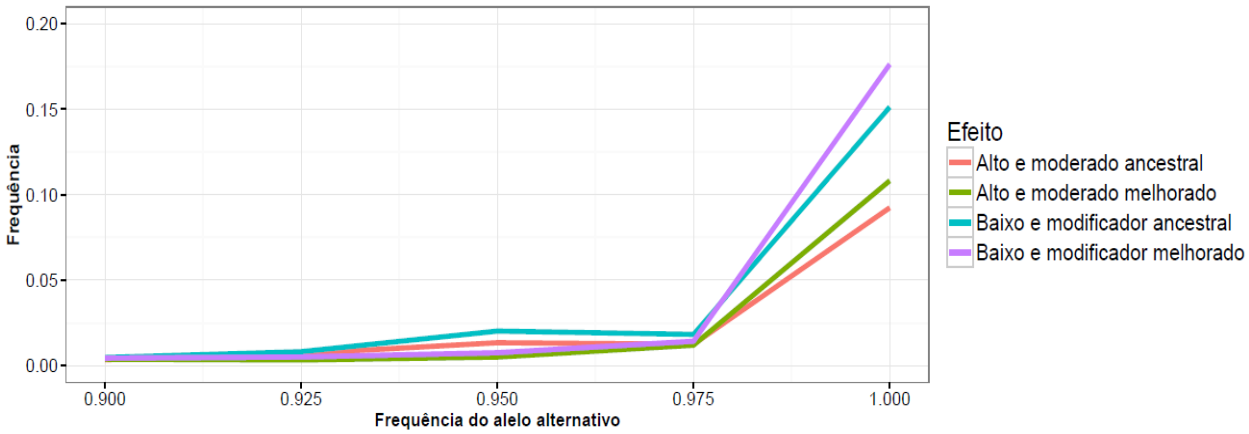


FIGURA 7 – Distribuição da frequência do alelo alternativo de 90 a 100% para locos variantes classificados de acordo com a magnitude de seus efeitos. Dados referentes aos grupos genotípicos ancestral e melhorado do Painel Brasileiro de Genótipos de Cana-de-Açúcar.

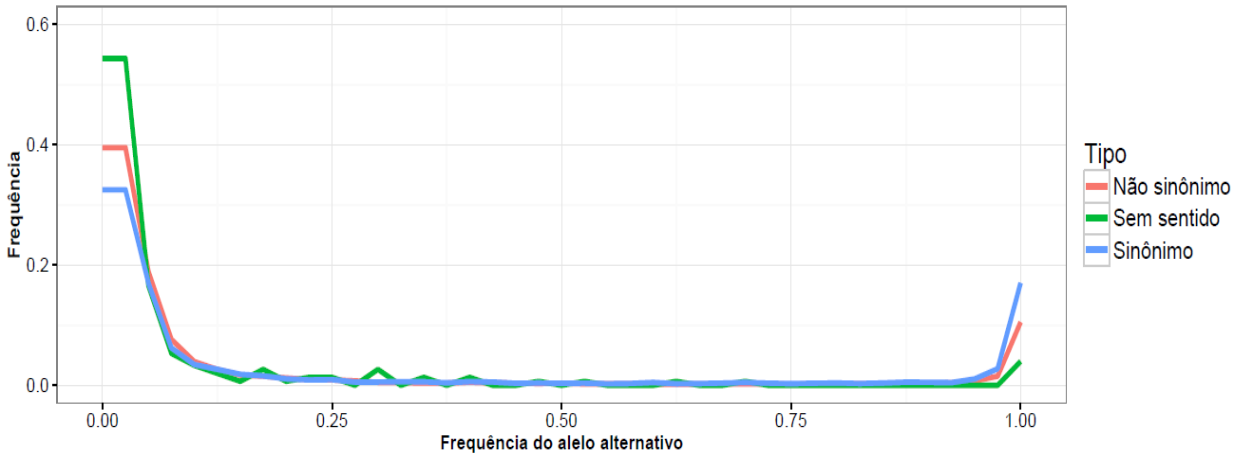
Os resultados parecem confirmar a ação da seleção artificial provocada pelo melhoramento genético da cana-de-açúcar sobre as mutações. Os genótipos melhorados apresentaram mais mutações de efeito alto fixadas. Para os efeitos baixo e modificador, nota-se que tanto mutações mais raras quanto as fixadas foram mais frequentes nos genótipos melhorados do que nos ancestrais, indicando perda de variabilidade genética. Ocorreu apenas pequena variação nas frequências intermediárias do alelo alternativo.

Considerando a tendência exposta, é lúcido dizer que as mutações de efeito alto e moderado que se fixaram ou apresentaram frequência próxima da fixação, embora representem uma pequena parcela do total dessas variações, provavelmente são evolutivamente favoráveis. Ao longo do melhoramento da cana-de-açúcar, essas variações foram positivamente selecionadas no genoma como um todo. Pelo mesmo raciocínio, procederam-se as mutações de efeitos mais brandos. Isso reforça a hipótese de que a seleção artificial apresentou o caráter duplo de selecionar em oposição aos estados de caracteres de maior agressividade, ao passo que apurou características vantajosas. Importante considerar que essa é uma análise do genoma como um todo, e não de regiões gênicas em específico.

5.2.2 Classificação quanto à alteração na estrutura da proteína

Os resultados apresentados na seção anterior são similares para as frequências alélicas dos polimorfismos em regiões codantes, classificados entre os tipos silencioso, de sentido trocado e sem sentido, embora mais marcantes nessa (Figura 8). Considerando-se em grau de potencial pertubação funcional em ordem decrescente as mutações sem sentido, não-sinônimas e sinônimas, esperava-se uma redução proporcional entre essas frequências ao longo da distribuição. De fato, isso aconteceu, sendo a maioria das mutações do tipo sem sentido de frequência rara, seguidas do tipo não-sinônimo e, por fim, do tipo sinônimo, alterando-se a ordem completamente para as fixadas.

Também era esperada uma pequena ocorrência das frequências alélicas intermediárias



do alelo alternativo, o que foi constatado, reafirmando os resultados da classificação dos locos quanto à magnitude do efeito.

FIGURA 8 - Distribuição da frequência do alelo alternativo para locos variantes classificados de acordo com a alteração na estrutura da proteína. Dados referentes aos 266 genótipos do Painel Brasileiro de Genótipos de Cana-de-Açúcar.

Contrastando-se os dois grupos genotípicos em questão quanto aos tipos sinônimo e não-sinônimo (Figuras 9 e 10), esperava-se encontrar evidência semelhante à descrita para os efeitos alto, moderado, baixo e modificador. Os resultados obtidos confirmaram essa premissa, fortificando o argumento evolutivo apresentado. Os genótipos melhorados apresentaram maior proporção de mutações fixadas, tanto para a classe de mutações sinônimas, como para de não sinônimas. Adicionalmente, os genótipos melhorados apresentaram mais mutações sinônimas de baixa frequência do que os ancestrais,

provavelmente refletindo perda de variabilidade genética. Esse fato tem importantes implicações para programas de melhoramento genético, devido à perda de variabilidade após diversos ciclos de seleção, muitas vezes a partir de cruzamentos envolvendo genitores de base genética estreita. Isto reforça a necessidade de introgressão de novos genótipos para uso como genitores.

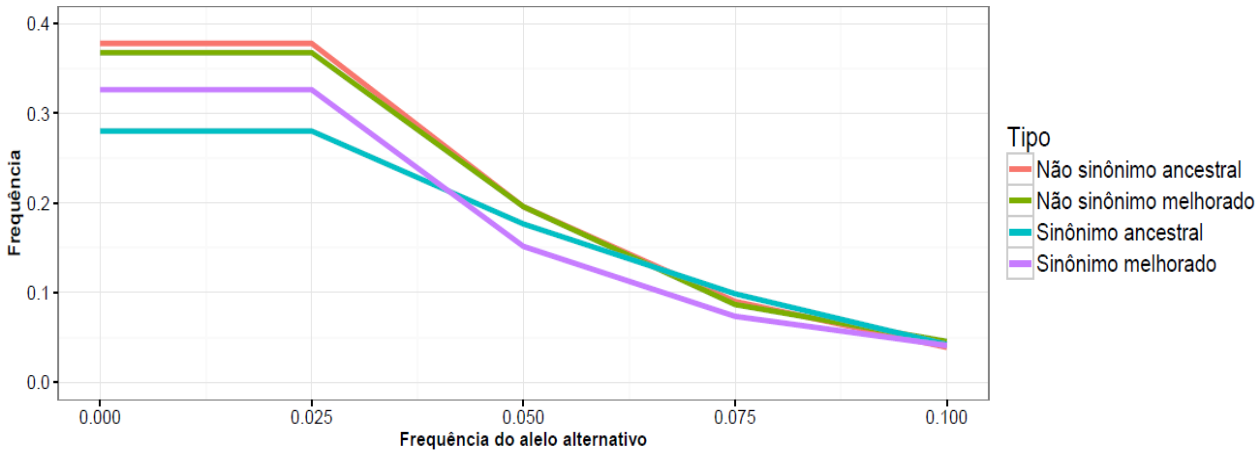


FIGURA 9 – Distribuição da frequência do alelo alternativo de 0 a 10% para locos variantes classificados de acordo com a alteração na estrutura da proteína. Dados referentes aos grupos genotípicos ancestral e melhorado do Painel Brasileiro de Genótipos de Cana-de-Açúcar.

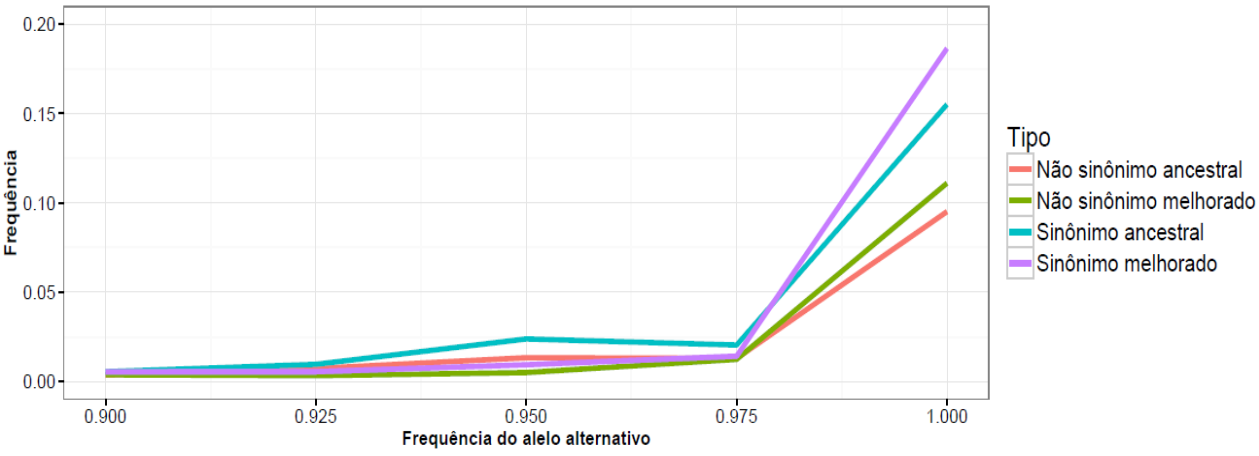


FIGURA 10 – Distribuição da frequência do alelo alternativo de 90 a 100% para locos variantes classificados de acordo com a alteração na estrutura da proteína. Dados referentes aos grupos genotípicos ancestral e melhorado do Painel Brasileiro de Genótipos de Cana-de-Açúcar.

5.2.3 Regiões genômicas

Tendo-se em vista que a análise considerou as anotações funcionais do genoma de forma global até o momento, mostra-se útil uma abordagem discriminando-se as regiões genômicas, seja apenas quanto à incidência das alterações (Figuras 11 e 12), seja contrastando-se os efeitos, com destaque para as regiões gênicas.

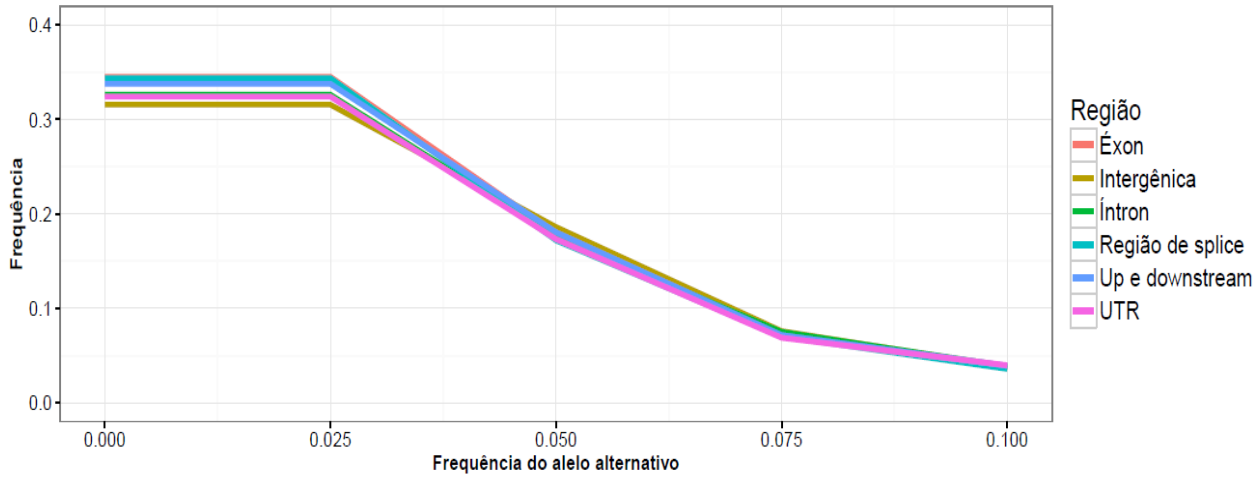


FIGURA 11 - Distribuição da frequência do alelo alternativo de 0 a 10% para locos variantes classificados de acordo com as regiões genômicas intergênica, intrônica, de *splice*, exônica, *up* e *dowstream* e UTR. Dados referentes aos 266 genótipos do Painel Brasileiro de Genótipos de Cana-de-Açúcar.

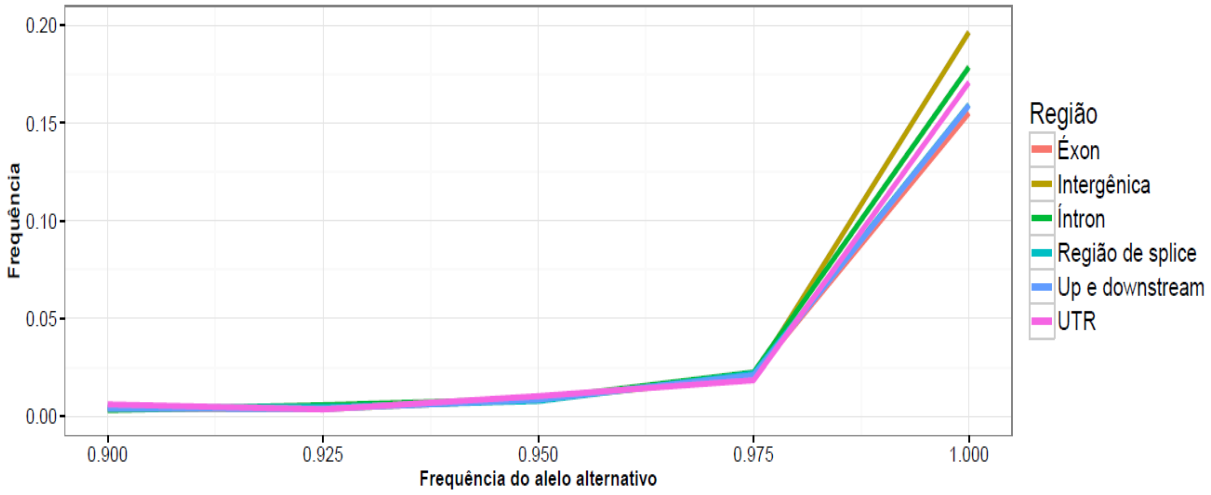


FIGURA 12 - Distribuição da frequência do alelo alternativo de 90 a 100% para locos variantes classificados de acordo com as regiões genômicas intergênica, intrônica, de *splice*, exônica, *up* e *dowstream* e UTR. Dados referentes aos 266 genótipos do Painel Brasileiro de Genótipos de Cana-de-Açúcar.

Analisando-se as mutações raras, é possível se discriminar principalmente três grupos posicionais quanto à passividade de evolução adaptativa. Em maior proporção estão os mais suceptíveis de sofrer pressões seletivas, as regiões codantes (exônica), de *splice* e *up* e *downstream*. Esse resultado contrasta-se com a menor frequência dessas mesmas regiões no caso das mutações fixadas. Observa-se, pois, a tendência evolutiva de seleção purificadora para as alterações que ocorrem em regiões mais diretamente envolvidas com a síntese proteica. Justifica-se o agrupamento dessas por duas delas serem gênicas de ação direta no transcrito (a exônica afeta a codificação do peptídeo enquanto a de *splice* afeta seu processamento) e uma ser intimamente ligada à expressão gênica (*up* e *downstream*, por seu possível papel regulatório).

Um segundo grupo, de comportamento intermediário, é o que reúne duas regiões de sensibilidade igualmente intermediária quanto à evolução adaptativa. Essas regiões são UTR e intrônica, ambas gênicas, porém não traduzidas (a primeira apenas não é traduzida e a segunda não permanece no RNA processado). Ao longo do gráfico, ambas apresentam distribuição regular, sem comportamentos antagônicos, indicando uma mais provável neutralização das ações seletivas contrárias.

Por fim, uma única região (intergênica) apresentou o comportamento oposto da primeira. Por não estar envolvida, nem direta, nem indiretamente propriamente com a

codificação, sua susceptibilidade à ação seletiva é muito reduzida. Seu comportamento antagônico ao primeiro grupo confirma a tendência apresentada.

Por fim, frisa-se que os locos variantes com frequências intermediárias do alelo alternativo, isto é, aqueles com frequência entre 10 e 90%, foram encontrados em pequena quantidade. Essa intensa redução demonstra uma possível intensa ação da seleção no genoma da gramínea, o que reforça as análises exploratórias realizadas. Por sua vez, a mais rara existência de alelos totalmente fixados em relação ao total justifica-se pelo fato simples da maioria das mutações ser deletéria ou neutra e o genoma da cana-de-açúcar ser poliploide, implicando em latentes dificuldades de se atingir a homozigose. Considera-se, também, que grande parte da ação seletiva encontrada é artificial e os programas de melhoramento genético aplicados à cultura são bastante recentes sob o ponto de vista histórico.

5.3 Teste de evolução adaptativa por classe de ontologia gênica

5.3.1 Análise descritiva e termos de componente celular

A aplicação do teste de McDonald & Kreitman resultou em 28 termos GO significativos após correção de múltiplos testes, dos quais dois mostraram-se redundantes e dois apresentaram funcionalidade semelhante, totalizando 26 classes funcionais. Os resultados dos índices de neutralidade foram maiores do que um para todos os grupos gênicos identificados, o que impossibilitou o cálculo do parâmetro α. Além disso, é importante ressaltar que quatro índices não puderam ser calculados devido ao pequeno número de SNPs, visto que dois apresentaram ausência de mutações fixadas não-sinônimas somente e dois de mutações fixadas não-sinônimas e polimórficas sinônimas. Uma relação dos termos da base *Gene Ontology* identificados no teste, com os valores de NI e a discriminação dos polimorfismos componentes de seu cálculo encontra-se presente na Tabela 7.

Tabela 7 – Termos do *Gene Ontology* significativos ($\alpha = 0{,}05$) para aplicação do Teste de McDonald & Kreitman no Painel Brasileiro de Genótipos de Cana-de-Açúcar. São mostrados os índices de neutralidade (*NI*), as frequências de mutações fixadas sinônimas (*Ds*), polimórficas sinônimas (*Ps*), fixadas não-sinônimas (*Dn*) e polimórficas não sinônimas (*Pn*) e os valores *q*. Os locos filtrados correspondem àqueles com frequência do alelo alternativo superior a 1%, enquanto que os fixados têm frequência alélica superior a 99%. O valor *p* foi corrigido por meio do método FDR (*False Discovery Rate*).

Termo	Nome	NI	Ds	Ps	Dn	Pn	Valor *q*
GO:0005622	Intracelular	2,089063	42	160	24	191	$4{,}4\ x\ 10^{-2}$
GO:0005634	Núcleo	2,51799	31	369	7	97	$1{,}9\ x\ 10^{-3}$
GO:0016021	Componente integral da membrana	2,747087	41	236	16	253	$8{,}0\ x\ 10^{-3}$
GO:0005524	Ligação de ATP	1,835165	167	1092	106	1272	$1{,}1\ x\ 10^{-4}$
GO:0005515	Ligação proteica	1,872011	182	1106	117	1331	$2{,}3\ x\ 10^{-5}$
GO:0004672	Atividade de proteína quinase	2,371635	111	743	56	889	$1{,}9\ x\ 10^{-5}$
GO:0003700	Atividade de fator de transcrição	2,329907	36	214	20	277	$2{,}8\ x\ 10^{-2}$
GO:0043565	Atividade de fator de transcrição	3,294147	29	168	12	229	$7{,}0\ x\ 10^{-3}$
GO:0003677	Ligação ao DNA	2,089279	87	415	59	588	$1{,}2\ x\ 10^{-3}$
GO:0004553	Atividade de hidrolase	3,553114	20	78	7	97	$3{,}2\ x\ 10^{-2}$
GO:0016706	Atividade de oxirredutase	6,439024	12	41	3	66	$2{,}2\ x\ 10^{-2}$
GO:0016616	Atividade de oxirredutase	5,824405	19	84	4	103	$6{,}7\ x\ 10^{-3}$
GO:0003824	Atividade catalítica	2,988027	40	226	17	287	$3{,}2\ x\ 10^{-3}$
GO:0043531	Ligação de ADP	2,146807	35	221	27	366	$3{,}4\ x\ 10^{-2}$
GO:0050662	Ligação de coenzima	4,635417	15	72	4	89	$3{,}3\ x\ 10^{-2}$

Termo	Nome	NI	Ds	Ps	Dn	Pn	Valor q
GO:0008168	Atividade de metiltransferase	Infinito*	7	45	0	39	$3{,}8\ x\ 10^{-2}$
GO:0003854	Atividade de desidrogenase de 3-beta-hidroxi-delta-5-esteroide	10,56604	14	53	2	80	$3{,}7\ x\ 10^{-3}$
GO:0009378	Atividade de helicase de junção de quatro fitas	Infinito*	5	16	0	48	$8{,}9\ x\ 10^{-3}$
GO:0003849	Atividade de sintase de 3-desoxi-fosfoheptulonato	Infinito*	3	0	0	4	$3{,}6\ x\ 10^{-2}$
GO:0006468	Fosforilação de proteína	2,379639	111	743	56	892	$1{,}9\ x\ 10^{-6}$
GO:0006629	Processo metabólico de lipídeo	3,076923	20	91	9	126	$4{,}1\ x\ 10^{-2}$
GO:0055085	Transporte transmembranar	2,341629	45	286	17	253	$2{,}6\ x\ 10^{-2}$
GO:0055114	Processo redox	2,145541	87	546	43	579	$1{,}7\ x\ 10^{-3}$
GO:0005975	Processo metabólico de carboidratos	2,702564	31	130	12	136	$3{,}2\ x\ 10^{-2}$
GO:0044237	Processo metabólico celular	6,615385	15	65	3	86	$1{,}1\ x\ 10^{-2}$
GO:0006694	Processo biossintético de esteroide	10,56604	14	53	2	80	$3{,}7\ x\ 10^{-3}$
GO:0006457	Dobramento proteico	Infinito*	4	22	0	41	$4{,}8\ x\ 10^{-2}$
GO:0009073	Processo de biossíntese de aminoácidos aromáticos	Infinito*	3	0	0	4	$3{,}6\ x\ 10^{-2}$

***** Resultado da divisão por zero em condições de baixa contagem de polimorfismos.

A partir desses fatos expostos, observa-se que os termos apresentaram somente seleção purificadora. Isso se deve, muito provavelmente, ao fato de esses corresponderem a funções importantes ao funcionamento da planta, de forma que modificações tenderam a ser minimizadas dado o provável efeito deletério. Esse fato foi abordado anteriormente por Ahloowalia *et al.* (2004), diferindo-se aqui pela correspondência não se relacionar ao impacto na proteína, mas no metabolismo como um todo. É importante salientar que, mesmo nos termos com NI infinito, houve uma indicação de seleção negativa, embora sua significância possa ser questionada pela violação dos princípios do teste. Esses termos relacionaram-se à função molecular de atividade de helicase de junção de quatro fitas (GO:0009378), à função de sintase de 3-desoxi-7-fosfo-heptulonato (GO:0003849), ao processo biológico da biossíntese de aminoácidos aromáticos (GO:0009073) e à função de atividade da metiltransferase (GO:0008158).

Sobre o último termo, é necessário destacar que foram identificadas 91 mutações, com ausência apenas de SNPs fixados não-sinônimos, de maneira que uma análise exploratória revela uma possível importância do termo. Saini *et al.* (1995) descrevem essa grande classe de enzimas como importante para diversos processos metabólicos, com destaque em plantas para a biossíntese de halometanos e metanotiol, importantes nos processos de combustão celular. Em cana-de-açúcar, a metiltransferase desempenha atividade relacionada à biossíntese de metabólitos secundários (FRANÇA *et al.*, 2001) e à lignificação (WHETTEN & SEDEROFF, 1995) - essa tocante ao metabolismo do silício e a questões de porte e proteção do vegetal.

Das 22 ontologias gênicas de interesse, onze foram relativas a funções moleculares (50,0%), oito a processos biológicos (36,4%) e três a componentes celulares (13,6%). Os três termos de componentes celulares detectados corresponderam a níveis amplos na hierarquia, a saber: componentes intracelular (GO:0005622), nuclear (GO:0005634) e integral da membrana (GO:0016021). Assim, não será feita aqui uma descrição detalhada.

5.3.2 Termos de função molecular

São especialmente interessantes os termos de função molecular GO:0016706/ GO:0016616 (atividade de oxirredutase), GO:0004553 (atividade de hidrólise de ligações O-glicosídicas) e GO:0003854 (atividade de desidrogenase de 3-beta-hidroxi-delta-5-esteroide) dentre os termos de função molecular (Tabela 7).

A atividade de oxirredutase citada relaciona-se à ação do 2-oxoglutarato (2OG), que possui importância na assimilação de amônio em plantas. O nitrogênio possui importante

função em quase todos os seres vivos, sendo constituinte de aminoácios livres e proteicos, bases nitrogenadas, que perfazem os ácidos nucleicos, e outros compostos importantes, como a síntese de hormônios, coenzimas, alcaloides e hexosaminas. Em plantas, esse nutriente apresenta ligação iônica com o magnésio na formação da clorofila, responsável pela acepção e distribuição de energia luminosa na fase fotoquímica da fotossíntese. Nesse contexto, o amônio (NH_4^+) é a forma iônica de alta mobilidade no solo e no interior dos tecidos vegetais cujo contato íon-raiz ocorre majoritariamente por fluxo de massa. A assimilação do amônio é uma etapa sensivelmente importante, especialmente considerando sua competição com o nitrato (NO_3^-). Esse íon também é absorvido pelas plantas, embora com papel muito reduzido perante o primeiro no caso da maioria das plantas cultivadas, como é o caso da cana-de-açúcar (ROBINSON *et al.*, 2011b). Além disso, sua assimilação é importante para que a conversão e permanência da amônia (NH_3) nos tecidos vegetais não perdure, uma vez que esse composto pode ser tóxico metabolicamente (LEA & IRELAND, 1999).

A síntese de glutamato e glutamina é muito importante para a maioria dos compostos nitrogenados, sendo, portanto, a cadeia de ação das enzimas sintetase de glutamina (GS) e sintase de glutamato (GOGAT) especialmente relevante. O ciclo GS/GOGAT exige energia na forma de ATP com GOGAT utilizando agentes redutores sob a forma de 2OG. Esse último também se apresenta inerente no Ciclo de Krebs, envolvendo-se com a respiração e com a síntese de ácidos orgânicos, possui relevância na síntese de giberelina (um importante hormônio de crescimento de ramos e folhas) e seus níveis podem refletir o estado C/N dentro da célula, relacionado ao monitoramento e sinalização para a maquinaria reguladora vegetal (HODGES, 2002; LACIEN *et al.*, 2000).

No que se refere à hidrolase, o termo relaciona-se aos componentes da hidrólise de ligações O-glicosil, ligada a processos diversos na planta. Genericamente, pode-se elencar sua influência na biossíntese de metionina, que, segundo Ravanel *et al.* (1998), é um precursor de poliaminas (estando envolvida, portanto, na produção de aminoácidos) e de etileno. Esse último composto é um fitormônio de grande importância na senescência de plantas, possuindo autocatálise e atuando em diversos processos, como abscisão de órgãos, expansão celular horizontal, inibição do florescimento, formação de raízes, germinação de sementes, dentre outros (DAVIES, 2010).

Mais abrangentemente, Davies & Herinssat (1995) pontuam que a hidrólise seletiva de ligações glicosídicas é crucial para a absorção de energia, a expansão da parede celular e a degradação de celulose. Nesse sentido, o O-glicosil exerce presença fundamental em uma variedade significativa de funções biológicas, incluindo armazenamento e transdução de vias

de sinalização em vegetais – reforçando a ação dos genes participantes dessa ontologia em relação ao sistema de membranas (envoltórios celulares e endomembranas). Nesse caminho, Henrissat *et al.* (1998) abordam um esquema para designar a hidrólise de polissacarídeos nas paredes celulares, que inclui a ação de enzimas do domínio de hidrolases glicosídicas e seu papel nas reações catalíticas.

Por fim, no tocante à atividade de desidrogenase de 3-beta-hidroxi-delta-5-esteroide, nota-se uma farta abrangência de trabalhos em animais, como no metabolismo hepático de esteroides, cujo exemplo é um trabalho de andrógeno em porcos de Kim *et al.* (2013). Em plantas, sua relevância relaciona-se à biossíntese de fitoesteroides, cuja abordagem será realizada a seguir.

5.3.3 Termos de processo biológico

É menos genérico o termo GO:0006694 (processo biossintético de esteroide) dentre os de processo biológico (Tabela 7). O trabalho de Heftmann (1975) resume minuciosamente as funções dos esteroides em plantas, de forma que sua biossíntese está substancialmente ligada a questões como precursão de metabólitos secundários e ação em vias de sinalização e resposta. Assim, esses compostos estão envolvidos na recepção de informações pela célula e em mecanismos de defesa planta-patógeno. Mais especificamente, destacam-se os mecanismos bioquímicos pós-formados para o combate de infecções fúngicas.

Esses fatores de resistência são bastante importantes em cana-de-açúcar, tendo em vista a já citada ausência de grande diversidade de produtos fitossanitários para a cultura e sendo a infestação fúngica algo bastante problemático em canaviais, comprometendo a produção de mosto e sua qualidade. Nesse sentido, segundo Thaler *et al.* (1999), as fitoalexinas são os principais compostos antimicóticos, possuem baixo peso molecular e acumulam-se nas células em resposta às infecções, sendo produzidos em função de estímulos resultantes de elicitores. Esses são compostos de diferentes naturezas que sinalizam o ataque patogênico, originando-se da planta em forma de moléculas degradadas ou do próprio fungo.

São ainda bastante relevantes os termos GO:0005975 (metabolismo de carboidratos) e GO:0006468 (fosforilação de proteínas), embora sejam bastante gerais. Em relação ao metabolismo de carboidratos, a importância em cana-de-açúcar é clara, uma vez que essa é a característica valorada na cultura, além de ser básica em mecanismos de resistência e defesa contra patógenos e outros processos em geral do metabolismo (SCARPARI & BEAUCLAIR, 2008; JONES & JONES, 1997). Em relação à foforilação de proteínas, aborda-se sua

importância no metabolismo de enzimas de um espectro bastante grande, incluindo processos energéticos e regulatórios da maquinaria celular (DENANCÉ *et al.*, 2014).

5.4 Assinaturas de seleção

5.4.1 Panorama geral dos resultados

Utilizando-se o parâmetro *Hp*, foram identificadas 19 regiões com evidências de assinaturas de seleção. Oito dessas regiões eram sobrepostas, de modo que a junção das adjacentes resultou em 15 regiões distintas. Essas janelas localizaram-se em todos os cromossomos do genoma do sorgo, à exceção do cromossomo seis (Tabela 8). Uma relação detalhada das regiões com seus SNPs está disponível na Tabela 9.

TABELA 8 - Regiões com assinaturas de seleção detectadas a partir do parâmetro de heterozigosidade combinada (*Hp*) no Painel Brasileiro de Genótipos de Cana-de-Açúcar. O índice de significância empregado foi de 5% e os cromossomos estão representados pela letra *C*.

Correspondência	Regiões (cromossomo e posição em pb)	Tipo de evidência
Metabolismo/biossíntese de lipídeos/ácidos graxos	C1(54.680.001 a 54.720.000), C3(71.780.001 a 71.840.000), C4(6.560.001 a 6.600.000) e C5(620.001 a 660.000)	Transcritos
Caracteres de altura de planta	C2(61.800.001 a 61.860.000), C3(70.100.001 a 70.140.000 e 11.780.000 a 11.820.000) e C9(57.280.001 a 57.320.000)	QTLs
Processos energéticos	C1(54.680.001 a 54.720.000), C2(61.800.001 a 61.860.000), C3(11.780.000 a 11.820.000), C4(6.560.001 a 6.600.000), C5(620.001 a 660.000), C7(580.001 a 59.620.000) e C8(480.001 a 520.000)	Transcritos
Transdução de sinal, via hormonal e constituição da membrana	C2(61.800.001 a 61.860.000), C3(71.780.001 a 71.840.000), C5(2.300.001 a 2.360.000) e C7(120.001 a 160.000 e 580.001 a 59.620.000)	Transcritos

Referência	Regiões (cromossomo e posição em pb)	Tipo de evidência
Caracteres de dormência de semente	C3(70.100.001 a 70.140.000)	QTL deduzido de transcrito
Resistência a doenças	C3(70.100.001 a 70.140.000), C5(620.001 a 660.000) e C10(54.820.001 a 54.880.000)	Transcritos
Marcação/conformação de proteínas	C3(70.100.001 a 70.140.000 e 71.780.001 a 71.840.000)	Transcritos
Metabolismo/biossíntese de sacarídeos	C3(71.780.001 a 71.840.000) e C9(57.280.001 a 57.320.000)	Transcritos e QTLs
Desintoxicação	C3(71.780.001 a 71.840.000)	Transcritos
Biossíntese de alcoóis graxos e translocação de solutos	C3(71.780.001 a 71.840.000) e C8(480.001 a 520.000)	Transcritos
Parede celular	C4(6.560.001 a 6.600.000), C7(120.001 a 160.000) e C8(480.001 a 520.000)	Transcritos
Floração	C3(70.100.001 a 70.140.000) e C5(620.001 a 660.000)	Transcritos
Processamento de RNA	C5(620.001 a 660.000)	Transcrito
Morte celular programada	C7(120.001 a 160.000)	Transcrito
Processos da cromatina/ expressão gênica	C7(580.001 a 59.620.000)	Transcritos
Crescimento e desenvolvimento/ resposta a estímulo ambiental	C7(580.001 a 59.620.000) e C10(54.820.001 a 54.880.000)	Transcritos
Ação da polifenoloxidase	C8(2.620.001 e 2.660.000)	Transcrito
Tradução de proteínas	C9(57.280.001 a 57.320.000)	Transcritos

Quatro das regiões identificadas estavam próximas a QTLs já mapeados, de acordo com a base Comparative Saccharine Genome Resource (ZHANG *et al.*, 2016), sendo uma dessas regiões situada no cromossomo dois e uma no cromossomo três com tamanho de 60 Kpb. Os caracteres quantitativos correspondentes foram teor de açúcar e altura da planta. Genericamente, foram encontrados diversos transcritos nas regiões de variabilidade reduzida, cujas possíveis funções incluíam processos energéticos celulares (sete regiões), biossíntese e metabolismo de lipídeos e/ou ácidos graxos (quatro), rotas hormonais, vias de transdução de sinal e/ou constituição das membranas celulares (quatro), resistência a patógenos (três), biossíntese e metabolismo de sacarídeos (dois), biossíntese de alcoóis graxos e translocação de solutos (dois), constituintes da parede celular (dois), floração (dois), dentre outros. Apenas em uma região das 15, localizada no cromossomo oito, não foram encontradas evidências de QTL ou transcritos anotados.

Tabela 9 – Janelas com evidência de assinatura de seleção detectadas através da heterozigosidade combinada (*Hp*) no Painel Brasileiro de Genótipos de Cana-de-Açúcar. Os locos foram filtrados para frequência do alelo alternativo superior a 1% e o valor *p* foi corrigido por meio do método FDR (*False Discovery Rate*). O valor de significância empregado foi de $\mathbf{0,05}$. Os efeitos alto, moderado, baixo e modificador são descritos por Cingolani *et al.* (2005).

Cromossomo	Início da janela	Tamanho da janela	Alto	Moderado	Baixo	Modificador	Sinônimo	Não-sinônimo	Hp	Valor *q*
1	54.680.001	40 Kpb	0	2	2	2	2	2	0,008	0,047
2	61.800.001	60 Kpb	0	1	1	14	0	1	0,008/0,014	0,000/0,047
3	11.780.001	40 Kpb	0	4	1	0	1	4	0,008	0,047
3	70.100.001	40 Kpb	0	0	0	22	0	0	0,004	0,031
3	71.780.001	60 Kpb	0	3	9	29	9	3	0,009	0,000
4	6.560.001	40 Kpb	0	0	0	16	0	0	0,013	0,000
5	620.001	40 Kpb	0	9	9	39	9	9	0,014	0,000
5	2.300.001	60 Kpb	0	0	5	14	5	0	0,006/0,008	0,000/0,031
7	120.001	40 Kpb	0	3	6	36	5	3	0,024	0,047
7	59.580.001	40 Kpb	2	4	10	89	10	4	0,024	0,000
8	480.001	40 Kpb	1	1	5	34	5	1	0,026	0,031
8	2.620.001	40 Kpb	0	0	1	51	0	0	0,033	0,047
8	54.260.001	40 Kpb	0	12	11	57	11	12	0,038	0,031
9	57.280.001	40 Kpb	1	2	3	32	3	2	0,016	0,000
10	54.820.001	60 Kpb	0	0	0	15	0	0	0,009	0,000

5.4.2 Cromossomo 1

No cromossomo um, apenas uma região com assinatura de seleção foi identificada, situada entre as posições 54.680.001 e 54.720.000 pb. Não foi identificado qualquer QTL próximo a essa janela.

A análise das sequências genômicas revelou uma proteína similar a uma codificada conhecida, a fosfatidilcolina-esterol O-aciltransferase, relativa ao metabolismo de lipídeos em folhas, essencial à biossíntese de ésteres. Como apontam Totton & Lardy (1949), os ésteres são componentes de ácidos graxos, com funções de hidrólise e saponificação, sendo essenciais à atividade dos lipídeos, que desempenham funções muito abrangentes em todas as células. Destaca-se a presença desses compostos orgânicos nas membranas, funcionando como meio de proteção interna, transdução de sinais, material energético (incluindo reserva) e dissolução de vitaminas. Hunsigi (1993) relaciona essa proteína à síntese de clorofila, cuja importância para a fotossíntese já foi abordada.

5.4.3 Cromossomo 2

Duas janelas adjacentes no cromossomo dois com evidência de varredura seletiva foram encontradas, compreendendo, portanto, uma única região contínua com 60 Kpb, da posição 61.800.001 à 61.860.000 pb. Funcionalmente, os SNPs localizados dentro desta região foram majoritariamente modificadores (87,5%), o que, considerando haver somente 16 polimorfismos identificados, comprometeu análises mais aprofundadas relativas ao quociente ω (com um SNP não-sinônimo e nenhum sinônimo anotados).

O QTL Q.Ming2002a.PLHT.31.2_4-1, relativo à altura da planta, foi identificado entre as posições 61.550.813 e 61.551.763. Segundo Moore (1987), a diferenciação de altura do colmo durante a domesticação da cana-de-açúcar foi um fator bastante pronunciado.

No relativo à sequência genômica, foram identificadas oito proteínas codificadas nessa região, sendo sua função totalmente desconhecida para três delas e sendo uma dessas presente em milho, porém sem detalhes a respeito de sua anotação funcional. Das quatro restantes, uma sequência é similar à cinesina KP1, que, de acordo com Yang *et al.* (2011), está relacionada à respiração e à cadeia energética. Nesse sentido, em outra sequência foi encontrada similaridade à proteína dissulfeto-isomerase SCO2, envolvida no desenvolvimento de cloroplastos, com destaque para a formação de tilacoides – estruturas membranosas envolvidas na captação e na distribuição de energia na fase fotoquímica da fotossíntese. Foi

descrita também uma relação com germinação e cloroplastos cotiledonares (TANZ *et al.*, 2012).

Das duas proteínas restantes, uma não está totalmente descrita, mas presume-se um envolvimento no grupo proteico 12b de ligação de cálcio. Tal fato é reforçado pela similaridade da outra sequência à proteína calmodulina, que, segundo Magnan (2008), apresenta importância em plantas ligada à recepção de Ca^{2+} em vias de transdução de sinal importantes, em que atuam hormônios e fatores de estresse abiótico.

5.4.4 Cromossomo 3

Foram identificadas três regiões significativas no cromossomo três, sendo uma de 60 Kpb, entre 71.780.001 e 71.840.000 pb, uma relativamente próxima a essa, no intervalo entre 70.100.001 e 70.140.000 pb, e uma na região entre 11.780.001 e 11.820.000. A primeira região apresentou 41 SNPs anotados, dos quais 70,7% foram de efeito modificador, 22,0% de efeito baixo e 7,3% de efeito moderado. A taxa ω foi de 0,33, fornecendo indícios de haver seleção purificadora. A região das posições 70.100.001 a 70.140.000 apenas apresentou SNPs de efeito modificador (22) e a de 11.780.001 a 11.820.000 não conteve uma quantidade representativa que possa ser discutida funcionalmente.

A última região em questão não apresentou proximidade a QTL algum da base consultada. No que se refere à investigação de suas sequências, foram identificadas três, ao passo que uma não apresentou qualquer semelhança a uma proteína conhecida. As outras duas mostraram relação com a eficiência energética, sendo uma relativamente similar a uma protease dependente de ATP e outra semelhante à subunidade proteolítica da protease Clp dependente de ATP. Essa é uma enzima de ação na mitocôndria vegetal, com relações bem definidas de ancestralidade com bactérias (JANSKA, 2005), além de também estar relacionada com cloroplastos, controlando atividades metabólicas importantes nesses plastídios, notoriamente a fotossíntese (SJÖGREN *et al*., 2006).

O QTL Q.Ming2002a.PTHT.32_2-1, também relativo à altura da planta, abrangeu a extensão de 44.970.786 a 74.441.160 pb, na qual as duas regiões restantes estavam contidas. Em relação à menor região, foram identificadas sete sequências, das quais uma apresentou possível homologia com uma proteína não caracterizada e outra apenas apresenta similaridade à proteína *Streptococcal hemagglutinin*, cuja homologia pode ser de uma fitohemaglutinina, considerando que as lectinas se ligam a carboidratos, podendo ser relavantes para a resistência a patógenos. Das cinco sequências restantes, duas associam-se a à dormência de sementes,

sendo uma dessas relativa a uma proteína com repetições polimórficas ricas em leucina PLRR-4, com poucas informações na literatura, assim como uma outra sequência. É importante frisar que essa tipologia proteica possui atuação diversa, com exemplos de resistência a patógenos e importância durante a floração (JONES & JONES, 1997; TORTI *et al.*, 2012). Por fim, uma única sequência mostrou-se similar à proteína quinase 1-PTI1, associada à resistência bacteriana no trabalho de Zhou *et al.* (1995) com tomateiro.

Finalmente, a região entre 71.780.001 e 71.840.000 pb do cromossomo três conteve nove sequências, sendo uma similar a uma proteína não caracterizada. Uma sequência apresentou semelhança com a proteína ubiquitina-ligase E3, que desempenha um papel importante na marcação e eliminação de proteínas defeituosas (ARDLEY & ROBINSON, 2005), e duas tiverem semelhança a proteínas não totalmente descritas relacionadas à ligação de membranas de vesículas do citoplasma. Duas sequências codantes foram relacionadas à glucoronosiltransferase, atuando na cadeia biossintética do ácido glocurônico, que, segundo Douglas & King (1952), é um constituinte dos polissacarídeos.

Ainda, uma sequência foi semelhante à proteína Sec 14, que foi amplamente caracterizada no trabalho de Saito *et al.* (2007), descrevendo-a como um componente de importante papel no metabolismo de fosfolipídeos. Dessa maneira, essa possui papel de auxílio em fatores como a marcação proteica, a transdução de sinal, a biossíntese, o transporte e o metabolismo de lipídeos, além de sua importância na preservação das membranas e compartimentos celulares. Paterman *et al.* (2004) descreveram anteriormente sua importância junto ao conjunto de proteínas do complexo de Golgi no transporte intermembranar de moléculas. Há, ainda, descrições relativas a seu papel na floração, como observado no trabalho de Upadhyaya *et al.* (2015).

Uma sequência apresentou similaridade à proteína glutationa S-transferase, cujas funções incluem a conjugação e desintoxicação de herbicidas (Dixon *et al.*, 2002). Tal resultado mostra-se possivelmente interessante do ponto de vista agronômico, sendo relevantes investigações mais aprofundadas futuramente.

Uma última sequência relacionou-se à omega-hidroxipalmitato O-feruloil transferase, afetando especificamente o acúmulo de ferulato na suberina de raízes e sementes e a biossíntese de alcoóis graxos, que afetam a translocação de solutos nesses órgãos (YADAV *et al.*, 2014).

5.4.5 Cromossomo 4

O cromossomo quatro apresentou apenas uma região com indício de varredura seletiva, que compreende a extensão da posição 6.560.001 à posição 6.600.000 pb. Seus 16 SNPs identificados foram todos de efeito modificador e não foi encontrada evidência de QTL na região.

Cinco sequências foram identificadas, sendo que todas mostraram-se possivelmente homólogas a uma proteína conhecida. Um transcrito foi similar à proteína de transporte de cátions ATPase, relacionada ao transporte energético. Um outro relacionou-se à poligalacturonase, envolvida na síntese de pectina, que desempenha um papel importante na planta em relação à formação da parede celular, referindo-se à proteção da planta e ao porte e à ereção do caule (MICHELI, 2001). Ainda, duas sequências alinharam-se à proteína transportadora de beta-cetoacil-acil sintase I, participando intensamente da síntese de ácidos graxos (GARWIN *et al.*, 1980). Por fim, uma última sequência mostrou provável homologia com a fumarilacetoacetase, que, como exposto por Han *et al.* (2013), está envolvida na via de degradação da tirosina.

5.4.6 Cromossomo 5

Foram encontradas duas regiões com evidência de assinatura de seleção nesse cromossomo, sendo uma de 40 Kpb, da posição 620.001 à posição 660.000 pb, e outra de 60 Kpb, entre as posições 2.300.001 e 2.360.000. Na menor região foram identificados 57 SNPs, dos quais 68,4% de efeito modificador e igualmente 15,8% de efeitos baixo e moderado, de forma que seu índice ω foi igual a 1,0. É importante salientar que a quantidade de mutações sinônimas e não-sinônimas foi reduzida, podendo ter prejudicado a aferição da taxa em questão. Ademais, essa relação é apenas um indício e, embora nenhum QTL tenha sido evidenciado, foram encontradas oito sequências genômicas, cuja importância está descrita a seguir.

Primeiramente, uma das sequências foi semelhante a uma proteína contendo repetições ricas em leucina, cuja função em plantas foi associada à resistência a patógenos e à floração, como já descrito. Uma segunda sequência foi similar à proteína fosfatase 2c, família que atua na transdução de sinal, por exemplo pela via do ácido abscísico (SHEEN, 1998; MEYER *et al.*, 1994). Funções mais gerais foram relatadas para dois transcritos. Uma sequência alinhou-se à proteína aldo-ceto redutase, que, segundo Jez *et al.* (1997), está associada ao

NAD/NADP, envolvido no Ciclo de Krebs. Outra relacionou-se à proteína LSM36B, encontrada em algodão, cuja função é relativa ao processamento de RNA (GOLISZ *et al.*, 2013). Por fim, três sequências alinharam-se a sequências não identificadas e uma à proteína com repetições de pentatricopeptídeos dPPR-U8g2, cuja família possui grande espectro de atuação durante a germinação e o desenvolvimento de vegetais, atuando em aspectos como a expressão gênica e tradução extranucleares (BARKAN, A. & SMALL, I., 2014).

Por sua vez, na maior região foram identificados apenas 19 SNPs, sendo 14 de efeito modificador e cinco de efeito baixo. Não foi detectado QTL algum próximo à região e foram encontrados quatro sequências, sendo três não alinhadas a proteínas descritas e uma similar à hidroximetil glutaril-CoA liase, envolvida na síntese de fitoesteroides, com importância na formação de membranas e, consequentemente, no crescimento celular e em mecanismos de defesa da planta (ALEX *et al*., 2000; VAN DER HEIJDEN, 1994).

5.4.7 Cromossomo 7

Duas regiões distantes, entre as posições 120.001 e 160.000 e 59.580.001 e 59.620.000 pb do cromossomo em questão, foram significativas. Foram encontradas 45 mutações na primeira região e 105 na segunda região, sendo 80,0% e 84,8% de efeito modificador, respectivamente, com o adendo de que duas alterações de efeito alto foram detectadas na última. Ambas regiões apresentaram evidência preliminar de seleção purificadora, com taxas respectivas ω de 0,5 e 0,4. Nenhum QTL próximo foi identificado.

As duas localidades genômicas mostraram-se relacionadas a processos de metabolismo celular. Em relação à primeira, foram identificadas cinco sequências, sendo duas referentes a proteínas não identificadas e as três demais ao sistema membranar. Uma primeira sequência foi similar a uma proteína da família sinaptobrevina, que, segundo Edamatsu & Toyoshima (2003), atua essencialmente no transporte vesicular. Outra sequência mostrou-se semelhante à proteína *casparian strip* II (CASPII), cujo complexo é relativo à deposição de parede celular (ROPPOLO *et al.*, 2011). Finalmente, uma última foi associada à proteína do grupo *Harpin-induced* 1 (HIN1), envolvida na inibição de ATP e/ou na morte celular programada, como demonstrado por Xie & Chen (2000) em tabaco.

A segunda região conteve onze sequências alinhadas com genes, com seis não associadas a proteínas descritas e uma alinhada a uma proteína não descrita, tida como associada à membrana peroxissomal. Dos demais, uma sequência mostrou-se similar à proteína de *manutenção estrutural de cromossomos 1* (SMC1), de importância na coesão

entre os cromossomos e na formação de seus centrômeros em plantas (LAM *et al.*, 2005; STEINER & HENIKOFF, 2015), enquanto outra revelou similaridade com proteína relativa à atividade de metiltransferase das histonas na eucromatina, estando, portanto, conectada à expressão gênica (RICE *et al.*, 2003). Também foram presentes uma sequência similar à proteína *proton gradient regulation* 5 (PGR5) e uma que correspondeu à proteína da grupo de carboidrato quinase PfkB. A primeira desempenha, segundo Nandha *et al.* (2005), papel redox no transporte de elétrons, auxiliando a preservar a fotossíntese, inclusive com relatos em cana-de-açúcar presentes no trabalho de Andrade *et al.* (2010). A segunda, embora não descrita para cana-de-açúcar, possui funções bem conhecidas para outras plantas, como no trabalho de Gilkerson *et al*. (2012). Nesse há a descrição de importantes contribuições dessa família proteica para expressão de genes cloroplásticos e de outros plastídeos semelhantes. Também se evidencia importância na reciclagem de adenilato/ciclase e metil, com contribuição no crescimento e desenvolvimento do vegetal, além de fatores correlatos, como o geotropismo.

5.4.8 Cromossomo 8

Foram encontradas três regiões com evidência de seleção no cromossomo oito, nas janelas entre 480.001 e 520.000, 2.620.001 e 2.660.000 e entre 54.260.001 e 54.300.000 pb. Na primeira, foram identificadas 41 mutações, sendo 82,9% de efeito modificador e apenas uma de efeito alto. Embora pela baixa frequência de SNPs sob as classificações sinônima e não-sinônima (seis no total), a taxa ω foi de 0,2 (seleção negativa).

Dos nove genes candidatos identificados nessa região, cinco não se alinharam a qualquer proteína caracterizada. Duas sequências associaram-se à família das quinases, havendo uma relacionada ao processo de fosforilação, que é um processo importante na regulação da atividade de outras proteínas, como já discutido. A segunda sequência relaciona-se à quinase de riboflavina, com conexão com a reação de fosforilação oxidativa. Uma sequência apresentou homologia contra *trichome birefringence*, que, segundo Bischoff *et al.* (2010), relaciona-se à síntese de celulose e formação da parede celular. Por fim, uma última sequência associou-se a uma proteína nodulina, cuja função para plantas não-noduladoras, como a cana e outras gramíneas, está relacionada ao transporte de nutrientes, solutos, aminoácidos ou hormônios e à maioria dos aspectos do desenvolvimento da planta, com destaque para situações de estresse nutricional (DENANCÉ *et al.*, 2014).

Por sua vez, a segunda região em questão apresentou dificuldades para uma análise mais acurada. Foram identificadas 52 mutações, sendo 51 de efeito modificador e uma de

efeito baixo, não havendo SNPs não-sinônimos e, consequentemente, impossibilitando o cálculo do parâmetro ω. Não houve qualquer evidência de QTL nessa região e, apesar de seis sequências identificadas, cinco não puderam ser anotadas. Contudo, a única sequência que apresentou alinhamento a uma proteína caracterizada associou-se à polifenoloxidase cloroplastídica (PPO), de importância em muitos vegetais. Seu papel é especialmente relevante para características agronômicas, por questões de toxidez e escurecimento dos órgãos comestíveis, mas também na proteção contra patógenos (YORUK & MARSHAL, 2007; LEITE *et al.*, 2013). Em cana-de-açúcar, resultados de Qudiseh *et al.* (2002) demonstraram ação da PPO no decréscimo dos conteúdos totais de tanino e clorofilas *a* e *b* durante a maturidade fisiológica, assim como ação da enzima no processo de escurecimento do mosto.

Finalmente, apesar da quantidade de mutações (80) encontradas na última região, sendo que menos de 72% eram de efeito modificador, a taxa ω foi de 1,1, não sendo indicativo de seleção. Ademais, não foi encontrada evidência para QTL algum em relação à base utilizada e só foram detectados dois genes candidatos, ambos alinhados a proteínas não descritas. Relevante destacar que o valor de *Hp* desta região foi de, aproximadamente, 0,38, o mais alto entre os valores significativos, e seu *q-valor* de 0,031, inferior ao de outras cinco regiões aqui analisadas e que não apresentaram resultados semelhantes. Frisa-se que a correção empregada foi moderadamente conservativa, de forma que mais investigações dessa região genômica mostram-se oportunas.

5.4.9 Cromossomo 9

No cromossomo nove apenas uma região com indício de varredura seletiva foi detectada, entre as posições 57.280.001 e 57.320.000 pb. Foram observadas 38 mutações, com 84,2% de efeito modificador e havendo uma de efeito alto. O parâmetro ω foi igual a 0,67, calculado a partir de três e dois SNPs dos tipos sinônimo e não-sinônimo, respectivamente. De maneira a exemplificar a varredura seletiva que essa e outras regiões sofreram, apresenta-se uma figura com o perfil de diversidade nucleotídica, a qual inclui a região detectada (Figura 13).

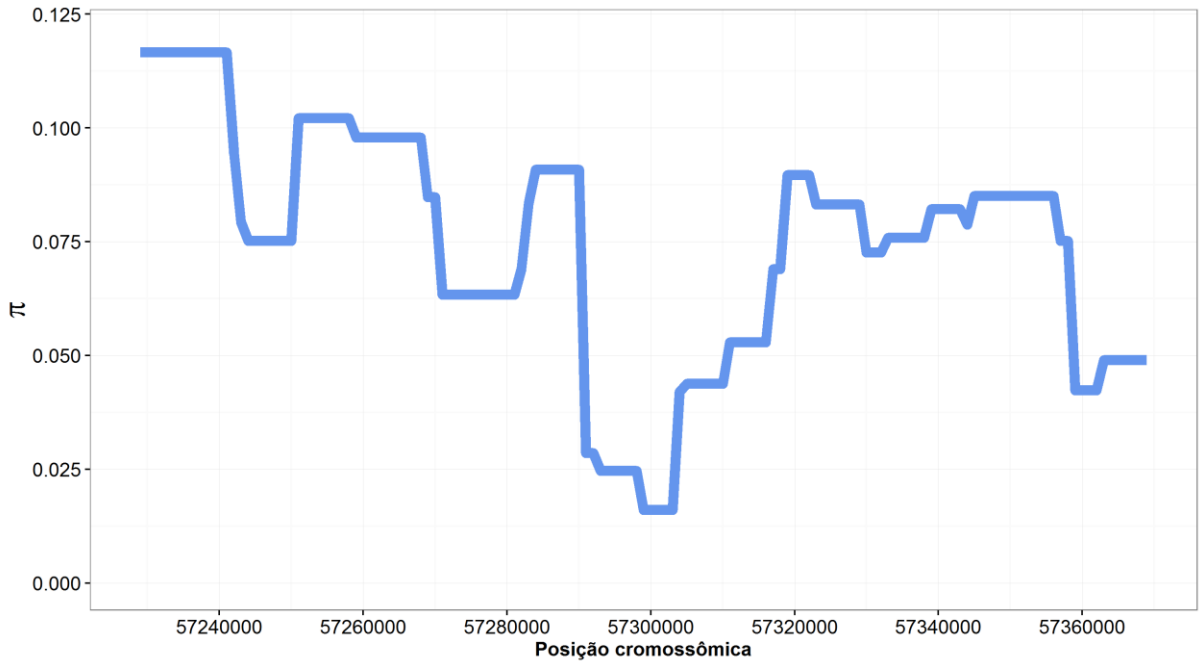


FIGURA 13 – Perfil de diversidade nucleotídica dos polimorfismos detectados no Painel Brasileiro de Genótipos de Cana-de-Açúcar, para a região entre as posições 57.229.001 e 57.369.000 pb do cromossomo nove do sorgo. Os polimorfismos foram ancorados na versão 2.1 do genoma de *Sorghum bicolor*).

Sete evidências gênicas foram encontradas, das quais quatro puderam ser alinhadas a sequências já anotadas, e sinal de quatro QTLs. Um dos QTLs foi Q.Ming2002a.PLHT.5.2_2-1, indicado para altura da planta e predito entre as posições 58.649.042 e 58.651.214 pb. Duas sequências alinharam-se a proteínas da subunidade ribossomal 60S (L18 e L18a), que, segundo Lo & Johnson (2009), constitui a partícula 80S nos eucariotos. Sua presença na ponta aminoacil conclui a síntese de proteínas traduzindo a cadeia de mRNA em cadeia peptídica com a saída do RNA transportador (tRNA) livre. Também, uma sequência associou-se à proteína G acessória da urease, com atuação em sua síntese, sendo essa enzima responsável pela reciclagem de nitrogênio proveniente do catabolismo de ureídeo, purina e arginina em plantas (WITTE *et al*., 2005).

Por sua vez, os outros três QTLs identificados, Q.Ming2001b.ISUGARCN.5.2-1 (também de 58.649.042 a 58.651.214 pb), Q.Ming2001b.DSUGARCN.4-1 (idem) e Q.Ming2001b.DSUGARCN.5.1-1 (ibidem), foram associados a variações no acúmulo de açúcar. Esse resultado é reforçado pela relação da outra sequência com a enzima sacarose-fosfato sintase. A reação catalisada por esta enzima é de grande importância na biossíntese de

sacarose, como atesta o trabalho de Verma *et al.* (2011) em cana-de-açúcar. Discute-se, pois, que a sacarose é um açúcar de translocação para a gramínea em questão, acumulando-se no colmo sob situação de estresse e sua extração é o objetivo principal da cultura canavieira, sendo utilizada diretamente ou indiretamente (MOORE, 1987).

5.4.10 Cromossomo 10

No cromossomo dez apenas uma região de 60 Kpb, da posição 54.820.001 à posição 54.880.000 pb, foi identificada como potencial alvo de seleção. Foram encontradas 15 mutações, embora todas de efeito modificador, não havendo alguma de efeito sinônimo ou não sinônimo, impossibilitando o cálculo da estimativa ω. Um QTL foi detectado proximamente, embora sem intersecções com a região e, considerando que seu intervalo era extremamente amplo (de 1 a 52.482.087 pb), abrangendo quase todo o cromossomo, torna-se difícil associá-lo funcionalmente à janela detectada.

Entre os dois genes candidatos encontrados, uma sequência foi vinculada a uma proteína putativa ligada à resistência a doenças. A outra sequência assemelhou-se à proteína H^+-pirofosfatase vacuolar. Raza *et al.* (2016), estudando a expressão heteróloga do gene de *Arabidopsis* que codifica essa enzima na cana-de-açúcar, verificaram efeitos de resposta fotossintética, crescimento da planta e resistência à seca.

5.4.11 Visão geral das assinaturas de seleção

Em relação à magnitude dos efeitos preditos dos polimorfismos identificados nas 15 regiões, houve uma predominância do modificador (80,4%) perante os demais e uma baixíssima ocorrência de efeito alto (0,01%). Esses resultados eram esperados, primeiramente pela grande quantidade de classes de polimorfismo que abarcam o tipo modificador. Em segundo lugar, pela tendência notada de seleção negativa em regiões gênicas, os efeitos altos foram bastante minimizados por questões já expostas anteriormente. Igualmente, esperava-se uma tendência à seleção negativa nas regiões como um todo, o que de fato ocorreu, com uma taxa aproximada ω de 0,65. Também se nota a completa ausência de mutações sem sentido em todas as regiões.

É lúcido dizer que um resultado de interesse agronomicamente foi o relativo à região entre as posições de 57.280.001 e 57.320.000 do cromossomo nove, em que tanto a base de QTLs, como a sequência de um transcrito relacionaram-se ao acúmulo de sacarose no colmo,

de relevância para a cultura canavieira. Os resultados de resistência a doenças também merecem destaque pelo fato enunciado de o melhoramento genético da gramínea ter sido iniciado, e por muito tempo basicamente focado, nessa característica.

6 CONCLUSÕES

Como resultado proeminente, a técnica de GBS foi bem sucedida em encontrar regiões genômicas de interesse. O uso do genoma do sorgo como referência para aferições funcionais dos polimorfismos também se delineou como solução dentro das possibilidades existentes para esse poliploide complexo. Adicionalmente, observou-se elevada proporção de polimorfismos de baixa frequência alélica, potencialmente como consequência da natureza poliploide da cana-de-açúcar, bem como da predominância de genótipos de melhoramento recente.

Há indícios da intensidade de um processo seletivo ser influenciada pela severidade de uma mutação. Além disso, mutações de efeito mais drástico tendem a sofrer seleção negativa. Foram observados indicativos desse fenômeno tanto para polimorfismos individuais, com menor frequência dos variantes de efeito supostamente mais agressivo, quanto para SNPs anotados em diferentes posições nos modelos gênicos preditos. O teste de McDonald & Kreitman forneceu indícios de seleção purificadora, com destaque para genes potencialmente relacionados à sobrevivência da planta ou a caracteres de interesse agronômico, embora outras análises sejam indicadas para confirmação desse fato. No genoma como um todo, a partir dos polimorfismos fixados, foi detectada possível evidência de seleção negativa, cuja conclusão é limitada pelo entendimento da metodologia de GBS ter favorecido a amostragem de determinadas regiões genômicas.

Foram detectados 22 termos funcionais da ontologia gênica e 15 regiões com evidência de seleção, ressaltando-se que os resultados são moderadamente conservativos e novas investigações com outras abordagens fazem-se proveitosas. Em linhas gerais, mostraram-se possivelmente relevantes os resultados de componentes de membranas de organelas celulares e seus constituintes (incluindo suas vias de comunicação bioquímica), integrantes de rotas energéticas (como fotossíntese e respiração), constutuintes da parede celular e processos moleculares relativos à transcrição e à oxirredução. Mais especificamente, foram detectadas regiões envolvidas com caracteres de altura da planta (evidência de QTLs), mecanismos de resistência a patógenos e a biossíntese de esteroides e carboidratos, em particular a sacarose.

Tratando-se da cultura da cana-de-açúcar, os resultados são motivadores de programas de seleção assistida. Mais essencialmente, há um destaque para as características de acúmulo de sacarose no colmo e resistência a doenças. No primeiro caso há uma relação com a produtividade da cultura. O segundo caso, por sua vez, é relevante porque considera o restrito

rol de alternativas de controle químico contra patologias em canaviais, bem como o fato que o melhoramento genético da cultura concentrou-se fortemente nesse quesito durante décadas. Também é relevante dizer que esse estudo tem potencial de auxiliar a escolha de genitores de cruzamentos em programas de melhoramento genético, considerando a relatada dificuldade de formar grupos heteróticos verdadeiros.

Em suma, esse trabalho avaliou possíveis consequências de processos evolutivos em regiões polimórficas ao longo do melhoramento genético da gramínea com o uso da genotipagem-por-sequenciamento, caracterizando a variabilidade dos genótipos. Seus resultados são valorosos sob o ponto de vista de melhoristas e de geneticistas de populações, considerando as novas tecnologias em estudos genéticos com plantas cultivadas. Nesse contexto, encoraja-se a continuidade de investigações genômicas desse poliploide e seus possíveis desdobramentos científicos e tecnológicos.

REFERÊNCIAS

AHLOOWALIA,B.S.; MALUSZYNSKI,M.; NICHTERLEIN,K. Global impact of mutation-derived varieties. **Euphytica**, New York, v. 135, p. 187-204, 2004.

ALEX,D.; BACH,T.J.; CHYE,M.L. Expression of *Brassica juncea* 3-hydroxy-3-methylglutaryl CoA synthase is developmentally regulated and stress-responsive. **Plant Journal**, London, v. 22, p. 415-426, 2000.

ANDRADE,J.C.F.; ANDRADE,T.G.; SILVA,J.V.; CAETANO,L.C.; ALMEIDA,C.C.S. **Análise de expressão gênica diferencial em genótipo de cana-de-açúcar tolerante ao estresse hídrico usando RT-PCR**. 2010. 59 p. Dissertação (Mestrado em Agronomia – Produção Vegetal) – Centro de Ciências Agrárias, Universidade Federal de Alagoas, Maceió. 2010.

ARDLEY,H.C.; ROBINSON,P.A. E3 ubiquitin ligases. **Essays In Biochemistry**, London, v. 41, p. 15-30, 2005.

BARRETO,F.Z. **Caracterização fenotípica e molecular do Painel Brasileiro de Genótipos de Cana-de-açúcar.** 2016. 78 p. Dissertação (Mestrado em Produção Vegetal e Bioprocessos) – Centro de Ciências Agrárias, Universidade Federal de São Carlos, Araras, SP, 2016.

BARKAN,A.; SMALL,I. Pentatricopeptide repeat proteins in plants. **Annual Review of Plant Biology**, London, v. 65, 415-442, 2014.

BARROS,G.S.C.; ADAMI,A.C.O.; ZANDONÁ,N.F. **Faturamento e volume exportado do agronegócio brasileiro são recordes em 2013**. Piracicaba, SP: Centro de Estudos Avançados em Economia Aplicada, Universidade de São Paulo, 2014. 10 p.

BESSE,P.; MCINTYRE,C.L.; BERDING,N. Characterisation of *Erianthus* sect. *Ripidium* and *Saccharum* germplasm (Andropogoneae – Saccharinae) using RFLP markers. **Euphytica**, New York, v.93, p. 283-292, 1997.

BISCHOFF,V.; NITA,S.; NEUMETZLER,L.; SCHINDELASH,D.; URBAIN,A.; ESHED,R.; PERSSON,S.; DELMER,D.; SCHEIBLE,W.R. Thicome birefringence and its homolog *AT5G01360* encode plant-specific DUF231 proteins required for cellulose biosynthesis in *Arabidopsis*. **Plant Physiology**, Rockville, USA, v. 153, p. 590-602, 2010.

BLACKBURN,F. **Sugar-cane**. London: Longman, 1984. 414 p.

BOMBLIES,K.; DOEBLEY,J. Molecular Evolution of FLORICAULA/LEAFY Orthologs in the Andropogoneae (Poaceae). **Molecular Biology and Evolution**, Oxford, v. 22, p. 1082-1094, 2016.

BOS,I; CALIGARI,P. **Selection methods in plant breeding**. San Diego: Chapman & Hall, 1995. 347 p.

BULL,J.K.; HOGARTH,D.M.; BASFORD,K.E. Impact of genotype x environment interactions on response to selection in sugarcane. **Australian Journal of Experimental Agriculture**, Adelaide, AUS, v. 32, p. 731-737, 1992.

CHARLESWORTH,J.; WALKER,A.E. The McDonald–Kreitman Test and slightly deleterious mutations. **Molecular Biology and Evolution**, Oxford, v. 25, p. 1007-1015, 2008.

CINGOLANI,P.; PLATTS,A.; WANG,L.L.; COON,M.; NGUYEN,T.; WANG,L.; LAND,S.J.; LU,X.; RUDEN,D.M. SnpEff manual. Available in <http://snpeff.sourceforge.net/SnpEff_manual.html>. Acesso em: 6 abril 2016.

CINGOLANI,P.; PLATTS,A.; WANG,L.L.; COON,M.; NGUYEN,T.; WANG,L.; LAND,S.J.; LU,X.; RUDEN,D.M. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w 1118 ; iso-2; iso-3. **Fly**, Abingdon, UK, v. 6, p. 80-92, 2012.

CLAYTON,W. D.; DANIELS,C.A. Geographical, historical and cultural aspects of origin of the Indian and Chinese sugarcane *S. barberi* e *S. sinesis*. **ISSCT Sugarcane Breed**, Queensland, v. 36, p. 4-23, 1975.

COLLARD,B.C.Y.; JAHUFER,M.Z.Z.; BROWER,J.B.; PANG, E.C.K. An introdution to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: the basic concepts. **Euphytica**, New York, v. 142, p. 169-196, 2005.

CONWAY,G. **The doubly Green Revolution**: food for all in the 21st. century. Ithaca, USA: Cornell University Press, 1997. 84 p.

COSTA,C. Primeiras canas e primeiros açúcares no Brasil. **Brasil Açucareiro**, Rio de Janeiro, v. 3, p. 160-168, 1958.

DANECEK,P.; AUTON,A.; ABECASIS,C.; ALBERTS,C.A.; BANKS,E.; DEPRISTO,M.A.; HANDSAKER,R.E.; LUNTER,G.; MARTH,G.T.; SHERRY,S.T.; MCVEAN,G.; DURBIN,R. The variant call format and VCFtools. **Bioinformatics**, Oxford, v. 27, p. 2156-2158, 2011.

DANIELS,J.; SIMITH,P.; PATON,N. The origin of sugarcane and centers of genetic diversity in *Saccharum*. **Sugarcane Breeding Newsletter**, Queensland, v. 35, p.4-18, 1975.

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

DAVIES,G.; HENRISSAT,B. Structures and mechanisms of glycosyl hydrolases. **Structure**, Amsterdam, v. 3, p. 853-859, 1995.

DAVIES,P.J. **Plant Hormones**. New York: Springer, 2010. 348 p.

DENANCÉ,N.; SZUREK,B.; NOËL,L.D. Emerging functions of nodulin-like proteins in non-nodulating plant species. **Plant Cell Physiology**, Oxford, v. 55, p. 469-474, 2014.

DEPRISTO,M.; BANKS,E.; POPLIN,R.; GARIMELLA,K.V.; MAGUIRE,J.R.; HARTL,C.; PHILIPPAKIS,A.A.; ANGEL,G.; RIVAS,M.A.; HANNA,M.; MCKENNA,A.; FENNELL,T.J.; KERNYTSKY,A.M.; SIVACHENCO,A.Y.; CIBULSKIS,K.; GABRIEL,S.B.; ALTSHULER,D.; DALY,M.J. A framework for variation discovery and genotyping using next-generation DNA sequencing data. **Nature Genetics**, New York, v. 43, p. 491-498, 2011.

DILLON,S.K.; NOLAN,M.F.; WU,H.; SOUTHERTON,S.G. Association genetics reveals candidate gene SNPs affecting wood properties in *Pinus radiata*. **Australian Forestry**, Adelaide, AUS, v. 73, p. 185-190, 2010.

DIXON,D.P.; LAPTHORN,A.; EDWARDS,R. Plant glutatione transferases. **Genome Biology**, Londres, v. 3, p. 3004.1-3004.2, 2002.

DOUGLAS,J.F.; KING,C.G. The metabolism of uniformly labeled D-glucoric acid in the Guinea pig. **Journal of Biological Chemistry**, Rockville, USA, v. 198, p. 187-194, 1952.

EATHINGTON,S.R.; CROSBIE,T.M.; EDWARDS,M.D.; REITER,R.S.; BULL,J.K. Molecular Markers in a commercial breeding program. **Crop Science Society of America**, Madison, USA, v. 47, p. 154-163, 2007.

EDAMATSU,M.; TOYOSHIMA,Y.Y. Fission yeast synaptobrevin is involved in cytokinesis and cell elongation **Biochemical and Biophysical Research Communications**, Amsterdam, v. 30, p. 641-645, 2003.

ELSHIRE,R. J.; GLABUBITZ,J. C.; POLAND,J. A.; KAWAMOTO,K.; BUCKLER,E. S.; MITCHELL,S. E. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. **PLoS ONE**, San Francisco, v. 6, e19379, 2011.

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA. **Clima para cana-de-açúcar** (2000). Disponível em <https://www.agencia.cnptia.embrapa.br/Repositorio/clima_para_cana_000fhc5hpr702wyiv80efhb2aul9pfw4.pdf>. Acesso em: 4 abril 2016.

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA. **Cana-de-açúcar:** plantas daninhas. Disponível em <http://www.agencia.cnptia.embrapa.br/gestor/cana-de-acucar/arvore/CONTAG01_52_711200516718.html>. Acesso em: 6 abril 2016.

ERNEST,F. The difference between spontaneous and base-analogue induced mutations of phage T4. **Proceedings of National Academy of Sciences of the United States of America**, Washington, D.C, v. 45, p. 622-633, 1959a.

ERNEST,F. The specific mutagenic effect of base analogues on phage T4. **Journal of Molecular Biology**, Amsterdam, v. 1, p. 87-105, 1959b.

FIGUEIREDO,P. Breve história da cana-de-açúcar e o papel do Instituto Agrônomico no seu estabelecimento no Brasil. In: MIRANDA,L.L.D.; VASCONCELOS,A.C.M.; LANDELL,M.G.A. (Ed.). **Cana-de-açúcar**. Campinas, SP: Instituto Agronômico, 2008. p. 31-44.

FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS. **Statistics at FAO.** Disponível em <http://www.fao.org/statistics/en/>. Acesso em: 4 abril 2016.

FRANÇA,S.C.; ROBERTO,P.G.; MARINS,M.A.; PUGA,R.D.; RODRIGUES,A.; PEREIRA,J.O. Biosynthesis of secondary metabolites in sugarcane. **Genetics and Molecular Biology**, Ribeirão Preto, SP, v. 24, p. 243-450, 2001.

FREEBAYES. Disponível em <https://github.com/ekg/freebayes>. Acesso em: 11 maio 2016.

GALLO,D.; NAKANO,O.; SILVEIRA NETO,S.; CARVALHO,R.P.L.; BAPTISTA,G.C.; BERTI FILHO,E.; PARRA,J.R.P.; ZUCCHI,R.A.; ALVES,S.B.; VENDRAMIM,J.D.; MARCHINI,L.C.; LOPES,J.R.S.; OMOTO,C. Pragas das plantas e seu controle. In: NAKANO, O. (Org.). **Entomologia Agrícola**. Piracicaba, SP: Fundação de Estudos Agrários "Luiz de Queiroz", 2002. p. 397-898.

GARCIA,A.A.F.; MOLLINARI,M.; MARCONI,T.G.; SERANG,O.R.; SILVA,R.R.; VIEIRA,M.L.C.; VICENTINI,R.; COSTA,E.A.; MANCINI,M.C.; GARCIA,M.O.S.; PASTINA,M.M.; GAZAFFI,R.; MARTINS,E.R.F.; DAHMER,N.; SFORÇA,D.A.; SILVA,C.B.C.; BUNDOCK,P.; HENRY,R.J.; SOUZA,G.M.; SLUYS,M.A.; LANDELL,M.G.A.; CARNEIRO,M.S.; VINCETZ,M.A.G.; PINTO,L.R.; VENCOVSKY,R; SOUZA,A.P. SNP genotyping allows na in-depth characterisation of the genome of the sugarcane and other complex autopolyploids. **Scientific Reports**, New York, v. 3, e3399, 2013.

GARRISON,E.; MARTH,G. Haplotype-based variant decetion from short-read sequencing. **arXiv**, Ithaca, USA, v. 2, arXiv:1207.3907, 2012.

GARWIN,J.L.; KLAGES,A.L.; CRONAN JR.,J.E. Structural, enzymatic, and genetic studies of beta-ketoacyl-acyl carrier protein synthases I and II of Escherichia coli. **Journal of Biological Chemistry**, Rockville, USA, v. 255, p. 11949-11956, 1980.

GENE ONTOLOGY CONSORTIUM. The Gene Ontology (GO) database and informatics resource. **Nucleic Acids Research**, Oxford, v. 32, p. D258-D261, 2004.

GHEYAS,A.A.; BOSCHIERO,C.; EORY,L.; RALPH,H.; KUO,R.; WOOLLIAMS,J.A.; BURT,D.W. Functional classification of 15 million SNPs detected from diverse chicken populations. **DNA Research**, Oxford, v. 22, p. 1-13, 2015.

GIBSON,N.J. The use of real-time PCR methods in DNA sequence variation analysis. **Clinica Chimica Acta**, Amsterdam, v. 363, p. 32-47, 2006.

GILKERSON,J.; PEREZ-RUIZ,J.M.; CHORY,J.; CALLIS,J. The plastid-localized pfkB-type carbohydrate kinases frutokinase-like 1 and 2 are essential for growth and development of *Arabidopsis thaliana*. **BMC Plant Biology**, London, DOI: 10.1186/1471-2229-12-102, 2012.

GOLDEMBERG,J.; COELHO,S.T.; GUARDABASSI,P. The sustentability of ethanol production from sugarcane. **Energy Policy**, Amsterdam, v. 36, p. 2086-2097, 2008.

GOLISZ,A., SIKORSKI,P.J., KRUSZKA,K., KUFEL,J. Arabidopsis thaliana LSM proteins function in mRNA splicing and degradation. **Nucleic Acids Research**, Oxford, v. 41, p. 6232-6249, 2013.

GOODSTEIN,D.M.; SHU,S.; HOWSON,R.; NEUPANE,R.; HAYES,D.R.; 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, D1178-D1186, 2012.

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

GUARINO,N. **Formal ontology in information systems**. Amsterdam: IOS Press, 341 p., 1998.

HARA,K.; WATABE,H.; SASAZAKI,S.; MUKAI,F.; MANNEN,H. Development of SNP markers for individual indetification and parentage test in Japonese Black cattle population. **Animal Science Journal**, Tokyo, v.18, p. 152-157, 2010.

HAN,C.; REN,C.; ZHI,T.; ZHOU,Z.; LIU,Y.; CHEN,F.; PENG,W.; XIE,D. Disruption of fumarylacetoacetate hydrolase causes spontaneous cell death under short-day conditions in Arabidopsis. **Plant Physiology**, Rockville, USA, v. 162, p. 1956-1964, 2013.

HARLAN,J.R. **Crops and man**. Madison, USA: American Society of Agronomy, 1975. 306 p.

HEFTMANN,E. Functions of steroids in plants. **Phytochemistry**, Amsterdam, v. 14, p. 891-901, 1975.

HENRISSAT,B.; TEERI,T.T.; WARREN,R.A.J. A scheme for designating enzymes that hydrolyse the polyssacarides in the cell wall of plantas. **Federation of European Biochemical Societies Letters**, Zurich, v. 425, p. 352-354, 1998.

HODGES,M. Enzyme redundancy and the importance of 2-oxoglutarate in plant ammonium assimilation. **Journal of Experimental Botany**, Oxford, v. 53, p. 905-916, 2002.

HUNSIGI,G. Production of sugarcane: theory and practice. **Advanced Series in Agricultural Science**, New York, v. 21, 244 p., 1993.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Censo Agropecuário 2006.** Disponível em <http://www.ibge.gov.br/home/estatistica/economia/agropecuaria/censoagro/default.shtm>. Acesso em: 04 abril 2016.

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**, London, v. 50, p. 574-585, 2007.

JANSKA,H. ATP-dependent proteases in plant mitochondria: What do we know about them today? **Physiologia Plantarum**, London, v. 123, p. 399-405, 2005.

JESWIET,J. The development of selection and breeding of the sugarcane in Java. **Journal of Proceedings of the International Society of Sugarcane Technologists**, Adelaide, AUS, v. 3, p. 44-57, 1930.

KIM, J.M.; AHN, J.H.; LIM, K.S.; LEE, E.A.; CHUN, T.; HONG, K.C. Effects of hydroxyl-delta-5-steroid dehydrogenase, 3 beta- and steroid delta-isomerase 1 polymorfisms of fat androsterone level and gene expression in Duroc pigs. **Animal Genetics**, London, v. 44, p. 592-595, 2013.

KIMURA,M. Evolutionary rate at the molecular level. **Nature**, New York, v. 217, p. 624-626, 1968.

JEZ,J.M.; BENNETT,M.J.; SCHLEGEL,B.P.; LEWIS,M.; PENNING,T.M. Comparative anatomy of the aldo–keto reductase superfamily. **Biochemical Journal**, London, v. 326, p. 625-636, 1997.

JONES,D.A.; JONES,J.D.A. The role of leucine-rich repeat proteins in plant defences. **Advances in Botanical Research**, Amsterdam, v. 24, p. 89–167, 1997.

KELLER,I.; BENSASSON,D.; NICHOLS,R.A. Transition-transversion bias is not universal: a counter example from grasshopper pseudogenes. **PLoS Genetics**, San Francisco, v.3, e22, 2007.

LACIEN,M.; GADAL,P.; HODGES,M. Enzyme redundance and the importance of 2-oxoglutarate in higher plant ammonium assimilation. **Plant Physiology**, Rockville, USA, v. 132, p.817-824, 2000.

LAM,W.S.; YANG,X.; MAKAROFF,C.A. Characterization of *Arabidopsis thaliana* SMC1 and SMC3: evidence that AtSMC3 may function beyond chromosome cohesion. **Journal of Cell Science**, Cambridge, USA, v. 118, p. 3037-3048, 2005.

LEA,P.J.; IRELAND,R.J. Nitrogen metabolism in higher plants. In: SINGH, B.K. (Org.). **Plant Amino Acids**: Biochemistry and Biotechnology, Princeton: Marcel Dekker Inc., 1999. p. 1-47.

LEITE,M.E.; SANTOS,J.B.; RIBEIRO JÚNIOR,P.M.; SOUZA,D.A.; LARA,L.A.C.; RESENDE,M.L.V. Biochemical responses associated with common bean defence against *Sclerotinia sclerotiorum*. **European Journal of Plant Pathology**, Amsterdam, v. 138, p. 391–404, 2014.

LI,H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ***arXiv***, Ithaca, USA, arXiv:1303-3997, 2013.

LO,K.Y.; JOHNSON,A.W. Reengineering ribosome export. **Molecular Biology of the Cell**, Bethesda, USA, v. 20, p. 1545-1554, 2009.

LUIKART,G.; ALLENDORF,F.W.; CORNUET,J.M.; SHERWIN,W.B. Distortion of allele frequency distributions provides a test for recent population bottlenecks. **Journal of Heredity**, Oxford, v. 89, p. 238-247, 1998.

MAQUAT,L.E. The power of point mutations. **Nature Genetics, Rochester, USA, v. 27, p. 5-6, 2001.**

MARIN,F.R.; CARVALHO,G.L. Spatio-temporal variability of sugarcane yield efficiency in the state of São Paulo, Brazil. **Pesquisa Agropecuária Brasileira**, Brasília, v. 47, p. 149-156, 2012.

MAYER,E. Mutation pressure. **The evolutionay synthesis*:*** *Perspectives on the Unification of Biology*. Cambridge, USA: Harvard University Press, p. 21-22, 1998.

MAZOYER,M.; ROUDART,L. **A history of world agriculture:** from de Neolitic age to the current crisis. Oxford: Earthscan Press, 2006. 512 p.

MCDONALD,J.H.; KREITMAN,M. Adaptative protein evolution at the *Adh* locus in *Drosophila*. **Nature**, New York, v. 351, p. 652-654, 1991.

MCKEY,D.; ELIAS,M.; PUJOL,B.; DUPUTIÉ,A. The evolutionary ecology of clonally propagated domesticated plants. **New Phytologist**, London, v. 186, p. 318-332, 2010.

MESSER,P.W.; PETROV,D. Frequent adaptation and the McDonald-Kreitman test. **Proceedings of National Academy of Sciences of the United States of America**, Washington, D.C., v. 110, p. 8615-8620, 2013.

MEYER,K.; LEUBE,M.P.; GRILL,E. A protein phosphatase 2C involved in ABA signal transduction. **Science**, New York, v. 264, p. 1452, 1994.

MICHELI,F. Pectin methylesterases: cell wall enzymes with important roles in plant physiology. **Trends in Plant Science**, Amsterdam, v. 6, p. 414-419, 2001.

MINISTÉRIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO. **Cana-de-açúcar.** Disponível em <http://www.agricultura.gov.br/vegetal/culturas/cana-de-acucar>. Acesso em: 4 abril 2016.

MOLLINARI,M.; SERANG,O. Quantitative SNP genotyping of polyploids with MassARRAY and other platforms. **Methods in Molecular Biology**, New York, v. 1245, p. 215-241, 2014.

MOORE,P.H. Morphology and anatomy. Sugarcane improvement through breeding, **Developments in Crop Science Society of America**, Madison, USA, v. 11, p. 85-142, 1987.

NANDHA,B.; FINAZZI,G.; JOLIOT,P.; HALD,S.; JOHSON,G.N. The role of PGR5 in the redox poising of photosynthetic electron transport. **Bioenergetics**, Amsterdam, v. 1767, p. 1252–1259, 2007.

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. Database resources of the National Center for Biotechnology Information. **Nucleic Acid Research**, Oxford, v. 41, p. D8-D20, 2013.

NEI,M. Selectionism and neutralism in molecular evolution. **Molecular Biology and Evolution**, Oxford, v. 22, p. 2318-2342, 2005.

NEPH,S.; KHEHN,M.S.; REYNOLDS,A.P. BEDOPS: high-performance genomic feature operations. **Bioinformatics**, Oxford, v. 28, p. 1919-1920, 2012.

NIELSEN,R. Statistical tests of selective neutrality in the age of genomics. **Heredity**, New York, v. 86, p. 641-647, 2001.

NIELSEN,R. Molecular signatures of natural selection. **Annual Review of Genetics**, Palo Alto, USA, v. 39, p. 197-218, 2005.

OBENCHAIN,V; LAWRENCE,M; CAREY,V; GOGARTEN,S; SHANNON,P; MORGAN,M. VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. **Bioinformatics**, Oxford, 30, p. 2076-2078, 2014.

OHTA,T. Slightly deleterious mutant substitutions in evolution. **Nature**, New York, v. 246, p. 96-98, 1973.

OLEKSYK,T.K.; SMITH,M.W.; O'BRIEN,S.J. Genome-wide scans for footprints of natural selection. **Philosophical Transactions of the Royal Society**, London, v. 365, p. 185-205, 2010.

PATERMAN,T.K.; YAMINI,M.O.; MCREYNOLDS,L.J.; LUNA,E.J. Patellin1, a Novel Sec14-Like Protein, Localizes to the Cell Plate and Binds Phosphoinositides. **Plant Physiology**, Rockville, USA, v. 136, p. 3080-3094, 2004.

PATERSON,A.H.; BOWERS,J.E.; BUROWA,M.D.; DAYEB, X.; ELSIK,C.G.; JIANG,C.X.; KATSAR,C.S.; LAN,T.H.; LIN,Y.R.; MING,R.; WRIGHT,R,J. Comparative genomics of plant chromosomes. **The Plant Cell**, Rockville, USA, v. 12, p. 1523-1539, 2000.

PATERSON,A.H.; BOWERS,J.E.; BRUGGMANN,R.; DUBCHAK,I.; GRIMWOOD,J.; GUNDLACH,H.; HABERER,G.; HELLSTEN,U.; MILTROS,T.; POLIAKOV,A.; SCHMUTZ,J.; SPANNANG,M.; TANG,H.; WANG,X.; WICKER,T.; BHARTI,A.K.; CHAPMAN,J.; FELTRUS,F.A.; GOWIK,U.; GRIGORIEV,I.V.; LYONS,E.; MAHER,C.A.; MARTINS,M.; NARECHANIA,A.; OTILLAR,R.P.; PENNING,B.W.; SALAMOV,A.A.; WANG,Y.; ZHANG,L.; CARPITA,N.C.; FREELING,M.; GINGLE,A.R.; HASH,C.T.; KELLER,B.; KLEIN,P.; KRESOVICH,S.; MCCANN,M.C.; MING,R.; PETERSON,D.G.; RAHMAN,M.; WARE,D.; WESTHOFF,P.; MAYER,K.F.X.; MESSING,J.; ROKHSAR,D.S. The Sorghum bicolor genome and the diversification of grasses. **Nature**, New York, v. 457, p. 551-556, 2009.

PIPEREDIS,G.; CHRISTOPHER,M.J.; CAROLL,B.J.; BERDING,N.; D'HONT,A. Molecular contribution to selection of intergeneric hybrids between sugarcane and the wild species *Erianthus arundinaceus*. **Genome**, Amsterdam, v. 43, p. 1033-1037, 2000.

PRZEWORSKI,M.; CROOP,G.; WALL,J.D. The signature of positive selection on standing genetic variation. **Evolution**, Oxford, v. 59, p. 2312-2323, 2005.

QUDSIEH,H.Y.M.; YUSOF,S.; OSMAN,A.; RAHMAN,R.A. Effect of maturity on chlorophyll, tannin, color, and polyphenol oxidase (PPO) activity of sugarcane juice (*Saccharum officinarum*). **Journal of Agricultural and Food Chemistry**, Washington, D.C., v. 50, p. 1615–1618, 2002.

R. **The R project for statistical computing.** Disponível em <http://www.r-project.org>. Acesso em: 12 maio 2016.

RAVANEL,S.; GAKIÈRE,B.; JOB,D.; DOUCE,R. The specific features of methionine biosynthesis and metabolism in plants. **Proceedings of National Academy of Sciences of the United States of America**, Washington, D.C., v. 95, p. 7805-7812, 1998.

RAZA,G.; ALI,K.; ASHRAF,M.Y.; MANSOOR,S.; JAVID,M.; ASAD,S. Overexpression of an H+-PPase gene from *Arabidopsis* in sugarcane improves drought tolerance, plant growth, and photosynthetic responses. **Turkish Journal of Biology**, Ancara, v. 40, p. 109-119, 2016.

RICE,J.C.; BRIGGS,S.D.; UEBERHEIDE,B.; BARBER,C.M.; SHABANOWITZ,J.; HUNT,D.F.; SHINKAI,Y.; ALLIS,C.D. Histone methyltransferases direct different degrees of methylation to define distinct chromatin domains. **Molecular Cell**, Amsterdam, v. 12, p. 1591–1598, 2003.

ROACH,B.T.; DANIELS,J. **A review of the origin and improvement of sugarcane**. Piracicaba, SP: Copersucar, 1987. 31 p.

ROBINSON,J.T.; THORVALDSDÓTTIR,H.; WINCLER,W.; GUTTMAN,M.; LANDER,E.S.; GETZ,G.; MESIROV,J.P. Interative Genomics Viewer. **Nature Biotechnology**, New York, v. 29, p. 24-26, 2011a.

ROBINSON,N; BRACKIN,R; VINALL,K; SOPER,F; HOLST,J; GAMAGE,H. Nitrate paradigm does not hold up for sugarcane. **PLoS ONE**, San Francisco, v. 6, e19045, 2011b.

ROPPOLO,D.; RYBEL,B.; DENERVAUD,T.V.; PFISTER,A.; ALASSIMONE,J.; VERMEER,J.E.M.; YAMAZAKI,M.; STIERHOF,Y.D.; BEECHMAN,T.; GELDNER,N. A novel protein family directs Casparian strip formation in the endodermis. **Nature**, New York, v. 473, p. 380-383, 2011.

RUBIN,C.J.; ZODY,M.C.; ERIKSSON,J. Whole-genome resequencing reveals loci under selection during chicken domestication. **Nature**, New York, v. 464, p. 587-591, 2010.

SAINI,H.S.; ATTIEH,J.M.; HANSON,A.D. Biosynthesis of halomethanes and methanethiol by higher plants via a novel methyltransferase reaction. **Plant, Cell and Environment**, London, v.18, p. 1027-1033, 1995.

SAITO,K.; TAUTZ,L.; MUSTELIN,T. The lipid-binding SEC14 domain. **Molecular and Cell Biology of Lipids**, Amsterdam, v. 1771, p. 719-726, 2007.

SCARPARI,M.S.; BEAUCLAIR,E.G.F. Anatomia e botânica. In: MIRANDA, L.L.D.; VASCONCELOS, A.C.M.; LANDELL, M.G.A. (Ed.). **Cana-de-açúcar**. Campinas, SP: Instituto Agronômico, 2008. p. 47-56.

SERANG,O.; MOLLINARI,M.; GARCIA,A.A.F. Efficient exact maximum a posteriori computation for Bayesian SNP genotyping in polyploids. **PLoS ONE**, San Francisco, v. 7, e101371, 2012.

SHEEN,J. Mutational analysis of protein phosphatase 2C involved in abscisic acid signal transduction in higher plants. **Proceedings of National Academy of Sciences of the United States of America**, Washington, D.C., v. 95, p. 975-980, 1998.

SIM,S.C.; ROBBINS,M.D.; DEYNZE,A.V.; MICHEL,A.P; FRANCIS,D.M. Population structure and genetic differentiation associated with breeding history and selection in tomato (*Solanum lycopersicum* L.). **Heredity**, New York, v. 106, p. 927-935, 2011.

SJÖGREN, L.L.E.; STANNE, T.M.; ZHENG, B.; SUTINEN, S.; CLARKE, A.K. Structural and functional insights into the chloroplast ATP-dependent Clp protease in *Arabidopsis*. **Plant Cell**, Rockville, USA, v. 18, p. 2635-2649, 2006.

SOURCE FORGE. **GBS barcode splitter**. Disponível em <https://sourceforge.net/projects/gbsbarcode/>. Acesso em: 08 ago. 2016.

STEINER,F.A.; HEINIKOFF,S. Diversity in the organization of centromeric chromatin. **Current Opinion in Genetics & Development**, Amsterdam, DOI: 10.1016/j.gde.2015.03.010, 2015.

STOREY,J.D. A direct approach to false discovery rates. **Journal of the Royal Statistical Society**, London, v. 64, p. 479-498, 2002.

STRIMMER,K. *fdrtool*: a versatile R package for estimating local and tail area-based false discovery rates. **Bioinformatics**, Oxford, v. 24, p. 1461-1462, 2008.

SUBRAMANIANA,A.; TAMAYOA,P.; MOOTHA,V.K.; MURHERIEED,S.; EBERTA,B.; GILLETTEA,M.A.; PAULOVICHG,A.; POMEROYH,S.L.; GOLUBA,T.R., LANDERA,E.S.; MESIROVA,J.P. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. **Proceedings of the National Academy of the United States of America**, Washington, D.C., v. 102, p. 15545-15550, 2005.

SUTRISNO,H. Molecular phylogeny of Indonesian armyworm Mathimna Guenée (Leidoptera: Hadeninae) based on CO I gene sequences. **Journal of Biosciences**, Bangalore, India, v. 19, p. 65-72, 2012.

SZMRECSÁNYI,T.; MOREIRA,E.P. O desenvolvimento da agroindústria canavieira do Brasil desde a Segunda Guerra Mundial. **Estudos Avançados**, São Paulo, v. 5, p. 57-79, 1991.

TANZ,S.; KILIAN,K.; JOHNSSON,C.; APEL,L.; SMALL,I.; HARTER,K.; WANKE,D.; POGSON,B.; ALBRECHT,V. The SCO2 protein disulphide isomerase is required for

thylakoid biogenesis and interacts with LCHB1 chlorophyll a/b binding proteins which affects chlorophyll biosynthesis in Arabidopsis seedlings. **The Plant Journal**, London, v. 69, p.743-754, 2012.

THALER,J.S.; FIDANTSEF,A.L.; DUFFEY,S.S.; BOSTOCK,R.M. Trade-offs in plant defense against pathogens and herbivores: a field demonstration of chemical elicitors of induced resistance. **Journal of Chemical Ecology**, New York, v. 25, p. 1597-1609, 1999.

TOKESHI,H.; RAGO,A. Doenças da cana-de-açúcar. In: KIMATI,H.; AMORIM,L.; REZENDE,J.A.M.; BERGAMIN FILHO,A.; CAMARGO,L.E.A. (Ed.). **Manual de Fitopatologia**: Volume 2 - Doenças das Plantas Cultivadas. São Paulo: Editora Agronômica Ceres, 2005. p. 185-196.

TORTI,S.; FORNARA,F.; VICENT,C.; ANDRÉS,F.; NORDSTRÖM,K.; GÖBEL,U.; KNOLL,D.; SCHOOF,H.; COUPLAND,G. Analysis of the *Arabidopsis* Shoot Meristem Transcriptome during Floral Transition Identifies Distinct Regulatory Patterns and a Leucine-Rich Repeat Protein That Promotes Flowering. **The Plant Cell**, Rockville, USA, v. 24, p. 444-462, 2012.

TOTTON,E.L.; LARDY,H.A. Phosporic esters of biological importance. **The Journal of Biological Chemistry**, Rockville, USA, v. 181, p. 701-706, 1949.

UENO,O. Structural characterization of photosynthetic cells in amphibious sedge, *Eleocharis vivipara*, in relation to C3 e C4 metabolism. **Planta**, New York, v. 199, p. 382-393, 1996.

UNITED STATES DEPARTAMENT OF AGRICULTURE. Disponível em <http://plants.usda.gov/core/profile?symbol=MISCA>. Acesso em 05 may 2016.

UPADHYAYA,H.D.; BAJAJ,D.; DAS,S.; SAXENA,M.S.; BADONI,S.; KUMAR,V.; TRIPATHI,R.; GOWDA,C.L.L.; SHARMA,S.; TYAGI,A.K.; PARIDA,S.K. A genome-scale integrated approach aids in genetic dissection of complex flowering time trait in chickpea. **Plant Molecular Biology**, New York, v 89, p. 403-420, 2015.

VAN DER HEIJDEN,R.; BOER-HLUPÁ,V.; VERPOORT,R.; DUINE,J.A. Enzymes involved in the metabolism of 3-hydroxy-3-methylglutaryl-coenzyme A in *Catharanthus roseus*. **Journal of Plant Biotechnology**, Amsterdam, v. 38, p. 345-349, 1994.

VAN DILLEWIJN,C. **Botany of sugarcane**. New York: Stechert-Hafner, 1952. 371 p.

VEIGA,J.E. **O desenvolvimento agrícola**: uma visão histórica. São Paulo: Editora da Universidade de São Paulo, 1991. 243 p.

VERMA,A.K.; UPADHYAY,S.K.; VERMA,P.C.; SOLOMON,S.; SINGH,S.B. Functional analysis of sucrose phosphate synthase (SPS) and sucrose synthase (SS) in sugarcane (*Saccharum*) cultivars. **Plant Biology**, London, v. 13, p. 325-232, 2011.

WALKER,A.E. Changing effective population size and the McDonald-Kreitman Test. **Genetics**, Rockville, USA, v. 162, p. 2017-2024, 2002.

WANG,J.; ROE.B.; MACMIL,S.; YU,Q.; MURRAY,J.E.; TANG,H.; CHEN,C.; NAJAR,F.; WILEY,G.; BOWERS,J.; SLUYS,M.A.; ROKHSAR,D.S.; HUDSON,M.E.; MOOSE,S.P.; PATERSON,A.H.; MING,R. Microcollinearity between autopolyploid sugarcane and diploid sorghum genomes. **BMC Genomics**, London, DOI: 10.1186/1471-2164-11-261, 2010.

WELER,J.I.; GLICK,G.; EZRA,E.; ZERON,Y.; SEROUSSI,E.; RON,M. Paternity validation and estimation of genotyping error rate for the BovineSNP50 BeadChip. **Animal Genetics**, London, v. 41, p. 551-553, 2010.

WHETTEN,R.; SEDEROFF,R. Lignin Biosynthesis. **The Plant Cell**, Rockville, USA, v. 7, p. 1001-1013, 1995.

WICKHAM,H. *ggplot2*. **Computational Statistics**, New York, v. 3, p. 180-185, 2011.

WITTE,C.P.; ROSSO,M.G.; ROMEIS,T. Identification of three urease accessory proteins that are required for urease activation in Arabidopsis. **Plant Physiology**, Rockville, USA, v. 139, p. 1155-1162, 2005.

XIE,Z.; CHEN,Z. Harpin-induced hypersensitive cell death is associated with altered mitochondrial functions in tobacco cells. **Molecular Plant-Microbe Interactions**, Saint Paul, USA, v. 13, p. 183-190, 2000.

YADAV,V.; MOLINA,I.; RANATHUNGE,K.; QUERALTA,I.; CASTILLO,Q.; ROTHSTEIN,S.J.; REEDA,J.W. ABCG Transporters Are Required for Suberin and Pollen Wall Extracellular Barriers in *Arabidopsis*. **The Plant Cell**, Rockville, USA, v. 26, p. 3569–3588, 2014.

YANG,X.Y.; CHEN,Z.W.; XU,T.; QU,Z.; PAN,X.D.; QIN,X.H.; REN,D.T.; LIU,G.Q. Arabidopsis kinesin KP1 specifically interacts with VDAC3, a mitochondrial protein, and regulates respiration during seed germination at low temperature. **The Plant Cell**, Rockville, USA, v. 23, p. 1093-1106, 2011.

YORUK,R.; MARSHALL,M.R. Physicochemical properties and function of plant polyphenol oxidase: a review. **Journal of Food Biochemistry**, London, v. 27, p. 361–422, 2003.

ZHANG,H.; XIAOXIANG, H.; WANG,Z.; ZHANG,W.; WANG,S.; WANG,N; MA,L.; LENG,L.; WANG,S.; WANG,Q.; WANG,Y.; TANG,Z.; LI,N.; DA,Y.; LI,H. Selection signature analysis implicates the *PC1/PCSK1* region for chicken abdominal fat content. PLoS ONE, San Francisco, DOI: 10.1371/journal.pone.0040736, 2012.

ZHANG,D.; GUO,H.; KIM,C.; LEE,T.-H. CSGRqtl, a comparative quantitative trait locus database for Saccharinae grasses. **Plant physiology**, Rockville, USA, 161, 594–599, 2013.

ZHOU,J.; LOH,Y.T.; BRESSAN,R.A.; MARTIN,G.B. The tomato gene Pti1 encodes a serine/threonine kinase that is phosphorylated by Pto and is involved in the hypersensitive response. **Cell**, Amsterdam, v. 83, p. 925-935, 1995.

ZOHARY,D.; HOPF,M.; WEISS,D. **Domestication of plants in Old World**: the origin and spread of domesticated plants in south-west Asia, Europe, and Mediterranean Basin. Oxford: Oxford University Press, 2012. 237 p.

ANEXO I – Anotação posicional e funcional dos polimorfismos detectados em todos os locos (dados brutos), em locos filtrados com frequência superior a 1% e em locos fixados (frequência alélica superior a 99%) do Painel Brasileiro de Genótipos de Cana-de-Açúcar. A anotação foi realizada através do *software* SnpEff e utilizou como referência a versão 2.1 do genoma do sorgo (*Sorghum bicolor*).

Tipo	Frequência		
	Dados brutos	Dados filtrados	Dados fixados
3_prime_UTR_variant	6.955 (2,769%)	2.430 (2,966%)	333 (3,081%)
5_prime_UTR_premature_start_codon_gain_variant	503 (0,200%)	193 (0,236%)	29 (0,268%)
5_prime_UTR_variant	3.892 (1,549%)	1.435 (1,752%)	183 (1,693%)
disruptive_inframe_deletion	83 (0,033%)	36 (0,044%)	3 (0,028%)
disruptive_inframe_deletion+splice_region_variant	1 (0,000%)	1 (0,001%)	0 (0,000%)
disruptive_inframe_insertion	75 (0,030%)	43 (0,052%)	3 (0,028%)
disruptive_inframe_insertion+synonymous_variant	4 (0,002%)	2 (0,002%)	2 (0,019%)
downstream_gene_variant	71.259 (28,369%)	23.303 (28,445%)	3.136 (29,016%)
frameshift_variant	813 (0,324%)	250 (0,305%)	1 (0,009%)
frameshift_variant+missense_variant	26 (0,010%)	13 (0,016%)	0 (0,000%)
frameshift_variant+splice_region_variant	11 (0,004%)	2 (0,002%)	0 (0,000%)
frameshift_variant+start_lost	8 (0,003%)	1 (0,001%)	0 (0,000%)
frameshift_variant+stop_gained	24 (0,010%)	8 (0,010%)	0 (0,000%)
frameshift_variant+stop_gained+missense_variant	1 (0,000%)	1 (0,001%)	1 (0,009%)
frameshift_variant+stop_lost	1 (0,000%)	0 (0,000%)	0 (0,000%)
frameshift_variant+synonymous_variant	8 (0,003%)	3 (0,004%)	1 (0,009%)

* Nomes originais fornecidos pelo SnpEff

Tipo*	Frequência		
	Dados brutos	Dados filtrados	Dados fixados
inframe_deletion	88 (0,035%)	48 (0,059%)	4 (0,037%)
inframe_deletion+splice_region_variant	1 (0,000%)	0 (0,000%)	0 (0,000%)
inframe_deletion+synonymous_variant	4 (0,002%)	2 (0,002%)	0 (0,000%)
inframe_insertion	182 (0,072%)	75 (0,092%)	10 (0,093%)
inframe_insertion+splice_region_variant	1 (0,000%)	0 (0,000%)	0 (0,000%)
inframe_insertion+synonymous_variant	3 (0,001%)	3 (0,004%)	0 (0,000%)
initiator_codon_variant	11 (0,004%)	2 (0,002%)	0 (0,000%)
intergenic_region	14.888 (5,927%)	5.309 (6,480%)	857 (7,929%)
intron_variant	25.524 (10,161%)	9.380 (11,450%)	1.463 (13,536%)
missense_variant	45.497 (18,113%)	11.671 (14,246%)	1.011 (9,354%)
missense_variant+disruptive_inframe_deletion	7 (0,003%)	4 (0,005%)	1 (0,009%)
missense_variant+disruptive_inframe_insertion	2 (0,001%)	2 (0,002%)	1 (0,009%)
missense_variant+inframe_deletion	6 (0,002%)	6 (0,007%)	0 (0,000%)
missense_variant+inframe_insertion	11 (0,004%)	8 (0,010%)	0 (0,000%)
missense_variant+splice_region_variant	726 (0,289%)	120 (0,146%)	12 (0,111%)
splice_acceptor_variant+inframe_deletion+splice_region_variant+splice_region_variant+intron_variant	1 (0,000%)	0 (0,000%)	0 (0,000%)
splice_acceptor_variant+intron_variant	140 (0,056%)	18 (0,022%)	0 (0,000%)
splice_acceptor_variant+splice_region_variant+intron_variant	3 (0,001%)	3 (0,004%)	0 (0,000%)
splice_donor_variant+disruptive_inframe_deletion+splice_region_variant+intron_variant	1 (0,000%)	0 (0,000%)	0 (0,000%)

* Nomes originais fornecidos pelo SnpEff

Tipo*	Frequência		
	Dados brutos	Dados filtrados	Dados fixados
splice_donor_variant+intron_variant	321 (0,128%)	59 (0,072%)	2 (0,019%)
splice_donor_variant+splice_region_variant+intron_variant	7 (0,003%)	6 (0,007%)	1 (0,009%)
splice_region_variant	102 (0,041%)	24 (0,029%)	3 (0,028%)
splice_region_variant+intron_variant	151 (0,601%)	524 (0,640%)	67 (0,620%)
splice_region_variant+splice_region_variant+intron_variant	2 (0,001%)	1 (0,001%)	0 (0,000%)
splice_region_variant+stop_retained_variant	3 (0,001%)	0 (0,000%)	0 (0,000%)
splice_region_variant+synonymous_variant	354 (0,141%)	116 (0,142%)	15 (0,139%)
start_lost	64 (0,025%)	13 (0,016%)	1 (0,009%)
start_lost+inframe_deletion	1 (0,000%)	0 (0,000%)	0 (0,000%)
start_lost+inframe_insertion	1 (0,000%)	0 (0,000%)	0 (0,000%)
stop_gained	1.104 (0,440%)	184 (0,225%)	4 (0,037%)
stop_gained+disruptive_inframe_deletion	1 (0,000%)	0 (0,000%)	0 (0,000%)
stop_gained+disruptive_inframe_insertion	2 (0,001%)	(0,000%)	0 (0,000%)
stop_gained+inframe_insertion	1 (0,000%)	(0,000%)	0 (0,000%)
stop_gained+splice_region_variant	33 (0,013%)	2 (0,002%)	0 (0,000%)
stop_lost	47 (0,019%)	6 (0,007%)	4 (0,037%)
stop_lost+inframe_deletion	1 (0,000%)	1 (0,001%)	0 (0,000%)
stop_lost+splice_region_variant	9 (0,004%)	1 (0,001%)	0 (0,000%)
stop_retained_variant	31 (0,012%)	10 (0,012%)	2 (0,019%)
synonymous_variant	26.573 (10,579%)	10.166 (12,409%)	1.453 (13,444%)
transcript	1 (0,000%)	0 (0,000%)	0 (0,000%)
upstream_gene_variant	50.262 (20,010%)	16.438 (20,065%)	2.205 (20,402%)