CARLOS ALEXANDRE DOS SANTOS SILVA

SIMILARIDADE GENÉTICA ENTRE CLONES DE CANA-DE-AÇÚCAR (SACCHARUM SPP) USANDO GENOTIPAGEM DE ALTO RENDIMENTO E PEDIGREE

LAVRAS - MG
2021

CARLO ALEXANDRE DOS SANTOS SILVA

SIMILARIDADE GENÉTICA ENTRE CLONES DE CANA-DE-AÇÚCAR (*Saccharum spp*) USANDO GENOTIPAGEM DE ALTO RENDIMENTO E PEDIGREE

Trabalho de Conclusão de Curso apresentado à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação, Mestrado Profissional em Genética e Melhoramento de Plantas, para obtenção do título de Mestre Profissional.

Orientador: Prof. Dr. José Airton Rodrigues Nunes – DBI/UFLA

Coorientadora: Dra. Aurinelza Batista Teixeira Condé - EPAMIG

Coorientador: Dr. João Ricardo Bachega Feijó Rosa – Centro de Tecnologia Canavieira – CTC.

LAVRAS - MG
2021

CARLO ALEXANDRE DOS SANTOS SILVA

SIMILARIDADE GENÉTICA ENTRE CLONES DE CANA-DE-AÇÚCAR (*Saccharum spp*) USANDO GENOTIPAGEM DE ALTO RENDIMENTO E PEDIGREE

GENETIC SIMILARITY AMONG SUGARCANE GENOTYPES (*Saccharum spp*) USING HIGH-THROUGHPUT GENOTYPING AND PEDIGREE

20/08/2021

Trabalho de Conclusão de Curso apresentado à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação, Mestrado Profissional em Genética e Melhoramento de Plantas, para obtenção do título de Mestre Profissional.

APROVADA em 20/08/2021

Dr. José Airton Rodrigues Nunes – DBI/ICN/UFLA

Dr. João Ricardo Bachega Feijó Rosa – Centro de Tecnologia Canavieira – CTC

Dr. Evandro Novaes – DBI/ICN/UFLA

Dr. Itaraju Junior Baracuhy Brum – CTC

Orientador: Prof. Dr. José Airton Rodrigues Nunes – DBI/ICN/UFLA

LAVRAS - MG
2021

RESUMO

O conhecimento da variabilidade genética do germoplasma disponível é um dos requisitos importantes para que se tenha êxito na condução de um programa de melhoramento. Este trabalho teve por objetivo avaliar asimilaridade genética entre os acessos ou genótipos de cana-de-açúcar do banco de germoplasma da empresa Centro de Tecnologia Canavieira (CTC) por meio de genotipagem densa e pedigree; e investigar a correlação entre diferentes coeficientes de similaridade calculadas a partir dessas informações de relacionamento genético. Foram genotipados 1.230 genótipos usando 81.309 marcadores SNPs, e, então, estimados os coeficientes de similaridade de Jaccard e de parentesco genômico aditivo de Van Raden. A partir do pedigree foi estimado o coeficiente de parentesco de Malecot. A similaridade genética estimada pelo coeficiente de Jaccard variou de 0,14 à 0,92, com uma média de 0,46. O coeficiente de parentesco de Malecot variou de 0 à 0,75, com média equivalente a 0,02. Pela otimização de Tocher, os acessos/genótipos foram agrupados em 83 grupos com base no coeficiente de parentesco de Malecot e 6 grupos quando utilizado coeficiente de similaridade de Jaccard. A análise de componentes principais com base nos dados de similaridade genética resultou na estruturação dos genótipos em consonância com o histórico evolutivo da cana-de-açúcar. As estimativas dos coeficientes de Jaccard e de Van Raden foram mais elevadas, enquanto as correlações destes com o coeficiente de parentesco de Malecot foram baixas, demonstrando que o pedigree foi menos informativo sobre a real similaridade genética no germoplasma avaliado. Os acessos mais divergentes foram o MOL1032 (*Saccharum spontaneum)* e IK7635 *Saccharum officinarum*.

Palavras-chave: Cana-de-açúcar, Divergência genética, genotipagem, coeficiente de parentesco.

ABSTRACT

The knowledge of the genetic variability of the available germplasm is an important requirement for the success of the plant breeding program. This study aimed to evaluate the genetic similarity among accessions/genotypes of the sugarcane germplasm bank of the CTC company through dense genotyping and pedigree data; and to investigate the correlation between different similarity coefficients calculated based on this genetic relationship information. There were genotyped 1,230 genotypes using 81,309 SNPs markers, and then estimated Jaccard similarity coefficient and Van Raden additive genomic kinship coefficient from the marker data. The Malecot kinship coefficient was computed based on the pedigree data. The genetic similarity estimated by the Jaccard coefficient ranged from 0.14 to 0.92, with a mean of 0.46. The Malecot coefficient ranged from 0 to 0.75, and mean equal to 0.02, The Tocher optimization grouped the genotypes into 83 groups based on the Malecot coancestry coefficient and 6 groups when Jaccard's similarity coefficient was used. The principal component analysis using the marker-based genetic dissimilarity structured the genotypes in consonance to the sugarcane evolutionary history. The Jaccard and Van Raden coefficients presented higher correlation (r = 0.45) than those marker-based coefficients with the Malecot kinship coefficient, indicating pedigree data was less informative about the actual genetic similarity in the assessed germplasm. The most divergent genotypes were MOL1032 (*Saccharum spontaneum*) and IK7635 (*Saccharum officinarum*).

Key words: Sugarcane, Genetic divergence, genotyping, kinship coefficient.

1. INTRODUÇÃO

A cana-de-açúcar é muito importante para a economia mundial devido à produção de açúcar e etanol, bem como energia a partir de sua biomassa. No Brasil, o cultivo da cana ocupa uma área total de 9,8 milhões de hectares, com uma produção de 739 milhões de toneladas (https://observatoriodacana.com.br). Para que o país atingisse tal importância no setor sucroenergético, o melhoramento genético da cana-de-açúcar foi fundamental. Os programas de melhoramento vêm elevando os rendimentos agroindustriais, com ganhos de 1% a 2% ao ano e desenvolvendo novas variedades resistentes a fatores bióticos e abióticos, o que têm contribuído para a expansão do cultivo. Entretanto, ainda há a necessidade de aumento da variabilidade genética das populações, com o intuito de obter ganhos de produtividade, reduzindo o tempo despendido para obtenção de novas variedades.

O uso generalizado de variedades comerciais adaptadas ao sistema agrícola, na maioria das vezes oriundas de ancestrais muito próximos, com pequena distância genética entre si, pode levar ao estreitamento da base genética, à depressão por endogamia e ao fenômeno genético denominado vulnerabilidade genética. Assim, o conhecimento da diversidade genética entre um grupo de genitores é importante para o melhoramento genético, sobretudo na identificação de combinações híbridas de maior heterozigose e maior efeito heterótico e, portanto, na recuperação de genótipos superiores em gerações segregantes.

Escolha dos genitores para os cruzamentos é o gargalo nesse processo seletivo, devendo ser planejada de maneira a maximizar a probabilidade de seleção de genótipos superiores com potencial de serem liberados como cultivares comerciais, e ainda serem utilizados em um processo de recorrência.

Os cruzamentos devem ainda ser orientados de forma a evitar a ocorrência de cruzamentos entre indivíduos aparentados, empregando preferencialmente clones elites e cultivares desenvolvidos no país ou região, e ainda observando a associação entre importantes caracteres agroindustriais.

Vários métodos têm sido utilizados para investigar a variação genética. Os métodos tradicionais, que combinam recursos agronômicos e características morfológicas, foram usadas no início. Porém, muitas das caracteristicas vegetativas são influenciadas por fatores ambientais, apresentando variação contínua e alta grau de plasticidade. Com isso, esses caracteres muitas vezes não refletem a verdadeira diversidade de *Saccharum spp*. Com isto, é de vital importância o estudo de parâmetros genéticos e de marcadores moleculares, que podem ser valiosas ferramentas a favor do melhoramento genético da cana-de-açúcar. O uso de técnicas moleculares permite analisar a variabilidade em nível de DNA, resultando em um agrupamento de indivíduos com características semelhantes, permitindo o planejamento dos cruzamentos e a obtenção dos melhores resultados em um tempo menor.

Atualmente, aproximadamente 4478 genótipos compõem o banco de germoplasma na estação de hibridação do CTC – Centro de Tecnologia Canavieira, localizada no município de Camamu – BA, onde são realizados anualmente ~ 1200 cruzamentos, dirigidos para cada uma das regiões canavieiras do país. Combinações são orientadas com base nas informações do *breeding value* dos genitores, de forma a evitar a ocorrência de cruzamentos entre indivíduos aparentados, por meio do coeficiente de parentesco, e ainda observando a associação entre importantes caracteres agroindustriais. Tais informações são geradas através da caracterização dos clones de acordo com dados de Tonelada de Colmo por Hectare (TCH), teor de açúcar, Tonelada de Pol por Hectare (TPH), assim como comportamento em relação as doenças. Este trabalho teve por objetivo avaliar o nível de similaridade genética (SG) entre os genótipos de cana-de-açúcar presente no Banco de Germoplasma da empresa CTC (Centro de Tecnologia Canavieira) por meio de genotipagem densa e pedigree, e investigar a correlação entre as similaridades calculadas com base nessas informações de relacionamento genético.

2. REFERENCIAL TEÓRICO

2.1. Variabilidade Genética em Cana-de-açúcar

A cana-de-açúcar é uma espécie alógama, com ciclo perene e própria de climas tropical e subtropical, apresentando três prováveis regiões de origem, segundo vários relatos: Nova Guiné, China e Índia. Esta cultura e espécies afins são membros da tribo Andropogoneae, pertencente à família Poaceae, gênero Saccharum, a qual é representada atualmente por seis espécies: *Saccharum officinarum* L., *Saccharum spontaneum* L., *Saccharum robustum* J., *Saccharum barberi* Jeswiet, *Saccharum sinense* Roxb, e *Saccharum edule*. Membros do gênero *Saccharum* são altamente poliplóides e possuem uma elevada variabilidade interespecífica e muita aneuploidia, caracterizando indivíduos de várias constituições genéticas, oscilando de 2n = 40 até 2n = 205, o que implica numa base genética complexa (Cesnik, 2005).

A espécie *S. officinarum* (2n = 80 cromossomos) é também conhecida por cana-nobre, termo criado pelos melhoristas holandeses em 1920 pelo seu elevado teor de açúcar. Esta é uma das principais espécies que contribuíram com alelos para genótipos de cana-de-açúcar cultivados atualmente no mundo.

Apesar do não cultivo de *S. spontaneum* (2n = 40 a 2n = 128), ela proporcionou, no início do século XX a retomada do crescimento da indústria do açúcar no mundo após severas epidemias de doenças ocorridas naquela época.

S. spontaneum é autopoliplóide, altamente polimórfica e contribuiu para o melhoramento da cana-de-açúcar com seu vigor, capacidade de perfilhamento e de rebrota da soqueira, além de tolerância a estresses abióticos, pragas e doenças (Gheysa Coelho Silva & Cana-de-açúcar, 2012).

As cultivares modernas de cana-de-açúcar são em grande parte híbridos interespecificos compostos de aproximadamente 80% do genoma de *S. officinarum,* 10% de *S. spontaneum* e 10% recombinantes entre os dois genomas (Marta et al., 2007). A elevada contribuição de *S. officinarum* nos genomas dessas cultivares deve-se ao processo conhecido como “nobilização”, através de retrocruzamentos com genitores de *S. officinarum* para aumentar a

proporção desse genome de "cana-nobre" nas cultivares. Essas cultivares possuem genomas altamente poliploides e aneuploides, com número de cromossomos variando de 100 a 130.

Considerando que a variabilidade genética é o ponto de partida para qualquer programa de melhoramento genético, sua caracterização e avaliação são indispensáveis aos trabalhos ligados ao melhoramento de plantas. Para isso, o melhorista lança mão de vários métodos de análise de dados. A escolha do método mais adequado deve levar em consideração o nível de precisão desejada, facilidade de análise, forma de obtenção e natureza dos caracteres avaliados (Cavalcante et al., 2010). Em programas de melhoramento genético, a variabilidade genética para a seleção de novos clones é obtida através da realização de cruzamentos entre variedades comerciais e pré-comerciais que apresentem em sua constituição genes relacionados a caracteres de interesse, aumentando a probabilidade de obtenção de uma nova variedade (Dutra Filho et al., 2011).

Nesse contexto, o conhecimento da diversidade genética entre variedades comerciais em programas de melhoramento de plantas é de vital importância para os melhoristas na identificação e organização dos recursos genéticos disponíveis. Esse conhecimento deve racionalizar a utilização destes recursos genéticos na produção de novas variedades promissoras (Silva et al., 2011).

Autores retratam que para a cana-de-açúcar, a estimativa da divergência genética entre diferentes genótipos vem sendo estudada, visando a seleção de genitores para formação de híbridos ou mesmo a formação de novas populações segregantes, oriundas do intercruzamento de genótipos divergentes com características agronômicas superiores.

Silva et al.(2011) estudaram a divergência genética entre sete variedades-padrão e onze clones RB de cana-de-açúcar, por meio de técnicas uni e multivariadas, com base em dez caracteres agroindustriais. Eles constataram que os caracteres Pol % cana (porcentagem em massa de sacarose aparente contida em uma solução), toneladas de cana por hectare, brix % cana e altura de colmos foram os principais determinantes na

quantificação da divergência genética. Por fim, eles concluíram que a metodologia multivariada permite identificar genótipos de maior divergência genética para utilização em programas de melhoramento da cana-de-açúcar.

Dutra Filho et al.(2011) avaliaram a divergência genética em progênies de cana-de-açúcar oriundas de autofecundação e variedades comerciais por meio de técnicas multivariadas com base em oito caracteres agroindustriais. Observou-se variabilidade genética entre as progênies, para os caracteres toneladas de Pol por hectare (TPH) e toneladas de cana por hectare (TCH), que são considerados os mais importantes componentes de produção em cana-de-açúcar. Utilizando o método de otimização de Tocher, com base na distância generalizada de Mahalanobis, os autores agruparam as seis progênies avaliadas em quatro grupos distintos. Neste tipo de análise, é comum que os primeiros grupos tenham um maior número de indivíduos que os últimos. Elias et al. (2007), avaliando a variabilidade genética em germoplasma tradicional de feijão preto, afirmaram que o método de otimização de Tocher tem como princípio manter a homogeneidade dentro dos grupos e a heterogeneidade entre os grupos. Assim sendo, o maior número de indivíduos em um determinado grupo indica que eles apresentam maior similaridade genética e os indivíduos enquadrados no último grupo apresentam maior divergência em relação àqueles que estão no primeiro grupo.

Lopes et al.(2014) estudaram divergência genética, avaliando 138 clones da cana-de-açúcar usando a metodologia de dispersão gráfica por componentes principais associada aos modelos lineares mistos. Com isso, eles identificaram quais são os genótipos mais divergentes e os mais produtivos de forma mais precisa, para posterior recombinação. Constatou-se que há variabilidade genética no material estudado, sendo possível o estudo da divergência, e que os mesmos podem ser utilizados em cruzamentos com o objetivo de gerar variabilidade nas gerações segregantes. Os autores concluíram que a metodologia de componentes principais associada aos modelos lineares mistos mostrou que há variabilidade genética entre os clones de cana-de-açúcar estudados, em todos os ambientes. O caráter que mais contribuiu para a divergência genética total foi quilograma de brix por parcela, e o que menos contribuiu foi o caráter número de colmos por parcela. Os clones

mais produtivos e mais divergentes podem ser combinados em novas etapas de cruzamentos e seleções visando à obtenção de genótipos mais produtivos e com variabilidade.

Por meio de marcadores moleculares do tipo microssatélite, Duarte, (2012) avaliou os níveis de diversidade genética das cultivares. Ele concluiu que os microssatélites SSR05, SSR06 e SSR93 foram eficientes em determinar perfis genéticos únicos e em discriminar o grau de parentesco e a diversidade genética entre as 21 cultivares avaliadas.

Resende et al. (2013) utilizaram RAPD e EST's SSR como ferramentas para avaliar a variabilidade, bem como para estimar a divergência genética entre 23 variedades comerciais de cana-de-açúcar oriundos de autofecundação. Eles constataram que os marcadores RAPD detectaram um alto grau de polimorfismo genético, produzindo 61 bandas, das quais 58 foram polimórficas. Os marcadores EST's SSR amplificaram 38 alelos, sendo 34 polimórficos. Os marcadores identificaram três grupos na população estudada. A maior parte da variação genética foi mantida dentro das progênies, evidenciando a ocorrência de variabilidade genética entre os genótipos de cada progênie para fins de melhoramento. Através da divergência genética estimada foi possível identificar parentais divergentes para trabalhos de hibridação, visando a obtenção de clones superiores com caracteres de interesse à agroindústria canavieira. Os marcadores moleculares RAPD e EST's SSR foram igualmente eficientes para estimar a variabilidade genética nos genótipos testados e elaborar os cruzamentos a serem realizados nos programas de melhoramento.

Crystian et al. (2018) analisaram a base genética do banco de germoplasma da Serra do Ouro nas últimas décadas usando marcadores moleculares microssatélites. Constatou-se o uso de apenas alguns genótipos nos programas de melhoramento da cana durante as décadas de 1970 e 1980. Esses genótipos visavam principalmente aumentar os níveis de sacarose, o que elevou os níveis de similaridade genética no germoplasma dos programas de melhoramento. A similaridade média foi de 0,448, indicando valores moderados. A partir da década de 1990, houve uma mudança na estratégia de melhoramento da cana, devido ao surgimento de novas pragas e doenças e à

necessidade de cultivares mais adaptadas que as existentes. Dado o novo cenário, os programas de melhoramento incluíram novas variedades progenitoras, contribuindo para um menor índice de similaridade genética entre as variedades desenvolvidas após 1990. Dessa forma, Crystian et al. (2018) destacam que os níveis de diversidade genética no melhoramento da cana-de-açúcar permaneceram praticamente os mesmos nas últimas cinco décadas e que as variedades desenvolvidas resultaram da variabilidade genética dos primeiros híbridos. Estes foram obtidos por cruzamento entre *S. officinarum* e *S. spontaneum*, resultando em híbridos com diferentes razões de constituição genômica, devido a diferentes proporções cromossômicas (10-20% de *S. spontaneum* e 80-85% de *S. officinarum*).

2.2. Uso do pedigree e Marcadores Moleculares no Melhoramento Genético da Cana-de-açúcar

O surgimento das técnicas de marcadores de DNA com capacidade de detectar variação genética adicional, trouxe novos avanços para o melhoramento genético de plantas, sendo utilizadas com êxito em várias culturas. A principal colaboração que as técnicas trouxeram foi a possibilidade de analisar intrinsecamente o genótipo de um indivíduo sem a necessidade da ocorrência da expressão fenotípica e, consequentemente, excluindo-se a influência do ambiente. O uso de marcadores baseados no genótipo do indivíduo tem recebido atenção especial para a caracterização de cultivares, decorrente do seu potencial de distinção de genótipos morfologicamente similares e geneticamente aparentados. Já a utilização da tecnologia de marcadores moleculares no processo seletivo ocorre através da procura de alelos desejáveis indiretamente por meio do uso de marcadores ligados a locos controladores de caracteres quantitativos (QTL). Essa técnica de apoio ao melhoramento é conhecida como seleção assistida por marcadores moleculares (SAM)(Toppa & Jadoski, 2013).

O uso de técnicas moleculares na cana começou no final da década de 1980, quando geneticistas e melhoristas examinaram vários tipos de marcadores de DNA e sua utilidade no melhoramento genético da cana-de-açúcar. No entanto, a cana-de-açúcar ficou atrás de outras culturas na

utilização de marcadores moleculares, por causa de sua complexidade genética, ao contrário de seus parentes diploides como o milho, que apresenta genoma mais simples e pode ser explorado com maior eficácia. No complexo poliploide da cana-de-açúcar, a análise dos dados gerados pela amplificação de marcadores do tipo microssatélite (SSR) é tecnicamente mais exigente do que para os organismos mais simples, devido ao padrão de bandas obtidas como resultado de múltiplos alelos presentes nos cromossomos homólogos (Duarte, 2012).

Os marcadores moleculares são ferramentas valiosas no estudo de genomas complexos como o da cana-de-açúcar. A utilização de marcadores na seleção de características agronômicas desejáveis durante os estágios iniciais do melhoramento, ou mesmo na escolha de genitores em um cruzamento, pode reduzir significantemente o tempo de desenvolvimento de novas variedades comerciais. Diversidade genética, escolha de genitores e identificação de variedades também são algumas das aplicações dos marcadores moleculares em cana-de-açúcar (Rosa, 2011).

Marcadores moleculares associados a características agronômicas relevantes podem reduzir significativamente o tempo e o custo envolvidos no desenvolvimento de novas variedades, pois podem ajudar na seleção dos melhores progenitores e acelerar a taxa de ganho genético no programa de melhoramento (Racedo et al., 2016).

Garcia et al. (2013) avaliaram o uso de SNPs e novos métodos estatísticos para estimativa de nível de ploidia em cana-de-açúcar usando procedimentos baseados em espectrometria de massa e o software SuperMASSA. Eles demonstraram que é possível estimar o nível de ploidia e a dosagem de SNPs, fornecendo informações úteis sobre a interpretação do genoma da cana-de-açúcar. A cana-de-açúcar é um excelente caso de teste para poliploides, pois possui genoma complexo com nível de ploidia desconhecido e aneuploidia frequente. Para realização do trabalho, foram explorados dois cenários diferentes. Primeiro, 271 SNPs gerados usando a tecnologia Sequenom iPLEX MassARRAY foram usados para analisar uma população de 180 indivíduos de um cruzamento biparental entre as variedades IACSP95-3018 e IACSP93-3046. Segundo, 1034 SNPs foram analisados em

um painel de 142 genótipos relevantes de cana-de-açúcar. O painel consistiu em importantes variedades comerciais, além de genótipos ancestrais e parentais que têm sido frequentemente usados em um amplo espectro de programas de melhoramento.

R. R. Silva. (2013) conclui através do estudo da estrutura populacional em cana-de-açúcar usando marcadores do tipo SNP, que há evidências de um possível estreitamento da base genética da cana-de-açúcar ao longo das gerações, devido ao cruzamento recorrente de indivíduos aparentados. Isto pode ser afirmado porque o material melhorado geneticamente agrupou-se quando sua estrutura genética foi avaliada com base em SNPs.

2.3. Estimativa do Coeficiente de Parentesco com Base em Marcadores Moleculares

Embora o coeficiente de parentesco entre indivíduos possa ser calculado a partir de uma linhagem conhecida e o seu pedigree, na ausência dessa informação, a relação pode ser estimada usando dados de marcadores genéticos. Vários estimadores foram desenvolvidos para esse fim e geralmente podem ser classificados em duas categorias: estimadores de método do momento e de probabilidade máxima.

Os estimadores de método de momento substituem os momentos da amostra pelo momento desconhecido da população para estimar vários parâmetros populacionais. Esses métodos podem gerar uma estimativa imparcial da relação, a probabilidade de que os dois genes de um indivíduo são idênticos por descendência (IBD) aos dois genes de outro indivíduo. O segundo método, probabilidade máxima, estima a probabilidade de observar um dado padrão alélico (probabilidade de um único alelo em um indivíduo ser IBD a outro em outro indivíduo), e as frequências do alelo. Embora existam muitas estimativas baseadas no método do momento ou nos métodos de máxima verossimilhança para organismos diplóides, poucos estimadores existem para organismos com vários níveis de ploidia, como o caso da cana-de-açúcar.

HUANG et al. (2015) descrevem um estimador de probabilidade máxima para autopoliplóides e quantificaram seu desempenho estatístico sob uma

variedade de condições biologicamente relevantes. Os desempenhos estatísticos de cinco estimadores poliplóides adicionais de parentesco também foram quantificados em condições idênticas. Concluiu–se que, no geral, o estimador de probabilidade máxima que desenvolveram, oferece várias vantagens sobre os métodos existentes. Primeiro, geralmente exibe RMSE mais baixo em comparação com outros estimadores. Segundo, todas as estimativas se enquadram em uma faixa biologicamente significativa, e os coeficientes de "ordem superior" podem ser explicados como probabilidades. Assim, a interpretação biológica das estimativas individuais é direta. Terceiro, fornece uma solução para situações em que a dose do alelo não pode ser determinada. Embora o estimador de probabilidade com máxima probabilidade tenha tido bom desempenho em simulações, existem condições nas quais outros estimadores tiveram um desempenho melhor, de acordo com métricas específicas. Não existe um estimador único com desempenho superior em todas as condições e em todas as métricas. Para aplicações específicas sob condições específicas de pesquisa, é possível identificar um estimador ideal. O pacote de software P OLYRELATEDNESS fornece uma função de simulação que ajuda os pesquisadores a avaliar o desempenho de cada estimador sob suas condições.

Loci (2018) apresenta recomendações de diretrizes para programas de melhoramento na estimativa da relação molecular com autopoliploidia, simulando vários cenários com diferentes números de loci e alelos e nível de ploidia. Com base nessas populações, estimou - se a relação com vários métodos disponíveis avaliando seu poder estatístico. Os autores concluíram que é crucial considerar a dosagem poliploide para que um determinado método estatístico atinja um bom poder estatístico. Em apenas alguns cenários específicos com frequência de alelo altamente desequilibrada, métodos pseudo-diplóides podem ser satisfatórios. Portanto, métodos especificamente desenvolvidos para poliploides deve ser usado. Inferir relação com alto poder estatístico em uma população autopoliplóide altamente endogâmica, é mais difícil do que uma população sem endogamia, métodos e marcadores disponíveis não são satisfatórios neste caso.

2.4. Coeficiente de Parentesco com Base no Pedigree

Informações sobre a base genética do germoplasma e as relações entre o material elite que será melhorado, com a escolha de parentais para obtenção de híbridos, é essencial em programas de melhoramento. Uma ferramenta proposta para auxiliar o melhorista nessa tarefa é o conhecimento a priori do grau de parentesco entre os possíveis genitores. O grau de parentesco entre as variedades pode ser estimado pelo coeficiente de parentesco (fxy), também chamado de coancestralidade ou coeficiente de kinship, fornecendo uma estimativa das relações genéticas entre dois genótipos baseado na análise de genealogia. O coeficiente de parentesco entre dois genótipos é a probabilidade de que um alelo, ao acaso, de um indivíduo seja IBD a um alelo, ao acaso, do mesmo loco de outro indivíduo (MALECOT, 1948).

O coeficiente de parentesco (f) é um método importante para estimar a distância genética com base em pedigree. De forma indireta, ele mede a distância genética entre cultivares, estimando, a partir de registros genealógicos, a probabilidade de que alelos, em um locus, sejam idênticos por descendencia. No entanto, suposições feitas ao calcular f em relação ao parentesco de ancestrais, pressão de seleção e deriva genética geralmente não são atendidos. Seu uso tem sido difundida entre espécies de auto-fertilização, como soja, trigo e cevada.

A adoção do coeficiente de parentesco como uma medida de similaridade talvez seja a alternativa mais fácil e barata para estimar a dissimilaridade genética entre um grupo de genótipos, desde que existam informações referentes a sua genealogia.

O cálculo além de requerer a genealogia detalhada de todos os genótipos, admite que os ancestrais originais do germoplasma em estudo não são relacionados (fxy = 0). Também admite-se que cada parental contribui com igual proporção de alelos para sua progênie, fato este que muitas vezes não corresponde à realidade. O cálculo do coeficiente de parentesco usado para espécies autógamas teve que ser adaptado para sua utilização na cultura da cana-de- açúcar. Como todos os ancestrais, cultivares e parentais em cana-de-açúcar são heterozigotos foi admitido o valor de fxy = 0,5 no parentesco do

genótipo consigo mesmo, diferentes das espécies autógamas, em que este valor é 1,0. Muitas vezes a cana-de-açúcar apresenta genealogia incompleta ou desconhecida. Além disso, as pressuposições assumidas para cálculo do fxy normalmente não são atendidas. Portanto, seria de muita utilidade algum método que pudesse determinar o grau de parentesco entre variedades de cana-de-açúcar, independentemente do conhecimento de sua genealogia. Ao contrário do parentesco medido a partir da análise genealógica, a análise por marcadores moleculares fornece uma medida direta da diversidade (Gheysa et al., 2014).

2.5. Medidas para Estudo de Divergência Genética usando Marcadores Moleculares

As diferenças entre os seres vivos baseiam-se na diversidade genética que está codificada nos genes. Porém, apenas uma pequena porção da variabilidade genética total dentro de cada espécie é utilizada em plantas comerciais. Geralmente uma das primeiras preocupações de um melhorista é a existência de variabilidade genética no germoplasma, que pode aumentar as chances de encontrar indivíduos superiores nas gerações segregantes.

A utilização generalizada de variedades comerciais adaptadas aos sistemas agrícolas, frequentemente oriundos de ancestrais muito próximos, com pequena distância genética entre si, pode levar ao estreitamento da base genética e ao fenômeno genético denominado vulnerabilidade genética. Assim, o conhecimento da diversidade genética, entre variedades comerciais em programas de melhoramento de plantas, é de fundamental importância para os melhoristas na identificação e organização dos recursos genéticos disponíveis, visando a utilização desses na produção de novas variedades promissoras (Gheysa Coelho Silva & Cana-de-açúcar, 2012).

As técnicas de aferição da divergência têm sido utilizadas pelos melhoristas de plantas há várias décadas principalmente para investigar a natureza e suas relações de divergência genética envolvendo caracteres morfo-agronômicos e moleculares, parentesco, diversidade de origem geográfica, capacidade de combinação e heterose.

A divergência genética pode ser estimada com base em métodos preditivos a partir de marcadores agronômicos, bioquímicos, morfológicos e moleculares. Os três primeiros apresentam limitações em relação à influência do meio ambiente e ao número limitado de descritores fenotípicos. Os marcadores moleculares, em contrapartida, apresentam a vantagem de representar todo o genótipo, mantendo consistência nos resultados, evitando o problema da expressão do fenótipo (Almeida et al., 2009). Além disso, aceleram a obtenção de genótipos desejáveis e podem estar ligados a locos que determinam características de interesse, não sendo influenciado por variações ambientais (Marta et al., 2007).

No caso de variáveis quantitativas, essa variabilidade pode ser acessada utilizando-se medidas de dissimilaridade, destacando-se, entre elas, a distância Euclidiana e a distância generalizada de Mahalanobis. Essa última leva em consideração as variâncias e covariâncias residuais existentes entre as características mensuradas, quando o experimento se encontra sob delineamento experimental.

Os métodos de agrupamento têm por finalidade separar um grupo original de observações em vários subgrupos, de forma a se obter homogeneidade dentro e heterogeneidade entre os subgrupos. No entanto há vários métodos de agrupamento, sendo os hierárquicos e os de otimização os mais utilizados no melhoramento de plantas. Nos métodos hierárquicos, os indivíduos são agrupados por um processo que se repete em vários níveis, estabelecendo-se um dendrograma, sem preocupação com o número ótimo de grupos. Nos métodos de otimização, por sua vez, os grupos são estabelecidos otimizando-se determinado critério de agrupamento, e difere dos métodos hierárquicos pelo fato de os grupos formados serem mutuamente exclusivos (Gheysa et al., 2014).

2.6. Análises de Agrupamento

As técnicas de agrupamentos têm por objetivo agrupar indivíduos em classes. Portanto, dado um conjunto de n indivíduos, todos avaliados para p variáveis, tais indivíduos devem ser agrupados em classes, de forma que os

mais semelhantes permaneçam na mesma classe. De forma geral, o número de classes não é conhecido inicialmente. Porém, quando essas técnicas geram grupos não esperados, isso pode sugerir que as relações entre os objetos precisam ser mais bem estudadas.

Inicia-se o processo definindo-se os indivíduos e os objetivos desejados para a aplicação da análise, além dos critérios que irão definir as semelhanças entre eles. Obtidos esses dados, eles são dispostos na forma de uma matriz, em que as colunas representam os indivíduos de interesse e as linhas representam as variáveis. Obtida a matriz de dados padronizados, o próximo passo é a escolha de uma medida que quantifique o quanto dois indivíduos são parecidos. Tais medidas são denominadas coeficientes de similaridade e são elas que vão gerar a matriz D de similaridade. Tais coeficientes podem ser divididos em duas categorias: medidas de similaridade e medidas de dissimilaridade. Para a primeira categoria, quanto maior o valor observado, mais parecidos são os indivíduos; para a segunda, quanto maior o valor observado, menos parecidos são os indivíduos. Para cada tipo de variável (quantitativas, qualitativas ordinais, qualitativas nominais e mistas), são definidos diferentes coeficientes de similaridade.

Os marcadores moleculares geram variáveis qualitativas binárias, caracterizadas pela presença ou ausência da marca (banda) após as análises laboratoriais. Dunn & Everitt (1980) designaram por zeros e uns os estados do caráter para caracteres binários, que serão submetidos posteriormente à análise. Como definido anteriormente, o número 1 indica a presença da marca e 0, a ausência. Genótipos que possuam maior coincidência de presenças e/ou ausências das marcas serão considerados mais similares entre si.

De um modo geral, as medidas de similaridade e de dissimilaridade são interrelacionadas e facilmente transformáveis entre si. Há muitos coeficientes de similaridade e/ou de dissimilaridade para caracteres binários disponíveis na literatura, tais coeficientes podem ser divididos em cinco diferentes classes, apresentadas a seguir: i) Coeficientes de Similaridade, ii) Coeficientes de Associação, iii) Distância Euclidiana, iv) Conteúdo de Informações ou Medidas de Diversidade, v) Medidas de Similaridade Dependentes da Probabilidade Estimada.

2.7. Classificação das técnicas de agrupamento

Vários são os tipos de técnicas de agrupamento encontradas na literatura. Manly (1994) destacou duas abordagens particulares, destacando, a primeira delas: i) técnicas que produzem dendrogramas, em que o primeiro passo é calcular as medidas de dissimilaridade (ou similaridade) entre todos os pares possíveis de indivíduos e, assim, formar os grupos por processos aglomerativos ou divisivos; ii) técnicas que envolvem partições, em que os indivíduos podem se mover fora e dentro dos grupos em diferentes estágios da análise.

Tais métodos podem ser classificados em diversas categorias contrastantes, apresentadas a seguir: a) Métodos aglomerativos e divisivos, b) Métodos hierárquicos e métodos não hierárquicos, c) Métodos sem sobreposição e métodos com sobreposição, d) Métodos sequenciais e métodos simultâneos, dentre outros.

Meyer (2002), por meio da comparação de coeficientes de similaridade, relata que dentre todas as análises de agrupamento empregadas (dendrogramas, método de otimização de Tocher e projeção da matriz de similaridade no plano bidimensional), a que apresentou resultados mais próximos da natureza dos dados foi a que fornece dendrogramas, pois ela refletiu a constituição dos grupos esperados a priori. Porém, essa coerência nos agrupamentos foi maior quando considerados os dados provenientes do marcador AFLP. Além disso, com base nas correlações cofenéticas, distorções e estresses, nota-se que os dendrogramas permitiram uma representação das matrizes de similaridade superior aos demais métodos.

3. MATERIAIS E METODOS

3.1 Descrição do germoplasma

O germoplasma pertencente ao Programa de Melhoramento Genético do CTC – Centro de Tecnologia Canavieira, localizado no Munícipio de Camamu – BA, possui atualmente aproximadamente 4478 genótipos. Para realização deste trabalho, foram consideradas amostras (para extração de DNA) de 1329 acessos, dos quais 32 referem – se a genótipos de *Saccharum Spontaneum*, 27 *S. Sinense*, 34 *S. Officinarum*, 26 *S*. *Barberi*, 2 *S. Robustum* e 1208 Híbridos interespecíficos (sendo os genótipos desenvolvidos pelos atuais programas de melhoramento genético no Brasil (CTC, IAC e RIDESA), classificados como Híbridos Modernos (total de 1142), e Híbridos Antigos que compreendem genótipos oriundos dos primeiros programas de melhoramento genético como os clones “NAs” – Norte da Argentina, “CP” – Campos-RJ, “Q” – Australia (total de 66).

3.2 Genotipagem

3.2.1 Coleta do material vegetal e extração e quantificação do DNA

As amostras foram constituídas por três folhas jovens de cada genótipo. Para padronização da coleta, deu-se preferência para a retirada da folha +1, ou seja, a primeira folha com “colarinho” visível (Top Visible Dewlap ou TVD), excluída a nervura central. Com o auxílio de um furador foram recortadas amostras das folhas de 80mg de tecido vegetal (cerca de 12 discos).

Para extração do DNA, utilizou-se o KIT Nuc Prep. O DNA de cada clone foi extraído em duplicata, para garantir a quantidade suficiente para genotipagem de marcadores SNPs.

3.2.2 Genotipagem por Sequenciamento

Para obtenção dos dados de marcadores moleculares, foi aplicado a metodologia de Genotipagem por Sequenciamento. O serviço foi realizado pela empresa Rapid Genomics (Florida Hub Innovation, Flórida - EUA).

A Rapid Genomics desenvolveu uma plataforma de genotipagem de alto rendimento para cana-de-açúcar baseada na captura de sequência direcionada seguida de sequenciamento de última geração (metodologia de captura de sequência). Resumidamente, o DNA genômico é cortado, adaptadores com código de barras são ligados aos fragmentos, um enriquecimento de PCR é realizado e os indivíduos são agrupados para captura. Em seguida, um processo de captura usando sondas é executado e as bibliotecas são enviadas para sequenciamento (Illumina HiSeq 2000; com modo 100nt emparelhado). Depois que o Rapid Genomics recebe os dados de sequenciamento, o arquivo é demultiplexado por amostra código de barras individual, filtrado por qualidade e alinhado à sequência de referência usada para projetar as sondas. As leituras emparelhadas são alinhadas e SNPs bialélicos (polimorfismos de nucleotídeo único) são identificados e a genotipagem é realizada.

As sondas de captura são selecionadas e projetadas para hibridizar com regiões únicas e específicas de interesse no genoma, como forma de reduzir a complexidade do genoma para genotipagem. Como um genoma de referência da cana-de-açúcar não está disponível, o desenvolvimento de sondas de captura contou com um conjunto unigênico e sequências shotgun de leitura curta ancoradas no genoma intimamente relacionado ao Sorgo. As sondas foram projetadas para serem complementares a 40.000 loci do genoma da cana-de-açúcar. Esses loci foram selecionados com critérios rigorosos para (i) representar o maior número de genes possível, com base em ESTs (Expressed Sequence Tags) exclusivos de cana-de-açúcar disponíveis em bancos públicos de dados genômicos; (ii) representar uniformemente o genoma em regiões intergênicas; (iii) contêm uma cinética de hibridação favorável; e (iv) evitar repetições e regiões de baixa complexidade do genoma. Embora o genoma do sorgo tenha sido usado como uma referência de ancoragem para várias etapas do pipeline de design da sonda, a sequência das próprias sondas foi projetada com base nos dados da sequência genômica da cana-de-açúcar. A análise de duas amostras realizadas como duplicatas independentes mostram uma repetibilidade da chamada genotípica acima de 98%, sugerindo que a abordagem é viável para genotipagem em grande escala e reproduzível.

A distribuição dos marcadores moleculares foi avaliada com base no genoma do Sorgo (cromossomos de referência).

3.3 Mapeamento dos marcadores SNPs

Após o processo de genotipagem, foram identificados um conjunto de 126.591 SNPs nos 1329 genótipos. É importante salientar inicialmente que, as informações de genotipagem SNP utilizada neste trabalho não fornece a dosagem com que cada alelo ocorre em cada indivíduo (sabidamente poliploide).

Para controle de qualidade, existem diferentes critérios que podem ser usados, envolvendo cada SNP individualmente e para as amostras. Para os SNPs, normalmente, são utilizados como critérios o escore de genotipagem (GCscore), a taxa de genotipagem (Call rate), a menor frequência alélica (MAF), o desvio do equilíbrio de Hardy-Weinberg (HWE) e a correlação (r2) entre SNPs. Para o controle de qualidade da amostra, os critérios mais comumente empregados são o Call rate e o nível de heterozigose. No entanto, não há um consenso entre os pesquisadores quanto aos valores de limiar de exclusão atribuído para cada critério de controle de qualidade que, em geral, são determinados de acordo com os objetivos do trabalho, permitindo maior ou menor rigor (Bresolin, 2015).

Para o presente trabalho, os marcadores foram filtrados com base em Call Rate (>0,80) e MAF (>0,05). O número de genótipos foi atualizado para 1230 e marcadores SNPs foi reduzido para 81.309.

3.4 Genealogia entre os acessos

O registro dos genitores envolvidos nos cruzamentos, é uma prática importante dentro dos programas de melhoramento genético. Isso é possível quando realiza – se cruzamentos biparental (cruzamento envolvendo um genitor feminino e um genitor masculino) onde os genitores são conhecidos. Estas informações são armazenadas em um banco de dados, onde a consulta é possível sempre que necessário. Por outro lado, quando são realizados

cruzamentos multiparental ou policruzamentos (cruzamento envolvendo um genitor feminino conhecido com mais de um genitor masculino), apenas a informação do genitor conhecido fica registrado, com isso, não é possível obter a genealogia completa de genótipos oriundos destes cruzamentos.

Os dados de genealogia utilizados neste trabalho, foram obtidos através da consulta do banco de dados do CTC.

3.5 Análise dos dados moleculares

3.5.1 Similaridade de Jaccard usando os dados moleculares

A similaridade genética com base nos dados de marcadores, foi determinada através dos coeficientes de similaridade de Jaccard, amplamente utilizado em estudos de variabilidade genética por excluir a coincidência do tipo “0-0” como fator de similaridade. Apenas a coincidência “1-1” deve ser levada em consideração na similaridade entre dois acessos.

Equação Coeficiente similaridade de Jaccard:

$$SGij = a/(a+b+c),$$

Onde:

SGij - é a similaridade genética entre os genótipos i e j;

a - é o número de marcadores presentes em ambos i e j;

b - é o número de marcadores presentes em i e ausentes j;

c - é o número de marcadores presentes em j e ausentes em i.

A estimativa do índice de similaridade Jaccard, foi calculada por meio do software R utilizando o pacote NetCoin (Escobar, Barrios, Prieto e Martinez-Uribe 2020). Para computar a dissimilaridade, utilizamos a diferença do valor observado na matriz de similaridade, onde Dissimilaridade = 1 – Matriz de Similaridade.

3.5.2 Coeficiente de parentesco genômico aditivo

A matriz de parentesco genômico aditivo (Matriz G) pode ser calculada por diversos métodos utilizando dados de marcadores do tipo SNP. VanRaden (2008) apresentou três métodos utilizando a matriz M, que especifica quais marcadores de alelos cada indivíduo herdou. A dimensão de M é o número de indivíduos (n) pelo número de locos (m). As equações podem incluir as informações de marcadores usando a matriz n x n, MM', ou, a matriz m x m, M' M. Se os elementos de M forem -1, 0 e 1 para os locos, homozigoto, heterozigoto e o outro homozigoto, respectivamente, a diagonal MM' conta o número de locos homozigotos para cada indivíduo, e os elementos fora da diagonal medem o número de alelos compartilhados entre os indivíduos aparentados. Já a matriz M' M indica o número de indivíduos homozigotos para cada alelo na diagonal e o número de vezes que os alelos de diferentes locos foram herdados pelo mesmo indivíduo.

Para o presente trabalho, foi estimada a matriz G (matriz de parentesco genômico aditivo) por meio do pacote AGHmatrix (AMADEU et al., 2016). O parentesco foi calculado usando o estimador de Van Raden admitindo o nível de ploidia diploide.

3.6 Coeficiente de parentesco

Através do programa computacional “Prócruza” (programa estatístico CTC), foram estimados os coeficientes de parentesco de Malecot (Malecot. 1948), o qual reflete a similaridade genética por descendência, a partir da informação de genealogia. Para construção do pedigree foram utilizadas informações de genealogia disponíveis no Banco de Dados CTC.

3.7 Análise de Componentes Principais

A análise de componentes principais foi realizada a partir da matriz de dissimilaridade genética (1 – Similaridade de Jaccard) das combinações dos

1230 genótipos. Para realizar a análise, foi utilizada a função *prcomp* do pacote stats do software estatístico R (R CORE TEAM 2020).

A escolha do número de componentes foi feita baseada na inspeção visual dos gráficos de dispersão dos scores dos componentes principais e intepretação do ponto de vista prático de cada componente.

Para comparar os resultados da estrutura de população com as informações a respeito da genealogia dos cultivares, foi feita uma análise de componentes principais com a matriz de distâncias cujo elementos foram definidos como 1− (coeficiente de parentesco de Malecot). O coeficiente de parentesco, é a probabilidade de que um alelo tomado ao acaso no indivíduo A seja idêntico por descendência a um alelo também tomado ao acaso no indivíduo B. Em outras palavras, é a probabilidade de que um alelo tomado ao acaso do indivíduo A tenha o mesmo ancestral do alelo do indivíduo B também tomado ao acaso. Dessa forma, a matriz de parentesco representa o grau de relacionamento entre cada possível par de indivíduos em uma amostra.

3.9 Análise de agrupamento

Com base nas matrizes de similaridade genética de Jaccard e parentesco com base no pedigree, foi realizado o agrupamento dos genótipos pelo método de agrupamento de Otimização de Tocher (função tocher, pacote biotools no R Core Team 2020), e coeficiente de correlação cofenética (função cophenetic do pacote stats no R Core Team 2020) para verificar a consistencia do agrupamento.

3.9 Coeficiente de correlação

Para determinar o nível de correlação entre a similaridade genética baseada nos marcadores moleculares e no coeficiente de parentesco via genealogia, realizou-se uma análise de correlação de matrizes, através do teste de Mantel (função cophenetic do pacote vegan no R Core Team 2020). Este teste faz inúmeras permutações nos valores das duas matrizes e calcula o valor da correlação (r) em cada permutação de matrizes. Após

isso, é realizado um teste estatístico buscando comparar a correlação real dos dados com quantil da correlação de 0,95 originado das inúmeras permutações para avaliar se a correlação entre as matrizes realmente existe, ou pode ser resultado simplesmente do acaso.

4 RESULTADOS E DISCUSSÃO

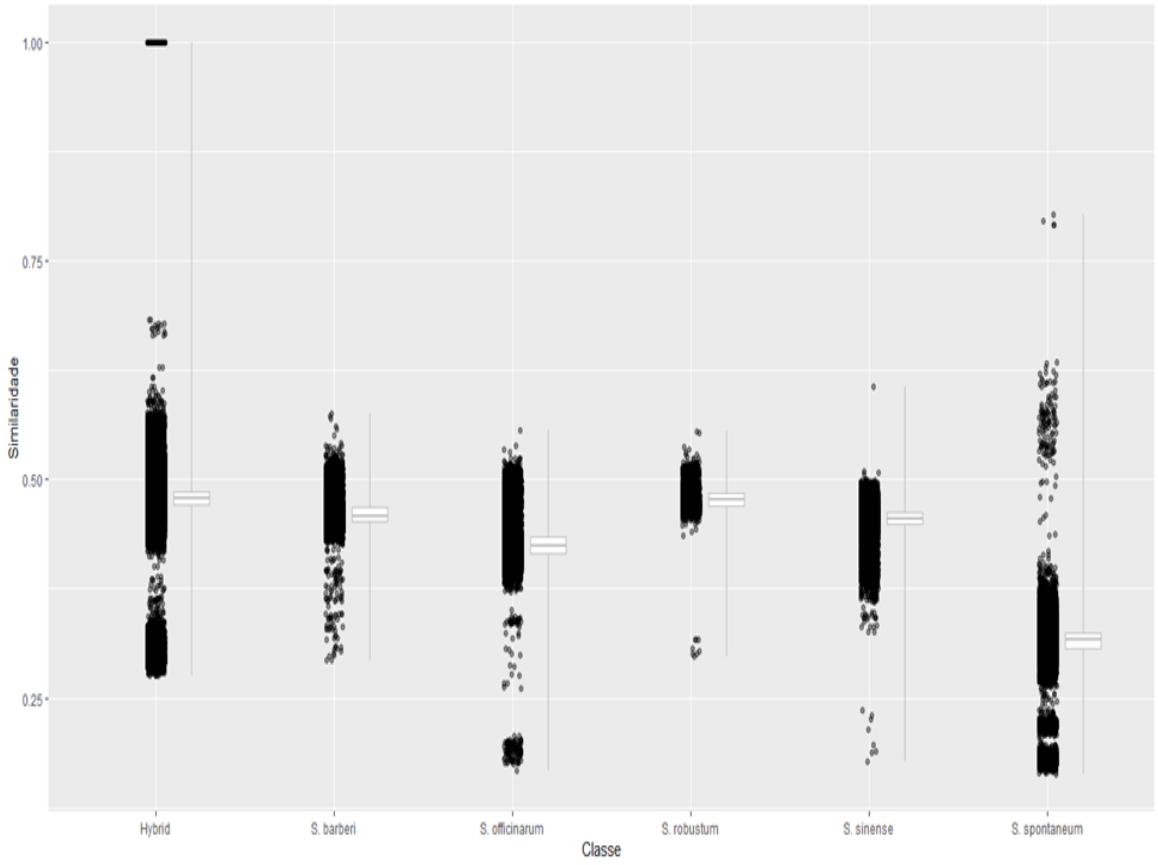
4.1. Similaridade genética de Jaccard

A similaridade genética estimada pelo índice de Jaccard apresentou ampla variação. A maior divergência genética (0,14) foi estimada entre os acessos Kletak (*S. spontaneum*) e NG5136 (*S. officinarum*), enquanto os acessos NG7770 (*S. officinarum*) e NG57118 (*S. officinarum*) se mostraram menos divergentes (0,89). Em média, o coeficiente de Jaccard foi de 0,46, considerando as similaridades entre os 1230 acessos utilizando as informações de marcadores moleculares, indicando valor médio próximo ao relatado por Crystian et al. (2018), de 0,448.

Silva (2012) analisando a diversidade genética de clones elites de cana-de-açúcar, observou um coeficiente de Jaccard médio de 0,49, variando de 0,32 a 0,71 nas 1953 combinações obtidas, usando as informações de marcadores moleculares EST-SSR.

Na Figura 1 é evidenciado que os híbridos modernos apresentam maior similaridade média com as espécies *S. robustum* (0,48), seguida por *S. sinense* (0,48), *S. barberi* (0,47), *S. officinarum* (0,43) e *S. spontaneum* (0,32). Esses resultados são coerentes com o histórico evolutivo da espécie, dado que as espécies *S. sinense* são híbridos interespecíficos entre *S. officinarum* e *S. spontaneum* com maior contribuição no genoma da espécie *S. officinarum* (Silva; et al., 2002).

Figura 1 – Grau de relacionamento dos Grupos (Híbridos Modernos, S. officinrum, S. barberi, S. robustum, S. sinensi e S. spontaneum) com os Híbridos Modernos.



Foram observados valores acima 0,75 para acessos pertencentes a S. *spontaneum* (Figura 1). Em análise mais detalhada, detectou-se que estes valores se referem a clones oriundos de um projeto de Introgressão, onde o objetivo foi a incorporação de alelos de S. *spontaneum* ao acervo, através de retrocruzamentos entre um híbrido e sua original geração progenitora visando obter materiais para produção de biomassa. Diante das novas exigências para a cana-de-açúcar, como o uso de fibra para produção de etanol de segunda geração, cogeração de eletricidade, novos progenitores com diferentes proporções de *S. spontaneum* podem ser identificados e incorporado aos cruzamentos.

4.2 Coeficiente de parentesco de Malecot

Foram estimados os coeficientes de parentesco de Malecot entre os 1230 acessos de cana-de-açúcar. Os valores de parentesco variaram de 0 a 0,75, com média equivalente a 0,02. Os acessos SP911049 e CT002651 foram identificados como os mais aparentados, fato coerente, uma vez que o acesso CT002651 é oriundo de retrocruzamento envolvendo SP911049. Tais valores, se assemelham aos relatados por (Gheysa Coelho Silva., 2012) onde observou – se variação de 0 a 0,70, com média de 0,05 nas 1.953 combinações obtidas, usando informações de pedigree.

O uso dos coeficientes de parentesco e de endogamia tem sido frequente no gerenciamento de bancos de germoplasma associado aos programas de melhoramento de plantas e animais (PETERNELLI et al., 2009). Essas estimativas de coeficiente de parentesco têm sido usadas como indicadores preliminares das combinações mais divergentes, em programas de melhoramento nos estágios iniciais.

Uma característica inerente à estimação dos coeficientes de parentesco e de endogamia é a dependência sobre a informação do pedigree. Neste tocante, vale salientar que muitos dos registros genealógicos não são conhecidos, especialmente em se tratando de acessos mais antigos.

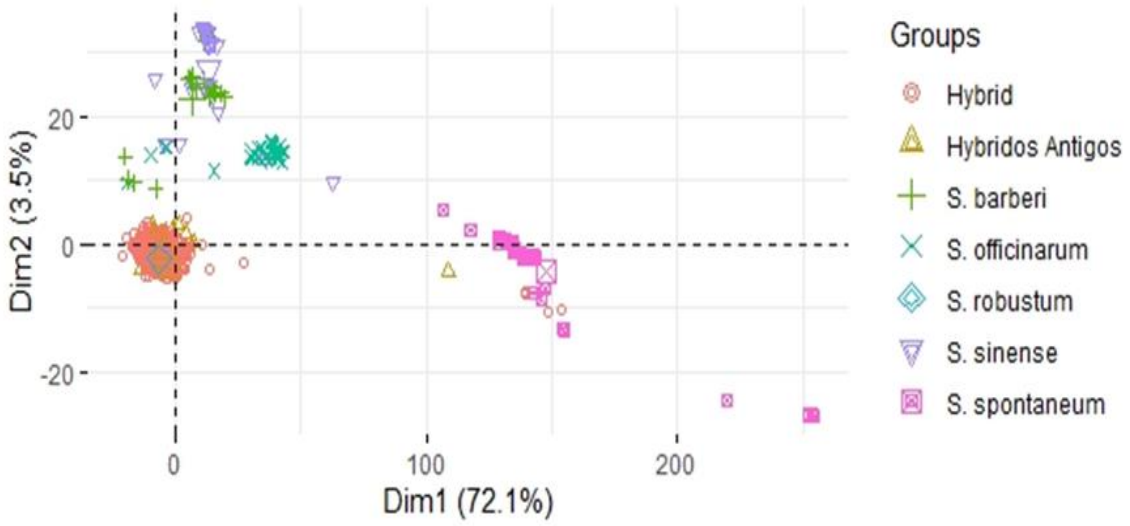
4.3 Estrutura populacional dos acessos pela análise de componentes principais

A análise de componentes principais das distâncias genéticas (1 – similaridade de Jaccard) revelou presença de estrutura nos dados referente aos acessos do banco de germoplasma de cana-de-açúcar do CTC. O primeiro componente (Dim 1) explicou 72,1% da variabilidade dos dados. Notou-se que os indivíduos de *S. spontaneum,* encontram-se com valores acima de 100, o grupo de *S. officinarum* com valores entre 0 e 50, e os demais grupos distribuídos entre os valores -25 e 25, sugerindo assim, a formação de 4 (quatro) subpopulações. A primeira, formada pelos acessos de híbridos modernos (Hybrid) e híbridos antigos, a segunda formada por acessos de *S.*

oficinarum, a terceira formada por acessos que são espécies não melhoradas de *S. barberi, S. robustum e S. sinensi,* e a quarta formada por *S. spontaneum* O segundo componente explicou 3,5 % da variabilidade dos dados e separou as espécies (*S. Sinense* e *S. barberi*) dos híbridos modernos (Figura 2).

Nas informações observadas (Figura 2), é evidenciado a história evolutiva da cana-de-açúcar. As S. barberi e S. sinense foram originadas a partir de cruzamentos interespecíficos entre S. officinarum e S. spontaneum. Genótipos cultivados atualmente, surgiram principalmente da hibridação das espécies S. officinarum e S. spontaneum realizada pelos primeiros programas de melhoramento genético de cana-de-açúcar (Landell e Bressiani, 2008; D'Hont et al., 2002). Após o primeiro cruzamento, foram realizados sucessivos retrocruzamentos entre híbrido F1 com S. officinarum, a fim de aumentar o teor de sacarose (Processo de Nobilização da cana-de-açúcar). No entanto a cada geração era observada aumento da frequência do genoma de S. officinarum em relação ao genoma de S.spontaneum. Como consequência dessa sucessão de cruzamentos, os cromossomos das espécies genitoras se recombinaram perdendo a sua estrutura inicial, levando a um aumento ainda maior da complexidade genômica (Crystian et al., 2018).

Figura 2 - Biplot análise de componentes principais dos dados de Dissimilaridade Genética (1 – Similaridade de Jaccard), dos 1230 acessos.



Por sua vez, a análise de componentes principais aplicada à matriz de distâncias definida por 1 − coeficiente de parentesco de Malecot mostra resultados contrastantes, onde não é possível observar uma estrutura nos dados, assim como observado na Figura 2. O primeiro componente (Dim 1) explicou apenas 14,7 % da variabilidade dos dados e a segunda coordenada explicou (Dim 2) 6,2 % (Figura 3). Provavelmente, as razões das diferenças dos resultados entre análises de componentes principais, deve-se aos registros faltantes e a algumas pressuposições irrealistas feitas para a determinação do coeficiente de parentesco. Por exemplo, assume-se ausência de deriva genética e seleção, ausência de falhas e de erros de anotação dos registros da genealogia (Silva, 2013), e não leva em consideração as segregações genéticas que ocorrem de uma geração a outra, baseando-se na média (esperança) do parentesco esperado entre os diferentes relacionamentos (p.ex. meios-irmãos e irmãos-completos).

O coeficiente de parentesco, embora altamente informativo em um programa de melhoramento, apresenta erros inerentes durante a sua estimativa, resultando em valores com alguns vieses. Isto é, em parte, devido a algumas suposições genéticas, que são assumidas no cálculo do coeficiente de

parentesco. A suposição de que o genótipo recebe a mesma quantidade de alelos de cada pai é questionável. A cana-de- açúcar é poliplóide e altamente heterozigótica, além do fato bem conhecido que quando se utiliza S. officinarum como fêmea, a sua meiose não é equivalente, resultando na vantagem de um parental sobre o outro (BREMER, 1961). Portanto, as estimativas de similaridade genética obtidas a partir de marcadores moleculares irá fornecer mais informações do que aqueles disponíveis através de informações de pedigree, quando este apresentar acentuado viés por falta de informações para construção do pedigree e consequentemente erro no cálculo destas estimativas.

Figura 3 - Biplot da análise de componentes principais dos dados de distância (1 – Coeficiente de Parentesco), de 1230 acessos.

4.4 Agrupamento de Tocher com informação de Genotipagem e pedigree.

Pelo método de agrupamento de Tocher a partir das distâncias (1 - coeficiente de Jaccard), os acessos foram distribuídos em 6 grupos, dos quais o Grupo 1 com 3 genótipos; Grupo 2 com 32 genótipos; Grupo 3 com 38

genótipos; Grupo 4 com 1154 genótipos; Grupo 5 com 2 genótipos; Grupo 6 com 1 genótipo. O grupo 4 concentrou 93% dos acessos.

Na Tabela 1, podemos observar as distancias médias entre os grupos, e na diagonal, encontram – se a distância média dentro dos grupos. De acordo com a tabela, podemos considerar os grupos 1 e 4 como mais distantes.

Tabela 1 – Distâncias observadas entre os grupos obtidos pelo método de Agrupamento de Tocher.

	Grupo 1	**Grupo 2**	**Grupo 3**	**Grupo 4**	**Grupo 5**	**Grupo 6**
Grupo 1	0,071871	8,904987	5,454882	10,33434	8,734475	1,646318
Grupo 2	8,904987	0,752044	5,039825	2,629158	2,642596	7,609636
Grupo 3	5,454882	5,039825	0,955082	5,82403	3,961491	4,146725
Grupo 4	10,33434	2,629158	5,82403	1,006431	2,177091	8,976846
Grupo 5	8,734475	2,642596	3,961491	2,177091	1,416951	7,367641
Grupo 6	1,646318	7,609636	4,146725	8,976846	7,367641	0

O valor de correlação cofenética (r = 0,95), indicando que a análise de agrupamento mostrou um ajuste muito bom para a matriz de distância genética.

Para os dados de coeficiente de parentesco com base no pedigree, a análise de agrupamento de Tocher gerou um total de 83 grupos, dos quais 23 contém apenas 1 genótipo. Valor de correlação cofenética encontrado nesta análise foi inferior, de (r = 0,59).

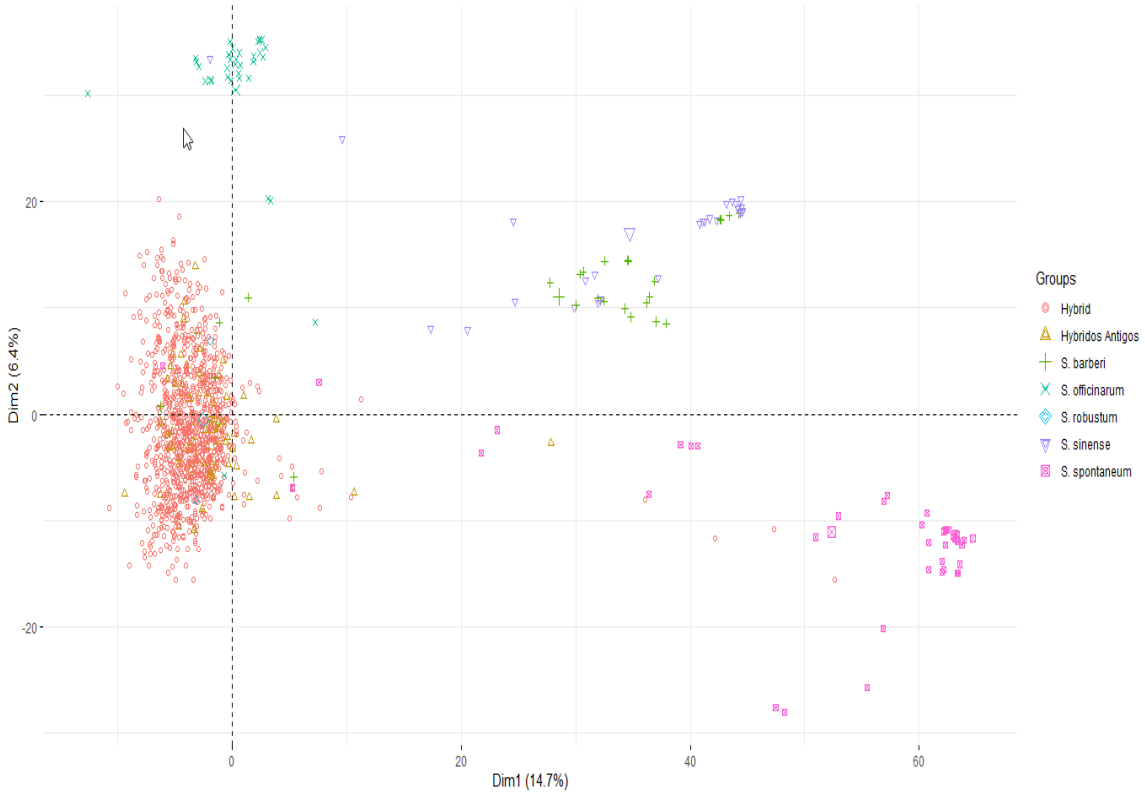
Observou - se grupos que apresentaram apenas 1 acesso cada, como pode ser visto nas Tabelas 1 (grupo 6). Por este método de otimização verifica-se que um número bastante reduzido de grupos reúne grande parte dos acessos. Vasconcelos et al. (2007) argumentam que este método apresenta uma inconveniência no agrupamento de genótipos com maior dissimilaridade: na maioria dos casos, cada genótipo forma um grupo específico (com apenas um genótipo) em virtude de este agrupamento ser influenciado pela distância dos genótipos já agrupados. Isso ocorre devido ao fato de o método se utilizar de um critério global de agrupamento, ou seja, baseia-se na maior entre as menores distâncias encontradas na matriz de dissimilaridade durante todo o

processo. Resultados desse tipo também foram encontrados por (Silva et al., 2014).

4.5 Matriz de parentesco genômico aditivo (G)

Com objetivo de avaliar a correlação da matriz de relacionamento genético estimada com base nos dados disponíveis, com a matriz de similaridade e coeficiente de parentesco, calculou- se a matriz G (matriz de parentesco genômico aditivo), através do pacote AGHmatrix (AMADEU et al. 2016). Na Figura 8 é representada o biplot dos genótipos com base na análise de componentes principais a partir das estimativas de parentesco da matriz G.

Figura 8 - Biplot da análise de componentes principais do relacionamento genético adtitivo genômico entre acessos de cana-de-açúcar.



Com a disponibilidade de dados moleculares, informações de marcadores de DNA para vários locos espalhados ao longo do genoma, podem

ser usados para medir a similaridade de indivíduos e calcular a matriz de parentesco genômico realizado com elementos que demonstram a verdadeira proporção do genoma em comum. O relacionamento genético aditivo pode melhor estimar a proporção genética compartilhada entre indivíduos. Assim, tornou-se possível à utilização de informações mais precisas sobre os alelos idênticos por descendência (IBD), por possuírem um ancestral comum no genoma de um indivíduo, e alelos idênticos independentemente de serem ou não herdados de um antepassado recente (IBS), que podem ser compartilhados por ancestrais comuns, ausentes no pedigree, tornando possível a utilização da matriz de parentesco genômico denominada Genomic Relationship Matrix (GRM) (Forni et al., 2011). O que contribui de forma significativa para obtenção de estimativas de parâmetros populacionais mais acurados.

O relacionamento genético variou de -0,3 (valores observados entre MOL1032 S. spontaneum e IK7635 S. officinarum), e 1,08 (S. spontaneum SES147B e MOL1032 S. spontaneum). Comparando estas relações, com as estimativas encontradas com no coeficiente de Jaccard, podemos observar os seguintes valores 0,16 (valores observados entre MOL1032 S. spontaneum e IK7635 S. officinarum), e 0,62 (S. spontaneum SES147B e MOL1032 S. spontaneum), evidenciando correlação entre as estimativas.

Pode ser visto que os elementos fora da diagonal das matrizes genômicas podem assumir valores negativos. Hayes et al. (2009) explicam que os coeficientes de parentesco são sempre relativos à uma população base, assim subtraindo os valores dos elementos contidos em M pela frequência média dos alelos, faz com que os relacionamentos contidos em G passem a ser relativos à atual população. Consequentemente, o parentesco médio se aproxima de 0 e alguns elementos podem assumir valores negativos.

4.6 Coeficiente de correlação

Para determinar o nível de correlação entre as matrizes geradas neste trabalho, realizou - se uma análise de correlação de matrizes, através do teste

de Mantel. Na tabela 2, podemos observar os valores de correlação obtidos e o resultado do teste de Mantel.

Tabela 2. Correlação entre matrizes de similaridade e teste de Mantel, (com significancia = 0.001; 999 permutações).

	Pedigree	**VanRaden**
Jaccard	0,14	0,45
Pedigree	-	0,17

Observou-se maior correlação entre as matrizes baseadas nas informações de marcadores moleculares. Quando comparadas as matrizes baseadas nos marcadores com a matriz de parentesco de Malecot, obtida a partir das informações de pedigree, foram observadas correlações de baixa magnitude. Tais discrepâncias, podem estar relacionadas ao fato de que os métodos utilizados foram desenvolvidos para diploide; além do fato de que na estimação do coeficiente de parentesco de Malecot, assume-se ausência de deriva genética e seleção, contribuição igual dos gametas oriundos dos genitores, ausência de falhas e de erros de anotação dos registros da genealogia, falta de registros.

5 CONCLUSÕES

Os marcadores SNPs foram eficientes para estudos de estrutura populacional em cana-de- açúcar. Os resultados das análises de componentes principais revelaram presença de estrutura de população que tem relação com a ordem taxonômica dos indivíduos, e histórico evolutivo da cultura-de-açúcar.

As informações de coeficiente de Jaccard, assim como a estimativa de relacionamento genômico podem auxiliar na determinação das combinações entre os genitores.

REFERÊNCIAS BIBLIOGRÁFICAS

Almeida, C. M. A. De, Lima, S. E. N. De, Lima, G. S. D. A., Brito, J. Z. De, Donato, V. M. T. S., & Silva, M. V. Da. (2009). Caracterização Molecular De Cultivares De Cana-De-Açúcar. *Ciência e Agrotecnologia*, *33*(spe), 1771–1776. https://doi.org/10.1590/S1413-70542009000700012

Amadeu, R. R., C. Cellon, J. W. Olmstead, A. A. F. Garcia, M. F. R. Resende, and P. R. Muñoz. 2016. AGHmatrix: R Package to Construct Relationship Matrices for Autotetraploid and Diploid Species: A Blueberry Example. The Plant Genome 9. doi:10.3835/plantgenome2016.01.0009

Bresolin, T. (2015). *Efeito da utilização de diferentes critérios de controle de qualidade dos genótipos em estudos de associação e seleção genômica ampla*. 53.

Cavalcante, M., Ferreira, P. V, & Pereira, R. G. (2010). Desempenho agronômico , dissimilaridade genética e seleção de genitores de batata doce para hibridização Agronomic performance , genetic dissimilarity and parental selection of sweet potato for hybridization. *Revista Brasileira de Ciências Agrárias*, *82*, 485–490. https://doi.org/10.5239/agraria.v5i4.816

Cesnik, R. (2005). Melhoramento da cana-de-açucar: marco sucro-alcooleiro no Brasil. *Revista Eletrônica de Jornalismo Científico*.

Crystian, D., Santos, J. M. dos, Barbosa, G. V. de S., & Almeida, C. (2018). Genetic diversity trends in sugarcane germplasm: Analysis in the germplasm bank of the RB varieties. *Crop Breeding and Applied Biotechnology*, *18*(4), 426–431. https://doi.org/10.1590/1984-70332018v18n4n62

Duarte, L. S. C. F. (2012). ANÁLISE DA VARIABILIDADE GENÉTICA E DE CARACTERÍSTICAS AGROINDUSTRIAIS EM CANA-DE-AÇÚCAR. *Psychology Applied to Work: An Introduction to Industrial and Organizational Psychology, Tenth Edition Paul*, *53*(9), 1689–1699. https://doi.org/10.1017/CBO9781107415324.004

Dutra Filho, J. de A., Melo, L. J. O. T. de, Resende, L. V., Anunciação Filho, C. J. da, & Bastos, G. Q. (2011). Aplicação de técnicas multivariadas no estudo da divergência genética em cana-de-açúcar. *Revista Ciência Agronômica*, *42*(1), 185–192. https://doi.org/10.1590/s1806-66902011000100023

E, P. D. E. P. E. M. G., Da, K., & Carneiro, S. (2017). *Caracterização genética de uma população base do programa de melhoramento de cana-de-açúcar da ridesa/ufg*.

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., … Souza, A. P. (2013). SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. *Scientific Reports*, *3*, 1–

10. https://doi.org/10.1038/srep03399

Gheysa, C., De, F. J., Filho, A., Clodoaldo, J., Neto, S., Djalma, E., & De, L. J. O. T. (2014). *Divergência genética entre genótipos de cana-de-açúcar*. *October*.

Huang, K., Guo, S. T., Shattuck, M. R., Chen, S. T., Qi, X. G., Zhang, P., & Li, B. G. (2015). A maximum-likelihood estimation of pairwise relatedness for autopolyploids. *Heredity*, *114*(2), 133–142. https://doi.org/10.1038/hdy.2014.88

Loci, M. pairwise relatedness in autopolyploids: a simulation study considering linkage and many. (2018). *Molecular pairwise relatedness in autopolyploids: a simulation study considering linkage and many loci*.

Lopes, V. R., Filho, J. C. B., Daros, E., Oliveira, R. A., & Guerra, E. P. (2014). Divergência genética entre clones de cana-de- açúcar usando análise multivariada associada a modelos mistos. *Semina:Ciencias Agrarias*, *35*(1), 125–134. https://doi.org/10.5433/1679-0359.2014v35n1p125

Marta, M., Laura, P. Æ., Teixeira, H. M., & Figueira, V. Æ. (2007). *Functional integrated genetic linkage map based on EST-markers for a sugarcane (Saccharum spp .) commercial cross Functional integrated genetic linkage map based on EST- markers for a sugarcane (Saccharum spp .) commercial cross*. *August*. https://doi.org/10.1007/s11032-007-9082-1

Meyer, A. da S. (2002). Comparação de Coeficientes de Similaridade usados em Análises de Agrupamento com Dados de Marcadores Moleculares Dominantes. *Usp*, 106.

Racedo, J., Gutiérrez, L., Perera, M. F., Ostengo, S., Pardo, E. M., Cuenya, M. I., Welin, B., & Castagnaro, A. P. (2016). Genome-wide association mapping of quantitative traits in a breeding population of sugarcane. *BMC Plant Biology*, *16*(1). https://doi.org/10.1186/s12870-016-0829-x

Resende, L. V., Bastos, G. Q., & Machado, P. R. (2013). Utilização de marcadores moleculares RAPD e EST ' s SSR para estudo. *Revista Ciência Agronômica*, *44*(1), 141–149. https://doi.org/10.1590/S1806-66902013000100018

Rosa, J. R. B. F. (2011). *Análise do desequilíbrio de ligação e da estrutura populacional do germoplasma brasileiro de cana-de-açúcar*. 97p.

Silva;, M. A., Gonçalves, P. S., Landell, M. G. A., & Bressiani, J. A. (2002). Estimates of genetic parameters and expected gains from selection of yield traits in sugarcane families. *Cropp Breeding and Applied Biotechnology*, *2*(4), 569–578. https://doi.org/10.12702/1984-7033.v02n04a10

Silva, Gheysa C., De Oliveira, F. J., Da Anunciação Filho, C. J., Neto, D. E. S., & De Melo, L. J. O. T. (2011). Divergência genética entre genótipos de cana-de-açúcar. *Revista Brasileirade Ciencias Agrarias*, *6*(1), 52–58. https://doi.org/10.5039/agraria.v6i1a848

Silva, Gheysa Coelho, & Cana-de-açúcar, C. E. M. (2012). *Diversidade Genética E Capacidade*.

Silva, R. R. (2013). *Estudo da estrutura populacional em cana-de-açúcar usando marcadores do tipo SNP*. 88p.

Toppa, E. V. B., & Jadoski, C. J. (2013). O Uso dos Marcadores Moleculares no Melhoramento Genético de Plantas. *Scientia Agraria Paranaensis*, *12*(1), 1–5. https://doi.org/10.18188/1983-1471/sap.v12n1p1-5