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

Estudo da estrutura populacional em cana-de-açúcar usando marcadores do tipo SNP

Renato Rodrigues Silva

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

Piracicaba
2013

Renato Rodrigues Silva
Engenheiro Agronômo

Estudo da estrutura populacional em cana-de-açúcar usando marcadores do tipo SNP

Orientador:
Prof. Dr. **ANTONIO AUGUSTO FRANCO GARCIA**

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

Piracicaba
2013

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

Silva, Renato Rodrigues
Estudo da estrutura populacional em cana-de-açúcar usando marcadores do tipo SNP / Renato Rodrigues Silva.- - Piracicaba, 2013.
88 p: il.

Tese (Doutorado) - - Escola Superior de Agricultura "Luiz de Queiroz", 2013.

1. Cana-de-açúcar 2. Estrutura populacional 3. Mapeamento genético 4. Marcador molecular 5. Melhoramento genético vegetal I. Título

CDD 633.61
S586e

AGRADECIMENTOS

A DEUS.

Ao Departamento de Genética da ESALQ/USP, pela oportunidade concedida.

Ao CNPq pela bolsa de estudos de doutorado.

Ao Prof. Dr. Antonio Augusto Franco Garcia, pela receptividade inicial, pelos ensinamentos, confiança e amizade, sempre com seu grande incentivo.

Aos meus pais Osmar Marinho Silva e Derci Rodrigues Brito Silva, pelo apoio, ensinamentos e carinho Se cheguei até aqui, foi por eles.

Ao meu irmão Rafael Rodrigues Silva e minha cunhada Cristiane Alves pela amizade e convivência.

À professora Drª. Anete Pereira de Souza, à Drª Monalisa Sampaio Carneiro e aos pesquisadores Dr. Thiago Marconi e Drª Melissa de Oliveira Santos Garcia pela parceria, disponibilizando os dados moleculares.

Ao pessoal do Laboratório de Genética-Estatística: Adriana, Carina, Edjane, Gabriel, Graciela, Guilherme, João, Maria Marta, Marcelo, Maria Izabel, Letícia, Luciano, Rafael, Rodrigo, Rodrigo Gazaffi, pelos valiosos ensinamentos e pela agradável convivência em todos os momentos.

À toda a minha família, amigos e colegas que estiveram presente em minha vida e contribuiram para meu crescimento pessoal e profissional, muito obrigado.

SUMÁRIO

RESUMO 7
ABSTRACT 9
1 INTRODUÇÃO 11
2 REVISÃO BIBLIOGRÁFICA 13
2.1 Aspectos da Cana-de-Açúcar 13
2.2 Origem, domesticação e melhoramento da cana-de-açúcar 15
2.3 Marcadores moleculares 16
2.3.1 Marcadores moleculares do tipo SNP 17
2.4 Classificação dos dados de SNPs 19
2.4.1 *Software* SuperMASSA 20
2.5 Mapeamento Associativo 22
2.7 Desequilíbrio de Ligação 24
2.8 Estrutura Populacional 25
2.8.1 Análise de Componentes Principais 27
2.8.1 *Software* STRUCTURE 28
3 MATERIAL E MÉTODOS 33
3.1 Material 33
3.1.1 Painel de Genótipos 33
3.1.2 Genotipagem dos indivíduos 33
3.2 Métodos 35
3.2.1 Classificação dos dados de SNPs 35
3.2.2 Imputação de dados 35
3.2.3 Estrutura Populacional 37
3.2.3.1 Análise de componentes principais e coordenadas principais 37
3.2.3.1 Análise de agrupamento 38
3.2.3.1 *Software* STRUCTURE 39

4 RESULTADOS 41
4.1 Classificação dos dados de SNPs 41
4.2 Imputação de dados 42
4.3 Estrutura Populacional 43
4.3.1 Análise de componentes principais e coordenadas principais 43
4.3.2 Análise de agrupamento 45
4.3.3 *Software* STRUCTURE 47
5 DISCUSSÃO 59
6 CONCLUSÃO 65
REFERÊNCIAS 67
APÊNDICE 83

RESUMO

Estudo da estrutura populacional em cana-de-açúcar usando marcadores do tipo SNP

Embora já existam estudos anteriores a respeito da estrutura de população em cana-de-açúcar, até o momento nenhum estudo foi feito usando marcadores SNPs gerados a partir de plataformas de genotipagem de larga escala, como por exemplo, Sequenom iPLEX MassARRAY. No presente trabalho, foi investigada a estrutura populacional no painel brasileiro de variedades de cana-de-açúcar. Esse painel é formado por materiais elites, ancestrais importantes e cultivares utilizados em programa de melhoramento. Um total de 1033 marcadores SNPs foram utilizados para genotipar os acessos do painel. A classificação dos dados feita usando o *software* SuperMASSA. A estrutura de população foi analisada por meio de análise de componentes principais (ACP), análise de agrupamentos e usando o *software* STRUCTURE. Devido ao fato que no software STRUCTURE não é possível dados de marcadores moleculares provenientes de espécies poliploides com aneuplodia frequente, o conjunto de dados foi separado e analisado de acordo com nível de ploidias dos SNPs. Com a finalidade de comparar os resultados, foi feita uma análise de coordenadas principais na matriz de distância, com os elementos definidos por 1 - coeficiente de parentesco. A análise de componentes principais revelou presença de estrutura de população. O primeiro componente separou o acesso IN84-58 (*S. spontaneum*) dos outros acessos que por sua vez estão separados em três grupos: o primeiro grupo formado pelos acessos que são *S. sinense*, o segundo grupo formado pelos cultivares modernos de cana-de-açúcar e o terceiro grupo formado por acessos que são espécies *S. officinarum*. Resultados da análise de agrupamento usando distância de alelos compartilhados são condizentes com resultados da ACP. Por outro lado, análise de coordenadas principais e método de agrupamento UPGMA usando o coeficiente de parentesco mostraram uma maior dissimilaridade genética entre os acessos separando as progênies do cultivar RB72454 do grupo formado pelos genitores e ou progenies do cultivar NA56-79. A diferença entre os resultados da análise de componentes principais e de coordenadas principais é devido principalmente a pressuposições nas estimativas do coeficiente de parentesco que são irrealísticas. Com relação a análise feita com o software STRUCTURE, o número de subpopulações e a matriz Q estimada variou de acordo com nível de ploidia dos marcadores. De um modo geral, as análises de estrutura de população mostrou que há evidências de estreitamento da base genética dos acessos devido a cruzamentos recorrentes de indivíduos aparentados. Espera-se que estas informações sejam importantes para o mapeamento associativo e melhoramento genético da espécie.

Palavras-chave: Estrutura Populacional; Marcadores SNP; Cana-de-açúcar; Componentes principais; STRUCTURE

ABSTRACT

Evaluation of population structure in sugarcane using SNP markers

Although there are several studies inferring population structure in sugarcane, none of them have yet used SNP markers generated from high-throughput platforms, such as, Sequenom iPLEX MassARRAY platform. In this study, it was investigated the population structure in a Brazilian panel of sugarcane varieties. This panel is comprised by elite breeding materials, important ancestors, and cultivars mostly used by breeding programs. 1,033 SNP markers were scored. SNP genotype calling was made using software SuperMASSA. The population structure was analyzed via principal components analysis (PCA), cluster analysis and using the software STRUCTURE. Due to the fact that STRUCTURE is not possible to analyze molecular markers data scored on species with mixed ploidy level, the dataset was separated and analyzed according to SNPs level ploidy estimates. With purpose of comparing the results, it was made a principal coordinate analysis (PCO) of distance matrix, with elements defined by 1 - kinship coefficient. The principal components analysis revealed some structure. The first component separated out IN84-58 (Saccharum spontaneum) from the others acessions, whereby these others acessions were allocated in three groups, comprised by *S. sinense* species, sugarcane modern cultivars and *S. officinarum* species. Results from cluster analysis using allele shared distance are in agreement with PCA results. On the other hand, principal coordinates analysis and UPGMA hierarchical clustering method based on kinship coefficient showed a broader genetic dissimilarity between acessions, allocating RB72454 progenies apart from its parents and/or progenies of NA56-79. The difference of results between PCA and PCO are mainly due to irrealistics assumptions in the calculation of kinship. Regarding analysis from the STRUCTURE, the number of subpopulations and Q matrix estimated varied with level ploidy. In general, the study of population structure showed some evidence of narrow genetic distances between accessions, due to recurrent crosses between related individuals. The information presented hereby could be important for association mapping and sugarcane breeding program.

Keywords: Population structure; SNP marker; Sugarcane; Principal components; STRUCTURE

1 INTRODUÇÃO

A cana-de-açúcar (*Saccharum spp.*) é uma cultura de grande importância econômica para o Brasil porque é muito utilizada para produção de açúcar e álcool. Dados disponibilizados pela FAO (2012), a respeito do ano agrícola de 2010, mostra que o Brasil é o maior produtor mundial, com uma produção de aproximadamente 717.462 milhões de toneladas de cana-de-açúcar, o que equivale aproximadamente ao valor de 23.362 bilhões de dólares no mercado internacional. Um dos fatores do sucesso produtivo do setor sucroalcooleiro é o melhoramento genético, pois, segundo Landell e Bressiani (2008), o melhoramento destaca-se por permitir constantemente a obtenção de cultivares com requisitos adequados a interesses agronômicos e industriais.

O melhoramento genético da cana-de-açúcar baseia-se na seleção e clonagem de genótipos superiores de populações segregantes obtidas por meio de cruzamentos biparentais ou multiparentais (MATSUOKA; GARCIA; ARIZONO, 1999). Embora os programas de melhoramento de cana-de-açúcar já possuam metodologias bem estabelecidas para seleção e liberação de novos cultivares, atualmente o tempo de obtenção de um novo cultivar, que ocorre em média de 10 a 15 anos, ainda é considerado longo (RABOIM, 2008). Portanto, um dos objetivos dos melhoristas é buscar métodos para a obtenção de uma seleção mais precoce dos genótipos superiores. Neste contexto, estudos de mapeamento de QTL (*Quantitative Trait Loci*) podem ser úteis ao melhoramento, uma vez que em uma perspectiva futura podem ser usados em programas de seleção assistida por marcadores (MALOSETTI et al. 2007, HOSPITAL et al. 1997). Assim, o tempo para obtenção de novos cultivares poderia ser reduzido.

Em cana-de-açúcar, normalmente o mapeamento de QTL é baseado em progênie de irmãos completos. Entretanto, em um programa de melhoramento por diversas vezes há o interesse de detectar associações genéticas válidas para populações naturais, coleções de germoplasma, conjunto de materiais elite, entre outros. Nestes casos, pode ser mais adequada a utilização do mapeamento associativo, pois esta abordagem permite detectar associações em populações sujeitas a uma maior quantidade gerações de recombinação entre o marcador e o QTL. Por outro lado, diferentemente do mapeamento de QTL baseado em populações experimentais, em uma análise de mapeamento associativo a ligação física não é a única fonte de desequilíbrio de ligação. O desequílibro de ligação observado pode ser causado, entre outros fatores, por meio de relações de parentesco entre indivíduos e/ou existência de estrutura de populações. Assim, identificar a estrutura da população é uma etapa importante para evitar a presença de falsos positivos (FLINT-GARCIA; THORNSBERRY; BUCKLER, 2003; YU et al., 2005).

Estudos de estrutura de população já foram realizados em diversas plantas cultivadas, tais como: batata (D'HOOP et al. 2010), arroz (ZHANG et al. 2011), soja (Li et al. 2008), trigo (ZORIC et al. 2012). Considerando a cana-de-açúcar, nota-se que há um aumento do número de trabalhos publicados nessa área nos últimos 5 anos (RABOIM et al., 2008, LOPES, 2011, ROSA, 2011). Entretanto, nenhum dos estudos de estrutura de população envolvendo cana-de-açúcar utilizaram marcadores que permitiam distinguir a dosagem alélica em indivíduos poliplóides, isto é, o número de cópias de um determinado alelo em um determinado loco. Os estudos citados foram realizados usando marcadores SSR, AFLP e RFLP. Tais marcadores têm herança dominante em espécies com alto grau de ploidia como a cana-de-açúcar, tornando seu uso relativamente limitado.

Com o surgimento das plataformas de genotipagem de larga escala, como por exemplo, o Sequenom iPLEX MassARRAY (OETH et al. 2009) e de novas metodologias estatísticas para classificação do genotipo de cada indivíduo em cada loco (*SNP calling*), é possível estimar a ploidia e dosagem alélica de indíviduos poliploides para cada loco. Consequentemente, esse fato pode mudar o paradigma de análises genético-estatísticas de dados de espécies como a cana-de-açúcar. Assim, neste trabalho propõe-se inferir a estrutura de populações das variedades comerciais e clones de interesse que constituem o painel brasileiro de cana de açúcar utilizando marcadores do tipo SNP, pois, para o caso de espécies poliploides, estes permitem levar-se em consideração o nível de ploidia e a dosagem alélica. As informações sobre a estrutura de população podem ser úteis ao desenvolvimento do mapeamento associativo e consequente implementação de métodos de seleção assistida por marcadores.

2 REVISÃO BIBLIOGRÁFICA

2.1 Aspectos gerais da cana-de-açúcar

A cana-de-açúcar é uma planta cultivada semi-perene que se desenvolveu principalmente em regiões tropicais e subtropicais. Embora seja uma planta alógama, nas lavouras comerciais normalmente reproduz-se por meio propagação vegetativa (MING et al., 2006).

Botanicamente, a cana-de-açúcar pertence à família *Poaceae*, tribo *Andropogoneae*, subtribo *Saccharine* que incluem os gêneros *Saccharum* e outros gêneros relacionados, tais como *Erianthus* e *Miscanthus* (HENRY, 2010; STEVENSON, 1965; ARTSCHWAGER; BRANDES, 1958). No genêro *Saccharum* estão incluídas duas espécies selvagens, *S. spontaneum* L. $(2n = 40-128)$ e *S. robustum* Brandes e Jeswiet ex Grassl $(2n = 60-205)$, e quatro espécies cultivadas *S. officinarum* L. $(2n = 80)$, *S. barberi* $(2n = 81 - 124)$, Jeswiet, *S. sinense* $(2n = 111 - 120)$ Roxb e *S. edule* Hassak $(2n = 60 - 80)$ (DANIELS; ROACH, 1987, AITKEN; McNEIL, 2010; IRVINE 1999, GRIVET 2006).

A espécie *S. spontaneum* possui o número básico de cromossomos igual a 8 e um nível de ploidia variável de 8 à 16, o que por consequência acarretou em uma grande variação nas suas características morfológicas (MUKHERJEE, 1957; AITKEN; McNEIL, 2010). Por exemplo, as folhas podem ser ausentes assim como podem ter largura de 4 centímetros, o caule é fino com diâmetro variando de 3 à 15 milímetros. Além disso, essa espécie apresenta uma maior distribuição geográfica dentre todas do gênero *Saccharum*. *S. robustum* possui caules pequenos a médios e baixo teor de sacarose. O habitat típico dessa espécie são margens de cursos d'águas e bancos de areia. Em relação as espécies cultivadas, a espécie *S. officinarum* tem como característica possuir colmos grossos e com alto teor de sacarose, por isso é denominada *cana nobre*. Além disso, possuem o número básico de cromossomos $x = 10$, indicando que a espécie é um autooctaplóide. As espécies *S. barberi* e *S. sinense* possuem caules médios a finos, com baixo a médio teor de sacarose, grande quantidade de fibras e alta tolerância a estresse comparada a *S. officinarum* (DANIELS; ROACH, 1987; STEVENSON, 1965)

2.2 Origem, domesticação e melhoramento da cana-de-açúcar

Os cultivares modernos de cana-de-açúcar são essencialmente híbridos interespecíficos entre *S. officinarum* e *S.spontaneum*, com contribuições das espécies *S. robustum*, *S. sinense*, *S. barberi* e gramíneas relacionadas ao genêro *Miscanthus*, *Narenga* e *Erianthus* (DANIELS; ROACH,

1987; MING et al., 2006). A espécie *S. robustum* surgiu a partir de hibridizações naturais entre *S. spontaneum* e espécies de outros gêneros relacionados (DANIELS; ROACH, 1987). Uma hipótese sobre a origem da espécie *S. officinarum* é que a mesma evoluiu a partir da espécie *S. robustum* e tem como centro de origem e de diversidade a região à leste da linha de Wallace na Nova Guiné. Posteriormente, migrou para Indonésia, Malásia, China, Índia, Micronésia e Polinésia durante o período pré histórico (ARTSCHWAGER; BRANDES, 1958; DANIELS; ROACH, 1987). Estudos baseados em hibridização genômica *in situ* (GISH) demonstraram que as espécies *S. barberi* e *S. sinense* foram origindas a partir dos cruzamentos interespecíficos entre *S. officinarum* e *S. spontaneum* no Norte da Índia e na China, respectivamente (D' HONT et al., 2002). Estes resultados corroboram os obtidos por Alix et al. (1999) utilizando a técnica *Southern Blot*. Por sua vez, *S. edule* é considerado um híbrido entre *S. officinarum* ou *S. robustum* e outros gêneros, tais como *Miscanthus* e *Erianthus*. (DANIELS; ROACH, 1987).

Os primeiros programas de melhoramento de cana-de-açúcar começaram em Java e Barbados em 1888. Em Java os melhoristas realizaram pela primeira vez a nobilização da cana-de-açúcar (STEVENSON, 1965; ARTSCHWAGER; BRANDES, 1958). A nobilização é o processo que consistiu na realização de cruzamentos interespecíficos, que envolveram as espécies *S. officinarum* e *S.spontaneum*, seguidos por sucessivos retrocruzamentos entre os híbridos e a espécie *S. officinarum*. A finalidade desses retrocruzamentos foi recuperar características agronômicas favoráveis, uma vez que a *S.officinarum* possui boas características gerais para a industrialização (MING et al., 2006). Entretanto, tal procedimento resultou em um aumento da complexidade do genoma da cana-de-açúcar, de modo que os genótipos atuais apresentam um genoma bastante complexo, com distintos números de cromossomos para cada grupo de homologia, ocorrência de mesmo cromossomo em diversos grupos de homologia e recombinação entre os cromossomos de diferentes espécies (GRIVET; ARRUDA. 2001; D' HONT, 2005; MATSUOKA; GARCIA; ARIZONO, 1999; HEINZ; TEW, 1987). Recentemente, foi feito um estudo baseado na técnica GISH com cultivares modernos de cana-de-açúcar. Descobriu-se que a porcentagem dos cromossomos desses cultivares que são derivados de *S.spontaneum* ou que são recombinates interespecíficos é de aproximadamente entre 15 % à 27,5 % (PIPERIDIS; PIPERIDIS; D'HONT, 2010).

Além dos trabalhos realizados em Java, um outro programa de destaque foi aquele estabelecido na estação de Coimbatore na Índia. Nesse programa, a noblização envolvia as espécies *S. spontaneum* e *S. barberi*. Os cultivares resultantes dessa estação foram indicados com a sigla *Co*, sendo que os cultivares Co213, Co281 e Co290 adaptaram-se muito bem as condições ambientais da região dos trópicos, tendo grande aceitação na região subtropical da Índia, Aus-

tralia, na Lousiana, Argentina e Brasil. A nobilização também foi praticada em Barbados e em outras estações produzindo vários cultivares. Esse método só entrou em desuso após 1930 (STEVENSON 1965; MING et al., 2006).

A partir de 1930, a estratégia dos melhoristas foi principalmente realizar cruzamentos entre cultivares que já tinham sido nobilizados. Por exemplo, o cruzamento entre POJ2878 com outros cultivares resultou no cultivar POJ3067, que ocupou mais de 85 % da área de cana-de-açúcar de Java. O cruzamento entre Co312 e POJ2978 deu a origem ao cultivar H32-8560, o mais importante do Havaí no ano de 1945. Pode-se citar ainda o cruzamento entre POJ2878 e Co290 que produziu o cultivar Co419, sendo que este foi importante na área tropical da Índia e para produção de açúcar no Brasil entre os anos 1940 e 1950 (SEGALLA, 1964). Dessa forma, pode-se dizer que ao longo dos anos o melhoramento genético da cana-de-açúcar aumentou significativamente a produtividade das lavouras de cana; contudo, nota-se que houve um estreitamento da base genética. Estima-se que os cultivares modernos foram desenvolvidos a partir de no máximo 15 a 20 genótipos ancestrais que foram nobilizados em Java ou em Coimbatore na Índia (ROACH, 1989; MING et al., 2006)

Atualmente, o melhoramento genético da cana-de-açúcar baseia-se na seleção e clonagem de genótipos superiores de populações segregantes obtidas por meio de cruzamentos sexuais entre indivíduos diferentes (MATSUOKA et al., 1999). Ainda segundo esses autores, nas etapas iniciais tem-se muito genótipos e poucas repetições de cada genótipo, o que permite apenas descarte e seleção de caracteres de alta herdabilidade. Conforme avançam as etapas do programa, diminui-se o número de genótipos e aumenta a disponibilidade de material de cada genótipo, o que permite avaliação de caracteres de média e baixa herdabilidade por meio de experimentos com repetições. Nas etapas finais, recomendam-se os cultivares baseando-se em resultados de experimentos realizados em diferentes locais e anos de cultivo. Em função disso, em um programa de melhoramento de cana-de-açúcar, a obtenção de um novo cultivar ocorre em média de 10 a 15 anos. Assim, um dos atuais desafios na área de melhoramento e genética de cana-de-açúcar é diminuir o tempo de obtenção de novos cultivares para reduzir o custos dos programas de melhoramento.

Neste contexto, tecnologias como marcadores moleculares aliados a métodos genético-estatísticos tais como mapeamento de QTL e mapeamento associativo podem ser muito úteis ao melhoramento. O recente desenvolvimento de técnicas e uso de marcadores moleculares permitiram um melhor entendimento dos processos do melhoramento, desde que marcadores moleculares propiciem informações a respeito da arquitetura genética dos caracteres quantitativos ao nível de DNA (PASTINA et al., 2010).

2.3 Marcadores moleculares

Marcador molecular é todo e qualquer fenótipo molecular oriundo de um fragmento de DNA codificante ou não de um gene. Tais marcadores são ferramentas úteis a estudo genômicos porque detectam polimorfimsmos ao nível de DNA e não sofrem qualquer tipo de influencia ambiental (FERREIRA; GRATTAPAGLIA, 1998; SOUZA; 2001).

Nos últimos anos, diversos tipos de marcadores moleculares foram desenvolvidos e normalmente podem ser classficados como marcadores dominantes e codominantes.

Quando utiliza-se marcadores dominantes o polimorfismo é detectado pela presença ou ausência de fragmentos, o que, por consequencia, impossibilita a distinção entre locos homozigotos e heterozigotos quando o fragmento está presente. Exemplos de marcadores dominantes são RAPD, *Random Amplified Polymorphic DNA*, (WILLIANS et al., 1990) e AFLP, *Amplified Fragment Length Polymorphism*, (ZABEAU; VOS, 1993). Os marcadores dessa classe não são dirigidos a uma região particular do genoma e revelam o polimorfismo presente em dezenas de locos simultaneamente (MOLLINARI, 2007; SOUZA, 2001).

O princípio do RAPD consiste em realizar uma reação de PCR utilizando-se o DNA do indivíduo em estudo e um *primer* único de sequência arbritrária. Dessa forma, mutações, deleções e ou inserções no sítio de iniciação podem causar a não complementariedade entre o DNA analisado e o *primer*, impedindo a amplificação do segmento (WILLIANS et al., 1990; SOUZA; 2001). Assim a detecção do polimorfismo é baseada na presença ou ausência do fragmento. AFLP é uma técnica baseada na amplificação seletiva de um subconjunto de fragmentos genômicos gerados após digestão do DNA realizada com auxílio de enzimas de restrição. Para essa técnica, o polimorfismo é revelado devido à presença ou ausência de um dado sítio de restrição entre dois genótipos (FERREIRA; GRATTAPAGLIA, 1998; SOUZA, 2001).

Os marcadores codominantes podem apresentar ambos alelos de um loco, o que permite distinguir locos homozigotos de heterozigotos. Os principais são os marcadores RFLP, *Restriction Fragment Length Polymorphism*, (BOTSTEIN et al., 1980); SSR *Simple Sequence Repeat*, (TAUTZ, 1989); EST-SSR, *Expressed Sequence Tag SSR*, (CATO et al., 2001) e SNP, *Single Nucleotide Polymorphisms*, (GUPTA, 2001). A técnica do RFLP é baseada na fragmentação do DNA por meio de enzimas de restrição e hibridização com sonda específica marcada radiotivamente ou com quimioluminescencia. Nesse caso, o polimorfismo é evidenciado devido a observação de fragmentos de restrição de diferentes tamanhos (FERREIRA; GRATTAPAGLIA, 1998; SOUZA; 2001). SSR ou microsatélites são sequências de DNA que contém poucos nucleotídeos, 1 à 6 pares de bases de comprimento, repetidas em tandem. Embora o número de

repetições em tandem de um microsatélite em um mesmo loco pode ser variável entre distintos genótipos, as sequências de bases adjacentes ao microsatélite pode ser única no genoma e conservada. Assim, pode-se desenhar um *primer* específico para as sequências adjacentes a um dado microsatélite de tal modo que por meio de uma reação de PCR é possível amplificar este loco em diferentes genótipos. Os produtos dessa amplicação mostrarão um polimorfismo de acordo com o tamanho do fragmento amplificado. Para visualização dos fragmentos, utiliza-se eletroforese de gel em poliacrilamida ou agarose especial de alta resolução. A visualização das bandas no gel pode ser feita diretamentre por coloração ou através de autoradiografia (FERREIRA; GRATTAPAGLIA, 1998; SOUZA, 2001; MARCONI, 2011). Geralmente, os alelos distintos de cada loco têm tamanhos suficientemente diferentes e migram para zonas separadas do gel. Com a disponibilização de banco de dados de sequências expressas, EST, tornou-se possível o desenvolvimento de SSR funcionais a partir dessas sequências. A busca por SSRs em ESTs é uma estratégia rápida e simples para o estudo da porção expressa do genoma, mesmo para os organismos com genomas grandes, complexos e redundantes, como a cana-de-açúcar (MARCONI, 2011). A vantagem da utilização desse tipo de marcador é que, se este estiver associado a uma característica de interesse, provavelmente o gene mapeado afetará diretamente essa característica. Além disso, pelo fato desses marcadores serem derivados de uma região codante, eles são mais conservados no genoma e, portanto, são mais facilmente transferíveis ao longo das gerações (CATO et al., 2001; MARCONI, 2011). Entretanto, no caso de organismos poliploides e aneuploides como a cana-de-açúcar, os marcadores SSR e EST-SSR comportam-se como dominantes uma vez que cada banda representa, para um determinado loco e indivíduo, pelo menos 1 cópia alélica de um determinado grupo de homologia. Sendo assim, a deteção do polimorfismo é baseada apenas na presença ou ausência de bandas, proporcionando apenas informações parciais do genoma da espécie (ROSA, 2011). Em outras palavras, tais marcadores não permitem determinar o número de cópias de cada alelo, mas sim apenas presença ou ausência do mesmo. Isto é, característico de locos com herança dominante.

2.3.1 Marcadores moleculares do tipo SNP

Polimorfismos de nucleotídeo único ou *Single Nucleotide Polymorphisms*, são variações na sequência de DNA que ocorrem quando um único nucleotídeo é alterado. Podem ser encontrados em regiões genoma com funções regulatórias e/ou regiões não codantes. Esse tipo de polimorfismo têm como característica serem bialélicos, amplamente distribuídos e conservados ao longo do genoma. Estima-se que um SNP ocorra a cada 100 a 300 pares de bases em qual-

quer genoma. Estudos feitos em aveia constataram a existência um SNP a cada 20 pares de bases (GUPTA, 2001; RAFALSKY, 2002). Além disso, a taxa de mutação normalmente é baixa, na ordem de 10^{-8} pares de bases por geração (HAMBLIN; WARBURTON; BUCKLER, 2007). Devido a essas características e aliado ao fato de que apresentam facilidades para a genotipagem em alta escala, (*"high-throughput"*), atualmente SNPs estão sendo muito utilizados como marcadores moleculares (GIANCOLA, 2006).

No caso da cana-de-açúcar, os SNPs possuem ainda a grande vantagem de apresentar herança codominante, conforme será apresentado. A identificação dos SNPs pode ser feita por meio de sequenciamento de nova geração, prospecção em bibliotecas genômicas *shotgun* e através de busca *in silico* em banco de dados de sequências de ESTs (RAFALSKY, 2002; SANTORO, 2010). A busca de SNPs nesses bancos de dados apresenta baixo custo relativo aos métodos que envolvem sequenciamento e, ainda, é possível escolher como alvo sequências homólogas a genes de interesse (MARCONI, 2011). Para o caso da cana-de-açúcar, tem-se utilizado o banco de dados SUCEST (Brazilian Sugarcane EST Project) que representa o maior e mais completo banco de dados de ESTs de cana-de-açúcar. Este banco contém 237.954 sequências ESTs, agrupadas em 43.141 clusters, o que poderia indicar cerca de 33.000 genes de cana-de-açúcar (GRIVET et al., 2001; MARCONI, 2011). Os principais métodos utilizados para a genotipagem de SNP em larga escala, têm-se, entre outros, o pirosequenciador (CORDEIRO, 2006) e o espectrômetro de massas (GABRIEL et al., 2002). Cordeiro et al. (2006) testaram o desempenho do pirosequenciador na genotipagem de SNPs obtidos a partir de genótipos de cana-de-açúcar oriundos da Austrália. Esses autores concluíram que a robustez do pirosequenciador é questionável para genotipagem de SNPs em locos pertencentes a grupo de homologia com elevados nível de ploidia e sugeriram o uso de plataformas de genotipagem baseada em espectrômetro de massas para tal finalidade.

Uma das plataformas de genotipagem em larga escala baseada em espectrometria de massas é o sistema Sequenom MassARRAY ® (OETH et al., 2009). Esse sistema consiste em, inicialmemnte, realizar uma reação de PCR para amplificação do fragmento de DNA que contém o SNP previamente identificado *in silico*, utilizando um par de primers. Após a reação de PCR, é realizado um tratamento com a enzima *shrimp alkaline phosphatase* (SAP), que tem como objetivo neutralizar os desoxinucletídeos trifosfato (dNTPs) não incorporados durante a PCR e que podem interferir nas etapas seguintes. A reação SAP cliva o grupo fosfato dos dNTPs não incorporados, convertendo-os para ddNTP dideoxinucletídeos, tornando-os nucleotídeos terminadores. Na etapa seguinte, ocorre uma reação de extensão primer denominada reação *iPLEX*. Um *cocktail* composto por primers, enzimas, buffers e nucleotídeos terminadores são adicionados.

Os produtos da reação de PCR e do *cocktail* são termociclados para processar a reação iPLEX, a qual envolve a adição enzimática dos nucleotídeos terminadores que são complementares a base presente na região do SNP. Dessa forma, a reação iPLEX Gold produz alelos específicos de diferentes massas, sendo que a massa de cada alelo é determinada por meio do espectrômetro de massa. Assim, o resultado da plataforma Sequenom MassARRAY ® consistem em os picos de alelos de menor e maior massa obtidos por de uma análise do espectro gerado pela amostra (MARCONI,2001; GABRIEL et al., 2002; SERANG; MOLLINARI; GARCIA, 2012). Com esses picos faz-se a inferência dos genótipos dos indivíduos para cada loco, que é denominada de "SNP *calling*" na literatura.

2.4 Classificação dos dados de SNPs

Para organismos diploides, já existem diversos métodos disponíveis para realizar tal classificação Ranade et al. (2001) e Olivier et al. (2002) utilizaram o método k-médias para tal propósito. Esse método pode prover resultados satisfatórios quando os grupos estão bem separados, mas nem sempre é efetivo, especialmente quando esses grupos possuem diferentes variâncias (FUJISAWA et al. 2004; Yan et al. 2008). Fujisawa et al. (2004) propuseram um método baseado no ajuste de misturas de distribuições normais com distintas variâncias usando uma função de verossimilhança penalizada. Teo et al. (2007) usaram uma mistura de distribuições t de Student truncada. No entanto, para organismos poliploides apenas alguns métodos foram desenvolvidos até o momento. Voorrips, Gort e Vosman (2011) apresentaram um método para autopoliploides baseada em ajuste de modelos de misturas Gaussianas, com o número de componentes igual ao número de classes genotípicas. Esse modelo pode ser expandido para permitir a inclusão de mais componentes, permitindo seu uso para outros tipos de autopoliploides. Entretanto, em certas situações o nível de ploidia é desconhecida, e consequentemente, o número de classes genotípicas também, havendo a necessidade de estimá-las. Serang, Mollinari e Garcia (2012) elaboraram um método de *calling* SNPs baseado em um modelo gráfico Bayesiano e implementaram no *software* SuperRMASSA. Esses modelos permitem fazer a inferência dos genótipos para os indivíduos de cada loco, mesmo se a ploidia seja desconhecida. A principal vantagem de tal abordagem consiste em classificar os genótipos de todos indivíduos simultaneamente, diferentemente dos demais métodos que faz a inferência do genótipos dos indivíduos um de cada vez. Além disso, esse método não é exclusivo para poliploides, ou seja, pode ser aplicado para espécies com qualquer nível de ploidia.

2.4.1 *Software* SuperMASSA

O uso dos modelos gráficos Bayesianos permite calcular a probabilidade dos dados observados dado os genótipos dos indivíduos para uma determinada ploidia. Esses modelos são construídos a partir do processo pelos quais os dados são gerados.

No software SuperMASSA estão implementados 3 modelos: modelo de população F_1; modelo de Hardy-Weinberg e modelo de população qualquer (sem pressuposições sobre a segregação). A seguir será descrito com maiores detalhes o modelo de Hardy-Weinberg, porque foi o utilizado no presente trabalho.

A figura 1 apresenta o modelo gráfico baseado nas pressuposições do equilíbrio Hardy-Weinberg, ou seja, população panmítica, ausência de mutação, migração, seleção e deriva genética. Nessa figura as letras dentro dos círculos representam as variáveis e/ou vetores aleatórios; as setas representam as dependências entre elas.

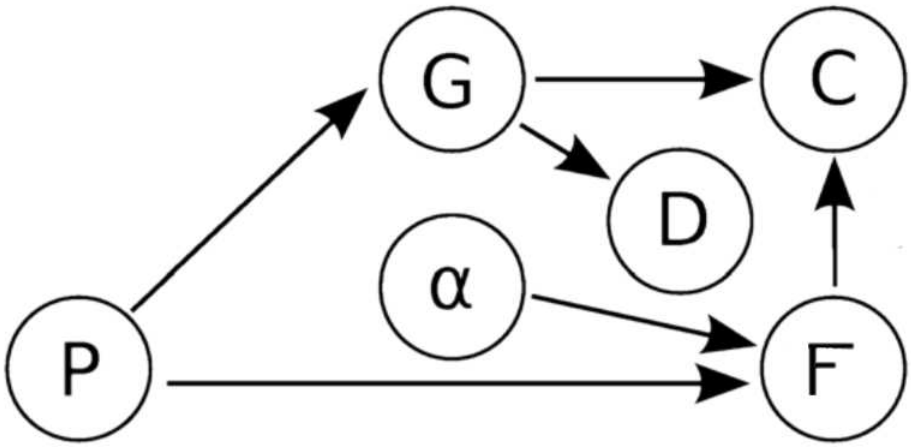


Figura 1 – Modelo gráfico bayesiano baseado no equilíbrio de Hardy Weinberg. **P** é a variável aleatória que representa um nível de ploidia, $\boldsymbol{\alpha}$ é o parâmetro que representa a frequência do alelo de referência na população, **G** o vetor aleatório que representa os genótipos atribuídos aos dados observados para dada uma ploidia, **D** o vetor dos dados observados, **C** o vetor aleatório que representa o número de genótipos pertencentes a cada classe genotípica, **F** o vetor aleatório que representam as frequências genotípicas na população e as setas representam a relação de dependência das variáveis e/ou vetores aleatórios

Neste modelo $Pr(P)$ é a distribuição *a priori* da variável aleatória nível de ploidia e $Pr(\alpha)$ é a distribuição *a priori* das frequencias alelicas do alelo de referência na população, sendo que assumiu-se que ambas seguem distribuição uniforme. Neste trabalho, o nível de ploidia esta sendo considerado como o número de cópias dos locos. Dado um nível de ploidia P , pode-se definir a distribuição condicional dos possíveis genótipos atribuídos a cada dado observado, $Pr(\mathbf{G}|P)$.

Por exemplo, para um loco diploide $P = 2$, o genótipo do i-ésimo indivíduo G_i pode assumir

equiprovavelmente os genótipos AA, Aa, aa, o que é equivalente aos valores $(2,0),(1,1),(0,2)$ tomando o alelo A como referência. Portanto, os números entre parênteses indicam o número de cópias de cada alelo. Por sua vez, para um loco tetraploide $P = 4$, G_i pode assumir equiprovavelmente os genótipos AAAA, AAAa, AAaa, Aaaa, aaaa, o equivale a $(4,0),(3,1),(2,2),(1,3),(0,4)$ e assim por diante.

Além disso, dado as frequências do alelo de referência α e um nível de ploidia P, obtêm-se a distribuição condicional dos parâmetros que representam as frequências genotípicas na população $F = \{f_1, \ldots, f_P\}$ que segue uma distribuição binomial, isto é,

$$Pr(F_j = f_j|\alpha, P) = \binom{P}{j}\alpha^j(1-\alpha)^{P-j}$$

Dado o vetor $\mathbf{F}$, pode-se obter a distribuição de probabilidade condicional do vetor aleatório $\mathbf{C}$ por meio da distribuição multinomial com parâmetros P e $\mathbf{F}$

$$Pr(C|F,G) = \frac{P!}{\prod_{j=0}^{P} c_j!}\prod_{j=0}^{P} f_j^{c_j}$$

em que c_j são o número de indivíduos pertencentes a uma determinada classe genotípica.

Por fim, têm-se a verossimilhança do modelo que é baseada na distribuição da projeção dos dados observados normalizados, $\hat{D}_i$, nas retas definidas pela configuração genotípica esperada G_i. A verossimilhança é definida por meio de

$$Pr(D_i|G_i = g_i) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp -\frac{1}{2}\left(\frac{\parallel \hat{D}_i - \hat{g}_i \parallel_2^2}{\sigma}\right)^2$$

em que o operador $\hat{\mathbf{u}} = \frac{\mathbf{u}}{\|\mathbf{u}\|}$ é usado para normalizar D_i e g_i usando a norma L_1, definida por $\parallel \mathbf{u}_i \parallel = \sum_{j=1}^{2} |\mathbf{u}_j|$, sendo $\parallel . \parallel_2^2$ é a norma Euclideana.

Dessa forma pode-se representar o modelo gráfico usando a distribuição de probabilidade conjunta que pode ser definida da seguinte forma:

$$Pr(P,\alpha,\mathbf{F},\mathbf{G},\mathbf{C},\mathbf{D}) = Pr(P)Pr(\alpha)Pr(\mathbf{F}|\alpha,P)Pr(\mathbf{G}|P)Pr(\mathbf{C}|\mathbf{G},\mathbf{F})Pr(\mathbf{D}|\mathbf{G}) \quad (1)$$

A inferência dos genótipos dos indivíduos para cada loco é feita enumerando-se todas as possíveis combinações de níveis de ploidias e classes genotípicas, e determinando qual a configuração maximiza a *probabilidade a posteriori* do modelo. Esta configuração é conhecida como

estimador *máximo a posteriori* e propricia a maior probabilidade possível. Para obtenção desse estimador, no software SuperMASSA foi implementada um método chamado inferência exata eficiente, cujos maiores detalhes podem ser encontrados em Serang, Mollinari e Garcia (2012).

Uma vez obtidos os valores dos parâmetros $\gamma = (\mathbf{P}, \alpha)$ que maximizam a probabilidade conjunta *a posteriori* dada por (1), é estimada a probabilidade *a posteriori* conjunta da configuração genotípica dos indivíduos e do vetor de parâmetros γ, condicionada aos dados observados:

$$Pr(\mathbf{G} = g, \gamma|\mathbf{D}) = \frac{Pr(\gamma)Pr(\mathbf{G} = g, \mathbf{D}|\gamma)}{\sum_{\gamma'} Pr(\gamma')Pr(\mathbf{G} = g, \mathbf{D}|\gamma')} \quad (2)$$

Essa probabilidade pode ser utilizada como um índicio de confiabilidade do ajuste do modelo aos dados pois quantifica o quão melhor é uma configuração genotípica comparadas com as outras. Concluindo, com emprego de tal método, é possível estimar a ploidia e classificar cada indivíduo de acordo com seu genótipo. Isto é essencial para que os dados de SNPs possam ser adequadamente usados em análises subsequentes.

2.5 Mapeamento Associativo

Mapeamento associativo é uma abordagem para identificar regiões cromossômicas associadas a caracteres poligênicos de interesse. Sua principal diferença com relação ao mapeamento de QTL's deve-se ao fato de não requerer a realização de cruzamentos biparentais. Essa método também é baseado no desequilíbrio de ligação existente entre marcadores polimorficos e os QTL's. Porém, diferentemente do mapeamento de QTL baseado em cruzamentos, este método detecta associações a partir de dados oriundos de populações naturais, coleções de germoplasma, conjunto de materiais elite, entre outros, de tal forma que a identificação de QTL's é feita em genomas constituídos por blocos gênicos menores. O mapeamento associativo explora portanto eventos de recombinações em um contexto histórico-evolutivo (YU; BUCKLER, 2006; ZHU et al., 2008). Por consequência, neste tipo de estudo há uma maior resolução do mapeamento, acarretando numa maior precisão na localização dos QTL's. Em contrapartida, são necessárias amostras de tamanhos maiores para detecção de associação uma vez que há uma maior variação na população (HUANG; AITKEN; GEORGE, 2010; ROSA, 2011). Obviamente, necessita também da disponibilidade de grande número de locos marcadores, já que as regiões em desequilíbrio são menores que aquelas presentes em cruzamentos biparentais.

Os estudos de mapeamento associativo começaram em humanos baseados em genes candidatos identificados por meio de mapeamento por análise de ligação (LANDER; SCHORK, 1994). Em plantas, um dos primeiros trabalhos teve como objetivo testar a associação do gene

candidato *dwarf8* com tempo de florescimento em 92 linhagens de milho (THORNSBERRY et al., 2001; HUANG; AITKEN; GEORGE, 2010). Os resultados desse trabalho indicaram uma série de polimorfismos, localizados próximos a região 5' do gene candidato *dwarf8*, associados com os fenótipos avaliados.

Entretanto, com o surgimento das plataformas de genotipagem de larga escala, estudos de mapeamento associativo no genoma como um todo (*"genome wide association studies"*, GWAS) estão sendo feitos rotineiramente em humanos, animais e plantas (HUANG; AITKEN; GEORGE, 2010). Por exemplo, Arazana et al., (2005) utilizaram 95 acessos de *Arabidopsis* para identificar associação entre SNPs genotipados ao longo do genoma para caracteres tais como tempo de florescimento e resistência a patógenos. Genes principais foram detectados demonstrando o potencial deste tipo de estudo em plantas. Huang et al. (2010) usaram essa mesma abordagem para estudar associações em 14 características agronômicas em um painel formado por 517 acessos de arroz genotipados com marcadores SNPs. Os locos identificados explicam 36% da variação fenotípica. Ainda, pode-se citar os estudos de GWAS feitos em soja (JUN et al., 2008), milho (TIAN et al., 2011), cevada (PASAN et al., 2012), trigo (GOUIS et al., 2012), sorgo (CANIATO et al. 2011), entre outros.

Com relação a cana-de-açúcar, poucos trabalhos envolvendo estudos de mapeamento associativo foram publicados. Uma das principais razões pode ser o fato de até o momento a maioria das plataformas de genotipagem em larga escala ainda não estarem adaptadas para trabalhar com espécies poliploides e com genoma tão complexo (HUANG; AITKEN; GEORGE, 2010). Pesquisadores da área tem trabalhado com marcadores dominantes como DArT ou com marcadores moleculares SSRs que se comportam como dominantes para essa espécie. McInyre et al. (2005) usaram mapeamento associativo para validar QTLs identificados para podridão radicular e ferrugem marrom baseado em 154 materais elites genotipados com aproximadamente 1000 marcadores de dose única incluindo RFLPs, AFLPs e SSRs. Para os dados do painel, 6 marcadores dos 13 que já haviam sido previamente identificados permaneceram em associação com a podridão radicular, e 7 de 15 permaneceram em associação com a ferrugem marrom. Esses resultados são promissores para o uso de marcadores moleculares e podem ser úteis para seleção de indivíduos resistentes a essas doenças em um programa de melhoramento. Wei et al. (2010) fizeram um estudo de mapeamento associativo para os caracteres teor de açúcar e produção em cana-de-açúcar, levando em consideração a estrutura de população, interação genótipo e ambiente e variação espacial. Esse estudo foi baseado em um painel de aproximadamente 480 acessos genotipados com marcadores DArT. Um grande número de marcadores foram significativos para os caracteres de teor de açúcar e produção. Todavia, não foram significativos os

efeitos de estrutura de população e a interação genótipo ambiente, o que pode ter inflacionado o número de associações positivas.

2.6 Desequilíbrio de Ligação

No contexto de análise de mapeamento associativo, o estudo do desequilíbrio de ligação ao longo do genoma é muito importante. O desequilíbrio ligação pode ser defindo como sendo uma associação não-aleatória, ou preferencial, de alelos de diferentes locos em uma população (FLINT-GARCIA; THORNSBERRY; BUCKLER, 2003; HEDRICK, 2010).

Normalmente, o desequilíbrio de ligação ocorre de forma aleatória no genoma e depende da espécie e da população estudada (ORAGUZIE et al., 2007; ROSA, 2011). Portanto, o estudo da extensão do desequilíbrio é muito importante, pois possibilita determinar a resolução de mapeamento e o número de marcadores necessários para identificar QTLs (ZHU et al., 2008; ROSA, 2011). Se houver desequilíbrio de ligação apenas em curtas distâncias no genoma, espera-se uma alta resolução de mapeamento e a necessidade de uma elevada quantidade de marcadores para identificar os QTLs. Caso contrário, se o desequilíbro se estender a maiores distâncias no genoma, espera-se uma menor resolução de mapeamento. Requerendo uma menor quantidade de marcadores (YU; BUCKLER, 2006; ORAGUZIE et al., 2007; ZHU et al., 2008; ROSA, 2011). Recentemente, estudos de extensão do desequilíbrio de ligação tem sido feitos em cevada (KRAAKMAN et al., 2004), milho (STICH et al., 2006), uva (BARNAUD; LACOMBE; DOLIGEZ, 2006), arroz (MATHER et al., 2007) entre outros.

Para o caso da cana-de-açúcar, Raboim et al. (2008) calcularam a extensão do desequilíbrio em um painel de 72 acessos de cana-de-açúcar genotipados com 1537 marcadores do tipo AFLP. Foi determinado a extensão do desequilíbrio até uma distância de 30 cM. Rosa (2011) estudou a extensão do desequilíbrio de ligação do painel brasileiro de variedades de cana-de-açúcar, que é composto por 135 acessos. O material foi genotipado com 1474 marcadores, incluindo SSRs e EST-SSRs. Foi constatada a extensão do desequilíbrio até 15 cM, confirmando que o genoma da cana-de-açúcar o desequílibrio permanece até mesmo em longas distâncias.

Apesar da ligação física ser um fator importante para causa do desequilíbrio. Vale ressaltar que o desequilíbrio de ligação pode ocorrer entre alelos de dois ou mais locos que estão bem distantes fisicamente (distância maiores de 50 cM) ou até mesmo em cromossomos diferentes. O desequilíbrio de ligação pode ocorrer nesta situações porque existem outros fatores que causam desequilíbrio de ligação, tais como: mutação, deriva, seleção, sistema reprodutivo, e estrutura de população. Destes fatores, a estrutura de população e a relação de parentesco são

os principais fatores para ocasionar falsos positivos em mapeamento associativo (PRITCHARD; PRZEWORSKI, 2001; FLINT-GARCIA; THORNSBERRY; BUCKLER, 2003; MALOSETTI et al. 2007; ROSA, 2011).

2.7 Estrutura Populacional

Estrutura de população refere-se a presença de subgrupos diferenciados geneticamente na população original, ou seja, uma população tem uma estrutura quando existem diferenças sistemáticas na ancestralidade dos indivíduos que as constituem (CANIATO et al. 2011, ASTLE; BALDING 2009). Para espécies domesticadas, a estrutura de população pode ser influenciada pela história natural das populações dos ancestrais pré-domesticados e pelos métodos de melhoramento empregados pelos humanos ao longo do tempo (GARRIS et al., 2005). Como já foi mencionado, esses tipos de estudos são importantes para o controle da presença de falsos positivos em mapeamento associativo e também em estudos evolutivos.

Thornsberry et al. (2001) fizeram um dos primeiros trabalho que considerou o efeito da estrutura populacional em estudos de mapeamento associativo. Mais tarde, Yu et al. (2006) apresentaram um modelo estatístico para mapeamento associativo que inclui como covariável vetores que representam a estrutura de população que posteriormente foi estimada em outros trabalhos (CASA et al., 2008; HANSEY et al., 2011).

Com relação a inferência de estudos de estrutura de população com aplicação em estudos evolutivos de plantas cultivadas, Garris et al. (2005) estudaram estrutura de população de arroz baseado em um painel de 224 acessos genotipados com 169 marcadores microsatélites. Utilizando o modelo mistura com frequências correlacionadas implementado no software STRUCTURE, foram encontradas 5 populações de arroz, coincidindo com os resultados obtidos por meio do método *neighbor-joining* e com a história evolutiva da espécie. Caniato et al. (2010) fizeram um estudo para testar se a tolerância de alumínio varia de acordo com a origem racial e geográfica de sorgos cultivados, concluindo que as subpopulações formadas por acessos da raça *Guínea* são mais tolerantes ao alumínio do que as subpopulações compostas por acessos da raça *Caudatum*.

Pode-se afirmar que estudos de inferência de estrutura de população podem ser aplicados na utilização e conservação de recursos genéticos coletados em bancos de germoplasma (ZHANG et al., 2011). Um número grande de acessos acumulados em bancos de semente reduz a eficiência com os quais estes recursos genéticos podem ser explorados. Assim, a determinação da estrutura de população das coleções de germoplasma pode ser importante para o processo de

amostragem dos acessos, visando a formação de coleções nucleares com menor repetitividade possível (BROWN, 1989; ODONG et al., 2011).

Existem diversas metodologias genético-estatísticas para inferência de estrutura de população. O método proposto por Pritchard, Stephens e Donnelly (2000) e implementado no software STRUCTURE é um dos mais usados em humanos, animais e em plantas (KAEUFFER et al. 2007). Aplicações desse método tem sido utlizado em estudo de estrutura de população de plantas cultivadas tais como batata (D'HOOP et al., 2010), tomate (SIM et al., 2011), cevada (CASTILLO et al., 2010) ente outras. Em cana-de-açúcar nota-se que os métodos mais utilizados para inferência de estrutura de população até o momento foram os métodos de agrupamento e o software STRUCTURE.

Para o caso da cana-de-açúcar, a maioria dos estudos de estrutura de população mostram que a presença de subgrupos coincide com a genealogia ou com a classificação botânica das espécies, tendo pouca relação com a origem geográfica dos acessos. Por exemplo, Jannoo et al. (1999) mostraram que não há evidências de que a diversidade genética entre uma coleção de clones de *S. officinarum* tenha relação com a distribuição geográfica. Por sua vez, Wei et al. (2006) estudaram o efeito de estrutura de populações em associações entre marcadores e resistência as doenças de podridão radicular, carvão e escaldadura em um painel de 154 acessos genotipados com 1068 marcadores AFLPs e 141 SSRs. Utilizando o software STRUCTURE, esses autores detectaram 8 subpopulações, sendo que esta classificação coincidiu com a estrutura da genealogia do painel analisado. Raboim et al. (2008) não encontraram estrutura de população no painel composto por 72 cultivares modernos de cana-de-açucar oriundos dos mais diversos programas de melhoramento do mundo. Esses autores especulam que a ausência de estruturação é devido a um intenso intercâmbio de cultivares entre os programas de melhoramento. Rosa (2011) estudou a estrutura de população do painel brasileiro de variedades de cana-de-açúcar e encontrou 4 subpopulações. Neste trabalho foi concluído que o os grupos formados estavam de acordo com as informações das genealogias.

Embora seja reconhecida a importância do software STRUCTURE em estudos de inferência de estrutura de população, diversos autores afirmam que o tempo de execução do algoritmo implementado no software STRUCTURE é muito alto quando se trabalha com grande quantidades de dados, podendo até inviabilizar o seu uso em alguns casos. Em contrapartida, análise de componentes principais ou ACP é um método com tempo de execução baixo, que não necessita de pressuposições sobre a estrutura de população e tem performance igual ou melhor que o software STRUCTURE (PATTERSON; PRICE; REICH, 2006; PRICE et al. 2006; ZHAO et al. 2007; MYLES et al. 2009; HUANG et al. 2010). Atualmente, está crescendo o número

de publicações sob este enfoque. Brown, Myles e Kresovich (2011) estudaram a congruência da classificação genética obtida com o uso de ACP, o software STRUCTURE e a classificação racial de sorgo. O material utilizado foi um painel de 216 acessos genotipados com 434 marcadores incluindo SNPs e SSRs. Ambos os modelos resultaram em uma classificação genética próxima a classificação racial. Myles et al. (2011) utilizaram ACP para verificar presença de subpopulações entre acessos selvagens e domesticados de uva oriundos de diferentes origens geográficas.

Além dos método de Pritchard Stephens e Donnelly (2000) e análise de componentes principais, métodos de agrupamentos também têm sido utilizados para se estudar estrutura de população. Esses métodos são baseados em matrizes de distâncias ou dissimilaridade. Entre os métodos de agrupamento pode-se citar os métodos *unweighted Pair Group Method with Arithmetic Mean* e Ward (ODONG et al. 2011) e *neighbor-joining* (SAITOU; NEI, 1987; SOKAL; MICHENER, 1958). No entanto a limitação desses métodos é que os resultados das análises são sensíveis à escolha do tipo de distância escolhida (PRITCHARD; STEPHENS; DONNELLY, 2000). Dada a importância dos métodos do STRUCTURE e ACP, detalhes sobre ambos serão apresentados em ítens específicos, a seguir.

2.7.1 Análise de Componentes Principais

A idéia central de análise de componentes principais é reduzir a dimensionalidade de um conjunto de dados capturando o máximo possível da variância das variáveis (JOLLIFFE, 1986, JOHNSON; WICHERN, 1992). No contexto de estrutura de população, análise de componentes principais pode ser definida como a projeção de indivíduos em um subespaço de menor dimensão de tal maneira que a localização desses indivíduos nesse espaço projetado revela similaridade genética entre eles (NOVEMBRE et al. 2008; ASTLE; BALDING 2009; ENGELHARDT; STEPHENS, 2010).

Considere uma matriz $\mathbf{M}$ em que as linhas representam os indíviduos, $i = 1, 2, \ldots, N$, as colunas representam os marcadores $l = 1, 2, \ldots, L$ e que cada elemento dessa matriz seja o número de cópia do alelo de referência, que pode variar de 0 até o nível de ploidia de cada loco.

Uma análise de ACP é baseada na decomposição por valores singulares da matriz de dados centralizada na média de cada coluna, ou seja,

$$\mathbf{M}_c = \mathbf{USV}^{'} = \sum_{l=1}^{L} s_{ii(l)} \mathbf{u}_l \mathbf{v}_l^{'} \tag{3}$$

em que $\mathbf{M_c} = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\mathbf{M}$; as N colunas da matriz $U_{N\times N}$ são os valores singulares à esquerda, as L linhas da matriz $V_{L\times L}$ são os valores singulares à direita equivalente aos *loadings* de uma análise ACP; S é uma matriz diagonal composta pelas raiz quadrada dos autovalores das matrizes $\mathbf{M}_c\mathbf{M}_c' = \mathbf{M}_c'\mathbf{M}_c$.

Os componentes principais $\mathbf{Y}$ podem ser definidos por meio de

$$\mathbf{Y} = \mathbf{M}_c\mathbf{V} = \mathbf{US} \tag{4}$$

sendo que as n-ésima coluna da matriz $\mathbf{Y}$ representa o n-ésimo componente.

Um fato que deve ser levado em consideração é que a determinação do número de componentes ainda é objeto de pesquisa e discussão na literatura. Existem diversos métodos propostos na literatura, tais como o *scree plot*, métodos baseados na proporção da variação explicada pelos componentes, entre outros (JOLLIFFE, 1986). Especificamente para estudos de estrutura de população, Patterson, Price e Reich (2006), propuseram um teste formal para autovalores baseado na distribuição Tracy-Windom; no entanto, normalmente esse método superestima o número de autovalores (SHRINER, 2011). Contudo, o mais importante nesse tipo de análise é verificar o significado do ponto de vista prático de cada componente (JOLLIFFE, 1986).

Deve-se mencionar que em situações nas quais se tem uma matriz de distâncias $\mathbf{\Theta}$ também pode ser aplicado o método de decomposição singular para reduzir a dimensionalidade dos dados, método denominado como análise de coordenadas principais.

Em síntese, a análise de coordenadas principais consiste em aplicar a decomposição de valores singulares na matriz $\mathbf{\Theta_c} = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\mathbf{\Theta}$, sendo que os autovetores obtidos pela decomposição são denominados como eixos de coordenadas principais. Aplicações de análise de coordendas principais em estudos de estrutura de populações em milho, batata e cana-de-açúcar podem ser visto em REIF et al., 2004; D'HOOP et al., 2010 e LOPES, 2011, respectivamente.

2.7.2 *Software* STRUCTURE

O método implementado na primeira versão do software STRUCTURE (PRITCHARD; STEPHENS; DONNELLY, 2000) é baseado nas pressuposições que cada subpopulação esteja em equilíbrio de Hardy-Weinberg e que dentro de cada subpopulação os locos estejam em equilíbrio de ligação. Em síntese, estima-se a frequência alélica das subpopulações e dado essas estimativas, atribui-se probabilisticamente os individuos nas suas respectivas subpopulações ou a mais de uma, quando assume-se que esses indivíduos possuem ancestrais oriundos de mais de uma população. Mais recentemente, o *software* STRUCTURE foi expandido permitindo o uso

dos modelos sem mistura, com mistura, modelo de ligação e modelo com uso de informação *a priori*, além de permitir o uso de alelos dominantes, alelos nulos e estimação de frequências alélicas correlacionadas (FALUSH, STEPHENS, PRITCHARD; 2003; 2007, HUBISZ et al. 2009).

O Modelo sem mistura ou *"No admixture"* pressupõe que o genoma como um todo dos indivíduos originou-se exclusivamente a partir de uma das K subpopulações. Nesse modelo, assume-se *a priori* que a probabilidade de um índividuo originar-se a partir uma determinada subpopulação seja $\frac{1}{K}$. Assume-se também que para cada subpopulação k e loco l a distribuição *a priori* das frequências alélicas $\boldsymbol{\pi}_{kl}$ segue uma distribuição de Dirichlet com parâmetros $(\lambda_1, \ldots, \lambda_{J_l})$. A verossimilhança do modelo sem mistura é análoga a distribuição de frequências alélicas sob Equílibrio de Hardy-Weinberg, ou seja,

$$Pr(\mathcal{G}|\mathbf{Z}, \mathbf{\Pi}) = \prod_{k=1}^{K}\prod_{l=1}^{L} \frac{Nc_{kl}!}{n_{kl1}! \ldots n_{klJ_l}!} \prod_{j=1}^{J_l} \pi_{klj}^{n_{klj}} \tag{5}$$

em que n_{klj} é o número de cópias do alelo j no loco l dos indivíduos com origem na população k, $Nc_{kl} = \sum_{j=1}^{J_l} n_{klj}$, $\mathcal{G}$ é o vetor aleatório que representam os genótipos e $\mathbf{Z}$ é o vetor aleatório que indica o origem dos indivíduos.

Dessa form,a por meio do Teorema de Bayes, tem-se a distribuição *a posteriori* conjunta, isto é,

$$\begin{aligned} Pr(\mathbf{Z}, \mathbf{\Pi}|\mathcal{G}) &\propto Pr(\mathbf{Z})Pr(\mathbf{\Pi})Pr(\mathcal{G}|\mathbf{Z}, \mathbf{\Pi}) \\ &= \prod_{i=1}^{N} \frac{1}{K} \prod_{k=1}^{K}\prod_{l=1}^{L} \frac{Nc_{kl}!}{n_{kl1}! \ldots n_{klJ_l}!} \prod_{j=1}^{J_l} \pi_{klj}^{n_{klj}} \prod_{k=1}^{K}\prod_{l=1}^{L} \frac{1}{\text{Beta}(\boldsymbol{\lambda})} \prod_{j=1}^{J_l} \pi_{klj}^{\lambda_j - 1} \end{aligned} \tag{6}$$

Como não é possível obter a distribuição a posteriori dada por (6) por meio de uma solução analítica, uma alternativa para se fazer inferência a respeito dos vetores aleatórios $\mathbf{\Pi}$ e $\mathbf{Z}$ é utilizar o algoritmo *Gibbs Sampling*. O algoritmo *Gibbs Sampling* tem como objetivo realizar simulações Monte Carlo da distribuição de interesse indiretamente, utilizando uma cadeia de Markov com probabilidades de transição igual o valor das probabilidades condicionais completas. Assim, após um número suficientemente grande de iterações, a cadeia converge para a distribuição de interesse (GELMAN; CARLIN, 1995).

Para o modelo sem mistura, o algoritmo *Gibbs Sampling* consiste em

- Simular $\mathbf{\Pi}^h$ a partir de $Pr(\mathbf{\Pi}^h|\mathcal{G}^{(h-1)}, \mathbf{Z}^{(h-1)}), \quad h = 1, 2, \ldots, H$

- Simular $\mathbf{Z}^h$ a partir de $Pr(\mathbf{Z}^h|\mathcal{G}^{(h-1)}, \mathbf{\Pi}^{(h-1)})$

em que $Pr(\mathbf{\Pi}|\mathcal{G},\mathbf{Z}) \propto \prod_{k=1}^{K}\prod_{l=1}^{L}\prod_{j=1}^{J_l}\pi_{klj}^{\lambda_j+n_{klj}-1}$ e

$$Pr(\mathbf{Z}|\mathcal{G},\mathbf{\Pi}) \propto \frac{Pr(\mathbf{\Pi})Pr(\mathbf{Z})Pr(\mathcal{G}|\mathbf{\Pi},\mathbf{Z})}{\sum_{k'=1}^{K}Pr(\mathbf{\Pi})Pr(\mathbf{Z})Pr(\mathcal{G}|\mathbf{\Pi},\mathbf{Z})}$$

.

O modelo com mistura ou *Admixture* permite que os ancestrais possam ser oriundos de diferentes populações. Isto é modelado declarando que o indivíduo i herdou alguma fração q_{ki} do genoma de um dos ancestrais de uma população k. Nesse caso, a distribuição conjunta *a posteriori* é dada por

$$\begin{aligned} Pr(\mathbf{Z},\mathbf{\Pi},\mathbf{Q}|\mathcal{G}) &\propto Pr(\mathbf{Q})Pr(\mathbf{\Pi})Pr(\mathbf{Z}|\mathbf{\Pi},\mathbf{Q})Pr(\mathcal{G}|\mathbf{\Pi},\mathbf{Q},\mathbf{Z}) \\ &= \prod_{i=1}^{N}\frac{1}{\text{Beta}(\boldsymbol{\nu})}\prod_{k=1}^{K}q_k^{\nu-1}\prod_{k=1}^{K}\prod_{l=1}^{L}\frac{1}{\text{Beta}(\boldsymbol{\lambda})}\prod_{j=1}^{J_l}\pi_{klj}^{\lambda_j-1}\prod_{i=1}^{N}\frac{\Omega_i!}{\omega_{1i}!\ldots\omega_{Ki}!}\prod_{k=1}^{K}q_{ki}^{\omega_{ki}} \times \\ &\quad \prod_{k=1}^{K}\prod_{l=1}^{L}\frac{Nc_{kl}!}{n_{kl1}!\ldots n_{klJ_l}!}\prod_{j=1}^{J_l}\pi_{klj}^{n_{klj}} \end{aligned}$$

em que ω_{ki} é o número de cópias alélicas de um indivíduo que são atribuídas a população k, $\Omega_i = \sum_{k=1}^{K} = \omega_{ki}$ e n_{klj} é o número de cópias do alelo j no loco l que são atribuídas a população k.

O algoritmo *Gibbs Sampling* consiste em

- Simular $\mathbf{\Pi}^h$ e $\mathbf{Q}^h$ a partir de

$$Pr(\mathbf{\Pi}^h,\mathbf{Q}^h|\mathcal{G}^{(h-1)},\mathbf{Z}^{(h-1)},\mathbf{Q}^{(h-1)}) = Pr(\mathbf{\Pi}^h|\mathcal{G}^{(h-1)},\mathbf{Z}^{(h-1)})Pr(\mathbf{Q}^{(h)}|\mathcal{G}^{(h-1)},\mathbf{Z}^{(h-1)},\mathbf{\Pi}^{h-1})$$

- Simular $\mathbf{Z}^h$ a partir de $Pr(\mathbf{Z}^h|\mathcal{G}^{(h-1)},\mathbf{\Pi}^{(h-1)},\mathbf{Q}^{(h-1)})$

sendo $Pr(\mathbf{Q}|\mathcal{G},\mathbf{Z},\mathbf{\Pi}) = Pr(\mathbf{Q}|\mathcal{G},\mathbf{Z}) \propto \prod_{i=1}^{N}\prod_{k=1}^{K}q_{ki}^{\omega_{ki}+\alpha-1}$ e

$$Pr(\mathbf{Z}|\mathcal{G},\mathbf{\Pi},\mathbf{Q}) = \frac{Pr(\mathbf{Q})Pr(\mathbf{\Pi})Pr(\mathbf{Z}|\mathbf{\Pi},\mathbf{Q})Pr(\mathcal{G}|\mathbf{\Pi},\mathbf{Q},\mathbf{Z})}{\sum_{k'=1}^{K}Pr(\mathbf{Q})Pr(\mathbf{\Pi})Pr(\mathbf{Z}|\mathbf{\Pi},\mathbf{Q})Pr(\mathcal{G}|\mathbf{\Pi},\mathbf{Q},\mathbf{Z})} \quad (7)$$

O modelo de ligação é uma expansão do modelo com mistura para levar em consideração o desequilíbrio de ligação causado pela mistura de populações. No entanto, esse modelo não

foi desenvolvido com o objetivo de modelar o desequilíbrio de ligação causado por ligação física entre os marcadores. A ideia é que os indivíduos dessas populações possuem cromossomos compostos por blocos que contêm alelos provenientes do mesmo ancestral gerando o desequilíbrio (FALUSH, STEPHENS, PRITCHARD; 2003). Portanto, há necessidade de que os locos não estejam ligados fisicamente para que possam ser utilizados.

Ainda, tem-se o modelo de uso da informação a priori, que permite o uso de informações de dados não genéticos para inferência da estrutura de população, como por exemplo a localização geográfica do local da amostragem (HUBISZ et al. 2009).

Com relação à inferência do número K de subpopulações, pode ser usado o cálculo de $Pr(K|\mathcal{G})$ dado por

$$Pr(K|\mathcal{G}) = Pr(K)Pr(\mathcal{G}|K) \tag{8}$$

Contudo, existem muitos problemas computacionais para obter estimativas acurada de $Pr(\mathcal{G}|K)$ tornando o problema difícil (PRITCHARD; STEPHENS; DONNELLY, 2000). Embora ainda não exista uma solução definitiva, alguns métodos foram propostos na literatura. Pritchard, Stephens e Donnelly (2000) propuseram uma solução *ad hoc* baseada em uma aproximação de $Pr(\mathcal{G}|K)$ dada por

$$Pr(\mathcal{G}|K) = \exp -\frac{\hat{\kappa}}{2} - \frac{\hat{\zeta}^2}{8} \tag{9}$$

sendo $\hat{\kappa} = \frac{1}{H}\sum_{h=1}^{H} -2\log Pr(\mathcal{G}|\mathbf{\Pi}, \mathbf{Q}, \mathbf{Z})$ e $\hat{\zeta^2} = \frac{1}{H}(\sum_{h=1}^{H} -2\log Pr(\mathcal{G}|\mathbf{\Pi}, \mathbf{Q}, \mathbf{Z}) - \hat{\mu}^2)$

O procedimento consiste em assumir um valor K para o número de subpopulações, aplicar o algoritmo *Gibbs Sampling* ao conjunto de dados obtendo um valor para $Pr(\mathcal{G}|K)$ definido em (9). O valor de K que maximiza (9) é o número de subpopulações mais provável. Evano, Regnaut e Goudet (2005) propuseram inferir K a partir de uma estatística ad hoc ΔK.

3 MATERIAL E MÉTODOS

3.1 Material

3.1.1 Painel de Genótipos

O estudo de estrutura de população foi feito utilizando um painel constituído por 142 acessos de cana-de-açúcar, selecionados a partir da sua importância para o germoplasma brasileiro e denominado painel brasileiro de variedades de cana-de-açúcar. Os critérios utilizados para elaboração do painel foram: cultivares mais plantados em lavouras comerciais brasileiras, como RB867515, SP81-3250 e RB855453; ancestrais importantes na história do melhoramento da cana-de-açúcar no Brasil tais como Co331, IAC48-65 e NA56-79; principais genitores utilizados em cruzamentos; genitores utilizados em programa de mapeamento como SP80-180 e SP80-4966; cultivares lançados recentemente e cultivares promissores dos programas de melhoramento brasileiro. No entanto, na estação experimental, dentre os 142 acessos disponíveis não foi possível a identificação de 13 indivíduos, portanto tem-se 129 indivíduos identificados corretamente e 13 acessos não identificados, conforme apresenta a Tabela 1.

3.1.2 Genotipagem dos indivíduos

Um total de 1033 marcadores do tipo SNP foram utilizados para genotipar 142 acessos de cana-de-açúcar. Dentre esses 1033 marcadores, 2 são monomórficos e foram desconsiderados das análises, resultando em um total de 1031 marcadores. A extração do DNA genômico foi feita a partir do meristema da cana-de-açúcar seguindo o protocolo descrito por Al-Janabi et al. (1999). Para quantificar o DNA genômico, usou-se o corante flourescente PicoGreen dsDNA quantitation kit (Invitrogen) e para leitura da flourescência utilizou-se o leitor de placas MWGT Sirius HT-TRF microplate reader (MWG). A identificação dos SNPs foi feita através de uma busca *in silico* em *clusters* selecionados do banco de dados SUCEST. Foram considerados como SNPs, a ocorrência de mais de um nucleotídeo em uma única posição dos *clusters* cujas as bases tiveram uma qualidade maior ou igual 20 (MARCONI, 2011). A genotipagem dos SNPs foi feita usando-se o sistema Sequenom MassARRAY ®. Foi realizada uma reação de PCR utilizando um par de *primers* específico para a região de interesse contendo o SNP, um tratamento com enzimas SAP para inativar os dNTPs não incorporados durante a PCR e uma outra reação de PCR denominada reação iPLEX na qual utilizou-se *primers* específicos e adjacentes ao SNPs com presença da enzima iPLEX, que amplificam nucleotídeos terminadores de massa modificadas.

Tabela 1 – Origem e nome dos 129 acessos de cana-de-açúcar que foram identificados e que são pertecentes ao painel brasileiro de variedades de cana-de-açúcar

Origem	Genótipos			
Ancestrais	Badila*	Ganda Cheni**	Maneria***	Chunee**
Java,	EK28	PO86-62		
Campos, Brasil	CB36-24	CB40-13	CB41-76	CB45-155
	CB46-47	CB47-355	CB49-260	CB53-98
Coimbatore, Índia	Co290	Co331	Co419	Co449
	Co740	Co997	NCo310	
Índia	IN84-58****			
Taiwan	F31-962	F36-819		
Queensland, Austrália	Q165			
Canal Point, EUA	CP70-1547	CP52-68	CP51-22	
Norte da Argentina	NA56-79			
Lousiana, EUA	L60-14			
Tucuman, Argentina	TUC71-7			
Hawaii	H59-1966	H53-3989		
Reunion	R570			
Campinas, Brasil	IAC48-65	IAC49-131	IAC50-134	IAC51-205
	IAC52-150	IAC64-257	IAC68-12	IAC82-2045
	IAC82-3092	IAC83-4157	IAC86-2210	IAC87-3396
	IAC91-1099	IAC93-2060	IAC93-3046	IAC93-3046
	IAC95-3018	IAC95-5000	IAC98-3022	
República do Brasil	RB721012	RB72199	RB72454	RB725053
	RB725828	RB732577	RB735220	RB735275
	RB739359	RB739735	RB75126	RB765418
	RB785148	RB815690	RB825317	RB825336
	RB83102	RB835019	RB835054	RB835089
	RB835205	RB835486	RB845197	RB845210
	RB845257	RB855002	RB855035	RB855036
	RB855077	RB855113	RB855156	RB855536
	RB855206	RB855350	RB855453	RB855463
	RB855563	RB855595	RB855511	RB867515
	RB92579	RB925211	RB925268	RB925268
	RB925345	RB935744	RB965902	RB965917
	RB966928			
São Paulo, Brasil	SP70-1005	SP70-1078	SP70-1143	SP70-1284
	SP70-1423	SP70-3370	SP71-799	SP71-1406
	SP71-6163	SP71-6949	SP72-4928	SP77-5181
	SP79-1011	SP79-2233	SP79-2312	SP79-2313
	SP79-6134	SP79-6192	SP80-1520	SP80-1816
	SP80-1836	SP80-1842	SP80-3280	SP81-3250
	SP83-2847	SP83-5073	SP89-1115	SP91-1049
	CTC2	CTC9	CTC15	

* *Saccharum officinarum*; ** *Saccharum barberi*; *** *Saccharum sinense*; **** *Saccharum spontaneum*

Os produtos dessas reações foram analisados pelo espectrômetro de massa MALDI-TOF, onde determinou-se quais são as bases presentes na posição do SNP comparando a massa do primer não extendido com as massas dos primers extendidos com cada um dos nucleotídeos de massa modificadas. Os resultados dessas analises feita no MALDI-TOF são picos de intensidade do espectro obtidos para cada uma desssas bases. Esses picos de intensidade representam alelos específicos de diferentes massas dependendo da sequência analisada. Ou seja, para cada SNPs tem-se um conjunto de dados formado por 2 picos de intensidade (maior e menor massa) para cada um dos indivíduos que compõem o painel.

3.2 Métodos

3.2.1 Classificação dos dados de SNPs

A inferência do genótipo dos indivíduos para cada loco foi feita usando o software SuperMASSA (SERANG; MOLLINARI; GARCIA, 2012). O modelo gráfico Bayesiano ajustado para cada um dos SNPs foi o chamado modelo Hardy-Weinberg. O espaço de busca do parâmetro P variou de 6 à 20 e do parâmetro σ variou de 0,01 à 1 com incremento de 0,05. Após o ajuste do modelo, obteve-se para cada SNP a probabilidade conjunta *a posteriori* da configuração genotípica ótima e nível de ploidia P, condicionada aos dados observados.

A estimativa do cálculo dessa probabilidade foi usada como uma medida de qualidade do ajuste do modelo, ou seja, quanto maior essa probabilidade maior é qualidade desse ajuste. Assim, utilizando um critério *ad hoc* foram selecionados apenas os SNPs cuja máxima probabilidade fosse maior ou igual a 0,7.

Estatísticas de alguns parâmetros genéticos foram calculadas, tais como estimativa da frequência alélica e variância da frequência alélica, cujo o estimador é dado por:

$$Var(\hat{F}_{obs}) = \frac{\hat{F}_{obs}(1-\hat{F}_{obs})}{NP} = \frac{\bar{m}_j(1-\bar{m}_j)}{NP} \tag{10}$$

em que $\hat{F}_{obs} = \frac{\sum_{j=1}^{N} m_{ij}}{NP}$; N é o número de acessos; P é o nível de ploidia em um determinado SNP; m_{ij} é a dosagem alélica do i-ésimo acesso no j-ésimo loco; $\bar{m}_j$ é a média da dosagem alélica para cada SNP.

3.2.2 Imputação de dados

Um aspecto importante em uma análise de componentes principais é a questão de dados perdidos. É conhecido que para-se aplicar a decomposição de valores singulares é necessário

que não haja a presença de dados perdidos na matriz de dados (PATTERSON; PRICE; REICH, 2006). Portanto, para solucionar esse problema, neste trabalho foram aplicados e comparados dois métodos de imputação, cujos detalhes serão apresentados a seguir

Inicialmente, foi feita uma análise descritiva dos dados para verificar quantos SNPs possuem pelo menos um dado perdido, bem como a porcentagem de dados perdidos por SNPs. Ainda como estatística descritiva dos dados, foram estimados os desvios padrão da dosagem alélica de cada SNP.

A imputação de dados foi feita através do método proposto por Stekhoven e Buhlmann (2012) baseado no algoritmo de *floresta aleatória* e que está implementado na biblioteca *MissForest* do software estatístico R (R DEVELOPMENT CORE TEAM, 2012). Utilizou-se 100 árvores e o número de variáveis selecionadas aleatoriamente foi igual à raiz quadrada do número de marcadores selecionados após o SNP *calling*. O algoritmo de floresta aleatória é definida por um conjunto de árvores de classificação e/ou de regressão. Essas árvores são modelos baseados em partições recursivas do espaço formado pelas cováriáveis e pela variável resposta, de tal forma que os subespaços gerados sejam de variância mínima (BREIMAN et al., 1984). Assim, baseando-se no método de floresta aleatória, Stekhoven e Buhlmann (2012) propuseram um algoritmo não paramétrico para imputação de dados. Esse algoritmo consiste em, para cada SNP, ajustar um modelo de floresta aleatória para os dados que são observados. Dessa forma, à partir desse modelo são preditos os dados perdidos. O algoritmo repete esses dois passos até atingir a convergência sob algum critério de parada.

Um estudo de validação cruzada foi realizado com a finalidade de comparar a acurácia do método de Stekhoven e Buhlmann (2012) com o método de imputação de dados baseado na média da frequência alélica de cada SNP. Esta segunda alternativa também foi empregada porque é a utlizada nos softwares TASSEL (BRADBURY et al., 2007) e *Eigensoft* (PATTERSON; PRICE; REICH, 2006).

Considerando apenas os marcadores que não apresentavam dados perdidos, esse estudo consistiu em dividir aleatoriamente o conjunto de dados em $T = 1000$ subconjuntos mutuamente exclusivos sendo que para cada iteração um único subconjunto foi utilizado para se fazer a validação de cada um dos métodos de imputação e os outros $T - 1$ subconjuntos remanescentes foram usados como entrada de dados para a implementação dos métodos de imputação (dados de treinamento). Assim, repetiu-se esse processo T vezes.

A estatística utilizada para mensurar a acurácia dos métodos de imputação foi a raiz quadrada do erro quadrático médio normalizado (*"normalized root mean squared error"* NRMSE), proposto por Stekhoven e Buhlmann (2012) e dada por

$$NRMSE = \sqrt{\frac{\frac{1}{NL}\sum\sum\left(m_{ij} - \hat{m}_{ij}\right)^2}{\frac{1}{(NL-1)}\sum\sum\left(m_{ij} - \bar{m}_{ij}\right)^2}}$$

em que $\hat{m}_{ij}$ é o dado imputado por cada um dos métodos e $\bar{m}_{ij}$ a média geral dos dados. Quando o método de imputação apresenta uma boa performance a raiz quadrada do erro quadrático médio normalizado tende ao valor 1; em caso de má performance, essa estatística tende ao valor 0.

Foram obtidas T estimativas do NRMSE referentes ao método de Stekhoven e Buhlmann (2012), e T estimativas referentes ao método de imputação pela média. Para verificar qual dentre dois métodos de imputação é o mais acurado, foi feito um teste de comparação de médias *bootstrap* composto pelas seguintes etapas:

- A partir da amostra original produziu-se $B = 1000$ estatísticas raiz quadrada do erro quadrático médio normalizado NRMSE *bootstrap* para ambos métodos de imputação.

- Para cada par de elementos das duas amostras *bootstrap* foi calculada a estatística

$$\tau^{boot} = \frac{\sqrt{n_{boot}}\phi(nmrse_1 - nmrse_2)}{\sigma_{boot}(nmrse_1 - nmrse_2)}$$

  em que $\phi(nmrse_1 - nmrse_2), \sigma_{boot}(nmrse_1 - nmrse_2)$ são a média e o desvio padrão *bootstrap* da diferença dos NRMSE obtidos a partir dos dois métodos de imputação.

3.2.3 Estrutura populacional

3.2.3.1 Análise de componentes principais e coordenadas principais

A análise de componentes principais foi feita a partir de uma matriz de dados M, em que as linhas representam os acessos de cana-de-açúcar e as colunas representam os marcadores SNPs. Para realizar a análise, foi utilizada a função *prcomp* do software estatístico R (R CORE TEAM 2012).

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.

Adicionalmente, foi feito um estudo para determinar qual é o número mínimo de marcadores necessários para inferir a presença de estrutura de populações via análise de componentes

principais. Esse estudo consistiu em, inicialmente, retirar-se $B_l = 100, 200, \ldots, L$ amostras *bootstrap* a partir do matriz original de dados M sendo que o sorteio foi feito apenas nas colunas. Para cada amostra *bootstrap*, estimou-se W componentes principais sendo que o número W de componentes foi determinado previamente na ACP dos dados originais.

Em seguida, para todo $w \in W$, determina-se a estimativa do quadrado da correlação de Pearson entre w-ésimo componente principal *bootstrap* e w-ésimo componente principal estimado pela ACP feita na matriz M original, de tal forma que para cada B_l tem-se a distribuição empírica *bootstrap* do quadrado da correlação de Pearson, entre os componentes principais observados e aqueles calculados nas estimativas *bootstrap*.

A ideia desse método é que pressupondo um número de marcadores genotipados suficientemente grande, conforme aumenta o número de marcadores utilizados, a estimativa *bootstrap* da média do quadrado da correlação de Pearson tende a 1 e a estimativa *bootstrap* da variância tende a zero, portanto, o objetivo é analisar a partir de quantos marcadores esse fato ocorre. Em outras palavras, isto permite avaliar se o número de SNPs utilizados é uma amostra suficiente do genoma.

Para comparar os resultados da estrutura de população com as informações a respeito da genealogia dos cultivares, foi feita uma análise de coordenadas principais com a matriz de distância cujo os elementos foram definidos como $1 - coeficientedeparentesco$. O coeficiente de parentesco θ_{AB} 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 (LYNCH; WALSH, 1998; HEDRICK, 2010).

Neste trabalho, a matriz de parentesco foi estimada através da função *kinship* disponível na biblioteca *kinship2* e a análise de coordenadas principais foi feita usando a função *cmdscale* ambos implementados no *software* estatístico R (R Development Core Team, 2012).

3.2.3.2 Análise de agrupamento

A Matriz de distâncias genéticas usando dados de marcadores moleculares SNPs foi obtida a partir das estimativas de distância de alelos compartilhados, calculadas 2 a 2 para os 143 acessos. A distância de alelos compartilhados é definida como um menos a proporção de alelos compartilhados por dois indivíduos quaisquer considerando-se L locos (BOWCOCK et al.,

1994).

A expressão do estimador de distância de alelos compartilhados é dada por

$$\Theta^{i,i^{'}}_{\text{alelos compartilhados}} = 1 - Fr^{i,i^{'}}_{\text{alelos compartilhados}} = 1 - \sum_{l=1}^{L}\sum_{j=1}^{2} \min\left(\breve{\pi}^{i}_{lj} - \breve{\pi}^{i^{'}}_{lj}\right) = \frac{\sum_{l=1}^{L} |\breve{\pi}^{i}_{lj} - \breve{\pi}^{i^{'}}_{lj}|}{L}$$

em que $\Theta^{i,i^{'}}_{\text{alelos compartilhados}}$ é o estimador da distância de alelos compartilhados entre os indivíduos i e $i^{'}$, $Fr^{i,i^{'}}_{\text{alelos compartilhados}}$ é a proporção de alelos compartilhados entre os indivíduos i e $i^{'}$, $\breve{\pi}^{i}_{lj}$ é a frequência alélica dentro do indivíduo i, no loco l do alelo j e L é o número de locos.

Foi verificado se o número de marcadores SNPs foi suficiente para obter uma estimativa precisa das distâncias genéticas, por meio do o método de Tivang, Nienhuis, Smith (1994). Em linhas gerais, esse método consiste em para cada número de marcadores $100, 200, \ldots, 805$ obter a distribuição empírica *bootstrap* do coeficiente de variação das distâncias genéticas. Considera-se um número de marcadores suficientes aqueles cuja a mediana da distribuição *bootstrap* do coeficiente de variação seja menor ou igual a 10 %.

As análises de agrupamento foram feitas utilizando-se o algoritmo *unweighted pair-group method using arithmetic average* (UPGMA) implemmentado na função *hclust* do *software* estatístico R (R Development Core Team, 2012).

Para avaliar a qualidade do agrupamento, foi estimada a correlação cofenética, definida como a correlação entre os valores estimados da matriz de distância genética de alelos compartilhados ou de parentesco e os valores da matriz de distâncias que originam o dendrograma. Por fim, foi estimada a correlação linear de Pearson entre o coeficiente de parentesco e a proporção de alelos compartilhados.

3.2.3.3 Software STRUCTURE

O software STRUCTURE versão 2.3.3 (PRITCHARD; STEPHENS; DONNELLY, 2000) foi utilizado para inferir a estrutura populacional. Uma vez que não é possível analisar dados que representam locos com níveis de ploidias diferentes, como ocorre em cana-de-açúcar, para cada nível ploidia retirou-se uma amostra de 50 SNPs distribuídos ao acaso no genoma, o que resultou em 8 conjuntos de dados. Assim, para cada um desses conjunto de dados, ajustou-se o modelo de mistura com frequências correlacionadas, pressupondo-se que para cada acesso pode-se ter uma fração do genoma oriunda de uma ou mais subpopulação e assumindo ainda que essas subpopulações podem estar correlacionadas. Utilizou-se um *burnin* de 100 000 iterações e 200 000 iterações após o *burnin*, sendo que cada modelo foi ajustado repetidamente 10 vezes. A

determinação do número ótimo de subpopulações foi feita por meio da estatśitica delta proposta por Evano, Regnaut e Goudet (2005). Dessa forma, para cada nível de ploidia, obteve-se uma matriz Q a qual fornece para cada indíviduo a fração do genoma que pode pertencer a cada uma das K subpopulações.

4 RESULTADOS

4.1 Classificação dos dados de SNPs

Os resultados do *software* SuperMASSA revelaram que, dentre 1031 marcadores SNPs analisados, 194 SNPs foram classficados como hexaploides, 94 como octaploides, 71 como decaploides, 89 como dodecaploides, 92 como tetradecaploides, 95 como hecadecaploides, 122 como octadecaploides e 274 como icosaploides. Um maior número de marcadores foram classificados nos níveis de ploidia 6 e 20 e um menor número de marcadores foram classficados nos níveis de ploidia 10 e 12. No entanto, a medida que aumenta o nível de ploidia de cada loco há uma maior ocorrência de máximas probabilidades *a posteriori* menores que 0,7, diminuindo a confiabilidade da classificação para esses marcadores (Figura 2 (a)).

Analisando apenas os marcadores cuja a máxima probabilidade *a posteriori* foi maior que 0,7 nota-se uma redução do número de marcadores classificados com ploidias 6 e 20, além de ter-se marcadores em que a confiabilidade da inferência do genótipos condicional aos dados observados seja maior (figura 2). Assim, utilizando o critério *ad hoc* de considerar apenas os marcadores SNPs com probabilidade *a posteriori* maior que 0,7, tem-se 138 marcadores SNPs com nível de ploidia estimada igual a 6, 85 igual a 8, 59 igual 10, 72 igual a 12, 64 igual a 14, 65 igual a 16, 81 igual a 18 e 241 com nível de ploidia igual a 20. Conforme esperado, nota-se que o genoma da espécie é de fato complexo e composto de uma mistura de níveis de ploidia.

A Figura 3 apresenta a distribuição das frequências estimadas de um alelo de referência dos SNPs utilizados no painel. Considerando todos os marcadores tem-se um maior número de SNPs com estimativa de frequência alelica de 0,01 e 0,96 o que indica uma grande presença de alelos raros na amostra. Por outro lado, para os SNPs com probabilidade *a posteriori* maior que 0,7, observa-se uma diminuição dos alelos raros tornando a distribuição da frequências do alelo de referência mais uniforme. Isto é interessante, uma vez que normalmente alelos com alta (ou baixa) frequência são eliminados em análises de mapeamento associativo.

Na figura 4 são apresentadas as estimativas da variância da frequência alélica observada para cada SNP. Nota-se que apenas três deles possuem uma variância da frequência alélica superior a 0,0025. Agora vale enfatizar que a maioria dos SNPs possuem frequências alélicas entre 0,01 e 0,31 ou entre 0,71 e 0,96. Analisando a expressão do estimador (equação 10) vê-se que a mesma é minimizada para valores de frequência alélica tendendo a 1 ou a 0, portanto o resultado obtido está de acordo com o esperado teoricamente.

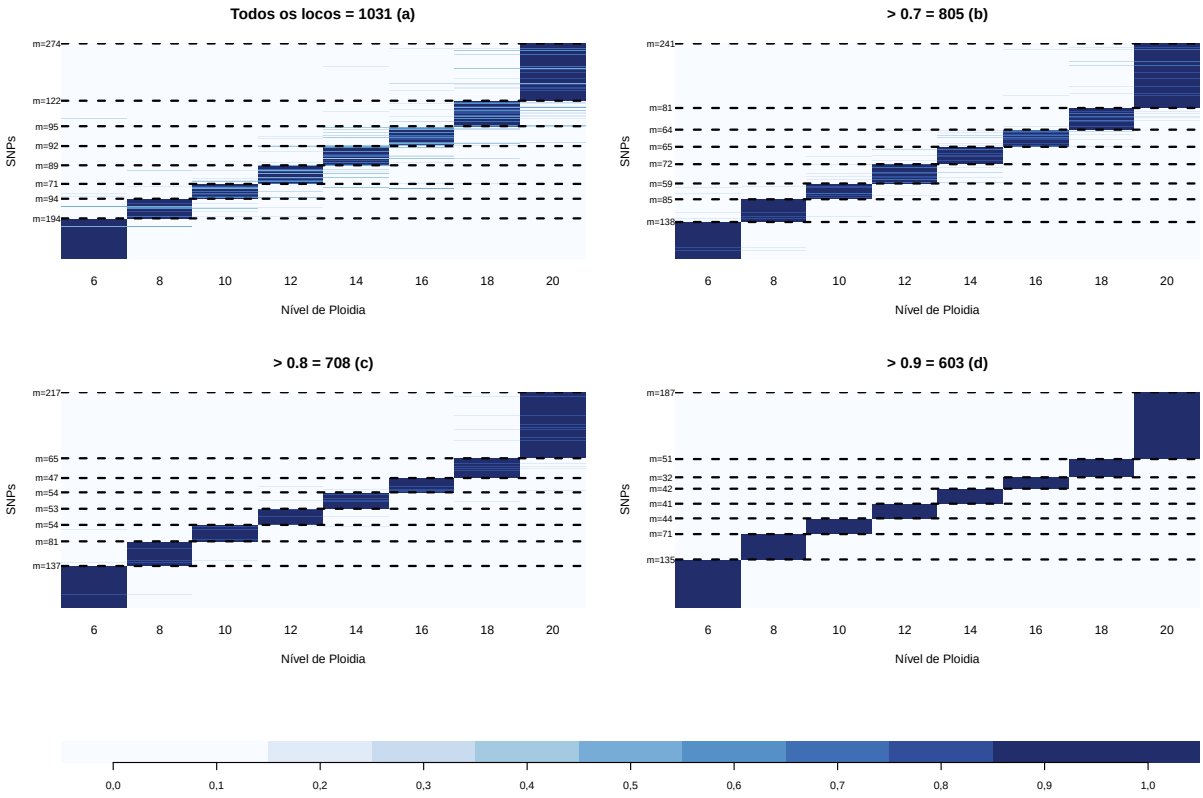


Figura 2 – Classificação dos dados de SNPs para o painel brasileiro de variedades de cana-de-açúcar. (a) Nesta classficação cada loco foi atribuído a uma determinada classe genotípica baseando-se na máxima probabilidade conjunta *a posteriori* dos genótipos e nível de ploidia condicional aos dados observados. Considerando apenas os marcadores cuja a probabilidade conjunta *a posteriori* dos genótipos e nível de ploidia condicional aos dados observados forem maior que 0,7 (b); probabilidade *a posteriori* $>$ 0,8 (c); probabilidade *a posteriori* $>$ 0,9 (d). Estimativa da probabilidade *a posteriori* estão representadas pela escala de valores em tom de azul variando de 0 à 1

4.2 Imputação de dados

O número de marcadores que apresentam pelo menos algum dado perdido foi 410, o que equivale a 51,93 % dos marcadores com probabilidade *a posteriori* maior que 0,7. Entretanto, apenas 22 SNPs apresentam uma porcentagem de dados perdidos superior a 25 % (Figura 5). Além disso, percebe-se que apenas 3 SNPs possuem variância da dosagem alélica maior ou igual 3, o que significa uma boa qualidade da genotipagem (Figura 6).

A Figura 7 apresenta as estimativas *bootstrap* da média da estatística raiz quadrada do erro quadrático médio normalizado obtidas pelos dos dois métodos de imputação e a distribuição empírica da estatística t do teste de hipótese *bootstrap*. O valor P do teste foi igual a 0.49, indicando que não existem evidências estatísticas para se rejeitar a hipótese de que a raiz quadrada do erro quadrático médio normalizado obtida por ambos os métodos sejam iguais. Ou seja, o método de imputação de dados usando o método de Stekhoven e Buhlmann (2012) não se mos-

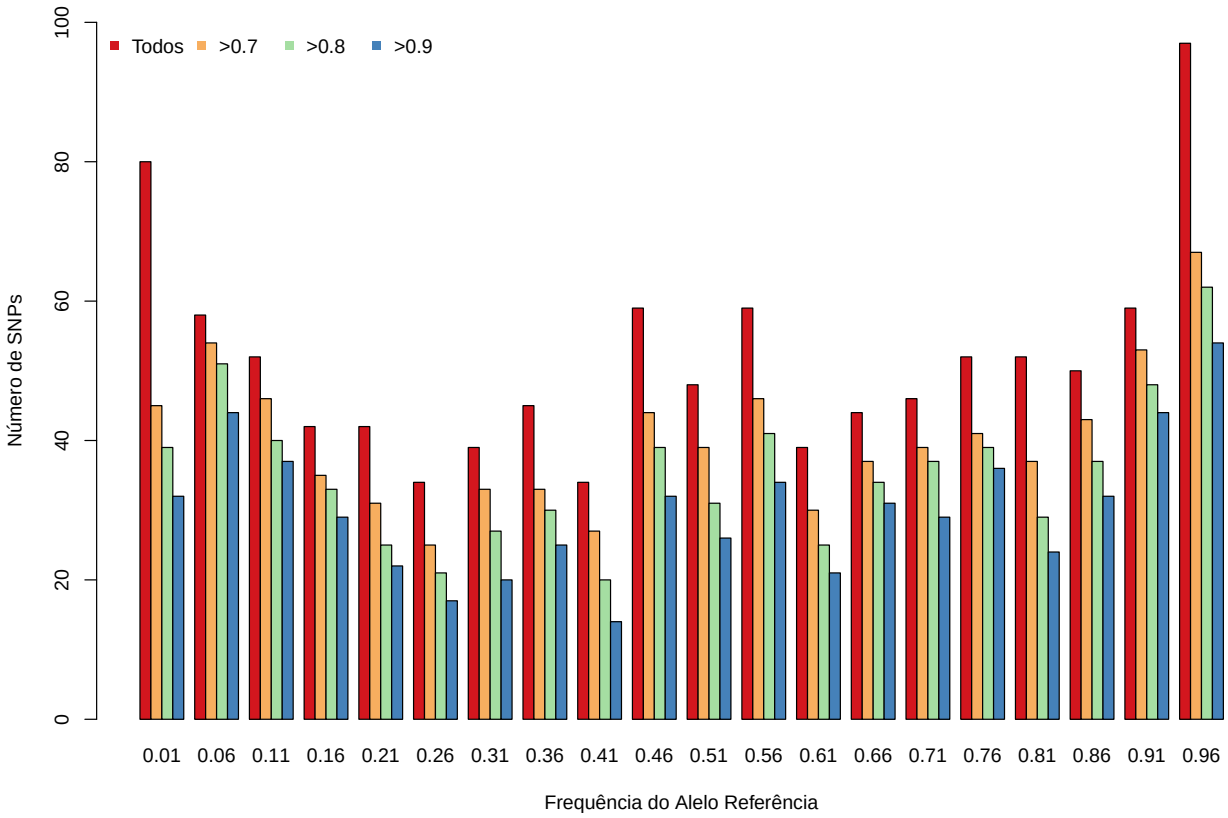


Figura 3 – Distribuição da frequência de um alelo de referência dos SNPs usados nos dados do painel brasileiro de variedades de cana-de-açúcar. Diversos cenários foram considerados: todos os locos (vermelho), apenas os marcadores com probabilidade conjunta *a posteriori* maior que 0,7 (laranja); marcadores com probabilidade conjunta *a posteriori* maior que 0,8 (verde); marcadores com probabilidade conjunta *a posteriori* maior que 0,9 (azul)

trou mais acurado do que o método de imputar os dados usando a média das frequências alélicas de cada SNP. Isto indica que o método mais fácil de implementar pode ser usado em situações práticas.

4.3 Estrutura populacional

4.3.1 Análise de componentes principais e coordenadas principais

A análise de componente principais revelou presença de estrutura nos dados do painel brasileiro de variedades de cana-de-açúcar. O primeiro componente (PC1) explicou 7,63 % da variabilidade dos dados. Nota-se que o acesso IN84-58 está localizado entre os valores de -100 e -50 do primeiro componente, os acessos Chunnee e Guanda Cheni estão entre -50 e 0, os cultivares modernos estão entre -25 e 25 e os acessos Badila e EK28 estão localizados entre 25 e

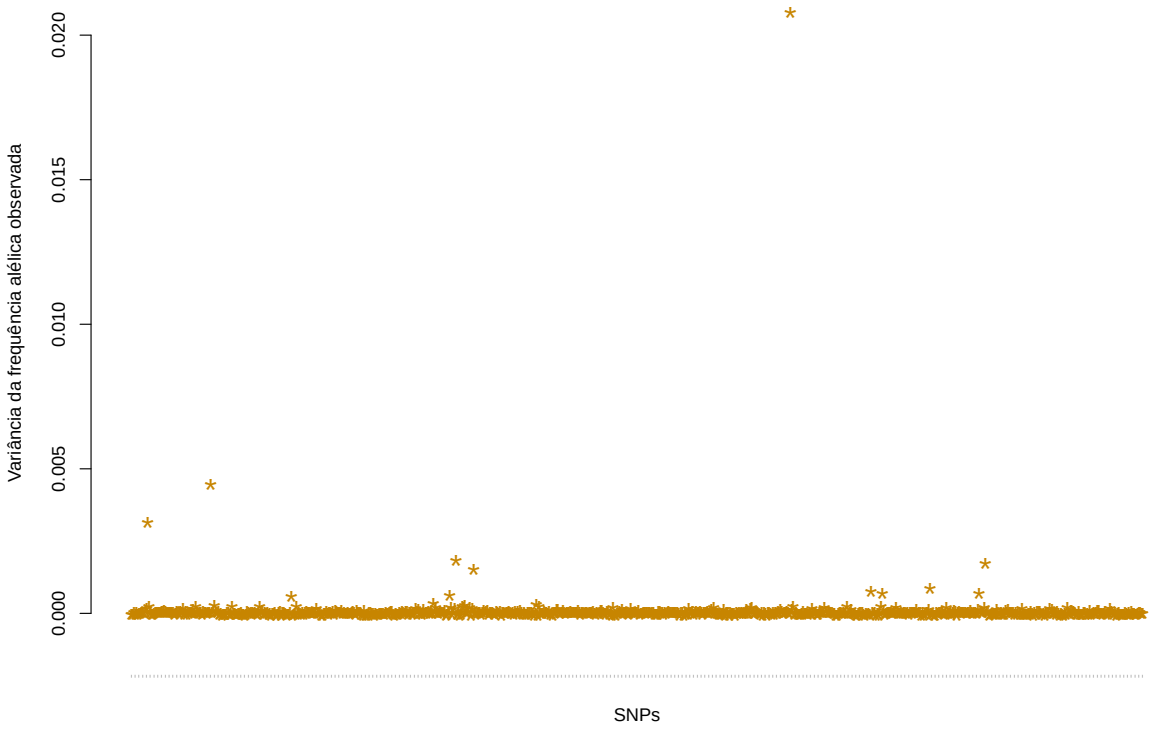


Figura 4 – Estimativas da variância das frequência alélica observada para cada SNP referente ao painel brasileiro de variedades de cana-de-açúcar

50, sugerindo que o primeiro componente separa o acesso IN84-58 (*S. spontaneum*) dos acessos restantes do painel que por sua vez estão separados em 3 subpopulações. A primeira, formada pelos acessos Chunnee, Gandacheni (*S. Sinense*); a segunda formada pelos cultivares modernos de cana-de-açúcar; a terceira formada por acessos que são espécies não melhoradas de (*S. officinarum*). O segundo componente explica 4,32 % da variabilidade dos dados e separa as espécies (*S. spontaneum*, *S. Sinense* e *S. officinarum*) dos cultivares modernos (Figura 8).

Por sua vez, a análise de coordenadas principais aplicada na matriz definida por $1-coeficiente\ de\ parentesco$ mostra resultados contrastantes. O primeiro eixo de coordenadas principais explica 4,49 % da variabilidade dos dados e separa a progênie dos cruzamentos RB72454 e SP70-1143 e RB72454 e TUC71-7 dos genitores e progênie do acesso NA56-79. Os acessos RB835089 e RB835019 estão destacados porque são progênies do cruzamento RB72454 e NA56-79 sugerindo uma variação contínua entre esses dois grupos. O segundo eixo de coordenadas principais explica 3,92 % da variabilidade dos dados e separa esses dois grupos anteriormente citados dos demais cultivares (Figura 9).

A Figura 10 mostra a distribuição empírica *bootstrap* do quadrado da correlação de Pearson

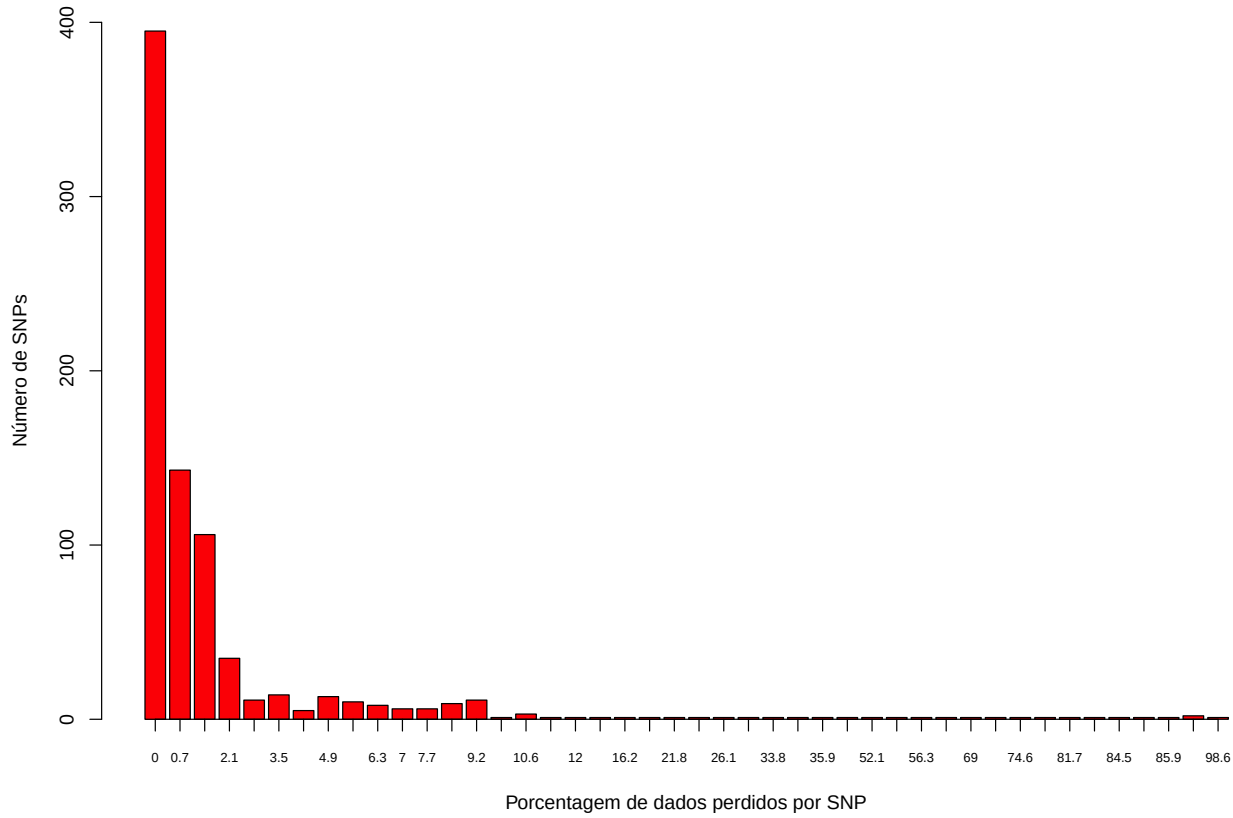


Figura 5 – Número de marcadores do tipo SNP e porcentagem de dados perdidos por SNP do painel brasileiro de variedades de cana-de-açúcar

entre o primeiro componente principal estimado a partir dos dados originais e o primeiro componente estimado a partir de amostras *bootstrap*, assim como do segundo componente principal original e o estimado a partir das amostras *bootstrap*. Observa-se que para o primeiro componente, a partir de 500 marcadores a variância da distribuição empírica estabiliza-se e a mediana da distribuição do quadrado da correlação de Pearson é aproximadamente igual a 0,95, indicando possivelmente que esse seja o número de SNPs necessário para detectar a estrutura de população revelada pelo primeiro componente. Por sua vez, para o segundo componente observa-se que a variância da distribuição empírica estabiliza-se somente após 700 marcadores SNPs utilizados e a mediana dessa distribuição é aproximadamente 0,9 para o número de marcadores SNPs maiores que 800.

4.3.2 Análise de agrupamento

A proporção de alelos compartilhados variou de 0,7750 (Badila e IN84-58) a 0,9633 (entre

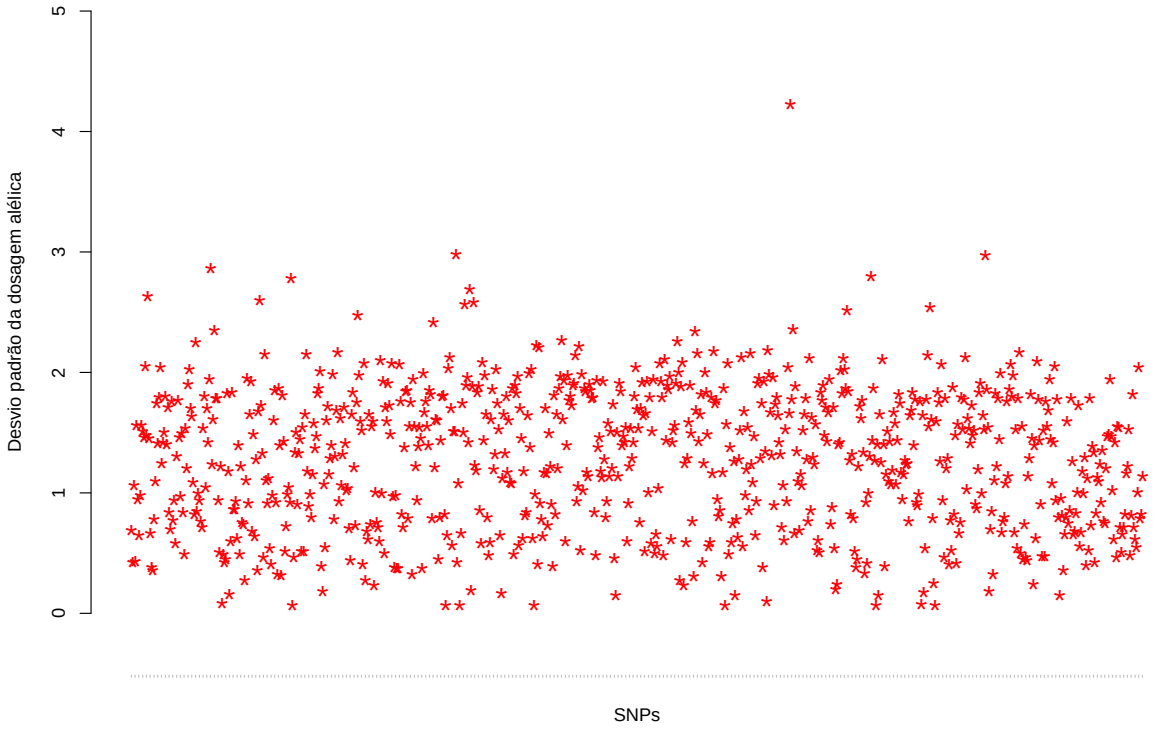


Figura 6 – Estimativas dos desvios padrão da dosagem alélica para cada SNP referente ao painel brasileiro de variedades de cana-de-açúcar

IAC83-4157 e RB73500) com média igual a 0,8973. O coeficiente de parentesco variou de 0 a 0,503 (entre NA56-79 e Co419) a 1 com média igual a 0,043. Gráficos *boxplot* da distribuição bootstrap do coeficiente de variação das distâncias de alelos comparitlhados obtida para um determinado número de marcador mostra que a partir de 200 marcadores a mediana do coeficiente de variação dessa distribuição é menor que 10%. Ou seja, a partir dessa quantidade de marcadores pode-se obter uma estimativa precisa da matriz de distância de alelos compartilhados (Figura 11).

A análise de agrupamento utilizando o método UPGMA baseado nos dados dos marcadores SNPs mostrou que os cultivares modernos de cana-de-açúcar pertecem ao mesmo grupo, e que a maior distância observada no dendrograma é entre os acessos IN84-58 (*S. spontaneum*) e Badila (*S. officinarum*), confirmando os resultados da análise de componentes principais. A correlação cofenética foi igual a 0.911, indicando uma boa qualidade do agrupamento (Figura 12).

Por sua vez, o dendrograma resultante da aplicação do método UPGMA à matriz cujos os elementos são $1 - coeficiente\ de\ parentesco$ apresentou uma maior dissimilaridade entre os cultivares modernos. A maior dissimilaridade foi observada entre os acessos NA56-79 e RB855206

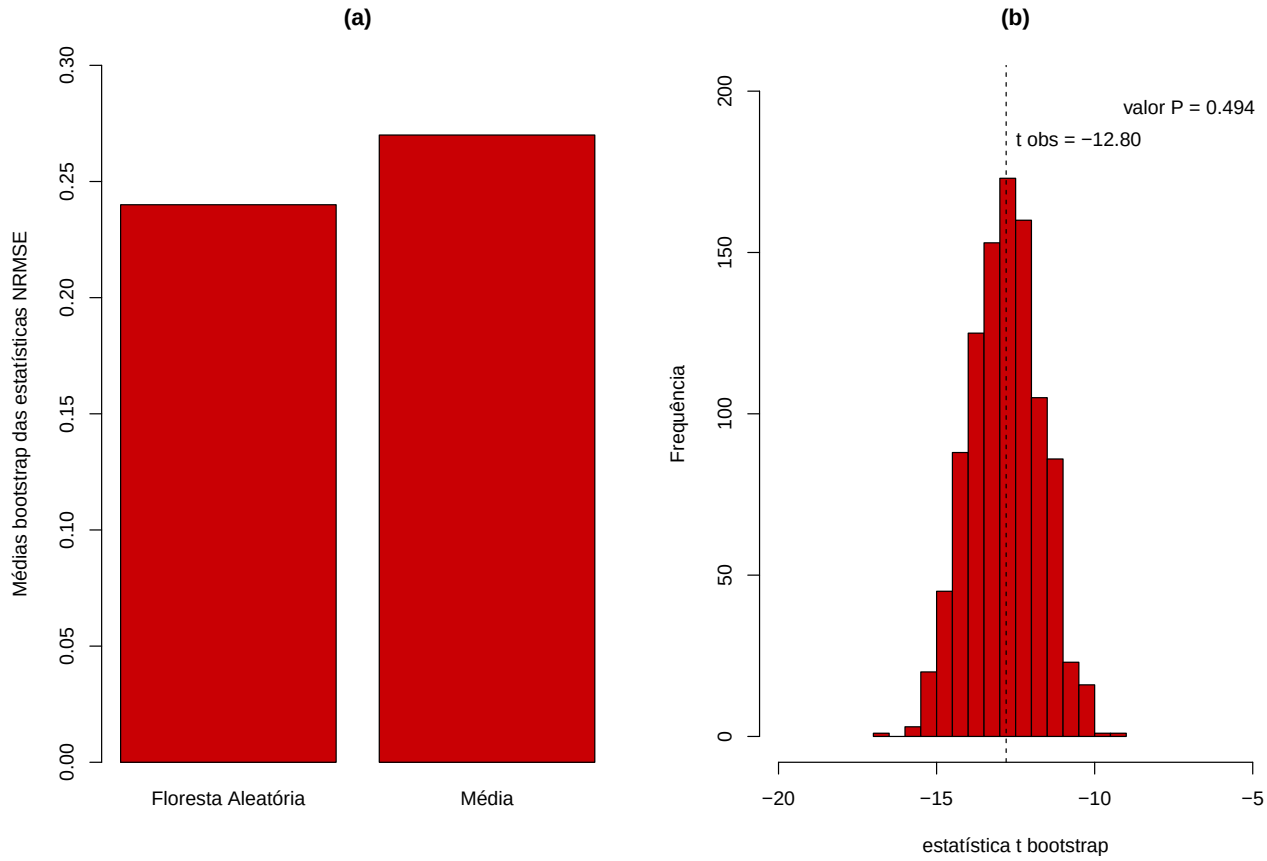


Figura 7 – Teste de hipótese *bootstrap* feito para comparar os métodos de imputação de dados baseados na média da frequência alelica ou em floresta aleatória. Estimativas *bootstrap* da média da estatística NRMSE obtida para os dois métodos de imputação (a). Distribuição empírica da estatística t *bootstrap* e valor p (b)

(progênie do cruzamento RB72454 e TUC71-7) confirmando os resultados da análise de coordenadas principais (Figura 13). O gráfico de dispersão entre a proporção de alelos compartilhados e a estimativa dos coeficientes parentesco mostra uma baixa correlação entre essas variáveis $(r = 0,30)$ (Figura 14).

4.3.3 Software STRUCTURE

Estudos de inferência de estrutura de populações usando o STRCTURE versão 2.3.3 (PRITCHARD; STEPHENS; DONNELLY, 2000) foram realizados para subconjuntos de dados de SNPs, separados de acordo com a estimativa do nível de ploidia. A Tabela 2 apresenta os resultados das estimativas das estatísticas propostas pelo método de Evano, Regnaut e Goudet (2005) para modelo de frequência correlacionada considerando locos para cada nível de ploidia. Resultados das estimativas da estatística delta mostra que o número ótimo de populações para o nível

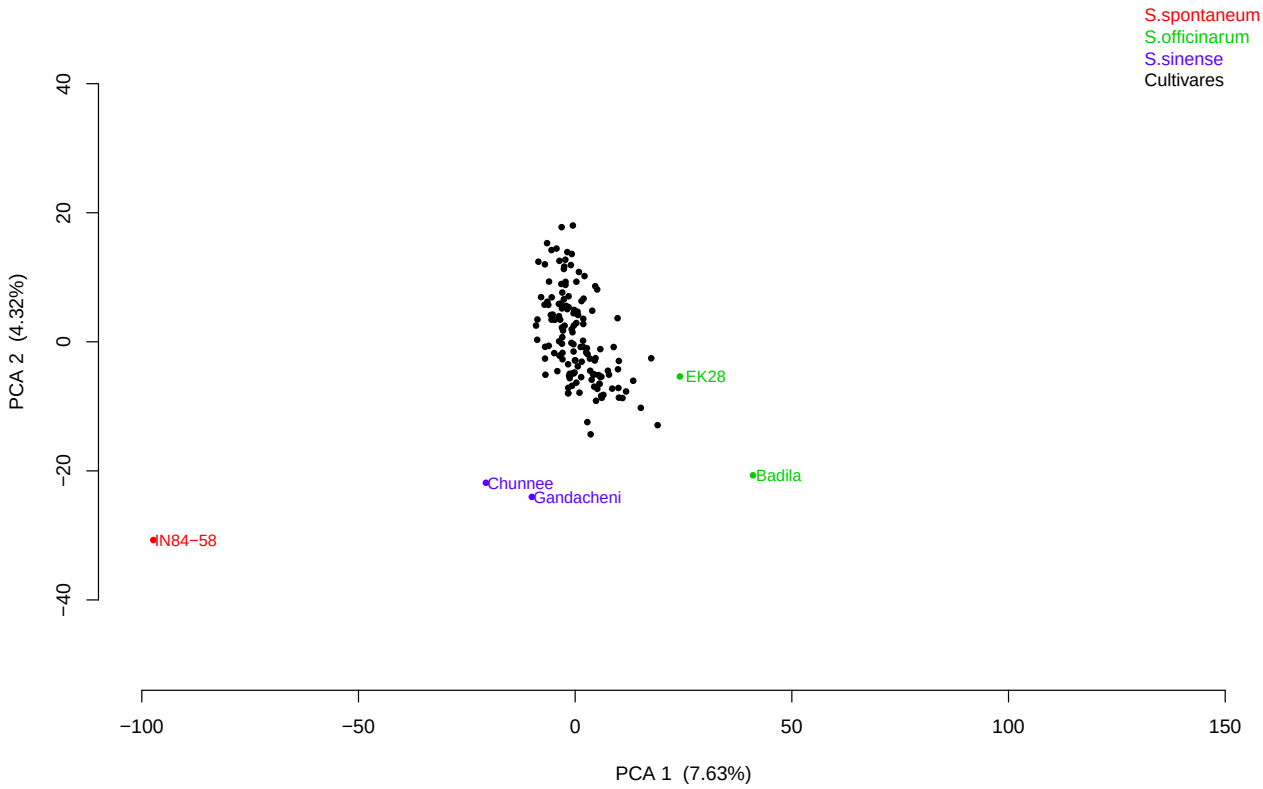


Figura 8 – Análise de componentes principais dos dados de marcadores SNPs que foram usados para genotipar o painel brasileiro de variedades de cana-de-açúcar. Componente principal 1 no eixo das abicssas e componente principal 2 no eixo das ordenadas

de ploidia 6 é igual a 6, para os níveis de ploidia 8 e 20 esse número ótimo de populações é igual a 4 e para os demais níveis de ploidia K ótimo é igual a 2. Ou seja, para os três conjuntos de dados de SNPs que possuem uma maior quantidade de dados classificados (níveis de plodia igual à 20, 6, e 8 respectivamente) observou-se $K > 2$.

No apêndice A são apresentadas as matrizes Q obtida para cada subconjunto de dados. Sendo que cada linha dessa matriz Q representa os acessos do painel e cada coluna representa a fração do genoma do indíviduo pertencente a umas das das K subpopulações. Analisando essas matrizes, de um modo geral, nota-se que para cada nível de ploidia houve um agrupamento dos acessos de uma forma diferente. Para o nível de ploidia 6, observou-se que a primeira subpopulação é composta principalmente por acessos originados do programa de melhoramento Campos Brasil (CBs) e Coimbatore Índia (Co); a segunda subpopulação composta por acessos proveniente do programa do instituto agronômico de Campinas (IACs) e por alguns cultivares RBs; a terceira e quarta compostas exclusivamente por cultivares RBs; a quinta composta por cultivares

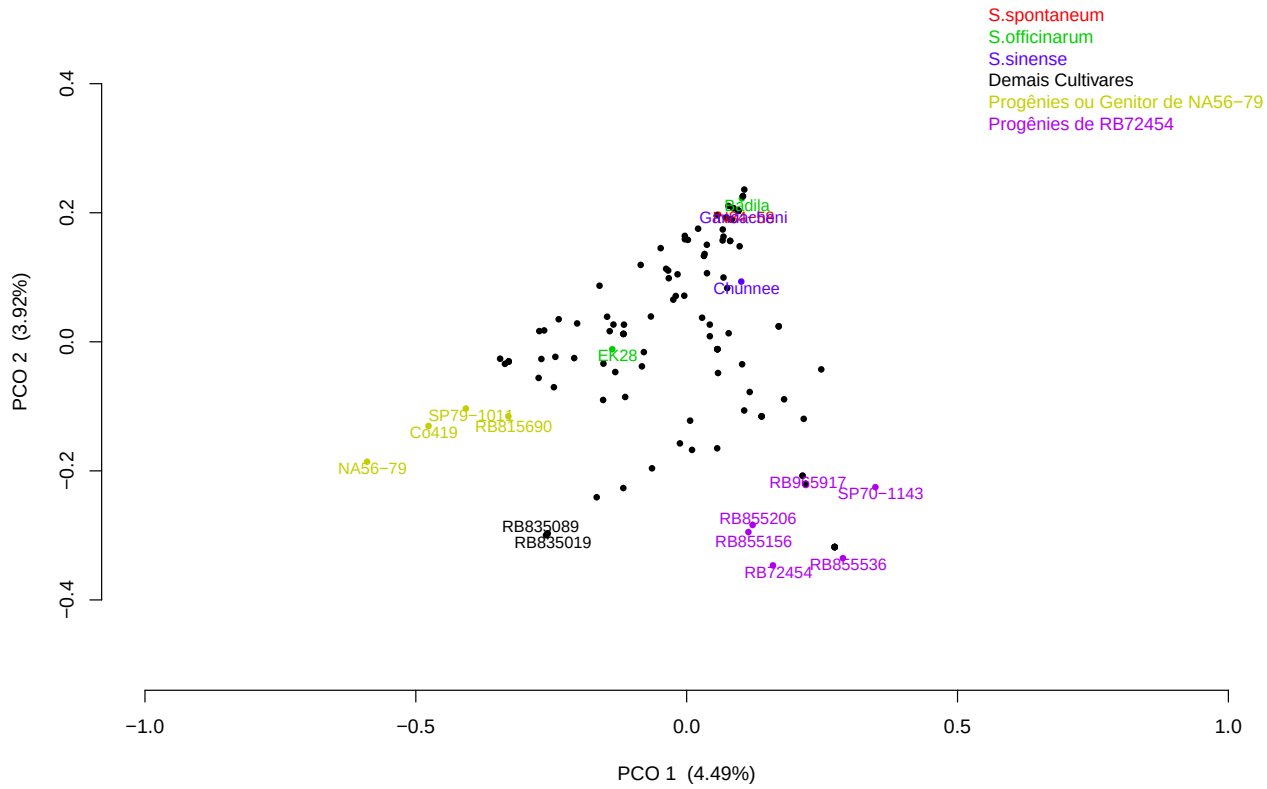


Figura 9 – Análise de coordenadas principais da matriz de distâncias cujo os elementos são definidos como $1 - coeficiente\ de\ parentesco$. Eixo de coordenada principal 1 no eixo das abicssas e eixo de coordenada principal 2 no eixo das ordenadas

RBs e SP70 e SP71 e a sexta subpopulação composta principalmente por cultivares SP77, SP79 e SP80. Para o nível de ploidia 8, a primeira subpopulação é composta principalmente por acessos oriundos do programa de Canal Point (CPs), Campos Brasil (CBs); a segunda subpopulação composta por acessos oriundos programa do instituto agronômico de Campinas (IACs) e de Coimbatore Índia (Co); a terceira composta por acessos RBs e quarta composta principalmente por acessos SPs e RBs. Por sua vez, para os níveis de ploidia 10 a 18, nota-se que uma subpopulação é composta por cultivares SPs e parte dos RBs e a segunda subpopulação composta pelos demais acessos. Para o nível de ploidia 20, embora o número ótimo de subpopulações seja igual a 4, vê-se que a maioria dos elementos da matriz Q é aproximadamente igual a 0,25, o que indica que o genoma desses indivíduos não é predominantemente originado de nenhuma das quatro populações.

De uma forma geral, esses resultados indicaram que a estratégia de subdividir o conjuntos de dados por nível de ploidia e ajustar um modelo de frequencias correlacionadas para cada um

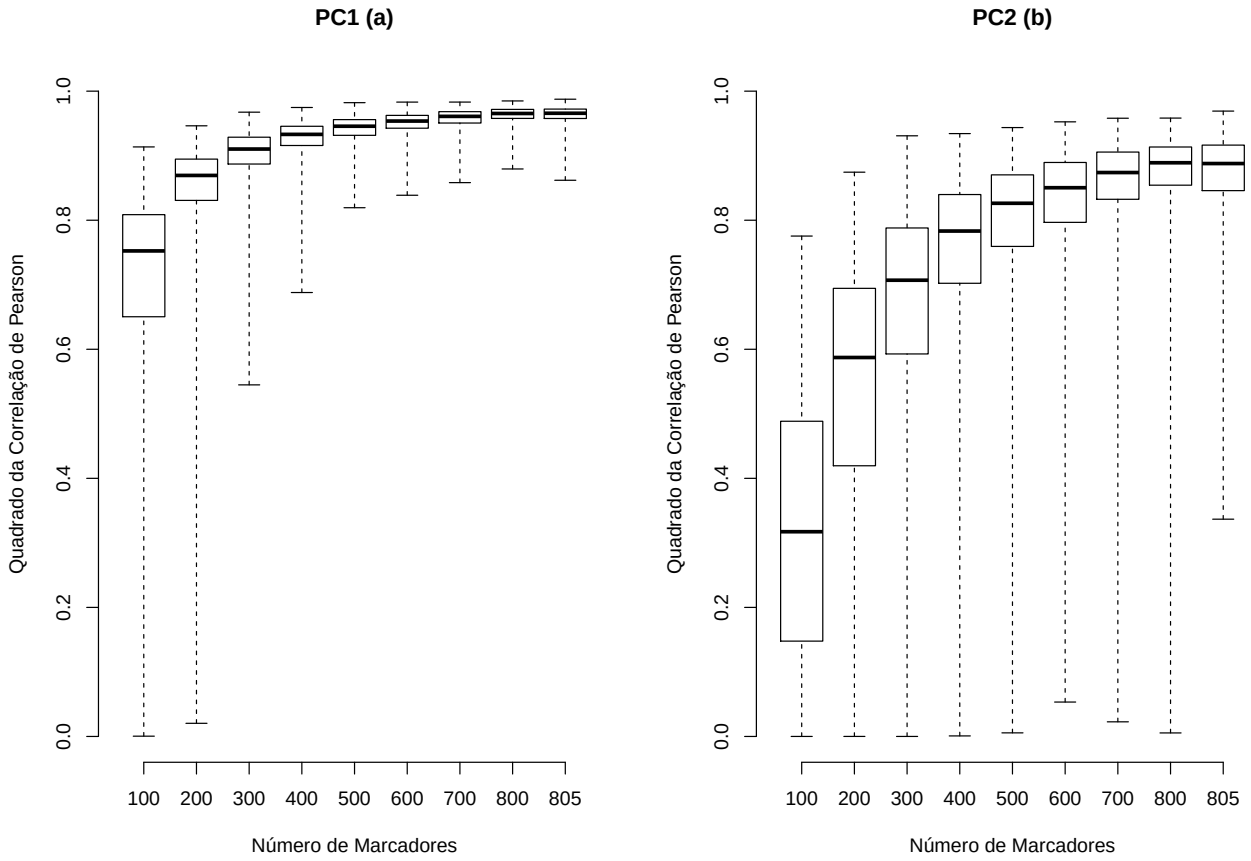


Figura 10 – Gráficos *boxplot* da distribuição do quadrado da correlação de Pearson entre PC1 e PC1 bootstrap (a) e entre PC2 e PC2 bootstrap versus o número de marcadores

deles, não possibilitou ter-se uma resposta única para o número mais provável de subpopulações. Ademais, a fração do genoma de cada indivíduo que é originada por cada uma das subpopulações variou de acordo com o nível de ploidia dos marcadores considerados. Claramente, isto revela a dificuldade em se realizar estudos para uma espécie com um genoma complexo como a cana-de-açúcar.

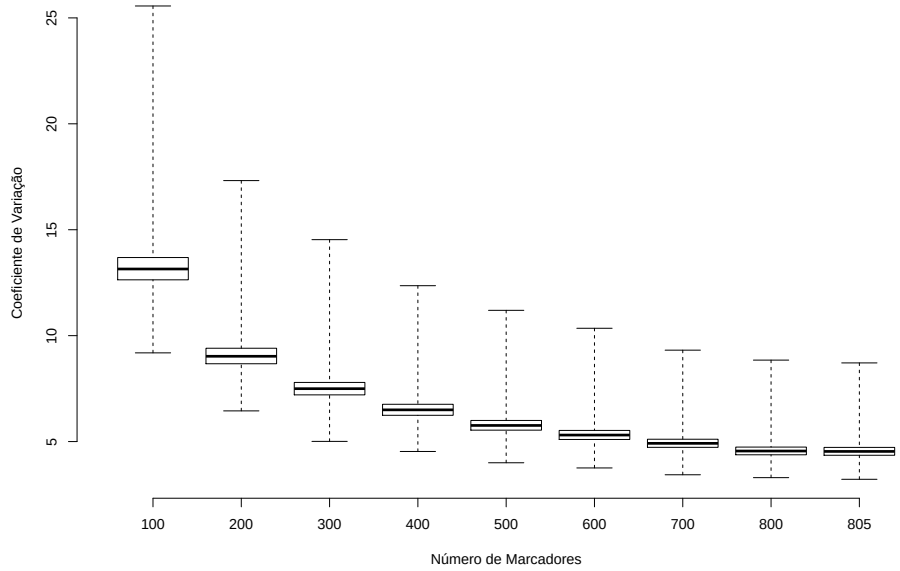


Figura 11 – Gráficos *boxplot* da distribuição *bootstrap* do coeficiente de variação das estimativas das distâncias de alelos compartilhados versus número de marcadores SNPs

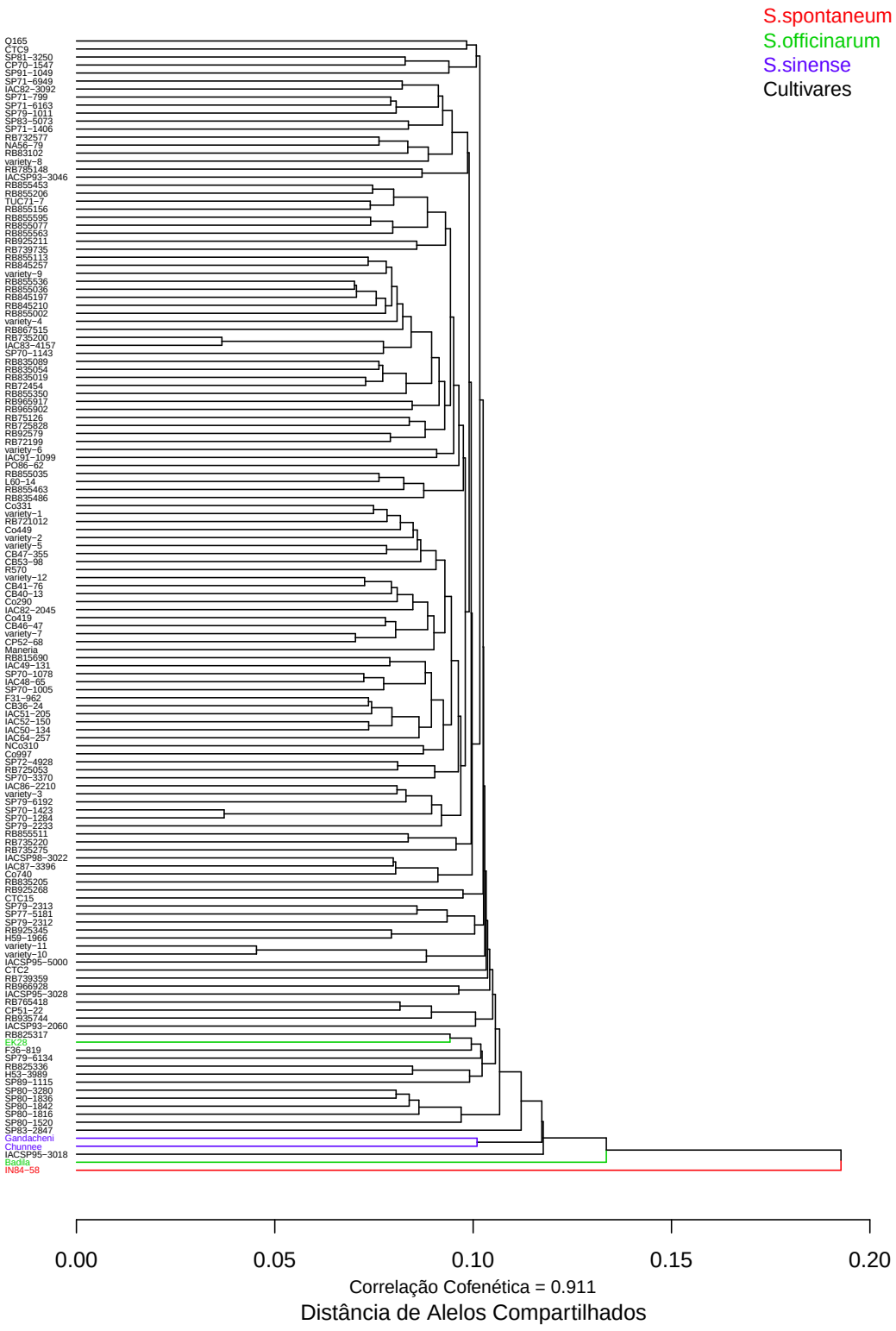


Figura 12 – Dendrograma resultante da análise de agrupamento pelo método UPGMA obtido com base na distância de alelos compartilhados

Figura 13 – Dendrograma resultante da análise de agrupamento pelo método UPGMA obtido com base na distância $1 - coeficiente\ de\ parentesco$

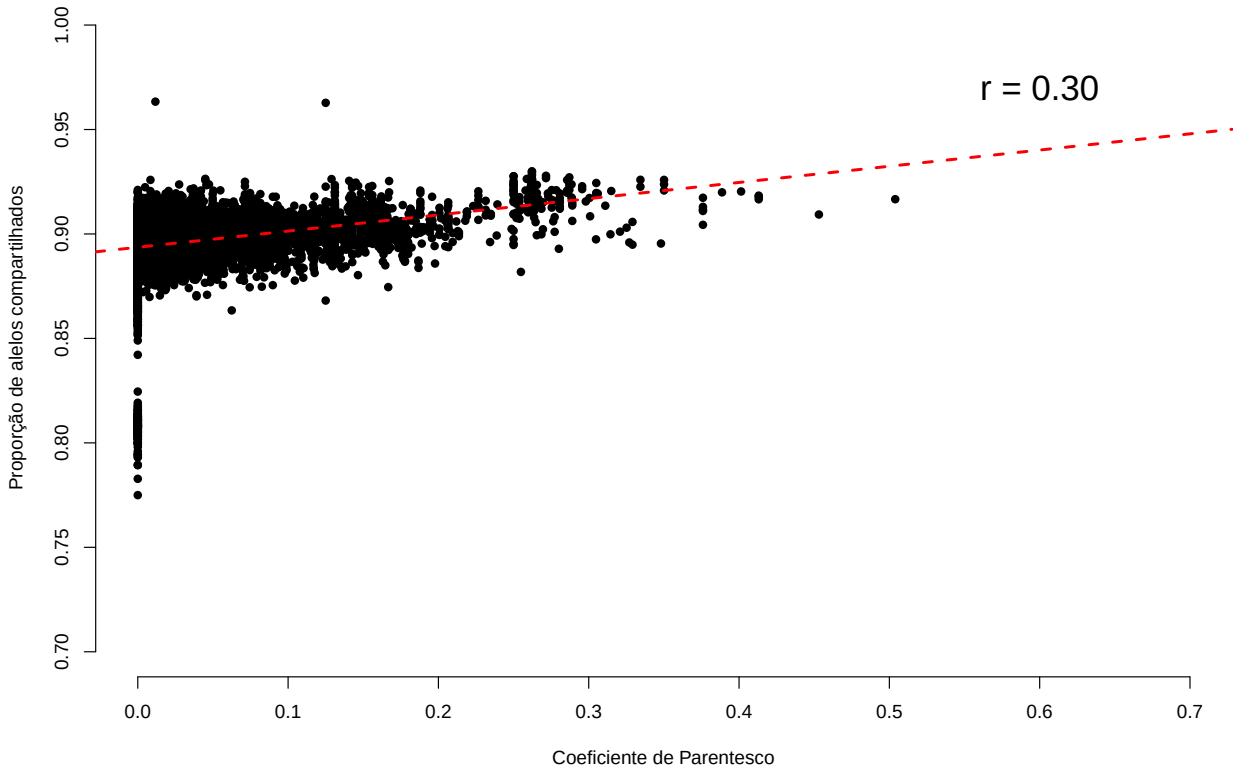


Figura 14 – Gráfico de dispersão entre proporção de alelos compartilhados obtida a partir dos dados de SNPs e estimativa de coeficiente de parentesco baseados na genealogia e correlação de Pearson r entre as variáveis

Tabela 2 – Estimativas das estatísticas propostas pelo método de Evano, Regnaut e Goudet (2005) para modelo de frequência correlacionada, considerando locos para cada nível de ploidia. K é o número de populações, $m(LnP(K))$ é a média do logaritmo da verossimilhança obtida pelas 10 repetições modelo, $s(LnP(K))$ é o desvio padrão, $m(Ln^{''}(K))$ é a média da primeira derivada do logaritmo da verossimilhança obtida pelas 10 repetições (Continua)

Nível de Ploidia	K	$m(LnP(K))$	$s(LnP(K))$	$m(Ln^{''}(K))$	Delta K
	1	-14148,18	0,0919		
	2	-13733,34	77,0172	91,85	1,192591
	3	-13410,35	36,5996	59,01	1,612312
	4	-13146,37	12,5493	75,09	5,983608
6	5	-12957,48	22,0255	22,53	1,022906
	6	-12746,06	3,6561	107,89	29,509538
	7	-12642,53	29,3250	11,28	0,384655
	8	-12550,28	11,0236	2,43	0,220436
	9	-12455,60	5,6403	24,38	4,322441
	10	-12385,30	23,3950		
	1	-19249,36	0,1506		
	2	-19148,14	3,2411	64,01	19,749318
	3	-19110,93	5,8262	12,99	2,229585
	4	-19060,73	4,5571	179,29	39,343366
8	5	-19189,82	34,5975	185,80	5,370337
	6	-19504,71	238,6222	249,03	1,043616
	7	-19570,57	194.7432	96,78	0,496962
	8	-19539,65	88.5284	89,32	1,008942
	9	-19598,05	108.6584	131,88	1,213712
	10	-19524,57	96.6392		
	1	-22131,11	0,0316		
	2	-22032,52	3,6453	22,16	6,079008
	3	-21956,09	9,0162	21,07	2,336901
	4	-21900,73	4,1695	17,23	4,132411
10	5	-21862,60	16,3961	32,00	1,951679
	6	-21856,47	27,2830	57,06	2,091415
	7	-21907,40	52,5030	56,83	1,082414
	8	-22015,16	114,0666	134,35	1,177821
	9	-22257,27	328,5467	295,16	0,898381
	10	-22204,22	160,1281		

Tabela 2 – Estimativas das estatísticas propostas pelo método de Evano, Regnaut e Goudet (2005) para modelo de frequência correlacionada, considerando locos para cada nível de ploidia. K é o número de populações, $m(LnP(K))$ é a média do logaritmo da verossimilhança obtida pelas 10 repetições modelo, $s(LnP(K))$ é o desvio padrão, $m(Ln^{''}(K))$ é a média da primeira derivada do logaritmo da verossimilhança obtida pelas 10 repetições (Continua)

Nível de Ploidia	K	$m(LnP(K))$	$s(LnP(K))$	$m(Ln^{''}(K))$	Delta K
	1	-38942,29	0,1197		
	2	-38837,52	2,1436	91,36	42,619474
	3	-38824,11	9,7533	9,42	0,965823
	4	-38801,28	6,7488	7,67	2,618246
12	5	-38796,12	39,2547	0,77	0,019615
	6	-38791,73	70,4330	8,08	0,114719
	7	-38779,26	22,8397	170,37	7,459382
	8	-38937,16	123,5080	83,44	0,675584
	9	-39178,50	183,7109	279,94	1,523807
	10	-39139,90	168,8862		
	1	-49019,94	0,0699		
	2	-48870,07	1,1016	97,87	88,84636
	3	-48818,07	2,6124	20,63	7,896998
	4	-48786,70	5,2351	42,23	8,066646
14	5	-48713,10	17,135	44,48	2,595853
	6	-48683,98	29,0413	90,99	3,133128
	7	-48745,85	69,1764	25,26	0,365153
	8	-48832,98	65,1204	7,56	0,116093
	9	-48927,67	130,1403	133,42	1,025201
	10	-48888,94	88,5196		
	1	-55944,11	0,0568	-	-
	2	-55785,23	2,4060	230,89	95,962897
	3	-55857,24	32,5725	571,29	17,539015
	4	-56500,54	237,1850	797,39	3,361890
16	5	-56346,45	269,2092	176,41	0,655290
	6	-56368,77	406,7774	315,13	0,774699
	7	-56075,96	353,1334	203,43	0,576071
	8	-55986,58	166,8723	51,92	0,311136
	9	-55949,12	53,3750	25,15	0,471194
	10	-55936,81	79,9514		

Tabela 2 – Estimativas das estatísticas propostas pelo método de Evano, Regnaut e Goudet (2005) para modelo de frequência correlacionada, considerando locos para cada nível de ploidia. K é o número de populações, $m(LnP(K))$ é a média do logaritmo da verossimilhança obtida pelas 10 repetições modelo, $s(LnP(K))$ é o desvio padrão, $m(Ln^{''}(K))$ é a média da primeira derivada do logaritmo da verossimilhança obtida pelas 10 repetições (Conclusão)

Nível de Ploidia	K	$m(LnP(K))$	$s(LnP(K))$	$m(Ln^{''}(K))$	Delta K
	1	-67101,79	0,0738		
	2	-67101,84	32,5617	499,22	15,331521
	3	-67601,11	792,2609	905,00	1,142300
	4	-67195,38	179,2434	967,52	5,397798
18	5	-67757,17	869,8173	970,39	1,115625
	6	-67348,57	118,9507	326,12	2,741641
	7	-67266,09	79,2618	31,61	0,398805
	8	-67152,00	85,5377	89,23	1,043166
	9	-67127,14	60,3542	2,74	0,045399
	10	-67099,54	48,0616		
	1	-80692,48	0,0632		
	2	-80719,30	69,7050	60,04	0,861344
	3	-80806,16	476,1825	277,43	0,582613
	4	-80615,59	4,2212	179,83	42,601333
20	5	-80604,85	1,9127	9,41	4,919806
	6	-80603,52	1,5569	1,56	1,001978
	7	-80600,63	2,1520	0,67	0,311334
	8	-80597,07	1,8355	4,73	2,576978
	9	-80588,78	14,7466	6,89	0,467227
	10	-80587,38	13,2473		

5 DISCUSSÃO

Para a realização de mapeamento associativo, um dos principais fatores que precisa ser considerado é a presença de estrutura populacional. Isto se deve ao fato de que painéis de genótipos formados a partir de coleções de germoplasma e/ou cultivares, a associação não aleatória entre alelos podem surgir por outras razões que não a ligação física.

No presente trabalho, a abordagem escolhida para estudar a estrutura populacional levou em consideração o fato de que a cana-de-açúcar é um poliploide bastante complexo, tendo inclusive diferente número de cromossomo homólogos por grupo de homologia. Para que isso fosse possível, uma nova classe de marcadores (SNPs) foi escolhida. Ao permitir que informações sobre a dosagem alélica seja usada, foi possível estimar a ploidia de cada loco, bem como realizar estudos sobre a estrutura populacional de uma forma inédita até o momento para essa espécie.

A primeira etapa foi relacionada à correta classificação dos dados de SNPs. Os resultados mostraram que locos com nível de ploidia 20 é o mais frequente no painel (29,94 %). Além disso, observou-se um total de 720 marcadores SNPs com ploidia diferente de 8, o que equivale a 89,44 %. Comparando a porcentagem de SNPs com ploidia igual a 8 encontradaas nesse trabalho (10,56 %) com a porcentagem do número de cromossomos que são derivados da *S. officinarum*, entre 72,5 % à 85 %, encontrada por Piperidis, Piperidis e D'Hont (2010) e D'Hont (2005), nota-se uma disparidade entre esses resultados. Entretanto a alta porcentagem de SNP cuja a estimativa da ploidia seja diferente de 8, pode ser explicada devida a existência de cromossomos que estão presentes em distintos grupos de homologias, cromossomos recombinantes entre duas ou mais espécies relacionadas ao genêro *Saccharum* e cromossomos oriundos de outras espécies que não sejam *S. officinarum*. Isto mostra que, pelo menos para o germoplasma brasileiro, não é correta a pressuposição de que a espécie geralmente tem comportamento auto-octaploide.

Uma possível explicação para a alta frequência de SNPs com estimativas de ploidia igual a 20 é a superparametrização do modelo, pois para estes casos o modelo implementado no *software* SuperMASSA minimiza o erro esperado entre a intensidade alélica observada e a esperada para dada uma ploidia. Portanto, para os locos que apresentaram significativa variância dentro dos grupos formados por indíviduos com mesmo genótipo, o modelo mais provável é aquele que ajusta o maior número de retas aumentando dessa forma a estimativa do nível de ploidia. Por essa razão, ainda é prematuro afirmar que existem grupos de ploidia 20. Tal hipótese precisa ser confirmada com mais evidências biológicas, como aquelas provenientes de estudos citogenéticos.

Um dos fatores para a ocorrência de SNPs que apresentam uma grande variância dentro dos grupos é uma possível falta de otimização da plataforma de genotipagem de larga escala Sequenom MassARRAY ® para o uso em espécies poliploides. Essa otimização da plataforma incluem mudanças nas condições de PCR, nas concentrações de DNA e de reagentes entre outras. Outras fontes de aumento da variância dentro desses grupos seriam SNPs duplicados no genoma e presença de SNPs nos primers (MOLLINARI; SERANG, 2013).

Convém ressaltar que a alta variabilidade dentro dos grupos é mais problemática conforme aumenta-se o nível de ploidia dos SNPs. Sob o ponto de vista da modelagem dos dados, para SNPs com ploidias mais altas, tais como 18 e 20, o ângulo das retas do modelo que definem a classe genotípica são menores em relação aos locos hexaploides ou octaploides, por exemplo, tornando essas retas mais próxima uma das outras. Logo, quanto maior variância dentro do grupo formado por indivíduos que estão em torno da mesma reta, maior será a dificuldade de distinguir e separar esses grupos.

Uma dificuldade adicional no contexto da classificação dos SNPs para os acessos de um painel de cana-de-açúcar é que, nessas circusntâncias, tem-se para o mesmo loco indivíduos de espécies diferentes e com níveis de ploidia diferentes. Por exemplo, para o mesmo loco tem-se indíviduos de espécies *S. officinarum*, que são autoctaploides, e cultivares comerciais, cujo o nível de ploidia é variável de acordo com grupo de homologia e desconhecido até o momento. Além disso, acessos provenientes de um painel não representam populações em equilíbrio de Hardy-Weinberg, sendo oriundos apenas de um único cruzamento biparental. Dessa forma, embora nesse trabalho e em outros tais como de Serang et al. (2012) foi utilizado o modelo de Hardy-Weinberg, em trabalhos futuros é interessante o desenvolvimento de modelos gráficos Bayesiano que levem em consideração pressuposições mais realísticas (MOLLINARI; SERANG, 2013).

Após o procedimento *ad hoc* de filtragem dos resultados, observou-se uma distribuição aproximadamente uniforme da frequência do alelo de referência dos SNPs. Normalmente, em estudos de mapeamento de QTLs e de mapeamento associativo, remove-se os SNPs que contém alelo de menor frequencia inferior a 0,05 por causa da falta de poder de detecção de QTL (LANDER, 1996; MYLES, 2009). Entretanto, nesse trabalho optou-se pela não exclusão de SNPs que apresentassem baixa frequência do alelo de referência. Isto foi feito pois para espécies poliploides como a cana-de-açúcar, podem haver SNPs que possuam alelos de menor frequência alélica inferior a 0,05 com frequências genotípicas maiores que 0,05. Por exemplo, para o SNP Contig1009#3b1 com ploidia estimada igual a 6, foram classificados 22 indivíduos com o genótipo A^1a^5 e 111 indivíduos com o genotipo A^0a^6. Isso resulta em uma frequência observada

do alelo A igual a $\frac{22}{6(111+22)} = \frac{22}{798} = 0,0275$, porém com frequência observada do genótipo $A^1a^5 = 0,165$ e $A^0a^6 = 0,835$. Ademais, a exclusão de SNPs com alelos de menor frequencia inferior a 0,05 podem ser a razão pela qual modelos de mapeamento associativos explicam muito pouco da variabilidade genética de caracteres de interesse, fato conhecido na literatura como *missing heritability* (MYLES, 2009).

Com relação a imputação dos dados, verificou-se para esse conjunto de dados o método de Stekhoven e Buhlmann (2012) não foi mais acurado do que imputar dados simplesmente através da média do número de cópias alélicas de cada loco. A explicação desse resultado deve-se a baixa variabilidade da dosagem alélica dentro de cada loco, como pode ser visto na Figura 6. Uma vez que a variância da dosagem alélica dentro de cada loco é baixa, os desvios da dosagem alélica de um indivíduo em relação a média dentro de cada loco são minimizados, tornando o estimador da média um preditor acurado para dados perdidos.

Quanto à presença de estrutura de população, principal objetivo do trabalho, a análise de componentes principais mostrou que o acesso IN84-58 (*S. spontaneum*) destaca-se dos demais acessos do painel que ainda estão separados entre si, formando três subpopulações. Uma dessas subpopulações é formada pelas espécies *S. sinense*, uma outra formada por acessos que são *S. officinarum* e, por fim, uma subpopulação formada pelos cultivares modernos que estão presentes no painel. Entretanto, essas subpopulações apresentam uma mudança gradual da frequência alélica, sugerindo que os indivíduos pertencentes à uma dessas subpopulações pode ter uma fração do seu genoma oriundo de outra subpopulação adjacente. Esses resultados são coerentes com o histórico evolutivo da espécie, dado que as espécies *S. sinense* são hibridos interespecíficos entre *S. officinarum* e *S. spontaneum* com maior contribuição no genoma da espécie *S. officinarum* (D' HONT et al., 2002), assim como os cultivares modernos, que ainda passaram pelo processo de nobilização (D'HONT et al. 1996, MING et al. 2006). Essa é provavelmente a razão pela qual os três grupos (*S. officinarum*, *S. sinense* e cultivares modernos) estão mais próximos entre si do que em relação ao acesso IN84-58. Um outro ponto importante é que o acesso IN84-58 não tem relação de parentesco com os demais acessos que compõe o painel; portanto, a dissimilaridade genética identificada nessa análise é também devido ao grau de parentesco entre eles.

Além disso, pode-se afirmar que a análise de componentes principais não revelou qualquer relação entre presença de estrutura populacional e a origem geográfica dos acessos. Esses resultados corroboram os obtidos por Lu et al. (1994), Aitken et al. (2006) e Raboim et al. (2008). Esses autores argumentam que esse fato ocorre porque ao longo dos anos houve um grande intercâmbio de germoplasma entre os programas de melhoramento, e que muitos programas de melhoramento espalhados pelo mundo utilizaram os mesmos ancestrais em seus respectivos

programas.

A análise de coordenadas principais aplicada a uma matriz de distância genética definda por $1 - coeficiente\ de\ parentesco$ mostrou uma maior dissimilaridade genética entre os acessos do painel. Provavelmente, as razões das diferenças dos resultados entre a análise de coordenadas principais e análise de componentes principais deve-se a algumas pressuposições irrealistas feitas para a determinação do coeficiente de parentesco. Por exemplo, 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. Além disso, para a estimativa do coeficiente de parentesco dos acesso desse painel, assumiu-se que os indivíduos fossem diploides, o que obviamente não é o caso da cana-de-açúcar (LIMA et al. 2002). A correlação baixa entre a matrizes de coeficientes de parentesco e de proporção de alelos compartilhados ($r = 0,30$), é um resultado que fortemente sugere essa hipótese.

Comparando os resultados da análise de componentes principais como os resultados obtidos por Rosa (2011), alguns pontos podem ser destacados. Rosa (2011) utilizou o modelo sem mistura do software STRUCTURE para analisar dados de microsatelites, e identificou existências de subpopulações entre os cultivares modernos de cana-de-açúcar. Assim, as diferenças encontradas podem ser principalmente devidas ao tipo de marcador utilizado, especialmente porque os microsatélites têm comportamento dominante e são neste caso menos informativos. Embora os SNPs levem em consideração a dosagem alélica e o nível de ploidia, podendo assim discriminar indivíduos que possuem doses únicas dos indivíduos que possuem doses mais altas, os SNPs possuem uma menor taxa de mutação, portanto são mais conservados ao longo do genoma do que outros marcadores (HAMBLIN; WARBURTON; BUCKLER; 2007). Além disso, por definição, SNPs são marcadores de polimorfismos em uma única base, ou seja, abrangem uma região menor do genoma em comparação aos microsatélites. Isto pode dificultar a identificação de subpopulações com frequências alélicas próximas.

Contudo, convém mencionar que diferentemente da análise de componentes principais, os modelos com e sem mistura implementados no software STRUCTURE são baseados nas pressuposições que dentro das subpopulações os locos estejam em equilíbrio de Hardy-Weinberg e equilíbrio de ligação. Portanto, no caso dos modelos implementados no *software* STRUCTURE, quaisquer desvios da pressuposição de Hardy Weinberg, causadas por exemplo, pelo grau de parentesco, podem influenciar na identificação das subpopulações. A tendência é que os modelos desse *software* aloquem os indvíduos com grau de parentesco mais próximos na mesma subpopulação e indivíduos com grau de parentesco mais distantes, em distintas subpopulações. No entanto, o objetivo da análise de componentes principais não é identificar subpopulações mas

sim projetar os indivíduos em um eixo de variação de acordo com a similaridade da frequência alélica dos mesmos. Nesse contexto, os resultados da análise de componentes principais desse trabalho pareceram ser satisfatórios.

Como análise complementar, foram feita análises de agrupamento aplicadas às matrizes de distâncias genéticas de alelos compartilhadas, estimadas a partir dos dados de marcadores SNPs. Sendo a distância genética definida como $1 - coeficiente\ de\ parentesco$. De uma forma geral, essas análises confirmaram os resultados obtidos nas análises de componentes principais e de coordenadas principais, respecitvamente. Comparando a média da proporção de alelos compartilhados dos dados de SNPs (0,8973) com a similaridade genética estimada através da distância de Jaccard usando marcadores AFLP (LIMA et al. 2002), 0,47, vê-se que os SNPs proporcionaram resultados que indicam uma maior similaridade genética entre os indivíduos dos que os marcadores AFLP. Isto possivelmente pode ser explicado pelo fato dos SNPs permitirem a exploração em maior profundidade do genoma, permitindo melhor mensuração da similaridade intra alélica.

Por fim, também foi ajustado o modelo com mistura e frequência correlacionada implementado no software STRUCTURE. Conforme apresentado, SNPs que foram divididos por nível de ploidia, o que fez sentido biológico. De um modo geral, as estatísticas delta indicaram que o número ótimo de subpopulações variou de acordo com o conjunto de dados analisados. Assim, as subpopulações formadas dependeram do nível de ploidia considerado, sugerindo que os resultados da análise de componentes principais são mais fáceis de ser interpretados do que os obtidos para o software STRUCTURE, tendo também maior significado biológico.

Comparando o número de subpopulações encontradas nesse trabalho para os conjuntos de SNPs com ploidias 8 e 20 com aqueles encontrados por Rosa (2011), vê-se que o número de subpopulações é o mesmo. No entanto, para os SNPs com ploidia 20 não foi possível verificar a partir de qual subpopulação o genoma de cada indivíduo é predominantemente originado, porque a maioria dos valores dos elementos da matriz Q está em torno de 0,25. Vale a ressalva que os SNPs cuja a estimativa do nível de ploidia foi igual a 20 foram os que apresentam menor probabilidade *a posteriori* na classificação. Assim, dados para ploidia 20 precisam ser interpretados com cautela.

Um fato interessante é que, por definição, usando análise de componentes principais, identifica-se o padrão de similaridade genética entre os indivíduos por meio dos primeiros componentes, descartando os que representam o ruído. Dessa forma, a análise de componentes principais talvez seja mais robusta à presença de SNPs com baixa probabilidade *a posteriori* do que os modelos implementados no software STRUCTURE. Com relação aos conjuntos de dados de

SNPs com ploidias 10 a 18, uma possível explicação para os resultados é que nesses casos as amostras de tamanho 50 representam uma porcentagem significativa da quantidade de dados (80,75 % 69,44 % , 76,92 %, 78,13 % 61,73%, respectivamente). Dessa forma, a chance dos locos ligados estarem e/ou desequilíbrio de ligação é muito grande, o que pode ter influenciado os resultados. Vale ressaltar que testes para verificar o desequilíbrio de ligação e ou a ligação entre os locos não foram feitos porque os métodos estatísticos para fazer isso com marcadores SNPs em cana-de-açúcar ainda não foram desenvolvidos.

Por outro lado, cada conjunto de dados de SNPs com seu respectivo nível de ploidia pode representar melhor uma espécie distinta do genêro *Saccharum*. Por exemplo, a matriz Q resultante da análise do software STRUCTURE com marcadores SNPs com ploidia igual a 8 podem representar em sua maioria a fração do genoma oriundas das espécies *S. officinarum*, enquanto a matriz Q obtida a partir de marcadores com ploidia mais altas podem representar em sua maioria a fração do genoma proveniente das espécies *S. spontaneum* ou dos híbridos interespecíficos *S. sinense* ou *S. barberi*. Sob essa hipótese, pode-se interpretar que a matriz Q obtida a partir dos marcadores SNPs com ploidia igual a 8 corresponde as informações obtidas com as genealogias. Para ploidia 8, nota-se que diversos irmãos completos estão na mesma subpopulação, como por exemplo, os cultivares SP80-1816, SP80-1842 e SP80-3280 (subpopulação 4) e os cultivares RB845197, RB845210 e RB845257 (subpopulação 3). Tais resultados são os mesmos que aqueles obtidos por Rosa (2011). Entretanto, convém mencionar que mesmo sob esse ponto de vista, as análises com o STRUCTURE têm desvantagem em relação à ACP, porque no contexto de mapeamento associativo, o objetivo da estrutura de população é controlar falsos positivos de uma forma geral, e não apenas ter uma estrutura de população para cada nível de ploidia. Assim sendo, as análises do STRUCTURE controlam apenas a estrutura populacional de uma parte do genoma, que pode evidentemente não conter os alelos dos QTLs de interesse no melhoramento. Portanto, as análises aqui apresentadas devem ser muito vantajosas.

Diante dos resultados da estrutura de população conclui-se 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. Isto tambpem foi relatado por Lima et al. (2002) e Roach (1989). Desse modo, espera-se que as informações aqui apresentadas sejam úteis para elaboração de trabalhos futuros de mapeamento associativo em cana-de-açúcar, bem como para auxiliar os melhoristas em obter um melhor conhecimento da cana-de-açúcar.

6 CONCLUSÃO

i.) Os marcadores SNPs foram eficientes para estudos de estrutura populacional em cana-de-açúcar.

ii.) Os resultados das analises de componentes principais possuem interpretação com maior significado biológico quando comparados com os resultados do software STRUCTURE. Isso é particularmente importante para espécies poliploides, como a cana-de-açúcar.

iii.) A análise de componentes principais revelou presença de estrutura de população que tem relação com a ordem taxonômica dos indivíduos, mas não com a origem geográfica dos acessos.

REFERÊNCIAS

AITKEN, K.S.; LI, J.C.; JACKSON, P.; PIPERIDIS, G.; McINTYRE, C.L.AFLP analysis of genetic diversity with *Saccharum Officinarum* and comparision to sugarcane cultivars. **Australian Journal Agricultural Research**, Queensland, v. 57, p. 1167-1184, 2006.

AITKEN, K.S.; McNEIL, M. Diversity analysis. In: HENRY, R.J.; KOLE, C. (Eds.). **Genetics, Genomics and Breeding of Sugarcane**. New Hampshire: Science Publishers, 2010. p. 19-42.

ALIX, K.; PAULET, F.; GLAZMANN, J.C.; D'HONT, A. Inter-Alu like species-specific sequences in the *Saccharum complex*. **Theoretical and Applied Genetics**, New York, v. 99, p. 962-968, 1999.

ARANZANA, M.J.; KIM, S.; ZHAO, K.; BAKKER, E.; HORTON, M.; JAKOB, K.; LISTER, C.; MOLITOR, J.; SHINDO, C.; TANG, C.; TOOMAIJAN, C.; TRAW, B.; ZHENG, H.; BERGELSON, J.; DEAN, C.; MARJORAM, P.; NORDBORG, M. Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. **PLoS Genetics**, San Francisco, v. 1, n. 5, 2005. Disponível em: <http://www.plos genetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.0010060>. Acesso em: 21 abr. 2012.

ARTSCHWAGER, E.; BRANDES, E. W. Sugarcane (S. officinarum L.). Origin, classification, characteristics and descriptions of representative clones. In: _____ Washington, 1958. Disponível em: <http://naldc.nal.usda.gov/catalog/CAT87208934/>. Acesso em: 20 abr. 2012.

ASTLE, W.; BALDING, D. Population structure and cryptic relatedness in Genetic association studies. **Statistical science**, Hayward, v. 24, p. 451-471, 2009

BARNAUD, A.; LACOMBE, T.; DOLIGEZ, A. Linkage disequilibrium in cultivated grapevine, *Vitis vinifera L.*. **Theoretical and Applied Genetics**, New York, v. 112, p. 708-716, 2010.

BOTSTEIN, D. WHITE, R.L.; SKOLNICK, M.; DAVIS, R.W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. **American Journal of Human Genetics**, Baltimore, v. 32, p. 314-331, 1980.

BOWCOCK, A.M.; RUIZ-LINARES, A.; TOMFOHRDE, J.; MINCH, E.; KIDD, J.R.; CAVALLI-SFORZA, L.L. High resolution of human evolutionary trees with polymorphic microsatellites. **Nature**, London, v. 368, p. 455-457, 1994.

BRADBURY, P.J.; ZHANG, Z.; KROON, D.E.; CASSTEVENS, T.M.; RAMDOSS, Y.; BUCKLER, E.S. TASSEL: software for association mapping of complex traits in diverse samples. **Bioinformatics**, Oxford, v. 23, n. 19, p. 2633-2635, 2007.

BREIMAN, L.; FRIEDMAN, J.H.; OLSHEN, R.A.; STONE, C.J. **Classification and Regression Trees**. Monterey: Wadsworth and Brooks/Cole. 1984. 358 p.

BROWN, A.H.D. Core collections a practical approach to genetic-resources management. **Genome**, Ottawa, v. 31, p. 818-824, 1989.

BROWN, P.J; MYLES, S.; KRESOVICH, S. Genetic support for phenotype-based racial classification in sorghum. **Crop Science**, Madison, v. 51, p. 224-230, 2011.

CASA, A.M.; PRESSOIR, G.; BROWN, P.J.; MITCHELL, S.H.; ROONEY, W.L.; TUINSTRA, M.R.; FRANKS, C.D.; KRESOVICH, S. Community resources and strategies for association mapping in sorghum. **Crop Science**, Madison, v. 48, p. 30-40, 2008.

CASTILLO, A.; DORADO, G.; FEUILLET, C.; SOURDILLE, P.; HERNANDEZ. Genetic structure and ecogeographical adaptation in wild barley (*Hordeum chilense Roemer et Schultes*) as revealed by microsatellite markers. **BMC plant**, London, v.10, n. 266, 2010. Disponível em: <http://www.biomedcentral.com/1471-2229/10/266>. Acesso em: 22 abr. 2012.

CANIATO, F.F.; GUIMARÃES, C.T.; HAMBLIN, M.; BILLOT, C.; RAMI, J.F.; HUFNAGELL, B.; KOCHIAN, L.V.; LIU, J. GARCIA, A.A.F.; HASH, T.; RAMU, P.; MITCHELL, S.; KRESOVICH, S.; OLIVEIRA, A.C.; AVELLARI, G.; BORÉM, A.; GLASZMANN, J.C.; SCHAFFERT, R.E.; MAGALHÃES, J.V. The relationship between population structure and aluminum tolerance in cultivated sorghum. **PLoS ONE**, San Francisco, v. 6, 2011. Disponível em: <http://www.plosone.org/article/info%3Adoi%2F10.1371 %2Fjournal.pone.0020830>. Acesso em: 25 abr. 2012.

CATO, S.A.; GARDNER, R.C; KENT, J.; RICHARDSON, T.E. A rapid PCR-based method for genetically mapping ESTs. **Theoretical and Applied Genetics**, New York, v. 102, p. 296-306, 2001.

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

D'HONT, A. Unravelling the genome structure of polyploids using FISH and GISH; examples of sugarcane and banana. **Cytogenetics and Genome Research**, Basel, v. 109, p. 27-33, 2005.

D'HONT, A.; PAULET, F.; GLASZMANN, J.C. Oligoclonal interspecific origin of "North Indian" and "Chinese" sugarcanes. **Chromossome Research**, London, v. 10, p. 253-262, 2002.

D'HOOP, B.B.; PAULO, M.J.; KOWITWANICH, K.; SENGERS, M.; VISSER, R.G.; van ECK, H.J.; van EEUWIJK, F.A. Population structure and linkage disequilibrium unravelled in tetraploid potato. **Theoretical and Applied Genetics**, New York, v. 121, p. 1151-1170, 2010.

EASTMENT, H.T.; KRZANOWSKI, W.J. Cross-valodatory choice of the number of components from a principal component analysis. **Technometrics**, Princeton v. 24, p. 73–77, 1982.

EEUWIJK, F.A. van; MALOSETTI, M.; YIN, X.; STRUIK, P.C.; STAM, P. Statistical models for genotype by environment data: from conventional ANOVA models to eco-physiological QTL models. **Australian Journal of Agricultural Research**, Melbourne, v. 56, p. 883-894, 2005.

ENGELHARDT, B.E.; STEPHENS, M.; Analysis of population structure: A unifying framework and novel methods based on sparse factor analysis. **PLoS Genetics**, San Francisco, v.6, 2010. Disponível em: <http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1001117>. Acesso em: 02 abr. 2012.

EUROPEAN PATENT OFFICE (Netherlands). Marc Zabeu; Pieter Vos. **Selective restriction fragment amplification**: a general method for DNA fingerprinting publication 0 534 858 A1, 01 abr 1993.

EVANNO, G.; REGNAUT, S.; GOUDET, J. Detecting the number of clusters of individuals using the software STRUCTURE : a simulation study. **Molecular Ecology**, West Sussex, v. 14, p. 2611-2620, 2005.

FAO. Food and agricultural commodities production: Disponível em <http://faostat.fao.org/site/339/default.aspx>. Acesso em: 01 jul. 2012.

FALUSH, D.; STEPHENS, M.; PRITCHARD, J.K.Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. **Genetics**, Bethesda, v. 164, p. 1567-1587, 2003.

FALUSH, D.; STEPHENS, M.; PRITCHARD, J.K.Inference of population structure using multilocus genotype data: dominant markers and null alleles. **Molecular Ecology Notes**, Oxford, v. 7, p. 574-578, 2007.

FERREIRA, M.E.; GRATTAPAGLIA, D. **Introdução ao uso de marcadores moleculares em análise genética**. 3.ed. Brasília: EMBRAPA-CENARGEN, 1998. 220 p.

FLINT-GARCIA, S.A.; THORNSBERRY, J.M.; BUCKLER, E.S.; Structure of linkage disequilibrium in plants. **Annuals Reviews in Plant Biology**, Palo Alto, v. 54, p. 357-374, 2003.

FUJISAWA, H.; EGUCHI, S.; USHIJIMA, M.; MIYATA, S.; MIKI, Y.; MUTO, T.; MATSUURA, M. Genotyping of single nucleotide polymorphisms using model based clustering. **Bioinformatics**, London, v. 20, p. 718-726, 2004.

GABRIEL, S.B.; SCHAFFNER, S.F.; NGUYEN, H.; MOORE, J.M.; ROY, J.; BLUMENSTIELL, B.; HIGGINS, J.; DEFELICE, M.; LOCHNER, A.; FAGGART, M.; LIU, C.S.N.; ROTIM, C.; ADEYEMO, A.; COOPER, R.; WARD, R.; LANDER, E.S.; DALY, M.J.; ALTSHULER, D. The structure of haplotype blocks in the human genome. **Science**, Washington, D.C., v. 296, p. 2225-2229, 2002.

GARRIS, A.J.; TAI, T.H.; COBURN, J.; KRESOVICH, S.; McCOUCH, S. Genetic Structure and Diversity in *Oryza sativa L.* **Genetics**, Bethesda, v.169, p. 1631-1638, 2005.

GELMAN, A.; CARLIN, J.B.; STERN,H.S.; RUBIN, D.B.; DUNSON, D.B. **Bayesian data analysis**. 3.ed. London, UK: Chapman & Hall, 1995. 526 p.

GRIFFITHS, A.J.F.; WESSLER, S.R.; LEWONTIN, R.C.; CARROLL, S.B. **An introduction to genetic analysis**, 9.ed, New York, W.H.: Freeman and Co, 2008, 712 p.

GRIVET, L.; ARRUDA, P. Sugarcane genomics: depicting the complex genome of an important tropical crop. **Current Opinion in Plant Biology**, Amsterdam, v. 5, p. 122-127, 2001.

GRIVET. L; GLAZMANN, J.C.; D'HONT, Molecular evidences for sugarcane evolution and domestication. In: MOTLEY, T.; ZEREGA, N.; CROSS, H. (Ed.). **Darwin's Harvest. New Approaches to the Origins, Evolution and Conservation of Crops**. New York: Columbia University Press, 2006. p. 49-66.

GUPTA, P.K.; ROY, J.K.; PRASAD, M. Single nucleotide polymorphisms: A new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. **Current Science**, Washington, D.C., v. 80, p. 524-535, 2001.

HAMBLIN, M.T; WARBUTON, M.L.; BUCKLER, E.S. Empirical comparison of simple sequence repeats and single nucleotide polymorphisms in assessment of maize diversity and relatedness. **PLoS ONE**, San Francisco, v.7, 2007. Disponível em: <http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0001367>. Acesso em: 05 abr. 2012.

HANSEY, C.N.; JOHNSON, J.M.; SEKHON, R.S.; KAEPPLER, S.M.; de LEON, N. Genetic diversity of a maize association population with restricted phenology. **Crop Science**, Madison, v.51, p.704-715, 2011.

HEDRICK, P.W. **Genetics of populations**. 4.ed. Sudbury, MA: Jones and Bartlett Publishers, 2010. 675 p.

HEINZ, D.J.; TEW, T.L. Hybridization procedures. In: HEINZ, D.J. (Ed.). **Sugarcane improvement through breeding**. Amsterdam: Elsevier, 1987. p. 313-342.

HENRY, R. Basic Information on the Sugarcane Plant. In: _____ **Genetics, genomics and breeding of sugarcane**. New Hampshire: Science Publishers, 2010. p. 43-68.

HOSPITAL, F.; MOREAU, L.; LACOUNDRE, F.; CHARCOSSET, A.; GALLAIS, A.; More on the efficiency of marker-assisted selection. **Theoretical and Applied Genetics**, New York, v. 95, p. 1181-1189, 1997.

HUBISZ, J.M.; FALUSH, D.; STEPHENS, M.; PRITCHARD, J.K. Inferring weak population structure with the assistance of sample group information. **Molecular Ecology Resources**, Oxford, v. 9, p. 1322-1332, 2009.

HUANG, E.; AITKEN, K.; GEORGE, A. Association studies. In: HENRY, R.J.; KOLE, C. (Eds.). **Genetics, genomics and breeding of sugarcane**. New Hampshire: Science Publishers, 2010. p. 43-68.

HUANG, X.; WEI, X.; SANG,T.; ZHAO, Q.; FENG, Q.; ZHAO, Y.; LI, C.; ZHU, C.; LU, T.; ZHANG, Z.; LI, M.; FAN, D.; GUO, Y.; WANG, A.; WANG, L.; DENG, L.; LI, W.; LU, Y.; WENG, Q.; LIU, K.; HUANG, T.; ZHOU, T.; JING, Y.; LI, W.; BUCKLER, E.S.; QIAN, Q.;

ZHANG, Q.F; LI, J.; HAN, B. Genome-wide association studies of 14 agronomic traits in rice landraces. **Nature Genetics**, New York, v. 42, p. 961–967, 2010.

IRVINE, J.E. *Saccharum* species as horticultural classes. **Theoretical and Applied Genetics**, New York, v. 98, p. 186-194, 1999.

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

JOLLIFFE, I. **Principal Component Analysis**. 3.ed. New York, Springer Verlag, 1986. 485 p.

JOHNSON, R.A,; WICHERN, D. W. **Applied multivariate statistical analysis**. 1.ed. Englewood Cliffs, N.J. Prentice Hall, 1992. 642 p.

JUN, T.H.; VAN, K.; KIN, M.Y.; LEE, S.H.; WALKER, D.R. Association analysis using SSR markers to find QTL for seed protein content in soybean. **Euphytica**, Dordrecht, v. 162, p. 179-191, 2008.

KRAAKMAN, A.T; NIKS, R.E.; van den BERG, P.M.M.M.; STAN, P.; van EEUWIJK, F.A. Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. **Genetics**, Bethesda, v.168, p. 435-446, 2004.

LANDELL, M.G.A.; BRESSIANI, J.A. Melhoramento genético, caracterização e manejo varietal. In: DINARDO-MIRANDA, L.L.; VASCONCELOS, A.C.M.; LANDELL, M.G.A. (Eds.). **Cana-de-açúcar**. Campinas: IAC, 2008. p. 101-155.

LANDER, E.S.; SCHORK, N.J. Genetic dissection of complex traits. **Science**, Washington, D.C., v. 265, p. 2037-2048, 1994.

LANDER, E.S.; The new genomics: global views of biology. **Science**, Washington, D.C., v. 274, p. 536-539, 1996.

LE GOUIS, J.; BORDES, J.; RAVEL, C.; HEUMEZ, E.; FAURE, S.; PRAUD, S.; GALIC, N.; REMOUÉ, C.; BALFOURIER, F.; ALLARD, V.; ROUSSET, M. Genome-wide association analysis to identify chromosomal regions determining components of earliness in wheat. **Theoretical and Applied Genetics**, New York, v. 124, p. 597-611, 2012.

LIMA, M.L.A.; GARCIA, A.A.F.; OLIVEIRA, K.M.; MATSUOKA, S.; ARIZONO, H.; SOUZA JUNIOR, C.L.; SOUZA, A.P. Analysis of genetic similarity detected by AFLP and coefficient of parentage among genotypes of sugar cane (Saccharum spp.). **Theoretical and Applied Genetics**, New York, v. 104, p. 30-38, 2002.

LI, Y.; GUAN, R.; LIU, Z.; MA, Y.; WANG, L.; LIN, F.; LUAN, W.; CHEN, P.; YAN, Z.; GUAN, Y.; ZHU, L.; NING, X.; SMULDERS, M.J.M.; LI, W.; PIAO, R.; CUI, Y.; YU, Z.; GUAN, M.; CHANG, R.; HOU, A.; SHI, A.; ZHANG, B.; ZHU, S.; QIU, L. Genetic structure and diversity of cultivated soybean (*Glicine max* (L.)) Merr) landraces in China. **Theoretical and Applied Genetics**, New York, v. 117, p. 857-871, 2008.

LOPES, F.C.C.L. **Mapeamento genético de cana-de-açúcar (*Saccharum spp*.) por associação empregando marcadores SSR e AFLP**. 2011. 144 p. Tese (Doutorado em Genética e Melhoramento de Plantas) - Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2011.

LU, Y.H.; D’HONT, A, PAULET, F.; GRIVET, L; ARNAUD, M.; GLASZMANN, J.C. Molecular diversity and genome structure in modern sugarcane varieties. **Euphytica**, Dordrecht, v. 78, p. 217-226, 1994.

LYNCH, M.; WALSH, B. **Genetics and analysis of quantitative traits**. Sunderland: Sinauer Associates, 1998. 980 p.

MALOSETTI, M.; VOLTAS, J.; ROMAGOSA, I.; ULLRICH, S.E.; EEUWIJK, F.A. van. Mixed models including environmental covariables for studying QTL by environment interaction. **Euphytica**, Wageningen, v. 137, p. 139-145, 2004.

MALOSETTI, M.; van der LINDEN, C.G.; VOSMAN, B.; EEUWIJK, F.A.; A mixed-model approach to association mapping using pedigree information with an illustration of resistance to *Phytophthora infestans* in potato. **Genetics**, Bethesda, v. 175, p. 879-889, 2007.

MARCONI, T.G. **Mapa Funcional em cana-de-açúcar utilizando marcadores moleculares baseados em SSR e SNP** . 2011. 147 p. Tese (Doutorado em Genética e Biologia Vegetal - Área de Concentração: Genética Vegetal e Melhoramento) - Instituto de Biologia, Universidade de Campinas, Campinas, 2011.

McINTYRE, C.L.; WHAN, V. A.; CROFT, B.; MAGAREY, R.; SMITH, G.R. Identification and validation of molecular markers associated with Pachymetra root rot and brown rust resistance in sugarcane using map- an association-based approaches. **Molecular Breeding**, Berlin, v. 16, p. 151-161, 2005.

MATSUOKA, S.; GARCIA, A.A.F.; ARIZONO, H. Melhoramento da cana-de-açúcar. In: BORÉM, A. (Ed.) **Melhoramento de espécies cultivadas**. Viçosa: UFV, 1999. p. 205-252.

MATHER, K.A.; CAICEDO, A.L.; POLATO, N.R.; OLSEN, K.M.; McCOUCH, S.; PURUGGANAN, M.D. The Extent of Linkage Disequilibrium in Rice (Oryza sativa L.). **Genetics**, Bethesda, v. 177, p. 2223-2232, 2007.

MING, R.; MOORE, P.H.; WU, K.K.; D'HONT, A.; GLASZMANN, J.C.; TEW, T.L.; MIRKOV, T.E.; DA SILVA, J.; JIFON, J.; RAI, M.; SCHNELL, R.J.; BRUMBLEY, S.M.; LAKSHMANAN, P.; COMSTOCK, J.C.; PATERSON, A.H. Sugarcane improvement through breeding and biotechnology. **Plant Breeding Reviews**, Westport, v. 27, p. 15-100, 2006.

MOLLINARI, M. **Comparação de algortimos usados na construção de mapas genéticos**. 2007. 76 p. Dissertação (Mestrado em Genética e Melhoramento de Plantas) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2007.

MOLLINARI, M.; SERANG, O. Quantitative SNP Genotyping of Polyploids with MassARRAY and Other Platforms, In: WALKER, J. M. (Ed.). **Methods in Molecular Biology**, New York: Springer, 2013. In Press.

MYLES, S.; PEIFFER, J.; BROWN, P.J.; ERSOZ, E.S.; ZHANG, Z.; COSTICH, D.E.; BUCKLER, E.S. Association Mapping: Critical Considerations Shift from Genotyping to Experimental Design. **Plant Cell**, Baltimore, v. 21, p. 2194-2202, 2009.

MYLES, S.; BOYKO, A.R.; OWENS, C.L.; BROWN, P.J.; GRASSI, F.; ARADHYA, M.K.; PRINS, B.; REYNOLDS, A.; CHIA, J.M.; WARE, D.; BUSTAMANTE,C.D.; BUCKLER, E.S. Genetic structure and domestication history of the grape. **Proceedings of the National Academy of Science of the United States of America**. Washington, v.108, p. 3530-3535, 2011.

NOVEMBRE, J.; JOHNSON, T.; BRYC, J.; KUTALIK, J.; BOYKO, A. R.; AUTON, A.; INDAP, A.; KING, K. S.; BERGMANN, S.; NELSON, M. R.; STEPHENS, M.; BUSTAMANTE, C. D. Genes mirror geography within Europe. **Nature**, London, v. 456, p. 98-101, 2008.

OETH, P.; de MISTRO, G.; MARNELLOS, G.; SHI, T. van DEN BOOM, D. Single nucleotide polymorphisms. In: _____ **Qualitative and quantitative genotyping using single base primer extension coupled with matrix-assisted laser desorption / ionization time of flight mass spectrofotometry (MassARRAY)**. New York: Humana Press, 2009. p. 307-343.

OLIVIER, M.; CHUANG, L.M; CHANG, M.S.; CHEN, Y.T.; PEI, D.; RANADE,K.; WITTE, A.; ALLEN, J.; TRAN, NGUYET, T.; CURB, D.; PRATT, R.; NEEFS, H.; INDIG, M.A.; LAW, S.; NERI, B.; WANG, L.; COX, D.R. High-throughput genotyping of single nucleotide polymorphisms using new biplex invader technology. **Nucleic Acids Research**, London, v.30, 2002. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC117295>. Acesso em: 06 mar. 2012.

ORAGUZIE, N.C.; WILCOX, P.L.; RIKKERINK, E.H.A.; SILVA, H.N. de. Linkage disequilibrium. In: ORAGUZIE, N.C.; RIKKERINK, E.H.A.; GARDINER, S.E.; SILVA, H.N. de. (Eds.) **Association mapping in plants**. New York: Springer, 2007. p. 11-39.

PASAM, R.K.; SHARMA, R.; MALOSETTI, M.; van EEUWIJK, F.A.; HASENEYER, G.; KILIAN, B.; GRANER, A. Genome-wide association studies for agronomical traits in a world wide spring barley collection. **BMC Plant Biology**, London: v.12, 2012. Disponível em: <http://www.biomedcentral.com/1471-2229/12/16>. Acesso em: 02 nov. 2012.

PASTINA, M.; PINTO, L.R.; OLIVEIRA, K.M.; SOUZA, A.P.; GARCIA, A.A.F. Molecular mapping of complex traits. In: HENRY, R.J.; KOLE, C. (Eds.). **Genetics, Genomics and Breeding of Sugarcane**. New Hampshire: Science Publishers, 2010. p. 117-148.

PATTERSON, N.; PRICE A. L.; REICH, D.; Population structure and eigenanalysis. **PLoS Genetics**, San Francisco, v.2, 2006. Disponível em: <http://www.plosgenetics.org/article/info:doi/10.1371/journal.pgen.0020190>. Acesso em: 05 jan. 2011.

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

PRICE, A.L.; PATTERSON, N.J.; PLENGE, R.M.; WEINBLATT, M.E.; SHADICK, N.A.; REICH, D. Principal components analysis corrects for stratification in genome-wide association studies. **Nature Genetics**, New York, v. 38, p. 904-909, 2006.

PRITCHARD, J. K.; STEPHENS, M.; DONNELLY, P.; Inference of population structure using multilocus genotype data. **Genetics**, Bethesda, v.155, p. 945-959, 2000.

PRITCHARD, J.K.; PRZEWORSKI, M. Linkage disequilibrium in humans: models and data. **American Journal of Human Genetics**, Chicago, v. 69, p. 1-14, 2001.

R DEVELOPMENT CORE TEAM. **R:** A language and environment for statistical computing. Vienna, 2012. Disponível em: <http://www.R-project.org>. Acesso em: 01 out. 2012.

RABOIN, L.M.; PAUQUET, J.; BUTTERFIELD, M.; D'HONT, A.; GLASZMANN, J.C. Analysis of genome-wide linkage disequilibrium in the highly polyploid sugarcane. **Theoretical and Applied Genetics**, New York, v. 116, p. 701-714, 2008.

RAFALSKI, A. Applications of single nucleotide polymorphisms in crops genetics. **Current Opinion in Biotechnology**, London, v. 5, p. 94-100, 2002.

RANADE, K. CHANG, M.S.; TING, C.T.; PEI, D.; HSIAO, C.F.; OLIVIER, M.; PESICH, R.; HEBERT, J.; CHEN, Y.D.; DZAU, V.J.; CURB, D.; OLSHEN, R.; RISCH, N.; COX, D.R.; BOTSTEIN, D. High-throughput genotyping with single nucleotide polymorphisms. **Genome Research**, Woodbury, v. 11, p. 1262-1268, 2001.

REIF, J.C.; XIA, X.C.; MELCHINGER, A.E.; WARBURTON, M.L.; HOISINGTON, D.A.; BECK, D.; BOHN, M.; FRISCH, M. Genetic Diversity Determined within and among CIMMYT Maize Populations of Tropical, Subtropical, and Temperate Germplasm by SSR Markers. **Crop Science**, Madison, v. 44, p. 326-334, 2004.

ROACH, B.T. Origin and improvement of the genetic base of sugarcane. In: PROCEEDINGS OF THE AUSTRALIAN SOCIETY OF SUGARCANE TECHNOLOGISTS, 1989, Queensland. **Proceedings...**, Queensland: Australian Society of Sugar Cane Technologists, 1989. p. 34-47.

ROSA, J.R.B.F. **Análise do desequilíbrio de ligação e da estrutura populacional do germoplasma brasileiro de cana-de-açúcar**. 2011. 97 p. Dissertação (Mestrado em Genética e Melhoramento de Plantas) - Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2011.

SAITOU, N.; NEI, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. **Molecular Biology and Evolution**, Chicago, v. 4, p. 406-425, 1987.

SANTORO,A. **Identificação de *Single Nucleotide Polymorphisms* SNPs no gene Nove-cis-epoxicarotenóide dioxigenase (NCED) em *Eucalyptus***. 2010. 70 p. Dissertação (Mestrado em Ciências Biológicas - Área de Concentração: Genética) - Instituto de Biociências, Universidade Estadual Paulista “Júlio Mesquita Filho”, Campinas, 2010.

SATAGOPAN,J.M.; YANDELL, B.S.; NEWTON, M.A.; OSBORN, T.C. A Bayesian approach to detect quantitative trait loci using Markov Chain Monte Carlo. **Genetics**, Bethesda, v. 144, p. 805-816, 1999.

SEGALLA, A. L.; ALVAREZ, R.. Melhoramento da cana-de-açúcar: I - Experiências com os "seedlingss" obtidos em 1947, 1948 e 1949. **Bragantia**, Campinas, v. 23, p. 187-223, 1964.

SERANG, O.R.; MOLLINARI, M.; GARCIA, A.A.F.; Efficient exact maximum a posteriori computation for bayesian SNP genotyping in polyploids. **PLoS ONE**, San Francisco, v.7, 2012. Disponível em: <http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal .pone.0030906>. Acesso em: 02 mar. 2012.

SHRINER, D. Investigating population stratification and admixture using eigenanalysis of dense genotypes. **Heredity**, London, v. 107, p. 413-420, 2011.

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

SOKAL, R. R.; MICHENER, C. D. A statistical method for evaluating systematic relationships. **University of Kansas Science Bulletin**, Kansas, v.38, p. 1409-1438, 1958.

STEVENSON, G.C. **Genetics and breeding of sugar cane.** London: Longman, 1965. 284 p.

STICH, B.; MOHRING, J.; PIEPHO, H.P.; HECKENBERGER, M.; BUCKLER, E.S.; MELCHINGER, A.E. Comparison of mixed-model approaches for association mapping. **Genetics**, Bethesda, v. 178, p. 1645-1754, 2008.

STICH B.; MAURER, H.P.; MELCHINGER, A.E.; FRISCH, M.; HECKENBERGER, M.; van der VOORT, J.R.; PELEMAN, J.; SORENSEN, A.P.; REIF, J.C. Comparison of Linkage Disequilibrium in Elite European Maize Inbred Lines using AFLP and SSR Markers. **Molecular Breeding**, Berlin, v. 17, p. 217-226, 2006.

STEKHOVEN, D. J.; BUHLMANN, P. MissForest non-parametric missing value imputation for mixed-type data. **Bioinformatics**, Oxford, v. 28, p. 112-118, 2012.

SOUZA, A.P. Biologia molecular aplicada ao melhoramento. In: NASS, L.L.; VALOIS, A.C.C.; MELLO, I.S.; VALADARES-INGLIS, M.C. (Eds.). **Recursos genéticos e melhoramento - plantas**. Rondonópolis: Fundação MT, 2001. p. 939-965.

TAUTZ, D. Hypervariability of simple sequences as a general source for polymorphic DNA markers. **Nucleic Acids Research**, London, v. 17, p. 6463-6471, 1989.

TEO, Y.Y.; INOUYE, M.; SMALL, K.S.; GWILLIAM, R.; DELOUKAS, P.; KWIATKOWSKI, D.P.; CLARK, T.G. A genotype calling algorithm for the Ilumina Bead Array plataform. **Bioinformatics**, London, v. 24, p. 2741-2746, 2008.

TIAN, F.; BRADBURY, P.J.; BROWN, P.J.; HUNG, H.; SUN, Q.; FLINT-GARCIA, S.; ROCHEFORD, T.R.; McMULLEN, M.D.; HOLLAND, J.B.; BUCKLER, E.S. Genome-wide association study of leaf architecture in the maize nested association mapping population. **Nature Genetics**, New York, v. 43, p. 159–162, 2011.

TIVANG, J.G.; NIENHUIS, J.; SMITH, O.S. Estimation of sampling variance of molecular marker data using the bootstrap procedure. **Theoretical and Applied Genetics**, New York, v. 89, p. 259-264, 1994.

van INGHELANDT, D.; MELCHINGER, A.E.; LEBRETON, C.; STICH, B. Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. **Theoretical and Applied Genetics**, New York, v. 120, p. 1289-1299, 2010.

VOORRIPS R.E.; GORT, G.; Vosman, B. Genotype calling in tetraploid species from bi-allelic marker data using mixture models. **BMC bionformatics**, London, v. 12, p. 172, 2011.

WEI, X.; JACKSON, P.A.; HERMANN, S.; KILIAN, A.; HELLER-USZYNSKA, K.; DEOMANO E. Simultaneously accounting for population structure, genotype by environment interaction, and spatial variation in marker-trait associations in sugarcane. **Genome**, Ottawa, v. 53, p. 973-981, 2010.

WEI, X.; JACKSON, P.A.; McINTREY, C.L.; AITIKEN, K.S.; CROFT, B. Associations between DNA markers and resistance to diseases in sugarcane and effects of population substructure **Theoretical and Applied Genetics**, New York, v.114, p. 155-164, 2006

WILLIANS, J.G.K.; KUBELIK, A.R.; LIVAK, K.J.; RAFALSKI, J.A; SCOTT, V. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. **Nucleic Acids Research**, London, v.18, p. 6531-6535, 1990.

YAN, L.; TSENG, G.C.; CHEONG, S.Y.; BEAN, L.J.; SHERMAN, S.L.; FEINGOLD, E. Smarter clustering methods for SNP genotype calling. **Bioinformatics**, London, v. 24, p. 2665-2671, 2008.

YU, J.; BUCKLER, E.S.; Genetic association mapping and genome organization of maize. **Current Opinion in Biotechnology**, London, v. 17, p. 155-160, 2006.

ZHAO, K.; ARANZANA, M.J.; KIM, S.; LISTER, C.; SHINDO, C.; TANG, C.; TOOMAJIAN, C.; ZHENG , H.; DEAN, C.; MARJORAM, P.; NORDBORG, M. An arabidopsis example of association mapping in structured samples. **PLoS ONE**, San Francisco, v. 3, 2007. Disponível em: <http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.0030004>. Acesso em: 03 dez. 2011.

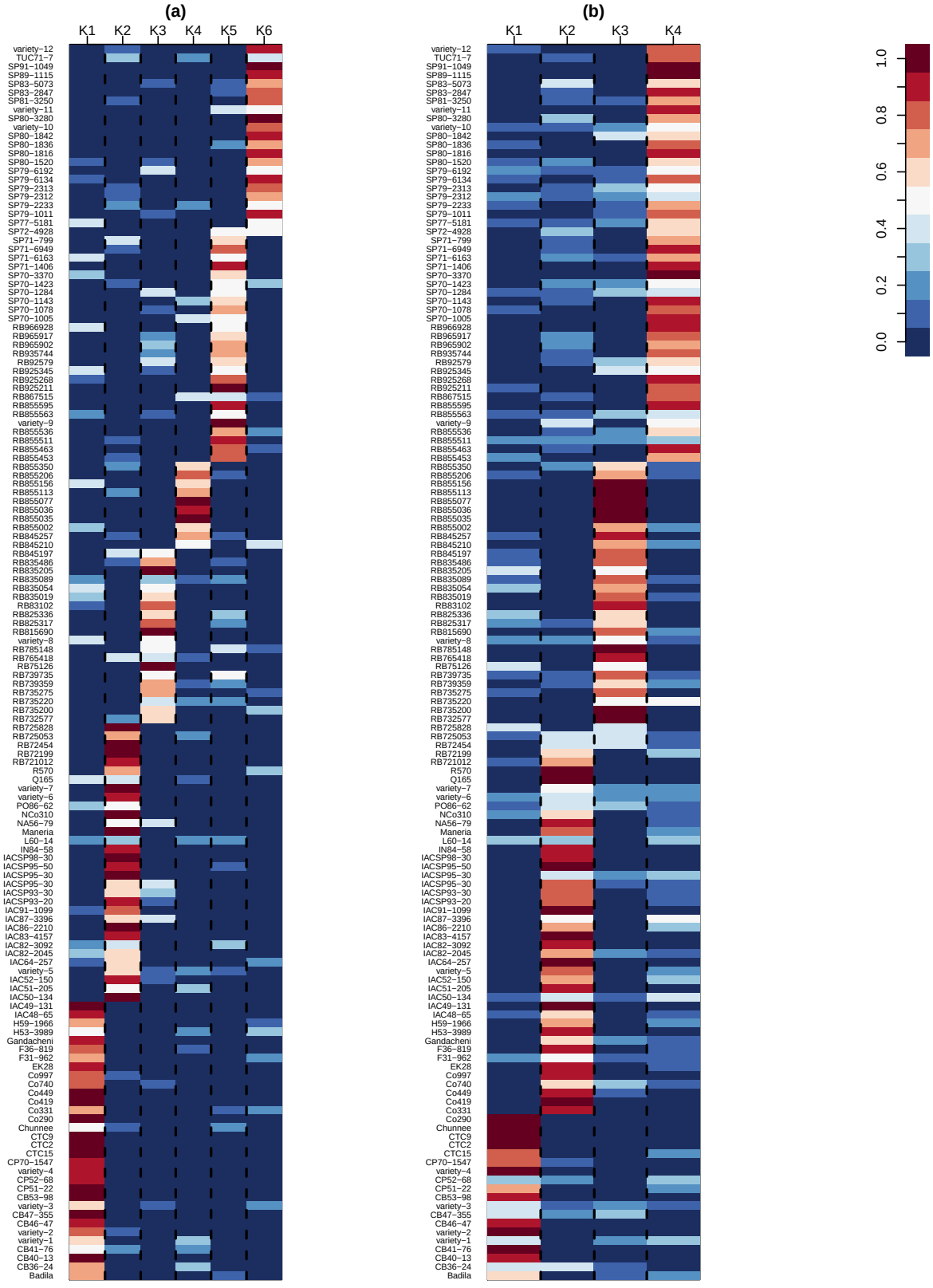
ZHANG, P.; LI, J.; LI, X.; LIU, X.; ZHAO, X.; LU, Y. Population structure and genetic diversity in a rice core collection (Oryza sativa L.) investigated with SSR markers. **PLoS ONE**, San Francisco, v. 6, 2011. Disponível em: <http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0027565>. Acesso em: 04 fev. 2012.

ZHU, C.; GORE, M.; BUCKLER, E.S.; YU, J. Status and prospects of association mapping in plants. **The Plant Genome**, Madison, v. 1, p. 5-20, 2008.

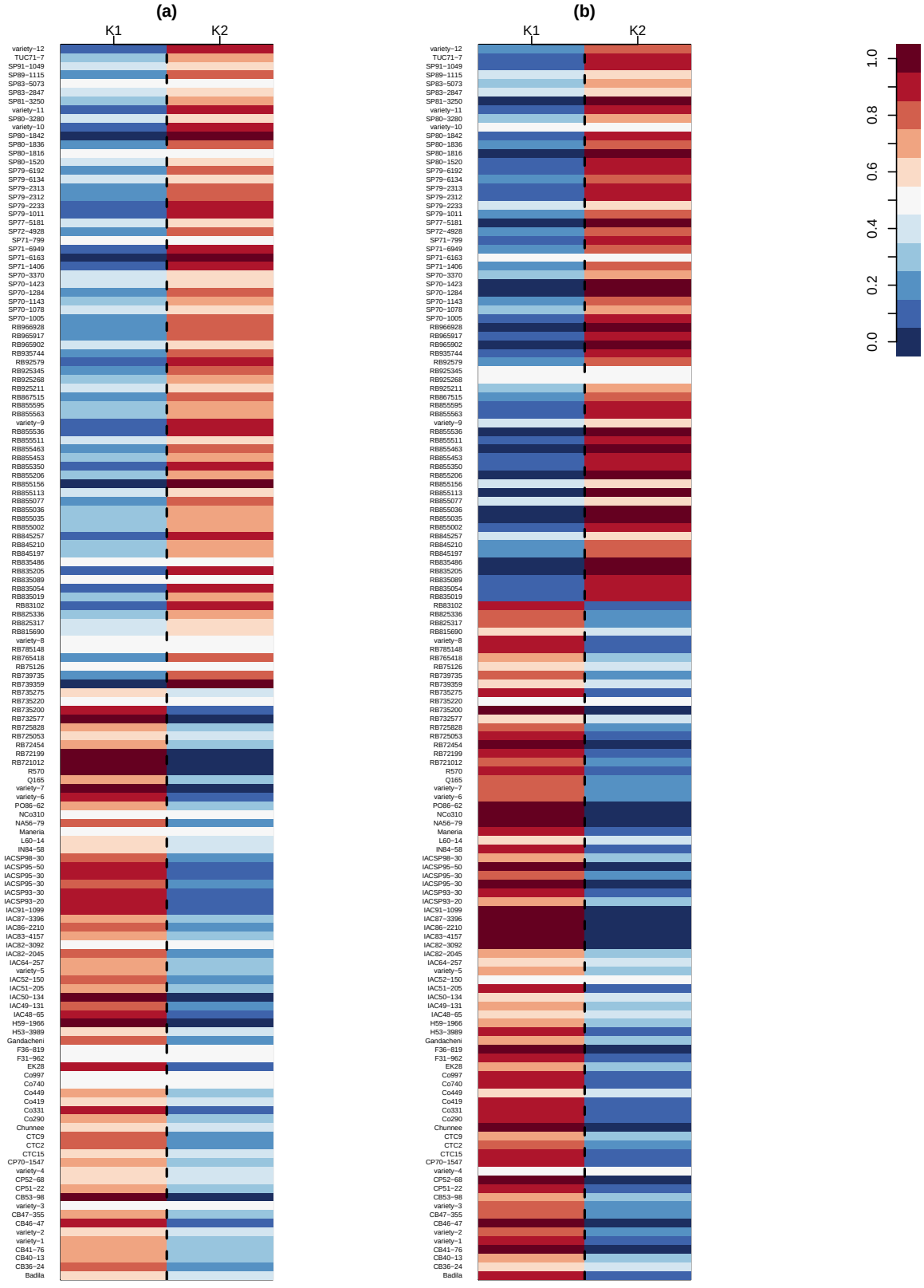
ZORIC, M.; DODIG, D.; KOBILJSKI, B.; QUARRIE, S.; BARNES, J. Population structure in a wheat core collection and genomic loci associated with yield under contrasting environments. **Genetica**, Dordrecht, v. 140, p. 259-275, 2012

APÊNDICES

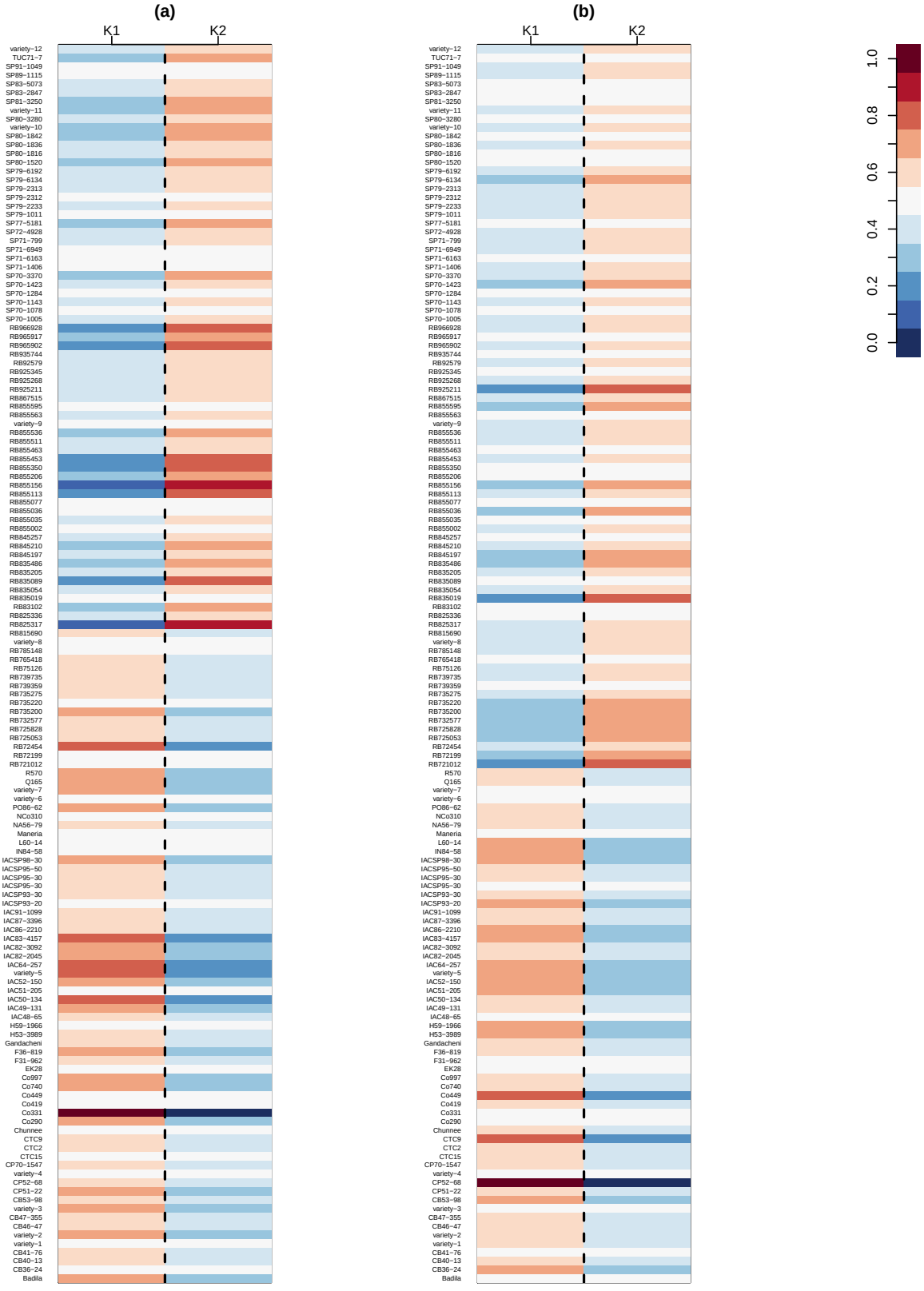
APÊNDICE A - Resultados das matrizes Q para os dados de SNPs com nível de ploidia 6 (a) e ploidia 8 (b)



APÊNDICE B - Resultados das matrizes Q para os dados de SNPs com nível de ploidia 10 (a) e ploidia 12 (b)



APÊNDICE C - Resultados das matrizes Q para os dados de SNPs com nível de ploidia 14 (a) e ploidia 16 (b)



APÊNDICE D - Resultados das matrizes Q para os dados de SNPs com nível de ploidia 18 (a) e ploidia 20 (b)