UFG

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

DESEQUILÍBRIO DE LIGAÇÃO E ANÁLISE DE SELEÇÃO GENÔMICA EM CANA-DE-AÇÚCAR

IVONE DE BEM OLIVEIRA

Orientador:
Prof. Dr. Alexandre Siqueira Guedes Coelho

Goiânia - GO
Brasil

Fevereiro – 2014

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

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei nº 9610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou *download*, a título de divulgação da produção científica brasileira, a partir desta data.

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

2. Identificação da Tese ou Dissertação

Nome completo do autor: Ivone de Bem Oliveira

Título do trabalho: Desequilíbrio de ligação e análise de Seleção Genômica em Cana-de-açúcar

3. Informações de acesso ao documento:

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

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

Ivone de Bem Oliveira
_____
Assinatura do (a) autor (a) [2]

Data: 21 / 02 / 2017

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. A extensão deste prazo suscita justificativa junto à coordenação do curso. Os dados do documento não serão disponibilizados durante o período de embargo.

[2] A assinatura deve ser escaneada.

IVONE DE BEM OLIVEIRA

DESEQUILÍBRIO DE LIGAÇÃO E ANÁLISE DE SELEÇÃO GENÔMICA EM CANA-DE-AÇÚCAR

Dissertação apresentada à Coordenação do Programa de Pós-Graduação em Genética e Melhoramento de Plantas, da Universidade Federal de Goiás, como requisito parcial à obtenção do título de Mestre em Genética e Melhoramento de Plantas.

Orientador: Prof. Dr. Alexandre Siqueira Guedes Coelho

Goiânia, GO - Brasil

2014

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

de Bem Oliveira, Ivone
Desequilíbrio de ligação e análise de Seleção Genômica em Cana-de açúcar [manuscrito] / Ivone de Bem Oliveira. - 2014.
LXXXI, 81 f.: il.

Orientador: Prof. Alexandre Siqueira Guedes Coelho.
Dissertação (Mestrado) - Universidade Federal de Goiás, Escola de Agronomia (EA), Programa de Pós-Graduação em Genética & Melhoramentos de Plantas , Goiânia, 2014.
Bibliografia.
Inclui gráfico, tabelas, algoritmos.

1. LD. 2. modelagem. 3. Melhoramento de Plantas. 4. Saccharum. I. Siqueira Guedes Coelho, Alexandre, orient. II. Título.

CDU 633

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

UFG

ATA DA DEFESA PÚBLICA DA DISSERTAÇÃO DE IVONE DE BEM OLIVEIRA. Aos vinte e sete dias do mês de Fevereiro do ano de dois mil e catorze (27.02.2014), às 14h00min, no Auditório PPGA da Escola de Agronomia e Engenharia de Alimentos, reuniram-se os componentes da Banca Examinadora, Prof. Dr. Alexandre Siqueira Guedes Coelho – Presidente/Orientador; Prof. Dr. Evandro Novaes e Drª. Tereza Cristina de Oliveira Borba. Sob a presidência do orientador, e em sessão pública, procedeu-se à avaliação da defesa de Dissertação intitulada: **"DESEQUILÍBRIO DE LIGAÇÃO E ANÁLISE DE SELEÇÃO GENÔMICA EM CANA-DE-AÇUCAR",** de autoria de **Ivone de Bem Oliveira**, discente do Programa de Pós-Graduação em Genética e Melhoramento de Plantas, no nível de Mestrado, da Universidade Federal de Goiás. A sessão foi aberta pelo presidente da Banca Examinadora, Prof. Dr. Alexandre Siqueira Guedes Coelho, que fez a apresentação formal dos membros da Banca. A palavra, a seguir, foi concedida ao autor da Dissertação que, em 40 minutos, apresentou o seu trabalho. Terminada a apresentação, cada membro da Banca argüiu o mestrando, tendo-se adotado o sistema de diálogo seqüencial. Ao final, a banca reunida em separado procedeu à avaliação da defesa. O título da dissertação foi alterado para "_____
_____
_____".

De acordo com a Resolução nº 1053/2011, do CEPEC - Conselho de Ensino, Pesquisa, Extensão e Cultura, que regulamenta o Programa de Pós-Graduação em Genética e Melhoramento de Plantas, e desde que procedidas às correções recomendadas, a Dissertação será considerada Aprovada pela Banca Examinadora, estando integralmente cumprido este requisito para fins de obtenção do título de MESTRE EM GENÉTICA E MELHORAMENTO DE PLANTAS, pela Universidade Federal de Goiás. O mestrando deverá efetuar as modificações eventualmente sugeridas pela Banca Examinadora e encaminhar a versão definitiva da Dissertação à Secretaria do PGMP, no prazo máximo de trinta dias após a data da Defesa. A conclusão do Curso e a emissão do Diploma dar-se-ão após o cumprimento do Artigo 52 da Resolução CEPEC nº 1053/2011. A Banca Examinadora recomenda a publicação de artigo(s) científico(s), oriundo(s) dessa Dissertação, em periódicos de circulação nacional e, ou, internacional, depois de procedidas as modificações sugeridas. Cumpridas as formalidades de pauta, às 17:30. A presidência da mesa encerrou esta sessão de defesa de Dissertação e, para constar eu, Jéssica Almeida, secretária PGMP, lavrei a presente Ata que depois de lida e aprovada, segue assinada pelos membros da Banca Examinadora, em duas vias de igual teor.

Alexandre Siqueira Guedes Coelho
Prof. Dr. Alexandre Siqueira Guedes Coelho
Presidente/Orientador

Evandro Novaes
Prof. Dr.Evandro Novaes
Membro Interno

Tereza Borba
Drª. Tereza Cristina de Oliveira Borba
Membro Externo

DEDICO este trabalho à minha família por todo apoio, orientação e amor.

OFEREÇO este trabalho a Deus e aos meus pais, Magali e Junior, por serem meus exemplos de vida e de moral.

"Deus nos concede, a cada dia, uma página de vida nova no livro do tempo. Aquilo que colocarmos nela, corre por nossa conta" (Chico Xavier).

AGRADECIMENTOS

Primeiramente à Deus, pela oportunidade, pelos desafios e pela coragem e determinação que me conferiu.

Aos meus pais, Magali e Júnior, por estarem sempre presentes, mesmo à quilômetros de distância, pelas conversas, pelas broncas, pelos mimos, pela confiança irrestrita, por me auxiliarem durante a formação intelectual e moral, por serem meus exemplos de vida e de determinação. Agradeço à vocês pela vida!

Ao meu irmão Gabriel, pelo amor, pela amizade, pelas risadas, pelas danças, pelas caminhadas, pelas conversas, pelos conselhos, mas especialmente por me ensinar o quanto a vida é simples de ser vivida.

À madrinha Nídia, por sempre acreditar, por me amar tanto e por todo o auxílio e carinho dispendidos durante toda a minha vida.

À minha avó Dulce por ser um anjo na terra e por tornar o mundo mais doce e as dificuldades mais simples, devido à sua simples existência. Por todo o apoio e principalmente por todo o amor e por todas as ótimas lembranças.

À Teté (Onélia) por ser mais uma mãe, pelo carinho, pelo apoio, pelas conversas, por estar presente em todos os momentos e por ser uma das únicas pessoas que realmente me conhecem e me entendem.

À Ester, à Renata e à Hizumi, simplesmente por existirem e por deixarem as batalhas da vida um pouco mais fáceis de serem vividas.

À minha irmã Iziara, pelas conversas, pelas batalhas compartilhadas, pelo apoio, pela amizade, pelas comidas deliciosas, pelo carinho e pelo companheirismo.

À tia Cideli e às primas Tati e Jéssica, pelas alegrias, pela amizade e pelo amor.

À tia Andreza, tio Renaldo e ao primo Bruno pelo auxílio durante esta fase da vida.

Às amigas Paula, Patrícia, Adrienny, Isabela Pavanelli, Stela e Isabela Cerri Bertolino pelas alegrias compartilhadas e pelo apoio durante os momentos bons e também nos não tão bons.

Aos professores do Programa de Pós-graduação em Genética e Melhoramento de Plantas da EA-UFG, pelos conhecimentos compartilhados.

Ao Professor Dr. Alexandre Siqueira Guedes Coelho, pela orientação, pelo auxílio diário e pela compreensão. Agradeço também por me aceitar novamente como orientada e por servir como exemplo na profissão que pretendo seguir.

À orientadora da graduação Profª Drª Rosete Pescador, além do incentivo e do encorajamento, agradeço pela amizade, pelos conselhos e pelo estímulo intelectual e moral.

Aos orientadores Prof. Dr. Marcio Elias Ferreira e Profª Drª Glaucia Salles Cortopassi Buso, que me permitiram chegar à UFG, agradeço pela oportunidade e pelo salto no aprendizado que me proporcionaram.

A todos os colegas do Laboratório de Genética e Genômica de Plantas da EA-UFG: Ludmila, Milena, Arthur, Clistiane, Bianca, Dienny e demais estagiários e colegas da Escola de Agronomia da UFG, muito obrigada pela ajuda e por compartilharem desta batalha.

À Escola de Agronomia da Universidade Federal de Goiás, pela oportunidade e incentivo e ao Programa de Pós-Graduação em Genética e Melhoramento de Plantas, por todo auxílio oferecido.

Ao Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES, pela concessão da Bolsa de Estudos.

À Rede Interuniversitária para o Desenvolvimento do Setor Sucroenergético (Ridesa), pela concessão de recursos para realização dos experimentos.

À Usina Centroálcool, por disponibilizar a área para a condução do experimento e mão-de-obra para auxílio durante o plantio e as avaliações fenotípicas.

SUMÁRIO

RESUMO

DE BEM, I. O. **DESEQUILÍBRIO DE LIGAÇÃO E ANÁLISE DE SELEÇÃO GENÔMICA EM CANA-DE-AÇÚCAR (*Saccharum* spp.)**. 2014. 85f. Dissertação (Mestrado em Genética e Melhoramento de Plantas) - Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2014.[1]

O desequilíbrio de ligação (LD) é um fenômeno genético que interfere diretamente na dinâmica genética das populações. Seu efeito é observado na segregação não independente dos alelos dos diferentes locos, resultando na correlação entre eles durante a formação dos haplótipos. Qualquer fator que altere as frequências alélicas interfere no LD. Assim, tanto os fatores evolutivos considerados em genética populacional, como a maioria dos eventos utilizados no melhoramento interferem na dinâmica equilíbrio/desequilíbrio de ligação. Como os fatores microevolutivos são específicos para cada população (natural ou de melhoramento), o LD deve ser avaliado em cada população, de maneira específica. Medidas de LD são desenvolvidas considerando os diferentes fatores evolutivos, afim de diminuir o viés gerado por eles durante sua mensuração. Com o intuito de mensurar o LD em populações de melhoramento da Ridesa foi implementado, no ambiente R, o modelo descrito por Raboin et al. (2008), que considera como fatores controlados para a diminuição de viés na mensuração do LD a polissomia e a poliploidia, características inerentes ao genoma da cana. Com o uso da modelagem implementada no ambiente R foram obtidos o perfil de decaimento para duas populações de melhoramento da Ridesa, a primeira compostas por 91 indivíduos provenientes do cruzamento biparental entre os clones-elite RB97327 e RB72454 e a segunda por 81 clones gerados por autofecundação do clone RB97327, a partir da análise de 850 e 470 locos DArT para cada população, respectivamente. As populações estudadas apresentaram alto LD, mostrando a existência de desequilíbrio mesmo entre locos separados por 30cM. Com base nos valores encontrados e nos estudos sobre LD existentes para cana na bibliografia foram discutidos os possíveis efeito dos processos de melhoramento sobre a dinâmica genética da cultura. Um segundo estudo foi realizado associando o padrão de segregação mendeliana dos locos e seu efeito no padrão de decaimento do desequilíbrio. Esta análise mostrou que os perfis de decaimento do desequilíbrio ao longo de distâncias crescentes entre locos pode estar associado à dosagem dos alelos sob análise. Além disso, afim de ilustrar um dos usos do LD no melhoramento de cana foi desenvolvido um estudo de seleção genômica ampla

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

(GWS) com o uso da metodologia RR-BLUP. Esta análise foi realizada como uma prova de conceito, com o intuito de verificar a viabilidade do desenvolvimento do método de seleção genômica ampla para cana-de-açúcar. Para isto, os 132 genótipos foram caracterizados para seis diferentes caracteres, sendo eles: peso do feixe de colmos (Kg), diâmetro médio do colmo (mm), comprimento médio de colmo (m), concentração de sólidos solúveis (^{o}Brix), número médio de internódios e número de colmos/touceira. Além de evidenciar o grande potencial de desenvolvimento de modelos desta natureza utilizando marcadores DArT em cana-de-açúcar, os resultados obtidos sugeriram a existência de forte efeito de estruturação intrapopulacional e existência de associações espúrias nas acurácias obtidas nos modelos. Mostrando que os efeitos desta estruturação, normalmente negligenciado durante as análises de calibração de modelos de seleção genômica ampla, devem ser melhor investigados em estudos posteriores.

Palavras-chave: LD, modelagem, *Saccharum*.

ABSTRACT

DE BEM, I. O. **LINKAGE DISEQUILIBRIUM AND GENOME-WIDE SELECTION ANALYSIS IN SUGARCANE (*Saccharum* spp.)**. 2014. 80f. Dissertation (Master in Genetic and Plant Breeding) - Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2014[2].

Linkage disequilibrium (LD) is a genetic phenomenon, since it directly interferes in the genetic dynamics of populations. Its effect is observed in the non independent segregation of alleles of different loci, resulting in a correlation between them during the haplotype formation. Any factor that alters allele frequencies can interfere in its dynamics. Both evolutionary factors considered in population genetics and the events used in breeding programs affect the linkage equilibrium/disequilibrium balance. As the microevolutionary factors are specific for each population (natural or from breeding), the LD should be assessed specifically in each population. Specific measures of LD were developed considering the different factors involved in order to mitigate their bias during measurement. Aiming to measure the LD in Ridesa breeding populations a model to predict the LD in polyploids was implemented, using the R platform, based on the Raboin et al. (2008) equations, that considers the sugarcane genome inherent attributes (polysomy and polyploidy). By using this model the LD decay profile was obtained for two breeding populations from Ridesa, the first one composed by 91 individuals from the cross between the RB97327 and RB72454 elite clones and the second from the 81 individuals derived from the selfing of the RB97327 clone. A total of 850 and 470 loci of DArT markers, respectively, were evaluated for each population. The populations showed a high LD, suggesting the existence of linkage disequilibrium even between loci that are 30cM apart. Based on the values found and on other sugarcane LD studies found in literature, the possible effect of the processes of genetic improvement on the dynamics of sugarcane LD were discussed. A second study was carried out associating the mendelian segregation analysis of loci and their effect on the LD pattern, that showed that the LD profiles over increasing distances between loci can be associated to the allelic dosage. Furthermore, aiming to illustrate one of the uses of LD in breeding programs a genome-wide selection (GWS) study was developed using RR-BLUP methodology. This study was carried out as

[2] Supervisor: Prof. Dr. Alexandre Siqueira Guedes Coelho. EA – UFG.

a proof of concept, in order to study the feasibility of the method of genome-wide selection in sugarcane. For this, the 132 genotypes were characterized for six different characters, namely: stalk weight (kg), stalk diameter (mm), stalk length (m), soluble solids concentration (Brix), internodes number and stalk number per plot. In addition to demonstrate the great potential of these studies in sugarcane using DArT, the results suggest the existence of a strong effect of intrapopulation structure and of the spurious associations between the loci in the accuracy of the model. These factors, usually ignored in this kind of analysis, should be further investigated in future studies.

Keywords: LD, model, *Saccharum*.

1 INTRODUÇÃO

A cana-de-açúcar é uma das principais culturas na economia mundial, sendo o Brasil o maior produtor e, consequentemente, o maior produtor de açúcar e álcool produzidos a partir do seu beneficiamento. Devido a esta condição, apresenta grande importância no mercado econômico mundial, visto que é responsável por mais de 60% das exportações (Brasil, 2012). Além disso, o setor sucroalcoleiro brasileiro está em plena expansão. As áreas em produção continuam aumentando e as perspectivas para as próximas safras são consideradas muito boas (CONAB, 2013).

A importância econômica desta cultura no país e no mundo exige dos programas de melhoramento a geração rápida, eficaz e barata de novas variedades. Neste contexto, as novas tecnologias envolvendo obtenção, análise e a interpretação de dados genéticos são ferramentas com grande potencial no auxílio aos melhoristas para a produção de novas variedades, pois aceleram, facilitam e diminuem os custos inerentes aos programas de melhoramento genético, eliminando entraves e expandindo limites, gerando grande quantidade de informações para o desenvolvimento de estudos voltados para a análise genética de populações (Faleiro, 2007; Ferreira & Grattapaglia, 1998).

O fenômeno genético intitulado desequilíbrio de ligação interfere na maneira com que os fatores evolutivos exercem seus efeitos na estrutura genética das populações, atuando na alocação dos genes durante a formação dos gametas (Hedrick, 2011). Diversos trabalhos, visando maior eficiência dos métodos de seleção em programas de melhoramento, nas mais diversas culturas, têm investigado este fenômeno (Jannoo et al., 1999b; Carneiro & Vieira, 2002; Resende, 2008; Raboin et al., 2008; Lopes, 2011). É consenso que a inexistência do desequilíbrio de ligação entre um marcador e um loco de interesse diminui a eficácia de seleção das técnicas baseadas em marcadores (Durães et al., 2004; Resende, 2008; Resende et al., 2010; Rosa, 2011).

A existência de desequilíbrio de ligação permite a inferência do comportamento de um loco baseado no comportamento de um segundo (Thoday, 1961). Este é o princípio de todas as metodologias de seleção baseadas em marcadores, entre elas a Seleção Genômica Ampla (GWS), que é uma das propostas mais recentes de utilização

de marcadores moleculares para auxilio da seleção em programas de melhoramento genético. Este processo se fundamenta na seleção de grande quantidade de marcas distribuídas no genoma como um todo, a fim de se capturar toda, ou pelo menos a maior parte, da variação genética responsável pela variabilidade fenotípica de caracteres de interesse agronômico. Para que isso seja possível, existe a premissa da existência de desequilíbrio de ligação entre parte das marcas utilizadas e os polimorfismos causais destas variações (Meuwissen et al., 2001; Zhang et al., 2011). Neste sentido, a GWS permite a inferência, com determinada acurácia, do efeito de vários locos responsáveis pela variação fenotípica de determinado caractere, com base no comportamento de outros locos. Trata-se portanto de uma expansão dos fundamentos teorizados por Thoday (1961).

Assim, tanto na Seleção Genômica Ampla, como nos demais métodos baseados em seleção assistida por marcadores, a existência de desequilíbrio de ligação entre marcadores genético-moleculares e os caracteres de importância agronômica interfere e condiciona a eficiência da seleção. Este fato justifica o desenvolvimento de estudos para o conhecimento do comportamento e da extensão do desequilíbrio de ligação em populações de melhoramento, visando o estabelecimento de estratégias mais eficazes de melhoramento genético.

É importante, em estudos de caracterização da magnitude e extensão do desequilíbrio de ligação, ressaltar o fato de que os parâmetros estimados são específicos de cada população estudada. Na interpretação do desequilíbrio de ligação encontrado, devem ser consideradas todas suas possíveis causas que, por sua vez, são específicas e influenciam os mecanismos envolvidos na geração, na permanência e no decaimento deste nas populações (Slatkin, 1994; Schaper et al., 2012). Acentuando-se assim a importância de estudos específicos do desequilíbrio de ligação em cada população de melhoramento.

Em cana-de-açúcar, muitos entraves são encontrados para a correta mensuração do desequilíbrio de ligação, como por exemplo o alto nível de ploidia, a origem interespecífica dos cromossomos (Bremer, 1961), a ocorrência frequente de aneuploidias e a grande extensão do genoma (Cuadrado et al., 2004; D'Hont, 2005; Piperidis et al., 2010; Zhang et al., 2012). A implementação de novas tecnologias de sequenciamento e o desenvolvimento de ferramentas computacionais avançadas têm contribuído para a solução deste problema, viabilizando a realização de estudos genéticos mais abrangentes e de maior precisão (Raboin et al., 2008; Ferreira & Grattapaglia, 1998).

Neste contexto, com o intuito de se realizar a caracterização de alta resolução do desequilíbrio de ligação em uma população-base de programa de melhoramento genético de cana-de-açúcar, objetivou-se: (1) Avaliar a magnitude e a extensão do desequilíbrio de ligação, em escala genômica, presente em uma população-base de um programa de melhoramento de cana-de-açúcar; (2) Pelo uso de modelagem, avaliar diferentes hipóteses para se explicar a magnitude do desequilíbrio de ligação encontrado na população em estudo e (3) Avaliar o potencial de utilização da estratégia de seleção genômica ampla em uma população de melhoramento genético de cana-de-açúcar, exemplificando um dos usos do LD no processo de melhoramento.

2 REVISÃO DE LITERATURA

2.1 A CANA-DE-AÇÚCAR

2.1.1. Aspectos taxonômicos

As plantas designadas coloquialmente como cana-de-açúcar pertencem ao gênero *Saccharum*. As espécies deste gênero, segundo Tzvelev (1989) podem ser classificadas sistematicamente como pertencentes à família *Poaceae*, subfamília *Panicoideae,* tribo *Andropogoneae*, subtribo *Saccharinae* e obviamente ao gênero *Saccharum.* Este gênero inclui seis espécies: *Saccharum officinarum* L., *Saccharum spontaneum* L., *Saccharum barberi* Jeswiet, *Saccharum sinense* Roxb., *Saccharum robustum* Brandes & Jeswiet ex Grassl e *Saccharum edule* Hassk. (Bremer, 1961; Daniels & Roach 1987; Jannoo et al., 1999a).

Devido à rara ocorrência de fósseis de plantas, toda a história da cana-de-açúcar é contada a partir do que ainda existe de diversidade dessa cultura (Grivet et al., 2006). Assim, vários locais do globo são classificados como centro de origem para as diferentes espécies de *Saccharum*, visto que podem ser encontradas formas selvagens desde o norte da África, Oriente Médio, Índia, Malásia, Thailândia, China, Indonésia, Japão, Filipinas, Polinésia e Nova Guiné (Bremer, 1961).

Estudos filogenéticos apontam que apenas três das seis espécies componentes do gênero podem realmente ser consideradas espécies filogenéticas: *S. officinarum*, *S. spontaneum* e *S. robustum*. As demais são híbridos naturais gerados, principalmente, por cruzamentos envolvendo *S. officinarum* e *S. spontaneum.* Desta maneira, *S. sinense*. *S. barberi* e *S. edule* não podem ser tratadas como *taxa* distintos, visto que a distância genética com as outras espécies não o permite. Por serem importantes grupos horticulturais na Índia e na China, no entanto, acabam sendo tratadas como espécies (Price, 1968; Bremer, 1961; D'Hont et al., 1996; Grivet et al., 2006).

Sobre estas "pseudo-espécies" pode-se dizer que *S. sinense* e *S. barberi* possuem como centro de origem Índia e China, respectivamente, e diferem de *S.*

officinarum na quantidade de açúcar presente nos caules, além da inflorescência, da maior quantidade de fibras no colmo e da maior adaptabilidade, principalmente a regiões subtropicais. Mesmo possuindo menor quantidade de açúcar, estas espécies também foram utilizadas para a extração deste componente, principalmente na Índia e na China (Daniels & Daniels, 1975; Grivet et al., 2006). Sobre *S. edule* é sabido que por conter baixa quantidade de açúcar é mais utilizada em Fiji e Nova Guiné para subsistência, principalmente no uso das inflorescências abortadas para alimentação (D'Hont, 2005; Chandran, 2011), mas sua origem continua obscura (Piperidis et al., 2010).

S. officinarum, possivelmente a primeira espécie domesticada desse gênero (Grivet et al., 2006), é chamada de cana nobre devido à alta produtividade de açúcar, sendo, porém muito suscetível a doenças. Seu centro de origem é apontado como Nova Guiné que também é apontada como seu centro de diversidade (D'Hont et al., 1996; Jannoo et al., 1999a).

Das seis espécies do gênero *Saccharum* duas são consideradas selvagens, a primeira *S. spontaneum* que possui colmos finos, firmes e fortes e apresenta grande variabilidade (Bremer, 1961; D'Hont et al., 1996). Essa espécie é considerada a mais polimórfica e adaptável das espécies, se desenvolve tanto nos trópicos como nas regiões subtropicais e tem como centro de origem a Índia (Daniels et al., 1975 in D'Hont, 1998). A segunda espécie considerada selvagem é *S. robustum*, que apresenta colmos com maior diâmetro que as demais espécies, além de serem grandes e fortes e com baixa quantidade de açúcar. Devido a estes fatores é utilizada por fazendeiros para fabricação de cercas (Grivet et al., 2006). Esta espécie pode ser encontrada desde Bornéu até Nova Guiné e New Hebrides (Price, 1965) e é considerada o progenitor selvagem de *S. officinarum* (Bremer, 1961). Como as demais espécies têm relação com *S. officinarum* e a mesma possui relação próxima com *S. robustum, S. spontaneum* é considerada, então, a espécie geneticamente mais distante do gênero (Daniels & Roach, 1987).

2.1.2. Aspectos citogenéticos

Saccharum é considerado um dos gêneros mais complexos entre as plantas cultivadas, possuindo genoma extenso e grande número de cromossomos (D'Hont et al., 1994; D'Hont et al., 1996; Piperidis et al., 2010; Zhang et al., 2012). A ocorrência de

múltiplas cópias de um mesmo loco em todos os cromossomos dão apoio à hipótese de autopoliploidia (Grivet et al., 1996), mas a espécie continua sem um parente diploide conhecido (D'Hont et al., 1994; Zhang et al., 2012). O sorgo (*Sorghum* sp.) é apontado como um modelo diploide adequado para o gênero por dois motivos, entre as gramíneas *Saccharum* spp. possui maior sintenia com essa espécie e, além desta grande colinearidade entre as duas espécies, há pouco tempo de divergência entre elas (Grivet & Arruda, 2002).

Muitos estudos sobre a citogenética do gênero *Saccharum* foram realizados. Diferentes técnicas foram consideradas para esclarecer e gerar informações sobre os genomas das espécies, disponibilizando informações básicas tanto para programas de melhoramento quanto para a conservação de recursos genéticos. Desta maneira chegou-se a conclusões sobre o número de cromossomos, sobre a ploidia, sobre o número básico de cromossomos de cada uma das espécies e também sobre o tamanho do genoma de algumas delas.

Assim, sobre as duas principais espécies do gênero, sabe-se que a maioria dos clones de *S. officinarum* apresentam 2n=80 cromossomos, enquanto os de *S. spontaneum* variam entre 2n=48 e 2n=128. Uma das suposições para o diferente número de cromossomos de *S. spontaneum* é o cruzamento entre duas formas estáveis desta espécie em seu centro de origem, gerando diferentes padrões de herança quanto ao número de cromossomos, além da presença de aneuploidias (Bremer, 1961).

O número básico de cromossomos (x) também difere entre estas espécies. Em *S. officinarum* x=10. Para *S. spontaneum* a determinação de x foi mais complexa. Muitas dúvidas existiam devido às diferenças no número de cromossomos encontrado na espécie. Estudos apontavam para x=6, 8 ou 10, mas x=8 foi proposto como o número básico mais provável após estudos com DNA *in situ hibridization* (Bremer, 1961; D'Hont et al., 1996; D'Hont et al., 1998).

Apesar das diferenças entre os genomas de *S. officinarum* e *S. spontaneum* os genomas destas espécies apresentam grande colinearidade (Grivet et al., 1996; D'Hont et al., 1998). Grivet et al. (1996) e Ming et al. (1998) discutem sobre o assunto e apontam que as diferenças entre os genomas destas espécies são resultado de simples fusões, translocações, ou eventos de fissão, o que corrobora a existência de colinearidade entre os genomas.

O tamanho dos genomas também apresenta certa diferenciação, visto que está diretamente correlacionado à quantidade de cromossomos na espécie (Zhang et al., 2012). Utilizando os dados de citometria de fluxo Zhang et al. (2012), pode verificar que o tamanho dos genomas de *S. officinarum* e de *S. spontaneum* diferem bastante, sendo de 7,5 até 8,6 Gb para a primeira e 3,4 a 12,6 Gb para *S. spontaneum*. Em média, o genoma básico (monoploide) assumido para a cultura compreende entre 760 a 985 Mpb, o que é duas vezes o genoma de arroz e similar ao genoma de sorgo (Menossi et al., 2008; Zhang et al., 2012).

2.1.3. O desenvolvimento das primeiras cultivares híbridas

Até o século XIX, *S. officinarum*, *S. sinense* e *S. barberi* representavam a maioria das cultivares comerciais no mundo (Jannoo et al., 1999a). Os primeiros cruzamentos artificiais com obtenção de sementes viáveis foram realizados em 1893 (Bremer, 1961). As primeiras hibridizações intencionais foram realizadas em Java, neste mesmo ano, na tentativa de se obter resistência às doenças que vinham surgindo.

Porém, sem o conhecimento dos estudiosos da época, os híbridos naturais já eram utilizados e mesmo o clone Kassoer, considerado como representante selvagem na época, utilizado para a primeira hibridização intencional de 1983 foi classificado posteriormente como um híbrido entre *S. spontaneum* e *S. officinarum* (após estudos citológicos realizados em 1921) e *S. spontaneum* foi considerada a selvagem "real" do gênero (D'Hont et al., 1996). Desta maneira, só em 1897 foi realizada a real hibridação artificial entre *S. officinarum* e *S. spontaneum* (Bremer, 1961).

Grandes ganhos em produtividade, resistência a doenças e adaptabilidade foram obtidos com as hibridações interespecíficas (D'Hont et al., 1996). Após a hibridação, sucessivos retrocruzamentos foram realizados com variedades de cana nobre para melhorar a produtividade, recuperando a produção de açúcar (Bremer, 1961; D'Hont et al., 1996). Poucos eventos meióticos foram realizados entre o primeiro cruzamento interespecífico e as cultivares existentes, até 1996 aproximadamente 6 meioses ocorreram (D'Hont et al., 1994; D'Hont et al., 1996) e calculam-se que apenas 10 eventos meióticos ocorreram desde a primeira hibridação até 2008 (Raboin et al., 2008).

As hibridações e retrocruzamentos realizados com *S. officinarum* foram

chamados eventos de nobilização, pois pretendia-se recuperar as características da cana nobre (*S. officinarum*). A primeira nobilização foi realizada no início do século 20, resultado dos híbridos F_1 obtidos entre *S. officinarum* e *S. spontaneum*. Os primeiros retrocruzamentos foram denominados de segunda nobilização, os segundos retrocruzamentos de terceira nobilização e assim por diante (Bremer, 1961; D'Hont et al., 1996). Como resultado dos processos de nobilização foram obtidas muitas cultivares comerciais e seus parentais (D'Hont et al., 1998), desde o primeiro processo de nobilização poucos parentais foram utilizados nos cruzamentos (Price, 1965), desta maneira as cultivares atuais descendem de poucos parentais e são fruto de poucos eventos meióticos.

É interessante citar que o sucesso das primeiras cultivares resultantes da hibridização foi em parte determinado pelo tipo de herança operante durante a formação do genoma dos híbridos. Um tipo diferente de herança se dá quando ocorre o cruzamento de *S. officinarum* e *S. spontaneum*. Não são herdados n + n = 2n cromossomos, e sim é verificada a presença de 2n + n cromossomos (Piperidis et al., 2010). Assim, 2x40 cromossomos de *S. officinarum* (gentior feminino) somados aos 1x56 de *S. spontaneum* (genitor masculino) resultam nos 136 cromossomos contidos nos híbridos (Bremer, 1961; D'Hont et al., 2004; D'Hont, 1996; Jannoo et al., 2004; Piperidis et al., 2010).

Uma das primeiras explicações para este fenômeno era a de não redução na meiose do parental *S. officinarum*, mas esta hipótese foi rejeitada, visto que as células reprodutivas apresentavam redução (Bremer, 1961). Posteriormente, verificou-se que a hipótese mais verossímil era a de endoduplicação ou fusão de dois núcleos após a segunda divisão meiótica, a herança n+n também pode acontecer, mas não é tão frequente (Bhat & Gill, 1985).

Como consequência da hibridação interespecífica, diferentes níveis de organização cromossômica coexistem nos híbridos. Devido à diferença entre os genomas que os compõem, os híbridos apresentam alto nível de ploidia e aneuploidias frequentes (Cuadrado et al., 2004). Com a utilização das técnicas de hibridização *in situ* de DNA foram identificados os grupos de co-segregação dos cromossomos, permitindo a classificação e a identificação de homólogos e homeólogos nos genomas dos híbridos (D'Hont et al, 1994; D'Hont et al., 1996). Com os avanços das técnicas genômicas foi possível verificar que entre 10 e 20% do genoma dos híbridos provém de *S. spontaneum*, de 70 a 80% são de herança materna, ou seja, de *S. officinarum*, e ainda, outros 10 a 20%

parecem ser provenientes de eventos de recombinação entre estes genomas (Figura 1) (Cuadrado et al., 2004; D'Hont, 2005; Piperidis et al., 2010).

Verificou-se também que a porção correspondente à herança de *S. spontaneum* é menos redundante e que, por isso, na maioria dos estudos desenvolvidos, a cobertura desta parte do genoma é maior (Grivet & Arruda, 2002). Além disso, a maior parte da variabilidade está contida na porção do genoma que corresponde à *S. spontaneum* (D'Hont et al., 2004), embora Jannoo et al. (1999a) tenham detectado que a porção referente à *S. officinarum* também confere certo grau de polimorfismo aos híbridos.

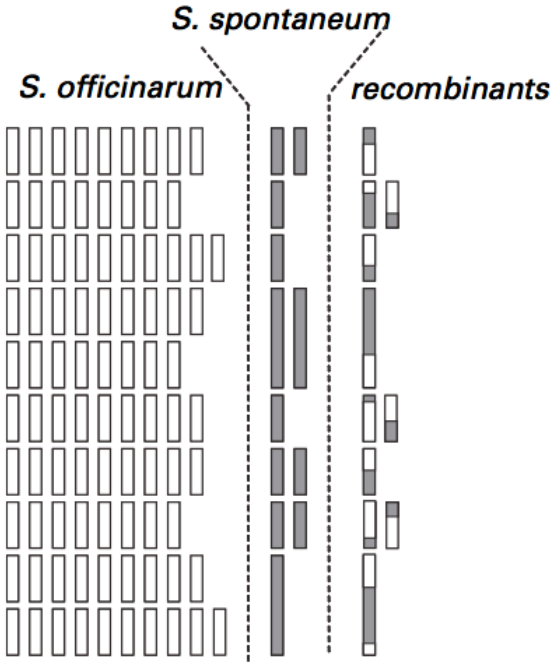

Figura 1. Figura representativa do genoma das cultivares híbridas, sendo cada cromossomo representado por uma barra e as colorações cinza e branca representam as frações do genoma dos cromossomo herdadas de *S. spontaneum* e *S. officinarum,* respectivamente. Fonte: D'Hont et al., 2005.

A coexistência dos dois genomas nas cultivares atuais não possui grande influência na meiose. Este fato é importante, visto que o pareamento cromossômico durante esta fase tem implicações extremamente importantes tanto na herança quanto na evolução das espécies (Jannoo et al., 2004). Em diferentes estudos, constatou-se uma meiose muito regular nos híbridos interespecíficos, com grande número de bivalentes observados na metáfase I. Estes estudos também verificaram que a presença de trivalentes ou de univalentes não induziu déficit na transmissão e que o pareamento espécie-específico

ocorre de maneira preferencial, mas não sistemática (D'Hont et al., 1996; D'Hont et al., 2005).

2.2 DESEQUILÍBRIO DE LIGAÇÃO

2.2.1. Considerações gerais

A partir dos estudos sobre a associação probabilística dos alelos nos diferentes locos dos indivíduos e seu comportamento nas populações, surgiu o conceito de equilíbrio/desequilíbrio de ligação. Estes estudos foram realizados a partir da averiguação dos desvios constatados entre as frequências gênicas que eram esperadas por simples acaso e aquelas observadas nas populações, quando considerados dois ou mais locos (Hartl & Clark, 2010).

Uma das definições de desequilíbrio de ligação mais utilizadas é a de Lewontin & Kojima (1960). Essa definição emerge da análise das frequências de pares de alelos durante a formação de gametas, por comparação entre as frequências observadas com aquelas esperadas sob independência, obtidas com base nas frequências de cada um dos alelos que compõem esses pares. Quando as frequências gaméticas observadas em determinada população são idênticas àquelas esperadas, obtidas pelo produto das frequências alélicas envolvidas, diz-se que os locos estão em equilíbrio de ligação (LE). Quando isto não ocorre têm-se a falta de equilíbrio de ligação, ou seja, o desequilíbrio de ligação (LD). Assim, quando há desvio nas frequências observadas dos haplótipos quando comparadas às esperadas tem-se o desequilíbrio de ligação (Lewontin & Kojima, 1960; Lewontin, 1964; Weir, 1979).

Por interferir diretamente na dinâmica genética das populações o LD deve ser entendido como um fenômeno genético. Seu efeito é visualizado na segregação não independente dos alelos dos diferentes locos (Hartl & Clark, 2010), resultando numa correlação entre eles durante a formação dos haplótipos (Flint-Garcia et al., 2003). Assim, sua existência acaba interferindo na formação dos genótipos dos indivíduos nas populações.

É importante ressaltar que qualquer fator que altere as frequências alélicas pode interferir na dinâmica do LD. Consequentemente, todos os fatores evolutivos

considerados nos estudos de genética populacional, bem como todas as ferramentas integrantes do processo de melhoramento genético que modifiquem estas frequências, atuam sobre esta dinâmica.

Assim, os fatores evolutivos, como o sistema de cruzamento, seleção, mutação, deriva, migração, ou fatores como coancestralidade podem ocasionar a associação não aleatória dos alelos nos indivíduos (Lewontin, 1988), gerando uma correlação entre eles na formação dos haplótipos, interferindo na dinâmica genética das populações e consequentemente na dinâmica equilíbrio-desequilíbrio de ligação.

Desta maneira, o estudo do desequilíbrio de ligação e a averiguação dos desvios em relação ao Equilíbrio de Hardy-Weinberg (EHW) pode mostrar o quanto e quais forças evolutivas estão atuando sobre determinada população. Sendo assim, o simples fato de assumirmos como falso um dos preceitos do EHW modifica as fórmulas utilizadas para mensuração do desequilíbrio de ligação. Para cada fator evolutivo são exigidas novas expressões de cálculo do LD. Por isso, em uma análise específica do desequilíbrio de ligação é de extrema importância a verificação cuidadosa de todos os fatores que nele interferem, pois pequenas alterações podem exigir mudanças importantes nos algoritmos utilizados.

2.2.2. Métodos de mensuração do LD

Existem diversas maneiras de aferir a magnitude do desequilíbrio de ligação. Aqui serão descritas as mais comumente utilizadas.

A medida básica de LD adotada pela maioria dos livros didáticos é chamada de *D* (1), foi proposta por Lewontin & Kojima em 1960 e consiste nas diferenças existentes entre os produtos das frequências dos haplótipos em associação e o produto das frequências dos haplótipos em repulsão (Lewontin & Kojima, 1960; Lewontin, 1964, Hedrick, 2011; Hartl & Clark, 2010).

$$D = f_{AB} \cdot f_{ab} - f_{aB} \cdot f_{Ab} \quad (1)$$

Por exemplo, quando são considerados dois locos bialélicos, *A* e *B*, com frequências alélicas *p* e *(1-p)*, para o primeiro loco e *r* e *(1-r)*, para o segundo loco, são

possíveis quatro diferentes haplótipos: *AB, Ab, aB* e *ab*, com frequências: f_{AB}, f_{aB}, f_{Ab}, e f_{ab}. Neste contexto, duas situações podem ser consideradas:

i. As frequências observadas dos diferentes haplótipos possíveis são condizentes com suas frequências esperadas sob independência, ou seja, estado de equilíbrio de ligação:

$$f_{AB} = p \,.\, r \qquad f_{aB} = (1\text{-}p) \,.\, r \qquad f_{Ab} = p \,.\, (1\text{-}r) \qquad f_{ab} = (1\text{-}p) \,.\, (1\text{-}r)$$

ii. As frequências observadas dos diferentes haplótipos possíveis diferem das suas frequências esperadas sob independência, ou seja, desequilíbrio de ligação (LD):

$$f_{AB} \neq p \,.\, r \qquad f_{aB} \neq (1\text{-}p) \,.\, r \qquad f_{Ab} \neq p \,.\, (1\text{-}r) \qquad f_{ab} \neq (1\text{-}p) \,.\, (1\text{-}r)$$

Na condição de LD (ii) são observados desvios entre as frequências esperadas e as observadas para os haplótipos.

$$f_{AB} = p \,.\, r + D \qquad f_{aB} = (1\text{-}p) \,.\, r - D \qquad f_{Ab} = p \,.\, (1\text{-}r) - D \qquad f_{ab} = (1\text{-}p) \,.\, (1\text{-}r) + D$$

Assim, quando os locos são considerados em equilíbrio de ligação têm-se $D = 0$ e quando os locos são considerados em desequilíbrio de ligação têm-se $D \neq 0$ (Figura 2).

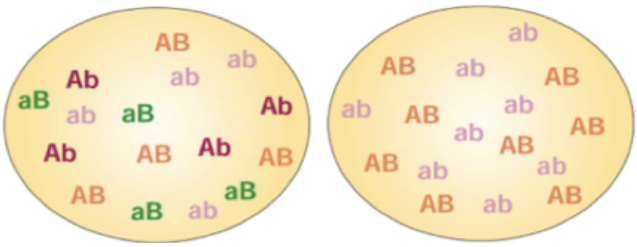

Figura 2. Esquema representativo dos haplótipos referentes a dois locos bialélicos (A e B). Na esquerda estão representados os quatro haplótipos esperados sob independência, equilíbrio de ligação. Na direita vê-se a formação preferencial dos haplótipos AB e ab, desequilíbrio de ligação. Fonte: Mackay (2001).

D é uma medida quantitativa da associação entre os alelos dos diferentes locos. Esta medida permite a inferência sobre a fase de ligação dos alelos nos haplótipos. Quando há excesso de haplótipos em associação observam-se desvios positivos (*D* positivo). O excesso de haplótipos em repulsão geram desvios negativos (*D* negativo) (Lewontin,

1964). Cuidado deve ser tomado devido a grande influência que as frequências alélicas têm sobre esta medida, visto que está diretamente associada a elas, refletindo assim, além da correlação entre os alelos dos diferentes locos o efeito das mudanças das frequências dos alelos que compõem os haplótipos (Lewontin, 1964; Hedrick, 1987; Gaut & Long, 2003).

Outra medida muito utilizada é denominada *D'* (2), ela reflete a razão entre o valor de *D* observado e o seu valor máximo possível (em módulo) para determinada frequência alélica. *D'* é uma medida normalizada (Lewontin, 1964; Hedrick, 1987; Zapata & Visedo, 1995) e esta normalização permite a avaliação da magnitude do desequilíbrio de ligação, sem considerar seu sinal, ou seja, é um valor relativo de *D* (Lewontin, 1964). A normalização também proporciona a diminuição da influência das frequências alélicas sobre o valor da medida de LD (Zapata, 2000). A magnitude dos valores de *D'* está sob forte influência da fração de recombinação entre os locos, refletindo seus efeitos (Gupta et al., 2005).

$$D' = \frac{D}{D_{máx}} \tag{2}$$

Em que $D_{máx}$ pode ser o menor valor obtido pelo produto das frequências dos alelos que compõem os haplótipos em repulsão ou o maior valor obtido pelo produto das frequências dos alelos que compõem os haplótipos em associação, com sinal negativo. O uso de um ou de outro valor está condicionado ao sinal encontrado na mensuração do *D,* se este for positivo usa-se o primeiro e se for negativo usa-se o segundo. Dessa forma o valor de *D'* não é influenciado pelos sinais do desequilíbrio calculado a partir da fórmula de *D* (Hartl & Clark, 2010).

O coeficiente de correlação *r* (3) é outra medida estandardizada do LD. Esta medida se baseia na correlação entre as frequências dos alelos de diferentes locos, podendo ser positiva ou negativa. Sua grandeza é fortemente afetada pela frequência dos alelos (Zapata & Visedo, 1995), sendo comumente expressa em termos quadráticos (r^2 - coeficiente de determinação). Esta parametrização (4) permite identificar apenas a magnitude do LD, sem considerar sua direção, assim como o *D'* (Waples & England, 2011).

$$r = \frac{D}{\sqrt{p\,.\,(1\text{-}p)\,.\,r\,.\,(1\text{-}r)}} \quad (3)$$

$$r^2 = \frac{D^2}{p\,.\,(1\text{-}p)\,.\,r\,.\,(1\text{-}r)} \quad (4)$$

A medida r^2 capta, além dos efeitos da fração de recombinação sobre o decaimento do LD, as histórias mutacionais ocorridas. Entre outras aplicações, ela permite o cálculo do tamanho efetivo populacional, parâmetro importante nos estudos genéticos populacionais (Gaut & Long, 2003; Gupta et al., 2005; Waples & England, 2011). A medida r^2 é uma das mais utilizadas para locos bialélicos (Mangin et al., 2011).

Outra alternativa para se avaliar a magnitude de desequilíbrio entre dois locos é a utilização dos testes de significância estatística. Segundo Slatkin (1994), a verificação estatística confirma a existência, ou não, de associação significativa entre alelos de diferentes locos, permitindo a inferência sobre o LD. Um dos testes mais utilizados é o Teste Exato de Fisher, que é altamente conservativo (Zapata & Alvarez, 1997) e quantifica a probabilidade de ocorrência de eventos tão extremos quanto os observados sob independência dos alelos de diferentes locos (dois-a-dois). Assim, quanto menor esta probabilidade, menos independentes os alelos se encontram e maior o LD (Gupta et al., 2005).

A melhor medida de LD deve ser escolhida considerando o objetivo do estudo a ser realizado, pois elas estão associadas a inferências com finalidades distintas. De modo geral, são levados em consideração na escolha da melhor medida de LD a interpretação biológica, a possibilidade de avaliação da significância estatística da correlação existente e a relação matemática entre as medidas obtidas e os fatores evolutivos (Hedrick, 1987; Gupta et al., 2005). Ainda que o objetivo da análise de LD seja a simples comparação entre locos ou populações, existem muitas controvérsias sobre qual medida é mais adequada para cada caso (Slatkin, 1994).

Grande cuidado deve ser tomado com a influência das frequências alélicas sobre as medidas obtidas. Ao se desconsiderar este fato, conclusões erradas podem ser sugeridas, principalmente na comparação entre amostras ou locos com diferentes frequências alélicas (Hedrick, 1987). Infelizmente não existe uma medida de desequilíbrio

totalmente independente das frequências alélicas (Lewontin, 1988), mas medidas estão sendo desenvolvidas para eliminar estes efeitos, bem como para eliminar a influência de associações espúrias, afim de obter medidas mais concretas de LD.

2.2.3. Desequilíbrio de ligação em espécies cultivadas

No início dos anos 90 os estudos de LD em animais já começavam a ganhar importância, mas poucos trabalhos tinham sido desenvolvidos em plantas. Os estudos disponíveis eram praticamente restritos às plantas-modelo, como milho e *Arabidopsis* (Flint-Garcia, 2003). Nos últimos anos, o estudo do desequilíbrio de ligação vem ganhando importância, principalmente devido à grande quantidade de dados genéticos à que as novas tecnologias dão acesso e devido à verificação da sua importância como fenômeno genético, principalmente em estudos microevolutivos e no desenvolvimento de ferramentas moleculares para avaliação genética.

Estudos tratando deste fenômeno genético e dos seus usos estão sendo desenvolvidos para as mais diferentes espécies. Em milho, eles mostram que o LD encontrado depende muito do tipo de população estudada, mas basicamente o LD possui baixa extensão, devido ao grande genoma e a alta taxa de recombinação, além da grande mobilidade existente no genoma desta espécie (pela ação de transposons e retrotransposons) (Flint-Garcia 2003; Gupta et al. 2005).

Assim, a extensão do LD em milho é tida como baixa, com rápido decaimento ao longo das distâncias físicas entre locos e com alcance variando, na maioria das populações, entre 0,5 e 7,0 kb (Tenaillon et al., 2001; Remington et al., 2001; Truntzler et al., 2012). Ching et al. (2002) relatam um alcance do LD de 100-500 kb em linhagens endogâmicas, mas estes valores não são comuns e resultam dos processos específicos de melhoramento a que esta população foi submetida.

Como o LD decai rapidamente com o aumento da distância física entre os locos no genoma do milho, para que sejam encontradas associações significativas entre locos marcadores e locos de interesse se faz necessário utilizar uma grande densidade de marcadores. Tal fato ilustra a importância do conhecimento da magnitude do LD, já que esta informação estabelece a quantidade de marcas necessárias para a realização de diversos estudos, incluindo os estudos de associação e de seleção assistida (Goldstein,

2001).

Em *Arabidopsis thaliana,* uma espécie considerada modelo do ponto de vista de análise genômica, o padrão do LD encontrado difere totalmente do padrão visualizado em milho, devido principalmente à alta taxa de autofecundação e à menor taxa de recombinação. Há na literatura diferentes estimativas do alcance do LD para esta espécie. Nordborg et al. (2002) mostram que o LD em *A. thaliana* tem alcance de cerca de 1 cM, o que corresponde à aproximadamente 250 kb. Já Nordborg et al. (2005) com uma amostra maior de indivíduos e maiores quantidades de marcadores, mostram que o LD é extensivo, como o esperado para espécies autógamas, com alcance entre 25 - 50 kb. Por outro lado, Kim et al. (2007), mostram que a extensão do LD pode ser menor do que se pensava, com estimativa de alcance de cerca de 10 kb.

É importante frisar que o LD pode diferir entre populações da mesma espécie. O desequilíbrio está condicionado a toda dinâmica demográfica, podendo, por exemplo, ser muito mais extenso em populações locais, como resultado da estruturação populacional gerada por eventos fundadores. Em *A. thaliana* este fenômeno é comum entre populações de diferentes regiões (Nordborg et al., 2002; Nordborg et al., 2005), justificando as diferenças observadas na mensuração. Outro fator que pode ter interferido nesta quantificação é a qualidade da informação genética utilizada, visto que as novas tecnologias disponibilizam informações em maior quantidade e qualidade.

Oryza sativa, outra espécie autógama, possui LD com ampla extensão. Esta espécie, no entanto, difere bastante de *A. thaliana* pelo seu histórico de domesticação e pelos processos de melhoramento genético, com ocorrência de inúmeros *bottlenecks*. Garris et al. (2003) observaram haplótipos de tamanho médio de mais de 100 kb. Resultados semelhantes também foram verificados por Agrama & Eizenga (2007). Em estudos mais recentes, Mather et al. (2007) compararam a extensão do LD nas diferentes espécies cultivadas de arroz com o LD encontrado no ancestral silvestre *Oryza rufipogon* que apresentou LD abrangendo regiões de 40 kb.

Outra informação relevante se deu na comparação entre o LD observado para as espécies cultivadas e a espécie selvagem. Sua extensão foi muito mais alta nas espécies cultivadas, sendo *O. japonica* de regiões temperadas a que apresentou maior LD (> 500 kb), seguida por *O. japonica* de regiões tropicais (aproximadamente 150 kb) e por *O. indica* (aproximadamente 75 kb), mostrando que a domesticação e os processos de

melhoramento interferiram fortemente no comportamento do LD.

Em cevada (*Hordeum vulgare*), também autógama e submetida a severos *bottlenecks* e processos de seleção, Zhang et al. (2009) encontraram rápido decaimento do LD após 3,5 cM. Comadran et al. (2011) verificaram que a maior quantidade de associações significativas se encontravam entre 5 e 10 cM. Os resultados obtidos por Stracke et al. (2003) sugerem um aumento deste limite para 20 cM. Resultados similares ao de Stracke et al. (2003) também foram obtidos por Krakmaan et al. (2004) e Zhou et al. (2012) para a cultura. Desta maneira percebe-se que o LD é extensivo no genoma desta espécie, mas que ainda encontram-se disparidades na sua mensuração.

Como inúmeros fatores alteram o LD, o estudo sobre este fenômeno deve ser específico para cada população, vistos que a extensão do LD depende fortemente da história demográfica da população sob estudo (Slatkin, 1994; Schaper et al., 2012) o que justifica as diferenças encontradas nas medidas de LD para as populações das diferentes espécies, visto que distintas populações foram utilizadas para estudo e cada uma delas possui características específicas.

Inúmeras outras espécies também estão sendo caracterizadas do ponto de vista do alcance do LD. Atualmente os estudos incluem desde espécies ornamentais, como orquídeas, até as mais diversas espécies cultivadas (Tabela 1).

Os estudos do desequilíbrio de ligação estão mudando com o desenvolvimento das novas tecnologias de sequenciamento e de interpretação de dados. Novos estudos mostram a diversidade do comportamento do desequilíbrio de ligação ao longo do genoma, e que o uso de um LD geral talvez seja pouco representativo. Hoje é sabido que o decaimento do LD é mais rápido em regiões propensas à recombinação (*hotspots*) (Jorde, 1995). Um exemplo desta diferença pode ser visto nos estudos de Würschum et al. (2013) em trigo, no qual foi verificada a disparidade de medidas de LD ao longo do genoma. A união de novas tecnologias e de novos métodos de mensuração deverá incrementar a compreensão deste fenômeno nos aspectos genético-evolutivos peculiares a cada uma das populações para as diferentes espécies.

Tabela 1. Exemplos de estudo do desequilíbrio de ligação em diferentes espécies vegetais. Fonte: Adaptado de Gupta et al. (2005).

Espécie	Sistema de Cruzamento	LD	Referências
Milho	alógama	0,5-7 kb	Remington et al. (2001); Tenaillon et al. (2001); Ching et al. (2002); Truntzler et al. (2012)
Arabidopsis	autógama	10-50 kb, podendo alcançar 250 kb	Nordborg et al. (2005); Kim et al. (2007); Nordborg et al. (2002)
Arroz	autógama	40-500 kb, dependendo da variedade	Garris et al. (2003); Mather et al. (2007)
Cevada	alógama	3,5-20 cM	Zhang et al. (2009); Comadran et al. (2011); Zhou et al. (2012)
Trigo	autógama	5-15 cM em variedades crioulas e 5-25 cM em cultivares modernas	Chao et al. (2010); Hao et al. (2011); Würschum et al. (2013)
Sorgo	misto (preferencialmente alógama)	10 -30 kb	Whang et al. (2013)
Soja	autógama	50-100 kb normalmente e até 600 kb em algumas cultivares	Zhu et al. (2003); Hyten et al. (2007); Li et al. (2008)
Batata	reprodução vegetativa	275 bp até mais de 5 cM	D'Hoop et al. (2010); Stich et al. (2013)
Phalaenopsis (orquídea)	alógama	80 cM	Gawenda et al. (2012)

2.2.4. Desequilíbrio de Ligação em Cana-de-Açúcar

Um caso interessante de análise do desequilíbrio de ligação é o da cana-de-açúcar (*Saccharum* spp.). Diversos fatores que compõem seu histórico de melhoramento geraram o estado de LD em que a cultura se encontra. Primeiramente é importante citar que o LD existente nas cultivares modernas foi considerado como resultado do forte efeito fundador, no qual poucos parentais de *S. officinarum* e *S. spontaneum* foram utilizados para compor os materiais atuais (Bremer, 1961; Jannoo et al., 1999b). Poucos eventos meióticos ocorreram desde as primeiras gerações de nobilização. O LD gerado durante o evento fundador persistiu, pois a quantidade de recombinações ocorridas não foi suficiente para dissipar o desequilíbrio por ele gerado (Raboin et al., 2008; Resende, 2008). Um dos fatores determinantes da reduzida quantidade de eventos meióticos a que a cultura foi submetida se dá pela forma de propagação da cultura. A facilidade de obtenção e de multiplicação clonal atenua a necessidade de cruzamentos e, consequentemente, de recombinações.

Desde então o histórico da cultura só reforça e existência de grandes blocos de haplótipos no genoma. Além do fator citado, a composição interespecífica do genoma (Cuadrado et al., 2004; D'Hont, 2005; Piperidis et al., 2010) e os sucessivos processos de retrocruzamentos com *S. officinarum*, favoreceram ainda mais a ocorrência de extensos blocos de ligação (Jannoo et al., 1999b; Raboin et al., 2008).

Assim, o histórico de melhoramento recente da cultura, com cruzamentos interespecíficos, uso de poucos parentais nos primeiros cruzamentos, além do consecutivo ciclo de retrogressão ao genoma de *S. officinarum* (Bremer, 1961) justificam o padrão de LD detectado, abrangendo blocos com grandes extensões, variando entre 10 e 30 cM (Jannoo et al., 1999b; Rosa, 2011; Lopes, 2011).

Muitos entraves são encontrados para o estudo deste fenômeno genético em cana-de-açúcar, entre eles pode-se citar a alta ploidia, a origem diferenciada dos cromossomos que compõem as gerações atuais (origem interespecífica), a constante presença de aneuploidias, além do genoma relativamente grande, com grande quantidade de cromossomos (>100). Devido a isso, poucos trabalhos sobre a mensuração e a compreensão da dinâmica do LD são encontrados para a cultura (Jannoo et al., 1999b; Lopes, 2011; Raboin et al., 2008; Rosa, 2011)

2.3 USOS DO DESEQUILÍBRIO DE LIGAÇÃO NO MELHORAMENTO DE ESPÉCIES CULTIVADAS

De modo geral, o desequilíbrio de ligação interfere grandemente no melhoramento genético, sendo fator determinante para a seleção com base em marcadores genéticos (Hartl & Clark, 2010; Resende et al., 2008; Resende et al., 2010). Auxiliando no desenvolvimento de estratégias, com vistas à redução de custos, contribuindo para o aumento da eficiência dos programas de melhoramento genético (Rosa, 2011), tanto para característica mono ou poligênicas. Ou seja, devido ao desequilíbrio de ligação, é possível prever o comportamento de um mono/poligene avaliando uma marca adjacente (monogene) e fisicamente ligada a ele (Thoday, 1961). Inúmeros trabalhos, visando maior eficiência dos métodos de seleção em programas de melhoramento, nas mais diversas culturas agronômicas, são embasados neste fenômeno (Carneiro & Vieira, 2002; Hedrick, 2011; Jannoo et al., 1999a; Lopes, 2011; Raboin, 2008; Resende, 2008).

Até muito recentemente, as seleções eram baseadas exclusivamente em observações fenotípicas. A metodologia BLUP (*Best Linear Unbiased Prediction*) é um dos principais métodos utilizados neste processo (Resende et al., 2008). Entretanto, o uso de dados moleculares nos processos de melhoramento envolvendo BLUP apresentaram grande eficácia (Fernando & Grossman, 1989). Mostrando que o uso da informação genética nos processos de seleção permitem a obtenção de um ganho genético maior do que os obtidos pela seleção baseada exclusivamente no fenótipo (Meuwissen et al., 2001; Resende et al., 2008).

Atualmente um dos métodos mais aclamados no âmbito do melhoramento genético é a Seleção Genômica Ampla. Jannink et al. (2010) e Heslot et al. (2012) a apontam como a grande revolução em curso no melhoramento genético, devido à alta acurácia, à diminuição dos custos do processo de seleção, à possibilidade de execução da seleção precoce direta e também por auxiliar na manutenção da variabilidade nos programas de melhoramento.

2.3.1. Seleção Genômica Ampla

Idealizado por Meuwissen et al. (2001) para aumentar a eficiência e acelerar o

processo de seleção nos programas de melhoramento, o método denominado Seleção Genômica Ampla (*Genome-Wide Selection* - GWS) não visa indicar locos individuais para a seleção e sim utilizar simultaneamente os efeitos de milhares de locos dispostos no genoma como um todo, de maneira a capturar seu efeito conjunto. O objetivo neste caso é explicar toda ou grande parte da variação genética causal de determinado caráter, ou seja, detectar a maior parte dos polimorfismos causais envolvidos na variação genética de um caráter quantitativo (Meuwissen et al., 2001; Resende et al., 2010; Endelman et al., 2011; Lorenz, 2013).

O processo admite a disponibilidade de grande quantidade de marcadores espalhados no genoma como um todo, de maneira que, devido à elevada densidade de marcadores, a probabilidade de se encontrar LD significativo entre eles e os locos que interferem na expressão do caráter quantitativo seja elevada (Meuwissen et al., 2001). Assim, na presença de LD, captura-se grande parte da variação genética responsável pela expressão fenotípica do caráter. A seleção baseada em marcadores neste caso atua no genoma como um todo, por isso é denominada ampla (Resende et al., 2008; Zhang et al., 2011).

Nos métodos clássicos a detecção de QTLs (*Quantitative Trait Loci*) é dificultada devido à natureza poligênica e à grande influência ambiental presente nos caracteres quantitativos. A estratégia tradicional de Seleção Baseada em Marcadores (SAM) perde informações sobre os polimorfismos causais, por se basear em um número pequeno de locos, perdendo informação de variação genética simplesmente pela impossibilidade de capturá-la. Por outro lado, a GWS permite a captura desta variação, sendo mais indicada para caracteres de natureza poligênica (Meuwissen et al., 2001; Resende et al. 2008).

A GWS faz uso tanto da fenotipagem dos caracteres quanto da genotipagem dos indivíduos e se baseia no desenvolvimento de um modelo de predição que, ao ser utilizado para a seleção, gera os valores genéticos de cada um dos indivíduos com base apenas no seu genótipo. Desta forma, pode-se fazer a predição e a seleção em fases muito precoces do programa, diminuindo o tempo de obtenção de materiais superiores (Meuwissen et al., 2001; Resende et al., 2010; De Los Campos et al., 2013).

É importante frisar que esse método relaciona os resultados das análises moleculares realizadas no genoma com as variações fenotípicas apresentadas em campo. O

modelo de GWS se baseia na detecção de associações entre estas informações, através do uso de ferramentas estatístico-computacionais (Rosa, 2011). Sendo assim, mesmo gerando modelos que exigem menor número de avaliações fenotípicas, a sua eficiência é extremamente dependente delas.

2.3.1.1. Populações utilizadas

Segundo Meuwissen et al. (2001), Goddard & Hayes (2007) e Resende et al. (2008), para a correta aplicação do método, são necessários três conjuntos de indivíduos ou populações. No primeiro conjunto, denominado população de descoberta, os indivíduos são fenotipados para diversos caracteres quantitativos e genotipados com grande quantidade de marcadores. Com estes dados é obtido o modelo matemático de predição dos valores genéticos genômicos (VGG) para cada um dos caracteres mensurados.

No segundo conjunto também são realizadas a genotipagem e a fenotipagem. O modelo anteriormente construído é aplicado a este segundo conjunto, afim de obter os VGG para cada indivíduo. Os valores obtidos com o uso do modelo são então comparados aos valores obtidos pela fenotipagem. Desta maneira, faz-se a validação e determina-se a acurácia do modelo construído. Esta população é denominada de população de validação.

O modelo validado, possuindo acurácia adequada, pode ser utilizado no terceiro conjunto de dados. Este conjunto é chamado população de seleção e nele somente os genótipos são obtidos. Com o uso do modelo validado são preditos os VGG e a seleção é realizada.

As equações de previsão da expressão fenotípica dos caracteres, obtidas por meio das populações de treinamento e de validação, são utilizadas para a obtenção de predições de fenótipos nas gerações subsequentes, sem que haja necessidade de novas análises fenotípicas nestas populações. Porém, no decorrer das gerações é indicada a realização de nova fenotipagem, em decorrência do decaimento do LD entre marcadores e QTLs. Por isso, com o passar das gerações, são necessárias recalibragens do modelo com novas fenotipagens (Resende et al., 2008).

2.3.1.2. Modelo

Inúmeros métodos para o cálculo do VGG têm sido sugeridos e, em geral, produzem resultados semelhantes (Zhang et al., 2011). Segundo Resende et al. (2010), o melhor método será sempre o que conseguir refletir de maneira mais acurada a natureza biológica do caráter quantitativo e sua natureza poligênica.

A necessidade de utilização de diferentes modelos surgiu da impossibilidade de estimação dos efeitos pelo falta de graus de liberdade, devido à grande quantidade de parâmetros (*p*) a serem calculados (efeitos de cada marca) e ao pequeno número de indivíduos avaliados na população (*n*), fenômeno conhecido como *large p and small n*. Por exemplo, pela inexistência de graus de liberdade para a estimação de todos os efeitos simultaneamente o método de quadrados mínimos se torna inadequado. Por outro lado, a estimação de cada um dos efeitos separadamente leva à superestimação da significância dos efeitos e gera baixa acurácia. Desta forma, boa parte da variação genética causal não é adequadamente capturada pelo método de quadrados mínimos tradicional (Resende et al., 2010; De Los Campos et al., 2013).

As metodologias de aplicação da GWS podem ser divididas entre paramétricas e não paramétricas. Entre os métodos paramétricos está a classe de regressão explícita. Nessa classe estão incluídos o BLUP (*Best Linear Unbiased Prediction* - Regressão Aleatória) e suas derivações, o LASSO (*Least Absolute Shrinkage and Selection Operator*) e os métodos Bayesianos, como o Bayes A e Bayes B, por exemplo, que diferem entre si no modelo genético assumido para a obtenção dos VGG (Resende et al., 2010; Jannink et al., 2010).

O método BLUP envolvendo regressão é chamado de regressão aleatória ou regressão de cumeeira (*Ridge Regression* - RR-BLUP). Esse método foi primeiramente proposto para uso na seleção baseada em marcadores por Whittaker et al. (2000) e foi um dos primeiros propostos para a seleção genômica (Whittaker et al., 2000; Endelman et al., 2011). Os preditores utilizados para o RR-BLUP são do tipo BLUP, mas são ajustados como covariáveis de efeitos aleatórios, ou seja, variáveis regressoras que geram efeitos de fenótipos (Resende et al., 2010; Jannink et al., 2010). Desta maneira são preditos os fenótipos a partir da regressão simultânea de grande quantidade de marcas espalhadas por todo o genoma (De Los Campos et al., 2013).

O método RR-BLUP utiliza um fator de regressão (*shrinkage*) para os efeitos aleatórios dos marcadores no modelo, denominado de λ, para a obtenção dos BLUPs. Este parâmetro atua como um fator de penalização. Como resultado de sua inserção na resolução do modelo têm-se o aumento do viés. Por outro lado, ocorre redução da variância do estimador, e como resultado final ocorre a melhora da acurácia do modelo (Resende et al., 2008).

No RR-BLUP o mesmo valor de λ é aplicado a todos os marcadores, o que o diferencia dos métodos Bayesianos, que utilizam valores específicos de λ para cada marca (Resende et al., 2010). Desta maneira assume-se a *priori* de que todas as marcas possuem o mesmo efeito para a explicação de determinado caráter. É interessante adicionar que a inserção de marcas não informativas não conduz a perdas significativas de acurácia durante o processo de seleção (Meuwissen et al., 2001).

Heslot et al. (2012) comparando onze diferentes modelos de predição dos VGG com o uso de seleção genômica observou que o RR-BLUP apresentou acurácia similar aos demais métodos de regressão, sendo ainda superior no tempo requerido na análise computacional. Tanto Jannink et al. (2010) como Lorenzana & Bernardo (2009) apontam que para cruzamentos biparentais o método é aconselhado, não apresentando menor acurácia que os métodos Bayesianos, além de ter fácil resolução e menor necessidade de marcas e indivíduos para a obtenção de alta acurácia na análise.

Um dos meios de aplicar a análise RR-BLUP aos dados se dá pelo uso de pacotes específicos que já estão disponíveis em algumas plataformas, como é o caso do rrBLUP, implementado na plataforma R (R Core Team, 2004), que permite a predição eficiente baseada em dados de treinamento não replicados, com o uso de um único componente de variância, além do erro residual, fazendo uso do algoritmo de máxima verossimilhança (*Restricted Maximum Likelihood* - REML) para a resolução de modelos mistos (Endelman et al., 2011).

O modelo matemático utilizado pelo método RR-BLUP (5) está bem descrito em Resende et al. (2008) e Resende et al. (2010). Nele assume-se que os valores preditos de *y* são obtidos com base no vetor de efeitos fixos (*b*), no vetor de efeitos aleatórios (*m*), que refere-se aos efeitos dos marcadores, e no vetor de resíduos aleatórios (*e*), sendo que os dois primeiros vetores são ponderados pelas matrizes de incidência *X* e *Z*, respectivamente. A matriz *Z* possui informações sobre as marcas e contém valores

relacionados à quantidade de alelos de referência em cada marcador. A matriz *X* é a matriz de incidência dos efeitos fixos.

$$y = Xb + Zm + e \tag{5}$$

2.3.1.3. Valores Genéticos Genômicos

No modelo de seleção genômica são obtidos valores genéticos para cada um dos marcadores. A soma destes valores genéticos para determinado caráter quantitativo é chamada de valor genético genômico. Para sua obtenção são utilizadas informações oriundas das fenotipagens e das genotipagens provenientes de um conjunto de indivíduos. Podem ser construídos modelos para um ou mais ambientes, dependendo do esquema utilizado para a obtenção do modelo estatístico (Resende et al., 2008).

Desta maneira, pelo somatório em *i* dos efeitos de cada um dos marcadores obtidos pela multiplicação entre a matriz de incidência *Z* e o vetor de efeitos de marcadores *m* são obtidos os VGG de cada indivíduo *j*, que se refere a cada valor fenotípico predito pelo modelo (6).

$$VGG = \hat{y}_j = \sum_i Z_i \hat{m}_i \tag{6}$$

2.3.1.4. Acurácia do modelo

Durante a validação do modelo é calculada a acurácia pela análise de correlação entre os VGG obtidos com o uso do modelo preditivo gerado na população de descoberta e os valores fenotípicos obtidos na fenotipagem da população de validação. Segundo Resende et al. (2010), como os conjuntos de dados utilizados na descoberta e na validação do modelo são independentes, toda a variação apresentada nos VGG também é independente. Assim, toda a correlação existente entre os dois valores são de cunho genético, equivalendo à própria acurácia.

Este mesmo trabalho mostra que a acurácia do modelo utilizado para a GWS depende de cinco fatores: da herdabilidade do caráter, da distribuição dos efeitos do QTLs nos locos do genoma, do número de indivíduos utilizados para a obtenção do modelo, do

tamanho efetivo da população e da distribuição das marcas ao longo do genoma.

A seleção e a avaliação do modelo preditivo da GWS pode ser realizada por procedimentos de validação cruzada. Nestes processos o mesmo modelo misto é analisado inúmeras vezes (Piepho et al., 2012). Uma das metodologias utilizadas para a validação cruzada em GWS é denominada *jack-knife*. Esta metodologia se baseia na divisão do conjunto original de (*N*) dados em (*g*) grupos de tamanho igual a *k*, de forma que $N = gk$, e na retirada de *n* observações em cada re-amostragem (Resende, 2002).

Trabalhos como o de Heslot et al. (2012) demostram a eficiência deste processo, no caso fazendo uso da validação cruzada do tipo *jack-knife 10-fold-validation*.Em que dividiram aleatoriamente o conjunto de dados dos indivíduos em 10 partes iguais e, posteriormente, para cada subgrupo obtido pela retirada de uma das partes recalcularam os VGG. Cada uma das partes remanescentes foi então utilizada para validar a acurácia do modelo.

2.3.1.5. Vantagens da Seleção Genômica Ampla

Segundo Resende et al. (2008), o método tem se mostrado eficiente e exequível, apresentando inúmeras vantagens. Entre elas pode-se citar (i) a redução do custo do processo, devido à seleção baseada apenas no genótipo; (ii) a plasticidade matemática dos modelos, que permitem que não só sejam preditos valores para populações de seleção em um ambiente, mas também capturar efeitos de interação genótipo x ambiente; (iii) a redução do efeito do ambiente sobre as estimativas, gerado pela repetição experimental devido a quantidade de indivíduos genotipados e fenotipados; (iv) a existência de diferentes metodologias; (v) com LD completo e ausência dos efeitos de epistasia e dominância, a herdabilidade é 1, ou seja, os efeitos permanecem constantes e a precisão tem acurácia máxima através das gerações em determinado ambiente; (vi) o método representa uma excelente alternativa para avaliação de caracteres de baixa herdabilidade e (vii) a GWS permite a execução da seleção direta precoce, que auxilia no ganho por unidade de tempo, sem a necessidade da espera pela idade ideal de análise de determinado caráter (Daetwyler et al., 2007; Resende et al., 2010).

Além das vantagens descritas, podem ser citadas o aumento da acurácia na seleção devido ao uso de uma matriz de parentesco real para cada caráter e o aumento da

acurácia ligado à repetição experimental dos alelos nos locos. Um dos pontos mais interessantes desta metodologia é que ela faz uso de toda a informação disponível, não só fenotípica mas também a genotípica e a genealógica (Resende et al., 2010).

3 MATERIAL E MÉTODOS

3.1 MATERIAL VEGETAL

A população utilizada no presente trabalho corresponde à um conjunto de clones de cana-de-açúcar (*S. officinarum*), composto por 172 indivíduos, incluindo 91 indivíduos resultantes do cruzamento entre os clones elite RB97327 e RB72454 e 81 indivíduos obtidos por autofecundação do clone RB97327, além de duas testemunhas experimentais (clones RB99395 e RB98710).

Os cruzamentos para obtenção destes clones foram realizados na Estação Experimental da Ridesa localizada na Serra do Ouro, em Murici (AL). As mudas, geradas a partir de sementes, foram primeiramente alocadas em casa de vegetação na Escola de Agronomia, da Universidade Federal de Goiás (UFG) e, posteriormente, transportadas e transplantadas em uma área experimental da Usina Centro-Álcool, localizada no município de Inhumas (GO) (16°20'50"S, 49°29'2"W). O experimento de campo foi realizado seguindo o delineamento de blocos aumentados de Federer (Federer, 1956), consistindo em nove linhas de 67,5 m subdivididas em 15 blocos cada. Os blocos foram constituídos por oito parcelas (touceira), das quais duas eram testemunhas (aleatoriamente distribuídas), presentes em todos os blocos.

3.2 OBTENÇÃO DO MATERIAL GENÉTICO E GENOTIPAGEM

O DNA genômico foi obtido com o uso do protocolo de Aljanabi et al. (1999) com adaptações. A genotipagem foi realizada em Yarralumla, Austrália, pela empresa *Diversity Arrays Tecnology Pty Ltd* (DArT P/L), seguindo a metodologia baseada na tecnologia de hibridização por microarranjos (DArT) proposta por Jaccoud et al. (2001).

Esta metodologia pode ser resumida nos seguintes passos: redução da complexidade genômica com o uso de enzimas de restrição, ligação de fragmentos adaptadores e posterior amplificação, isolamento e clonagem de cada um dos fragmentos

em vetores (insertos), amplificação dos insertos em base sólida (*spots*), obtenção dos painéis de diversidade (*slides*), escaneamento dos *slides* e visualização do padrão de intensidade de sinal de cada um dos *spots*. Esta técnica permite a obtenção de um padrão de sinais, cuja análise gera o padrão de diversidade para cada indivíduo (*fingerprinting*).

Para a obtenção dos DArTs foram utilizadas diferentes enzimas de restrição, afim de reduzir a complexidade genômica (scPth e scPtb). Todas as marcas foram nomeadas de acordo com a enzima utilizada. Como resultado da genotipagem, 7680 marcas passaram pelo controle de qualidade para discriminação de presença ou ausência da banda para os 172 genótipos, permitindo sua codificação em "1" e "0", respectivamente.

3.3 CARACTERIZAÇÃO FENOTÍPICA

Para a avaliação do método de Seleção Genômica Ampla, os 172 genótipos foram caracterizados para seis diferentes caracteres, sendo eles: peso do feixe de colmos (Kg), diâmetro médio do colmo (mm), comprimento médio de colmo (m), concentração de sólidos solúveis (^{o}Brix), número médio de internódios e número de colmos/touceira. Os caracteres foram avaliados tomando-se seis colmos do feixe completo, de maneira aleatória, exceto para peso e número de colmos/touceira, em que todos os colmos do feixe foram utilizados.

Para a obtenção de cada uma destas medidas foram utilizados, respectivamente, balança digital, paquímetro, realizando a medição no centro do entrenó localizado na porção mediana dos colmos, trena métrica, mensurando-se o comprimento a partir da base do colmo ao último entrenó, refratômetro analógico portátil, utilizando amostras de caldo obtidas a partir de corte na porção mediana do colmo e para os dois últimos contagem direta.

3.4 MODELAGEM DO DESEQUILÍBRIO DE LIGAÇÃO

Empregando como base as equações de previsão do comportamento de um haplótipo fundador em determinada população como descrito por Raboin et al. (2008), foi construído, com o uso de dados de simulação, um modelo de previsão do decaimento do desequilíbrio de ligação com o uso da plataforma R (R Core Team, 2004) e dos pacotes

nela implementados *ggplot2* (Wickham, 2009) e *resshape2* (Wickham, 2007). Este modelo foi ajustado utilizando a dinâmica observada nas duas populações de melhoramento sob estudo.

Assim, foi implementado o modelo teórico que permite a previsão da dinâmica do desequilíbrio de ligação em populações de melhoramento, controlando-se os seguintes fatores: taxa de recombinação por geração, frequências alélicas das marcas que compõem o haplótipo ao longo das gerações, na ausência de seleção e deriva, com combinação aleatória dos cromossomos (panmixia) e ausência de endogamia. Foram consideradas ainda duas características básicas do genoma de cana-de-açúcar, polissomia e alta poliploidia. Para utilização do modelo foram informados os seguintes parâmetros: a frequência de recombinação (r), a frequência inicial do haplótipo (gAB_0), o número de gerações transcorridas (t), a frequência do alelo que compõe o marcador (f) e que permanece constante ao longo das gerações, o nível de ploidia (n) e o número de indivíduos da população (*nind*).

Após a implementação do modelo, foram gerados gráficos de predição do comportamento de um haplótipo fundador no decorrer das gerações. Foram também estimadas a frequência dos indivíduos portadores dos diversos haplótipos e as frequências fenotípicas moleculares, pertinentes ao longo das gerações.

Para a construção dos gráficos com o uso do modelo previamente desenvolvido foram informados os mesmos parâmetros utilizados por Raboin et al. (2008): frequência de recombinação (r) variando entre 0,1 e 0,5, frequência inicial do haplótipo $gAA_0 = f$, número de gerações transcorridas (t) como 3 e 7, frequência do alelo que compõe o marcador (f) considerando 0,01, 0,05, 0,10 ou 0,20, ploidia n =10 e dois tamanhos populacionais, $nind_1$ = 72 e $nind_2$ = 200.

As equações de predição foram obtidas considerando-se "*A*" e "*B*" como presença do alelo e "*a*" e "*b*" como a ausência do alelo para cada marcador. Tomando-se como exemplo dois locos bialélicos, pode-se observar a formação dos seguintes haplótipos: *AB, Ab, aB* e *ab* e suas frequências haplotípicas *g(AB), g(Ab), g(aB)* e *g(ab),* respectivamente. Com base nesta notação foram implementadas as funções descritas no artigo de Raboin et al. (2008), como descrito a seguir.

A frequência da presença do alelo (f) na primeira (7) ou na segunda posição (8) do haplótipo foram calculadas da seguinte maneira:

$$g(A?) = g(AB) + g(Ab) = f \tag{7}$$

$$g(?B) = g(AB) + g(aB) = f \tag{8}$$

Considerando-se um haplótipo fundador *AB* a probabilidade de ele ter sido transferido para a geração seguinte se dá de duas maneiras: ou ele existia anteriormente na população e foi herdado sem que houvesse recombinação, ou ele ainda não existia e foi herdado por recombinação (Figura 3).

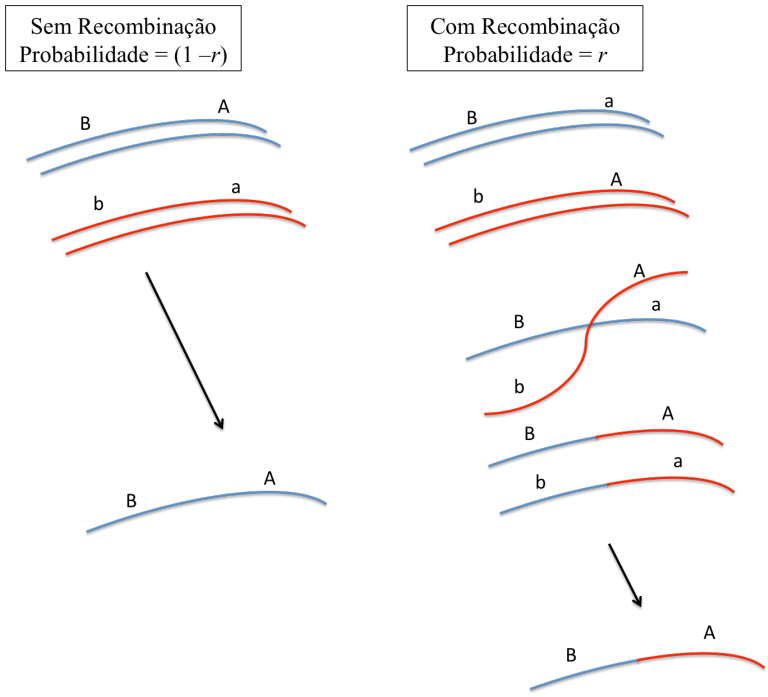

Figura 3. Ilustração dos dois eventos que podem gerar um haplótipo *AB* no transcorrer das gerações e suas probabilidades de acontecimento.

Nos dois casos, tanto a existência ou não do haplótipo, quanto a ocorrência ou não de recombinação, são eventos independentes, ou seja, a ocorrência de um evento não interfere na ocorrência do segundo evento (Sokal & Rohlf, 1981). A probabilidade de observarmos a ocorrência do haplótipo *AB* na geração *t* se dá então pela multiplicação das probabilidades associadas a cada um dos eventos. Ou seja, pela multiplicação da frequência do haplótipo *AB* pela probabilidade de não-recombinação, para a primeira

hipótese; ou pela multiplicação das frequências do alelo *A* na primeira posição e *B* na segunda posição pela probabilidade de ter ocorrido recombinação para a formação do haplótipo *AB*, para a segunda hipótese. Como a probabilidade de ocorrência de um evento pode ser considerada como a frequência deste evento, na estatística frequentista (9):

$$P(AB)_t = g(AB)_t \tag{9}$$

Assim:

$$g(AB)_t = g(AB)_t \,.\, (1\text{-}r) \;\; ou \;\; g(A?)_t \,.\, g(?B)_t \,.\, r$$

Considerando as equações de frequências descritas em (7) e (8):

$$g(AB)_t = g(AB)_t \,.\, (1\text{-}r) \;\; ou \;\; f^2 \,.\, r$$

Visto que as duas maneiras de obtenção do haplótipo *AA* na geração *t* são mutuamente exclusivas, ou seja, uma só pode ocorrer se a outra não tiver ocorrido (para uma mesma geração), obtém-se a seguinte fórmula:

$$g(AB)_t = g(AB)_t \,.\, (1\text{-}r) + f^2 \,.\, r$$

Como a fórmula de predição de comportamento do haplótipo fundador considera a frequência da geração anterior em seu cálculo, na geração t =2 o cálculo da geração 1 é utilizado, ou seja, ou o haplótipo *AB* existia na geração t =1 e foi herdado sem recombinação ou ele não existia na geração t = 1 e foi herdado por recombinação. O mesmo ocorre para as gerações t = 3 e para as demais gerações. Comparando as fórmulas para cada uma destas gerações é possível visualizar a fórmula de previsão da probabilidade do haplótipo *AB* no transcorrer das gerações como um somatório de termos de uma progressão geométrica (Figura 4).

Geração 1

$$g(AB)_1 = g(AB)_0 \,.\, (1\text{-}r) + f^2 \,.\, r$$

$$g(AB)_2 = g(AB)_1 \,.\, (1\text{-}r) + f^2 \,.\, r$$

Geração 2

$$g(AB)_2 = \left[g(AB)_0 \,.\, (1\text{-}r) + f^2 \,.\, r\right] \,.\, (1\text{-}r) + f^2 \,.\, r$$

$$g(AB)_2 = g(AB)_0 \,.\, (1\text{-}r)^2 + f^2 \,.\, r \,.\, (1\text{-}r) + f^2 \,.\, r$$

$$g(AB)_3 = g(AB)_2 \,.\, (1\text{-}r) + f^2 \,.\, r$$

Geração 3

$$g(AB)_3 = \left[g(AB)_0 \,.\, (1\text{-}r)^2 + f^2 \,.\, r \,.\, (1\text{-}r) + f^2 \,.\, r\right] \,.\, (1\text{-}r) + f^2 \,.\, r$$

$$g(AB)_3 = g(AB)_0 \,.\, (1\text{-}r)^3 + f^2 \,.\, r \,.\, (1\text{-}r)^2 + f^2 \,.\, r \,.\, (1\text{-}r) + f^2 \,.\, r$$

Figura 4. Exemplo da substituição ocorrida durante a obtenção das fórmulas de predição para a frequência do haplótipo *AB* ao longo das gerações, demonstrando o somatório de termos de uma progressão geométrica com índices *t*.

Desta maneira, a fórmula para a predição das frequências do haplótipo *AB* no passar das gerações foi generalizada (10). Para se obter a frequência genotípica do haplótipo *AB*, considerou-se que ele já exista na geração $t = 0$ e foi herdado sem recombinação, ou ele foi herdado pela ocorrência de recombinação no decorrer das gerações.

$$g(AB)_t = g(AB)_0 \,.\, (1\text{-}r)^t + f^2 \,.\, r \,.\, \left[\sum_{1}^{t} (1\text{-}r)^{t\text{-}1}\right] \tag{10}$$

Por substituição dos valores obtidos para gAB_t nas fórmulas (7) e (8), foram obtidas as frequências genotípicas dos haplótipos heterozigóticos, tanto com a presença do *A* na primeira, quanto na segunda posição do haplótipo (11):

$$g(Ab)_t = f - g(AB)_t \quad e \quad g(aB)_t = f - g(AB)_t \tag{11}$$

Fazendo uso da regra da complementariedade para a probabilidade, segundo a qual a soma das probabilidades de todos os eventos possíveis é 1, a frequência do haplótipo *ab* pode ser obtida por diferença considerando o somatório das demais probabilidades (12).

$$g(ab)_t = 1 - [\, g(AB)_t + g(aB)_t + g(Ab)_t \,] \tag{12}$$

Depois de obtidas as frequências genotípicas para cada um dos haplótipos foram derivadas as fórmulas para cada uma das frequências fenotípicas moleculares destes haplótipos (*fab, faB, fAb,* e *fAB*), no decorrer das gerações. Para isso foi admitido um nível de ploidia de 10 (n=10), que é considerada a ploidia média dos clones comerciais em uso. Assim, para que determinado fenótipo molecular tenha sido considerado produto da expressão do haplótipo *ab* nenhum dos cromossomos que carrega poderia apresentar o alelo *A* ou *B*, ou seja, todos os 10 hom(e)ólogos deveriam apresentar os alelos *a* e *b*. Como a probabilidade de determinado haplótipo ser encontrado se dá pelo produto das frequências dos alelos que o compõem, e na regra do produto para potenciações os expoentes são somados, elevamos a frequência genotípica do haplótipo *ab* à 10, obtendo a fórmula para esta primeira frequência fenotípica molecular (13).

$$fab_t = g(ab)_t^{10} \tag{13}$$

Da mesma maneira, para apresentar genótipo heterozigoto, pelo menos um alelo (no caso exemplificado pelo alelo *A*) precisa estar presente em pelo menos um dos hom(e)ólogos (Figura 5).

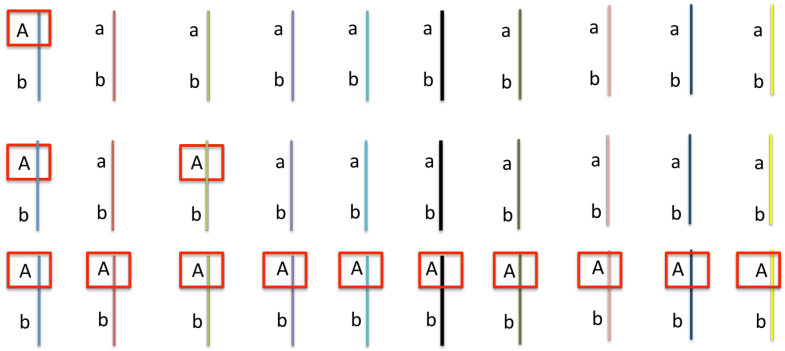

Figura 5. Ilustração do genótipo com presença do alelo apenas na primeira posição do haplótipo (*Ab*), mostrando que deve apresentar o alelo *A* na primeira posição em pelo menos um dos hom(e)ólogos.

Apresentando um, dois ou até dez doses da presença do alelo na primeira posição do haplótipo (*A*) o genótipo será *Ab*, o mesmo ocorre para o haplótipo *aB*, ou seja, na presença de um, dois, ou até dez doses da presença do alelo (*B)* na segunda posição do

haplótipo o genótipo será *aB*. Desta maneira para o cálculo das frequências fenotípicas moleculares de cada um destes dois haplótipos fez-se o somatório das combinações dos 10 homólogos *x* a *x,* com *x* sendo a presença do alelo na primeira posição do haplótipo (14) (ou a presença do alelo na segunda posição do haplótipo [15]), considerando a dosagem da presença do alelo variando entre 1 e 10. A fórmula foi então construída considerando as frequências genotípicas do haplótipo heterozigoto em questão e a do homozigoto para ausência dos alelos nas duas posições do haplótipo (*ab*). Assim,

$$fAb_t = \sum_{1}^{10} C_{10,x} \cdot g(Ab)_t^{x} \cdot g(ab)_t^{10-x} \tag{14}$$

$$faB_t = \sum_{1}^{10} C_{10,x} \cdot g(aB)_t^{x} \cdot g(ab)_t^{10-x} \tag{15}$$

Novamente, fazendo uso da regra da complementariedade das probabilidades, a frequência fenotípica molecular do haplótipo *AB* foi calculada pela diferença entre a frequência total da soma dos eventos possíveis (um) e a soma das frequências fenotípicas moleculares dos demais eventos possíveis (*fab, fAb* e *faB*).

$$fAB_t = 1 - [f(aB)_t + f(Ab)_t + f(ab)_t] \tag{16}$$

As frequências dos indivíduos carregando determinado haplótipo, ou seja, as frequências fenotípicas absolutas esperadas foram calculadas com base nas frequências fenotípicas moleculares teóricas admitindo-se o tamanho da população *nind* = 72 ou 200 indivíduos, como realizado por Raboin et al. (2008).

Os valores obtidos de frequências absolutas para cada haplótipo *(AB, Ab, aB e ab*) foram utilizados para a construção de uma tabela de contingência 2 x 2, permitindo o cálculo do p-valor associado pelo Teste Exato de Fisher (Figura 6). Como já comentado anteriormente, o uso de testes de significância também é realizado para constatação do LD (Gupta et al., 2005), sobretudo em situações em que não é possível obter as frequências alélicas diretamente. Assim, o Teste Exato de Fisher foi utilizado para avaliar a significância das associações entre as marcas, como sugerido por Raboin et al. (2008).

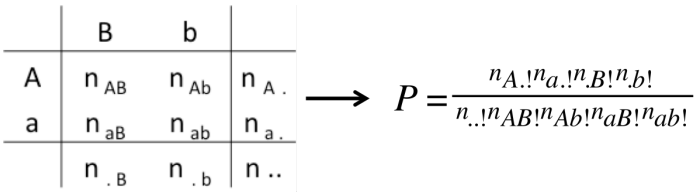

Figura 6. Exemplo da tabela de contingência construída com as frequências fenotípicas absolutas esperadas para tamanho populacional *n..* e fórmula utilizada para o cálculo das probabilidades associadas a cada um dos conjuntos utilizados para compor das tabelas de contingência.

As fórmulas descritas foram, então, implementadas em ambiente R, utilizando a linguagem e as funções nele disponíveis, de maneira à gerar os gráficos para avaliação do decaimento do desequilíbrio de ligação em diferentes cenários.

3.4.1. Obtenção do decaimento do LD para as populações de melhoramento

Para a avaliação do LD na população de melhoramento de cana-de-açúcar estudada, todas as informações genotípicas para os 172 indivíduos foram utilizadas, as duas populações foram subdivididas de acordo com a conformação dos locos nos parentais, exatamente como o realizado para a o teste de segregação mendeliana (Tabela 2). Foram utilizados 850 locos DArT nas análises da população de cruzamento biparental e 470 locos DArT para a população de autofecundação, estes locos foram selecionados por se apresentarem polimórficos.

O nível crítico de significância para identificação das associações significativas foi realizado de duas maneira. Primeiramente com o uso do FDR (*False Discovery Rate*), em que foram declarados significativos os valores de associação par-a-par com p-valores inferiores a 0,05. O segundo processo se deu pelo estabelecimento do limite empírico de -log p = 4,79588, o mesmo utilizado por Raboin et al. (2008), valor para o qual estes autores constataram a presença de associações verdadeiras (dentro do mesmo grupo de hom(e)ologia), descartando as associações espúrias (devidos à outros fatores que não a ligação física entre os locos). Com o uso deste limite foi traçada uma linha paralela à abscissa, indicando até onde as associações foram declaradas como significativas.

3.5 ANÁLISE DA SEGREGAÇÃO MENDELIANA

Uma das características básicas do genoma de cana é a poliploidia, ou seja, cada indivíduo carrega *n* (sendo $n \geq 4$) conjuntos de cromossomos do mesmo tipo e origem (mesmo grupo de hom(e)ologia). Desta maneira, como existem *n* cópias de um mesmo cromossomo carregando os mesmos locos, o número de cópias de um mesmo alelo (dado um loco) pode variar. Essa quantidade de cópias de um mesmo alelo é definida como a dosagem deste alelo (Garcia et al., 2013).

Devido a dificuldade de estabelecer as dosagens referentes aos alelos de cana-de-açúcar foi realizada uma análise de segregação mendeliana para cada loco utilizando o teste de Qui-quadrado (X^2).

Durante a análise do padrão de decaimento do LD na população estudada, foram utilizados somente os locos polimórficos, ou seja, foram utilizados os dados de 470 locos DArT para os 81 indivíduos originados por autofecundação do clone elite RB97327 e 850 locos DArT para os 91 indivíduos obtidos no cruzamento biparental (RB97327 x RB72454). Entre os locos escolhidos foram encontrados diferentes padrões de segregação, pois estes se apresentavam diferentemente nos parentais (presente [1] ou ausentes [0] nos parentais).

No caso da população de cruzamento biparental, por exemplo, a marca poderia estar presente (1) em um dos parentais e ausente (0) no outro, com duas combinações possíveis: 1x0 e 0x1, ou presentes nos dois parentais, no caso 1x1. Para a população de autofecundação, apenas uma conformação foi utilizada, a de presença na mãe, gerando o padrão 1x1. Estas conformações, quando produzidas por uma única dose do alelo são denominadas de *single dose*, ou como *double single dose* quando a presença em única dose é observada nos dois parentais. Assim, além de existirem duas populações (de cruzamento biparental e de autofecundação), existem divisões dentro de populações de acordo com a segregação esperada pela conformação das marcas observadas nos parentais (Tabela 2).

As segregações esperadas para marcadores *single dose* seguem o padrão mendeliano simples e o padrão de segregação observado pode ser testado pelo teste de Qui-quadrado (X^2) (Bearzoti, 2000). Desta maneira, para a análise do padrão de desequilíbrio encontrado, as marcas utilizadas foram submetidas individualmente a uma análise de padrão de segregação. Assim, foram realizados testes de X^2, considerando-se

como segregações mendelianas esperadas: 3:1 para padrões 1x1 nos parentais e 1:1 para os padrões 1x0 e 0x1 nos parentais.

Tabela 2. Populações e grupos dentro de populações utilizados para a análise do LD utilizando o modelo de Raboin et al. (2008).

Subpopulação	Grupo	Segregação esperada
Cruzamento biparental	1x0	1:1
	0x1	1:1
	1x1	3:1
Autofecundação	1x1	3:1

Para a visualização dos padrões de segregação obtidos com os testes de X^2 nos gráficos de LD, foram comparados par-a-par os resultados de X^2 obtidos, resultando em duas diferentes colorações nos gráficos de decaimento do LD para as populações de melhoramento, a coloração vermelha para associações contendo locos com segregações distorcidas (em que pelo menos um dos locos da comparação par-a-par apresentou distorção no teste de segregação) e coloração azul quando as associações envolviam locos com aderência à segregação mendeliana esperada.

3.6 SELEÇÃO GENÔMICA AMPLA

Para verificar o potencial de desenvolvimento da Seleção Genômica Ampla para cana-de-açúcar foi realizada a análise RR-BLUP com os dados fenotípicos e genotípicos das duas populações, formando uma única população, composta por 172 indivíduos.

Para o correto procedimento da análise todos os locos e indivíduos que possuíam dados faltantes em excesso foram desconsiderados, pois com dados faltantes não é possível se obter uma matriz de genótipos sob a qual o sistema de equações matriciais seja resolvido, inviabilizando a obtenção dos parâmetros necessários à construção do modelo preditivo. Desta maneira foram utilizados na análise os dados de 952 locos DArT, compondo a matriz de genótipos do modelo para 132 indivíduos da população original. Os seis caracteres fenotípicos foram analisados separadamente (peso, diâmetro, comprimento de colmo, brix, número de internódios e número de colmos).

A análise foi realizada com o uso do pacote rrBLUP (Endelman, 2011) implementado na plataforma R, com o uso do comando *mixed.solve*. O modelo de análise se deu como descrito na equação (5) deste documento, em que se assumem os VGG como os valores preditos de *y*, obtidos com base no vetor de efeitos fixos (*b*), no vetor de efeitos aleatórios (*m*) e no vetor de resíduos aleatórios (*e*), com os dois primeiros vetores ponderados pelas matrizes de incidência X e Z, respectivamente. A matriz Z possui informações sobre as marcas, utilizando valores para determinar a presença ou ausência de alelos no marcador e a matriz X é a matriz de incidência dos efeitos fixos do modelo.

$$y = Xb + Zm + e \tag{5}$$

A estrutura de médias e variâncias é dada por $m \sim N(0, G)$; $E(y = Xb)$; $e \sim N(0, R = I\sigma^2_e)$; Var ($y$) = V = ZGZ' + R; e $G = I\sigma^2_A / n$, como descrito por Fritsche-Neto et al. (2012).

As equações matriciais (17) utilizadas para a obtenção dos efeitos dos marcadores e predição dos VGG (7) foram:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \sigma^2_e \,.\, \frac{n}{\sigma^2_A} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{m} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix} \tag{17}$$

$$VGG = \hat{y}_j = \sum_i Z_i \hat{m}_j \tag{6}$$

Sendo (σ^2_A / n) a variância genética aditiva (σ^2_A) explicada por cada um dos *n* locos, tendo sido estimada por REML, com o uso dos dados fenotípicos (eBLUPs). Assumiu-se a *priori* de que todas as marcas possuem o mesmo efeito para a explicação de cada caráter (Meuwissen et al., 2001). O fator de regressão λ pode ser obtido por relações matemáticas (18) em função da herdabilidade ajustada (h^2) e o número de locos que controlam o caráter (*n*), sendo este desconhecido previamente, mas podendo ser obtido com o uso da equação (19):

$$\lambda = n \, \frac{(1 - h^2)}{h^2} \tag{18}$$

$$n = \left[\sum_{i}^{n} p_i (1 - p_i) \right] \tag{19}$$

Assim, a partir do uso das frequências dos marcadores (p_i) e da substituição da equação (19) na equação (18) pôde-se obter o fator de regressão λ (20), responsável pelo efeito de *shrinkage*.

$$\lambda = n \frac{(1 - h^2)}{h^2} = \left[\sum_{i}^{n} p_i (1 - p_i) \right] \frac{(1-h^2)}{h^2} \tag{20}$$

3.6.1. Validação Cruzada Por *jack-knife* - Obtenção da Acurácia do Modelo

A validação cruzada foi utilizada para a verificação da acurácia do modelo de GWS obtido, a partir da correlação entre os VGG preditos por *jack-knife* e os eBLUPs de cada um dos genótipos.

Uma das metodologias utilizadas para a validação cruzada em GWS é denominada *jack-knife*. Esta metodologia se baseia na divisão do conjunto original de (*N*) dados em (*g)* grupos de tamanho igual a *k*, de forma que *N* = *gk*, com a retirada de *n* observações em cada reamostragem (Resende, 2002). Desta forma o grupo de indivíduos *N* - *n* é utilizado para a previsão dos resultados do conjunto *n,* sendo este procedimento realizado *n* vezes. Comparando os resultados obtidos para *n* pela previsão realizada com o uso das *N* - *n* observações e os resultados observados para cada *n* (no caso os eBLUPs) foi obtida a acurácia da predição dos VGG e, assim, foi verificado o potencial de aplicação do método para a cultura.

Duas metodologias foram utilizadas para validar a acurácia do modelo na análise do conjunto de dados: a primeira pela decomposição do conjunto total *N*, deixando apenas uma observação fora da análise de cada vez (*leave one out* = *N* - 1), de maneira a prever o valor de um dos genótipos utilizando os demais 119 para predição do seu VGG. E a segunda foi realizada dividindo-se aleatoriamente o grupo original em 10 partes iguais (*10-fold-validation*), gerando 10 grupos com 12 indivíduos, realizando a predição dos valores de VGG para um dos grupos a partir do modelo de predição gerado com o uso dos outros 9 grupos. Assim, a predição dos VGG foi realizada para cada subgrupo, re-treinando o modelo com base nos demais nove subgrupos.

4 RESULTADOS E DISCUSSÃO

4.1 MODELAGEM DO DESEQUILÍBRIO DE LIGAÇÃO PARA ESPÉCIES POLIPLOIDES

Foi gerado um modelo para o cálculo do desequilíbrio de ligação considerando fatores intrínsecos ao genoma de cana-de-açúcar. Foram obtidos gráficos relacionando a medida de LD (−log p) à frações de recombinação teoricamente impostas. Além da suposição de LD completo entre os haplótipos fundadores para a análise, também foram consideradas frequências estáveis dos alelos ao longo das gerações e associação aleatória durante a meiose, com ausência de estrutura populacional como realizado por Raboin et al., 2008.

Após a obtenção dos gráficos com o uso da modelagem foi realizada a comparação com os gráficos gerados por Raboin et al. (2008) (Figura 7). Considerando os gráficos gerados pelo modelo implementado e os disponíveis no artigo de Raboin et al. (2008), pode-se afirmar que o modelo foi implementado com sucesso. Visto que as curvas apresentadas para as diferentes frequências alélicas em relação à taxa de recombinação, para três e sete gerações, considerando os dois tamanhos populacionais (72 e 200 indivíduos) foram compatíveis às curvas apresentadas no artigo base.

A análise dos gráficos gerados permite inferências sobre alguns fatores evolutivos que podem interferir na dinâmica do LD, como o evidente efeito da recombinação por exemplo (Lewontin, 1964; Mackay & Powel, 2007). Segundo Lewontin & Kojima (1960) uma população qualquer possui uma taxa de aproximação do equilíbrio (decaimento do desequilíbrio) relativa à taxa de recombinação existente (*r*), ou seja, relativa ao número de eventos de recombinação ocorridas ao longo das gerações (*t*) (21).

$$D_t = (1\text{-}r)^t \, . \, D_0 \tag{21}$$

Assim, como visualizado nos gráficos, quanto maior a taxa de recombinação mais rápido o decaimento do LD (Hartl & Clark, 2010). É interessante citar que, se a fração de recombinação for reduzida à metade da original, o número de gerações

necessárias para a dissipação do LD praticamente dobra (Mackay & Powel, 2007).

Outro fator importante nesta dinâmica é o número de processos meióticos (gerações) transcorridos (*t*). O mesmo efeito visualizado para recombinação pode ser visualizado para a quantidade de gerações transcorridas, visto que estes fatores estão intrinsecamente ligados. Para 3 gerações o LD encontrado é superior ao LD estimado para 7 gerações transcorridas (7 processos meióticos), tanto quando são considerados 72 ou 200 indivíduos, ou seja, quanto maior o número de processos meióticos, maior o decaimento do LD (Hartl & Clark, 2010; Hedrick, 2011).

É importante salientar que, no contexto em que foram realizadas as simulações, o tamanho das populações interferiu nos valores de LD detectados. Como já comentado, o LD foi avaliado utilizando-se a medida −log p, sabidamente dependente do tamanho amostral. Nos gráficos obtidos é possível visualizar que valores mais baixos de LD foram detectados na população de 72 indivíduos do que na população contendo 200 indivíduos, como esperado pelo maior poder estatístico inerente à amostras maiores.

Segundo Hedrick (1987) e Gupta et al. (2005) a melhor medida de LD deve ser escolhida considerando o objetivo do estudo a ser realizado, devido às diferentes interpretações que elas podem proporcionar. Além disso, a maior parte das medidas existentes se resume a extensões simples de modelos que consideram poucos locos (Lewontin, 1964). Cada uma das novas medidas sugeridas para a mensuração do LD sugere apenas alguma modificação para a diminuição de viés ocasionado por determinado fator (Mangin et al., 2011). Sendo que cada um destes diferentes algoritmos podem gerar inferências sobre as forças atuantes na população analisada (Weir, 1979), ou seja podem ser indicadores do efeito de vários fatores evolutivos aos quais as populações estão submetidas (Hedrick, 1987).

Assim, as medidas existentes são apenas "adaptações" considerando algum dos fatores evolutivos que influenciam a dinâmica do LD. O método utilizado neste trabalho considerou como fatores interferentes na dinâmica do LD a polissomia e a alta poliploidia, características inerentes ao genoma de cana. Embora ainda existam controvérsias sobre qual medida é mais adequada para cada caso (Slatkin, 1994), nenhuma medida além da sugerida por Raboin et al. (2008) foi desenvolvida para a cultura da cana-de-açúcar.

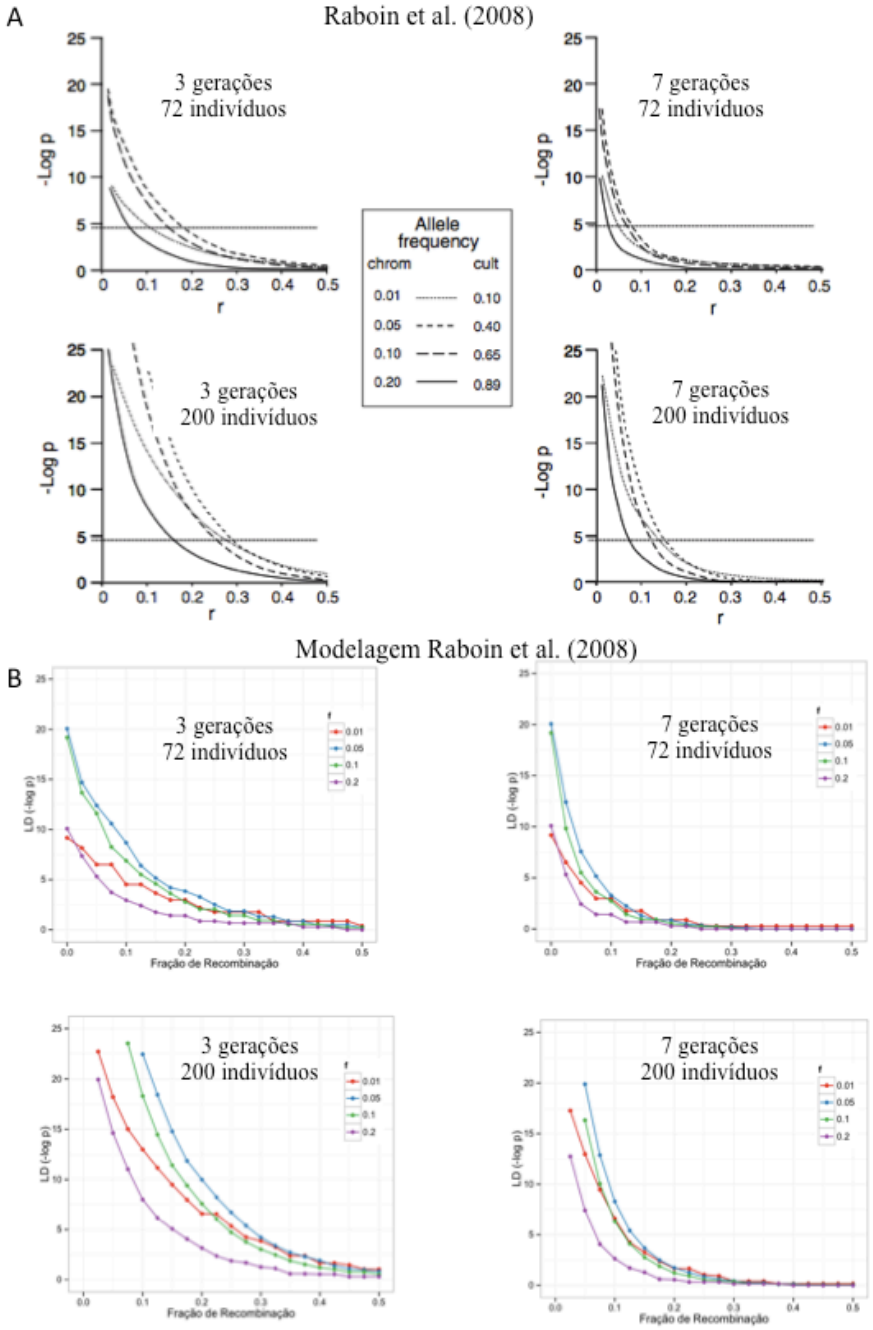

Figura 7. Comparação da dinâmica esperada do LD entre os gráficos obtidos por Raboin et al. (2008) e os obtidos com base no modelo teórico após sua implementação na plataforma R, considerando a relação entre a medida adotada (-log p) e a taxa de recombinação entre os marcadores. A) Gráficos disponíveis no artigo base Raboin et al. (2008); B) Gráficos gerados pela modelagem implementada no ambiente R a partir das fórmulas de Raboin et al. (2008). Ambos considerando 3 ou 7 gerações transcorridas após o estabelecimento da população fundadora considerando 72 ou 200 indivíduos.

4.1.1. Obtenção do perfil do decaimento do LD para as duas populações de melhoramento sob estudo

A obtenção da magnitude do desequilíbrio de ligação tem grande valor por si só, visto que se trata de um fenômeno genético. Além disso , seu estudo é essencial para a compreensão da estrutura do genoma (Lopes, 2011) e para as análises e processos envolvendo associação entre marcas e características de interesse agronômico (Durães et al., 2004; Resende, 2008; Resende et al., 2010; Rosa, 2011).

A metodologia de Raboin et al. (2008) foi utilizada para avaliar o decaimento do desequilíbrio de ligação para as população de cruzamento biparental e de autofecundação. Foram obtidos gráficos relacionando a medida de LD (-log p) à distância genética estimada entre os locos (cM). Duas significâncias estão envolvidas nesta análise, sendo a primeira a significância do desequilíbrio em si e sua extensão no genoma, considerando qual seria o tamanho dos blocos de ligação nas populações em estudo. E a segunda envolvendo a aderência dos locos ao teste de X^2 considerando a segregação mendeliana esperada para locos *single dose* e as observadas no conjunto de dados analisados.

Graficamente estas duas significâncias foram representadas da seguinte maneira: a significância que envolve o decaimento do desequilíbrio em si, é mostrada no gráfico a partir da plotagem dos valores de LD *versus* as frequências de recombinação. A segunda significância envolve as cores dos gráficos, ou seja, em azul as combinações que correspondem à associações entre locos par-a-par que apresentaram aderência ao teste de X^2, cuja conformação se deu de acordo com a esperada (=1:1 ou =3:1) e em vermelho as associações entre locos que apresentaram segregações diferentes das esperadas, que não apresentaram aderência quando submetidas ao teste de X^2 ($\neq$1:1 ou $\neq$3:1).

Os valores obtidos pela estatística *-log p* foram representados graficamente, de modo a permitir a análise da estrutura do LD no genoma desta população e gerar inferências sobre o comportamento deste fenômeno no genoma de cana.

Para a eliminação das associações espúrias (falso-positivas) devido ao grande número de comparações realizadas, foi utilizada a correção FDR, com o qual foram obtidos os números de associações significativas (LD significativo) para cada grupo dentro de população.

Considerando a análise da população de cruzamento biparental, a quantidade de associações significativas representou em média 80% do total de comparações realizadas dentro de cada um dos grupos, Para a análise do desequilíbrio no grupo 1x0 desta população foram realizadas 23005 testes de associação par-a-par. Deste total 18651 se apresentaram significativos ou seja aproximadamente 81% das comparações geraram resultado significativo após correção FDR (Figura 8). Já para o segundo grupo desta população (0x1), foram realizadas 72010 comparações par-a-par entre os locos e aproximadamente 80% delas (57520) envolviam locos com associações significativas (Figura 9). Para os indivíduos do grupo 1x1, 32385 comparações foram realizadas, sendo 17931 significativas (55,37%) (Figura 10).

Além da correção pelo FDR, foi considerado como padrão para a estipulação da significância do LD os padrões obtidos para o genoma de cana nos demais trabalhos existentes na literatura, assumindo como limite empírico para a assunção de LD significativo o valor de -log p = 4,79588 (equivalente a um valor de $r^2 = 0,2$) indicado nos gráficos como uma linha paralela à abcissa. Este valor é indicado por Raboin et al. (2008) como o limite que contém as associações que se devem à ligação física, ou seja, que contem os locos que estão em um mesmo grupo de hom(e)ologia. Tomando este limite o desequilíbrio de ligação encontrado para os grupos da população de cruzamento biparental abrangem o intervalo de 25 a 30 cM.

Considerando os indivíduos resultantes da autofecundação da variedade RB97327 (subgrupo 1x1), a relação entre o desequilíbrio de ligação e a distância física (cM) existente foi similar aos da população do cruzamento biparental (Figura 11), a significância considerando a correção FDR aponta que das 65535 associações par-a-par testadas, 47,1% mostraram-se significativas (30869), também é apontada a significância considerando o limite empírico (-log p = 4,79588), que indica um LD de 25 cM, valores altos como este já eram esperados, tanto pelo histórico de melhoramento da cultura, como pela existência de autofecundação, visto que o sistema reprodutivo interfere no LD.

Segundo Bennet & Binet (1956) existe influência do sistema de acasalamento na dinâmica do desequilíbrio de ligação. Estes autores evidenciaram que não só a recombinação, mas também as taxas de autofecundação (*s*) interferem no decaimento do LD ao longo das gerações (*t*). Devido à autofecundação, o desequilíbrio de ligação é em parte preservado, mantendo-se ao longo das gerações, mesmo entre locos não ligados

(Ritland & Hedrick, 1990).

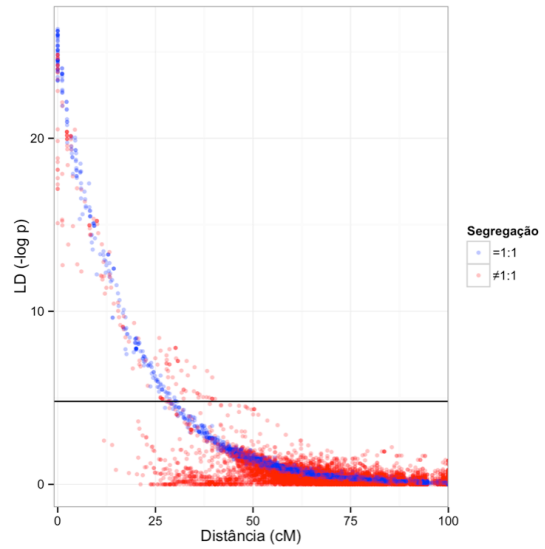

Figura 8. Relação entre o desequilíbrio de ligação e a distância, em cM, entre locos, na população de indivíduos oriundos do cruzamento entre os clones-elite RB97327 e RB72454 (*single dose* com banda no genitor feminino). Em azul, pares de locos que aderiram à segregação esperada 1:1. Em vermelho, pares em que pelo menos um dos locos apresenta desvios de segregação significativos, em relação à segregação mendeliana esperada,1:1.

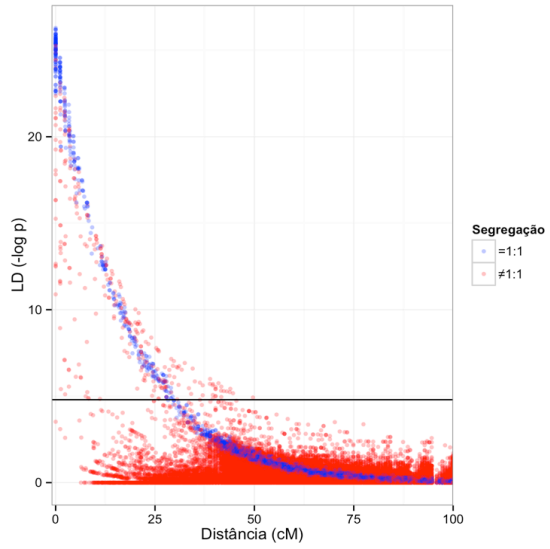

Figura 9. Relação entre o desequilíbrio de ligação e a distância, em cM, entre locos, na população de indivíduos oriundos do cruzamento entre os clones-elite RB97327 e RB72454. Em azul, pares de locos que aderiram à segregação esperada 1:1 (*single dose* com banda no genitor masculino). Em vermelho, pares em que pelo menos um dos locos apresenta desvios de segregação significativos, em relação à mendeliana esperada, 1:1.

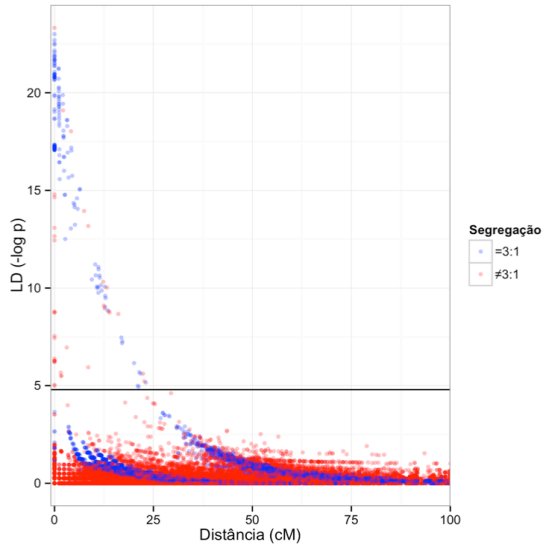

Figura 10. Relação entre o desequilíbrio de ligação e a distância, em cM, entre locos, na população de indivíduos oriundos do cruzamento entre os clones-elite RB97327 e RB72454. Em azul, pares de locos que aderiram à segregação esperada 3:1 (*double single dose* com bandas nos dois genitores). Em vermelho, pares em que pelo menos um dos locos apresenta desvios de segregação significativos, em relação à mendeliana esperada, 3:1.

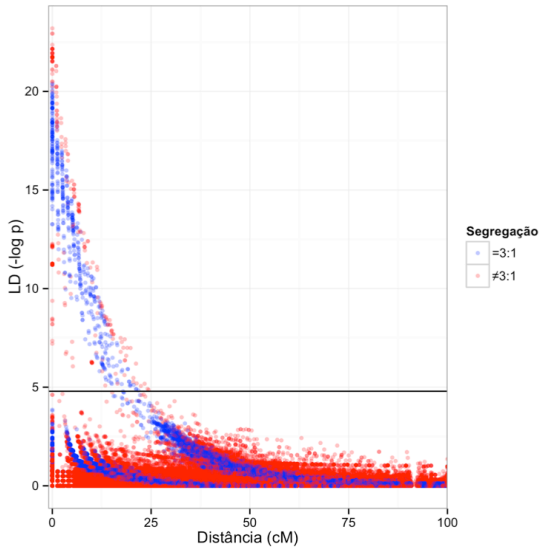

Figura 11. Relação entre o desequilíbrio de ligação e a distância, em cM, entre locos, na população de indivíduos oriundos da autofecundação do clone-elite RB97327. Em azul, pares de locos que aderiram à segregação esperada 3:1. Em vermelho, pares em que pelo menos um dos locos apresenta desvios de segregação significativos, em relação à 3:1.

Os resultados obtidos para as duas populações corroboraram os visualizados

para outras populações de cana-de-açúcar, porém é importante citar que o trabalho desenvolvido foi o primeiro a verificar o LD em uma população de autofecundação de cana.

O primeiro trabalho desenvolvido acerca deste fenômeno genético em cana foi o de Jannoo et al. (1999b). Em seu estudo com 59 cultivares de cana utilizando 72 marcadores RFLPs estes autores averiguaram que a maior parte das associações significativas entre os locos (75%) se encontravam a distâncias de até 10cM. Além disso, eles puderam corroborar a hipótese de que algumas das associações observadas derivavam de cruzamentos antigos, utilizando marcadores específicos em genótipos utilizados nos cruzamentos que deram origem às cultivares estudadas, mostrando a magnitude do efeito fundador no LD apresentado nas variedades atuais. Além disso, esse trabalho também detectou algumas associações entre locos extremamente distantes (90 cM), mostrando que existem no genoma associações que não se devem à ligação física.

Outro estudo relatando a dinâmica do LD em cana foi elaborado por Raboin et al. (2008). Utilizando marcadores dominantes, eles corroboraram os resultados obtidos por Jannoo et al. (1999b), verificando o rápido decaimento do LD em distâncias maiores do que 30 cM.

Rosa (2011) e Lopes (2011) obtiveram resultados condizentes aos previamente citados. Rosa (2011) verificou forte presença de LD a distâncias de 15 cM, reforçando ainda a existência ostensiva de associações significativas em regiões abaixo de 5 cM. Como Jannoo et al. (1999b), também encontrou associações a grandes distâncias (65 cM), mostrando a presença de associações referentes a outros fatores que não ligação física. Lopes (2011), em seu estudo, também constatou que a maioria das associações significativas (43,9%) estava contida na janela de 0 a 10 cM, e quase a totalidade delas (98%) encontravam-se até 30 cM.

De acordo com os resultados obtidos nos trabalhos envolvendo a estimação do LD em populações de cana-de-açúcar e os encontrados neste trabalho, pode-se inferir sobre a dinâmica do LD no genoma da cultura e a influência do melhoramento na sua dinâmica. Visto que os processos microevolutivos interferiram e interferem nesta dinâmica (Lewontin, 1988).

Assim, como efeitos do melhoramento, um dos fatores que pode ter interferido no padrão de LD é a deriva. Dois processos do melhoramento que podem ter gerado efeito

sobre o LD ligados a este fator são o uso de populações finitas e o forte efeito fundador envolvido na produção das variedades cultivadas. Kojima (1971) estudou o efeito do tamanho efetivo finito sobre o desequilíbrio de ligação nas populações e concluiu que os efeitos de deriva genética aleatória e distúrbios estocásticos são considerados equivalentes aos efeitos de segregação e mutação genética nas frequências alélicas em populações finitas com tamanhos variados. Weir (1979), estudando teorias e análises de LD, mostra que dentro de populações finitas, mesmo com a população sendo submetida a inúmeros processos meióticos, o LD não se desfaz. Segundo esse autor existem algoritmos específicos para a determinação exata deste valor, como por exemplo o derivado por Weir & Cockerham (1974).

O segundo efeito da deriva, consoante ao anteriormente citado, desenvolve-se como o efeito fundador, como o encontrado por Jannoo et al. (1999b), interferindo no LD pela diminuição no tamanho efetivo populacional. Quanto menor o tamanho efetivo da população fundadora, maior o LD gerado e, consequentemente, maior a quantidade de processos meióticos necessários para a quebra deste desequilíbrio (Jorde, 1995; Nordborg & Weigel, 2008).

Outro fator que pode ter interferido no LD é a seleção, tanto a natural quanto a artificial. A ocorrência de seleção provoca alterações nas frequências alélicas aumentando a frequência dos alelos ligados a determinadas características na população. Tomando como exemplo a seleção natural, ao dirigir determinada população a máximos adaptativos locais, os genótipos que apresentam combinações alélicas de maior valor adaptativo são favorecidos, em consequência, eles serão observados em maior quantidade na população, sendo responsáveis pela composição das demais gerações, favorecendo a manutenção do LD (Lewontin & Kojima, 1960; Lewontin, 1964). O processo exemplificado é exatamente o mesmo para a seleção artificial que é utilizada no processos de melhoramento, desta maneira, este fator deve ter contribuído para a geração do LD visualizado em cana-de-açúcar.

Todavia, a hipótese de que a ligação é a mais aceita para os valores de LD observados em cana. A partir dos valores de LD citados na bibliografia e dos obtidos na análise realizada pode-se inferir sobre o tamanho dos blocos de ligação herdados pelas variedades modernas como sendo 10 cM, com a maioria das associações significativas encontradas até a 5 cM. E o intervalo de 0 a 30 cM contendo praticamente a maioria das

associações significativas. Estes valores corroboram a hipótese de que a ligação é um dos principais fatores de origem do LD em cana, mesmo com a grande influência dos demais fatores microevolutivos na formação das variedades atuais.

O LD apresentado pela cultura se aproxima mais dos valores apresentados por espécies autógamas, mesmo a cana sendo considerada uma espécie alógama. Entretanto os altos valores observados de LD para todos os grupos analisados são compreensíveis, visto que além do sistema reprodutivo, somam-se aos fatores que podem ter afetado a magnitude do desequilíbrio observado a seleção, o efeito fundador gerado nos primeiros cruzamentos e o baixo número de meioses realizadas, processos inerentes aos programas de melhoramento da cultura. Além disso, características do próprio genoma, como as enumeradas anteriormente, como a origem interespecífica, a baixa taxa de recombinação entre os diferentes genomas, também devem ter contribuído para esta elevada magnitude de LD nesta população, sendo, contudo, o principal efeito do LD encontrado a ligação física entre os alelos dos diferentes locos.

Vê-se que este fenômeno é comum no genoma de cana, estudos aprofundados podem permitir grandes avanços, não só para o melhoramento da cultura, mas também para o esclarecimento dos processos de obtenção das variedades modernas e para o entendimento da estrutura e da dinâmica do genoma desta cultura (Jannoo et al., 1999; Raboin et al., 2008; Lopes, 2011).

É importante considerar que a alta ploidia dificultou, para todos os trabalhos citados, inclusive este, a mensuração precisa das associações. A complexidade do genoma da cultura é ainda um forte desafio a ser ultrapassado. Novos estudos e maior quantidade de informação genética devem ser utilizados para melhorar a qualidade da informação obtida, permitindo inferências mais acuradas sobre o LD nas diferentes populações.

Considerando agora a segunda significância envolvida na interpretação dos gráficos de decaimento do LD, ou seja, a relacionada à aderência das comparações par-a-par às segregações mendelianas esperadas para locos *single dose*, observa-se que os locos que segregaram de acordo com o esperado apresentaram um comportamento definido e diferenciado daquele observado para locos com segregação diferente da esperada, provavelmente por esses locos não apresentarem realmente a conformação *single dose.*

Cumpre destacar que este comportamento abre caminho para a possibilidade teórica de se identificar a dosagem alélica nos genitores com base na relação entre o

desequilíbrio de ligação e a distância estimada entre os locos. A pressuposição de segregação *single dose* baseada no fenótipo molecular dos parentais, quando não é verificada, produz um viés nesta relação. Segundo Garcia et al. (2013) a dosagem de alelos é um dos problemas mais sérios durante as análises envolvendo genomas poliploides. A maioria das técnicas utilizadas (AFLP, SSR, entre outras) não permitem a averiguação da quantidade de doses de um mesmo alelo. A presença de determinado alelo só determina na verdade a presença de pelo menos uma cópia deste alelo para determinado loco. No caso dos marcadores utilizados vê-se a dificuldade clara de indicação da dosagem correta do alelo pela interpretação do fenótipo molecular dos marcadores nos parentais, visto que grande parte das segregações apresentaram-se diferentemente das mendelianas esperadas. Neste contexto, o desenvolvimento de metodologias de dosagem alélica assume grande importância.

Os resultados aqui obtidos sugerem que as diferentes segregações inerentes a diferentes dosagens alélicas podem apresentar diferentes perfis de decaimento do LD, embora seja necessário um maior conhecimento sobre as diferentes dosagens alélicas existentes nos genitores avaliados para que conclusões mais seguras possam ser estabelecidas.

4.2 SELEÇÃO GENÔMICA AMPLA

Pelo uso do pacote rrBLUP (Endelman et al., 2011), implementado na plataforma R, foi obtido um modelo de GWS para a população de 132 indivíduos com o uso de informações de 952 locos DArTs. Com base nos modelos produzidos pôde-se obter os VGG de cada indivíduo, para cada um dos seis caracteres avaliados. Valores de acurácia de predição foram obtidos para cada um dos caracteres e gráficos de dispersão foram construídos evidenciando a relação entre os VGG preditos para os indivíduos e os valores por eles apresentados durante a fenotipagem (eBLUPs).

Inicialmente foi realizada a análise com o "modelo cheio" (full), ou seja, considerando as informações de todos os indivíduos para o cálculo dos VGG (full), e posteriormente foram realizadas validações por dois diferentes métodos *Leave One Out* (LOO) e *Ten Fold Validation* (10-fold), ambos por *jack-knife*.

A partir da correlação entre os VGG obtidos nas validações e os valores de

eBLUPs foram obtidos os valores de acurácia para os seis caracteres fenotípicos avaliados. Para a inferência sobre os valores obtidos foram considerados os valores obtidos pela validação LOO, visto que esta gera valores mais confiáveis.

Os maiores valores de acurácia (considerando o modelo LOO) foram obtidos para diâmetro (0,88), peso (0,84) e comprimento (0,74). Para ^{o}Brix e número de internódios, as acurácias obtidos foram baixas, 0,09 e -0,018, respectivamente. Segundo Heffner et al. (2009) valores de acurácia de 0,85, mesmo para caracteres de baixa herdabilidade, são suficientes para a seleção baseada apenas nas informações de marcadores. Tradicionalmente, com uma herdabilidade média (h^2= 0,5), a acurácia de seleção a partir dos fenótipos (BLUPs, por exemplo) é aproximadamente 0,7 (Goddart, 2009). Entretanto, para grande parte dos caracteres de importância agronômica, valores tão ou mais altos de herdabilidade são difíceis de se obter, dado seu caráter quantitativo e a alta influência do ambiente que sofrem (Chaves, 2001).

Visto que o sucesso de seleção está diretamente relacinado à acurácia obtida com os modelos (Meuwissen et al., 2001; Resende et al., 2008; Desta & Ortiz, 2014), acurácias como as obtidas com o uso da abordagem GWS trazem ótimas perspectivas de ganho para a cultura. Além disso, 0,18 pode ser considerado um ganho considerável em acurácia, quando considerados valores de acurácia de 0,7 para seleção fenotípica (como por exemplo para diâmetro - acurácia= 0,88). Esta abordagem também se mostra superior à seleção de genes candidatos, ou ao uso de baixa quantidade de marcadores para a seleção quando a proporção da variabilidade explicada por estes marcadores é baixa, o que acontece na grande maioria dos casos, principalmente para caracteres quantitativos (Meuwissen & Goddart, 1996; Goddart & Hayes, 2007).

Além disso, valores de acurácia entre são apontados como úteis para a seleção em milho. Valores ainda mais baixos, são utilizados para GWS em arroz, girassol e gado. Assim, os altos valores de acurácia, demostram o potencial de previsão do modelo obtido, ou seja o potencial de previsão do VGG dos indivíduos pela análise dos seus genótipos. De forma que as correlações dos VGG preditos e dos valores de eBLUPs mostraram-se altas independente do método de validação utilizado (Tabela 3). Os valores obtidos servem, então, como prova de conceito, de que a abordagem é possível para cana e de que podem ser obtidos altos valores de acurácia com o uso de modelos de GWS para cana-de-açúcar Porém, maior validação é requerida, visto que os efeitos dos marcadores foram estimados

com o uso de uma população diferente daquelas em que a seleção é realmente realizada para a cultura e este modelo pode gerar baixa acurácia quando aplicados a outras populações (Goddard & Hayes, 2007)

Tabela 3. Resumo dos resultados obtidos na análise de Seleção Genômica Ampla para a população de melhoramento da Ridesa, contendo 120 indivíduos, genotipados para 255 locos e caracterizados para seis diferentes caracteres.

Variável fenotípica	Acurácia full	Acurácia LOO	Acurácia 10Fold
Brix	0,5545324	0,0892427	0,0764571
Comprimento	0,8746668	0,7483619	0,7472396
Diâmetro	0,9511555	0,8867674	0,8835798
Nº colmos	0,9585569	0,667462	0,6523655
Peso	0,9577003	0,8474245	0,8345618
Nº internódios	0,4779828	-0,018138	-0,1001041

.

Na análise dos gráficos de dispersão, que retratam a relação entre os valores predito e os valores observados (Figura 12), observou-se a presença de dois diferentes aglomerados de pontos. Verificou-se que quanto maior a acurácia mais concisos e determinados estes grupos se encontravam, como pode ser visto claramente no gráfico que retrata a dispersão dos valores para diâmetro. Os demais caracteres apresentaram este efeito tão menos evidenciado quanto menor a acurácia do modelo. A análise minuciosa dos dois diferentes grupos mostrou que esta separação se devia ao efeito da estruturação devido ao sistema de acasalamento na população original.

A análise das estatísticas descritivas evidenciou o efeito de *shrinkage* nos valores preditos pelo modelo, mais evidente nos valores preditos para os quais as acurácias são mais baixas, como é o caso do ºBrix, por exemplo. Mas mesmo com os valores de acurácia altos pode-se ver a "contração" dos dados, retratando o efeito da penalização (Figuras 13 e14).

Poderia-se cogitar o baixo número de marcas ou o pequeno tamanho populacional como motivos para efeitos tão evidentes quanto os obtidos, visto que o aumento na densidade de marcadores proporciona ganhos na acurácia (Solberg et al., 2008), bem como o aumento do tamanho das populações utilizadas para obtenção e validação do modelo (Desta & Ortiz, 2014). Entretanto, outros trabalhos foram desenvolvidos considerando números ainda menores tanto de indivíduos como de marcas utilizadas, apresentando resultados semelhantes aos obtidos para cana, também com

acurácias elevadas. Fritsche-Neto et al. (2012), estudando 41 combinações híbridas de milho utilizando 80 marcadores microssatélites, fazendo uso do mesmo método utilizado neste trabalho (RR-BLUP) obteve resultados satisfatórios de acurácia. Cavalcanti et al. (2012), também escolheram como método para a seleção genômica ampla em cajueiro o RR-BLUP, na análise de 74 indivíduos de uma família de meios irmãos utilizando dados de 238 marcadores. A eficácia destas análises mostra que o método pode ser eficiente, mesmo com baixo número de marcas e de indivíduos. Mostrando o potencial desta abordagem para o melhoramento da cultura, o que pode permitir grandes ganhos por unidade de tempo, visto que reduz o tempo dos ciclos de seleção (Heffner et al., 2009).

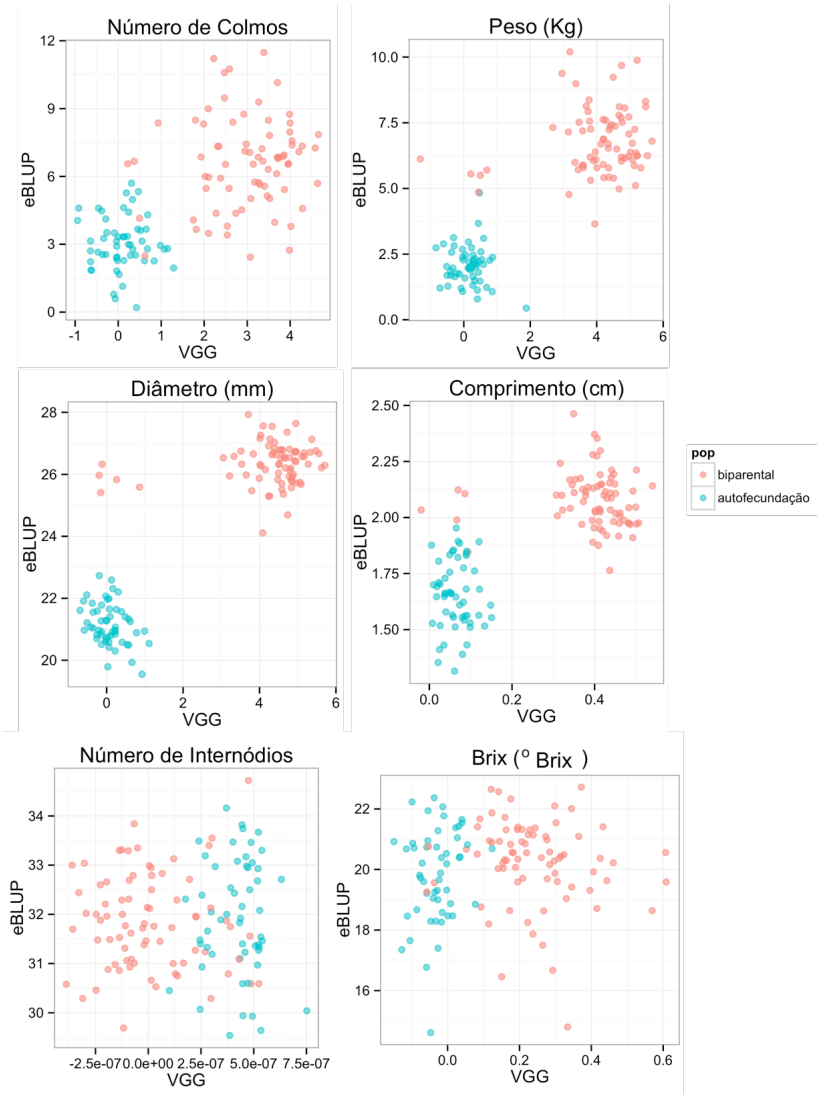

Figura 12. Gráficos de dispersão relacionando os VGG preditos obtidos com o uso do pacote rrBLUP e os valores de eBLUPS observados para cada um dos seis caracteres avaliados.

É importante salientar que os valores de acurácia obtidos devem ser considerados como potenciais em populações de base genética mais ampla, já que na população avaliada o LD apresentado foi bastante elevado. O alto LD apresentado na população em parte se contrapõe ao baixo tamanho populacional utilizado. Sabe-se que o número de indivíduos utilizados para uma análise de GWS deve ser suficiente para que todos os possíveis haplótipos estejam com seus fenótipos devidamente representados. A frequência dos haplótipos está diretamente relacionada à frequência dos alelos e à distância dos marcadores, ou seja, depende da taxa de recombinação (Schaeffer, 2006). Assim, ela é afetada pelo LD existente. Sendo o LD elevado, um menor número de indivíduos pode representar o número de haplótipos possíveis, ou se aproximar dele, justificando novamente o número de indivíduos utilizados na análise.

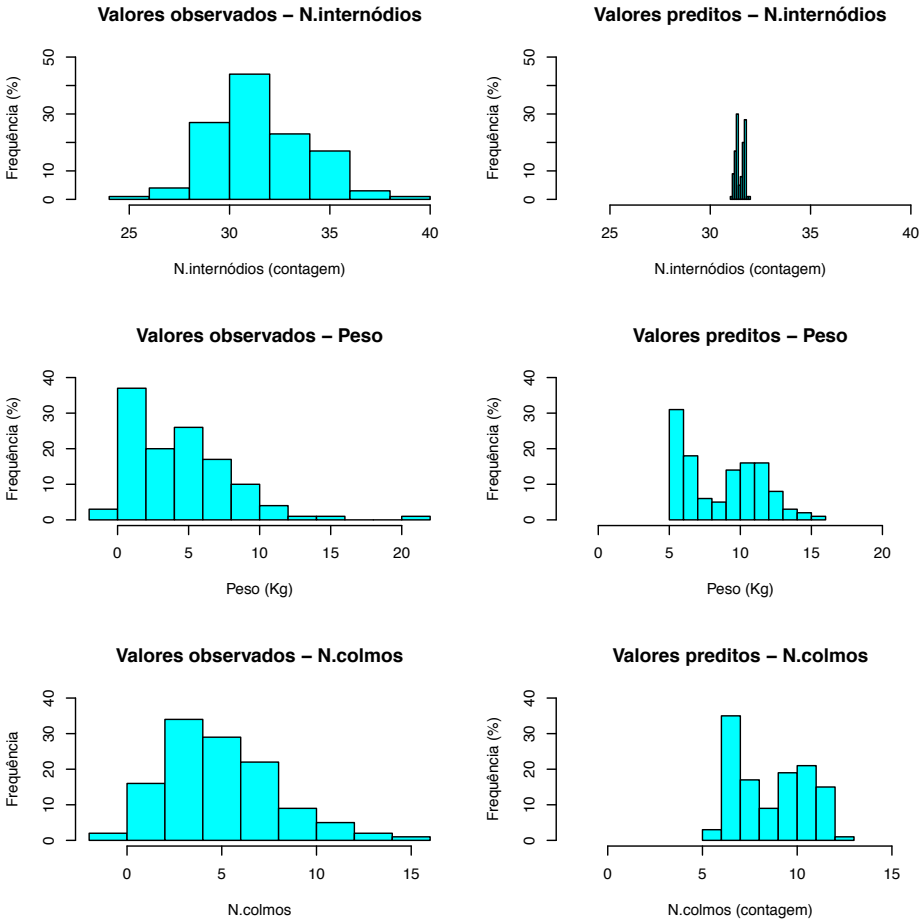

Figura 13. Histogramas representando a distribuição de frequência dos valores preditos pelo modelo de GWS aplicado e dos valores observados (eBLUPs) para as

características número de internódios, peso e numero de colmos, evidenciando o efeito *shrinkage*.

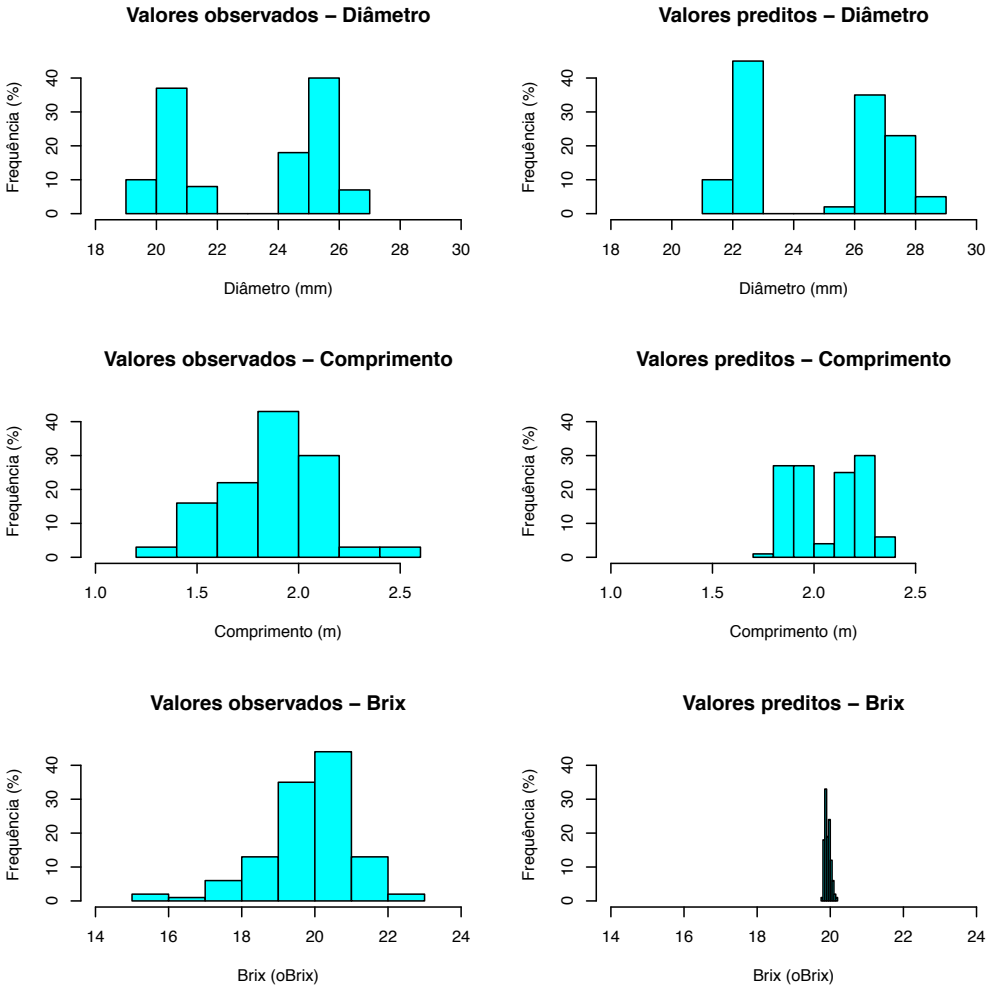

Figura 14. Histogramas representando a distribuição de frequência dos valores preditos pelo modelo de GWS aplicado e dos valores observados (eBLUPs) para as características diâmetro, comprimento e ^{o}brix, evidenciando o efeito *shrinkage*.

Os diferentes métodos para a obtenção de modelos de Seleção Genômica Ampla fazem uso de diferentes matrizes de parentesco. O RR-BLUP não requer o cálculo de uma matriz de parentesco médio, visto que para a resolução do modelo é necessária a obtenção da matriz Z'Z e esta é a matriz de parentesco verdadeira, em que Z é a matriz de incidência para o efeito dos marcadores, ou seja, para os efeitos aleatórios (Schaeffer, 2006; Resende et al., 2008).

Mesmo a matriz Z'Z sendo uma matriz real de parentesco, considerada mais

eficiente que do que uma matriz arbitrária, por capturar os reais valores de parentesco para cada marca e não assumir uma matriz de parentesco médio relacionada ao pedigree (Resende et al., 2010), o efeito da estruturação devido ao sistema de cruzamento interferiu fortemente nos resultados obtidos. Isto evidencia que a matriz utilizada para o desenvolvimento do modelo com o uso de programas estabelecidos, como é o caso do rrBLUP, pode não ser eficiente na retirada de efeitos de estruturação tão fortes como os existentes na população em estudo.

Mesmo tendo sido obtidos modelos com altas acurácias para a maior parte dos caracteres estudados, é importante ressaltar que estas acurácias podem estar infladas com o efeito da estruturação existente na população de treinamento. Neste caso, elas devem refletir também o efeito de associações falso-positivas entre os alelos dos diferentes locos.

Se este for o caso, estas associações deverão se desfazer rapidamente ao longo das próximas gerações de recombinação, reduzindo drasticamente a utilidade do modelo de predição em gerações futuras. O que exigiria uma nova etapa de fenotipagem e re-calibração do modelo de GWS.

Segundo Gupta et al. (2005) e também Nordborg & Weigel (2008) é necessário considerar a estruturação populacional para inferências sobre o LD, pois este fator interfere grandemente na sua mensuração. A existência de associações falso-positivas entre locos não ligados é um dos problemas que ocorrem quando não é realizado o controle efetivo da influência deste fator sobre o LD (Rafalski, 2002), exatamente como pode acontecer nas análises tradicionais de desenvolvimento de modelos de GWS. A formação de grupos concisos nos gráficos de distribuição relacionando os valores de eBLUPs observados e os VGG preditos sugere claramente este efeito. As diferenças existentes são visíveis entre os eBLUPs dos indivíduos da população de cruzamento e os dos obtidos por autofecundação e há a possibilidade de existirem locos capazes de captar exatamente estas diferenças. Neste caso, os alelos destes locos, presentes em uma população e ausentes na outra, podem capturar os efeitos de grande parte dessa variação, permitindo sua “explicação”, ou seja, a explicação da variação entre as populações, mesmo na ausência de ligação física com os polimorfismos causais.

5 CONCLUSÕES

Com base nos resultados obtidos, as seguintes conclusões podem ser estabelecidas:

(i) Foi implementado o modelo de simulação de Raboin et al. (2008), para avaliação do decaimento do desequilíbrio de ligação em diferentes cenários. A utilização deste modelo deverá permitir a realização de estudos posteriores de caracterização do desequilíbrio de ligação em espécies poliploides.

(ii) A população de cana-de-açúcar do presente estudo apresenta uma elevada magnitude de desequilíbrio de ligação, que se estendeu até 30cM.;

(iii) Os estudos desenvolvidos abriram a possibilidade teórica de se identificar a dosagem alélica nos genitores com base na relação entre o LD e a distância estimada entre os locos;

(iv) A estratégia de seleção genômica ampla tem grande potencial de aplicação em programas de melhoramento genético de cana-de-açúcar.

(v) Estudos posteriores devem ser desenvolvidos para investigar o efeito da estruturação genética presente nas populações de treinamento na acurácia de modelos de seleção genômica ampla e sua persistência ao longo das gerações.

6 REFERÊNCIAS BIBLIOGRÁFICAS

AGRAMA, H. A. & EIZENGA, G. C. Molecular diversity and genome-wide linkage disequilibrium patterns in a worldwide collection of *Oryza sativa* and its wild relatives. **Euphytica,** Wageningen, v. 160, n. 3, p. 339-355, 2007.

ALJANABI, S. M.; FORGET, L.; DOOKUN, A. An improved and rapid protocol for the isolation of polysaccharide-and polyphenol-free sugarcane DNA. **Plant Molecular Biology Reporter**, Amsterdam, v. 17, n. 3, p. 281-281, 1999.

BEARZOTI, E. Mapeamento de QTL. In: PINHEIRO, J. B. & CARNEIRO, I. F **Análise de QTL no melhoramento de plantas. Goiânia: FUNAPE**, 2000, p. 63-223.

BENNETT, J. H.; BINET, F. E. Association between Mendelian factors with mixed selfing and random mating. **Heredity**, New York, v. 10, n. 1, p. 51-55, 1956.

BHAT, S. R. & GILL, B. S. The implication of 2n egg gametes in nobilisation and breeding of sugarcane. **Euphytica,** Wageningen**,** n.34, p. 377–384, 1985.

BRASIL, MINISTÉRIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO (MAPA). **Produção e produtividade médias de lavouras temporárias e permanentes – Cana-de-Açúcar**. 2012.Disponível em: http://www.agricultura.gov.br. Acesso em: 23 jul. 2012.

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

BURR, G. O.; HARTT, C. E.; BRODIE, H. W.; TANIMOTO, T.; KORTSCHAK, H. P.; TAKAHASHI, D.; ASHTON, F. M.; COLEMAN, R. E. The sugarcane plant. **Annual Review of Plant Physiology**, Califórnia, v. 8, n. 1, p. 275-308, 1957.

CARNEIRO, M. S. & VIEIRA, M. L. C. Mapas genéticos em plantas. **Bragantia**, Campinas - SP, v. 61, n. 2, p. 89-100, ago. 2002.

CESNIK, R.; MIOCQUE, J. **Melhoramento da cana-de-açúcar: marco sucro-alcooleiro no Brasil**. Brasília: Embrapa Informações Tecnológicas, 2004. 307 p.

CHANDRAN, K. In vitro regeneration of Saccharum edule from immature inflorescence. **Sugar Tech**, New York v. 13, n. 2, p. 170-173, 2011.

CHAO, S.; DUBCOVSKY, J.; DVORAK, J.; LUO, M. C.; BAENZIGER, S. P.; MATNYAZOV, R.; CLARK, D. R.; TALBERT, L. E.; ANDERSON, J. A.; DREISIGACKER, S.; GLOVER, K.; CHEN, J.; CAMPBELL, K.; BRUCKNER, P. L.; RUDD, J. C.; HALEY, S.; CARVER, B. F. PERRY, S.; SORRELLS. M. E.;

AKHUNOV, E. D. Population-and genome-specific patterns of linkage disequilibrium and SNP variation in spring and winter wheat (*Triticum aestivum* L.). **BMC genomics**, Londres, v. 11, n. 1, p. 727, 2010.

CHAVES, L. J. Interação de genótipos com ambientes. In: NASS, L. L., **Recursos genéticos e melhoramento de plantas. Rondonopólis: Fundação MT**, p. 675-712, 2001.

CHEAVEGATTI-GIANOTTO, A.; DE ABREU, H. M. C.; ARRUDA, P.; BESPALHOK FILHO, J. C.; BURNQUIST, W. L.; CRESTE, S.; DI CIERO, L.; FERRO, J. A.; FIGUEIRA, A. V. O.; FILGUEIRAS, T. S.; GROSSI-DE-SÁ, M. F.; GUZZO, E. C.; HOFFMANN, H. P.; LANDELL, M. G. A.; MACEDO, N.; MATSUOKA, S.; REINACH, F. C.; ROMANO, E.; SILVA, W. J.; SILVA FILHO, M. C.; ULIAN, E. C. Sugarcane (Saccharum X officinarum): a reference study for the regulation of genetically modified cultivars in Brazil. **Tropical plant biology**, Aiea, Hawaii, v. 4, n. 1, p. 62-89, 2011.

CHING, A. D. A.; CALDWELL, K. S.; JUNG, M.; DOLAN, M.; SMITH, O. S. H.; TINGEY, S.; MORGANTE, M.; RAFALSKI, A. J. SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. **BMC genetics**, Londres, v. 3, n. 1, p. 19, 2002.

COMADRAN, J.; RAMSAY, L.; MACKENZIE, K.; HAYES, P.; CLOSE, T. J.; MUCHLBAUER, G.; STEIN, N.; WAUGH, R. Patterns of polymorphism and linkage disequilibrium in cultivated barley. **Theoretical and applied genetics**, Heidelberg, v. 122, n. 3, p. 523-531, 2011.

CONAB, **Companhia Nacional de Abastecimento. Acompanhamento de safra brasileira : Cana-de-açúcar, quarto levantamento, abril/2013** - Companhia Nacional de Abastecimento. Brasília: Conab 2013. Disponível em:< http://www.conab.gov.br>. Acesso em: 20 nov. 2012.

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

DAETWYLER, H.D.; VILLANUEVA, B.; BIJMA, P., WOOLLIAMS, J. A. Inbreeding in genome-wide selection. **Journal of Animal Breeding and Genetics**, Massachusetts, v. 124, n. 6, p. 369-376, 2007.

DANIELS, J.; DANIELS, C. A. Georgraphical, historical and cultural aspects of the origin of the Indian and Chinese sugarcanes *Saccharum barberi* and *Saccharum sinense.* **Sugarcane Breeders' newsletter**, Austrália, v. 36, p. 4 - 23, 1975.

DANIELS, J.; SMITH, P.; PATON, N. The origin of sugarcanes and centers of genetic diversity in *Saccharum.* **Sugarcane Breeder's Newsletter**, Austrália, n. 35, p. 418. In: D'HONT, A., ISON, D., ALIX, K., ROUX, C., & GLASZMANN, J. C. Determination of basic chromosome numbers in the genus Saccharum by physical mapping of ribosomal RNA genes. **Genome**, Birmingham, v. 41, n. 2, p. 221-225,

1998.

DANIELS, J.; ROACH, B. T. Taxonomy and evolution. In: Heinz DJ (ed) **Sugarcane improvement through breeding**. Elsevier, Amsterdam, v. 7, p. 7-84, 1987.

DESTA, Z. A.; ORTIZ, R. Genomic selection: genome-wide prediction in plant improvement. **Trends in plant science**, v. 19, n. 9, p. 592-601, 2014.

DE LOS CAMPOS, G.; HICKEY, J. M.; PONG-WONG, R.; DAETWYLER, H. D.; CALUS, M. P. L. Whole-genome regression and prediction methods applied to plant and animal breeding. **Genetics**, Pittsburg, v. 193, n. 2, p. 327-345, 2013.

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

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

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

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

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**, Heidelberg, v. 121, n. 6, p. 1151-1170, 2010.

DURÃES, F. O. M.; SANTOS, M. X. dos; GAMA, E. G.; MAGALHÃES, P. C.; ALBUQUERQUE, P. E. P.; GUIMARÃES, C. T. 2004. **Fenotipagem associada à tolerância a seca em milho para uso em melhoramento, estudos genômicos e seleção assistida por marcadores.** Sete Lagoas: Embrapa Milho e Sorgo, 17 f. Diponível em:< http://www.cnpms.embrapa.br/publicacoes/publica/2004/circular/Circ_39.pdf>. Acesso em: 10 ago. 2012.

EARL, D. A. & VONHOLDT, B. M. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. **Conservation Genetics Resources**, Holanda, v. 4, n. 2, p. 359-361, 2012.

ENDELMAN, Jeffrey B. Ridge regression and other kernels for genomic selection with R package rrBLUP. **The Plant Genome**, Madison, v. 4, n. 3, p. 250-255, 2011.

EVANNO, G.; REGNAUT, S.; GOUDET, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. **Molecular ecology**, Massachusetts, v. 14, n. 8, p. 2611-2620, 2005.

FALEIRO, F. G. **Marcadores Genético-Moleculares aplicados a programas de conservação e uso de recursos genéticos**. Planaltina: Embrapa Cerrados, 2007. 102 p.

FAO – Food and Agriculture Organization. **FAO Statistical Yearbook 2013 - World food and agriculture**. Roma, 2013. Disponível em: < http://www.fao.org>. Acesso em: 12 nov. 2012.

FEDERER, W. T. Augmented (or hoonuiaku) designs. **Hawaiian Planter's Records**, Hawaii, v. 55, p. 191- 208, 1956.

FERNANDO, R.L.; GROSSMAN, M. Marker Assisted Selection using Best Linear Unbiased Prediction. **Genetics Selection Evolution**, Londres, n. 21, p. 467-477, 1989.

FERREIRA, M.E.& GRATTAPAGLIA, D. **Introdução ao uso de marcadores moleculares em análise genética.** Brasília: Embrapa Recursos Genéticos e Biotecnologia, 1998. 220 p.

FIGUEIREDO, P. Breve história da cana-de-açúcar e do papel do instituto agronômico no seu estabelecimento no Brasil. In: DINARDO-MIRANDA, L. L.; VASCONCELOS, A. C. M.; LANDELL, M. G. A. **Cana-de-açúcar**. 1. ed. Campinas: IAC, 2008, cap. 1, p. 31-46.

FLINT-GARCIA, S. A.; THORNSBERRY, J. M.; BUCKLER, I.V. Structure of Linkage Disequilibrium in Plants. **Annual Review of Plant Biology**, Califórnia, v. 54, n. 1, p. 357-374, 2003.

FRITSCHE-NETO, R.; RESENDE, M. D. V.; MIRANDA, G. V.; DO VALE, J. C. Seleção genômica ampla e novos métodos de melhoramento do milho. **Revista Ceres**, Viçosa – MG, v. 59, n. 6, 2012.

GARCIA, A. A .; MOLLINARI, M.; MARCONI, T. G.; SERANG, O. R; SILVA, R. R.; VIEIRA, M. L. C.; VICENTINI, R.; COSTA, E. A.; MANCINI, M. C.; GARCIA, M. O. S.; PASTINA, M. M.; GAZAFFI, R.; MARTINS, E. R. F.; DAHMER, N.; SFORÇA, D. A.; SILVA, C. B. C.; BUNDOCK, P.; HENRY, R. J.; SOUZA, G. M.; SLUYS, M.-A.; LANDELL, M. G. A.; CARNEIRO, M. S.; VICENTZ, M. A. G.; PINTO, L. R.; VENCOVSKY, R.; SOUZA, A. P. SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. **Scientific reports**, Londres, v. 3, n. 3399, Dez. 2013.

GARRIS, A. J.; MCCOUCH, S. R.; KRESOVICH, S. Population structure and its effect on haplotype diversity and linkage disequilibrium surrounding the xa5 locus of rice (Oryza sativa L.). **Genetics**, Pittsburg, v. 165, n. 2, p. 759-769, 2003.

GAUT, B. S. & LONG, A. D. The lowdown on linkage disequilibrium. **The Plant**

Cell, Waterbury, v. 15, p. 1502-1506, Jul. 2003.

GAWENDA, I.; SCHRÖDER-LORENZ, A.; DEBENER, T. Markers for ornamental traits in Phalaenopsis orchids: population structure, linkage disequilibrium and association mapping. **Molecular Breeding**, Holanda, v. 30, n. 1, p. 305-316, 2012.

GODDARD, M. E.; HAYES, B. J. Genomic selection. **Journal of Animal Breeding and Genetics,** Massachusetts, v. 124, p. 323-330, 2007.

GOLDSTEIN, D. B. Islands of linkage disequilibrium. **Nature Genetics**, New York, n. 29, p. 109-11, 2001.

GRIFFITHS, A. J. F.; WESSLER, S. R.; LEWONTIN, R. C.; GELBART, W. M.; SUZUKI, D. T. **Introdução à genética**. Tradução de Paulo Armando Motta. 9. ed. Rio de Janeiro: G. Koogan, 2008. 726 p.

GRIVET, L. & ARRUDA, P. Sugarcane genomics: depicting the complex genome of an important tropical crop. **Current Opinion in Plant Biology**, Philadelphia, v. 5, n. 2, p. 122-127, 2002.

GRIVET, L.; D'HONT, A.; ROQUES, D.; FELDMANN, P.; LANAUD, C.; GLASZMANN, C. RFLP mapping in cultivated sugarcane (Saccharum spp.): genome organization in a highly polyploid and aneuploid interspecific hybrid. **Genetics**, Pittsburg, v. 142, n. 3, p. 987-1000, 1996.

GRIVET, L.; GLASZMANN, J. C.; D'HONT, A. Molecular evidence of sugarcane evolution and domestication. **Darwin's harvest: New approaches to the origins, evolution and conservation of crops.** Columbia University Press: New York, p. 49-66, 2006.

GUPTA, P. K.; RUSTGI, S.; KULWAL, P. L. Linkage disequilibrium and association studies in higher plants: Present status and future prospects. **Plant molecular Biology**, Amsterdam, n. 57, p. 461- 485, jan. 2005.

HAO, C.; WANG, L.; GE, H.; DONG, Y.; ZHANG, X. Genetic diversity and linkage disequilibrium in Chinese bread wheat (*Triticum aestivum* L.) revealed by SSR markers. **PLoS One**, Califórnia, v. 6, n. 2, p. e17279, 2011.

HARTL, D. L. & CLARK, A. G. **Princípios de genética de populações**. Porto Alegre: Editora Artmed, 4ª edição, 659pp, 2010.

HEDRICK, P. W. **Genetics of Population**. 4a ed. USA: Jones and Bartlett Publishers, Inc., 2011. 675 p.

HEDRICK, P. W. Gametic Disequilibrium Measures: Proceed With Caution. **Genetics,** Pittsburg, n. 117. p. 331-341. Oct, 1987.

HEFFNER, E. L.; SORRELLS, M. E.; JANNINK, J.. Genomic selection for crop improvement. **Crop Science**, Madison, v. 49, n. 1, p. 1-12, 2009.

HESLOT, N.; YANG, H-P.; SORRELLS, M. E.; JANNINK, J-L. Genomic selection in plant breeding: a comparison of models. **Crop Science**, Madison, v. 52, n. 1, p.

146-160, 2012.

HYTEN, D. L.; CHOI, I. Y.; SONG, Q.; SHOEMAKER, R. C.; NELSON, R. L.; COSTA, J. M.; SPECHT, J. E.; CREGAN, P. B. Highly variable patterns of linkage disequilibrium in multiple soybean populations. **Genetics**, Pittsburg, v. 175, n. 4, p. 1937-1944, 2007.

JACCOUD, D.; PENG, KAIMAN; FEINSTEIN, D.; KILIAN, A. Diversity arrays: a solid state technology for sequence information independent genotyping. **Nucleic acids research**, Oxford, v. 29, n. 4, p. e25-e25, 2001.

JAKOBSSON, M. & ROSENBERG, N. A. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. **Bioinformatics,** Oxford, n. 23, p. 801-1806, 2007.

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

JANNOO, N. et al. Linkage disequilibrium among modern sugarcane cultivars. **Theoretical and Applied Genetics**, Heidelberg, v. 99, n. 6, p. 1053-1060, 1999b.

JANNOO, N., GRIVET, L., DAVID, J., D'HONT, A., & GLASZMANN, J. C. Differential chromosome pairing affinities at meiosis in polyploid sugarcane revealed by molecular markers. **Heredity**, New York, v. 93, n. 5, p. 460-467, 2004.

JANNOO, N., GRIVET, L., SEGUIN, M., PAULET, F., DOMAINGUE, R., RAO, P. S.; GLASZMANN, J. C. Molecular investigation of the genetic base of sugarcane cultivars. **Theoretical and Applied Genetics**, Heidelberg, v. 99, n. 1-2, p. 171-184, 1999a.

JORDE, L. B. Reply to Zapata and Visedo. **American Journal of Human Genetics**, Massachusetts, v. 57, n. 1, p. 191, 1995.

KIM, S.; PLAGNOL, V.; HU, T. T.; TOOMAJIAN, C.; CLARK, R. M.; OSSOWSKI, S.; ECKER, J. R.; WEIGEL, D.; NORDBORG, M. Recombination and linkage disequilibrium in Arabidopsis thaliana. **Nature genetics**, New York, v. 39, n. 9, p. 1151-1155, 2007.

KIMURA, M. A model of a genetic system which leads to closer linkage by natural selection. **Evolution**, Massachusetts, p. 278-287, 1956.

KOJIMA, K. The distribution and comparison of “genetic loads” under heterotic selection and simple frequency-dependent selection in finite populations. **Theoretical Population Biology**, Philadelphia, v. 2, n. 2, p. 159-173, 1971.

KRAAKMAN, A.T. W.; NIKS, R. E.; VAN DEN BERG,P. M. M. M.; STAM P.; VAN EEUWIJK, F. A. Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. **Genetics**, Heidelberg, v. 168, n. 1, p. 435-446, 2004.

LEWIN, B. **Genes VIII**. Upper Saddle River: Pearson Prentice Hall, 2004. 988 p.

LEWONTIN, R. C. & KOJIMA, K. The Evolutionary Dynamics of Complex Polymorphisms. **Evolution**, Massachusetts, v. 14, n. 4, p. 458-472, dez. 1960.

LEWONTIN, R. C. On measures of gametic disequilibrium. **Genetics**, Heidelberg, v. 120, n. 3, p. 849-852, 1988.

LEWONTIN, R. C. The interaction of selection and linkage. I. General considerations; Heterotic Models. **Genetics**, Heidelberg, n. 49, p. 49-67, jan. 1964.

LI, Y.; GUAN, R.; LIU, Z.; MA, Y.; WANG, L.; LI, 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 (*Glycine max* (L.) Merr.) landraces in China. **Theoretical and Applied Genetics**, Heidelberg, 117, n. 6, p. 857-871, 2008.

LOPES, F. C. C. **Mapeamento genético de cana-de-açúcar (*Saccharum* spp.) por associação empregando marcadores SSR e AFLP**. 2011. 140 f. Tese (Doutorado em Genética e Melhoramento de Plantas)–Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2011.

LORENZ, A.J. Resource Allocation for Maximizing Prediction Accuracy and Genetic Gain of Genomic Selection in Plant Breeding: A Simulation Experiment. **G3: Genes, Genomes, Genetics**, Maryland, v. 3, n. 3, p. 481-491, 2013.

LORENZANA, R. E.; BERNARDO, R. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. **Theoretical and applied genetics**, Heidelberg, v. 120, n. 1, p. 151-161, 2009.

MACKAY, I. & POWELL, W. Methods for linkage disequilibrium mapping in crops. **Trends in Plant Science**, Massachusetts, v. 12, n. 2, p. 57-63, 2007.

MANGIN, B.; SIBERCHICOT, A.; NICOLAS, S.; DOLIGEZ, A.; THIS, P.; CIERCO-AYROLLES, C. Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. **Heredity**, New York, v. 108, n. 3, p. 285-291, 2011.

MATHER, 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**, Heidelberg, v. 177, n. 4, p. 2223-2232, 2007.

MENOSSI, M., SILVA-FILHO, M. C., VINCENTZ, M., VAN-SLUYS, M. A., & SOUZA, G. M. Sugarcane functional genomics: gene discovery for agronomic trait development. **International journal of plant genomics**, New York, v. 2008, n. 458732, 11 p. 2008.

MEUWISSEN, T. H. E; GODDART, M. E. The use of marker haplotypes in animal breeding schemes. **Genetics Selection Evolution**, v. 28, n. 2, p. 161-176, 1996.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDART, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, Heidelberg, v. 157, n. 4, p.

1819-1829, 2001.

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

NORDBORG, M.; HU, T. T.; ISHINO, Y.; JHAVERI1, J.; TOOMAJIAN, C.; ZHENG, H.; BAKKER, E.; CALABRESE, P.; GLADSTONE, J.; GOYAL, R.; JAKOBSSON, M.; KIM, S.; MOROZOV, Y.; PADHUKASAHASRAM, B.; PLAGNOL, V.; ROSENBERG, N. A.; SHAH, C.; WALL, J. D.; WANG, J.; ZHAO, K.; KALBFLEISCH, T.; SCHULZ, V.; KREITMAN, M.; BERGELSON, J. The pattern of polymorphism in Arabidopsis thaliana. **PLoS biology**, Grã-Bretanha, v. 3, n. 7, p. e196, 2005.

NORDBORG, M.; BOREVITZ, J. O.; BERGELSON, J.; BERRY, C. C.; CHORY, J.; HAGELBLAD, J.; KREITMAN, M.; MALOOF, J. N.; NOYES, T.; OEFNER, P. J.; STAHL, E. A.; WEIGEL, D. The extent of linkage disequilibrium in Arabidopsis thaliana. **Nature genetics**, New York, v. 30, n. 2, p. 190-193, 2002.

NORDBORG, M.; WEIGEL, D. Next-generation genetics in plants. **Nature**, New York, v. 456, n. 7223, p. 720-723, 2008.

PIEPHO, H. P.; OGUTU, J. O.; SCHULZ-STREECK, T.; ESTAGHVIROU, B.; GORDILLO, A.; TECHNOW, F. Efficient computation of ridge-regression best linear unbiased prediction in genomic selection in plant breeding. **Crop Science**, Madison, v. 52, n. 3, p. 1093-1104, 2012.

PIMENTEL, D. & PATZEK, T. W. Ethanol production: energy and economic issues related to US and Brazilian sugarcane. In: **Biofuels, Solar and Wind as Renewable Energy Systems**. Springer: Holanda, p. 357-371, 2008.

PIMENTEL, D.; PATZEK, T. W. Ethanol production using corn, switchgrass, and wood; biodiesel production using soybean and sunflower. **Natural resources research**, New York, v. 14, n. 1, p. 65-76, 2005.

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

PRICE, Sam. Cytogenetics of modern sugar canes. **Economic Botany**, New York, v. 17, n. 2, p. 97-106, 1963.

PRICE, S. **Cytology of Saccharum robustum and related symapatric species and natural hybrids**. Washington,: US Dept. of Agriculture, n. 1337, 47 p., jun. 1965. Disponível em:< http://books.google.com.br/books>. Acesso em: 22 dez. 2012

PRITCHARD, J. K.; STEPHENS, M.; DONNELLY, P. Inference of Population Structure Using Multilocus Genotype Data. **Genetics**, Heidelberg, v. 155, p. 945-959.

2000.

R CORE TEAM. R: **A language and environment for statistical computing. R Foundation for Statistical Computing.** Vienna, Austria, 2014. Disponível em: <http://www.R-project.org/>. Acesso em 8 jun. 2012.

RABOIN, L. M.; PAUQUET, J.; BUTTERFIELD, M.; D'HONT, A.; GLASZMANN, J. C. Analysis of genome-wide linkage disequilibrium in the highly polyploidy sugarcane. **Theoretical and Applied Genetics**, Heidelberg, v. 116, n. 5, p. 701-714, jan. 2008.

REMINGTON, D. L.; THORNSBERRY, J. M.; MATSUOKA, Y.; WILSON, L. M.; WHITT, S. R.; DOEBLEY, J.; KRESOVICH, S.; GOODMAN, M. M.; BUCKLER, E. S. Structure of linkage disequilibrium and phenotypic associations in the maize genome. **Proceedings of the National Academy of Sciences**, Washington, v. 98, n. 20, p. 11479-11484, 2001.

RESENDE, M. D. V. **Genética biométrica e estatística no melhoramento de plantas perenes.** Brasília: Embrapa Informação Tecnológica, 2002. 975 p.

RESENDE, M. D. V. **Genética Quantitativa e Seleção no Melhoramento de Plantas Perenes e Animais**. Colombo: Embrapa Florestas, 2008. 330 p.

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

RESENDE, M. D.V.; RESENDE JÚNIOR, M. F. R.; AGUIAR, A. M.; ABAD, J. I. M.; MISSIAGGIA, A. A.; SANSALONI, C.; PETROLI, C.; GRATTAPAGLIA, D. **Computação da seleção genômica ampla (GWS).** 1ª ed. Colombo: Embrapa Florestas 2010. 79p.

RITLAND, K. & HEDRICK, P. W. Gametic disequilibrium and multilocus estimation of selfing rates. **Heredity**, New York, v. 65, p. 343-347, 1990.

RODRIGUES, D. & ORTIZ, L. Em direção à sustentabilidade da produção de etanol de cana de açúcar no Brasil. **Porto Alegre, Brasil: Amigos da Terra Brasil**, 2006. Disponível em: <http://www.ecoa.org.br/arquivos/444052181.pdf>. Acessado em: 24 de ago. 2012.

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

ROSENBERG, Noah A. DISTRUCT: a program for the graphical display of population structure. **Molecular Ecology Notes**, Nova Jersey, v. 4, n. 1, p. 137-138, 2004.

SCHAEFFER, L. R. Strategy for applying genome-wide selection in dairy cattle. **Journal of Animal Breeding and Genetics**, Massachusetts, v. 123, n. 4, p. 218-223,

2006.

SCHAPER, E.; ERIKSSON, A.; RAFAJLOVIC, M.; SAGITOV, S.; MEHLIG, B. Linkage disequilibrium under recurrent bottlenecks. **Genetics**, Heidelberg, v. 190, n. 1, p. 217-229, 2012.

SCHAPER, E.; ERIKSSON, A.; RAFAJLOVIC, M.; SAGITOV, S.; MEHLIG, B. Linkage disequilibrium under recurrent bottlenecks. **Genetics**, Heidelberg, v. 190, n. 1, p. 217-229, 2012.

SEBASTIANI, Paola; ABAD-GRAU, María M. Bayesian estimates of linkage disequilibrium. **BMC genetics**, New York, v. 8, n. 1, p. 36, 2007.

SIMS, R. E., HASTINGS, A., SCHLAMADINGER, B., TAYLOR, G., & SMITH, P. Energy crops: current status and future prospects. **Global Change Biology**, Massachusetts, v. 12, n. 11, p. 2054-2076, 2006.

SLATKIN, M. Linkage disequilibrium in growing and stable populations. **Genetics**, Heidelberg, v. 137, n. 1, p. 331-336, 1994.

SOKAL, R. R. & ROHLF, F. J. Biometry: the principles and practice of statistics in biological research. 2a ed., New York: **WH Feeman and Company**, 1981.

SOLBERG, T. R.; SONESSON, A. K.; WOOLLIAMS, J. A.; MEUWISSEN, T. H.. Genomic selection using different marker types and densities. **Journal of Animal Science**, v. 86, n. 10, p. 2447-2454, 2008.

SOUZA, C. B. & MIZIARA, F. Políticas de Financiamento à expansão do setor sucroalcooleiro em Goiás versus Políticas Ambientais**. In: Congresso Latino-americano de Sociologia Rural**, Porto de Galinhas, 2010. Disponível em: <http://www.alasru.org/wp-content/uploads/2011/08/GT12-Cleonice-Borges-de-Souza.pdf>. Acesso em: 25 ago. 2012.

STICH, B.; URBANY, C.; HOFFMANN, P.; GEBHARDT, C. Population structure and linkage disequilibrium in diploid and tetraploid potato revealed by genome-wide high-density genotyping using the SolCAP SNP array. **Plant Breeding**, Massachusetts, v. 132, n. 6, p. 718-724, 2013.

STRACKE, S.; PEROVIC, D.; STEIN, N.; THIEL, T.; GRANER, A. Linkage disequilibrium in barley. In: **11th Molecular Markers Symposium of the GPZ.** 2003. 11th Molecular Markers Symposium of the GPZ. Disponível em:<http://meetings.ipkgatersleben.de/ moma2003/index.php>. Acessado em 10 jan. 2013.

TENAILLON, M. I.; SAWKINS, M. C.; LONG, A. D.; GAUT, R. L.; DOEBLEY, J. F. Patterns of DNA sequence polymorphism along chromosome 1 of maize (Zea mays ssp. mays L.). **Proceedings of the National Academy of Sciences**, Washington, v. 98, n. 16, p. 9161-9166, 2001.

THEODORO, A. D. **Expansão da cana-de-açúcar no Brasil: ocupação da cobertura vegetal do Cerrado**. 2001. 60 f. Trabalho (Graduação: Curso de

Tecnologia em Biocombustíveis)–Faculdade de Tecnologia de Araçatuba, Araçatuba, 2001.

TRUNTZLER, M.; RANC, N.; SAWKINS, M. C.; NICOLAS, S.; MANICACCI, D.; LESPINASSE, D.; RIBIÈRE, V.; GALAUP, P.; SERVANT, F.; MULLER, C.; MADUR, D.; BETRAN, J.; CHARCOSSET, A.; MOREAU, L. Diversity and linkage disequilibrium features in a composite public/private dent maize panel: consequences for association genetics as evaluated from a case study using flowering time. **Theoretical and Applied Genetics**, Heidelberg, v. 125, n. 4, p. 731-747, 2012.

TZVELEV, N. N. The system of grasses (Poaceae) and their evolution. **The Botanical Review**, New York, v. 55, n. 3, p. 141-203, 1989.

USDA. Foreign Agricultural Service: Sugar: **World Markets and Trade. Nov, 2013**. Disponível em: <http://usda01.library.cornell.edu/usda/current/sugar/sugar-11-21-2013.pdf> Acesso em: 3 de Dez. 2013.

VASCONCELOS, Paulo Henique Castanheira. Uma História feita de Açúcar e Álcool: apontamentos para uma discussão. **Revista Crase. edu**, v. 1, n. 1, p. 61 - 68, 2010.

WAACK, R. S.; NEVES, M. F.; MARINO, M. K.; MAMONE, A.; SZASZ, A. H. Competitividade do Sistema Agroindustrial da Cana-de-Açúcar. **Competitividade no agribusiness brasileiro: versão final. São Paulo: Instituto de Pesquisa Econômica Aplicada/USP**, v. 5, p. 1-194, jul. 1998.

WANG, Y.-H.; UPADHYAYA, H. D.; BURREL, A. M.; SAHRAEIAN, S. M. E.; KLEIN, R. R.; KLEIN, P. E. Genetic structure and linkage disequilibrium in a diverse, representative collection of the C4 model plant, Sorghum bicolor. **G3: Genes, Genomes, Genetics**, Maryland, v. 3, n. 5, p. 783-793, 2013.

WAPLES, R. S.; ENGLAND, P. R. Estimating contemporary effective population size on the basis of linkage disequilibrium in the face of migration. **Genetics**, Heidelberg, v. 189, n. 2, p. 633-644, 2011.

WEIR, B. S. & COCKERHAM, C. C. Behavior of pairs of loci in finite monoecious populations. **Theoretical population biology**, Philadelphia, v. 6, n. 3, p. 323-354, 1974.

WEI, X.; JACKSON, P. A.; MCINTYRE, C. L.; AITKEN, K. S.; CROFT, B. Associations between DNA markers and resistance to diseases in sugarcane and effects of population substructure. **Theoretical and Applied Genetics**, Heidelberg, v. 114, n. 1, p. 155-164, 2006.

WEIR, B. S. Inferences about linkage disequilibrium. **Biometrics**, Texas, v. 35, n. 1, p. 235-254, mar. 1979.

WHITTAKER, J.C.; R. THOMPSON; DENHAM, M.C. Marker-assisted selection using ridge regression. **Genetics Research Cambrige**, Inglaterra, n. 75, p. 249–252, 2000.

WICKHAM, H.. Reshaping Data with the reshape Package. **Journal of Statistical Software,** vol. 21, n. 12, p. 1-20, 2007. Disponível em: <http://www.jstatsoft.org/v21/i12/>. Acesso em: 20 out. 2013.

WICKHAM, H. ggplot2: elegant graphics for data analysis. **Springer, New York**, 2009.

WÜRSCHUM, T.; LANGER, S. M.; LONGIN, F. H.; KORZUN, V.; AKHUNOV, E.; EBMEYER, E.; SCHACHSCHNEIDER, R.; SCHACHT, J.; KAZMAN, E.; REIF, J. C. Population structure, genetic diversity and linkage disequilibrium in elite winter wheat assessed with SNP and SSR markers. **Theoretical and Applied Genetics**, Heidelberg, p. 1-10, 2012.

ZAPATA, C. & ALVAREZ, G. On Fisher's exact test for detecting gametic disequilibrium between DNA polymorphisms. **Annals of human genetics**, Massachusetts, v. 61, n. 1, p. 69-75, 1997.

ZAPATA, C. & VISEDO, G. Gametic disequilibrium and physical distance. **American journal of human genetics**, Massachusetts, v. 57, n. 1, p. 190, 1995.

ZAPATA, C. The D'measure of overall gametic disequilibrium between pairs of multiallelic loci. **Evolution**, Massachusetts, v. 54, n. 5, p. 1809-1812, 2000.

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

ZHANG, L. Y.; MARCHAND, S.; TINKER, N. A.; & BELZILE, F. Population structure and linkage disequilibrium in barley assessed by DArT markers. **Theoretical and applied genetics**, Heidelberg, v. 119, n. 1, p. 43-52, 2009.

ZHANG, Z.; DING, X.; LIU, J.; KONING, D-J.; ZHANG, Q. Genomic selection for QTL-MAS data using a trait-specific relationship matrix. In: **BMC proceedings**. BioMed Central, Londres, 2011. p. S15.

ZHOU, Hao; MUEHLBAUER, Gary; STEFFENSON, Brian. Population structure and linkage disequilibrium in elite barley breeding germplasm from the United States. **Journal of Zhejiang University SCIENCE B**, v. 13, n. 6, p. 438-451, 2012.

ZHU, Y. L.; SONG, Q. J.; HYTEN, D. L.; VAN TASSELL, C. P.; MATUKUMALLI, L. K.; GRIMM, D. R.; HYATT, S. M.; FICKUS, E. W.; YOUNG, N. D.; CREGAN, P. B. Single-nucleotide polymorphisms in soybean. **Genetics**, Heidelberg, v. 163, n. 3, p. 1123-1134, 2003.