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

# Análise do desequilíbrio de ligação e da estrutura populacional do germoplasma brasileiro de cana-de-açúcar

**João Ricardo Bachega Feijó Rosa**

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

**Piracicaba**
**2011**

**João Ricardo Bachega Feijó Rosa**
**Engenheiro Agrônomo**

# Análise do desequilíbrio de ligação e da estrutura populacional do germoplasma brasileiro de cana-de-açúcar

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

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

**Piracicaba**
**2011**

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

Rosa, João Ricardo Bachega Feijó
Análise do desequilíbrio de ligação e da estrutura populacional do germoplasma brasileiro de cana-de-açúcar / João Ricardo Bachega Feijó Rosa. - - Piracicaba, 2011.
97 p. : il.

Dissertação (Mestrado) - - Escola Superior de Agricultura "Luiz de Queiroz", 2011.

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

CDD 633.61
R788a

À minha avó materna, **Amélia** (*em memória*),
por todo seu amor, carinho e dedicação, além da
inesquecível convivência. À minha mãe, **Fátima**, pelo amor
incondicional e exemplo de perseverança e luta. À minha irmã, **Daniela**,
pela força e determinação nos momentos difíceis. Às três, por
fazerem o possível e o impossível para que eu chegasse até aqui,
e por sempre acreditarem em meus sonhos.
**Dedico**.

“Talvez não tenhámos conseguido fazer o melhor
Mas, lutamos para o que o melhor fosse feito
Não somos o que deveríamos ser, não somos o que iremos ser
Mas, graças a Deus, já não somos mais o que éramos.”
**Martin Luther King** (15/01/1929 - 04/04/1968)

“To every complex problem
there is a simple solution
And it is usually wrong.”
**Henry Louis Mencken**
(12/09/1880 - 29/01/1956)

“Se as coisas são inatingíveis...Ora!
Não é motivo para não querê-las.
Que tristes os caminhos se
não fora a presença distante das estrelas.”
**Mário Quintana** (30/07/1906 - 05/05/1994)

# Agradecimentos

Ao nosso Criador.

Ao Prof. Dr. Antonio Augusto Franco Garcia, que desde o princípio depositou total confiança em meu trabalho e me recebeu de braços abertos em seu laboratório. Sua orientação e amizade em tão pouco tempo foram cruciais para o meu caminhamento até aqui. Serei eternamente grato!

A todos os professores do Departamento de Genética, pelos ensinamentos e estímulos ao longo de todo o curso de mestrado.

À minha namorada Natália Masiero Volpe, que sempre me deu força e apoio nesses mais de 6 anos de convivência. Todo seu amor e carinho foram muito importantes nos momentos difíceis e decisivos para realização deste trabalho.

Ao meu cunhado Gilberto Bueno de Oliveira Junior, que foi um grande irmão ao longo de mais de 10 anos de convivência. Seu apoio incondicional e exemplo de humildade e perseverança foram muito importantes para a minha vida e realização desse trabalho.

À Profª. Drª. Monalisa Sampaio Carneiro, pelo fornecimento dos dados aqui utilizados e pela grandiosa amizade. Aos pesquisadores Fernanda Zatti Barreto e Thiago William de Almeida Balsalobre, que trabalharam incansavelmente, por noites e mais noites, na geração dos dados.

Aos amigos do Programa de Melhoramento Genético da Cana-de-açúcar da RIDESA/UFSCar: Hermann Hoffmann, Marcos Sanches, Antônio Bassinello, Marineide Aguillera, Antonio Ribeiro, Roberto Chapola, Danilo Cursi, Carlos Loureiro, Plínio Zavaglia, Claúdio Mendes, José Ciofi, José Adalberto da Cruz e outros que eventualmente possa ter esquecido. Trabalhar com vocês foi um grande estímulo para ter vindo estudar Genética e Melhoramento de Plantas na ESALQ/USP.

Ao Departamento de Genética da Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, pelo ensino brilhante e de grande qualidade.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq, pela bolsa concedida nos primeiros 5 meses do mestrado, e principalmente à Fundação de Amparo à Pesquisa do Estado de São Paulo - FAPESP, pela concessão da bolsa e demais recursos ao longo dos últimos 19 meses do curso até a conclusão deste trabalho.

Aos amigos do Laboratório de Genética Estatística: Rodrigo, Maria Marta, Gabriel, Marcelo, Graciela, Renato, Edjane, Carina, Luciano, Guilherme, Adriana e Rodrigo, pelos valiosos ensinamentos e pela agradável convivência em todos os momentos.

À Edjane, que abriu portas para o meu crescimento na época em que trabalhávamos juntos no Melhoramento de Cana-de-açúcar da UFSCar.

Ao Marcelo, que tanto me ajudou ao chegar em Piracicaba, um ambiente totalmente novo para mim. Não me esqueço das nossas conversas quase sempre na companhia de um ótimo “cafézinho”.

Ao Renato, pela extrema dedicação e paciência no momento da realização das análises deste trabalho. As agradáveis discussões sobre o assunto foram muito importantes para a construção das ideias aqui apresentadas.

Ao Marcelo, Rodrigo, Maria Marta e Gabriel, por terem desenvolvido e aprimorado o arquivo do LaTeX contendo todas as normas da dissertação.

À Carina, que tão companheira foi ao longo desses quase 2 anos de convívio. Disciplinas, provas, trabalhos, relatórios, mais provas...ufa! Mas, certamente valeu, e muito!

Ao grande amigo Fernando Guerra, companheiro desde os tempos da graduação. Ao Antônio Nogueira, que em tão pouco tempo se mostrou ser um grande parceiro.

Aos professores do Centro de Ciências Agrárias da Universidade Federal de São Carlos, que sempre me deram total apoio para continuar os estudos.

Aos funcionários do Departamento de Genética da ESALQ/USP: Seu Zé, Seu Antônio, Valdir, Berdan, Léia, Macedônio, Fernandinho, Maídia, Márcia e Glória, pelos grandiosos auxílios nos momentos decisivos.

A todas as pessoas, familiares e amigos, que direta ou indiretamente contribuíram para que eu chegasse até aqui, e pudesse realizar esse grande sonho. Muito obrigado!

# SUMÁRIO

# RESUMO

**Análise do desequilíbrio de ligação e da estrutura populacional do germoplasma brasileiro de cana-de-açúcar**

A cana-de-açúcar (*Saccharum* spp.) é uma cultura muito importante para a produção de açúcar e álcool no Brasil. Um importante avanço a respeito do seu genoma tem surgido com o emprego de marcadores moleculares, os quais têm sido extensamente utilizados para diversas finalidades, como o mapeamento de QTL's a partir de cruzamentos bi-parentais. Entretanto, o uso de populações oriundas de cruzamentos controlados tende a limitar os eventos de recombinação genética, gerando menor resolução de mapeamento. Nesse sentido, o mapeamento associativo surge como ferramenta promissora no estudo de QTL's, uma vez que podem ser utilizadas populações naturais, coleções de germoplasmas, conjunto de materiais elite, entre outros, possibilitando detectar eventos de recombinação em um contexto histórico-evolutivo. Para definir as melhores estratégias de mapeamento associativo, é fundamental conhecer o desequilíbrio de ligação (DL) e a estrutura genética presentes em determinada população, de modo a verificar o número de marcadores necessários e promover o controle de falsos positivos. Assim, este trabalho teve como objetivo analisar o DL e a estrutura populacional ao longo do genoma de variedades comerciais e clones de interesse para o melhoramento da cana-de-açúcar no Brasil, com base em 135 indivíduos do painel brasileiro de variedades de cana-de-açúcar (PBVCA). Esse painel, que foi desenvolvido principalmente para a realização do mapeamento associativo, reúne ancestrais importantes, variedades mais cultivadas, clones promissores, principais genitores em cruzamentos e variedades utilizadas em programas de mapeamento. Um total de 1.474 marcadores polimórficos foi obtido, considerando 86 EST-SSRs e 14 SSRs. Um mapa de ligação prévio foi utilizado para verificar as distâncias entre marcadores mapeados do PBVCA. O teste exato de Fisher foi realizado entre todos os possíveis pares de locos e usado como uma medida do DL. A estrutura populacional foi analisada através do método probabilístico implementado no STRUCTURE e do método Neighbor-Joining, este baseado na dissimilaridade genética, calculada pelo "simple matching", entre todos os pares de indivíduos. Forte DL foi observado em uma distância de até 15 cM, principalmente nos primeiros 5 cM. Quatro subpopulações foram detectadas no PBVCA através de ambos os métodos, que parecem estar de acordo com informações oriundas de *pedigree*. Esses resultados podem fornecer direcionamentos importantes para futuros estudos de mapeamento genético, os quais certamente precisarão considerar o elevado nível de ploidia da cana-de-açúcar.

Palavras-chave: *Saccharum* spp.; Poliplóides; Marcadores dominantes; Associação não-aleatória; Mapeamento por Desequilíbrio de Ligação

# ABSTRACT

**Analysis of linkage disequilibrium and population structure from the sugarcane Brazilian germplasm**

Sugarcane (*Saccharum* spp.) is a very important crop for sugar and alcohol production in Brazil. A great progress on its genome has emerged through molecular markers, which have been widely used for several purposes, such as QTL mapping based on bi-parental crosses. However, the use of populations derived from designed crosses tend to limit the genetic recombination events, resulting in a lower mapping resolution. In this sense, association mapping has emerged as a promising tool for QTL detection, since may be used natural populations, germplasm collections, elite set of materials, and others, allowing to detect recombination events in a historical and evolutionary context. To define the best strategies of association mapping, it is fundamental to account linkage disequilibrium (LD) and genetic structure existing in a certain population, so to verify the number of molecular markers and to promote the control of false positives. Here we measured LD and population structure across the genome of commercial varieties and clones of interest for sugarcane improvement in Brazil, based on 135 individuals from the Brazilian collection of sugarcane varieties (BCSV). This collection, which was mainly developed for association mapping, brings together important ancestral species, most planted varieties, promising clones, most used varieties as progenitors and varieties from the genetic mapping programs. A total of 1,474 polymorphic markers was scored, considering 86 EST-SSRs and 14 SSRs primers. A previous genetic map was used to verify distances between mapped BCSV markers. Fisher exact test was performed for all pairs of markers and used as a LD measure. Population structure was assessed by the probabilistic model implemented in STRUCTURE and the Neighbor-Joining method, based on simple matching dissimilarity between all pairs of individuals. Strong LD was observed up to 15 cM, mainly within the first 5 cM. Four subpopulations were detected in the BCSV with both methods, which is in agreement with *pedigree* informations. These results can provide important directions for future studies of genetic mapping, which would certainly need to consider high ploidy level of sugarcane.

Keywords: *Saccharum* spp.; Polyploids; Dominant markers; Non-random association; Linkage Disequilibrium Mapping

# 1 INTRODUÇÃO

Os caracteres agronômicos de importância econômica, que são selecionados pelos programas de melhoramento genético, são em sua grande maioria controlados por muitos genes e sofrem grande influência ambiental (FALCONER; MACKAY, 1996; WEIR, 1996; LYNCH; WALSH, 1998; MACKAY, 2001a). A herança complexa e pouco conhecida desses caracteres dificulta a seleção de genótipos e a adoção de novas estratégias pelos melhoristas para a exploração dos germoplasmas existentes (SOUZA, 2001). Essa dificuldade é pronunciada em espécies com grande complexidade genética, como é o caso da cana-de-açúcar (HOGARTH, 1987; HEINZ; TEW, 1987; MATSUOKA; GARCIA; CALHEIROS, 1999; D'HONT, 2005; OLIVEIRA et al., 2007; RABOIN et al., 2008), contribuindo com aumento no tempo de desenvolvimento de novas variedades pelos programas de melhoramento genético (WU et al., 2006; RABOIN et al., 2008).

Com o recente advento dos marcadores moleculares, um importante progresso no conhecimento da estrutura genômica e na genética de variedades modernas de cana-de-açúcar tem sido alcançado (GAZAFFI et al., 2010). Diversidade genética (JANNOO et al., 1999a; LIMA et al., 2002), escolha de genitores, identificação de variedades, mapeamento genético (AITKEN; JACKSON; MCINTYRE, 2005, 2007; RABOIN et al., 2006; GARCIA et al., 2006; OLIVEIRA et al., 2007) e de QTL's - *Quantitative Trait Loci* - (MCINTYRE et al., 2006; AL-JANABI et al., 2007; PIPERIDIS et al., 2008; AITKEN et al., 2008; PINTO et al., 2010) são algumas das aplicações dos marcadores moleculares em cana-de-açúcar. A construção de mapas genéticos e o mapeamento de QTL's com base em cruzamentos biparentais têm sido estratégias importantes no estudo de caracteres quantitativos em cana-de-açúcar, fornecendo informações valiosas rumo à seleção assistida por marcadores moleculares (GAZAFFI et al., 2010; PASTINA et al., 2010; MARGARIDO, 2011).

Entretanto, o uso de populações experimentais oriundas de cruzamentos controlados tende a limitar os eventos de recombinação genética, gerando menor resolução de mapeamento (ZHU et al., 2008). Nesse sentido, o mapeamento associativo surge como ferramenta promissora no estudo de QTL's, uma vez que podem ser utilizadas populações naturais, coleções de germoplasma, conjunto de materiais elite, entre outros, possibilitando detectar eventos de recombinação em um contexto histórico-evolutivo (ZHU et al., 2008; MACKAY et al., 2009). Para definir as melhores estratégias de mapeamento associativo, é fundamental conhecer a extensão genômica do de-

sequilíbrio de ligação, o qual representa a associação não-aleatória, ou preferencial, entre alelos de diferentes locos em uma população (LEWONTIN; KOJIMA, 1960; LEWONTIN, 1964; HILL; ROBERTSON, 1968; HILL, 1974; WEIR, 1979; LEWONTIN, 1988; PRITCHARD; PRZEWORSKI, 2001; ARDLIE; KRUGLYAK; SEIELSTAD, 2002; MCVEAN, 2007; FLINT-GARCIA; THORNSBERRY; BUCKLER, 2003; GUPTA; RUSTGI; KULWAL, 2005; HARTL; CLARK, 2007; SLATKIN, 2008; HAMILTON, 2009; HEDRICK, 2010). Além disso, é imprescindível verificar a estrutura genética presente nesta população, de modo que falsas associações, representadas por desequilíbrio de ligação entre locos não ligados, sejam desconsideradas nas análises de QTL's e, consequentemente, na seleção assistida por marcadores moleculares (FLINT-GARCIA; THORNSBERRY; BUCKLER, 2003; GUPTA; RUSTGI; KULWAL, 2005; ABDURAKHMONOV; ABDUKARIMOV, 2008; ZHU et al., 2008).

Assim, o presente trabalho teve como objetivo analisar o desequilíbrio de ligação e a estrutura populacional ao longo do genoma de variedades comerciais e clones de interesse para o melhoramento da cana-de-açúcar no Brasil, através de 135 indivíduos do painel brasileiro de variedades de cana-de-açúcar, servindo de base para a futura realização do mapeamento associativo.

# 2 REVISÃO BIBLIOGRÁFICA

## 2.1 Aspectos Gerais e Melhoramento da Cana-de-açúcar

A cana-de-açúcar (*Saccharum* spp.) está entre as espécies mais antigas e cultivadas do mundo (JAMES, 2004). Destaca-se pela produção de açúcar e etanol, podendo ser utilizada para outros fins, como produção de aguardente e alimentação bovina (MATSUOKA; GARCIA; ARIZONO, 1999). O Brasil é o maior produtor mundial, com área estimada para colheita de aproximadamente 8,43 milhões de hectares e produção ao redor de 588,92 milhões de toneladas na safra 2011/12, volume inferior estimado em 5,6% ao da safra anterior (CONAB, 2011). A redução dos subsídios da União Européia ao açúcar de beterraba, bem como o aquecimento global e a diminuição das reservas mundiais de petróleo colocam o Brasil em posição de destaque com a disponibilidade de açúcar e etanol (FNP – Consultoria & Comércio, 2008). Isso porque, além de ser o principal produtor de cana-de-açúcar, possui áreas agricultáveis para a sua expansão, condições ambientais favoráveis ao seu cultivo e variedades melhoradas com alta produtividade, desenvolvidas a partir dos trabalhos realizados pelos programas de melhoramento genético (MATSUOKA; GARCIA; CALHEIROS, 1999; BERDING; HOGARTH; COX, 2004; LANDELL; BRESSIANI, 2008).

A cana-de-açúcar é uma planta alógama, da família *Poaceae*, tribo Andropogoneae, subtribo Saccharinae, gênero *Saccharum* (STEVENSON, 1965; MATSUOKA; GARCIA; ARIZONO, 1999). Daniels e Roach (1987) consideram a ocorrência de seis espécies no gênero *Saccharum*, a saber: *S. officinarum* L. (2n = 80), *S. robustum* Brandes e Jeswiet ex Grassl (2n = 60 – 205), *S. barberi* Jeswiet (2n = 81 – 124), *S. sinense* Roxb. (2n = 111 – 120), *S. spontaneum* L. (2n = 40 – 128) e *S. edule* Hassk (2n = 60-80). Acredita-se que o complexo *Saccharum* spp., o qual compreende as variedades modernas de cana-de-açúcar, tenha sido originado da hibridação interespecífica entre estas espécies do gênero *Saccharum* e outras selvagens relacionadas (STEVENSON, 1965; DANIELS; ROACH, 1987; SREENIVASAN et al., 1987; IRVINE, 1999; MATSUOKA; GARCIA; ARIZONO, 1999; BERDING; HOGARTH; COX, 2004; MING et al., 2006; LANDELL; BRESSIANI, 2008). Por isso, considera-se que a cana-de-açúcar é uma das espécies cultivadas de maior complexidade genética (HOGARTH, 1987; D'HONT, 2005; RABOIN et al., 2008), com alto grau de ploidia e frequente aneuploidia (MATSUOKA; GARCIA; CALHEIROS, 1999), o que muito dificulta o seu melhoramento (HEINZ; TEW, 1987).

O melhoramento genético da cana-de-açúcar baseia-se na seleção e clonagem de genótipos superiores presentes em populações segregantes, as quais são formadas por milhares de plântulas obtidas através de cruzamentos sexuais entre indivíduos diferentes (MATSUOKA; GARCIA; ARIZONO, 1999). Várias etapas são necessárias para se obter poucos genótipos promissores em fases finais, os quais são avaliados com base em muitas repetições, diferentes locais e anos de cultivo. Com isso, é possível obter razoável segurança nas recomendações e, assim, utilizar os genótipos como variedades comerciais (MATSUOKA; GARCIA; ARIZONO, 1999). No entanto, a liberação comercial de uma nova variedade ocorre, em média, entre 10 e 15 anos, sendo bastante onerosa a um programa de melhoramento genético (WU et al., 2006; RABOIN et al., 2008).

Nesse sentido, estudos recentes têm sido realizados com o objetivo de reduzir significativamente o tempo de desenvolvimento de novas variedades de cana-de-açúcar (GAZAFFI et al., 2010). A análise da estrutura genômica desta espécie através do emprego de marcadores moleculares vem sendo uma estratégia bastante utilizada para tal finalidade. Esta análise envolve basicamente estudos de diversidade genética (JANNOO et al., 1999a; LIMA et al., 2002), mapas genéticos de ligação (AITKEN; JACKSON; MCINTYRE, 2005, 2007; RABOIN et al., 2006; GARCIA et al., 2006; OLIVEIRA et al., 2007), mapeamento de QTL's por análise de ligação (MCINTYRE et al., 2006; AL-JANABI et al., 2007; PIPERIDIS et al., 2008; AITKEN et al., 2008; PINTO et al., 2010) e, mais recentemente, mapeamento por associação (WEI; JACKSON; MCINTYRE, 2006; BUTTERFIELD, 2007; WEI et al., 2010). A análise por associação surge como alternativa promissora para seleção assistida por marcadores moleculares em cana-de-açúcar, a qual poderá contribuir para a redução no tempo de obtenção de novas variedades (FLINT-GARCIA; THORNSBERRY; BUCKLER, 2003; GUPTA; RUSTGI; KULWAL, 2005; YU; BUCKLER, 2006; ZHU et al., 2008).

## 2.2 Marcadores Moleculares, Mapas Genéticos de Ligação e Mapeamento de QTL's

Os marcadores moleculares são uma ferramenta rápida e eficaz para estudos genômicos, uma vez que detectam o polimorfismo diretamente ao nível do DNA e não sofrem qualquer tipo de influência ambiental (FERREIRA; GRATTAPAGLIA, 1998; SOUZA, 2001; AVISE, 2004; SCHLÖTTERER, 2004; GAZAFFI et al., 2010). Para a cana-de-açúcar, que apresenta elevada complexidade genética, a sua utilização é bastante valiosa (DAUGROIS et al., 1996). RFLP (Restriction Fragment Length Polymorphism), EST-RFLP (Expressed Sequence Tag RFLP), RAPD (Random Amplified Polymorphic DNA), AFLP (Amplified Fragment Length Polymorphism), SSR (Simple Sequence

Repeat) e EST-SSR (Expressed Sequence Tag SSR) são marcadores moleculares amplamente utilizados em cana-de-açúcar na construção de mapas genéticos (AITKEN; JACKSON; MCINTYRE, 2005, 2007; RABOIN et al., 2006; GARCIA et al., 2006; OLIVEIRA et al., 2007) e no mapeamento de QTL's por análise de ligação (MCINTYRE et al., 2006; AL-JANABI et al., 2007; PIPERIDIS et al., 2008; AITKEN et al., 2008; PINTO et al., 2010; PASTINA, 2010; MARGARIDO, 2011).

A análise de ESTs é uma estratégia simples para o estudo da porção expressa do genoma, mesmo em organismos com genomas grandes e complexos, como a cana-de-açúcar. Os marcadores derivados das ESTs oferecem oportunidades para a descoberta de genes, já que se determinada EST estiver geneticamente associada a uma característica de interesse, é provável que a mesma afete diretamente a característica (CATO et al., 2001). Além disso, como encontram-se em regiões transcritas do genoma, os sítios de *primers* são mais conservados (SCHLÖTTERER, 2004; VARSHNEY; GRANER; SORRELLS, 2005), tornando-os mais transferíveis entre espécies e aumentando o seu valor em programas de melhoramento. Os microssatélites derivados de ESTs (EST-SSRs) ocorrem em alta frequência no genoma (VARSHNEY; GRANER; SORRELLS, 2005) e têm sido avaliados em estudos envolvendo várias espécies, como a cana-de-açúcar (CORDEIRO et al., 2000; PINTO et al., 2004; OLIVEIRA et al., 2007; PASTINA, 2010; MARGARIDO, 2011).

A construção de mapas genéticos de ligação a partir de dados moleculares em cana-de-açúcar tem sido realizada com populações $F_1$ segregantes oriundas de cruzamentos entre indivíduos não endogâmicos (LIN et al., 2003; ALWALA; KIMBENG, 2010). Isso porque para esta espécie a obtenção de linhagens endogâmicas é impraticável, principalmente em função da grande depressão por endogamia gerada quando ocorrem autofecundações (GAZAFFI et al., 2010; PASTINA et al., 2010). Neste caso, uma alternativa para a construção de mapas genéticos é denominada duplo *pseudo-testcross*, a qual consiste basicamente na obtenção de um mapa para cada genitor através da identificação de polimorfismos em marcadores de dosagem única (GRATTAPAGLIA; SEDEROFF, 1994). Vários estudos já foram realizados com base nesta estratégia em cana-de-açúcar (SOBRAL; HONEYCUTT, 1993; AL-JANABI et al., 1993; DA SILVA et al., 1993, 1995; GRIVET et al., 1996; MUDGE et al., 1996; GUIMARÃES et al., 1999; HOARAU et al., 2001).

No entanto, é desejável a integração das informações contidas nesses mapas individuais em um único mapa integrado, através da utilização de diferentes marcadores moleculares e segregações (GARCIA et al., 2006; GAZAFFI et al., 2010). É uma grande vantagem para espécies poliplóides, como a cana-de-açúcar, pois permite aumentar a saturação do mapa de ligação e estender a caracteri-

zação do polimorfismo para todo o genoma (DA SILVA et al., 1993; GRIVET et al., 1996). Porém, Garcia et al. (2006) e Oliveira et al. (2007), ao construírem mapas integrados para cana-de-açúcar, obtiveram pouca cobertura do genoma, a qual também foi verificada no mapa integrado multiponto utilizado por Pastina (2010) e Margarido (2011). Isso mostra a necessidade de uma quantidade maior de marcadores para a localização mais precisa de QTL's nesta espécie, que podem estar relacionados com características agrícolas economicamente importantes.

O termo QTL tem sido utilizado para indicar regiões cromossômicas que contêm locos gênicos que controlam caracteres poligênicos, ou seja, caracteres quantitativos que sofrem grande influência ambiental e apresentam variação contínua (FALCONER; MACKAY, 1996; LYNCH; WALSH, 1998; MACKAY, 2001a; ZENG, 2001; DOERGE, 2002). O estudo de QTL's tem sido bastante importante no entendimento da herança complexa de características agronômicas importantes. Com o recente advento dos marcadores moleculares, tornou-se possível estudar com maior intensidade esses QTL's (FERREIRA; GRATTAPAGLIA, 1998; GAZAFFI et al., 2010).

O mapeamento de QTL's possibilita mensurar o número de locos quantitativos envolvidos na herança complexa, as suas localizações cromossômicas, o modo de ação gênica (aditividade, dominância e epistasia) e efeitos pleiotrópicos, além de possibilitar a decomposição da interação genótipos por ambientes ao nível de cada QTL (FERREIRA; GRATTAPAGLIA, 1998; AUSTIN; LEE, 1998; ZENG, 2001; CARNEIRO; VIEIRA, 2002; ZHU et al., 2008; PASTINA et al., 2010; GAZAFFI et al., 2010). Com isso, é possível realizar inferências em todo o genoma sobre as relações entre genótipo e fenótipo de caracteres quantitativos, o que em última análise permite aumentar a eficiência dos programas de melhoramento (GAZAFFI et al., 2010).

Em cana-de-açúcar, vários trabalhos de mapeamento de QTL's foram realizados com base em populações experimentais oriundas do cruzamento entre linhagens não endogâmicas e mapas individuais não integrados (GRIVET et al., 1996; GUIMARÃES et al., 1999; HOARAU et al., 2001), bem como mapas integrados (PASTINA, 2010; MARGARIDO, 2011). Entretanto, até o momento, poucos QTL's foram detectados para características agronômicas importantes, o que também tem ocorrido em estudos envolvendo outras espécies vegetais.

## 2.3 Desequilíbrio de Ligação

### 2.3.1 Definição e Considerações Iniciais

Desequilíbrio de Ligação é qualquer desvio das frequências alélicas em relação às frequências esperadas sob independência, indicando a existência de associação não-aleatória, ou preferencial, entre alelos de diferentes locos em uma população. A Figura 1 apresenta de maneira comparativa duas situações em que alelos de dois locos encontram-se associados aleatoriamente, ou em equilíbrio de ligação, e associados preferencialmente, ou em desequilíbrio de ligação (DL).

Apesar dessas ideias terem sido amplamente estudadas por vários trabalhos antigos (WEINBERG, 1909; JENNINGS, 1917; ROBBINS, 1918; GEIRINGER, 1944; BENNETT, 1954; KIMURA, 1956), o termo Desequilíbrio de Ligação apenas foi utilizado pela primeira vez no artigo publicado por Lewontin e Kojima (1960). Esses autores assim o designaram devido ao fato que a ligação entre locos era fator determinante para uma população atingir o equilíbrio (HEDRICK, 2010). Na época, os cientistas não se preocuparam com o quanto essa designação era apropriada, uma vez que suas definições matemáticas ainda não estavam claras (SLATKIN, 2008). Quando o DL passou a ser melhor compreendido, até mesmo fora da genética de populações, sua designação encontrava-se bastante consolidada para ser substituída (SLATKIN, 2008). De qualquer forma, considerações importantes devem ser apresentadas com relação a essa terminologia.

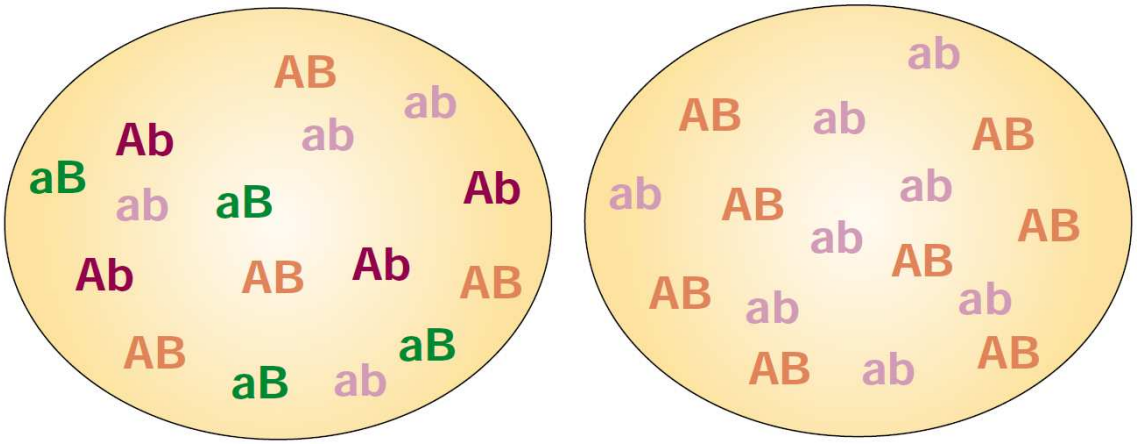


Figura 1 – Considere dois locos $A$ e $B$, ambos bialélicos ($A$ e $a$; $B$ e $b$). Quatro tipos gaméticos são possíveis: $AB$, $Ab$, $aB$ e $ab$. Se esses locos estiverem em equilíbrio de ligação (à esquerda), os quatro gametas são observados. No entanto, se esses locos estiverem em máximo desequilíbrio de ligação (à direita), apenas os gametas em associação ($AB$ e $ab$) são observados (MACKAY, 2001b)

DL e ligação física são termos frequentemente confundidos, pois são, muitas vezes, encarados como sinônimos. Isso é um equívoco, e normalmente ocorre porque locos intimamente ligados podem estar em elevado DL (FLINT-GARCIA; THORNSBERRY; BUCKLER, 2003). O fato de dois

locos quaisquer estarem em DL não significa que estão fisicamente ligados. Do mesmo modo, o fato de dois locos estarem fisicamente ligados em determinado cromossomo ou haplótipo não significa que estão em DL (PRITCHARD; PRZEWORSKI, 2001; ARDLIE; KRUGLYAK; SEIELSTAD, 2002; SLATKIN, 2008; HEDRICK, 2010). É exatamente por essa confusão que muitos autores recomendam a utilização dos termos Desequilíbrio da Fase Gamética ou Desequilíbrio Gamético para o DL (CROW; KIMURA, 1970; HEDRICK, 1987; TEMPLETON, 2006; HARTL; CLARK, 2007; ALLENDORF; LUIKART, 2007; HAMILTON, 2009; HEDRICK, 2010), assim como foi inicialmente proposto por Jain e Allard (1966). De qualquer forma, é fundamental que essa confusão seja devidamente esclarecida com os seguintes conceitos corretos: i) DL pode ocorrer entre alelos de dois ou mais locos fisicamente ligados, que estão no mesmo cromossomo ou haplótipo, sendo de interesse para o mapeamento de QTL's; e ii) DL pode ocorrer entre alelos de dois ou mais locos que não estão fisicamente ligados, ou seja, que estão muito distantes em um cromossomo ou que estão em cromossomos diferentes, não sendo de interesse para o mapeamento genético, no contexto da seleção assistida por marcadores moleculares. Esse último DL pode surgir, entre outros fatores, pela mistura de indivíduos de subpopulações ou de diferentes populações, a qual pode causar uma série de perturbações que acabam culminando em associações preferenciais (NEI; LI, 1973; JORDE, 1995; XIONG; GUO, 1997; KRUGLYAK, 1999; PRITCHARD; PRZEWORSKI, 2001; CARDON; BELL, 2001; ARDLIE; KRUGLYAK; SEIELSTAD, 2002; FLINT-GARCIA; THORNSBERRY; BUCKLER, 2003; SLATKIN, 2008; ABDURAKHMONOV; ABDUKARIMOV, 2008; ZHU et al., 2008; HEDRICK, 2010). Fica evidente portanto que as causas do DL precisam ser determinadas para que o mesmo possa ser usado corretamente.

Para entender detalhadamente essas ideias, são apresentados a seguir alguns conceitos importantes acerca do DL, principalmente no contexto de organismos diplóides. Oportunamente, esses conceitos são expandidos para o contexto de organismos poliplóides, como é o caso da cana-de-açúcar, apesar do comportamento do DL em maiores ploidias ser praticamente desconhecido.

### 2.3.2 Desequilíbrio de Ligação em Diplóides - Fase de Ligação Conhecida

#### 2.3.2.1 Pares de Locos Bialélicos

Considere dois locos fisicamente ligados $A$ e $B$, ambos bialélicos ($A$ e $a$; $B$ e $b$), em uma geração $N$ qualquer de determinada população. Quatro diferentes gametas, ou haplótipos, podem

ser identificados: $AB$, $Ab$, $aB$ e $ab$. Se esses locos estiverem em equilíbrio de ligação (DL = 0), ou seja, numa situação de independência ou associação aleatória entre os seus alelos, as frequências observadas dos haplótipos serão exatamente equivalentes às suas frequências esperadas, as quais correspondem ao produto das respectivas frequências alélicas (LEWONTIN; KOJIMA, 1960):

$$\begin{aligned} \text{Haplótipo 1:} \quad f_{AB} &= f_A f_B \\ \text{Haplótipo 2:} \quad f_{Ab} &= f_A f_b \\ \text{Haplótipo 3:} \quad f_{aB} &= f_a f_B \\ \text{Haplótipo 4:} \quad f_{ab} &= f_a f_b \end{aligned}$$

No entanto, se esses locos não estiverem em equilíbrio de ligação (DL $\neq$ 0), de modo que seus alelos estejam combinados de maneira não-aleatória, ou preferencial, as frequências dos haplótipos apresentarão um desvio ($D$) em relação às frequências esperadas (LEWONTIN; KOJIMA, 1960):

$$\begin{aligned} \text{Haplótipo 1:} \quad f_{AB} &= f_A f_B + D_{AB} \\ \text{Haplótipo 2:} \quad f_{Ab} &= f_A f_b - D_{Ab} \\ \text{Haplótipo 3:} \quad f_{aB} &= f_a f_B - D_{aB} \\ \text{Haplótipo 4:} \quad f_{ab} &= f_a f_b + D_{ab} \end{aligned}$$

sendo que $D_{AB} = -D_{Ab} = -D_{aB} = D_{ab} = D$ (MCVEAN, 2007; SLATKIN, 2008; HEDRICK, 2010). Com simples manipulações algébricas, é possível expressar esse DL (módulo $D$), dito de menor ordem, de forma geral, sendo a diferença entre as frequências observada e esperada dos haplótipos (LEWONTIN; KOJIMA, 1960):

$$D = D_{AB} = f_{AB} - f_A f_B$$

Ainda, é possível expressar o DL de menor ordem como a diferença entre o produto das frequências dos haplótipos em associação ($AB$ e $ab$) e o produto das frequências dos haplótipos em repulsão ($Ab$ e $aB$) (LEWONTIN; KOJIMA, 1960):

$$D = f_{AB} f_{ab} - f_{Ab} f_{aB}$$

sendo este, portanto, o conceito básico do DL (LEWONTIN; KOJIMA, 1960) num contexto de menor ordem, quando a fase de ligação entre os locos é conhecida. Vários autores podem ser consultados para maiores detalhes a respeito desses conceitos, sendo julgados alguns relevantes: Jennings (1917), Robbins (1918), Geiringer (1944), Bennett (1954), Kimura (1956), Lewontin e

Kojima (1960), Lewontin (1964), Hill e Robertson (1968), Ohta e Kimura (1969), Wright (1969), Franklin e Lewontin (1970), Crow e Kimura (1970), Hill (1974), Weir (1979), Hedrick (1987), Lewontin (1988), Devlin e Risch (1995), Falconer e Mackay (1996), Weir (1996), Slatkin e Excoffier (1996), Lynch e Walsh (1998), Templeton (2006), McVean (2007), Hartl e Clark (2007), Allendorf e Luikart (2007), Weir (2008), Slatkin (2008), Hamilton (2009) e Hedrick (2010).

É importante notar que o máximo valor de $D$ ($D = 0{,}25$) ocorre somente quando haplótipos em associação ($AB$ e $ab$) estão presentes numa amostra de determinada população, sendo que a soma das suas frequências deve ser igual a 1. Teoricamente, na ausência de outros processos, essa situação ocorre em populações recentes com poucas gerações, quando os eventos de recombinação entre os locos não foram suficientes para o surgimento de outros haplótipos. Neste caso, as frequências dos haplótipos em repulsão ($Ab$ e $aB$) são iguais a zero. Do mesmo modo, o mínimo valor de $D$ ($D = -0{,}25$) ocorre somente quando haplótipos em repulsão ($Ab$ e $aB$) estão presentes numa determinada população, sendo que a soma das suas frequências deve ser equivalente a 1. Após muitas gerações, e com elas muitas oportunidades de recombinação, espera-se que esse cenário seja possível para muitos locos. No entanto, isso tudo dependerá, entre outros fatores, da fração de recombinação entre os locos, já que quanto menor a distância entre eles menor a probabilidade de ocorrer um evento de recombinação e, com isso, o surgimento de novos haplótipos. Essas ideias são melhor discutidas no item 2.3.5.

Neste tópico, foi mostrado que dois locos bialélicos $A$ e $B$ estão em DL quando $D \neq 0$. Na verdade, esses locos podem estar em DL, pois o fato deste parâmetro apresentar um valor diferente de zero não é suficiente para afirmar que estão associados preferencialmente. Em outras palavras, os locos apenas estarão associados de maneira não-aleatória se o valor calculado para o DL for estatisticamente significativo (SLATKIN; EXCOFFIER, 1996; WEIR, 1996; MCVEAN, 2007; SLATKIN, 2008; HEDRICK, 2010). Nesse sentido, vários testes estatísticos têm sido utilizados, sendo que os mais comuns são o teste de qui-quadrado e o teste exato de Fisher (SLATKIN, 1994; LEWONTIN, 1995; SLATKIN; EXCOFFIER, 1996; WEIR, 1996; FLINT-GARCIA; THORNSBERRY; BUCKLER, 2003; GUPTA; RUSTGI; KULWAL, 2005; SLATKIN, 2008). O teste de qui-quadrado, que foi bastante utilizado pela simplicidade computacional exigida, é tido como adequado quando o número esperado de haplótipos não é tão pequeno numa amostra populacional (SLATKIN, 1994; WEIR, 1996; MCVEAN, 2007; HARTL; CLARK, 2007). Em contrapartida, o teste exato de Fisher, que passou a ser facilmente utilizado com o surgimento de algoritmos sofisti-

cados (MEHTA; PATEL, 1983; GUO; THOMPSON, 1992), pode ser empregado para maiores e menores quantidades de haplótipos decorrentes de uma amostra populacional (SLATKIN, 1994; MCVEAN, 2007; FOULKES, 2009). Por esta razão, pode apresentar maior poder na detecção do DL quando comparado ao teste de qui-quadrado (LEWONTIN, 1995). No entanto, ambos são considerados inapropriados quando frequências tão baixas são obtidas em função da presença de alelos raros oriundos de locos polimórficos, o que muito dificulta a detecção de DL (LEWONTIN, 1995). Neste caso, é difícil determinar qual desses testes é o mais adequado (LEWONTIN, 1995).

Detalhes a respeito dos procedimentos estatísticos envolvidos na detecção de DL podem ser encontrados em Weir (1996).

#### 2.3.2.2 Pares de Locos Multialélicos

No item 2.3.2.1, o cálculo do DL foi apresentado considerando que os locos $A$ e $B$ são bialélicos, de modo que quatro diferentes haplótipos poderiam ser identificados numa amostra populacional. Grande parte da teoria e do entendimento envolvendo o DL permanece nesse contexto até os dias de hoje (SLATKIN; EXCOFFIER, 1996). No entanto, se esses locos fossem multialélicos, com $X$ e $Y$ alelos, respectivamente, sendo ambos maior que 2, $X \times Y$ haplótipos poderiam ser identificados numa amostra da população. Assim, um número de haplótipos acima de quatro poderia ser encontrado, aumentando a probabilidade da amostra não conter todas as possíveis combinações alélicas (SLATKIN; EXCOFFIER, 1996). Neste caso, onde o cenário se torna mais complexo, a detecção de associações preferenciais entre os locos $A$ e $B$ poderá não ocorrer, o que não garante que seus alelos estejam associados aleatoriamente, ou em equilíbrio de ligação. Portanto, por mais que o poder na detecção de associações preferenciais seja maior considerando locos multialélicos (SLATKIN, 1994), a falta de representatividade amostral pode comprometer essa verificação e, com isso, gerar conclusões equivocadas a respeito do DL para determinada população.

Mesmo que a diversidade de haplótipos de uma população estivesse totalmente presente em uma amostra, o grau do DL entre dois locos multialélicos é bastante complexo para ser representado em um único coeficiente, como acontece no caso de dois locos bialélicos (ZAPATA, 2000; MCVEAN, 2007; HEDRICK, 2010). Assim, até o momento, nenhum conceito satisfatório foi desenvolvido para o cálculo do DL neste contexto (ARDLIE; KRUGLYAK; SEIELSTAD, 2002; SLATKIN, 2008), apesar de alguns já terem sido propostos. Hill (1975) propôs uma medida proporcional a uma distribuição de qui-quadrado, enquanto Hedrick (1987) propôs uma extensão da

medida $D'$ inicialmente desenvolvida para dois locos bialélicos (ver item 2.3.6). Apesar de estudos terem mostrado que a última sugestão pode ser adequada (ZAPATA, 2000), a mesma não parece ter tido grande aceitação (ARDLIE; KRUGLYAK; SEIELSTAD, 2002; SLATKIN, 2008). Isso pôde ser observado, por exemplo, em trabalhos realizados com humanos no início dos estudos de associação, uma vez que foram utilizados testes estatísticos como medidas do DL entre pares de locos microssatélites (ARDLIE; KRUGLYAK; SEIELSTAD, 2002). Porém, como esses testes não são medidas próprias do DL, muitas incertezas começaram a surgir a partir da sua utilização na detecção de associações preferenciais. Assim, essas limitações, aliadas às altas taxas de mutação dos microssatélites e ao rápido surgimento dos polimorfismos de base única (Single Nucleotide Polymorphisms - SNPs), parecem ter reduzido o interesse no uso dos microssatélites nos estudos envolvendo DL e mapeamento associativo em humanos (ARDLIE; KRUGLYAK; SEIELSTAD, 2002; SLATKIN, 2008). Em plantas, vários desses estudos ainda têm sido realizados com esses marcadores, sendo muitos deles derivados de ESTs, as quais representam regiões mais conservadas.

#### 2.3.2.3 Múltiplos Locos Bialélicos

Numa situação mais complexa, considere três locos fisicamente ligados $A$, $B$ e $C$, todos bialélicos ($A$ e $a$; $B$ e $b$; $C$ e $c$), em uma geração $N$ qualquer de determinada população. Oito diferentes gametas, ou haplótipos, podem ser identificados: $ABC$, $ABc$, $AbC$, $Abc$, $aBC$, $aBc$, $abC$ e $abc$. Se esses locos estiverem em equilíbrio de ligação (DL = 0), ou seja, numa situação de independência ou associação aleatória entre os seus alelos, as frequências observadas dos haplótipos serão exatamente equivalentes às suas frequências esperadas, as quais correspondem ao produto das respectivas frequências alélicas:

$$
\begin{aligned}
\text{Haplótipo 1:} \quad f_{ABC} &= f_A f_B f_C \\
\text{Haplótipo 2:} \quad f_{ABc} &= f_A f_B f_c \\
\text{Haplótipo 3:} \quad f_{AbC} &= f_A f_b f_C \\
\text{Haplótipo 4:} \quad f_{Abc} &= f_A f_b f_c \\
\text{Haplótipo 5:} \quad f_{aBC} &= f_a f_B f_C \\
\text{Haplótipo 6:} \quad f_{aBc} &= f_a f_B f_c \\
\text{Haplótipo 7:} \quad f_{abC} &= f_a f_b f_C \\
\text{Haplótipo 8:} \quad f_{abc} &= f_a f_b f_c
\end{aligned}
$$

Porém, se esses locos não estiverem em equilíbrio de ligação (DL $\neq$ 0), de modo que seus alelos estejam combinados de maneira não-aleatória, ou preferencial, as frequências dos haplótipos apresentarão desvios ($D$) em relação às suas frequências esperadas, os quais referem-se aos DL's entre os possíveis pares de locos ($A$ e $B$; $A$ e $C$; $B$ e $C$) e entre a interação dos três locos ($A$, $B$ e $C$) (GEIRINGER, 1944; BENNETT, 1954):

$$\begin{array}{lrcccccccccc}
\text{Haplótipo 1:} & f_{ABC} & = & f_A f_B f_C & + & f_A D_{BC} & + & f_B D_{AC} & + & f_C D_{AB} & + & D_{ABC} \\
\text{Haplótipo 2:} & f_{ABc} & = & f_A f_B f_c & - & f_A D_{BC} & - & f_B D_{AC} & + & f_c D_{AB} & - & D_{ABC} \\
\text{Haplótipo 3:} & f_{AbC} & = & f_A f_b f_C & - & f_A D_{BC} & + & f_b D_{AC} & - & f_C D_{AB} & - & D_{ABC} \\
\text{Haplótipo 4:} & f_{Abc} & = & f_A f_b f_c & + & f_A D_{BC} & - & f_b D_{AC} & - & f_c D_{AB} & + & D_{ABC} \\
\text{Haplótipo 5:} & f_{aBC} & = & f_a f_B f_C & + & f_a D_{BC} & - & f_B D_{AC} & - & f_C D_{AB} & - & D_{ABC} \\
\text{Haplótipo 6:} & f_{aBc} & = & f_a f_B f_c & - & f_a D_{BC} & + & f_B D_{AC} & - & f_c D_{AB} & + & D_{ABC} \\
\text{Haplótipo 7:} & f_{abC} & = & f_a f_b f_C & - & f_a D_{BC} & - & f_b D_{AC} & + & f_C D_{AB} & + & D_{ABC} \\
\text{Haplótipo 8:} & f_{abc} & = & f_a f_b f_c & + & f_a D_{BC} & + & f_b D_{AC} & + & f_c D_{AB} & - & D_{ABC}
\end{array}$$

De forma geral, é possível expressar o DL de maior ordem, ou seja, devido à interação entre os três locos ($A$, $B$ e $C$), como a diferença entre as frequências observada e esperada do haplótipo juntamente com as diferenças entre os DL's dos possíveis pares de locos, ponderados pelas frequências dos alelos que constituem o haplótipo (WEIR, 1996; MCVEAN 2007; SLATKIN, 2008):

$$D_{ABC} = f_{ABC} - f_A f_B f_C - f_A D_{BC} - f_B D_{AC} - f_C D_{AB}$$

sendo este, portanto, o conceito básico do DL (LEWONTIN; KOJIMA, 1960) num contexto de maior ordem, quando a fase de ligação entre os locos é conhecida. Essas ideias foram inicialmente apresentadas por Geiringer (1944), sendo posteriormente abordadas por outros autores, tais como: Bennett (1954), Wright (1969), Franklin e Lewontin (1970), Crow e Kimura (1970), Slatkin (1972), Robinson et al. (1991), Long et al. (1995), Weir (1996), Gorelick e Laubichler (2004), Mueller (2004), Nielsen et al. (2004), McVean (2007), Kim et al. (2008) e Weir (2008).

Em comparação ao item 2.3.2.1, é importante notar que apenas um loco bialélico a mais foi suficiente para duplicar a quantidade de haplótipos possíveis numa amostra de determinada população. Se os locos $A$, $B$ e $C$ fossem multialélicos ao invés de bialélicos, certamente a complexidade seria ainda maior na detecção de associações significativas, visto que a quantidade de haplótipos numa amostra poderia crescer exponencialmente. Em outras palavras, quanto maior a quantidade

de haplótipos, em função das diferentes combinações alélicas dos locos, maior a chance de uma amostra não ser representativa da população e, com isso, não favorecer a detecção de associações preferenciais significativas. A Tabela 1 mostra o crescimento da quantidade de haplótipos com o aumento do número de locos e alelos, evidenciando a complexidade referida anteriormente.

Tabela 1 – Número de haplótipos de acordo com o número de locos e alelos

| Nº de Locos | Nº de Alelos | | | |
|---|---|---|---|---|
| | Dois | Três | Quatro | Cinco ... |
| Dois | 4 | 9 | 16 | 25 ... |
| Três | 8 | 27 | 64 | 125 ... |
| Quatro | 16 | 81 | 256 | 625 ... |
| Cinco | 32 | 243 | 1024 | 3125 ... |
| ⋮ | | | | |

Franklin e Lewontin (1970) e Slatkin (1972), entre outros autores, já mencionavam que a teoria baseada em dois locos poderia subestimar seriamente o DL, considerando que a real segregação cromossômica ocorre no contexto de múltiplos locos. Apesar de pesquisas recentes em humanos mostrarem que o DL entre pares de locos pode gerar informações muito úteis do genoma (WEIR, 2008), é pouco provável que sua complexidade esteja sendo verificada de maneira satisfatória.

Poucas abordagens têm sido propostas para o cálculo do DL entre múltiplos locos/sítios (KIM et al., 2008), muito provavelmente pela dificuldade em testar associações entre alelos nesse contexto (WEIR, 2008). De qualquer forma, Slatkin (2008) menciona que, apesar do pouco uso prático do DL de maior ordem, essa abordagem poderá ajudar no maior entendimento dos padrões do DL encontrados em humanos, complementando a análise entre pares de locos (WEIR, 1996). Espera-se que o DL multiloco também possa ser útil para maior entendimento do contexto histórico-evolutivo de organismos poliplóides, como a cana-de-açúcar, contribuindo para um mapeamento mais refinado de características de interesse.

### 2.3.3 Desequilíbrio de Ligação em Diplóides - Fase de Ligação Desconhecida

Muitas vezes, dependendo da espécie e da finalidade do estudo, não é possível obter amostras contendo diretamente os haplótipos para um conjunto de locos de determinada população (HILL, 1974; EXCOFFIER; SLATKIN, 1995; SLATKIN; EXCOFFIER, 1996; WEIR, 1996; SLATKIN,

2008; HEDRICK, 2010). Neste caso, o DL não pode ser calculado pelos conceitos anteriores, visto que a fase de ligação não é conhecida. Com o surgimento das técnicas moleculares, onde muitos dados passaram a ser obtidos com o uso de marcadores polimórficos, o problema de fase desconhecida tornou-se bastante frequente (EXCOFFIER; SLATKIN, 1995). Para o caso de dois locos bialélicos, por exemplo, a fase de ligação não pode ser diretamente obtida devido ao confundimento dos indivíduos duplamente heterozigotos (SLATKIN; EXCOFFIER, 1996; SLATKIN, 2008; HEDRICK, 2010), independentemente se os marcadores são dominantes ou codominantes (HILL, 1974). Assim, é possível estimar o DL apenas entre os locos (HILL, 1974; WEIR, 1979; WEIR, 1996; SLATKIN, 2008; HEDRICK, 2010), e não entre suas combinações alélicas, o que não é interessante pela perda de informação a respeito da estrutura e distribuição do DL ao longo do genoma (PRITCHARD; PRZEWORSKI, 2001; SLATKIN, 2008). Com fase de ligação desconhecida, menor é a probabilidade de detectar associações significativas entre os locos, aumentando-se as chances de inferir erroneamente a respeito do DL de determinada população.

Em função disso, diferentes métodos têm sido propostos para obtenção de haplótipos e suas frequências a partir de dados genotípicos, tais como: i) amplificação de cada cromossomo por PCR (Polymerase Chain Reaction), permitindo a identificação direta da fase entre os locos; ii) utilização de genealogias e informações de parentesco, onde é possível, muitas vezes, inferir a fase entre os locos na progênie a partir dos genótipos dos parentais; e iii) utilização de métodos estatísticos, os quais podem ser baseados em diferentes abordagens, como máxima verossimilhança (HILL, 1974; EXCOFFIER; SLATKIN, 1995; LONG et al., 1995; SLATKIN; EXCOFFIER, 1996) e coalescência. Os métodos estatísticos têm sido os mais utilizados, principalmente quando se tem grandes quantidades de dados, como é o caso dos estudos de associação (SLATKIN, 2008). O método baseado em máxima verossimilhança considera que todos os locos encontram-se em equilíbrio de Hardy-Weinberg, o que pode não ser verdade (EXCOFFIER; SLATKIN, 1995; PRITCHARD; PRZEWORSKI, 2001). Além disso, não parece ser apropriado quando se tem amostras pequenas, gerando incertezas quanto às estimativas obtidas, principalmente em relação aos haplótipos mais raros (EXCOFFIER; SLATKIN, 1995; SLATKIN, 2008). Slatkin (2008) recomenda a utilização de um método que leva em consideração esses aspectos, pois do mesmo modo que o DL pode ser viciado quando a fase de ligação não é conhecida, pode também o ser com fase de ligação equivocada.

### 2.3.4 Desequilíbrio de Ligação em Poliplóides

Ao contrário dos diplóides, os gametas podem não ser haplótipos em poliplóides, como é o caso da cana-de-açúcar. A quantidade de cromossomos em um gameta poliplóide dependerá da ploidia da espécie ou até mesmo da ploidia do grupo de homologia, já que eventos aneuplóides durante a segregação podem ser frequentes. Em uma espécie tetraplóide, como a batata por exemplo, desconsiderando inicialmente a aneuploidia e a variação de ploidia entre indivíduos, quatro cópias do número básico de cromossomos são esperadas. Neste caso, os gametas são formados por dois cromossomos, sendo, portanto, diplótipos. Já em uma espécie decaplóide, também desconsiderando os aspectos anteriores, dez cópias do número básico de cromossomos são esperadas. Assim, os gametas são formados por cinco cromossomos, sendo, portanto, pentaplótipos. A quantidade de gametas possíveis nessas espécies dependerá dos diferentes cromossomos, os quais são proporcionais ao número de locos e alelos. O exemplo a seguir mostra essas ideias.

Considere dois locos fisicamente ligados $A$ e $B$, ambos bialélicos ($A$ e $a$; $B$ e $b$), numa geração $N$ qualquer de determinada população. Se essa população for diplóide, quatro gametas poderão ser encontrados, sendo estes equivalentes ao número de haplótipos possíveis ($AB$, $Ab$, $aB$ e $ab$). No entanto, se essa população for tetraplóide ou decaplóide, uma quantidade maior de gametas poderá ser encontrada, não sendo esta, portanto, equivalente aos quatro haplótipos anteriores. Nestes casos, os gametas são resultados de diferentes combinações desses haplótipos. Se os locos $A$ e $B$ fossem multialélicos ao invés de bialélicos, uma quantidade ainda maior de combinações poderia ser encontrada em função do aumento do número de haplótipos possíveis (FISHER, 1947; CROW; KIMURA, 1970), o que também ocorreria no contexto de múltiplos locos, bialélicos ou não. Portanto, através desses e outros aspectos, como o fenômeno da dupla redução em autopoliplóides (GALLAIS, 2003), é possível notar maior complexidade para estimação dos gametas e suas frequências em espécies com ploidias mais elevadas, influenciando diretamente no cálculo do DL entre locos.

Mesmo que os gametas e suas frequências pudessem ser obtidos a partir de dados genotípicos, nenhuma medida que capitalize a complexidade do DL em poliplóides foi desenvolvida até o momento. Devido à herança polissômica, onde vários cromossomos segregam conjuntamente para a formação de um gameta (GALLAIS, 2003), o DL em poliplóides pode ocorrer de diferentes formas: i) entre alelos de diferentes locos, tanto num mesmo haplótipo quanto em haplótipos distintos de determinado grupo de homologia; e ii) entre alelos de um mesmo loco, estando estes em haplótipos distintos de determinado grupo de homologia, como mostrado por Gallais (2003). Esse

autor designou os DL's do primeiro e segundo casos como desequilíbrios de ligação e panmítico, respectivamente, e mostrou em detalhes esses conceitos para espécies tetraplóides. Para espécies com ploidias mais elevadas, como a cana-de-açúcar, nenhuma abordagem teórica sobre o DL, a nível gamético, foi desenvolvida até o momento. Em função disso, o DL nessa espécie tem sido normalmente estudado apenas entre locos, quando a fase de ligação é desconhecida.

### 2.3.5 Redução do Desequilíbrio de Ligação ao Longo das Gerações

Nos itens anteriores, o cálculo do DL foi apresentado considerando haplótipos com dois e três locos fisicamente ligados, no contexto de locos bialélicos em diplóides. A maior complexidade na detecção do DL devido à existência de locos multialélicos apenas foi comentada, de modo que nenhuma abordagem mais detalhada foi apresentada. Assim, neste tópico, será discutido o comportamento do DL ao longo de gerações de cruzamentos numa abordagem mais simples, levando em consideração pares de locos bialélicos em uma população diplóide. Oportunamente, esses conceitos são expandidos para o contexto de organismos poliplóides, como é o caso da cana-de-açúcar.

Considere dois locos fisicamente ligados $A$ e $B$, ambos bialélicos ($A$ e $a$; $B$ e $b$), numa geração $N$ qualquer de determinada população. Quatro diferentes haplótipos podem ser identificados: $AB$, $Ab$, $aB$ e $ab$. Com as diferentes combinações desses haplótipos, que apresentam, respectivamente, as frequências $f_{AB}$, $f_{Ab}$, $f_{aB}$ e $f_{ab}$, dez possíveis genótipos podem ser encontrados na população. Esses genótipos, suas respectivas frequências e as frequências esperadas dos haplótipos na geração $N+1$, levando em consideração uma fração de recombinação $r$ entre os locos $A$ e $B$, são mostrados na Tabela 2 (HARTL; CLARK, 2007; HAMILTON, 2009; HEDRICK, 2010). Cada um desses locos pode estar em equilíbrio de Hardy-Weinberg, de modo que seus alelos estejam associados aleatoriamente (loco $A$: $f_A = p$, $f_a = q$, $f_{AA} = p^2$, $f_{Aa} = 2pq$, $f_{aa} = q^2$, $p + q = 1$; loco $B$: $f_B = r$, $f_b = s$, $f_{BB} = r^2$, $f_{Bb} = 2rs$, $f_{bb} = s^2$, $r + s = 1$). Caso forem detectados desvios em relação a essas frequências, indicando a influência de forças evolutivas, esses locos não estarão em equilíbrio de Hardy-Weinberg. De qualquer forma, essa situação é novamente estabelecida com apenas uma geração de recombinação (CROW; KIMURA, 1970; SLATKIN, 2008).

Tabela 2 – Genótipos e suas frequências na geração $N$ e frequências esperadas dos haplótipos na geração $N+1$

| Genótipos | Frequências | $AB$ | $Ab$ | $aB$ | $ab$ |
|---|---|---|---|---|---|
| $AB/AB$ | $f_{AB}^2$ | $f_{AB}^2$ | – | – | – |
| $AB/Ab$ | $2f_{AB}f_{Ab}$ | $f_{AB}f_{Ab}$ | $f_{AB}f_{Ab}$ | – | – |
| $Ab/Ab$ | $f_{Ab}^2$ | – | $f_{Ab}^2$ | – | – |
| $AB/aB$ | $2f_{AB}f_{aB}$ | $f_{AB}f_{aB}$ | – | $f_{AB}f_{aB}$ | – |
| $AB/ab$ | $2f_{AB}f_{ab}$ | $(1-r)f_{AB}f_{ab}$ | $(r)f_{AB}f_{ab}$ | $(r)f_{AB}f_{ab}$ | $(1-r)f_{AB}f_{ab}$ |
| $Ab/aB$ | $2f_{Ab}f_{aB}$ | $(r)f_{Ab}f_{aB}$ | $(1-r)f_{Ab}f_{aB}$ | $(1-r)f_{Ab}f_{aB}$ | $(r)f_{Ab}f_{aB}$ |
| $Ab/ab$ | $2f_{Ab}f_{ab}$ | – | $f_{Ab}f_{ab}$ | – | $f_{Ab}f_{ab}$ |
| $aB/aB$ | $f_{aB}^2$ | – | – | $f_{aB}^2$ | – |
| $aB/ab$ | $2f_{aB}f_{ab}$ | – | – | $f_{aB}f_{ab}$ | $f_{aB}f_{ab}$ |
| $ab/ab$ | $f_{ab}^2$ | – | – | – | $f_{ab}^2$ |

De acordo com a Tabela 2, é possível obter as frequências dos haplótipos ($f^{'}$) na geração $N+1$ através da soma das respectivas colunas:

$$\begin{aligned}
\text{Haplótipo 1:}\quad & f_{AB}^{'} = f_{AB}^2 + f_{AB}f_{Ab} + f_{AB}f_{aB} + (1-r)f_{AB}f_{ab} + (r)f_{Ab}f_{aB} \\
& f_{AB}^{'} = f_{AB} - rD_N \\
\text{Haplótipo 2:}\quad & f_{Ab}^{'} = f_{AB}f_{Ab} + f_{Ab}^2 + (r)f_{AB}f_{ab} + (1-r)f_{Ab}f_{aB} + f_{Ab}f_{ab} \\
& f_{Ab}^{'} = f_{Ab} + rD_N \\
\text{Haplótipo 3:}\quad & f_{aB}^{'} = f_{AB}f_{aB} + (r)f_{AB}f_{ab} + (1-r)f_{Ab}f_{aB} + f_{aB}^2 + f_{aB}f_{ab} \\
& f_{aB}^{'} = f_{aB} + rD_N \\
\text{Haplótipo 4:}\quad & f_{ab}^{'} = (1-r)f_{AB}f_{ab} + (r)f_{Ab}f_{aB} + f_{Ab}f_{ab} + f_{aB}f_{ab} + f_{ab}^2 \\
& f_{ab}^{'} = f_{ab} - rD_N
\end{aligned}$$

onde $D_N$ é o DL referente à geração anterior $N$ e $r$ é a fração de recombinação entre os locos $A$ e $B$. Com as frequências dos haplótipos e dos alelos ($A$ e $a$; $B$ e $b$) na geração $N+1$, é possível calcular o DL resultante:

$$D_{N+1} = f_{AB}^{'}f_{ab}^{'} - f_{Ab}^{'}f_{aB}^{'}$$

Substituindo as expressões equivalentes a estes haplótipos, e após manipulações algébricas, tem-se que:

$$D_{N+1} = (1 - r)D_N$$

Generalizando:

$$D_{N+T} = (1 - r)^{N+T} D_N$$

Pela expressão anterior, é possível notar que o DL tende a se reduzir ao longo das gerações com cruzamentos aleatórios, sendo que essa redução é função do tempo ($T$) e da fração de recombinação ($r$) entre os locos. Assim, quanto menor é a fração de recombinação entre dois locos quaisquer, menor é a probabilidade de um evento de recombinação ocorrer, e portanto maior é a associação preferencial entre eles devido à ligação física, fazendo com que a queda do DL ocorra mais lentamente (LEWONTIN; KOJIMA, 1960; CROW; KIMURA, 1970; FALCONER; MACKAY, 1996; LYNCH; WALSH, 1998; GALLAIS, 2003; TEMPLETON, 2006; HARTL; CLARK, 2007; ORAGUZIE et al., 2007; SLATKIN, 2008; HAMILTON, 2009; HEDRICK, 2010).

No início, com $N$ e $T$ iguais a zero, os locos $A$ e $B$ encontram-se em máxima associação preferencial, de modo que se tem o máximo valor possível para o DL ($D_0 = 0{,}25$; frequências alélicas iguais: $f_A = f_a = f_B = f_b = 0{,}5$). Neste momento, apenas haplótipos em associação ($AB$ e $ab$; $f_{AB} = f_{ab} = 0{,}5$; $f_{Ab} = f_{aB} = 0$) constituem os genótipos de uma população, sendo que a soma das suas frequências é equivalente a 1. Com o decorrer das gerações, e com elas muitas oportunidades de recombinação, novos haplótipos começam a surgir na população, e o DL máximo inicial vai se reduzindo gradualmente. A Figura 2 ilustra essas ideias. É possível notar que a queda do DL ao longo das gerações, na ausência de outros fatores que afetam seu comportamento, é influenciada diretamente pela fração de recombinação $r$ entre os locos. Por exemplo, para locos não ligados ($r = 0{,}5$), os quais se encontram separados por maior distância, pouco mais de 5 gerações seria suficiente para quebrar o DL máximo inicial. Porém, para locos intimamente ligados ($r = 0{,}01$), onde é rara a ocorrência de um evento de recombinação, mais de 300 gerações seriam necessárias para que o DL máximo fosse totalmente rompido, o que implicaria uma situação de independência ou equilíbrio de ligação.

Com o surgimento de novos haplótipos, é possível chegar a uma situação na qual apenas haplótipos em repulsão ($Ab$ e $aB$; $f_{AB} = f_{ab} = 0$; $f_{Ab} = f_{aB} = 0{,}5$) estão presentes numa população, sendo que a soma das suas frequências é igual a 1. Neste caso, o DL atinge seu mínimo valor possível ($D_0 = -0{,}25$; frequências alélicas iguais: $f_A = f_a = f_B = f_b = 0{,}5$). É importante notar que esse valor, apesar de mínimo, corresponde à máxima associação preferencial entre os alelos de

dois locos quaisquer. Em outras palavras, a intensidade do DL é exatamente a mesma quando se tem apenas haplótipos em associação ou apenas haplótipos em repulsão numa população. A queda do DL numa população que se iniciou com haplótipos em repulsão acontece de maneira semelhante à anterior, tendo-se o gráfico da Figura 2 na parte negativa do eixo correspondente ao DL.

A redução do DL não é função apenas da ligação física entre locos (PRITCHARD; PRZEWORSKI, 2001; MCVEAN, 2007). Outras forças evolutivas ao longo das gerações podem influenciar no sentido de aumentá-lo, retardando sua redução e, assim, a situação de equilíbrio de ligação numa população. Algumas dessas forças são deriva genética, seleções natural e artificial, migração, estrutura populacional, endogamia, entre outras (ver item 2.3.7). A ação desses fatores promove o surgimento de DL tanto entre locos não ligados quanto ligados, os quais permanecem em associação preferencial mesmo após muitas gerações.

A Figura 3 mostra a queda do DL numa população sob forte influência de seleção natural. É possível observar a existência de DL entre locos não ligados ($r = 0{,}5$) e ligados ($r = 0{,}05$) mesmo após 25 gerações de cruzamentos, com magnitudes diferentes em função das diferentes frações de recombinação, mostrando claramente a interferência do processo de seleção. No entanto, para locos intimamente ligados ($r = 0{,}01$), esse fator parece não alterar de maneira significativa a queda do DL, uma vez que seu comportamento, neste caso, se mantém próximo àquele verificado na ausência de forças. Portanto, após muitas gerações de cruzamentos, com ou sem forças evolutivas, espera-se que o DL remanescente seja devido a locos intimamente ligados, ou seja, muito próximos fisicamente. Uma vez que esses locos compartilham um mesmo contexto histórico-evolutivo, o DL existente entre eles é de grande importância para o mapeamento de QTL's e a seleção assistida por marcadores moleculares.

Em poliplóides, mais especificamente tetraplóides, a queda do DL ao longo das gerações foi estudada em detalhes por Gallais (2003). Por meio de abordagem fundamentalmente teórica, esse autor mostrou que os padrões do DL em uma população tetraplóide devem apresentar maior complexidade quando comparados ao de uma diplóide, visto que os gametas, em geral, não são haplótipos. Nesse sentido, considerando que os gametas em tetraplóides são constituídos por dois cromossomos, os quais segregam conjuntamente, o DL pode ocorrer entre alelos de um mesmo loco, situados em haplótipos distintos, e entre alelos de locos diferentes, situados num mesmo haplótipo (associação) ou em haplótipos diferentes (repulsão). Assim, a caracterização completa do DL em tetraplóides, ou mesmo em outros poliplóides, deve considerar essa complexidade. Contudo,

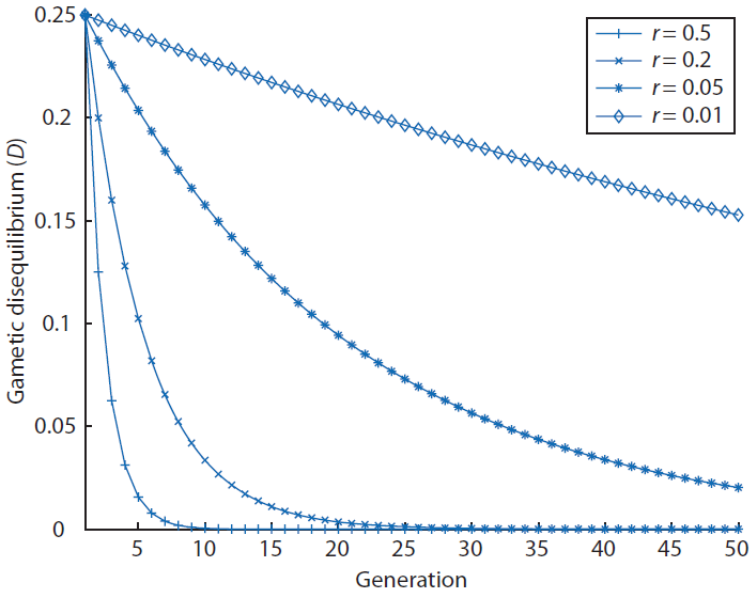


Figura 2 – Redução do desequilíbrio de ligação (DL) ao longo das gerações considerando quatro frações de recombinação ($r$). No início, apenas haplótipos em associação são observados na população ($D = 0{,}25$). A queda do DL acontece em função do tempo e da fração de recombinação ($D_{N+T} = (1-r)^{N+T} D_N$), assumindo cruzamentos aleatórios e ausência de forças evolutivas (HAMILTON, 2009)

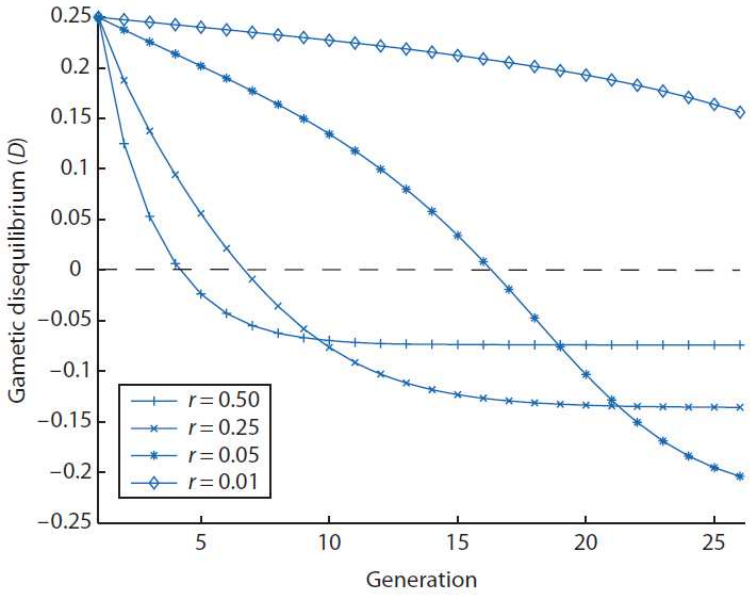


Figura 3 – Redução do desequilíbrio de ligação (DL) ao longo das gerações sob forte influência de seleção natural. Neste caso, força evolutiva está atuando juntamente com a fração de recombinação na queda do DL. Ao contrário da Figura 2, o DL não é reduzido a zero devido à ação desta força (HAMILTON, 2009)

Gallais (2003) apresentou a queda do DL para esta ploidia apenas considerando alelos de diferentes locos em associação, conforme pode ser observado na Figura 4. É possível notar, neste caso, a semelhança na queda do DL entre diplóide e tetraplóide.

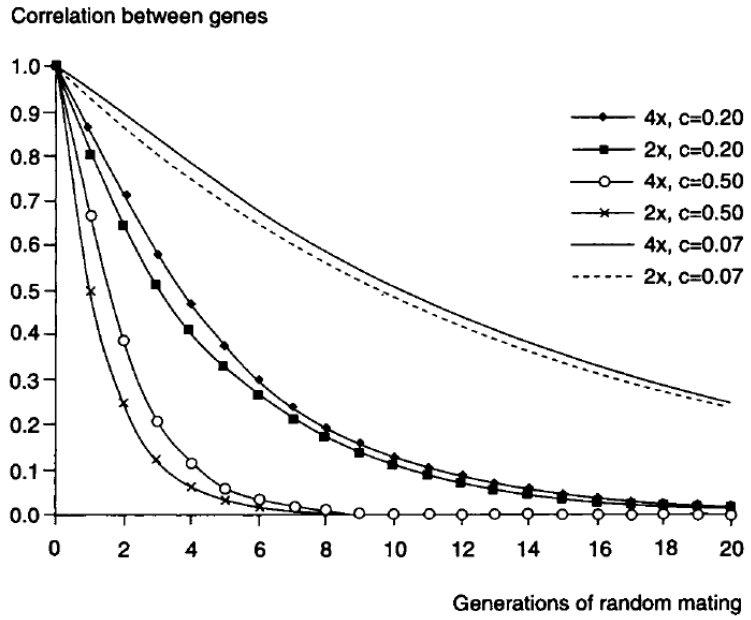


Figura 4 – Redução do desequilíbrio de ligação (DL) ao longo das gerações em populações diplóide e tetraplóide. Assumem-se: diferentes frações de recombinação entre um possível par de locos, as quais, neste caso, são indicadas pela letra $c$; cruzamentos aleatórios; e ausência de forças evolutivas (GALLAIS, 2003)

### 2.3.6 Medidas Relativas do Desequilíbrio de Ligação

O conceito básico do DL (LEWONTIN; KOJIMA, 1960), conforme exposto em tópicos anteriores, pode não ser procedimento adequado quando o objetivo é comparar a magnitude das associações preferenciais entre diferentes pares de locos. Isso acontece porque esse DL é calculado com base nas frequências alélicas de determinado par de locos, as quais podem naturalmente variar quando outro par de locos está sendo considerado. O exemplo a seguir mostra essas ideias.

Considere dois pares de locos fisicamente ligados: $A$ e $B$, ambos bialélicos ($A$ e $a$; $B$ e $b$); $P$ e $Q$, também ambos bialélicos ($P$ e $p$; $Q$ e $q$). Para esses pares, quatro diferentes haplótipos podem ser identificados numa geração qualquer de determinada população: $AB$, $Ab$, $aB$ e $ab$, para os locos $A$ e $B$; $PQ$, $Pq$, $pQ$ e $pq$, para os locos $P$ e $Q$. Esses locos podem estar em DL, e os respectivos valores desse parâmetro podem ser calculados através das frequências dos seus haplótipos:

$$\text{Locos } A \text{ e } B\text{: } D_{AB} = f_{AB}f_{ab} - f_{Ab}f_{aB}$$

$$\text{Locos } P \text{ e } Q\text{: } D_{PQ} = f_{PQ}f_{pq} - f_{Pq}f_{pQ}$$

Através desses cálculos, supondo significância estatística, é possível verificar qual dos pares de locos apresenta maior grau de DL, ou seja, maior associação preferencial entre os seus alelos. No

entanto, essa comparação não é apropriada, uma vez que as frequências alélicas referentes a um par de locos podem não ser as mesmas referentes a outro par de locos (HEDRICK, 1987; LEWONTIN, 1988). Em outras palavras, o máximo DL que o par de locos $AB$ pode apresentar não é necessariamente o mesmo máximo que o par de locos $PQ$ pode apresentar, dificultando afirmar, pela medida do DL acima, qual dos pares está mais fortemente associado devido à ligação física.

Assim, com o objetivo de possibilitar a comparação do DL entre pares de locos numa população, ou até mesmo de um loco em diferentes populações, contendo diferentes frequências alélicas, outras medidas foram propostas. Essas medidas foram revisadas em detalhes por Hedrick (1987) e Devlin e Risch (1995). Hedrick (1987) e Lewontin (1988) chamam atenção às dificuldades para definir a mais apropriada para o cálculo do DL, uma vez que, em partes, são sensíveis a diferentes processos populacionais que geram associações preferenciais (MCVEAN, 2007; SLATKIN, 2008). Na sequência, são apresentadas as medidas relativas mais comuns utilizadas para o cálculo do DL.

Lewontin (1964) propôs uma medida relativa do DL, designada por $D^{'}$, que possibilita comparar os DL's de diferentes pares de locos, uma vez que os máximos valores teóricos são levados em consideração, conforme mostra a expressão a seguir.

$$D^{'} = \frac{\mid D \mid}{D^{max}} \tag{1}$$

onde $D^{'}$ é a medida relativa do DL, $D$ é o conceito básico calculado com base nas frequências dos haplótipos, e $D^{max}$ é o máximo valor teórico do DL entre um possível par de locos.

Para os pares de locos considerados ($A$ e $B$; $P$ e $Q$), tém-se que:

$$D^{'}_{AB} = \frac{\mid D_{AB} \mid}{D^{max}_{AB}} \tag{2}$$

$$D^{'}_{PQ} = \frac{\mid D_{PQ} \mid}{D^{max}_{PQ}} \tag{3}$$

Quando $D_{AB}$ e $D_{PQ}$ forem $> 0$, os máximos teóricos $D^{max}_{AB}$ e $D^{max}_{PQ}$ serão equivalentes aos menores valores dentre os produtos $f_A f_b$ e $f_a f_B$, e $f_P f_q$ e $f_p f_Q$, respectivamente. Já, quando $D_{AB}$ e $D_{PQ}$ forem $< 0$, os máximos teóricos $D^{max}_{AB}$ e $D^{max}_{PQ}$ serão correspondentes aos menores valores dentre os produtos $f_A f_B$ e $f_a f_b$, e $f_P f_Q$ e $f_p f_q$, respectivamente. De outra forma, tém-se que:

$$D_{AB}^{max} = \begin{cases} min(f_A f_B, f_a f_b), & \text{se } D_{AB} \text{ é < 0} \\ min(f_A f_b, f_a f_B), & \text{se } D_{AB} \text{ é > 0} \end{cases}$$

$$D_{PQ}^{max} = \begin{cases} min(f_P f_Q, f_p f_q), & \text{se } D_{PQ} \text{ é < 0} \\ min(f_P f_q, f_p f_Q), & \text{se } D_{PQ} \text{ é > 0} \end{cases}$$

Ao contrário do conceito básico do DL, que pode variar entre $-0{,}25$ e $0{,}25$, o $D^{'}$, que é uma medida relativa, pode variar entre 0 e 1 ($0 \leq D^{'} \leq 1$). Com $D^{'} = 1$, indicando que a associação preferencial entre alelos é equivalente à máxima teórica possível, pode-se dizer que pelo menos um dos quatro haplótipos está ausente, o que caracteriza uma situação de perfeito DL (ARDLIE; KRUGLYAK; SEIELSTAD, 2002; SLATKIN, 2008; HEDRICK, 2010). Quando todos os haplótipos, em associação e repulsão, estão presentes, $D^{'}$ é menor que 1, indicando que eventos de mutação ou recombinação ocorreram ao longo das gerações, o que não caracteriza uma situação de máxima associação preferencial (FLINT-GARCIA; THORNSBERRY; BUCKLER, 2003; SLATKIN, 2008). Assim, um procedimento direto para saber se os quatro haplótipos estão presentes numa população é verificar o valor do $D^{'}$. De qualquer maneira, quando $D^{'}$ é menor que 1, a magnitude dos valores obtidos não apresenta clara interpretação (ARDLIE; KRUGLYAK; SEIELSTAD, 2002).

Outra medida relativa do DL é o $r^2$, ou $\Delta^2$, proposta por Hill e Robertson (1968). É normalmente indicada pelo quadrado do coeficiente de correlação, e mede o grau de associação entre locos (*covariância*) de acordo com a variação dos seus alelos. Assim como a anterior, essa medida depende do conceito básico do DL, o qual é baseado nas frequências dos haplótipos. A expressão a seguir mostra essas ideias.

$$r_{XY}^2 = \frac{Cov(X,Y)}{V(X)V(Y)} = \frac{D_{XY}^2}{f_X f_x f_Y f_y} = \frac{D_{XY}^2}{f_X(1-f_X)f_Y(1-f_Y)} \tag{4}$$

onde $r_{XY}^2$ (coeficiente de determinação, ou quadrado do coeficiente de correlação) é a medida relativa do DL entre dois locos bialélicos quaisquer $X$ e $Y$, $Cov(X,Y)$ é a *covariância* entre esses locos, $V(X)$ e $V(Y)$ são suas respectivas variâncias, $D_{XY}^2$ é o conceito básico do DL entre X e Y, e $f_{X,x,Y,y}$ são as frequências dos alelos referentes aos locos X e Y ($X$: $f_X + f_x = 1$, $V(X) = f_X f_x$; $Y$: $f_Y + f_y = 1$; $V(Y) = f_Y f_y$)

Para os locos considerados ($A$ e $B$; $P$ e $Q$), tém-se que:

$$r^2_{AB} = \frac{Cov(A,B)}{V(A)V(B)} = \frac{D^2_{AB}}{f_A f_a f_B f_b} = \frac{D^2_{AB}}{f_A(1-f_A)f_B(1-f_B)} \quad (5)$$

$$r^2_{PQ} = \frac{Cov(P,Q)}{V(P)V(Q)} = \frac{D^2_{PQ}}{f_P f_p f_Q f_q} = \frac{D^2_{PQ}}{f_P(1-f_P)f_Q(1-f_Q)} \quad (6)$$

Do mesmo modo que $D'$, $r^2$ pode variar entre 0 e 1 ($0 \leq r^2 \leq 1$). Com $r^2 = 1$, indicando associação preferencial máxima entre alelos de pares de locos, pode-se dizer que dois dos quatro haplótipos estão ausentes na amostra, o que caracteriza uma situação de perfeito DL (ARDLIE; KRUGLYAK; SEIELSTAD, 2002). Neste caso, apenas haplótipos em associação podem estar presentes, mostrando que eventos de mutação, recombinação ou migração não ocorreram no sentido de gerar haplótipos em repulsão. No contexto da seleção assistida por marcadores moleculares, informações a respeito de um loco fornecem completa informação a respeito do outro loco, sendo redundante a utilização de ambos no mapeamento de QTL's.

$D'$ e $r^2$ são as medidas mais comuns utilizadas para o cálculo do DL entre pares de locos bialélicos (ARDLIE; KRUGLYAK; SEIELSTAD, 2002; FLINT-GARCIA; THORNSBERRY; BUCKLER, 2003; GUPTA; RUSTGI; KULWAL, 2005; HARTL; CLARK, 2007; HEDRICK, 2010). Embora ambas não sejam apropriadas para verificá-lo em amostras pequenas, apresentando ou não baixas frequências alélicas, cada qual possui suas vantagens. Enquanto $r^2$ capitaliza eventos de mutação e recombinação ocorridos historicamente, $D'$ apenas capitaliza eventos de recombinação, sendo medida mais adequada para detectá-los. No entanto, $D'$ é fortemente afetada por amostras pequenas, de modo que a comparação entre locos contendo baixas frequências alélicas torna-se inadequada com essa medida em função das estimativas altamente viciadas do DL (ARDLIE; KRUGLYAK; SEIELSTAD, 2002; FLINT-GARCIA; THORNSBERRY; BUCKLER, 2003). Portanto, no caso dos estudos de associação, onde o tamanho da amostra é limitado por questões de tempo e custo envolvidos nos processos de genotipagem e fenotipagem, deve-se priorizar o $r^2$ para verificar a extensão do DL (JORDE, 1995; ARDLIE; KRUGLYAK; SEIELSTAD, 2002; FLINT-GARCIA; THORNSBERRY; BUCKLER, 2003; ZHU et al., 2008).

A Figura 5 mostra três situações de como locos fisicamente ligados podem exibir diferentes níveis de DL através das medidas $D'$ e $r^2$. Contudo, essas medidas foram desenvolvidas no contexto de pares de locos bialélicos e, por isso, não são apropriadas para o cálculo do DL entre pares de locos

multialélicos ou vários locos simultaneamente, sendo estes bialélicos ou não. Algumas medidas até chegaram a ser propostas para o primeiro (HEDRICK, 1987) e segundo (ROBINSON et al., 1991) casos, mas parece que não tiveram grande aceitação. Assim, outros métodos têm sido propostos principalmente para o cálculo do DL entre múltiplos locos/sítios (KIM et al., 2008), o qual está sendo alvo de intensa pesquisa a partir do sequenciamento de SNPs em humanos, como parte do Projeto Internacional HapMap.

### 2.3.7 Fatores que Afetam o Desequilíbrio de Ligação

Décadas atrás, quando a disponibilidade de dados moleculares ainda era limitada, o DL detectado em uma população era normalmente atribuído à seleção simultânea de locos. Nos dias de hoje, em especial para locos intimamente ligados, é reconhecido que mutação e deriva genética são fatores importantes no surgimento de DL, assim como recombinação na sua queda ao longo das gerações (HEDRICK, 2010). No entanto, esses fatores não são os únicos. Tamanho populacional, estrutura de população, endogamia, sistema reprodutivo, inversão e conversão de genes, entre outros, também podem afetar o DL de uma população (PRITCHARD; PRZEWORSKI, 2001; ARDLIE; KRUGLYAK; SEIELSTAD, 2002; GALLAIS, 2003; HARTL; CLARK, 2007; SLATKIN, 2008; HEDRICK, 2010). Alguns desses fatores são discutidos a seguir.

As seleções natural e artificial podem gerar DL entre locos ligados e não ligados. No primeiro caso, a seleção estabelece associações preferenciais entre um alelo favorável de determinado loco e alelos de locos adjacentes, de modo que haplótipos inteiros contendo o alelo selecionado permanecem na população. Consequentemente, haplótipos inteiros contendo o alelo deletério são eliminados da população ("bottleneck" loco-específico). No segundo caso, a seleção a favor ou contra um caráter controlado por locos não ligados promove DL entre seus alelos (FLINT-GARCIA; THORNSBERRY; BUCKLER, 2003), o qual não é de interesse para o mapeamento de QTL's.

A mutação exerce papel fundamental na criação de DL (SLATKIN, 2008). Considere dois locos ligados $A$ e $B$, sendo o primeiro polimórfico, com alelos $A$ e $a$, e o segundo monomórfico, com alelo $B$. Dois haplótipos - $AB$ e $aB$ - constituem os genótipos de uma população. Supondo a ocorrência de mutação no loco $B$, este se torna polimórfico, com alelos $B$ e $b$. No início, o alelo $b$ surge em apenas um cromossomo, o qual pode conter os alelos $A$ ou $a$. Em função disso, a frequência do haplótipo contendo $b$ é claramente inferior à esperada, mostrando que esse alelo encontra-se em DL com os demais alelos do cromossomo. Três haplótipos - $AB$, $aB$ e ($Ab$ ou $ab$) - passam a

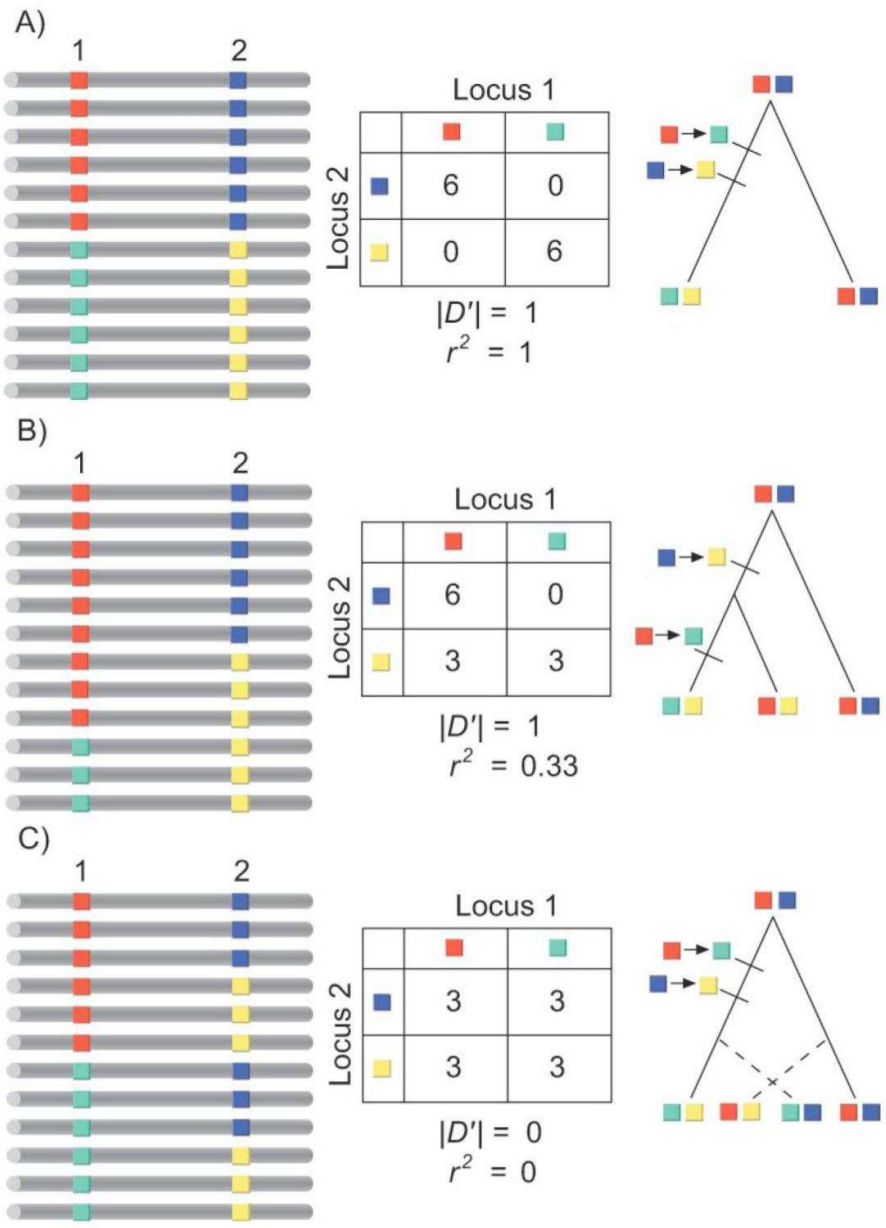


Figura 5 – Três cenários hipotéticos de desequilíbrio de ligação (DL) entre dois locos causados por eventos de mutação e recombinação históricos, mostrando o comportamento das medidas $D^{'}$ e $r^2$. Imagens na coluna esquerda mostram os estados alélicos dos locos 1 e 2. A coluna do meio representa a tabela de contingência 2 x 2 de haplótipos e os resultados de $D^{'}$ e $r^2$. A coluna da direita representa uma possível árvore mostrando o DL presente. **(A)** DL absoluto existente quando dois locos compartilham o mesmo histórico de mutações na ausência de recombinação. Ambos $D^{'}$ e $r^2$ são equivalentes a 1. **(B)** DL existente entre os locos mesmo quando os eventos de mutação não são totalmente compartilhados; ausência de recombinação. Note que os diferentes eventos de mutação são detectados apenas por $r^2$. **(C)** Equilíbrio de ligação (DL = 0) entre os locos com a ocorrência de recombinação, independente dos eventos de mutação. Ambos $D^{'}$ e $r^2$ são iguais a 0. (FLINT-GARCIA; THORNSBERRY; BUCKLER, 2003)

constituir os genótipos, sendo que o grau do DL entre $A$ e $B$ é considerado máximo. Contudo, o DL é máximo apenas considerando a medida $D^{'}$, visto que esta não capitaliza eventos de mutação,

os quais podem reduzir o DL. Assim, a mutação é um fator importante não apenas por gerar DL, mas também por reduzi-lo ao longo das gerações (ARDLIE; KRUGLYAK; SEIELSTAD, 2002; HARTL; CLARK, 2007; HAMILTON, 2009; HEDRICK, 2010). Por exemplo, SNPs com altas taxas de mutação podem apresentar nenhum ou pouco DL com marcadores adjacentes, mesmo na ausência de recombinação. Porém, como mutação é rara em SNPs (PRITCHARD; PRZEWORSKI, 2001; ARDLIE; KRUGLYAK; SEIELSTAD, 2002; SLATKIN, 2008), não há evidências que esse fator contribua de maneira significativa para romper o DL entre esses marcadores.

A deriva genética promove a flutuação aleatória de alelos e haplótipos ao longo das gerações, fazendo com que locos se associem de maneira não-aleatória (ARDLIE; KRUGLYAK; SEIELSTAD, 2002). Segundo Slatkin (2008), o efeito deste fator é similar à retirada de uma amostra pequena a partir de uma grande população. Neste caso, até mesmo locos em equilíbrio podem apresentar algum DL, devido à flutuação gerada por uma amostra deficiente. Apesar de estudos mostrarem que a deriva genética, ao interagir com mutação e recombinação, poderia ser ignorada como causa do DL (HILL; ROBERTSON, 1968; OTHA; KIMURA, 1969), sua importância pode ser verificada em diferentes situações. Por exemplo, em populações que apresentam menor tamanho efetivo ($N_e$), a deriva genética tende a ser pronunciada, promovendo aumento do DL em função da perda de alelos e haplótipos raros (FLINT-GARCIA; THORNSBERRY; BUCKLER, 2003). Porém, em populações que possuem maior tamanho efetivo, a deriva tende a ser reduzida, de modo que o DL esperado entre locos também tende a ser. Assim, a influência da deriva no comportamento do DL pode ser maior ou menor dependendo do tamanho da população em estudo.

Nesse contexto, o tamanho populacional é um fator importante que afeta o DL. Em populações que se encontram em expansão, derivadas de eventos demográficos recentes, ou que permanecem isoladas, tendo-se menor tamanho efetivo, é esperado encontrar DL em maiores extensões ao longo do genoma (MACKAY, 2001a). No entanto, em populações maiores, as quais possuem maior tamanho efetivo e provavelmente passaram por muitas gerações e oportunidades de recombinação, é esperado encontrar DL em menores distâncias no genoma (SLATKIN, 1994; KRUGLYAK, 1999; PRITCHARD; PRZEWORSKI, 2001; MACKAY, 2001a). Até mesmo em grandes populações que apresentam tamanho efetivo constante é esperado encontrar DL reduzido, presente entre locos intimamente ligados (PRITCHARD; PRZEWORSKI, 2001).

Contudo, grandes populações podem sofrer reduções bruscas no seu tamanho (“bottleneck”), como por exemplo um isolamento geográfico, acarretando na intensa perda de alelos e haplótipos.

Essa perda promove alterações nas frequências alélicas e dos haplótipos, culminando em associações preferenciais. Assim, locos que encontravam-se em equilíbrio passam a estar em DL, e este eleva-se pela ação da deriva genética na população subsequente, a qual apresenta-se com menor tamanho efetivo. A ocorrência de um "bottleneck" inicial tem sido a explicação do DL detectado entre locos distantes em vários estudos com humanos (SLATKIN, 2008).

Outro fator que afeta substancialmente o DL é a estrutura de população. Possivelmente, esta é a principal causa da ocorrência de falsos positivos em estudos de mapeamento associativo (ABDURAKHMONOV; ABDUKARIMOV, 2008). A mistura de indivíduos oriundos de diferentes fragmentos homogêneos (subpopulações) de uma população pode culminar em associações preferenciais. O efeito dessa mistura, a qual pode ou não ser intencional, é evidente em um caso extremo (SLATKIN, 2008). Por exemplo, para dois locos bialélicos $A$ e $B$, considere que uma subpopulação é fixada para os alelos $A$ e $B$, enquanto outra é fixada para os alelos $a$ e $b$. Dentro de cada subpopulação, o DL entre os locos $A$ e $B$ é zero. Com a migração de indivíduos de diferentes subpopulações, dois haplótipos em associação - $AB$ e $ab$ - passam a constituir a população misturada, de modo que o máximo DL é estabelecido. No entanto, como esse é um caso extremo, é provável que outros locos estejam em DL dentro de cada subpopulação, mesmo não estando fisicamente ligados. Isso pode ocorrer devido à ação de outras forças ou poucas oportunidades de recombinação. De qualquer forma, como DL entre locos não ligados (falsas associações) pode surgir pela mistura de diferentes subpopulações, é imprescindível que a estrutura de população seja considerada nos estudos de associação (FLINT-GARCIA; THORNSBERRY; BUCKLER, 2003).

Em geral, o DL ocorre de forma aleatória no genoma e depende da espécie e população sob estudo (ORAGUZIE et al., 2007). Assim, o estudo da extensão do DL é muito importante, e possibilita determinar a resolução de mapeamento e a densidade de marcadores necessários na identificação de QTL's (NORDBORG et al., 2002; WEI et al., 2006; ZHU et al., 2008). Se o DL permanecer em curtas distâncias no genoma, uma alta resolução de mapeamento será esperada, mas uma elevada quantidade de marcadores será necessária. Em contrapartida, se o DL se estender a maiores distâncias no genoma, a resolução de mapeamento tenderá a ser baixa, mas uma quantidade menor de marcadores será requerida (YU; BUCKLER, 2006; ORAGUZIE et al., 2007; ZHU et al., 2008).

### 2.3.8 Desequilíbrio de Ligação em Plantas

A extensão do DL têm sido estudada em diversas espécies de plantas. Em *Arabidopsis*, uma

espécie autógama, o DL se estende até cerca de 50 Kb (NORDBORG et al., 2002, 2005). Alguns estudos entre etnovariedades de arroz e sorgo, espécies predominantemente autógamas, mostram uma extensão de DL de até 100 Kb (GARRIS et al., 2003; HAMBLIN et al., 2005). As culturas que passaram por um processo de domesticação sofreram gargalos severos na diversidade genética, aumentando a extensão do DL. Além disso, o melhoramento moderno tem elevado o DL pelo uso de um número restrito de parentais nos processos de hibridação. Em milho, uma espécie alógama, há um declínio rápido no DL em algumas centenas de pares de bases (TENAILLON et al., 2001), o que muito dificulta o mapeamento por associação, já que milhares de marcadores são necessários. Em contrapartida, os materiais de melhoramento de milho podem apresentar DL em distâncias de até 33 cM (STICH et al., 2005), sendo que, entre materiais elite, não foi observada queda no DL até 500 pb (CHING et al., 2002). O mesmo foi observado entre etnovariedades de algodão, onde o DL se estendeu até 10 cM; entre variedades melhoradas, o DL se estendeu até 30 cM (ABDURAKHMONOV; ABDUKARIMOV, 2008). Em batata, uma espécie alógama e poliplóide, o DL permaneceu ao redor de 5 cM (D'HOOP et al., 2010), sendo necessária uma quantidade maior de marcadores para o mapeamento. Detalhes sobre a extensão do DL em outras espécies de plantas podem ser encontrados em Flint-Garcia, Thornsberry e Buckler (2003), Gupta, Rustgi e Kulwal (2005), Oraguzie et al. (2007), Abdurakhmonov e Abdukarimov (2008) e Zhu et al. (2008).

#### 2.3.8.1 Desequilíbrio de Ligação em Cana-de-açúcar

O primeiro trabalho sobre extensão do DL em cana-de-açúcar foi publicado por Jannoo et al. (1999b). Esses autores, ao estudarem 59 variedades cultivadas nas ilhas Maurício, detectaram DL entre marcadores RFLP em uma distância de até 10 cM, alegando como causa a ocorrência de "bottleneck" fundador entre as variedades ancestrais. Anos mais tarde, Raboin et al. (2008), ao estudarem 72 clones de diversas estações de melhoramento do mundo, detectaram forte DL nos primeiros 5 cM de distância, através da análise utilizando 1537 marcadores AFLP. No entanto, associações em uma distância de até 50 cM também foram detectadas, apesar de uma queda ter sido observada após 30 cM. Os autores discutiram que esse cenário era o esperado em função da história recente do melhoramento de cana-de-açúcar, a qual é caracterizada por poucas gerações segregantes (10 ou menos) devido à propagação vegetativa da espécie. Além disso, o uso recorrente de variedades superiores nos cruzamentos também deve ter contribuído para encontrar DL em maiores extensões. Contudo, como grande parte do DL permaneceu em menores distâncias, os au-

tores propuseram uma quantidade maior de marcadores para aumentar a resolução de mapeamento. Esses resultados também foram verificados por Wei et al. (2010), os quais calcularam a extensão do DL com base em 16891 marcadores DArT (Diversity Arrays Technology) e uma grande população (480 clones) de cana-de-açúcar da Austrália, a fim de realizar mapeamento por associação.

No entanto, esses trabalhos foram realizados com base em germoplasma diferente do que vem sendo explorado pelos programas de melhoramento brasileiros, mostrando a importância do presente estudo como base para o mapeamento associativo no Brasil.

## 2.4 Mapeamento Associativo

### 2.4.1 Mapeamento por Análise de Ligação vs Mapeamento Associativo

Em plantas, o mapeamento por análise de ligação baseia-se em populações experimentais oriundas de cruzamentos controlados, sendo que a busca por QTL's acontece em genomas que apresentam blocos gênicos maiores, em função das poucas oportunidades de recombinação (Figura 6a). Em contrapartida, o mapeamento associativo baseia-se em populações naturais, coleções de germoplasmas, conjuntos de materiais elite, entre outros, de modo que a busca por QTL's acontece em genomas que possuem blocos gênicos menores, devido aos eventos de recombinação histórico-evolutivos (Figura 6b). Neste caso, a resolução de mapeamento tende a ser elevada, gerando maior precisão na localização dos QTL's (ZHU et al., 2008; MACKAY et al., 2009).

Embora ambas as abordagens sejam baseadas no DL por ligação física entre marcador e QTL, bem como característica (LYNCH; WALSH, 1998), no mapeamento associativo é possível que DL entre locos não ligados seja considerado caso a estrutura da população em estudo não seja controlada (FLINT-GARCIA; THORNSBERRY; BUCKLER, 2003; YU; BUCKLER, 2006; ZHU et al., 2008). Esse controle significa considerar possíveis subpopulações ou fragmentos homogêneos dentro de uma população heterogênea, de modo que a busca por QTL's, neste caso, seja realizada no interior dessas subpopulações. Além disso, quando possível, é fundamental controlar outros fatores que também podem gerar falsas associações (YU; BUCKLER, 2006; ZHU et al., 2008).

Em geral, o mapeamento associativo, quando comparado às análises de ligação, pode apresentar como vantagens: i) detecção de associações genéticas válidas para toda a população e não somente para um cruzamento específico; ii) maior precisão e resolução de mapeamento na identificação de QTL's; e iii) redução dos custos e do tempo da pesquisa, pois não há necessidade do desenvolvi-

mento de populações experimentais (YU; BUCKLER, 2006; ABDURAKHMONOV; ABDUKARIMOV, 2008; ZHU et al., 2008).

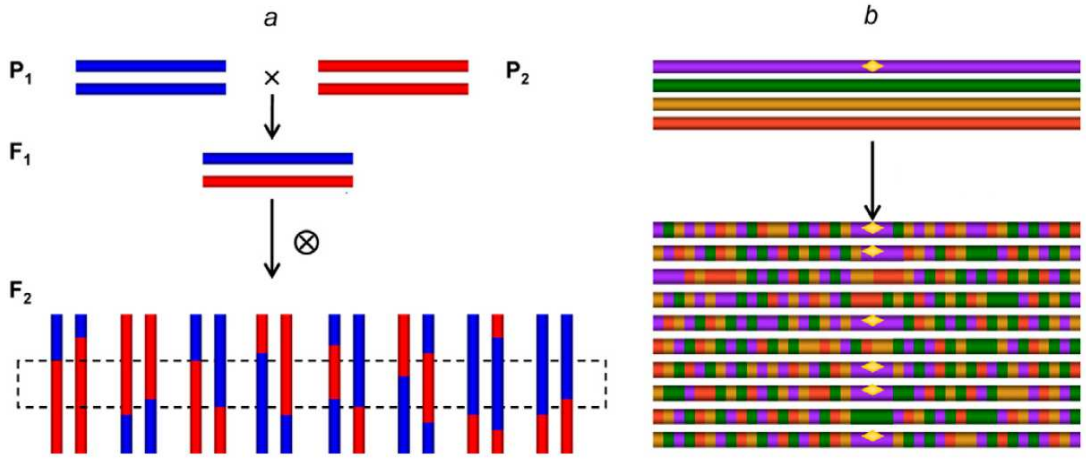


Figura 6 – Comparação entre os mapeamentos por análise de ligação ($a$) e associação ($b$). O primeiro é baseado em populações experimentais (no caso, $F_2$) e poucas oportunidades de recombinação, enquanto o segundo é baseado em populações com várias gerações e muitas oportunidades de recombinação. O diamante amarelo à direita representa o DL entre um alelo mutante e marcadores intimamente ligados, os quais permanecem em um mesmo bloco gênico ao longo das gerações, mostrando que é rara a ocorrência de recombinação quando locos estão muito próximos fisicamente (ZHU et al., 2008)

### 2.4.2 Mapeamento Associativo em Plantas

A crescente utilização do mapeamento associativo em humanos, principalmente para doenças importantes, tem estimulado seu uso em diversas espécies de plantas (BUTTERFIELD, 2007; ZHU et al., 2008). Em milho e *Arabidopsis*, espécies precursoras nesses estudos, o mapeamento associativo foi inicialmente realizado com base em genes candidatos, os quais mostraram-se associados com características importantes, como tempo de florescimento e peso de plantas. Nos últimos anos, com o surgimento de plataformas mais eficientes e menos onerosas ao processo de genotipagem, o mapeamento associativo passou a ser realizado ao longo de todo o genoma. A primeira espécie mapeada com o uso dessa abordagem, a qual consiste na genotipagem de uma série de indivíduos com marcadores moleculares, foi a beterraba. Posteriormente, outras espécies, como milho, *Arabidopsis*, arroz, trigo, aveia, eucalipto e batata também foram mapeadas para características importantes. Detalhes a respeito do mapeamento dessas e outras espécies podem ser encontrados em Flint-Garcia, Thornsberry e Buckler (2003), Oraguzie et al. (2007), Butterfield (2007), Abdurakhmonov e Abdukarimov (2008) e Zhu et al. (2008).

#### 2.4.2.1 Mapeamento Associativo em Cana-de-açúcar

O primeiro trabalho sobre mapeamento associativo em cana-de-açúcar foi publicado por Wei et al. (2006). Esses autores, baseado em um painel contendo 154 clones, detectaram 8 subpopulações com o uso do software STRUCTURE e, a partir delas, possíveis associações entre marcadores AFLP e SSR e resistência às principais doenças na Austrália. No entanto, como os próprios autores discutiram, essas subpopulações podem não ter sido fidedignas, considerando a natureza dos dados e a complexidade genética da espécie, aumentando as chances de falsas associações terem sido observadas. Além disso, discutiram que não seria necessária grande densidade de marcadores para o mapeamento associativo em cana-de-açúcar, ressaltando que um número superior ao utilizado (1068 AFLPs e 141 SSRs) seria fundamental para encontrar associações mais consistentes.

Em um trabalho similar, Butterfield (2007) utilizou 275 RFLPs e 1.056 AFLPs para detectar associações com doenças importantes da cana-de-açúcar na África do Sul, como carvão e broca do caule, baseado em um painel constituído por 77 variedades do programa de melhoramento estabelecido naquele país. Apesar de associações significativas terem sido observadas para ambas as doenças, nenhuma estrutura de população foi detectada, a qual foi verificada com base na dissimilaridade de Dice e na análise de agrupamento com o método Neighbor-Joining. Assim, por mais que o autor tenha discutido o contrário, falsas associações podem ter sido consideradas, principalmente pela pouca informatividade dos dados em função da natureza poliplóide da cana-de-açúcar.

Em um estudo mais recente, Wei et al. (2010) utilizaram 16.891 marcadores DArT (1.531 dominantes e 15.360 codominantes) e uma grande população (480 clones) da Austrália para realizar mapeamento associativo em cana-de-açúcar. Por meio da abordagem de modelos mistos, esses autores detectaram diversos marcadores associados com características agroindustriais importantes, como toneladas de cana por hectare e teor de sacarose. No entanto, nenhuma estrutura de população consistente foi detectada com base no software STRUCTURE e em informações de *pedigree*, mesmo utilizando marcadores mais informativos, mostrando que muitas das associações observadas poderão não ser úteis para a seleção assistida por marcadores em cana-de-açúcar.

No Brasil, o primeiro estudo sobre mapeamento associativo em cana-de-açúcar foi realizado por Lopes (2011). Com base em um grupo de 103 indivíduos estabelecidos em diferentes regiões de cultivo, esse autor detectou marcadores SSRs e AFLPs, que apresentaram comportamento dominante, associados com características importantes, como número de colmos, toneladas de cana por hectare e teor de sacarose. Essa detecção foi realizada através de modelos mistos, levando em consideração

informações a respeito da estrutura populacional, que foi verificada através dos métodos baseados em dissimilaridade genética, componentes principais e inferência bayesiana, este último com o uso do modelo de mistura implementado no STRUCTURE. Apesar do autor ter detectado 3 subpopulações no grupo de indivíduos considerado, acredita-se que o uso dos componentes principais e do modelo de mistura podem não ser apropriados com base em marcadores dominantes, visto que não é possível estimar frequências alélicas no contexto poliplóide da cana-de-açúcar.

Com base nesses e nos estudos sobre DL, dois aspectos são imprescindíveis para o mapeamento associativo em cana-de-açúcar: i) maior saturação do genoma através do uso de milhares de marcadores, aumentando a resolução de mapeamento e, assim, a detecção entre marcadores e QTL's intimamente ligados; e ii) controle eficiente da estrutura de população, de modo que apenas associações entre locos ligados sejam consideradas para a seleção assistida por marcadores.

## 2.5 Métodos para o Controle da Estrutura Populacional

Vários métodos baseados em marcadores moleculares têm sido desenvolvidos para o controle da estrutura populacional, visando principalmente estudos de mapeamento associativo (ZHU et al., 2008; ODONG et al., 2011). Entre esses métodos, podem ser enumerados aqueles que têm sido os mais utilizados: i) método baseado em inferência bayesiana, implementado no software STRUCTURE (PRITCHARD; STEPHENS, DONNELLY, 2000; FALUSH; STEPHENS; PRITCHARD, 2003; FALUSH; STEPHENS; PRITCHARD, 2007; HUBISZ et al., 2009); ii) método baseado em modelos mistos (YU et al., 2006); e iii) método baseado em componentes principais (PATTERSON; PRICE; REICH, 2006; PRICE et al., 2006). Além desses métodos, que são específicos, agrupamentos hierárquicos clássicos, baseados em distância ou dissimilaridade genética, também têm sido utilizados para o controle da estrutura populacional (ODONG et al., 2011).

O método probabilístico implementado no STRUCTURE tem sido muito utilizado para detectar estrutura populacional em humanos e plantas. Seu princípio é baseado na utilização de marcadores não ligados (PRITCHARD; STEPHENS, DONNELLY, 2000) ou fracamente ligados (FALUSH; STEPHENS; PRITCHARD, 2003) para inferir de maneira iterativa o número de populações mais provável a partir de um determinado grupo de indivíduos. Essa inferência pode ser realizada através de um modelo de não-mistura, onde indivíduos são atribuídos exclusivamente a uma determinada população, ou um modelo de mistura, onde os indivíduos podem apresentar proporções genômicas de diferentes populações. Porém, por depender de fatores que são inerentes

ao tipo de análise realizada, é prudente considerar que os resultados desse método podem não ser apropriados (PRITCHARD; STEPHENS, DONNELLY, 2000). Além disso, seu uso pode não ser satisfatório para grandes quantidades de dados, visto que um tempo muito elevado normalmente é necessário para obtenção das estimativas (PATTERSON; PRICE; REICH, 2006).

O método baseado em modelos mistos possibilita não apenas estimar a estrutura populacional de um determinado grupo de indivíduos, mas também a matriz de parentesco existente entre os mesmos. Essas estimativas são posteriormente inseridas como efeitos de um modelo misto típico, o qual é utilizado para a realização do mapeamento associativo. Face às suas propriedades, esse método pode ser uma estratégia interessante para complementar outros existentes (YU et al., 2006), visto que a grande maioria das populações de plantas, por exemplo, apresenta certo nível de estruturação e parentesco entre os seus indivíduos (YU; BUCKLER, 2006; ZHU et al., 2008).

O método dos componentes principais têm sido considerado uma rápida e importante ferramenta para a detecção da estrutura populacional (PATTERSON; PRICE; REICH, 2006; PRICE et al., 2006). Em linhas gerais, esse método possibilita sumarizar, em poucos componentes, a variação existente em um determinado conjunto de marcadores, obtido para um determinado grupo de indivíduos, de modo que é possível utilizar esses componentes para inferir a participação de cada um dos indivíduos em determinadas populações. Por ser um método baseado em frequências alélicas, não pode ser utilizado para detectar estrutura populacional em cana-de-açúcar através de marcadores dominantes, visto que estes não disponibilizam informações de doses alélicas superiores.

O método do Neighbor-Joining (SAITOU; NEI, 1987), que é um agrupamento clássico baseado em dissimilaridade genética, vem sendo utilizado para verificar estrutura populacional em vários estudos com cana-de-açúcar (BUTTERFIELD, 2007; RABOIN et al., 2008; D'HOOP et al., 2010). Apesar de ser uma análise exploratória, acredita-se que esse método poderá ser muito útil para tal finalidade, uma ideia que se reforça com os resultados obtidos por Odong et al. (2011). Esses autores, precursores na avaliação de métodos hierárquicos com o uso de marcadores moleculares, verificaram que tais métodos podem fornecer informações satisfatórias a respeito da estrutura de determinada população. Isso porque, em geral, exigem menor tempo computacional em comparação aos outros métodos, estão estabelecidos em muitos pacotes estatísticos e sua compreensão é bastante simples. Ademais, não exigem pressuposições de marcadores não ligados ou em equilíbrio de Hardy-Weinberg e ligação, algo que é verificado em outros métodos.

# 3 MATERIAL E MÉTODOS

## 3.1 Material

### 3.1.1 Painel Brasileiro de Variedades de Cana-de-açúcar

O desequilíbrio de ligação e a estrutura populacional foram analisados com base no painel brasileiro de variedades de cana-de-açúcar (PBVCA), o qual é constituído por 155 indivíduos, selecionados a partir da sua importância para o germoplasma brasileiro. Porém, como algumas dessas variedades não estiveram disponíveis até o momento da realização desse trabalho, as análises foram realizadas considerando 135 indivíduos do PBVCA, os quais estão apresentados na Tabela 1. Esse painel foi desenvolvido principalmente para realização do mapeamento associativo no Brasil, e certamente será útil para outros estudos relacionados ao seu germoplasma. A construção do PBVCA foi realizada com base nos seguintes critérios: i) variedades mais cultivadas em lavouras comerciais brasileiras, como RB867515, SP81-3250 e RB855453; ii) variedades ancestrais importantes, como POJ2878, Co331, IAC48-65 e NA56-79, as quais contribuíram para obtenção dos híbridos atuais e foram definidas a partir das genealogias desses híbridos; iii) principais genitores utilizados em cruzamentos; iv) genitores utilizados em programas de mapeamento, como SP80-180, SP80-4966, IAC93-3046 e IAC95-3018; v) variedades comerciais recentemente liberadas; e vi) variedades/clones promissores dos programas de melhoramento brasileiros.

### 3.1.2 Marcadores Moleculares e Genotipagem

Um total de 100 marcadores microssatélites (pares de *primers*), sendo 86 derivados de sequências expressas (EST-SSRs) e 14 genômicos (SSRs), foi utilizado para genotipar 135 indivíduos do PBVCA (Tabela 1), possibilitando a amplificação de 1.474 bandas polimórficas. Os microssatélites obtidos de ESTs foram desenvolvidos por Pinto et al. (2004) e Oliveira et al. (2007), enquanto os genômicos foram desenvolvidos por Cordeiro et al. (2000) e uma instituição de pesquisa denominada CIRAD (La Recherche Agronomique Pour Le Developpement), estabelecida na França.

A genotipagem foi realizada a partir do DNA extraído dos primórdios foliares das plantas, seguindo protocolo descrito por Al-Janabi et al. (1999). As reações de amplificação foram feitas com base nos procedimentos descritos por Oliveira et al. (2007). A eletroforese dos fragmentos

Tabela 3 – Painel brasileiro de variedades de cana-de-açúcar (PBVCA) constituído por 135 indivíduos

| Origem | Variedades | | | | |
|---|---|---|---|---|---|
| Ancestrais | Badila[1] | White Transp.[1] | Ganda Cheni[2] | Maneria[3] | EK28 |
| Java, Indonésia | POJ2364 | POJ2878 | IN84-58[4] | | |
| Campos, Brasil | CB36-24 | CB40-13 | CB41-76 | CB45-155 | CB45-3 |
| | CB46-47 | CB47-355 | CB49-260 | CB53-98 | |
| Coimbatore, India | Co290 | Co331 | Co419 | Co449 | Co740 |
| | Co997 | NCo310 | | | |
| Taiwan | F31-962 | F36-819 | | | |
| Canal Point, EUA | CP70-1547 | CP52-68 | CP51-22 | CP53-76 | |
| Norte da Argentina | NA56-79 | | | | |
| Louisiana, EUA | L60-14 | | | | |
| Tucumán, Argentina | TUC71-7 | | | | |
| Hawaii | H59-1966 | H53-3989 | | | |
| Reunion, França | R570 | | | | |
| Campinas, Brasil | IAC48-65 | IAC49-131 | IAC50-134 | IAC51-205 | IAC52-150 |
| | IAC58-480 | IAC64-257 | IAC68-12 | IAC82-2045 | IAC82-3092 |
| | IAC83-4157 | IAC86-2210 | IAC87-3396 | IAC91-1099 | |
| República do Brasil | RB721012 | RB72199 | RB72454 | RB725053 | RB725828 |
| | RB732577 | RB735200 | RB735220 | RB735275 | RB739359 |
| | RB739735 | RB75126 | RB765418 | RB785148 | RB806043 |
| | RB815690 | RB825317 | RB825336 | RB83102 | RB835019 |
| | RB835054 | RB835089 | RB835205 | RB835486 | RB845197 |
| | RB845210 | RB845257 | RB855002 | RB855035 | RB855036 |
| | RB855077 | RB855113 | RB855156 | RB855206 | RB855350 |
| | RB855453 | RB855463 | RB855465 | RB855536 | RB855546 |
| | RB855563 | RB855589 | RB855595 | RB855511 | RB867515 |
| | RB92579 | RB925211 | RB925268 | RB925345 | RB935744 |
| | RB965902 | RB965917 | RB966928 | | |
| São Paulo, Brasil | SP70-1005 | SP70-1078 | SP70-1143 | SP70-1284 | SP70-1423 |
| | SP70-3370 | SP71-799 | SP71-1406 | SP71-6163 | SP71-6949 |
| | SP72-4928 | SP77-5181 | SP79-1011 | SP79-2233 | SP79-2312 |
| | SP79-2313 | SP79-6134 | SP79-6192 | SP80-180 | SP80-185 |
| | SP80-1520 | SP80-1816 | SP80-1836 | SP80-1842 | SP80-3280 |
| | SP80-4966 | SP81-1763 | SP81-3250 | SP83-2847 | SP83-5073 |
| | SP89-1115 | SP91-1049 | | | |

[1] *Saccharum officinarum*; [2] *Saccharum barberi*; [3] *Saccharum sinense*; [4] indivíduo inserido posteriormente.

amplificados foi realizada conforme descrição feita por Creste et al. (2001). A interpretação dos géis foi realizada utilizando-se um sistema binário, ou seja, presença (1) ou ausência (0) de bandas em um determinado indivíduo. Quando houve falhas na amplificação, utilizou-se NA (Não Amplificado) para indicar os dados perdidos. Os géis foram lidos manualmente, com o auxílio de um transluminador, e os seus dados foram transferidos para uma planilha para posterior construção de uma matriz, formada pela combinação dos marcadores detectados com os indivíduos analisados.

No contexto poliplóide da cana-de-açúcar, a presença de bandas sugeriu que determinado alelo, para um determinado loco, esteve presente em pelo menos um dos cromossomos que formam um grupo de homologia, enquanto a ausência de bandas sugeriu que esse mesmo alelo não se fez presente em nenhum dos cromossomos. Assim, os microssatélites forneceram informações parciais a respeito do genoma da cana-de-açúcar, refletindo diretamente nos procedimentos para estimação do DL, conforme especificado no item 3.2.2.

## 3.2 Métodos

### 3.2.1 Mapa de Ligação

Para análise da extensão do DL ao longo do genoma, foi utilizado como referência o mapa de ligação publicado por Oliveira et al. (2007), construído a partir da genotipagem da população experimental oriunda do cruzamento entre as variedades de cana-de-açúcar SP80-180 e SP80-4966. Apenas marcadores dominantes com segregações do tipo 1:1 e 3:1 foram utilizados na construção desse mapa. No entanto, como o mesmo foi obtido, na época, por meio de uma abordagem de dois pontos, menos precisa, optou-se por reestimá-lo com abordagem multiponto, a qual está implementada na nova versão do software *OneMap* (MARGARIDO; SOUZA; GARCIA, 2007). Essa abordagem, baseada em Cadeias de Markov Ocultas, fornece maior precisão na ordenação dos locos e na estimação das suas distâncias dentro dos grupos de co-segregação. Portanto, posições e distâncias mais confiáveis foram obtidas para muitos dos microssatélites do PBVCA que estiveram presentes no referido mapa, refletindo diretamente nos padrões do DL encontrados.

### 3.2.2 Análise do Desequilíbrio de Ligação ao Longo do Genoma

Como foi apresentado, a utilização de 100 microssatélites (86 EST-SSRs e 14 SSRs) possibilitou a amplificação de 1.474 bandas polimórficas. Cada uma dessas bandas representou, para determi-

nado loco e indivíduo, todas as possíveis cópias de um grupo de homologia, de modo que as mesmas não puderam ser visualizadas no padrão molecular gerado. Assim, os marcadores microssatélites se comportaram como dominantes (presença e ausência) em cana-de-açúcar, gerando informações parciais acerca do seu genoma poliplóide e altamente complexo. Em função disso, não foi possível inferir sobre doses alélicas superiores de modo a identificar diferentes classes genotípicas nos dados das variedades, o que teoricamente possibilitaria a reconstrução de haplótipos e, a partir destes, a estimação de gametas e suas frequências por meio de métodos estatísticos.

Nesse contexto, medidas comumente utilizadas para a estimação do DL ($D^{'}$ e $r^2$), as quais são baseadas nas frequências dos alelos e dos haplótipos, não puderam ser aplicadas, sendo que a associação não-aleatória entre cada possível par de marcas foi verificada utilizando-se a probabilidade de Fisher, obtida pela aplicação do teste exato (FISHER, 1935). Para cada possível par de marcas, foi construída uma tabela de contingência $2 \times 2$ contendo as frequências dos fenótipos moleculares ($1-1$, $1-0$, $0-1$, $0-0$) e, desta, calculada uma probabilidade $p$, a qual teve significância verificada pela probabilidade de Fisher ($P$), apresentada a seguir.

$$P = \sum_{x} p, \text{ sendo } p = \frac{\binom{n_{1.}}{n_{11}}\binom{n_{2.}}{n_{21}}}{\binom{n_{..}}{n_{.1}}} = \frac{n_{1.}!n_{2.}!n_{.1}!n_{.2}!}{n_{..}!n_{11}!n_{12}!n_{21}!n_{22}!} \tag{7}$$

sendo $x$ o conjunto de todas as tabelas que poderiam ter sido verificadas com probabilidade $p$ menor ou igual à da tabela observada, $n_{ij}!$ as frequências dos fenótipos moleculares, $n_{i.}!$ e $n_{.j}!$ as frequências marginais correspondentes à linha $i$ e coluna $j$ da tabela de contingência $2\times2$, e $n_{..}!$ a frequência total da tabela observada. Os testes de Fisher entre os pares de marcas foram realizados através do pacote "exact2x2" (FAY, 2010), implementado no programa R (R Development Core Team, 2011).

Para verificar os pares de marcas que foram estatisticamente significativos, o que indica a presença de DL, foi estabelecido um "threshold", ou limiar, de $5\%$ ($P = 5 \times 10^{-2}$), considerando todos os testes simultaneamente. No entanto, um limiar dessa magnitude pode fazer com que hipóteses de nulidade ($H_0$) de falsas associações sejam indevidamente rejeitadas (Erro Tipo I). Isso quer dizer que um ponto de corte de $5 \times 10^{-2}$, neste caso, não é apropriado quando múltiplos testes estão sendo realizados, uma vez que falsas associações podem ser consideradas significativas. Assim, a correção de Bonferroni (WEIR, 1996) foi utilizada como um procedimento de controle, de modo que um limiar equivalente a cada teste foi estabelecido, conforme especificações a seguir.

$\alpha = P(\text{pelo menos um teste rejeita } H_0 \mid H_0 \text{ é verdadeira})$

$\alpha = 1 - P(\text{nenhum teste rejeita } H_0 \mid H_0 \text{ é verdadeira})$

$\alpha = 1 - [1 - P(\text{um teste rejeita } H_0 \mid H_0 \text{ é verdadeira})]^L$

$\alpha = 1 - (1 - \alpha^{'})^L$

$\alpha \approx L\alpha^{'}$

onde $\alpha$ é a probabilidade máxima (limiar) para os testes serem considerados significativos (no caso, $0{,}05$ ou $5$ x $10^{-2}$), $\alpha^{'}$ é a probabilidade máxima corrigida através do Bonferroni, e $L$ é o número de testes realizados. Logo,

$$\alpha^{'} \approx \frac{\alpha}{L} \tag{8}$$

No entanto, a correção de Bonferroni é um procedimento conservativo, pois ao mesmo tempo que controla o Erro Tipo I pode cometer o Erro Tipo II. Isso significa que, por apresentar limitações ao considerar a independência de testes múltiplos, essa abordagem poderá desprezar associações significativas entre marcadores ligados, as quais são de interesse para o contexto deste trabalho. Nesse sentido, foi utilizado o FDR - *False Discovery Rate* (BENJAMINI; HOCHBERG, 1995) como um procedimento alternativo de controle do Erro Tipo I, que considera a proporção esperada de hipóteses nulas indevidamente rejeitadas. Assumindo um FDR inicial de $5 \times 10^{-2}$, ou seja, $5\%$ de falsas associações dentre àquelas que foram significativas, foi possível estabelecer um limiar alternativo ($\alpha^{''}$) de modo a verificar quantas associações foram realmente significativas ($1 - FDR$), o que indica a presença de DL. O valor de $\alpha^{''}$ foi obtido pela seguinte expressão:

$$\alpha^{''} = \frac{FDR(\text{inicial})i}{\pi_0 L} \tag{9}$$

onde $\pi_0$ é a proporção de testes que não rejeitam a hipótese de nulidade, $i$ é o vetor com os índices referentes às probabilidades dos testes, e $L$ é o número de testes realizados. O valor de $\pi_0$, que é desconhecido a priori, foi obtido através do pacote "qvalue" (STOREY, 2003) implementado no programa R (R Development Core Team, 2011), o qual possibilita obter os q-valores a partir dos p-valores (probabilidades de Fisher) dos testes.

Para análise da extensão do DL ao longo do genoma, foram considerados todos os marcadores do PBVCA que estiveram presentes em um mesmo grupo de co-segregação no mapa de ligação, após a localização dos mesmos nos padrões moleculares envolvidos. Os logaritmos das probabilidades

de Fisher ($-LogP$) das associações entre os marcadores ligados foram plotados com as respectivas distâncias genéticas em centimorgan (cM).

### 3.2.3 Análise da Estrutura Populacional

Como foi visto, no mapeamento associativo, que é baseado em populações com várias gerações, é possível que locos ligados e não ligados estejam em DL. Como o segundo caso não é de interesse para o mapeamento de QTL's e a seleção assistida por marcadores moleculares, é fundamental que seja controlado, de modo que apenas o primeiro caso seja considerado. Nesse sentido, a estrutura de população, sendo talvez o fator que maior contribui para o surgimento de falsas associações, tem sido comumente analisada em um painel de variedades.

Até o momento, nenhum método foi desenvolvido para detectar estrutura de população em um poliplóide complexo como a cana-de-açúcar, visto que a grande maioria dos dados moleculares disponíveis apresenta pouca informatividade a respeito do seu genoma. Assim, neste trabalho, como análise geral, foram utilizados dois possíveis métodos na tentativa de detectar estrutura genética no PBVCA, que são: i) método baseado em inferência bayesiana, implementado no software STRUCTURE (PRITCHARD; STEPHENS; DONNELLY, 2000); e ii) método baseado em dissimilaridade genética, através do cálculo da correspondência simples, ou "simple matching", entre todos os indivíduos 2 a 2 e posterior agrupamento com o algoritmo Neighbor-Joining (SAITOU; NEI, 1987).

No primeiro método, foram utilizados 66 marcadores do PBVCA localizados em grupos de co-segregação diferentes no mapa de ligação, conforme a pressuposição de locos não ligados do modelo probabilístico. As análises foram realizadas de acordo com os seguintes parâmetros: a) nível de ploidia 1, o qual desconsidera a complexidade da cana-de-açúcar pela pouca informatividade dos dados; b) modelo de não-mistura ("no admixture"), o qual atribui os indivíduos a uma ou outra população sem considerar a presença de proporções genômicas de diferentes populações, sendo recomendado no caso de marcadores dominantes; c) 20.000 iterações para o comprimento do período de aquecimento, denominado "Burnin"; d) 200.000 iterações para a convergência da Cadeia de Markov via Monte Carlo; e) frequências alélicas independentes; e f) ajuste do modelo considerando, a priori, o número de populações variando de 1 à 10 ($K = 1, 2, ..., 10$), cada qual com probabilidade inicial de $1/K$. As análises foram repetidas por 10 sucessivas vezes, a fim de obter probabilidades mais confiáveis para cada uma das populações (PRITCHARD; STEPHENS; DONNELLY, 2000), as quais foram comparadas entre si.

No segundo método, foram utilizados todos os 1.474 marcadores polimórficos detectados no PBVCA, de acordo com a ausência de pressuposição de locos não ligados. A matriz de dissimilaridade genética foi obtida através do pacote “scrime”, implementado no programa R (R Development Core Team, 2011), e a Neighbor-Joining foi gerada no software Darwin (PERRIER et al., 2003).

# 4 RESULTADOS

## 4.1 Desequilíbrio de Ligação ao Longo do Genoma

Um total de 1.085.601 testes exatos de Fisher foi realizado considerando todas as possíveis combinações 2 × 2 entre os 1.474 marcadores polimórficos detectados no PBVCA. Com um "threshold" de $5\%$ ($P = 5 \times 10^{-2}$), 132.842 ($12{,}24\%$) associações foram estatisticamente significativas, indicando a existência de DL entre os locos envolvidos. No entanto, como muitas delas poderiam ser falsas em função da multiplicidade de testes, indicando, neste caso, equilíbrio de ligação entre os locos, foram estabelecidas as correções de Bonferroni e FDR, encontrando-se $P = 4{,}6 \times 10^{-8}$ e $P = 1{,}4 \times 10^{-3}$ como limiares, respectivamente. Com Bonferroni, 2.780 ($0{,}26\%$) associações foram consideradas significativas, as quais permaneceram abaixo do primeiro limiar. Com FDR, 30.314 ($2{,}79\%$) associações foram consideradas significativas ao permanecerem abaixo do segundo limiar, sendo um número bastante superior ao primeiro, mostrando o quanto aquele procedimento pode ser conservativo na detecção de associações preferenciais entre locos. Do total de 1.474 marcadores, 1.359 ($92{,}2\%$) estiveram envolvidos com as 30.314 associações consideradas.

Apesar de muitos marcadores terem alguma associação em DL, os resultados mostram que a grande maioria das associações encontra-se em equilíbrio de ligação. Isso representa 1.055.287 ($97{,}21\%$) associações envolvendo locos independentes, as quais não são de interesse para o contexto desse trabalho. Além disso, grande parte das associações em DL deve envolver marcadores não ligados, de modo que um número inferior a $2{,}79\%$ poderá ser, de fato, útil.

Com o objetivo de detectar DL entre locos ligados e estudar sua extensão ao longo do genoma, foram localizados 102 marcadores (EST-SSRs) do PBVCA no mapa multiponto utilizado como referência (ver item 3.2.1), correspondendo a $6{,}92\%$ do total de polimórficos. Esses marcadores distribuíram-se entre 66 grupos de co-segregação dos 198 obtidos, de modo que estiveram presentes em até 16 grupos de homologia. Foram considerados 5.151 ($0{,}47\%$) testes exatos de Fisher do total inicialmente realizado, correspondendo às possíveis combinações $2 \times 2$ entre os 102 marcadores mapeados. Desse número, apenas 60 ($1{,}17\%$) associações envolveram marcadores ligados presentes em um mesmo grupo de co-segregação, compreendendo diferentes grupos de homologia. A multiplicidade de testes foi novamente corrigida através dos procedimentos de Bonferroni e FDR, obtendo-se $P = 9{,}7 \times 10^{-6}$ e $P = 6{,}3 \times 10^{-4}$ como limiares, respectivamente. Com Bonferroni, 18

($0{,}35\%$) associações foram consideradas significativas, das quais apenas 5 envolveram marcadores ligados. Com FDR, a quantidade de associações significativas subiu para 66 ($1{,}28\%$), acréscimo este que não contribuiu para detectar mais DL entre marcadores ligados.

Na Figura 7 são mostrados os logaritmos das probabilidades de Fisher, aqui utilizados como uma medida aproximada do DL, em função da distância genética em cM. Apenas associações entre marcadores ligados são apresentadas nesse gráfico, totalizando 60 das 5.151 possíveis envolvendo locos mapeados. Como é possível observar, apenas 5 ($8{,}33\%$) associações, envolvendo 9 diferentes marcadores, foram significativas considerando os limiares de Bonferroni e FDR, indicando a existência de DL por ligação física. O aumento no limiar através do FDR não contribuiu no sentido de detectar outras associações em DL, como pode ser visto pela ausência de pontos no trecho que separa ambos os procedimentos. Apesar de poucas associações terem apresentado DL, o que pode comprometer uma verificação confiável dos seus padrões em relação à distância, é possível inferir a presença de forte DL nos primeiros 15 cM, tendo-se uma queda após essa distância.

A Figura 8, em complementação à anterior, mostra o número de associações significativas e não-significativas a cada 5 cM. De fato, forte DL é observado até 15 cM, principalmente nos primeiros 5 cM, como também tem sido verificado em outros estudos (JANNOO et al., 1999b; RABOIN et al., 2008; WEI et al., 2010). Apesar de uma queda ter sido observada após a primeira distância, é possível notar a presença de DL entre marcadores que permanecem a 65 cM em um mesmo grupo de co-segregação, indicando a existência de associações preferenciais em maiores extensões. Jannoo et al. (1999b), Raboin et al. (2008) e Wei et al. (2010) também detectaram DL em distâncias superiores a 50 cM, tendo-se uma queda a partir de 30 cM.

## 4.2 Estrutura Populacional

Os logaritmos das probabilidades a posteriori ($\log Pr(X|K)$) obtidos através do STRUCTURE, referentes ao conjunto $X$ dos 135 indivíduos do PBVCA de acordo com 10 subpopulações ($K$) estabelecidas a priori, são apresentados na Tabela 4. Foi possível observar que da primeira até a quarta subpopulação as probabilidades permaneceram quase que inalteradas ao longo de 10 análises realizadas, mostrando uma convergência confiável da Cadeia de Markov via Monte Carlo. A partir da quinta subpopulação, probabilidades variáveis passaram a ser obtidas, as quais, em geral, oscilaram de maneira gradativa com o aumento das subpopulações.

Uma inspeção da Figura 9, que apresenta as probabilidades mínimas, médias e máximas das

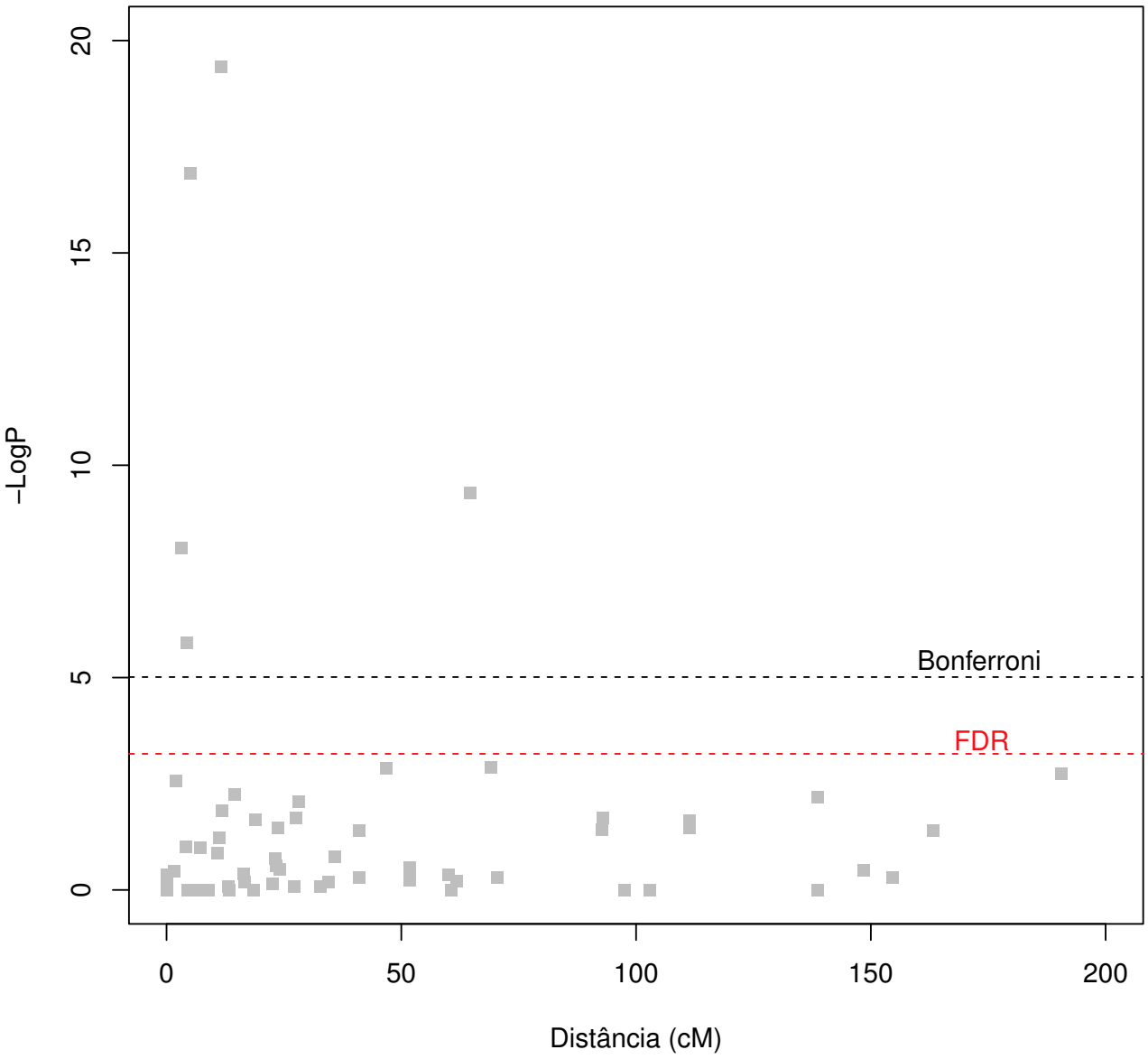


Figura 7 – Desequilíbrio de ligação ($-LogP$) em função da distância genética em cM. Os limiares correspondentes às correções de Bonferroni e FDR estão indicados em escala logarítmica, mostrando as associações significativas (em DL) e não-significativas acima e abaixo dos seus valores, respectivamente. As distâncias genéticas foram obtidas através da função de mapeamento Kosambi (KOSAMBI, 1944)

análises para cada uma das 10 subpopulações, mostra que o modelo com 4 subpopulações ($K = 4$) foi o que apresentou a maior probabilidade média dentre todos os modelos avaliados. A partir do modelo com 5 subpopulações ($K = 5$), as probabilidades médias foram decrescendo gradativamente, ao passo que as variações entre probabilidades mínimas e máximas foram aumentando. Nesse sentido, o modelo com 10 subpopulações ($K = 10$), apesar de ter apresentado uma elevação na probabilidade média em relação ao exatamente anterior ($K = 9$), foi aquele que apresentou a maior variação considerando as 10 análises realizadas.

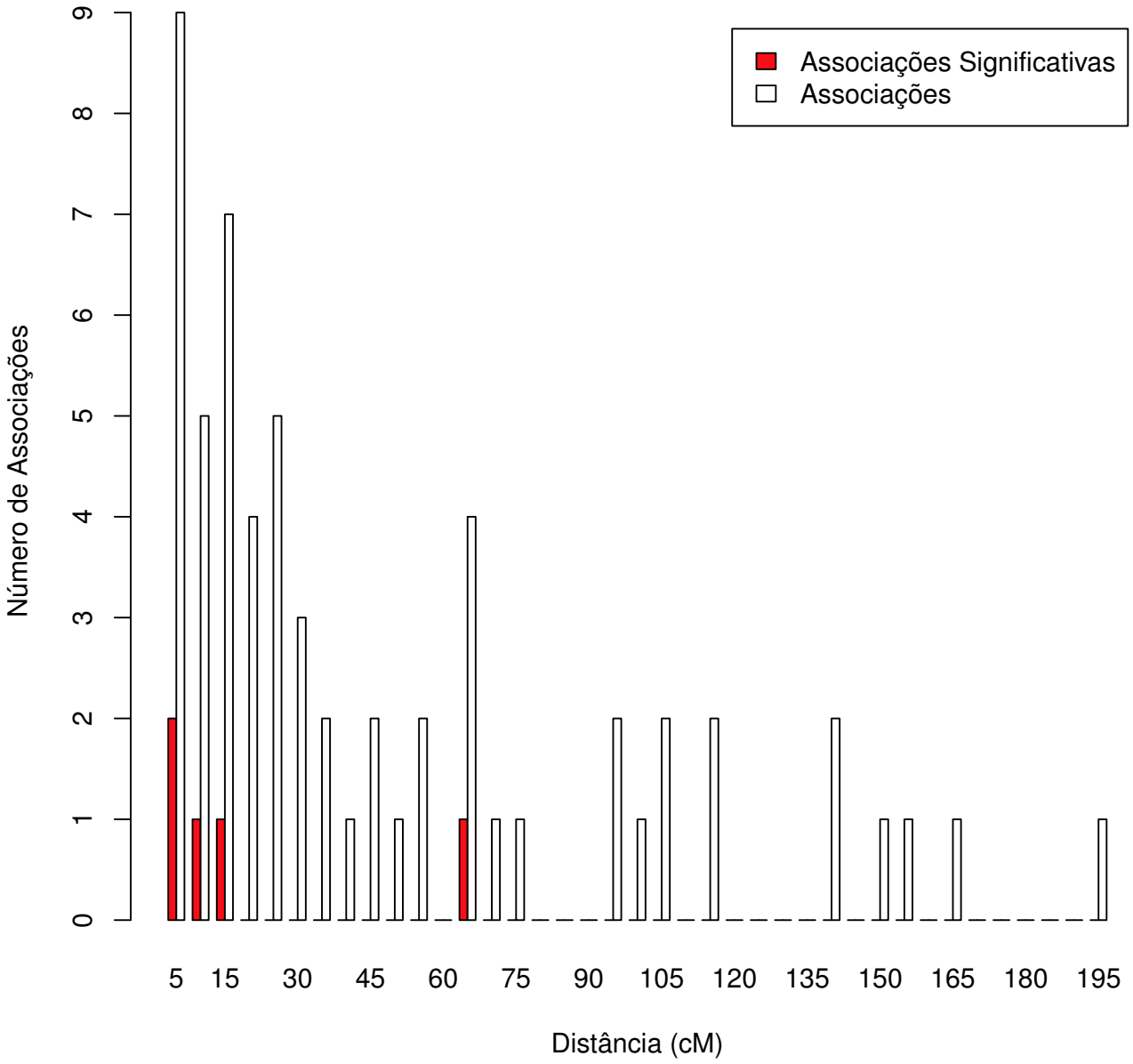


Figura 8 – Número de associações significativas e não-significativas a cada 5 cM entre 60 associações envolvendo pares de marcadores no mesmo grupo de co-segregação

Apesar dos resultados anteriores permitirem a inferência de $K = 4$, sugerindo a existência de 4 subpopulações no PBVCA, isso não deve ser considerado de maneira incontestável. Pritchard, Stephens e Donnelly (2000) mencionam que a inferência populacional através do STRUCTURE deve ser encarada com certo cuidado, visto que depende de vários fatores, como o número de populações estabelecido a priori, a quantidade de análises realizadas para cada população, o comprimento do período de aquecimento (“Burnin”), as iterações para convergência da Cadeia de Markov Via Monte Carlo, e entre outros. Assim, os resultados obtidos com esse tipo de análise, a qual é baseada em inferência bayesiana, são inerentes às suas propriedades, podendo, muitas vezes, não fornecer

estimativas realistas do $K$ para um determinado conjunto de indivíduos. Portanto, é importante sempre verificar se a inferência populacional obtida está de acordo com o esperado biologicamente (PRITCHARD; STEPHENS; DONNELLY, 2000).

Tabela 4 – Logaritmos das probabilidades a posteriori ($\log Pr(X|K)$), referentes a 135 indivíduos do PBVCA, de 10 análises realizadas no STRUCTURE (PRITCHARD; STEPHENS; DONNELLY, 2000) para cada uma das 10 subpopulações ($K$) estabelecidas a priori. A linha destacada em negrito refere-se ao número de subpopulações mais provável do PBVCA e suas probabilidades obtidas nas análises

| $K$ | Análises | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | −4.522,1 | −4.522,2 | −4.522,2 | −4.522,1 | −4.522,2 | −4.522,2 | −4.522,1 | −4.522,2 | −4.522,2 | −4.522,2 |
| 2 | −4.343,9 | −4.343,5 | −4.343,9 | −4.343,9 | −4.343,9 | −4.343,6 | −4.343,8 | −4.343,7 | −4.343,9 | −4.343,6 |
| 3 | −4.205,7 | −4.205,7 | −4.205,6 | −4.205,4 | −4.205,2 | −4.205,9 | −4.205,3 | −4.205,7 | −4.205,7 | −4.205,4 |
| **4** | **−4.129,8** | **−4.129,0** | **−4.128,5** | **−4.129,0** | **−4.128,9** | **−4.129,5** | **−4.128,8** | **−4.129,0** | **−4.129,1** | **−4.128,9** |
| 5 | −4.123,3 | −4.138,7 | −4.145,0 | −4.124,0 | −4.123,5 | −4.122,2 | −4.123,1 | −4.127,9 | −4.123,3 | −4.188,9 |
| 6 | −4.199,0 | −4.123,0 | −4.231,1 | −4.149,3 | −4.123,9 | −4.169,4 | −4.295,1 | −4.187,8 | −4.155,2 | −4.123,0 |
| 7 | −4.395,6 | −4.238,8 | −4.322,2 | −4.411,3 | −4.341,9 | −4.305,0 | −4.427,5 | −4.414,7 | −4.138,9 | −4.260,5 |
| 8 | −4.302,3 | −4.407,0 | −4.229,6 | −4.403,3 | −4.172,8 | −4.430,5 | −4.449,8 | −4.451,6 | −4.466,8 | −4.204,4 |
| 9 | −4.475,1 | −4.322,7 | −4.466,7 | −4.467,4 | −4.416,6 | −4.422,6 | −4.384,6 | −4.312,6 | −4.462,2 | −4.431,8 |
| 10 | −4.483,6 | −4.462,9 | −4.163,8 | −4.109,1 | −4.302,3 | −4.143,9 | −4.356,5 | −4.322,1 | −4.169,5 | −4.463,3 |

A Figura 10 apresenta graficamente as probabilidades a posteriori ($P(X|K)$) dos 135 indivíduos do PBVCA, considerando o modelo com 4 subpopulações ($K = 4$) anteriormente selecionado. Como tal modelo apresentou probabilidades consistentes ao longo das análises realizadas (Tabela 4), selecionou-se ao acaso os resultados da análise 2 para representação na Figura em questão. Nesta, é possível visualizar os 4 grupos obtidos, através de cores atribuídas conforme maior ou menor probabilidade. Com base na escala apresentada, a cor vermelha indicou a máxima probabilidade ($P(X|K) = 1{,}0$) de um indivíduo estar em um determinado grupo, de modo que a cor verde-clara indicou a menor probabilidade ($P(X|K) = 0{,}0$) disso ocorrer. Assim, a subpopulação 1, indicada na primeira coluna, foi constituída por 49 indivíduos concentrados na região mediana do gráfico (exceto RB855465 e SP80-180, que permaneceram na parte superior), dos quais 42 ($85{,}72\%$) apresentaram probabilidades acima de $0{,}9$. A subpopulação 2, situada na segunda coluna, foi formada por 32 indivíduos localizados na região superior do gráfico, dos quais apenas 2 ($6{,}25\%$)

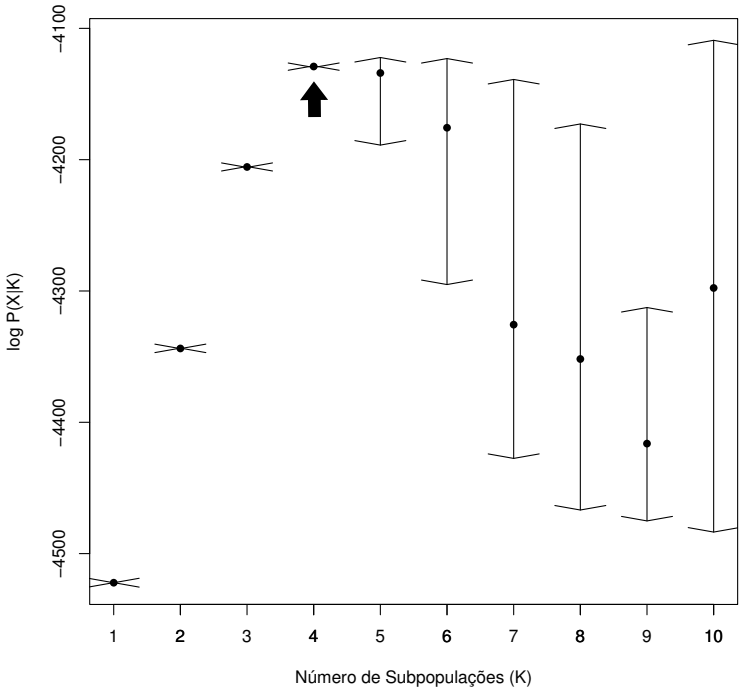


Figura 9 – Valores mínimos, médios e máximos dos logaritmos das probabilidades a posteriori ($\log Pr(X|K)$), referentes a 135 indivíduos do PBVCA, de 10 análises realizadas para cada uma das 10 subpopulações ($K$) estabelecidas a priori. A seta representada indica o número de subpopulações mais provável do PBVCA

tiveram probabilidades abaixo de $0{,}9$. A subpopulação 3, presente na terceira coluna, foi composta por apenas 7 indivíduos espalhados na região superior do gráfico, sendo que somente 1 ($14{,}3\%$) não apresentou probabilidade máxima. E, por fim, a subpopulação 4, mostrada na última coluna, foi constituída pelos 47 indivíduos restantes situados na região inferior do gráfico, sendo que apenas 3 ($6{,}3\%$) mostraram probabilidades inferiores a $0{,}9$.

A Tabela 5 apresenta as 4 subpopulações do PBVCA detectadas através do STRUCTURE. Foi possível notar que, com exceção à subpopulação 3, todas as demais apresentaram indivíduos dos diferentes programas de melhoramento brasileiros, bem como de outros germoplasmas existentes pelo mundo. Além disso, foram observadas variedades ancestrais nas 3 primeiras subpopulações.

Visando comparar com os resultados anteriores, foi construído um gráfico de Neighbor-Joining, o qual está apresentado na Figura 11, para os 135 indivíduos do PBVCA, a partir da dissimilaridade genética obtida pelo método "simple matching". Ao contrário do método probabilístico, que é baseado em análise inferencial, o agrupamento do Neighbor-Joining é realizado de maneira exploratória, e normalmente fornece uma ideia da relação entre os indivíduos (PRITCHARD; STEPHENS; DONNELLY, 2000). Assim, pelos padrões verificados na Figura 11, foi possível notar grande semelhança com o agrupamento realizado pelo STRUCTURE, diferindo deste apenas quanto aos 6

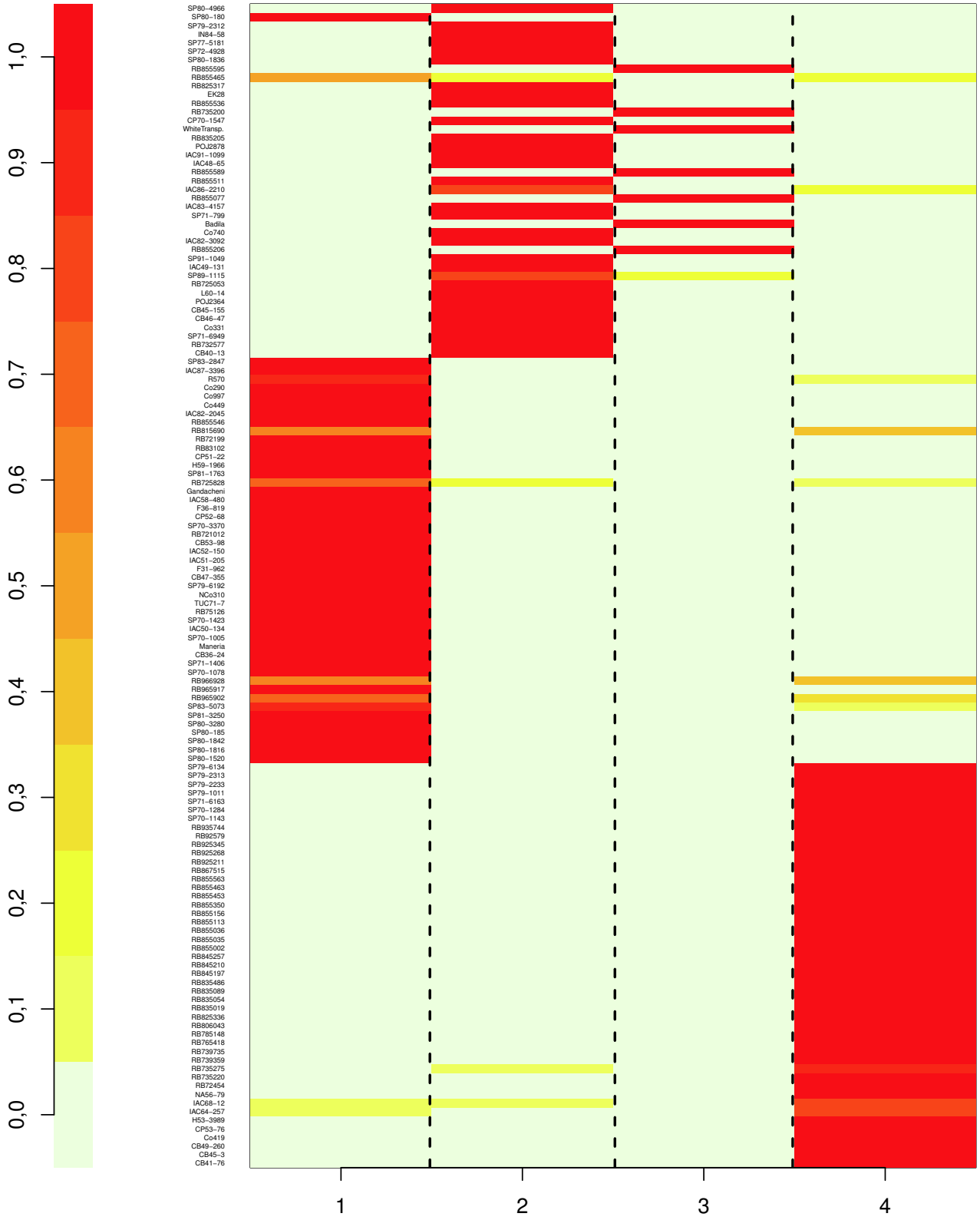


Figura 10 – Gráfico dos 135 indivíduos do PBVCA (eixo y) e 4 diferentes subpopulações (eixo x) obtidas pelo método probabilístico implementado no STRUCTURE. Quatro grandes colunas numeradas e separadas por retas pontilhadas podem ser observadas, correspondendo à cada uma das 4 subpopulações. Regiões em vermelho no interior dessas colunas indicam o conjunto de indivíduos agrupados na subpopulação correspondente, de acordo com as probabilidades a posteriori mostradas na escala de cores à esquerda, as quais variam de $0{,}0$ (probabilidade zero) até $1{,}0$ (probabilidade máxima)

Tabela 5 – Estrutura Populacional do PBVCA detectada pelo método probabilístico implementado no STRUCTURE. As 4 subpopulações detectadas estão representadas com diferentes cores

| Subpopulações | Indivíduos | | | | |
|---|---|---|---|---|---|
| Subpopulação 1 | SP80-1520 | SP80-1816 | SP80-1842 | SP80-185 | SP80-3280 |
| | SP81-3250 | SP83-5073 | RB965902 | RB965917 | RB966928 |
| | SP70-1078 | SP71-1406 | CB36-24 | Maneria | SP70-1005 |
| | IAC50-134 | SP70-1423 | RB75126 | TUC71-7 | NCo310 |
| | SP79-6192 | CB47-355 | F31-962 | IAC51-205 | IAC52-150 |
| | CB53-98 | RB721012 | SP70-3370 | CP52-68 | F36-819 |
| | IAC58-480 | Gandacheni | RB725828 | SP81-1763 | H59-1966 |
| | CP51-22 | RB83102 | RB72199 | RB815690 | RB855546 |
| | IAC82-2045 | Co449 | Co997 | Co290 | R570 |
| | IAC87-3396 | SP83-2847 | RB855465 | SP80-180 | |
| Subpopulação 2 | CB40-13 | RB732577 | SP71-6949 | Co331 | CB46-47 |
| | CB45-155 | POJ2364 | L60-14 | RB725053 | SP89-1115 |
| | IAC49-131 | SP91-1049 | IAC82-3092 | Co740 | SP71-799 |
| | IAC83-4157 | IAC86-2210 | RB855511 | IAC48-65 | IAC91-1099 |
| | POJ2878 | RB835205 | CP70-1547 | RB855536 | EK28 |
| | RB825317 | SP80-1836 | SP72-4928 | SP77-5181 | IN84-58 |
| | SP80-4966 | SP79-2312 | | | |
| Subpopulação 3 | RB855206 | Badila | RB855077 | RB855589 | White Transp. |
| | RB735200 | RB855595 | | | |
| Subpopulação 4 | CB41-76 | CB45-3 | CB49-260 | Co419 | CP53-76 |
| | H53-3989 | IAC64-257 | IAC68-12 | NA56-79 | RB72454 |
| | RB735220 | RB735275 | RB739359 | RB739735 | RB765418 |
| | RB785148 | RB806043 | RB825336 | RB835019 | RB835054 |
| | RB835089 | RB835486 | RB845197 | RB845210 | RB845257 |
| | RB855002 | RB855035 | RB855036 | RB855113 | RB855156 |
| | RB855350 | RB855453 | RB855463 | RB855563 | RB867515 |
| | RB925211 | RB925268 | RB925345 | RB92579 | RB935744 |
| | SP70-1143 | SP70-1284 | SP71-6163 | SP79-1011 | SP79-2233 |
| | SP79-2313 | SP79-6134 | | | |

indivíduos indicados pelas setas representadas. Portanto, baseado na semelhança com um método inferencial, pode-se dizer que o Neighbor-Joining também sugeriu a existência de 4 subpopulações no PBVCA, as quais podem ser observadas através das diferentes cores atribuídas aos ramos, que correspondem às subpopulações com as mesmas cores apresentadas na Tabela 5. Esses resultados,

obtidos com base nos 1.474 marcadores polimórficos do PBVCA, mostram que esse agrupamento, sendo um método hierárquico clássico, pode fornecer informações valiosas quanto à estrutura de determinada população.

De maneira geral, os agrupamentos apresentados por ambos os métodos mostraram estar de acordo com informações oriundas de *pedigree*. Isso pôde ser constatado, por exemplo, pela existência de irmãos completos dentro das subpopulações detectadas, como foi o caso das variedades SP80-1816, SP80-1842 e SP80-3280 na subpopulação 1, oriundas do cruzamento entre as variedades SP71-1088 e H57-5028, e das variedades RB845197, RB845210 e RB845257 na subpopulação 4, originárias do cruzamento entre as variedades RB72454 e SP70-1143. No entanto, em ambos os métodos, alguns irmãos completos permaneceram em subpopulações diferentes, como foi o caso da variedade SP80-1836, que esteve na subpopulação 2 sendo irmã completa do grupo de variedades SP apresentado anteriormente, e da variedade RB855536, que também esteve na subpopulação 2 sendo irmã completa do grupo de variedades RB mostrado acima.

Ademais, foi possível observar que as variedades ancestrais Maneria (*Saccharum sinense*), Gandacheni (*Saccharum barberi*), Badila (*Saccharum officinarum*) e White Transp. (*Saccharum officinarum*) foram todas atribuídas, em ambos os métodos, para alguma das 4 subpopulações detectadas, o que não foi verificado, para as três primeiras variedades, no estudo inicial realizado por Lima et al. (2002). Apesar dos métodos aqui utilizados terem apresentado diferenças na atribuição subpopulacional das ancestrais, como ocorreu no caso da Badila, é possível especular que as mesmas possam ter contribuído com maior proporção de seus genomas para os indivíduos que permaneceram em seus respectivos grupos, informação esta que poderá ser muito útil para o mapeamento associativo de características de interesse.

Comparando-se os agrupamentos da Tabela 5 com aqueles apresentados na Figura 11, foi possível observar que 6 indivíduos, do total de 135 do PBVCA, não foram atribuídos às mesmas subpopulações em ambos os métodos. Apesar do STRUCTURE ser baseado em um modelo probabilístico, uma quantidade bem menor de marcadores (66/1.474) foi utilizada em comparação ao Neighbor-Joining, levando a imaginar que este último método possa ter indicado o real agrupamento daqueles indivíduos.

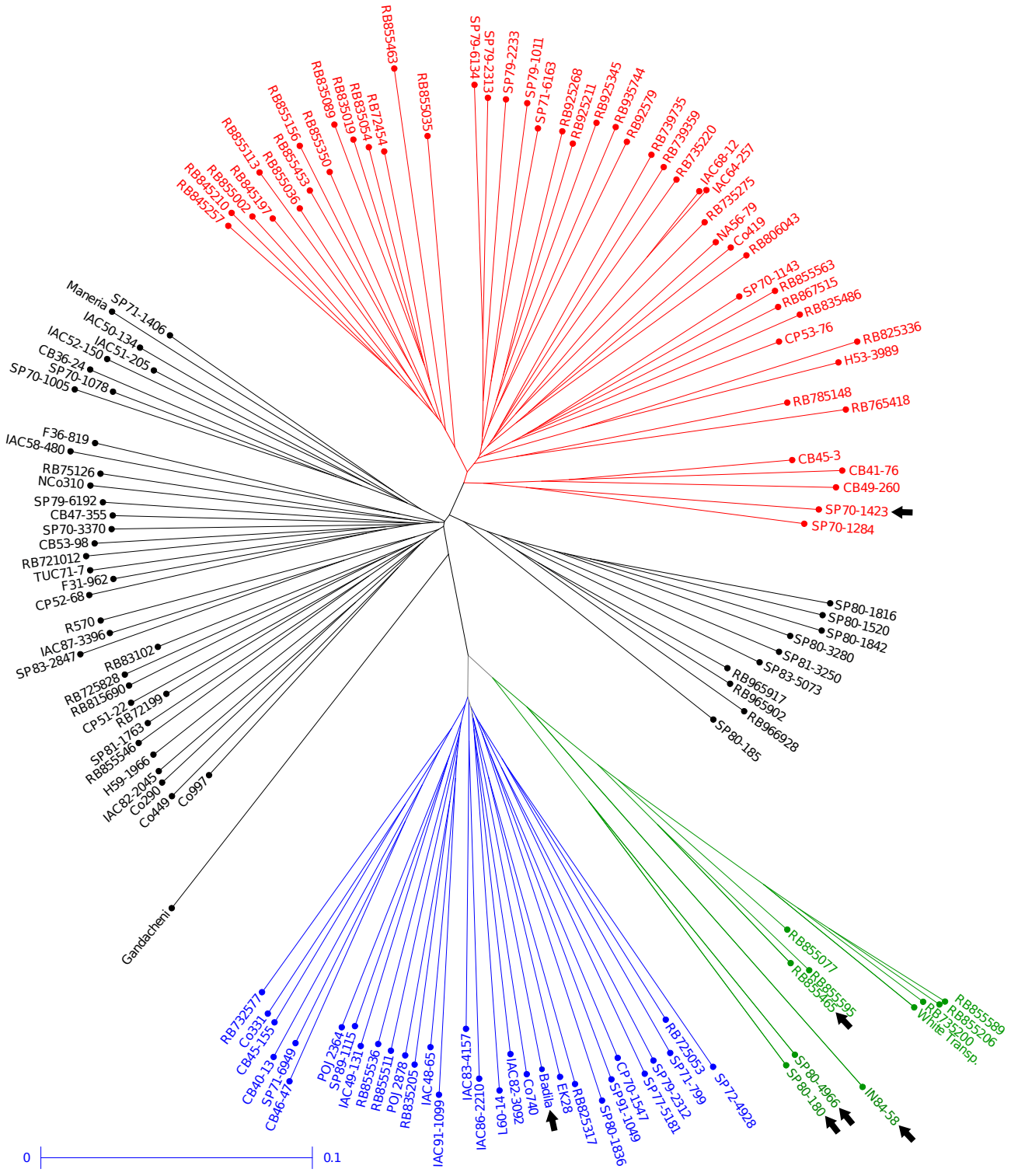


Figura 11 – Neighbor-Joining construída para 135 indivíduos do PBVCA a partir da dissimilaridade genética obtida pelo método “simple matching”. Os indivíduos indicados com a mesma cor pertencem à mesma subpopulação. Os agrupamentos representados pelas diferentes cores são equivalentes àqueles que tiveram as mesmas cores na Tabela 5. As setas indicadas representam os 6 indivíduos que não permaneceram no mesmo agrupamento em comparação com o método probabilístico

## 4.3 Desequilíbrio de Ligação Dentro de Subpopulações

Com a detecção de 4 subpopulações no PBVCA (Figura 10), foi possível estudar o comportamento do DL no interior de cada uma delas, com base na pressuposição do modelo probabilístico inicialmente utilizado. Essa pressuposição assume que locos não ligados devem estar em equilíbrio de Hardy-Weinberg e em completo equilíbrio de ligação no interior de cada subpopulação, considerando a estrutura populacional como único fator no surgimento de falsas associações. Assim, neste caso, qualquer DL detectado dentro de subpopulações poderia ser atribuído a ligação física, o que é de grande interesse no contexto deste trabalho.

Nesse sentido, foram realizados, dentro de cada subpopulação, testes exatos de Fisher considerando todas as possíveis combinações $2 \times 2$ entre marcadores polimórficos, os quais, neste caso, variaram de acordo com o agrupamento (população 1: 1325 marcadores; população 2: 1332 marcadores; população 3: 765 marcadores; população 4: 1148 marcadores). Em função da multiplicidade de testes, utilizou-se a correção de FDR para detectar associações em DL, obtendo-se os seguintes limiares: i) subpopulação 1: $P = 2{,}1 \times 10^{-6}$; ii) subpopulação 2: $P = 8{,}6 \times 10^{-7}$; iii) subpopulação 3: limiar não verificado; e iv) subpopulação 4: $P = 5{,}6 \times 10^{-6}$.

Na subpopulação 1, de um total de 877.150 testes, 42 associações foram significativas, envolvendo 77 diferentes microssatélites. Na subpopulação 2, 29 associações foram significativas de 886.446 possíveis testes, envolvendo 46 diferentes microssatélites. Na subpopulação 3, nenhuma das 292.230 associações testadas foi significativa, mostrando ausência de DL com base nos marcadores utilizados. E, por fim, na subpopulação 4, 75 associações mostraram-se em DL de um total de 658.378 testes, tendo sido observado 115 diferentes marcadores microssatélites.

Esses resultados sugerem a possibilidade de mapeamento de QTL's no interior das subpopulações 1, 2 e 4, claramente envolvendo uma quantidade bastante reduzida de marcadores aqui assumidos como fisicamente ligados.

# 5 DISCUSSÃO

A análise do DL envolvendo 1.474 microssatélites mostrou que a grande maioria das associações encontra-se em equilíbrio de ligação. Apenas $0{,}26\%$ (Bonferroni) e $2{,}79\%$ (FDR) do total de associações envolveram marcadores em DL, o que pode indicar, num primeiro momento, uma proporção bastante reduzida de associações preferenciais. No entanto, ao contrário do que possa parecer, a proporção de DL detectada no presente estudo mostrou-se bastante superior àquela verificada em outros trabalhos.

Jannoo et al. (1999b), ao utilizarem o teste de Fisher e a correção de Bonferroni para verificar DL entre 1.057 RFLPs em cana-de-açúcar, detectaram 59 associações preferenciais das 558.096 possíveis, correspondendo a $0{,}01\%$ de DL. Wei et al. (2006), baseados em 1.068 AFLPs gerados a partir da genotipagem de 154 clones de cana-de-açúcar, encontraram, após também utilizarem o Bonferroni, 736 associações em DL de um total de 569.778 testes de Fisher, correspondendo a $0{,}13\%$ de associações preferenciais. Raboin et al. (2008), ao utilizarem 1.537 AFLPs obtidos através de um painel mundial de cana-de-açúcar, obtiveram $0{,}02\%$ e $0{,}04\%$ de associações em DL com o uso do Bonferroni e de um limiar empírico ($1\%$ de falsas associações), respectivamente, a partir de um total de 1.180.416 testes de Fisher. Com exceção ao último, todos esses estudos utilizaram apenas o Bonferroni para uma detecção apropriada de DL, o qual tem se mostrado conservativo com o aumento no número de testes. No presente trabalho, apesar de uma proporção considerável de DL ter sido verificada com o Bonferroni, a correção através do FDR promoveu um acréscimo substancial no número de associações preferenciais, mostrando que esse procedimento pode ser decisivo quando milhares de testes estão envolvidos. Por mais que falsas associações possam ser consideradas significativas, maiores são as chances de detectar DL entre locos ligados, o que é essencial para o mapeamento de QTL's.

No entanto, uma das possíveis razões do DL ter sido muito superior no presente estudo, até mesmo com o Bonferroni, foi a utilização de todos os marcadores independentemente das suas frequências no PBVCA. O uso de marcadores que apresentaram muitas (maior que $95\%$) e poucas (menor que $5\%$) bandas, considerando os 135 indivíduos do PBVCA, pode ter contribuído para inflacionar determinados fenótipos moleculares, acarretando em DL devido às menores probabilidades obtidas com relação aos testes de Fisher. Esses marcadores não foram aqui retirados basica-

mente por dois motivos: i) marcadores com poucas bandas podem conter alelos raros eventualmente importantes para o mapeamento de características de interesse; e ii) marcadores com muitas bandas podem não conter o alelo correspondente em elevada frequência na população, visto que não é possível inferir sobre doses alélicas em cana-de-açúcar a partir de dados dominantes. Ademais, o número de locos do presente trabalho, embora adequado, não é elevado o suficiente para permitir estudos confiáveis caso o número de marcadores fosse reduzido.

Através de 60 associações envolvendo marcadores ligados, das quais apenas 5 foram significativas, foi possível observar forte DL em uma distância de até 15 cM, principalmente nos primeiros 5 cM, tendo-se uma queda em maiores extensões. Essa queda não significou ausência de DL, visto que associações preferenciais foram detectadas entre marcadores que apresentaram distância superior a 60 cM. Apesar de poucos marcadores ligados terem sido localizados no mapa de referência, o DL extensivo aqui obtido, que também foi detectado em outros estudos (JANNOO et al., 1999b; RABOIN et al., 2008; WEI et al., 2010), ilustra a história recente do melhoramento de cana-de-açúcar no Brasil, a qual é constituída por poucas gerações de recombinação devido à propagação vegetativa da espécie, tendo sido formada a partir de uma população com tamanho efetivo reduzido em função de um "bottleneck" ancestral. Além disso, o forte processo de seleção e o uso recorrente de variedades superiores nos cruzamentos certamente devem explicar parte do DL detectado em maiores extensões (RABOIN et al., 2008). Isso porque a seleção é um fator de grande importância no surgimento de DL entre locos ligados e não ligados, e o uso de genitores superiores normalmente acarreta no cruzamento entre indivíduos aparentados, promovendo o aumento do coeficiente de endogamia e a redução do tamanho populacional efetivo.

Apesar dos resultados indicarem DL extensivo no PBVCA, maior concentração de associações preferenciais foi verificada em pequenas distâncias (até 15 cM) ao longo do genoma. Isso significa que, na verdade, associações mais consistentes entre marcadores e QTL's poderão ser detectadas, o que seria de grande interesse em um programa de seleção assistida no Brasil. No entanto, para tal finalidade, uma quantidade elevada de microssatélites seria necessária, o que demandaria maior tempo e investimento para o desenvolvimento de novos *primers* e a genotipagem.

A extensão do DL com base em poucas associações entre marcadores ligados não parece ter sido comprometida no presente estudo. Raboin et al. (2008) detectaram 163 associações em DL de um total de 1.488 possíveis entre locos ligados, correspondendo a uma proporção de aproximadamente $11{,}0\%$. Neste trabalho, como foi apresentado, 5 associações foram significativas de 60 possíveis,

correspondendo a aproximadamente $9{,}0\%$. Portanto, um número bastante próximo.

Quanto à estrutura de população, os resultados obtidos sugeriram a presença de 4 subpopulações no PBVCA. A análise realizada no STRUCTURE, baseada em um modelo probabilístico, claramente mostrou o comportamento instável das probabilidades a posteriori a partir da quinta subpopulação. Isso deve ter ocorrido pelo fato que as subpopulações ($K$) a priori são estabelecidas aleatoriamente, fazendo com que o modelo apresente probabilidades instáveis e não confiáveis no momento da convergência de subpopulações que, de fato, não devem representar o conjunto de dados. Em outras palavras, se tivessem sido assumidas, de maneira ocasional, 4 subpopulações a priori, esse comportamento não teria sido verificado. A inferência a favor de 4 subpopulações no PBVCA foi certificada com o gráfico de Neighbor-Joining, que apresentou padrões muito semelhantes com o método probabilístico. Ambos mostraram padrões que parecem estar de acordo com informações oriundas de *pedigree*.

Desses padrões, foi possível observar que houve uma tendência no agrupamento de indivíduos oriundos de um mesmo cruzamento, bem como indivíduos com apenas um dos parentais em comum. Porém, vários irmãos completos permaneceram em subpopulações diferentes, mostrando o contrário do que seria esperado do ponto de vista teórico. Na verdade, é bastante possível que indivíduos genealogicamente aparentados de cana-de-açúcar apresentem grandes distâncias genéticas, fazendo com que, eventualmente, permaneçam em subpopulações diferentes. Isso porque imagina-se que a elevada ploidia da espécie pode fazer com que padrões extremamente complexos de recombinação aconteçam, gerando um número considerável de genótipos diferentes.

No entanto, é pouco provável que indivíduos de determinada subpopulação não apresentem proporções genômicas de outras subpopulações, como foi assumido neste trabalho com o uso do modelo de não-mistura (PRITCHARD; STEPHENS; DONNELLY, 2000). Isso seria insatisfatório considerando que vários irmãos completos foram atribuídos a diferentes grupos, o que levaria a pensar que seus genomas não apresentam proporções comuns. Na verdade, aquele modelo foi utilizado em função do comportamento dominante dos microssatélites aqui analisados, os quais disponibilizaram pouca informatividade a respeito do genoma da cana-de-açúcar. Assim, por mais que o agrupamento aqui verificado corrobore com informações oriundas de *pedigree*, acredita-se que o modelo de mistura possa ser mais realista acerca da estrutura populacional em cana-de-açúcar. Não apenas considerando o PBVCA, o qual foi desenvolvido visando representar o germoplasma brasileiro, onde muitos indivíduos devem mesmo compartilhar proporções genômicas comuns. Mas

também germoplasmas diferentes, devido à ancestralidade comum a partir de poucos indivíduos, bem como ao intercâmbio de variedades existente entre os mesmos.

A análise do DL respeitando a estrutura de população revelou a possibilidade de mapeamento de QTL's dentro das subpopulações 1, 2 e 4. É possível especular que microssatélites podem estar fisicamente associados a QTL's que controlam características de interesse, o que seria de grande importância para um programa de seleção assistida por marcadores moleculares. No entanto, uma quantidade muito reduzida desses microssatélites mostrou-se envolvida com associações preferenciais de interesse, indicando baixa cobertura do genoma. Face à concentração de DL em pequenas distâncias, como foi visto no presente trabalho, uma cobertura genômica dessa natureza pode não contribuir para a detecção de marcadores e QTL's intimamente ligados. Essa contribuição se torna ainda mais limitada considerando o comportamento dominante dos marcadores obtidos, os quais não representam o genoma da cana-de-açúcar no seu contexto real e poliplóide.

A existência de DL por ligação física dentro de subpopulações deve ser vista com certa ressalva. Não apenas por limitações do presente estudo, mas também pelo contexto que envolve a cana-de-açúcar e o seu melhoramento. Assim, considerando que poucas gerações se passaram desde as variedades ancestrais até os híbridos atuais (JANNOO et al., 1999b; WEI et al., 2006), é bastante improvável que o DL detectado dentro de subpopulações seja apenas por ligação física. Além disso, o forte processo de seleção no melhoramento (HUANG; AITKEN; GEORGE, 2010) e a presença de parentesco entre indivíduos pertencentes à uma mesma subpopulação (YU; BUCKLER, 2006) reforçam a ideia de que falsas associações podem ter sido detectadas.

Os resultados obtidos no presente estudo representam um avanço no entendimento do processo histórico-evolutivo da cana-de-açúcar no Brasil, e certamente serão importantes para a futura realização do mapeamento associativo. No entanto, além daqueles que já foram discutidos, outros aspectos importantes devem ser apresentados com o objetivo de aprimorar pesquisas subsequentes.

O painel de variedades utilizado neste estudo foi desenvolvido segundo critérios importantes (ver item 3.1.1). Através desses critérios, foi possível agrupar diversos materiais de grande importância para o germoplasma brasileiro, acreditando que este tenha sido representado de maneira adequada. No entanto, devido à prática rotineira de cruzamentos endogâmicos nos programas de melhoramento, esse painel acabou sendo naturalmente constituído por indivíduos que estreitam relações de parentesco entre si, as quais têm sido muito importantes no momento da realização do mapeamento associativo (YU; BUCKLER, 2006; YU et al., 2006; ZHU et al., 2008).

Yu e Buckler (2006) e Yu et al. (2006) mencionam que o mapeamento associativo pode ser tão mais vantajoso quanto maior for a independência entre os indivíduos de uma população, tendo-se a redução da endogamia e o aumento do tamanho efetivo. Portanto, além da estrutura populacional, que foi verificada no presente estudo, informações de parentesco do painel brasileiro, já disponíveis para muitos dos seus indivíduos (LIMA et al., 2002), bem como de similaridade genética baseada em marcadores moleculares também deverão ser consideradas para a futura realização do mapeamento associativo no Brasil.

Outro aspecto importante, como já foi comentado anteriormente, é a natureza dos dados utilizados. Em cana-de-açúcar, os microssatélites, que são marcadores loco-específicos do tipo codominante, se comportaram como marcadores dominantes (presença e ausência), gerando informações parciais acerca do seu genoma poliplóide e altamente complexo. Em função disso, não foi possível detectar doses alélicas superiores de maneira a verificar diferentes classes genotípicas no padrão molecular das variedades, o que possibilitaria a reconstrução de haplótipos e, a partir destes, a estimação das frequências alélicas e gaméticas. Nesse contexto, o DL poderia ser estimado com base no modelo teórico apresentado por Gallais (2003), o qual foi proposto para organismos tetraplóides. Assim, o DL não seria simplesmente obtido a partir da presença e ausência de bandas, mas sim das frequências alélicas e gaméticas, considerando: i) DL entre alelos de um mesmo loco, presentes em haplótipos diferentes; e ii) DL entre alelos de locos diferentes, podendo estar em um mesmo haplótipo (associação) ou em haplótipos diferentes (repulsão) (ver item 2.3.4 para detalhes).

Como o DL não pôde ser aqui verificado de acordo com a abordagem anterior, o teste exato de Fisher foi utilizado como uma medida aproximada para seu cálculo. Ardlie, Kruglyak e Seielstad (2002) mencionam que esse teste, muito usado no início dos estudos de associação em humanos, pode não fornecer estimativas confiáveis do DL pelo fato de não ser uma medida própria, como as tradicionais $D^{'}$ e $r^2$. De fato, o uso de p-valores para medir o grau de associação entre locos pode não ser, a rigor, apropriado, visto que não há força de evidência para tal. Além disso, pode ser inadequado comparar os resultados obtidos no presente estudo com outros que também utilizaram o teste de Fisher para calcular o DL, já que os p-valores encontrados dependem fortemente das amostras que estão sendo consideradas (ARDLIE; KRUGLYAK; SEIELSTAD, 2002).

É importante lembrar que a extensão do DL e a estrutura de população foram obtidas com informações de um mapa genético pouco preciso, o qual foi utilizado para obtenção das distâncias entre marcadores e suas posições. Apesar de um avanço ter sido alcançado com o uso de uma abordagem

multiponto, apenas marcadores em dose única do tipo 1:1 e 3:1 foram utilizados na construção desse mapa, o que certamente não representa o contexto poliplóide e complexo da cana-de-açúcar. Com o uso de marcadores mais informativos, apresentando outros tipos de segregação de maneira a estender o polimorfismo ao longo do genoma, mapas genéticos mais confiáveis certamente serão obtidos. Mesmo que o sequenciamento genômico se torne uma realidade em cana-de-açúcar em função do surgimento de plataformas cada vez mais acessíveis, o uso de mapas genéticos confiáveis será indispensável (LEWIN et al., 2009) para análise da extensão do DL e da estrutura populacional, quando esta necessitar de informações de marcadores mapeados.

No presente estudo, uma visão global sobre DL e estrutura populacional foi apresentada com base em 135 indivíduos do PBVCA, o qual foi desenvolvido para representar o germoplasma brasileiro. Apesar de microssatélites genômicos e de sequências expressas terem gerado 1.474 marcadores polimórficos, número este equivalente ou superior ao polimorfismo detectado em outros estudos (JANNOO et al., 1999b; RABOIN et al., 2008; WEI et al., 2010), informações parciais acerca de um genoma poliplóide e altamente complexo foram consideradas. Em função disso, acredita-se que os resultados aqui obtidos sejam válidos mais para futuros direcionamentos do que aplicações propriamente ditas, já que um longo caminho precisa ser percorrido para desvendar a complexidade do genoma da cana-de-açúcar. No momento em que essa dissertação está sendo confeccionada, avanços estão sendo obtidos na construção de mapas de ligação e no estudo do DL com base em marcadores SNPs, os quais são mais informativos por fornecerem doses alélicas superiores e, assim, identificarem muitas das classes genotípicas dos indivíduos. Espera-se com isso caracterizar de maneira precisa o DL e a estrutura de população referentes ao PBVCA, considerando a grande complexidade genética da cana-de-açúcar, o que será fundamental para maior entendimento sobre o contexto histórico-evolutivo dessa espécie no Brasil e mapeamento mais refinado de QTL's relacionados a características de interesse.

# 6 CONCLUSÃO

Basicamente, observou-se que o desequilíbrio de ligação descresce aproximadamente a partir de 5-15 cM. Além disso, notou-se que o painel brasileiro é constituído por 4 subpopulações. Assim sendo, um avanço.

## REFERÊNCIAS

ABDURAKHMONOV, I.Y.; ABDUKARIMOV, A. Application of association mapping to understanding the genetic diversity of plant germplasm resources. **International Journal of Plant Genomics**, Cairo, 2008, doi:10.1155/2008/574927, p. 1-18.

AITKEN, K.S.; JACKSON, P.A.; McINTYRE, C.L. A combination of AFLP and SSR markers provides extensive map coverage and identification of homo(eo)logous linkage groups in a sugarcane cultivar. **Theoretical and Applied Genetics**, New York, v. 110, p. 789-801, 2005.

AITKEN, K.S.; JACKSON, P.A.; McINTYRE, C.L. Quantitative trait loci identified for sugar related traits in a sugarcane (*Saccharum* spp.) cultivar x *Saccharum officinarum* population. **Theoretical and Applied Genetics**, New York, v. 112, p. 1306-1317, 2006.

AITKEN, K.S.; JACKSON, P.A.; McINTYRE, C.L. Construction of a genetic linkage map for *Saccharum officinarum* incorporating both simplex and duplex markers to increase genome coverage. **Genome**, Ottawa, v. 50, p. 742-756, 2007.

AITKEN, K.S.; HERMANN, S.; KARNO, K.; BONNETT, G.D.; McINTYRE, C.L.; JACKSON, P.A. Genetic control of yield related stalk traits in sugarcane. **Theoretical and Applied Genetics**, New York, v. 117, p. 1191-1203, 2008.

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

ALLENDORF, F.W.; LUIKART, G.H. **Conservation and the genetics of populations.** Oxford: Blackwell Publishing, 2007. 642 p.

AL-JANABI, S.M.; HONEYCUTT, R.J.; McCLELLAND, M.; SOBRAL, B.W.S. A genetic linkage map of *Saccharum spontaneum* (L.) 'SES 208'. **Genetics**, Bethesda, v. 134, p. 1249-1260, 1993.

AL-JANABI, 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**, Netherlands, v. 17, p. 1-8, 1999.

AL-JANABI, S.M.; PARMESSUR, Y.; KROSS, H.; DHAYAN, S.; SAUMTALLY, S.; RAMDOYAL, K.; AUTREY, L.J.C.; DOOKUN-SAUMTALLY, A. Identification of a major quantitative trait locus (QTL) for yellow spot (*Mycovellosiella koepkei*) disease resistance in sugarcane. **Molecular Breeding**, Berlin, v. 19, p. 1-14, 2007.

ALWALA, S.; KIMBENG, C.A. Molecular genetic linkage mapping in *Saccharum*: strategies, resources and achievements. In: HENRY, R.J.; KOLE, C. (Eds.). **Genetics, genomics and breeding of sugarcane**. New Hampshire: Science Publishers, 2010. p. 69-96.

ARDLIE, K.; KRUGLYAK, L.; SEIELSTAD, M. Patterns of linkage disequilibrium in the human genome. **Nature Reviews Genetics**, New York, v. 112, p. 876-884, 2002.

AUSTIN, D.F.; LEE, M. Detection of quantitative trait loci for grain yield and yield components in maize across generations in stress and nonstress environments. **Crop Science**, Madison, v. 38, p. 1296-1308, 1998.

AVISE, J.C. **Molecular markers, natural history, and evolution**. 2.ed. London: Chapman & Hall, 2004. 541 p.

BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. **Journal of the Royal Statistical Society Series B (Methodological)**, London, v. 57, p. 289-300, 1995.

BENNETT, H. On the theory of random mating. **Annals of Eugenics**, London, v. 18, p. 311-317, 1954.

BERDING, N.; HOGARTH, M.; COX, M. Plant improvement of sugarcane. In: JAMES, G.L. (Ed.) **Sugarcane.** 2.ed. Oxford: Blackwell Science, 2004. p. 1-19.

BUTTERFIELD, M.K. **Marker assisted breeding in sugarcane:** a complex polyploid. 2007. 182 p. Tese (Doutorado) - University of Stellenbosch, South African, Stellenbosch, 2007.

CARDON, L.R.; BELL, J.I. Association study designs for complex diseases. **Nature Reviews Genetics**, New York, v. 2, p. 91-99, 2001.

CARNEIRO, M.S.; VIEIRA, M.L.C. Mapas genéticos em plantas. **Bragantia**, Campinas, v. 61, p. 89-100, 2002.

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

CHING, A.; CALDWELL, K.S.; JUNG, M.; DOLAN, M.; SMITH, O.S.; TINGEY, S.; MORGANTE, M.; RAFALSKI, A.J. SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. **BMC Genetics**, London, v. 3, p. 19-32, 2002.

CONAB – COMPANHIA NACIONAL DE ABASTECIMENTO. **Acompanhamento da Safra Brasileira - Cana-de-açúcar: Safra 2011/2012, Segundo Levantamento.** Disponível em: <http://www.conab.gov.br>. Acesso em: 22 set. 2011.

CORDEIRO, G.M.; TAYLOR, G.O.; HENRY, R.J. Characterisation of microsatellite markers from sugarcane (*Saccharum* sp.), a highly polyploid species. **Plant Science**, Limerick, v. 155, p. 161-168, 2000.

CRESTE, S.; TULMANN NETO, A.; FIGUEIRA, A. Detection of single sequence repeat polymorphisms in denaturing polyacrylamide sequencing gel by silver satining. **Plant Molecular Biology Reporter**, Netherlands, v. 19, p. 299-306, 2001.

CROW, J.F.; KIMURA, M. **An introduction to population genetics theory.** New Jersey: The Blackburn Press, 1970. 591 p.

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

DA SILVA, J.A.G.; SORRELLS, M.E.; BURNQUIST, W.; TANKSLEY, S.D. RFLP linkage map of *Saccharum spontaneum.* **Genome**, Ottawa, v. 36, p. 782-791, 1993.

DA SILVA, J.A.G.; HONEYCUTT, R.J.; BURNQUIST, W.; AL-JANABI, S.M.; SORRELLS, M.E.; TANKSLEY, S.D.; SOBRAL, W.S. *Saccharum spontaneum* L. 'SES 208' genetic linkage map combining RFLP and PCR based markers. **Molecular Breeding**, Berlin, v. 1, p. 165-179, 1995.

DAUGROIS, J.H.; GRIVET, L.; ROQUES, D.; HOARAU, J.Y.; LOMBARDI, H.; GLASZMANN, J.C.; D'HONT A. A putative major gene for rust resistence linked with a RFLP marker in Sugarcane cultivar 'R570'. **Theoretical and Applied Genetics**, New York, v. 92, p. 1059-1064, 1996.

DEVLIN, B.; RISCH, N. A comparison of linkage disequilibrium measures for fine-scale mapping. **Genomics**, San Diego, v. 29, p. 311-322, 1995.

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

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

DOERGE, R.W. Mapping and analysis of quantitative trait loci in experimental populations. **Nature Reviews Genetics**, New York, v. 3, p. 43-52, 2002.

EXCOFFIER, L.; SLATKIN, M. Maximum-likelihood estimation of molecular ha-plotype frequencies in a diploid population. **Molecular Biology and Evolution**, Chicago, v. 12, p. 921-927, 1995.

FALCONER, D.S.; MACKAY, T.F.C. **Introduction to quantitative genetics.** 4.ed. Essex, UK: Longman, 1996. 464 p.

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

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

FAY, M.P. Confidence intervals that match Fisher's exact or Blaker's exact tests. **Biostatistics**, Oxford, v. 11, p. 373-374, 2010.

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

FISHER, R.A. The logic of inductive inference. **Journal of the Royal Statistical Society**, London, v. 98, p. 39-82, 1935.

FISHER, R.A. The theory of linkage in polysomic inheritance. **Philosophical Transactions of the Royal Society of London**, London, v. 233, p. 55-87, 1947.

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

FNP - Consultoria & Comércio. **AGRIANUAL 2008**: Anuário Estatístico da Agricultura Brasileira. São Paulo, 2008. 497 p.

FRANKLIN, I.; LEWONTIN, R.C. Is the gene the unit of selection. **Genetics**, Bethesda, v. 65, p. 707-734, 1970.

FOULKES, A.S. **Applied statistical genetics with R:** for population-based association studies. New York: Springer, 2009. 252 p.

GALLAIS, A. **Quantitative genetics and breeding methods in autopolyploid plants.** Paris: INRA Editions, 2003. 515 p.

GARCIA, A.A.F.; KIDO, E.A.; MEZA, A.N.; SOUZA, H.M.B.; PINTO, L.R.; PASTINA, M.M.; LEITE, C.S.; DA SILVA, J.A.G.; ULIAN, E.C.; FIGUEIRA, A.; SOUZA, A.P. Development of an integrated genetic map of a sugarcane (*Saccharum* spp.) commercial cross, based on a maximum-likelihood approach for estimation of linkage and linkage phases. **Theoretical and Applied Genetics**, New York, v. 112, p. 298-314, 2006.

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**, Bethesda, v. 165, p. 759-769, 2003.

GAZAFFI, R.; OLIVEIRA, K.M.; SOUZA, A.P.; GARCIA, A.A.F. Sugarcane: breeding and genetic mapping. In: CORTEZ, L.A.B. (Ed.) **Sugarcane bioethanol:** R&D for productivity and sustainability. São Paulo: Blucher, 2010. p. 333-344.

GEIRINGER, H. On the probability theory of linkage in Mendelian heredity. **Annals of Mathematical Statistics**, Ann Arbor, v. 15, p. 25-57, 1944.

GORELICK, R.; LAUBICHLER, M.D. Decomposing multilocus linkage disequilibrium. **Genetics**, Bethesda, v. 166, p. 1581-1583, 2004.

GRATTAPAGLIA, D.; SEDEROFF, R. Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross mapping strategy and RAPD markers. **Genetics**, Bethesda, v. 137, p. 1121-1137, 1994.

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

GUIMARÃES, C.T.; HONEYCUTT, R.J.; SILLS, G.R.; SOBRAL, B.W.S. Genetic maps of *Saccharum officinarum* L. and *Saccharum robustum* Brandes and Jew. Ex. Grassl. **Genetics and Molecular Biology**, Ribeirão Preto, v. 22, p. 125-132, 1999.

GUO, S.W.; THOMPSON, E.A. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. **Biometrics**, Washington, v. 48, p. 361-372, 1992.

GUPTA, P.K.; RUSTGI, S.; KULWAL, P.L. Linkage disequilibrium and association studies in higher plants: present status and future prospects. **Plant Molecular Biology**, Dordrecht, v. 57, p. 461-485, 2005.

HAMBLIN, M.T.; FERNANDEZ, M.G.S.; CASA, A.M.; MITCHELL, S.E.; PATERSON, A.H.; KRESOVICH, S. Equilibrium processes cannot explain high levels of short- and medium-range linkage disequilibrium in the domesticated grass sorghum bicolor. **Genetics**, Bethesda, v. 171, p. 1247-1256, 2005.

HAMILTON, M.B. **Population genetics.** Oxford: Wiley-Blackwell, 2009. 407 p.

HARTL, D.L.; CLARK, A.G. **Principles of population genetics.** 4.ed. Sunderland: Sinauer Associates, 2007. 565 p.

HEDRICK, P.W. Genetic disequilibrium measures: proceed with caution. **Genetics**, Bethesda, v. 117, p. 331-341, 1987.

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

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

HILL, W.G.; ROBERTSON, A. Linkage disequilibrium in finite populations. **Theoretical and Applied Genetics**, New York, v. 38, p. 226-231, 1968.

HILL, W.G. Estimation of linkage disequilibrium in randomly mating populations. **Heredity**, London, v. 33, p. 229-239, 1974.

HILL, W.G. Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. **Theoretical Population Biology**, New York, v. 8, p. 117-126, 1975.

HOARAU, J.Y.; OFFMAN, B.; D'HONT, A.; RISTERUCCI, A.M.; ROQUES, D.; GLASZMANN, J.C.; GRIVET, L. Genetic dissection of a modern sugarcane cultivar (*Saccharum* spp.). I. Genome mapping with AFLP markers. **Theoretical and Applied Genetics**, New York, v. 103, p. 84-97, 2001.

HOGARTH, D.M. Genetics of sugarcane. In: HEINZ, D.J. (Ed.) **Sugarcane improvement through breeding**. Amsterdam: Elsevier, 1987. p. 255-271.

HOLLAND, J.B. Genetic architecture of complex traits in plants. **Current Opinion in Plant Biology**, London, v. 10, p. 156-161, 2007.

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

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

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

JAIN, S.K.; ALLARD, R.W. The effects of linkage, epistasis, and inbreeding on population changes under selection. **Genetics**, Bethesda, v. 53, p. 633-659, 1966.

JAMES, G.L. An introduction to sugarcane. In: JAMES, G.L. (Ed.) **Sugarcane.** 2.ed. Oxford: Blackwell Science, 2004. p. 1-19.

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

JANNOO, N.; GRIVET, L.; DOOKUN, A.; D'HONT, A.; GLASZMANN, J.C. Linkage disequilibrium among modern sugarcane cultivars. **Theoretical and Applied Genetics**, New York, v. 99, p. 1053-1060, 1999b.

JENNINGS, H.S. The numerical results of diverse systems of breeding, with res-pect to two pairs of characters, linked or independent, with special relation to the effects of linkage. **Genetics**, Bethesda, v. 2, p. 97-154, 1917.

JORDE, L.B. Linkage disequilibrium as a gene-mapping tool. **American Journal of Human Genetics**, Chicago, v. 56, p. 11-14, 1995.

KIM, Y.; FENG, S.; ZENG, Z-B. Measuring and partitioning the high-order linkage disequilibrium by multiple order Markov chains. **Genetic Epidemiology**, New York, v. 32, p. 301-312, 2008.

KIMURA, M. A model of a genetic system which leads to closer linkage by natural selection. **Evolution**, Lawrence, v. 10, p. 278-287, 1956.

KOSAMBI, D.D. The estimation of map distances from recombination values. **Annual of Eugenics**, London, v. 12, p. 172-175, 1944.

KRUGLYAK, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. **Nature Genetics**, New York, v. 22, p. 139-144, 1999.

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

LEWIN, H.A.; LARKIN, D.M.; PONTIUS, J.; O'BRIEN, S.J. Every genome sequence needs a good map. **Genome Research**, Woodbury, v. 19, p. 1925-1928, 2009.

LEWONTIN, R.C.; KOJIMA, K. The evolutionary dynamics of complex polymorphisms. **Evolution**, Lawrence, v. 14, p. 458-472, 1960.

LEWONTIN, R.C. The interaction of selection and linkage. I. General considerations; heterotic models. **Genetics**, Bethesda, v. 49, p. 49-67, 1964.

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

LEWONTIN, R.C. The detection of linkage disequilibrium in molecular sequence data. **Genetics**, Bethesda, v. 140, p. 377-388, 1995.

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

LIN, M.; LOU, X.Y.; CHANG, M.; WU, R. A general statistical framework for mapping quantitative trait loci in nonmodel systems: issue for characterizing linkage phases. **Genetics**, Bethesda, v. 165, p. 901-913, 2003.

LONG, J.C.; WILLIAMS, R.C.; URBANEK, M. An E-M algorithm and testing strategy for multiple-locus haplotypes. **American Journal of Human Genetics**, Chicago, v. 56, p. 799-810, 1995.

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

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

MACKAY, T.F.C. The genetic architecture of quantitative traits. **Annual Review of Genetics**, Palo Alto, v. 35, p. 303-339, 2001a.

MACKAY, T.F.C. Quantitative trait loci in *Drosophila*. **Nature Reviews Genetics**, New York, v. 2, p. 11-20, 2001b.

MACKAY, T.F.C.; STONE, E.A.; AYROLES, J.F. The genetics of quantitative traits: challenges and prospects. **Nature Reviews Genetics**, New York, v. 10, p. 565-577, 2009.

MARGARIDO, G.R.A.; SOUZA, A.P.; GARCIA, A.A.F. OneMap: software for genetic mapping in outcrossing species. **Hereditas**, Lund, v. 144, p. 78-79, 2007.

MARGARIDO, G.R.A. **Mapeamento de QTL's em múltiplos caracteres e ambientes em um cruzamento comercial de cana-de-açúcar usando modelos mistos.** 2011. 107 p. Tese (Doutorado em Genética e Melhoramento de Plantas) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2011.

MATSUOKA, S.; GARCIA, A.A.F.; CALHEIROS, G.C. Hibridação em cana-de-açúcar. In: BORÉM, A. (Ed.) **Hibridação artificial em plantas.** Viçosa: UFV, 1999. p. 221-256.

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

MCINTYRE, C.L.; JACKSON, P.A.; CORDEIRO, G.M.; AMOUYAL, O.; HERMANN, S.; AITKEN, K.S.; ELIOTT, F.; HENRY, R.J.; CASU, R.E.; BONNETT, G.D. The identification and characterisation of alleles of sucrose phosphate synthase gene family III in sugarcane. **Molecular Breeding**, Berlin, v. 18, p. 39-50, 2006.

MCVEAN, G. Linkage disequilibrium, recombination and selection. In: BALDING, D.J.; BISHOP, M.; CANNINGS, C. (Eds.). **Handbook of statistical genetics.** 3.ed. New York: John Wiley & Sons, 2007. p. 909-944.

MEHTA, C.R.; PATEL, N.R. A network algorithm for performing Fisher's exact test in $r$ x $c$ contingency tables. **Journal of American Statistical Association**, Alexandria, v. 78, p. 427-434, 1984.

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

MUDGE, J.; ANDERSON, W.R.; KEHRER, R.L.; FAIRBANKS, D.J. A RAPD genetic map of *Saccharum officinarum.* **Crop Science**, Madison, v. 36, p. 1362-1366, 1996.

MUELLER, J.C. Linkage disequilibrium for different scales and applications. **Briefings in Bioinformatics**, London, v. 5, p. 355-364, 2004.

NEI, M.; LI, W-H. Linkage disequilibrium in subdivided populations. **Genetics**, Bethesda, v. 75, p. 213-219, 1973.

NIELSEN, D.M.; EHM, M.G.; ZAYKIN, D.V.; WEIR, B.S. Effect of two- and three-locus linkage disequilibrium on the power to detect marker/phenotype associations. **Genetics**, Bethesda, v. 168, p. 1029-1040, 2004.

NORDBORG, M.; BOREVITZ, J.O.; BERGELSON, J.; BERRY, C.C.; CHORY, J.; HAGENBLAD, 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, p. 190-193, 2002.

NORDBORG, M.; HU, T.T.; ISHINO, Y.; JHAVERI, 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.; SCHULTZ, V.; KREITMAN, M.; BERGELSON, J. The pattern of polymorphism in *Arabidopsis thaliana*. **PLoS Biology**, Cambridge, v. 3, 2005.

ODONG, T.L.; HEERWAARDEN, J. van.; JANSEN, J.; HINTUM, T.J.L. van.; VAN EEUWIJK, F.A. van. Determination of genetic structure of germplasm collections: are traditional hierarchical clustering methods appropriate for molecular marker data?. **Theoretical and Applied Genetics**, New York, v. 123, p. 195-205, 2011.

OHTA, T.; KIMURA, M. Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. **Genetics**, Bethesda, v. 63, p. 229-238, 1969.

OLIVEIRA, K.M.; PINTO, L.R.; MARCONI, T.G.; MARGARIDO, G.R.A.; PASTINA, M.M.; TEIXEIRA, L.H.M.; FIGUEIRA, A.M.; ULIAN, E.C.; GARCIA, A.A.F.; SOUZA, A.P. Functional genetic linkage map on EST-markers for a sugarcane (*Saccharum* spp.) commercial cross. **Molecular Breeding**, Berlin, v. 20, p. 189-208, 2007.

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

PASTINA, M.M. **Mapeamento de QTL's e estudo da interação entre QTL's, ambientes e cortes em cana-de-açúcar, usando a abordagem de modelos mistos.** 2010. 89 p. Tese (Doutorado em Genética e Melhoramento de Plantas) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2010.

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

PATTERSON, N.; PRICE, A.L.; REICH, D. Population structure and eigenanalysis. **PLoS Genetics**, Cambridge, v. 2, p. 2074-2093, 2006.

PERRIER, X.; FLORI, A.; BONNOT, F. Methods of data analysis. In: HAMON, P.S.; SEGUIN, M.; PERRIER, X.; GLASZMANN, J.C. (Eds.) **Genetic diversity of cultivated tropical plants.** Montpellier: Cirad, 2003. p. 31-63.

PINTO, L.R.; OLIVEIRA, K.M.; ULIAN, E.C.; GARCIA, A.A.F.; SOUZA, A.P. Survey in the sugarcane expressed sequence tag database (SUCEST) for simple sequence repeats. **Genome**, Ottawa, v. 47, p. 795-804, 2004.

PINTO, L.R.; GARCIA, A.A.F.; PASTINA, M.M.; TEIXEIRA, L.H.M.; BRESSIANI, J.A.; ULIAN, E.C.; BIDOIA, M.A.P.; SOUZA, A.P. Analysis of genomic and functional RFLP derived markers associated with sucrose content, fiber and yield QTLs in a sugarcane (*Saccharum* spp.) commercial cross. **Euphytica**, Wageningen, v. 172, p. 313-327, 2010.

PIPERIDIS, N.; JACKSON, P.A.; D'HONT, A.; BESSE, P.; HOARAU, J.Y.; COURTOIS, B.; AITKEN, K.S.; MCINTYRE, C.L. Comparative genetics in sugarcane enables structured map enhancement and validation of marker-trait associations. **Molecular Breeding**, Berlin, v. 21, p. 233-247, 2008.

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

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

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

R DEVELOPMENT CORE TEAM (2011). **R:** A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria, 2011. Disponível em: <http://www.r-project.org/>. Acesso em: 10 set. 2011.

RABOIN, L.M.; OLIVEIRA, K.M.; LECUNFF, L.; TELISMART, H.; ROQUES, D.; BUTTERFIELD, M.; HOARAU, J.Y.; D'HONT, A. Genetic mapping in sugarcane, a high polyploid, using biparental progeny: identification of a gene controlling stalk colour and a new rust resistance gene. **Theoretical and Applied Genetics**, New York, v. 112, p. 1382-1391, 2006.

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

ROBBINS, R.B. Some applications of mathematics to breeding problems II. **Genetics**, Bethesda, v. 3, p. 73-92, 1918.

ROBINSON, W.P.; ASMUSSEN, M.A.; THOMSON, G. Three-locus systems impose additional constraints on pairwise disequilibria. **Genetics**, Bethesda, v. 129, p. 925-930, 1991.

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

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

SLATKIN, M. On treating the chromosome as the unit of selection. **Genetics**, Bethesda, v. 72, p. 157-168, 1972.

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

SLATKIN, M.; EXCOFFIER, L. Testing for linkage disequilibrium in genotypic data using the Expectation-Maximization algorithm. **Heredity**, London, v. 76, p. 377-383, 1996.

SLATKIN, M. Linkage disequilibrium - understanding the evolutionary past and mapping the medical future. **Nature Reviews Genetics**, New York, v. 9, p. 477-485, 2008.

SOBRAL, B.W.S.; HONEYCUTT, R.J. High output genetic mapping of polyploids using PCR-generated markers. **Theoretical and Applied Genetics**, New York, v. 86, p. 105-112, 1993.

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

SREENIVASAN, T.V.; AHLOOWALIA, B.S.; HEINZ, D.J. Cytogenetics. In: HEINZ, D.J. (Ed.) **Sugarcane improvement through breeding.** Amsterdam: Elsevier, 1987. p. 211-253.

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

STICH, B.; MELCHINGER, A.E.; FRISCH, M.; MAURER, H.P.; HECKENBERGER, M.; REIF, J.C. Linkage disequilibrium in european elite maize germplasm investigated with SSRs. **Theoretical and Applied Genetics**, New York, v. 111, p. 723-730, 2005.

STOREY, J.D.; TIBSHIRANI, R. Statistical significance for genome-wide experiments. **Proceedings of the National Academy of Sciences**, Washington, v. 100, p. 9440-9445, 2003.

TENAILLON, M.I.; SAWKINS, M.C.; LONG, A.D.; GAUT, R.L.; DOEBLEY, J.F.; GAUT, B.S. 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, p. 9161-9166, 2001.

TEMPLETON, A.R. **Population genetics and microevolutionary theory.** New Jersey: John Wiley & Sons, 2006. 705 p.

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

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

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

WEINBERG, W. Über vererbungsgesetze beim menschen. **Z. Indukt Abstammungs Vererbungsl.**, Sem local, v. 1, p. 276-330, 1909.

WEIR, B.S. **Genetic Data Analysis II.** Sunderland, MA: Sinauer, 1996. 376 p.

WEIR, B.S. Linkage disequilibrium and association mapping. **Annual Review of Genomics and Human Genetics**, Palo Alto, v. 9, p. 129-142, 2008.

WEIR, B.S. Inferences about linkage disequilibrium. **Biometrics**, Washington, v. 35, p. 235-254, 1979.

WRIGHT, S. **Evolution and the genetics of populations**. v.2. theory of gene frequencies. Chicago: University of Chicago, 1969. 505 p.

WU, K.K.; MING, R.; MOORE, P.H.; PATERSON, A.H. Sugarcane genomics and breeding. In: LAMKEY, K.R.; LEE, M. (Eds.). **Plant breeding:** The Arnel R. Hallauer International Symposium. Oxford: Blackwell Publishing, 2006. p. 283-292.

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

YU, J.; PRESSOIR, G.; BRIGGS, W.H.; BI, I.V.; YAMASAKI, M.; DOEBLEY, J.F.; MCMULLEN, M.D.; GAUT, B.S.; NIELSEN, D.M.; HOLLAND, J.B.; KRESOVICH, S.; BUCKLER, E.S. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. **Nature Genetics**, New York, v. 38, p. 203-208, 2006.

XIONG, M.; GUO, S-W. Fine-scale genetic mapping based on linkage disequilibrium: theory and applications. **American Journal of Human Genetics**, Chicago, v. 60, p. 1513-1531, 1997.

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

ZENG, Z-B. **Statistical methods for mapping quantitative trait loci.** Raleigh: Department of Statistics, North Carolina State University, 2001. 128 p.

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