Universidade de São Paulo
Escola Superior de Agricultura Luiz de Queiroz

Análise da estrutura populacional e do desequilíbrio de ligação de um painel de acessos de sorgo: uma abordagem usando teoria da coalescência

João Ricardo Bachega Feijó Rosa

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

Piracicaba
2016

João Ricardo Bachega Feijó Rosa
Engenheiro Agrônomo

Análise da estrutura populacional e do desequilíbrio de ligação de um painel de acessos de sorgo: uma abordagem usando teoria da coalescência

versão revisada de acordo com a resolução CoPGr 6018 de 2011

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

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

Piracicaba
2016

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

Rosa, João Ricardo Bachega Feijó

Análise da estrutura populacional e do desequilíbrio de ligação de um painel de acessos de sorgo: uma abordagem usando teoria da coalescência / João Ricardo Bachega Feijó Rosa. – – versão revisada de acordo com a resolução CoPGr 6018 de 2011. – – Piracicaba, 2016 .

83 p.

Tese (Doutorado) – – USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Passado Evolutivo 2. Fluxo Gênico 3. Recombinação 4. Associações Não-aleatórias 5. Mapeamento Associativo 6. *Sorghum bicolor* . I. Título.

DEDICATÓRIA

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**.

À minha noiva, **Natália**, o grande amor da minha vida. Meu caminho se transformou completamente com a sua chegada. Pelo amor, carinho, companherismo, amizade, paciência, força e motivação em todos os momentos das nossas vidas: **Dedico**.

Ao meu cunhado, **Gilberto**, o grande irmão que não tive em minha infância. Pelo companherismo, amizade, força e humildade: **Dedico**.

AGRADECIMENTOS

À Deus, primeiramente. Enorme foi a graça concedida durante o esforço da caminhada. Força, luta, coragem, dedicação e persistência. Paciência, tranquilidade e serenidade. Faltam-me palavras para agradecer-lhe por toda honra e glória conquistadas.

Às pessoas mais importantes da minha vida: Maria Fátima (mãe), Natália (noiva), Daniela (irmã) e Gilberto (cunhado). Muito obrigado por vocês fazerem parte da minha vida e por terem estado ao meu lado, concedendo muita força, durante esta caminhada. Só vocês sabem o quanto foi difícil continuar neste caminho por uma série de razões que não cabe ser exposta neste momento. Muitíssimo obrigado! Serei eternamente grato por tudo!

Meus sinceros agradecimentos ao Departamento de Genética e ao Programa de Pós-Graduação em Genética e Melhoramento de Plantas da ESALQ/USP. Foi uma honra ter tido a oportunidade de cursar o doutorado (e o mestrado) nesta casa, que é digna de grande excelência e brilhantismo. Grande aprendizado foi-me concedido ao longo dos últimos seis anos, que certamente se eternizarão em minha memória.

Um agradecimento mais que especial ao Prof. Dr. Antonio Augusto Franco Garcia, pelo comprometimento em tornar-me um cientista. Lembro-me como se fosse hoje o nosso primeiro encontro em 2007, quando estive na ESALQ para manifestar interesse em sua orientação. Tive uma das melhores recepções da minha vida, que, de lá para cá, mudou e se transformou completamente. Faltam-me palavras para agradecer por tudo - acolhimento, abertura, orientação, ensinamentos, oportunidades, conversas extremamentes produtivas, compreensão, paciência e amizade. Sinto-me honrado e satisfeito por ter feito parte do seu grupo e da sua história.

Outro agradecimento especial ao meu sogro Orlando Volpe Júnior, à minha sogra Maria Alice Masiero Volpe e aos meus cunhados Henrique Volpe e Ana Cláudia Francelin. Vocês foram a minha segunda família ao longo de todos esses anos de muito esforço e empenho.

Aos pesquisadores Dr. Jurandir Vieira Magalhães e Drª. Maria Marta Pastina, da Embrapa Milho e Sorgo (Sete Lagoas/MG). O apoio de vocês e a troca de ideias foram muito importantes para a utilização dos dados de sorgo disponíveis e a realização deste trabalho. Foi uma imensa honra ter tido o apoio de profissionais como vocês, que possuem competência e brilhantismo inigualáveis.

Ao Prof. Dr. Geoffrey Morris, do *Department of Agronomy*, *Crop Genetics* & *Genomics*, *Kansas State University*, por ter tornado público o conjunto de dados genotípicos do painel mundial de acessos de sorgo utilizado neste trabalho. Sem tais dados e informações históricas relevantes acerca da espécie, que só podem ter sido elaboradas por cientistas renomados de sorgo, não teríamos tido a motivação de realizar o presente estudo.

À Profª. Drª. Anete Pereira de Souza, da Universidade Estadual de Campinas (UNICAMP), pela concessão da bolsa de doutorado sanduíche realizado na França. Embora o projeto desenvolvido neste país não esteve diretamente relacionado a este trabalho, a oportunidade e a experiência de ter vivido em outro país, com suas crenças e seus costumes, foram incríveis. A experiência profissional adquirida pelo desenvolvimento do estudo por lá realizado foi algo inestimável para a minha carreira. Muito obrigado por ter acreditado em meu trabalho.

Ao Dr. Vincent Le Guen, do Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD, Montpellier/França). Assim como aconteceu com o Prof. Dr. Antonio Augusto Franco Garcia, tive uma das melhores recepções da minha vida. Ter sido supervisionado por você durante o tempo em que

estive no CIRAD e na França foi uma honra e um grande prazer para mim. Muito obrigado por ter confiado em meu trabalho! E, além disso, tenho certeza que conquistei um grande amigo!

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela bolsa concedida durante o curso de doutorado (processo 141117/2012-5), e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela concessão da bolsa de doutorado sanduíche (processo 3243/15-0) realizado na França.

A todos os colegas que fizeram e fazem parte do Laboratório de Genética Estatística da ESALQ/USP: (i) Aos mais *antigos*: Rodrigo, Maria Marta, Marcelo, Graciela, Gabriel, Renato e Edjane; (ii) Aos *contemporâneos*: Carina, Luciano, Guilherme, Adriana, Rodrigo Júnior, Maria Izabel e Letícia; (iii) Aos mais *novos*: Rafael, Felipe, Marianella, Letícia Lara, Danilo, Cristiane e Jhonathan; e (iv) Aos *agregados*: Amanda e Fernando. Cada um de vocês foi muito importante de alguma forma durante a minha caminhada. Alguns foram importantes por fornecer apoio pessoal e profissional, enquanto que outros foram importantes pela agradável convivência e por tornar o nosso ambiente de trabalho mais alegre e divertido.

Um agradecimento especial aos colegas mais próximos do Laboratório: Carina, Renato, Guilherme, Adriana e Felipe. Pela agradável convivência em todos os momentos e pelas conversas científicas inteligentes e extremamente produtivas. À Carina, que esteve ao meu lado desde o princípio e se tornou uma grande amiga. Ao Renato, com quem aprendi muito durante o tempo em que esteve no laboratório e por quem tenho grande admiração. Ao Guilherme, pela conversas super inteligentes sobre ciência e pela competência. À Adriana, pelo exemplo de pessoa íntegra, amiga e profissional competente. Ao Felipe, pelo exemplo de dedicação e esforço em quaisquer circunstâncias.

Aos grandes amigos Fernando Guerra e Augusto Lima, com quem pude vivenciar grandes momentos durante o tempo em que estive em Piracicaba.

A todos os professores do Departamento de Genética e de Estatística e Experimentação Agronômica da ESALQ/USP. De uma forma ou de outra, vocês foram muito importantes pelos avanços alcançados durante o doutorado e pela confecção do presente trabalho.

Aos funcionários do Departamento de Genética da ESALQ/USP: Seu Antônio, Marcos, Valdir, Berdan, Léia, Macedônio, Fernandinho e Maídia, pelos grandiosos auxílios nos momentos decisivos. Aos funcionários da Biblioteca Central: Eliana, Cristina e Glória, pela ajuda na confecção final deste documento.

Finalmente, a todas as pessoas que direta ou indiretamente contribuíram para que eu chegasse até aqui e pudesse realizar este trabalho. Muito obrigado!

EPÍGRAFE

"Talvez não tenhamos 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)

SUMÁRIO

RESUMO

Análise da estrutura populacional e do desequilíbrio de ligação de um painel de acessos de sorgo: uma abordagem usando teoria da coalescência

A estrutura populacional e o desequilíbrio de ligação são dois processos fundamentais para estudos evolutivos e de mapeamento associativo. Tradicionalmente, ambos têm sido investigados por meio de métodos clássicos comumente utilizados. Tais métodos certamente forneceram grandes avanços no entendimento dos processos evolutivos das espécies. No entanto, em geral, nenhum deles utiliza uma visão genealógica de forma a considerar eventos genéticos ocorridos no passado, dificultando a compreensão dos padrões de variação observados no presente. Uma abordagem que possibilita a investigação retrospectiva com base no atual polimorfismo observado é a teoria da coalescência. Assim, o objetivo deste trabalho foi analisar, com base na teoria da coalescência, a estrutura populacional e o desequilíbrio de ligação de um painel mundial de acessos de sorgo (*Sorghum bicolor*). Para tanto, análises de mutação, migração com fluxo gênico e recombinação foram realizadas para cinco regiões genômicas relacionadas à altura de plantas e maturidade (*Dw1*, *Dw2*, *Dw4*, *Ma1* e *Ma3*) e sete populações previamente selecionadas. Em geral, elevado fluxo gênico médio ($\mathcal{M} = \frac{m}{\mu} = 41{,}78 - 52{,}07$) foi observado entre as populações considerando cada região genômica e todas elas simultaneamente. Os padrões sugeriram intenso intercâmbio de acessos e história evolutiva específica para cada região genômica, mostrando a importância da análise individual dos locos. A quantidade média de migrantes por geração ($\mathcal{M}$) não foi simétrica entre pares recíprocos de populações, de acordo com a análise individual e simultânea das regiões. Isso sugere que a forma pela qual as populações se relacionaram e continuam interagindo evolutivamente não é igual, mostrando que os métodos clássicos utilizados para investigar estrutura populacional podem ser insatisfatórios. Baixas taxas médias de recombinação ($\rho_L = 2N_e r = 0{,}030 - 0{,}246$) foram observadas utilizando o modelo de recombinação constante ao longo da região. Baixas e altas taxas médias de recombinação ($\rho_r = 2N_e r = 0{,}060 - 3{,}395$) foram estimadas utilizando o modelo de recombinação variável ao longo da região. Os métodos tradicional (r^2) e via coalescência ($E[r^2_{\texttt{rhomap}}]$) utilizados para a estimação do desequilíbrio de ligação mostraram resultados próximos para algumas regiões genômicas e populações. No entanto, o r^2 sugeriu padrões descontínuos de desequilíbrio em várias ocasiões, dificultando o entendimento e a caracterização de possíveis blocos de associação. O método via coalescência ($E[r^2_{\texttt{rhomap}}]$) forneceu resultados que pareceram ter sido mais consistentes, podendo ser uma estratégia eventualmente importante para um refinamento dos padrões não-aleatórios de associação. Os resultados aqui encontrados sugerem que o mapeamento genético a partir de um único *pool* gênico pode ser insuficiente para detectar associações causais importantes para características quantitativas em sorgo.

Palavras-chave: Passado Evolutivo, Fluxo Gênico, Recombinação, Associações Não-aleatórias, Mapeamento Associativo, *Sorghum bicolor*

ABSTRACT

Analysis of population structure and linkage disequilibrium of a sorghum accession panel: an approach using coalescent theory

Population structure and linkage disequilibrium are two fundamental processes for evolution and association mapping studies. Traditionally, both have been investigated using classical methods that are commonly used. These methods certainly provided important advances for the understanding of the evolution processes of the species. However, in general, none of them uses a genealogical view to consider genetic events occurred in the past, making difficult the understanding of the variation patterns observed in the present. An approach that enables the retrospective investigation based on the actual observed polymorphism is the coalescent theory. Here, we used the coalescent theory to analyze the population structure and linkage disequilibrium of a worldwide sorghum (*Sorghum bicolor*) accession panel. To reach this purpose, analyses of mutation, migration with gene flow and recombination were performed to five genomic regions related to plant height and maturity (*Dw1*, *Dw2*, *Dw4*, *Ma1* e *Ma3*) and seven previously selected populations. In general, high average gene flow ($\mathcal{M} = \frac{m}{\mu} = 41{,}78 - 52{,}07$) was observed between populations considering each genomic region and all the regions simultaneously. The patterns suggested a high exchance of accessions between populations and a specific evolutionary history for each genomic region, showing that the individual analysis of each locus was important. The average number of migrants per generation ($\mathcal{M}$) was not symmetric between reciprocal pairs of populations, according to the specific and simultaneous analyses of the regions. This result suggests that the historical and recent evolutionary relations between populations are not equal, showing that the classical methods to investigate population structure may be unsatisfactory. Low average recombination rates ($\rho_L = 2N_e r = 0{,}030 - 0{,}246$) were observed using a constant recombination model along the region. Low and high average recombination rates ($\rho_r = 2N_e r = 0{,}060 - 3{,}395$) were estimated using a variable recombination model along the region. Both traditional (r^2) and coalescent ($E[r^2_{\texttt{rhomap}}]$) methods for the estimation of linkage disequilibrium showed similar results for some genomic regions and populations. However, r^2 suggested discontinuous patterns of linkage disequilibrium in several cases, making difficult the understanding and definition of the association blocks. The coalescent method ($E[r^2_{\texttt{rhomap}}]$) provided results that seemed to be more consistent and could be an eventually important strategy to refine the non-random association patterns. The results detected here suggest that the genetic mapping from a unique gene *pool* may be insufficient to detect important causal associations for quantitative traits in sorghum.

Keywords: Evolutionary Past, Gene Flow, Recombination, Non-random Associations, Association Mapping, *Sorghum bicolor*

1 INTRODUÇÃO

Tradicionalmente, a estrutura populacional em painéis de associação e populações naturais tem sido investigada com base em métodos clássicos bastante conhecidos. Basicamente, tais métodos consistem em agrupamentos hierárquicos (Odong et al., 2011), inferência bayesiana (STRUCTURE) (Pritchard et al., 2000), componentes principais (Zhu et al., 2008), F_{ST} (Wright, 1931) e suas extensões (Nath and Griffiths, 1996). Certamente, grandes contribuições foram geradas pelo uso de tais abordagens. No entanto, em geral, todos eles utilizam as frequências alélicas dos locos para capturar níveis de diferenciação que podem ser o resultado de eventos de fluxo gênico recentes, pouco considerando eventos migratórios ocorridos no passado (Wilson and Rannala, 2003; Hübner et al., 2012). Além disso, esses métodos não fornecem uma estimativa de fluxo gênico com a qual é possível conhecer a magnitude da relação histórica entre populações, dificultando o entendimento do passado evolutivo. Assim, uma visão genealógica da estrutura populacional e do fluxo gênico é de grande motivação, sobretudo para fins de mapeamento associativo, que é o foco do presente trabalho.

Da mesma forma, o desequilíbrio de ligação presente em populações tem sido analisado com base em medidas relativas bastante conhecidas, tais como $D^{'}$ (Lewontin, 1964) e r^2 (Hill and Robertson, 1968). Normalmente, tais medidas apresentam grandes variações em suas estimativas (Nordborg and Tavaré, 2002), dificultando uma clara interpretação dos padrões de associação (Evans and Cardon, 2005). Além disso, $D^{'}$ e r^2 apresentam limitações por não considerarem associações entre múltiplos locos simultaneamente e por não permitirem explicações que mostrem as possíveis razões de locos estarem preferencialmente associados ao longo das gerações (Pritchard and Przeworski, 2001; Slatkin, 2008). Dessa forma, uma visão genealógica do desequilíbrio de ligação, com atenções voltadas ao mapeamento associativo, é bastante desejável (Nordborg and Tavaré, 2002; Zöllner and Pritchard, 2005).

Uma das estratégias que podem ser utilizadas para se obter uma visão genealógica de processos evolutivos é a teoria da coalescência (Kingman, 1982a,b,c). A abordagem proposta por esta teoria possibilita investigar diferentes genealogias caminhando-se do presente em direção ao passado, num processo retrospectivo (Fu and Li, 1999). Tais genealogias podem ser investigadas de forma a considerar processos evolutivos importantes, tais como mutação, fluxo gênico e recombinação (Hudson, 1990; Pritchard and Przeworski, 2001; Hein et al., 2004; Marjoram and Tavaré, 2006; Nordborg, 2007), fornecendo informações relevantes acerca do passado. No contexto do mapeamento associativo, a teoria da coalescência poderá ser importante para se ter maior compreensão dos padrões de associação entre locos, o que, em última análise, será fundamental para o entendimento das relações entre genótipo e fenótipo.

Em plantas domesticadas, o uso da coalescência parece ser bastante restrito. Em sorgo (*Sorghum bicolor*), tal abordagem foi utilizada para estimar taxas de recombinação populacional em regiões genômicas específicas (Hamblin et al., 2005) ou ao longo de todo o genoma (Morris et al., 2013). Contudo, embora esses trabalhos tenham gerado resultados importantes e fornecido uma ideia dos padrões de associação, nenhum deles direcionou atenções para o cálculo do desequilíbrio de ligação a partir da recombinação via coalescência. Veremos mais adiante que o desequilíbrio de ligação pode ser diretamente obtido a partir das taxas de recombinação. Além disso, até o nosso conhecimento, nenhum trabalho foi publicado para o sorgo considerando a estrutura populacional e o fluxo gênico estimados via coalescência.

Neste contexto, o objetivo deste trabalho foi analisar, com base na teoria da coalescência, a estrutura populacional e o desequilíbrio de ligação presentes em um painel mundial de acessos de sorgo. Para alcançar tal objetivo, análises de mutação, migração com fluxo gênico e recombinação foram realizadas para diferentes regiões genômicas e populações que constituem o presente painel. Buscou-se com isso um maior entendimento dos processos evolutivos ocorridos no passado, que podem ser cruciais para o mapeamento genético de características quantitativas importantes. Acredita-se que os resultados aqui gerados possam fornecer informações relevantes para estudos futuros de mapeamento associativo em sorgo.

2 REVISÃO BIBLIOGRÁFICA

Basicamente, a revisão bibliográfica está estruturada em três partes. Na primeira parte, serão abordados conceitos gerais e importantes sobre o desequilíbrio de ligação, que é um dos objetivos do presente trabalho. Não abordaremos os métodos clássicos utilizados para inferir estrutura populacional por acreditarmos que tais métodos estão amplamente discutidos na ciência. Na segunda parte, serão abordados conceitos importantes e os modelos evolutivos da teoria da coalescência, incluindo estrutura populacional (migração com fluxo gênico) e recombinação. Por fim, na terceira parte será apresentado um panorama geral sobre o sorgo, relatando desde a sua origem e domesticação até as estratégias adotadas pelo melhoramento genético, descrevendo alguns genes importantes que serão utilizados no presente estudo. Concluímos a revisão bibliográfica com uma rápida descricão sobre a genotipagem por sequenciamento, que também será utilizada neste trabalho.

2.1 Desequilíbrio de Ligação

Desequilíbrio de ligação (DL) é qualquer desvio das frequências genotípicas em relação às frequências esperadas sob independência, indicando associação preferencial entre alelos de diferentes locos em uma população (Lewontin and Kojima, 1960). Apesar do termo ter tido grande aceitação por parte da comunidade científica, DL não corresponde necessariamente à ligação física (Flint-Garcia et al., 2003; Templeton, 2006; Allendorf and Luikart, 2007; Hamilton, 2009; Hedrick, 2010). DL pode ocorrer tanto entre locos ligados quanto entre locos não ligados (Pritchard and Przeworski, 2001; Ardlie et al., 2002), enquanto que ligação física está relacionada a alelos que cosegregam no mesmo cromossomo ao longo das gerações (Flint-Garcia et al., 2003; Gupta et al., 2005). Assim, é fundamental que esses conceitos não sejam confundidos, já que abordagens importantes baseiam-se no DL por ligação física, como é o caso dos estudos de mapeamento genético.

Grande parte da teoria sobre DL é descrita para pares de locos bialélicos (Slatkin et al., 1996). Embora o DL esteja fundamentado no conceito proposto por Lewontin and Kojima (1960), designado por D, outras medidas, como D' (Lewontin, 1964) e r^2 (Hill and Robertson, 1968), têm sido muito utilizadas. No entanto, devido à obtenção frequente de amostras pequenas (Flint-Garcia et al., 2003), a segunda delas tem sido mais amplamente utilizada, como pode ser verificado nos estudos de mapeamento associativo em humanos (Pritchard and Przeworski, 2001; Slatkin, 2008) e plantas (Flint-Garcia et al., 2003; Gupta et al., 2005). De qualquer forma, espera-se que o DL seja melhor compreendido no contexto de pares de locos multialélicos e múltiplos locos bialélicos (Kim et al., 2008), o que certamente fornecerá informações mais refinadas para estudos de mapeamento genético (McVean, 2007a).

Vários fatores populacionais afetam o DL. Mutação, deriva genética, recombinação, estrutura populacional, fluxo gênico, endogamia e seleção podem ser considerados os principais fatores que influenciam no aumento ou na diminuição do DL entre locos (Ardlie et al., 2002; Flint-Garcia et al., 2003; Rafalski and Morgante, 2004; Gupta et al., 2005; Slatkin, 2008). Mutação, deriva genética, estrutura populacional, fluxo gênico, endogamia e seleção atuam no surgimento de associações preferenciais entre locos; já a recombinação atua na sua redução ao longo das gerações, algo que é proporcional à distância genética entre os mesmos (Hartl and Clark, 2007; Hedrick, 2010). Assim, quanto menor for essa distância, menor será a chance de recombinação e, com ela, a quebra do DL entre locos (Pritchard and Przeworski, 2001). No contexto do mapeamento associativo, o

objetivo é detectar locos intimamente ligados e que estejam em DL, sugerindo uma mesma genealogia ao longo das gerações (Yu and Buckler, 2006; Zhu et al., 2008; Abdurakhmonov and Abdukarimov, 2008).

De maneira 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 genes de interesse (Oraguzie et al., 2007). Nesse sentido, 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 (Zhu et al., 2008; Rosa, 2011).

Apesar dessas informações serem relevantes para a realização de estudos de mapeamento associativo, é desejável obter uma visão do DL sob o ponto de vista genealógico (Nordborg and Tavaré, 2002; Zöllner and von Haeseler, 2000; McVean, 2002; McVean and Cardin, 2005; McVean, 2007a), considerando as possíveis relações evolutivas entre sequências de DNA. Nesse sentido, a teoria da coalescência, que realiza inferências do presente em direção ao passado (Hudson, 1990; Fu and Li, 1999; Wakeley, 2009), poder ser bastante útil para a caracterização do DL em painéis de associação, principalmente para espécies que possuem poucas informações de genealogias. Mais do que isso, pode ser consideravelmente útil em relação às medidas tradicionais do DL, tais como D' e r^2, que têm mostrado grande variação nos padrões de associação (Nordborg and Tavaré, 2002; Evans and Cardon, 2005).

2.2 Teoria da Coalescência

2.2.1 Histórico e Conceitos Importantes

A teoria da coalescência foi formalmente proposta por John Frank Charles Kingman (Isaac Newton Institute of Mathematical Sciences, Cambridge University, Inglaterra) no início da década de 80. Três trabalhos seminais foram publicados em 1982 com os princípios teóricos sobre esta abordagem (Kingman, 1982a,b,c), que passou a ter enorme representatividade para os estudos envolvendo genética de populações e evolução. Com os avanços obtidos na área da biologia molecular e computação, a teoria da coalescência passou a ser amplamente utilizada em trabalhos recentes de evolução (Kingman, 2000). Por ser uma abordagem sofisticada do ponto de vista matemático e estatístico, a teoria da coalescência tem fornecido resultados bastante precisos e importantes em estudos evolutivos. Nos dias de hoje, é considerada por vários cientistas como a moderna (Palczewski and Beerli, 2013) e mais popular (Marjoram and Tavaré, 2006) genética de populações, a qual revolucionou o entendimento sobre os processos histórico-evolutivos (Holsinger and Weir, 2009).

Em linhas gerais, o modelo de coalescência é um processo estocástico com o qual genealogias históricas e desconhecidas podem ser modeladas a partir de um conjunto de sequências de determinada população (Fu and Li, 1999; Rosenberg and Nordborg, 2002; Nordborg and Tavaré, 2002; Marjoram and Tavaré, 2006). Ao contrário dos métodos clássicos utilizados na genética de populações, o método da coalescência é baseado em inferências retroprospectivas, partindo-se do presente em direção ao passado (Fu and Li, 1999; Hein et al., 2004; Marjoram and Tavaré, 2006). Assim, é possível desvendar os processos genéticos do passado que relacionam sequências de gerações presentes, compreendendo a história evolutiva ocorrida ao longo das gerações (Hudson, 1990; Fu and

Li, 1999; Wakeley, 2009). Portanto, a teoria da coalescência sugere que os fenômenos podem ser explicados e compreendidos com base nos eventos que já ocorreram e não naqueles que ainda poderão ocorrer (Marjoram and Joyce, 2011).

Para melhor compreender os conceitos e as motivações da teoria da coalescência, considere um conjunto de n sequências de DNA correspondentes à determinada região do genoma (Figura 2.1). Essas sequências representam uma amostra aleatória de indivíduos coletados recentemente em determinada população. Inicialmente, a história evolutiva que explica os padrões de variação envolvendo quatro sítios polimórficos é totalmente desconhecida. Assim, é de interesse desvendar esta história, sendo de grande motivação o uso da teoria da coalescência.

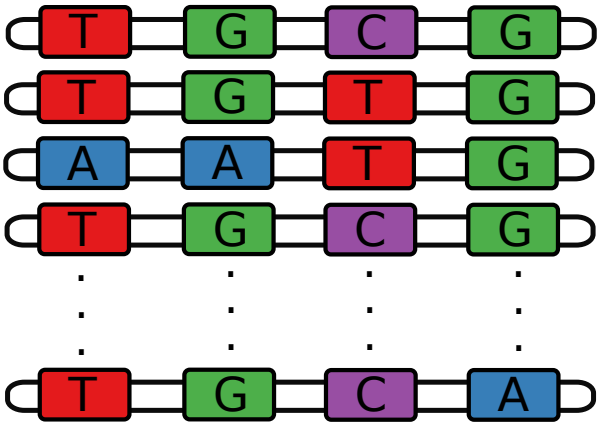

Figura 2.1: Amostra populacional contendo n sequências de DNA e quatro sítios polimórficos para determinada região do genoma

Para o modelo de coalescência proposto por Kingman (1982a,b,c), no caso de haver mutação, considera-se que os padrões de polimorfismo observados podem ser explicados pelo *modelo de sítios infinitos* (Karlin and McGregor, 1967; Kimura, 1969; Watterson, 1975), com o qual um número infinito de locos, sujeitos à mutação, pode ocorrer em posições contínuas no genoma. Neste caso, cada um dos quatro sítios polimórficos é resultado de um único evento de mutação ocorrido ao longo do processo evolutivo, gerando uma forma variante nunca antes observada (Griffiths and Tavaré, 1995; Wakeley, 2009). Trata-se de modelo apropriado para dados de sequências, cuja probabilidade de ocorrer mutações repetidas para uma mesma posição no genoma é negligenciável (Fearnhead and Donnelly, 2001).

A história evolutiva do segundo sítio polimórfico da amostra (Figura 2.1) é mostrada na Figura 2.2. Este sítio será utilizado para demonstração dos conceitos de coalescência, embora as inferências de genealogia aconteçam para todos os sítios simultaneamente, considerando os dados de sequências.

Ao retornar ao passado, é possível detectar ancestrais comuns, ou eventos *coalescentes*, envolvendo duas sequências quaisquer da amostra até alcançar o chamado *ancestral comum mais recente* (MRCA, do inglês *Most Recent Common Ancestor*), envolvendo todas as sequências da amostra (Fu and Li, 1999; Hein et al., 2004; Nordborg and Tavaré, 2002; Wakeley, 2009). Um evento coalescente corresponde à fusão de duas sequências quaisquer presentes no tempo t em uma única sequência qualquer presente no tempo $t-1$, que teve, numa visão prospectiva, seu DNA duplicado para trasmitir a informação genética para sequências de gerações atuais. Para uma amostra contendo n sequências de DNA, $n-1$ eventos de coalescência devem estar presentes caminhando-se do presente em direção ao passado, sendo que o último evento coalescente corresponderá sempre ao MRCA

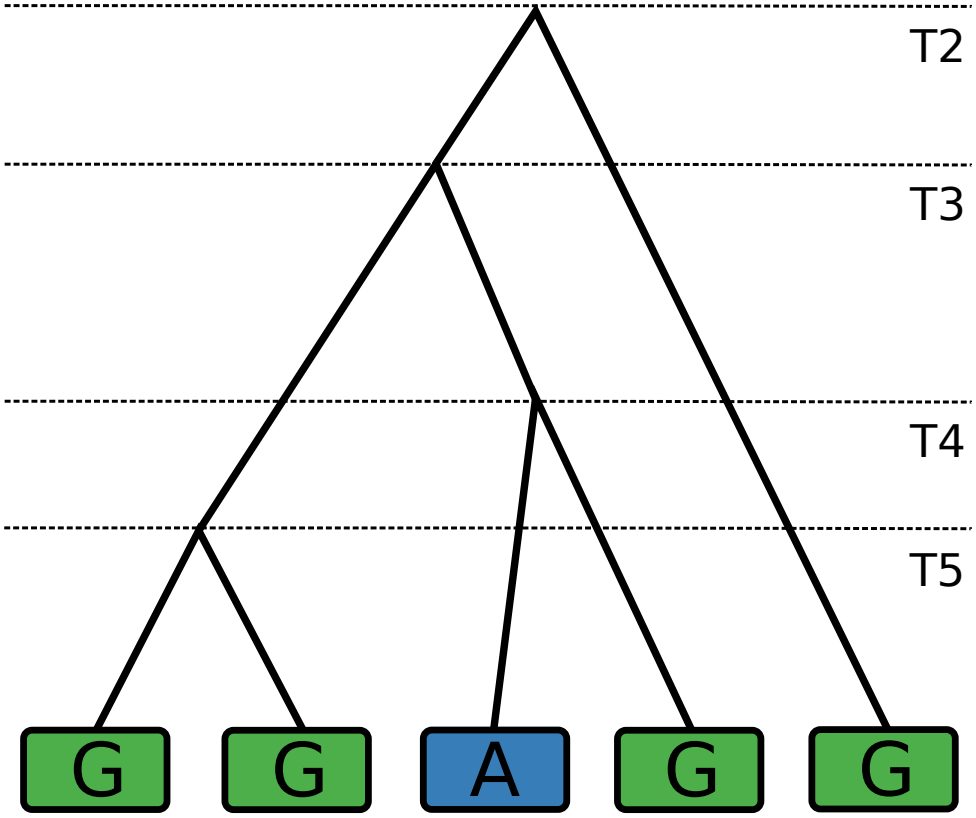

Figura 2.2: Possível história evolutiva do segundo sítio polimórfico (Figura 2.1) com base na teoria da coalescência. Cinco sequências de DNA estão representadas na geração atual, sendo que quatro delas possuem o nucleotídeo G (primeira, segunda, quarta e quinta) e uma única sequência possui o nucleotídeo A (terceira). Caminhando-se do presente em direção ao passado, o primeiro evento de coalescência acontece entre a primeira e a segunda sequências (à esquerda) que possuem o nucleotídeo G, no intervalo de tempo T5. O segundo evento de coalescência acontece entre a terceira (nucleotídeo A) e a quarta (nucleotídeo G) sequências no intervalo de tempo T4. O terceiro evento de coalescência acontece entre as sequências ancestrais obtidas a partir do primeiro e do segundo eventos de coalescência, no intervalo de tempo T3. E, por fim, o quarto e último evento de coalescência acontece entre a sequência ancestral das quatro primeiras sequências e a quinta sequência (nucleotídeo G), no intervalo de tempo T2. O último evento de coalescência corresponde ao ancestral comum mais recente da amostra (MRCA). Note que a terceira sequência é a única que possui o nucleotídeo A, mostrando que um evento de mutação deve ter ocorrido no ramo em que tal sequência se fundiu para formar o segundo evento de coalescência. O cálculo de uma probabilidade fornece o quanto que esta história evolutiva explica a variação dos dados da geração atual

(Nordborg, 2007; Wakeley, 2009). É importante notar que os eventos coalescentes correspondem aos eventos de divergência numa visão prospectiva da história evolutiva das populações (Fu and Li, 1999). No exemplo do segundo sítio polimórfico mostrado na Figura 2.2, quatro eventos coalescentes podem ser observados a partir de uma amostra contendo cinco sequências de DNA. Esses eventos estão situados nas posições exatas das linhas tracejadas mostradas horizontalmente no gráfico.

O primeiro evento coalescente ocorrerá entre duas das n sequências presentes na amostra. Uma vez que a genealogia histórica é totalmente desconhecida, o modelo de coalescência levará em consideração a possibilidade de todas as n sequências estarem envolvidas com o primeiro evento coalescente. Assim, considerando todos os possíveis pares envolvendo n sequências, um total de $\frac{n(n-1)}{2}$ combinações será possível para o primeiro evento coalescente, e o mesmo raciocínio pode ser utilizado para os demais eventos de coalescência. No modelo proposto, a probabilidade de duas sequências quaisquer coalescerem em alguma geração do passado é $\frac{1}{2N_e}$, sendo N_e o tamanho efetivo populacional. A probabilidade de duas sequências quaisquer não coalescerem em alguma geração

do passado é $1 - \frac{1}{2N_e}$. Dessa forma, a probabilidade de ocorrência do primeiro evento coalescente é $\frac{n(n-1)}{2}\frac{1}{2N_e}$, levando em consideração todas as possíveis combinações envolvendo pares de sequências. É importante notar que todos esses pares de sequências são *equiprováveis* neste modelo de coalescência, algo que nem sempre acontece em cenários mais complexos (HUDSON, 1990; WAKELEY, 2009).

Os eventos de coalescência são estatisticamente modelados seguindo um processo de Poisson (distribuição discreta), com taxa de 1 evento de coalescência ($\lambda = 1$) para cada par de sequências e intervalo de tempo (T_2,T_3,...,T_n). Os tempos de coalescência, que correspondem aos intervalos entre coalescentes sucessivos, são estatisticamente modelados seguindo uma distribuição exponencial (distribuição contínua) (HUDSON, 1990). Assim, a abordagem de coalescência é proposta de forma modificada em relação aos modelos que explicam uma população finita (HEIN ET AL., 2004; MARJORAM and TAVARÉ, 2006), na qual o tempo é modelado de forma discreta (em gerações) e não contínua. Neste caso, os coalescentes e os tempos de coalescência são modelados pelas distribuições binomial e geométrica, respectivamente (HUDSON, 1990; WAKELEY, 2009). Verifica-se que a inferência de certa genealogia a partir da teoria da coalescência baseia-se em diferentes modelos probabilísticos.

As diferentes formas de se obter eventos de coalescência determinam o número de genealogias (*NG*) ou configurações possíveis envolvendo sequências de gerações atuais. A equação 2.1 mostrada a seguir ilustra o processo de Poisson com o qual o número de genealogias possíveis pode ser modelado (WAKELEY, 2009):

$$NG = \prod_{i=2}^{n} \binom{i}{2} = \binom{2}{2} \times \binom{3}{2} \times \ldots \times \binom{n}{2} \tag{2.1}$$

em que o produtório ($\prod_{i=2}^{n}$) indica que as genealogias dos diferentes tempos de coalescência são modeladas independentemente (WAKELEY, 2009).

Como foi dito, cada tempo de coalescência pode ser explicado por uma distribuição exponencial. Um modelo para todos os tempos de coalescência, até o momento em que o MRCA da amostra é alcançado, segue, portanto, uma distribuição exponencial conjunta. Para este caso, a função densidade conjunta é obtida pelo produto das funções correspondentes aos intervalos de coalescência, assumindo que os mesmos sejam independentes (HUDSON, 1990; HEIN ET AL., 2004; WAKELEY, 2009). Portanto, a forma pela qual o $(n-1)$-ésimo evento coalescente ocorreu no tempo $t-1$ não influenciará na forma pela qual o $(n-2)$-ésimo evento ocorrerá no tempo $t-2$, e assim sucessivamente. As funções de densidade exponencial para cada tempo de coalescência e todos os tempos conjuntamente são apresentadas pelas equações 2.2 e 2.3, respectivamente (WAKELEY, 2009):

$$f_{T_i}(t_i) = \binom{i}{2} e^{-\binom{i}{2} t_i} \tag{2.2}$$

$$f_{T_2,\ldots,T_n}(t_2,\ldots,t_n) = \prod_{i=2}^{n} f_{T_i}(t_i) \tag{2.3}$$

$$= f_{T_2}(t_2) \times f_{T_3}(t_3) \times \ldots \times f_{T_n}(t_n) \tag{2.4}$$

$$= \binom{2}{2} e^{-\binom{2}{2} t_2} \times \binom{3}{2} e^{-\binom{3}{2} t_3} \times \ldots \times \binom{n}{2} e^{-\binom{n}{2} t_n} \tag{2.5}$$

em que T_i (t_i) é o tempo ou intervalo de coalescência quando i sequências estão presentes.

É possível obter a esperança ($E[T_i]$) e a variância ($Var[T_i]$) dos tempos de coalescência (equações 2.6 e 2.7, respectivamente):

$$E[T_i] = \frac{4N_e}{i(i-1)} \tag{2.6}$$

$$Var[T_i] = \left(\frac{4N_e}{i(i-1)}\right)^2 \tag{2.7}$$

Claramente, o tempo para ocorrência de um evento coalescente (T_i) é função direta do tamanho efetivo populacional e do tamanho da amostra ($i = 2,3,...,n$). Assim, os primeiros eventos coalescentes acontecerão, em média, muito rapidamente no passado, e os coalescentes mais antigos demorarão cada vez mais à medida que o MRCA se aproxima (WAKELEY, 2009). Para modelos de coalescência mais complexos, em que há alteração do tamanho efetivo populacional, outras situações podem ocorrer para a história evolutiva da amostra. Por exemplo, para uma população que sofreu expansão do seu tamanho efetivo ao longo das gerações, espera-se que os primeiros eventos coalescentes tenham ocorrido há muitas gerações no passado, e os coalescentes mais antigos tenham ocorrido em menor tempo em relação aos primeiros.

Duas medidas importantes do tamanho de uma genealogia são: (i) o tempo para alcançar o MRCA (T_{MRCA}, equação 2.8); e (ii) o comprimento total de todos os ramos que constituem a genealogia (T_{TOTAL}, equação 2.9):

$$T_{\mathrm{MRCA}} = \sum_{i=2}^{n} T_i \tag{2.8}$$

$$T_{\mathrm{TOTAL}} = \sum_{i=2}^{n} iT_i \tag{2.9}$$

Para ambas as medidas, é possível calcular a esperança (equações 2.10 e 2.11) e a variância (equações 2.12 e 2.13), considerando que diferentes genealogias podem ser formadas:

$$E[T_{\mathrm{MRCA}}] = \sum_{i=2}^{n} E[T_i] = \sum_{i=2}^{n} \frac{2}{i(i-1)} \tag{2.10}$$

$$E[T_{\mathrm{TOTAL}}] = \sum_{i=2}^{n} iE[T_i] = \sum_{i=2}^{n} i\frac{2}{i(i-1)} = 2\sum_{i=1}^{n-1} \frac{1}{i} \tag{2.11}$$

$$\mathrm{Var}[T_{\mathrm{MRCA}}] = 8\sum_{i=2}^{n} \frac{1}{i^2} - 4\left(1 - \frac{1}{n}\right)^2 \tag{2.12}$$

$$\mathrm{Var}[T_{\mathrm{TOTAL}}] = 4\sum_{i=1}^{n-1} \frac{1}{i^2} \tag{2.13}$$

KINGMAN (1982a,b,c) mostrou que uma população finita e muito grande (tendendo ao infinito), que segue o modelo proposto por FISHER (1930) e WRIGHT (1931), pode ser aproximadamente explicada pelo modelo de coalescência. Neste caso, tal modelo assume que as populações apresentam tamanho populacional finito e constante ao longo das gerações, bem como se reproduzem por meio de panmixia, que corresponde aos cruzamentos aleatórios que acontecem entre indivíduos de uma mesma geração (HEIN ET AL., 2004; WAKELEY, 2009). A coalescência também leva em consideração os princípios da teoria neutralista da evolução molecular (KIMURA, 1955; HUDSON,

1990; Fu and Li, 1999; Nordborg, 2007), com a qual alelos de determinado loco não apresentam diferenças em seus desempenhos e, portanto, são seletivamente neutros. Assim, apenas deriva genética é responsável pelas alterações alélicas ao longo das gerações (Hahn, 2008; Wakeley, 2009), de modo que seleção natural não seria força evolutiva preponderante nas populações (Hahn, 2008; Wakeley, 2009).

No entanto, os pressupostos de Fisher (1930) e Wright (1931) e da teoria neutralista nem sempre podem ser utilizados para explicar os padrões de variação. Certas populações podem apresentar tamanho efetivo populacional variável ao longo das gerações, subdivisão geográfica, recombinação, seleção e outros processos que alteram a estrutura das genealogias e causam desvios dos pressupostos iniciais. Assim, é comum o uso de testes estatísticos que podem detectar esses desvios e sugerir, pelo menos em princípio, cenários evolutivos mais complexos. São os chamados testes de neutralidade, que são baseados em estatísticas de diversidade que capitalizam a variação genética. Tais estatísticas serão apresentadas a seguir, e os testes de neutralidade mais conhecidos serão abordados na sequência.

2.2.2 Estatísticas de Diversidade

Historicamente, três importantes estatísticas foram propostas para explicar padrões da variação genética (Wakeley, 2009). A primeira delas é o número de sítios segregantes ou polimórficos (S) presentes em uma amostra de sequências (Watterson, 1975). A segunda é o número de diferenças nucleotídicas envolvendo pares de sequências (π) presentes nesta amostra (Tajima, 1983). E, por fim, a terceira é o número de sítios polimórficos (η_i) presentes na amostra de acordo com o número de cópias de suas bases nucleotídicas (i cópias de certa base e $n - i$ cópias da outra base), correspondendo à frequência nucleotídica dos sítios polimórficos (Wakeley, 2009). Para esta última, normalmente é impossível saber quais bases são mutantes (ξ_i) e quais são ancestrais (ξ_{n-i}) para os sítios, levando em consideração a ocorrência de um único evento de mutação para cada sítio (*modelo de sítios infinitos*) ao longo da história da amostra (Hein et al., 2004; Wakeley, 2009).

Embora S, π e η_i estejam diretamente relacionadas (as duas primeiras podem ser diretamente representadas a partir da terceira), cada qual sumariza a informação genética contida em uma amostra de diferentes formas. Assim, todas essas estatísticas têm sido historicamente utilizadas em diversos estudos de genética de populações, fornecendo evidências importantes acerca da variação. Além disso, pelo fato dos seus valores serem facilmente obtidos até mesmo para amostras grandes e com muitos sítios polimórficos, a motivação para o seu uso é ainda maior. Contudo, todas elas sumarizam a variação genética para um único valor, o que pode ser aceitável e desejável para certos casos, mas pode ser insuficiente quando o objetivo é compreender de que forma a variação genética foi estruturada diante de cenários evolutivos mais complexos (Kuhner, 2009; Wakeley, 2009).

2.2.3 Teoria Neutralista e Testes de Neutralidade

Os primeiros modelos desenvolvidos para explicar a variação genética presente em populações foram propostos por R. Fisher, S.Wright e J. Haldane em meados da década de 30 (Fisher, 1930; Wright, 1931). As ideias sugeridas por esses cientistas, com base nos modelos propostos, eram basicamente que as variações genéticas presentes nas populações aconteceriam em decorrência de eventos de mutação e seleção natural, sendo este último o fator que deveria, de fato, governar a evolução. Na época, esses cientistas revolucionaram pensamentos duvidosos e contrários a respeito do darwinismo e do mendelismo, mostrando que a seleção natural e a genética deveriam ser

encaradas conjuntamente (FUTUYMA, 2005).

No final dos anos 50, Motoo Kimura e colaboradores, com base em dados de proteínas, propuseram que grande parte da variação genética presente em populações finitas deveria ser neutra (KIMURA, 1955). Uma vez que mutações deletérias deveriam ser rapidamente eliminadas por seleção natural e mutações favoráveis mostrariam-se raras na população, esses cientistas concluíram que a grande maioria da variação não deveria exercer influências sobre o valor adaptativo dos indivíduos da população. Assim, alelos de locos neutros não mostrariam diferenças em seus desempenhos (HAHN, 2008) e, portanto, seriam seletivamente neutros, de modo que maiores ou menores proporções de um ou outro alelo não acarretariam maiores ou menores adaptações do ponto de vista evolutivo.

A teoria neutralista da evolução molecular, que passou a ser aplicada anos mais tarde, tornou-se um dos principais alicerces dos geneticistas e evolucionistas na tentativa de explicar a variação genética observada dentro e entre populações (HAHN, 2008). De maneira geral, o neutralismo fornece a base teórica para o entendimento da variação ao longo do genoma, possibilitando a elaboração de hipóteses testáveis e abordagens estatísticas que promovem a distinção entre seleção e deriva genética (KREITMAN, 2000; NIELSEN, 2001; HAHN, 2007).

Neste contexto, três diferentes testes estatísticos foram historicamente desenvolvidos para investigar desvios do modelo padrão baseado em neutralismo. O primeiro e mais conhecido teste de neutralidade foi proposto por TAJIMA (1989) no final da década de 80, sendo aqui designado como estatística D. Anos mais tarde, FU and LI (1993) propuseram dois outros testes de neutralidade, designados por D^* e F^*. Embora D, D^* e F^* apresentem suas particularidades metodológicas, todas essas estatísticas foram propostas a partir de uma mesma motivação, sendo baseadas nas estatísticas de diversidade S, π e η_i.

A estatística D pode ser representada pela seguinte equação:

$$D = \frac{\pi - \frac{S}{a_1}}{\sqrt{\widehat{\mathrm{Var}}[\pi - \frac{S}{a_1}]}} \tag{2.14}$$

em que $a_1 = \sum_{i=1}^{n-1} i$. Uma vez que a esperança das estatísticas π e S são iguais, respectivamente, a $E[\pi] = \theta$ e $E[S] = \theta a_1$, a diferença $\pi - \frac{S}{a_1}$ será igual a zero se ambas as estatísticas, cada qual com a sua forma de explicar a variação, forem iguais. Assim, no modelo padrão baseado em neutralismo, a estatística D é igual a zero, sugerindo que outros processos evolutivos, tais como estrutura populacional e seleção, podem não explicar os padrões de variação de certa amostra de sequências (WAKELEY, 2009). O parâmetro θ corresponde à taxa de mutação ao longo da história da amostra, que será abordada na seção 2.2.4 (Modelos Evolutivos) a seguir.

TAJIMA (1989) sugeriu que a estatística D poderia se aproximar de uma distribuição do tipo beta. Embora D não esteja distribuído exatamente de acordo com o modelo beta, valores de D acima de 2 e abaixo de -2 são estatisticamente significativos a 5% de nível de significância (WAKELEY, 2009). Isso significa que valores de D iguais a 2, -2 e entre ambos ($-2 \leq D \leq 2$) não são estatisticamente significativos e diferentes de zero a 5%, estando, portanto, de acordo com o modelo padrão baseado em neutralismo (HEIN ET AL., 2004).

Considerando todo o genoma, HEIN ET AL. (2004) argumentam que valores negativos de D (ramos externos longos) sugerem crescimento populacional ou severo *bottleneck*, enquanto que valores positivos (ramos internos longos) sugerem subdivisão geográfica ou leve *bottleneck*. Em contrapartida, considerando regiões específicas do genoma, valores negativos de D indicam seleção direcional e valores positivos indicam seleção balanceadora ou recente mistura populacional.

As estatísticas D^* e F^* (FU and LI, 1993) podem ser representadas pelas seguintes equações:

$$D^* = \frac{\frac{S}{a_1} - \frac{n-1}{n}\eta_1}{\sqrt{\widehat{\mathrm{Var}}[\frac{S}{a_1} - \frac{n-1}{n}\eta_1]}} \tag{2.15}$$

$$F^* = \frac{\frac{\pi}{a_1} - \frac{n-1}{n}\eta_1}{\sqrt{\widehat{\mathrm{Var}}[\frac{\pi}{a_1} - \frac{n-1}{n}\eta_1]}} \tag{2.16}$$

em que η_1 é o número de *singletons* ou sítios polimórficos nos quais uma de suas bases nucleotídicas possui apenas uma cópia na amostra de sequências. Ao invés de S e π estarem presentes em uma única medida de neutralidade, como é o caso da estatística D, ambas as estatísticas são consideradas separadamente na abordagem proposta por Fu and Li (1993). Dessa forma, os desvios do modelo padrão referentes às estatísticas D^* e F^* são, respectivamente, baseados nas diferenças entre S e π com quantidades proporcionais ao número de *singletons* presentes na amostra. Portanto, D^* e F^* podem fornecer valores diferentes e, de certo modo, complementares ao valor da estatística D em estudos de genética de populações (Wakeley, 2009).

No entanto, a estatística D parece ser mais poderosa para detectar desvios do modelo padrão baseado em neutralismo, devendo ser preferencialmente utilizada. Simonsen et al. (1995), ao utilizarem as estatísticas D, D^* e F^* para casos em que populações sofreram seleção, redução do tamanho populacional e subdivisão, verificaram que apenas a estatística D mostrou valores significativos e diferentes de zero. Contudo, esses valores foram observados somente quando processos evolutivos estiveram ocorrendo por muitas gerações, embora a detecção de tais processos tenha ocorrido considerando um passado relativamente recente. Portanto, é possível que a estatística D apresente valores significativos e diferentes de zero em situações mais drásticas de desvios do modelo padrão baseado em neutralismo.

Embora as estatísticas D, D^* e F^* apresentem comportamento diferentes, as respostas aos desvios do modelo padrão neutro são as mesmas: os valores se tornam negativos quando há excesso de polimorfismos de baixa e alta frequências e deficiência de polimorfismos de média frequência (Wakeley, 2009). A diferença é que apenas polimorfismos com frequências baixas e altas ao extremo contribuem para valores negativos das estatísticas D^* e F^*, enquanto que isso não acontece para D (Wakeley, 2009). Para esta última, a contribuição dos sítios polimórficos pode variar de positiva ou negativa simplesmente pelo tamanho da amostra.

Outro importante teste de neutralidade foi desenvolvido por Ramos-Onsins and Rozas (2002). Denominado de estatística R_2, este teste foi proposto para detectar desvios do modelo baseado em neutralismo especificamente causados por crescimento populacional. Os resultados mostraram que R_2 foi superior para detectar desvios em amostras pequenas e na presença de recombinação. Recentemente, Ramírez-Soriano et al. (2008), ao estudarem diferentes testes na presença de eventos demográficos e recombinação, concluíram que as estatísticas D e R_2 forneceram os resultados mais confiáveis, sendo, portanto, recomendadas.

Embora os testes de neutralidade possam sugerir desvios do modelo padrão baseado em neutralismo, evidências pouco conclusivas são normalmente fornecidas acerca do processo evolutivo (Fu and Li, 1999; Wakeley, 2009). Assim, modelos evolutivos específicos podem ser investigados com base na teoria da coalescência, utilizando-se abordagens mais precisas na exploração da variação genética e na estimação dos parâmetros de interesse. Veremos a seguir alguns desses modelos, que também serão investigados no presente trabalho.

2.2.4 Modelos Evolutivos

Como foi dito, vários modelos genético-populacionais podem ser assumidos e investigados com base na teoria da coalescência (Hudson, 1990; Pritchard and Przeworski, 2001; Marjoram and Tavaré, 2006; Nordborg, 2007). Esses modelos consideram forças evolutivas importantes, tais como mutação (constante e variável), fluxo gênico, recombinação e seleção (Fu and Li, 1999; Rosenberg and Nordborg, 2002). Em geral, esses modelos apresentam grande complexidade inferencial e demandam enorme tempo computacional, dificultando o uso da abordagem por parte dos pesquisadores. Contudo, essa complexidade dependerá da quantidade de dados disponíveis, da abordagem e do número de parâmetros a ser estimado, que, muitas vezes, é proporcional à complexidade do modelo evolutivo.

Iniciaremos a seguir com o modelo de mutação, que é considerado o mais simples. Alguns conceitos importantes serão inicialmente apresentados e, na sequência, o processo de inferência será mostrado. Depois disso, uma visão global dos modelos de fluxo gênico e recombinação será apresentada, visto que ambos serão aqui investigados. Tais investigações foram escolhidas com o propósito de fornecer informações relevantes para estudos futuros de mapeamento associativo em sorgo, que é a motivação do presente estudo.

2.2.4.1 Mutação

A extensão mais simples do modelo proposto por Kingman (1982a,b,c) é aquela que considera apenas mutação (Fu and Li, 1999; Nordborg and Tavaré, 2002; Wakeley, 2009). Para este caso, genealogias históricas são normalmente amostradas e eventos de mutação são simultaneamente incorporados por entre seus ramos, modelando-se genealogia e mutação de maneira independente (Hudson, 1990). Esta estratégia é totalmente válida pelo fato de que eventos de mutação são seletivamente neutros em populações de Fisher (1930) e Wright (1931), não interferindo na estrutura das genealogias (Wakeley, 2009). Se os estados alélicos influenciam diretamente a reprodução, não é possível distinguir entre sequências idênticas por descendência ou idênticas por estado (Nordborg, 2007).

Neste modelo, o objetivo é estimar o parâmetro designado por θ, que corresponde ao número esperado de mutações entre duas sequências quaisquer até o instante em que ambas se coalescem no passado (Hein et al., 2004). θ pode ser obtido a partir dos estimadores $\widehat{\theta} = 4N_e\mu$, para o caso de espécies diploides, ou $\widehat{\theta} = 2N_e\mu$, para o caso de espécies haploides (ou diploides quando os indivíduos são homozigotos para todos os locos).

Há métodos que estimam θ como sendo a taxa de mutação por sítio (Kuhner, 2006) e métodos que o estimam como sendo a taxa de mutação por região cromossômica, considerando os sítios polimórficos (Watterson, 1975). Neste contexto, é fundamental saber aquilo que de fato está sendo estimado, pois, no primeiro o caso, μ é a taxa de mutação por geração, indivíduo e para cada sítio, enquanto que, no segundo caso, μ é a taxa de mutação por geração, indivíduo e região cromossômica (Fu and Li, 1999; Nordborg, 2007; Nordborg and Tavaré, 2002; Hein et al., 2004; McVean and Cardin, 2005; Wakeley, 2009). De qualquer forma, normalmente o interesse está no parâmetro θ, pois é muito difícil distinguir N_e e μ a partir dos dados sem informações históricas prévias.

O parâmetro θ pode ser estimado com base em diferentes métodos. Dentre eles, tem-se: (i) estimadores clássicos baseados no método dos momentos (Watterson, 1975; Tajima, 1983; Felsenstein, 1992); (ii) estimadores baseados em máxima verossimilhança (Griffiths and Tavaré, 1994a,b,c, 1995; Kuhner et al., 1995; Fu, 1998); e (iii) estimadores baseados em inferência bayesiana (Beerli, 2006; Kuhner and Smith, 2007; Kuhner, 2009). O

estimador clássico proposto por WATTERSON (1975), comumente designado por θ_W, tem sido amplamente utilizado em estudos evolutivos até os dias de hoje. No entanto, apesar de θ_W ser facilmente obtido a partir de um conjunto de sequências, trata-se de um estimador que capitaliza apenas parte da variação dos dados. Isso é reflexo do método dos momentos, que, indiscutivelmente, possui propriedades inferiores quando comparado ao método da máxima verossimilhança (MYUNG, 2003). Assim, é desejável obter estimativas mais precisas e confiáveis para θ, e os estimadores baseados em máxima verossimilhança e inferência bayesiana podem trazer resultados bastante satisfatórios.

As principais abordagens de máxima verossimilhança para estimação de θ foram desenvolvidas por dois grupos: (i) Robert Griffiths e colaboradores (GRIFFITHS and TAVARÉ, 1994a,b,c, 1995); e (ii) Mary Kuhner e colaboradores (KUHNER ET AL., 1995; FELSENSTEIN ET AL., 1999). Na primeira abordagem, a estimação de θ é baseada na amostragem de importância (IS, do inglês *Importance Sampling*), com o uso do procedimento de Monte Carlo para simulação de genealogias por meio da integração computacional. Na segunda abordagem, a estimação de θ também é baseada na amostragem de importância, mas com o uso do procedimento de Monte Carlo via Cadeias de Markov (MCMC, do inglês *Markov Chain Monte Carlo*) para simulação de genealogias por meio da integração computacional. Segundo WAKELEY (2009) e HEY and NIELSEN (2007), Kuhner e colaboradores foram os precursores no uso do MCMC para simulação de genealogias e estimação do parâmetro θ no contexto de coalescência, mostrando que o uso de verossimilhança é possível para investigar modelos de mutação complexos.

Ambas as abordagens utilizam princípios semelhantes para a realização das inferências. Pelo fato dos conceitos coalescentes serem baseados em cadeias de Markov, a primeira abordagem também pode ser considerada como uma forma de MCMC. Contudo, talvez a grande diferença esteja na forma como a simulação é sequencialmente realizada a partir de uma distribuição de probabilidades. Enquanto a primeira abordagem utiliza Monte Carlo para simular genealogias *independentes*, a segunda utiliza Monte Carlo para simular genealogias *correlacionadas* entre si (FELSENSTEIN ET AL., 1999; WAKELEY, 2009). Assim, espera-se que a segunda forneça resultados mais precisos, visto que há tendência de que genealogias altamente prováveis sejam continuamente simuladas para explicar a variação dos dados.

Para compreender o processo de inferência da segunda abordagem (KUHNER ET AL., 1995; FELSENSTEIN ET AL., 1999), considere novamente a amostra de n sequências da Figura 2.1. É possível modelar o passado evolutivo dessas sequências com base na teoria da coalescência, utilizando-se os modelos das equações 2.1 e 2.2. Levando em consideração o tamanho efetivo populacional, é possível reescrever a equação 2.3 de forma a representar explicitamente N_e (FELSENSTEIN ET AL., 1999):

$$P(G' \mid N_e) = \prod_{i=2}^{n} \frac{2}{4N_e} \exp\left(\frac{-i(i-1)}{4N_e} u_i \right) \tag{2.17}$$

sendo G' o conjunto de genealogias possíveis na história, i o número de sequências ou linhagens do intervalo de coalescência, e u_i o comprimento do intervalo contendo i sequências.

A equação 2.17 corresponde à distribuição a priori do modelo de coalescência proposto inicialmente por KINGMAN (1982a,b,c). A partir deste e de um modelo evolutivo que considera mutações ao longo da história (FELSENSTEIN, 1981), é possível obter a distribuição de probabilidade dos dados D dado a genealogia, representada por $P(D \mid G', \mu)$. Combinando ambas as probabilidades, $P(G' \mid N_e)$ e $P(D \mid G', \mu)$, é possível obter a distribuição a posteriori dos dados representada pela equação 2.18 (FELSENSTEIN ET AL., 1999):

$$P(D \mid N_e,\mu) = \int_{G'} P(G' \mid N_e) P(D \mid G',\mu) \tag{2.18}$$

com a qual deve-se integrar para todas as genealogias possíveis (FELSENSTEIN, 1988).

As probabilidades $P(G' \mid N_e)$ e $P(D \mid G',\mu)$ podem ser facilmente calculadas. Em princípio, a probabilidade $P(D \mid N_e,\mu)$ também pode ser facilmente calculada pelo produto entre $P(G' \mid N_e)$ e $P(D \mid G',\mu)$, considerando coalescência e mutação como histórias independentes. No entanto, é complicado, ou totalmente inviável, obter $P(D \mid N_e,\mu)$ porque, normalmente, enorme quantidade de genealogias pode ser obtida para a história da amostra, visto que o passado evolutivo é totalmente desconhecido. Esta quantidade é representada pela integral $\int_{G'}$ na equação 2.18, sendo computacionalmente inviável integrar para todas as possíveis histórias com o objetivo de encontrar aquela que melhor explica a variação dos dados. É exatamente por conta disso que a segunda abordagem foi proposta no contexto do MCMC, para integrar sobre genealogias importantes e *correlacionadas*, o que é computacionalmente viável.

No entanto, a integração proposta inicialmente por KUHNER ET AL. (1995) não é realizada diretamente a partir da probabilidade $P(D \mid N_e,\mu)$. Como já foi dito, é muito difícil distinguir N_e e μ a partir dos dados, de modo que o foco permanece na combinação de ambos, ou seja, no parâmetro θ (KUHNER ET AL., 1995). Assim, alterando a escala de tempo para considerar θ, é possível reescrever a distribuição a priori do modelo de coalescência (equação 2.17) da seguinte forma (FELSENSTEIN ET AL., 1999):

$$P(G \mid \theta) = \prod_{i=2}^{n} \frac{2}{\theta} \exp\left(\frac{-i(i-1)}{\theta} \upsilon_i \right) \tag{2.19}$$

sendo $G = \mu G'$ e $\upsilon_i = \mu u_i$. G corresponde ao conjunto das genealogias G' com escala modificada. υ_i corresponde ao comprimento do intervalo u_i com i sequências também com escala modificada (FELSENSTEIN ET AL., 1999).

Da mesma forma, é possível escrever a distribuição a posteriori (equação 2.18) em função de θ pela seguinte equação (FELSENSTEIN ET AL., 1999; STEPHENS, 2007):

$$P(D \mid \theta) = \int_{G} P(G \mid \theta) P(D \mid G,\theta) \tag{2.20}$$

sendo $P(D \mid G,\theta)$ a distribuição dos dados a partir dos modelos de coalescência e mutação com escala modificada. Note que o objetivo é estimar o parâmetro θ a partir dos dados, sendo que a verossimilhança, $L(\theta)$, é proporcional à probabilidade $P(D \mid \theta)$.

A simulação de genealogias importantes proposta por KUHNER ET AL. (1995) é realizada a partir da distribuição escalada $P(D \mid \theta)$, e as razões para o uso de tal estratégia são claramente descritas em FELSENSTEIN ET AL. (1999). No entanto, para melhor direcionar o processo em função da grande quantidade de histórias ($\int_G$), KUHNER ET AL. (1995) propuseram a atribuição de um valor inicial para θ. A partir deste valor inicial (θ_0), o processo de simulação via MCMC é realizado com base na distribuição *corrigida* de $P(D \mid \theta)$. Isso significa que o MCMC visitará um espaço menor de histórias possíveis para as sequências dos dados, condicionado ao valor de θ_0 inicialmente atribuído e, portanto, à distribuição de $P(G \mid \theta_0)$. Claramente, o valor de θ mais provável estará condicionado ao valor de θ_0 e poderá ser insatisfatório se este não estiver próximo ao seu valor real (KUHNER ET AL., 1995, 1998; KUHNER, 2006, 2009; WAKELEY, 2009). Então, outros valores iniciais para θ_0, bem como repetições e maior quantidade de iterações poderão ser importantes para a obtenção de resultados mais precisos.

Com base no conjunto de genealogias simuladas via MCMC, outros possíveis valores de θ podem ser investigados. É possível calcular a razão entre as probabilidades de se obter cada genealogia simulada dado o valor investigado de θ e o valor inicialmente atribuído para θ_0, e obter uma média a partir da soma de todas as razões das genealogias para obtenção de uma superfície de razão de verossimilhanças. Assim, tem-se que:

$$\frac{L(\theta)}{L(\theta_0)} \approx \frac{1}{r}\sum_{i=1}^{r}\frac{P(D \mid G_i)P(G_i \mid \theta)}{P(D \mid G_i)P(G_i \mid \theta_0)} \tag{2.21}$$

$$\approx \frac{1}{r}\sum_{i=1}^{r}\frac{P(G_i \mid \theta)}{P(G_i \mid \theta_0)} \tag{2.22}$$

em que r é o número de genealogias simuladas via MCMC. Para cada valor de θ investigado, certo valor de $\frac{L(\theta)}{L(\theta_0)}$ será obtido, de modo que, ao final, um gráfico de razão de verossimilhanças e diferentes valores θ poderá ser construído (KUHNER ET AL., 1995).

Embora a distribuição a posteriori (equação 2.20) apresente a complexidade do modelo de coalescência com mutação, pode ser difícil, apenas analisando a mesma, compreender de que forma tal complexidade está estruturada no referido modelo. Com base nisso, WAKELEY (2009), ao invés de representar a quantidade enorme de genealogias possíveis por $\int_G$, desmembrou esta integração de modo a mostrar que algumas variáveis estão envolvidas com um processo discreto e outras com um processo contínuo no modelo de coalescência:

$$\begin{aligned}
L(\theta) &= \int_G P(G \mid \theta)P(D \mid G,\theta) \\
&= \sum_{\Lambda}\int_T P(D \mid \Lambda, \mathrm{T}, \theta)P(\Lambda)f(\mathrm{T})d(T) \\
&= \sum_{\Lambda}\sum_{U\in\Omega_\Lambda}\int_T P(D \mid U, \Lambda)P(U \mid \Lambda, \mathrm{T}, \theta)P(\Lambda)f(\mathrm{T})d(T) \\
&= \sum_{\Lambda}\sum_{U\in\Omega_\Lambda}\int_T P(D \mid U, \Lambda)P(U, \Lambda \mid \theta)
\end{aligned}$$

sendo G correspondente às possíveis genealogias envolvendo as sequências da amostra. G leva em consideração todas as possíveis configurações Λ de genealogias, todos os eventos de mutação U ($U \in \Omega_\Lambda$) na história da amostra, e todos os possíveis tempos de coalescência T ($T_2, T_3, \ldots, T_n$), proporcionais aos eventos de coalescência. Dessa forma, G, que pode ser designada conjuntamente por $G = (\Lambda, U, \mathrm{T})$, é o resultado de diferentes processos de Poisson (Λ e U) e distribuições exponenciais independentes (T) caminhando-se do presente em direção ao passado (WAKELEY, 2009; KUHNER, 2009).

2.2.4.2 Estrutura Populacional e Fluxo Gênico

Embora grande parte dos estudos envolvendo populações estruturadas se refira ao fluxo gênico como sendo a taxa de migração, acredita-se que a primeira designação é mais apropriada. Isso porque eventos de migração podem ser meramente geográficos, de forma que a contribuição genética e evolutiva dos migrantes não seja necessariamente garantida. Portanto, neste trabalho, o termo fluxo gênico será sempre utilizado para se referir ao presente modelo.

Tradicionalmente, o fluxo gênico entre populações tem sido estimado com base na estatística F_{ST} (NIELSEN and WAKELEY, 2001; PEARSE and CRANDALL, 2004) ou suas extensões (NATH and GRIFFITHS, 1996). Embora

avanços tenham sido alcançados no entendimento das relações entre populações com o uso desta medida, trata-se de uma estatística limitada (Kuhner, 2009; Sousa and Hey, 2013). Isso é justificável basicamente por quatro motivos: (i) um mesmo valor da estatística F_{ST} pode ser obtido para cenários bastante diferentes do ponto de vista evolutivo (Sousa and Hey, 2013); (ii) um único valor de F_{ST} é estimado considerando que todas as populações possuem o mesmo tamanho efetivo e que há um número infinitamente grande de populações; (iii) o fluxo gênico entre diferentes populações é inferido de forma simétrica, ou seja, tendo-se sempre o mesmo valor (Beerli and Felsenstein, 2001); e, por fim, (iv) a estatística F_{ST} é baseada no método dos momentos, que claramente não capitaliza toda a variação. Neste contexto, acredita-se que os métodos de fluxo gênico baseados em coalescência podem ser promissores para um maior entendimento das relações evolutivas envolvendo populações.

O modelo de coalescência proposto por Kingman (1982a,b,c) pode ser estendido para considerar fluxo gênico entre populações (Hudson, 1990; Beerli and Felsenstein, 1999, 2001). Caminhando-se do presente em direção ao passado, além dos eventos de coalescência que deverão ser inferidos para cada população, eventos de migração com fluxo gênico também poderão ser modelados entre essas populações (Beerli and Felsenstein, 2001). Assim, modelos de coalescência com fluxo gênico tendem a ser mais complexos que modelos de coalescência padrões, que apenas consideram mutação.

Em um modelo de coalescência com fluxo gênico, o interesse está em estimar os parâmetros $\mathcal{M}$ ou M. $\mathcal{M}$ corresponde à razão $\frac{m}{\mu}$, sendo m a taxa de fluxo gênico por geração e μ a taxa de mutação por geração e sítio. Em contrapartida, M pode corresponder à $4N_em$ (diploides) ou $2N_em$ (haploides ou diploides homozigotos). O produto de $\mathcal{M}$ pela taxa de mutação θ, que é conjuntamente estimada neste modelo, resulta no valor de M (Beerli, 1998; Beerli and Felsenstein, 1999, 2001). Então, $\mathcal{M}$ e M estão relacionados e correspondem às taxas de fluxo gênico ou ao número mínimo de eventos migratórios que acontecem entre populações de forma a resultar em contribuição genética. A diferença é que $\mathcal{M}$ depende diretamente da mutação, enquanto que M é independente deste processo (Beerli, 1998).

Diversas abordagens têm sido propostas para investigar o fluxo gênico entre populações via modelos de coalescência (Takahata, 1988; Takahata and Slatkin, 1990; Notohara, 1990; Nath and Griffiths, 1993, 1996; Bahlo and Griffiths, 2000; Beerli and Felsenstein, 1999, 2001; Nielsen and Wakeley, 2001; Wilson and Rannala, 2003; Hey and Nielsen, 2004, 2007; Hey, 2010). Tais abordagens forneceram avanços importantes no entendimento dos conceitos de coalescência envolvendo populações estruturadas. Contudo, apenas algumas delas parecem ser muito utilizadas até os dias de hoje, o que é de se esperar devido ao desenvolvimento de modelos mais realistas para o entendimento dos processos evolutivos. Duas dessas abordagens mais utilizadas foram propostas por Nielsen and Wakeley (2001) e Beerli and Felsenstein (2001), que serão apresentadas a seguir.

Nielsen and Wakeley (2001) desenvolveram métodos de máxima verossimilhança e inferência bayesiana para estimar seis parâmetros populacionais, a saber: (i) tempo de divergência entre certa população ancestral e duas populações recentes geradas a partir da primeira; (ii) taxa de mutação para cada uma das três populações consideradas; e (iii) taxas de fluxos assimétricas entre as duas populações recentes. Os autores propuseram o uso do MCMC para simular genealogias com elevadas probabilidades e *correlacionadas* entre si. No entanto, tal abordagem foi proposta apenas para estimar o fluxo gênico entre duas populações e sem considerar o uso simultâneo de múltiplos locos, o que pode ser pouco real para muitos casos.

Em contrapartida, Beerli and Felsenstein (2001) utilizaram os mesmos métodos de inferência para

estimar taxas de mutação e fluxo gênico envolvendo *n* populações, considerando um único loco ou múltiplos locos simultanea-mente. Com a presença de *n* populações em uma amostra, um total de n^2 parâmetros será estimado, correspondentes às *n* taxas de mutação para cada população e $n(n-1)$ taxas de fluxo gênico entre populações. Assim, o modelo leva em consideração a existência de diferentes taxas de mutação para as populações e taxas de fluxo gênico assimétricas entre as mesmas, o que pode ser um cenário real. Este método apresenta inferência similar com a abordagem proposta por Bahlo and Griffiths (2000).

Atualmente, cenários mais complexos têm sido considerados para a estimação de fluxo gênico entre populações. Beerli and Palczewski (2010) propuseram a investigação de diferentes modelos com fluxo gênico e a comparação dos mesmos com base em suas verossimilhanças, de forma a verificar aquele que melhor explica o passado evolutivo envolvendo populações. Palczewski and Beerli (2013) desenvolveram método aproximado para estimação de altas taxas de fluxo gênico em situações em que populações trocaram muitos alelos entre si. Além disso, taxas de fluxo gênico têm sido investigadas em modelos que consideram, pelo menos em princípio, outros processos, tais como recombinação (Becquet and Przeworski, 2007; Lohse et al., 2011; Mailund et al., 2012) e seleção (Sousa et al., 2013).

2.2.4.3 Recombinação

Outra força evolutiva importante que pode ser modelada via teoria da coalescência é a recombinação (Hudson, 1983; Griffiths and Marjoram, 1996). Ao contrário da mutação, a recombinação normalmente não deixa vestígios de ocorrência a partir de uma visualização inicial de um conjunto de sequências amostradas (Fu and Li, 1999; Nordborg and Tavaré, 2002). Assim, modelos de coalescência com recombinação tendem a ser mais complexos, algo que se torna ainda mais desafiador e intensivo pelo fato de que nenhuma configuração de genealogia pode ser trivialmente modelada diante deste cenário (Hein et al., 2004). Isso acontece porque um evento de recombinação é modelado de forma que uma única sequência mais recente é "quebrada" em posição aleatória para formar duas sequências mais antigas (Hein et al., 2004; McVean and Cardin, 2005; Nordborg, 2007), sendo o contrário de um evento de coalescência. Então, o processo evolutivo deste modelo é melhor representado por um gráfico e não por uma genealogia, tendo-se, portanto, tanto eventos de coalescência quanto eventos de recombinação (Hein et al., 2004; Nordborg, 2007).

O primeiro modelo de coalescência com recombinação foi proposto por Hudson (1983). Por contemplar eventos de coalescência e recombinação modelados aleatoriamente no passado, Griffiths and Marjoram (1996) propuseram a nomenclatura "gráfico de recombinação ancestral" (ARG, do inglês *Ancestral Recombination Graph*) anos mais tarde. Como alternativa ao modelo de Hudson (1983), Wiuf and Hein (1999) desenvolveram uma forma de simular ARGs percorrendo as sequências amostradas de um lado para outro, caracterizando uma complexa estrutura não-Markoviana (McVean and Cardin, 2005). Embora este método tenha tido sua importância, a primeira abordagem é considerada mais simples e útil (Hein et al., 2004).

Em um modelo de coalescência com recombinação, o interesse está em estimar os parâmetros C ou ρ. C corresponde à razão $\frac{r}{\mu}$, sendo r a taxa de recombinação por geração e para cada intersítio (Kuhner et al., 1995). ρ pode ser igual a $4N_e r$ (diploides) ou $2N_e r$ (haploides ou diploides homozigotos). C e ρ correspondem às taxas de recombinação populacional (Kuhner et al., 1995), sendo o número esperado de recombinações entre duas sequências quaisquer até o instante em que ocorre um evento de coalescência. É importante dizer que a taxa de

mutação θ também é estimada no presente modelo, com eventos de mutação sendo incorporados nas genealogias após a geração dos ARGs (Hudson, 1990; McVean and Cardin, 2005; Nordborg, 2007).

A inferência de recombinação via coalescência é claramente difícil e desafiadora (Wall, 2000; Stumpf and McVean, 2003; Nordborg, 2007). Estimadores baseados em momentos foram propostos para obtenção de ρ a partir das estimativas de θ (Hudson, 1987; Wakeley, 1997). Esses estimadores apenas utilizam parte da variação disponível, apresentando viés e elevada variância para as estimativas. Em contrapartida, estimadores baseados em máxima verossimilhança e inferência bayesiana utilizam toda a informação disponível nos dados, embora apresentem restrições por não possuírem nenhuma expressão analítica ou numérica para o cálculo de suas verossimilhanças (McVean and Cardin, 2005). Certamente, isso dificulta o entendimento de suas propriedades (Fu and Li, 1999) e a construção de métodos de simulação eficientes para a estimação de tal processo (McVean and Cardin, 2005).

Normalmente, a quantidade de ARGs que podem ser simulados a partir de certa distribuição é enorme. A simulação sobre todos eles, muitos dos quais possuem probabilidades irrisórias para as verossimilhanças (McVean and Cardin, 2005), é ingênua e intratável do ponto de vista computacional. Com base nisso, Kuhner et al. (2000) e Fearnhead and Donnelly (2001) desenvolveram métodos de Monte Carlo mais eficientes para a estimação de recombinação. Tais métodos estimam um único valor de recombinação para certa região genômica, com base em abordagem completa. Contudo, ambos podem ser ineficientes para grandes quantidades de dados, algo que se tornou uma realidade no final da década de 90.

Como alternativa, vários métodos de coalescência foram desenvolvidos para estimar recombinação a partir de muitos dados (Wall, 2000; Hudson, 2001; McVean et al., 2002; Fearnhead and Donnelly, 2002; Li and Stephens, 2003; McVean et al., 2004; McVean and Cardin, 2005; Auton and McVean, 2007). Em essência, todos esses métodos baseiam-se em alguma estratégia de aproximação na tentativa de reduzir a complexidade da abordagem completa. O mais recente deles (Auton and McVean, 2007) foi proposto para considerar taxas de recombinação variáveis e pontos quentes de recombinação (*hotspots*) entre pares de locos. Tal método é baseado em um tipo específico de MCMC no contexto bayesiano (rjMCMC, do inglês *reversible jump* MCMC) (Green, 1995), com o qual ARGs de elevadas probabilidades são simulados a partir de buscas em várias dimensões de probabilidade, aumentando as chances de se obter estimativas mais precisas de recombinação. Acredita-se que a abordagem proposta por Auton and McVean (2007) é bastante promissora, principalmente para grandes quantidades de dados.

Um aspecto interessante é que eventos de recombinação podem ser gerados tanto a partir de trocas entre cromátides homólogas (*crossing-overs*) quanto conversão de genes. Todos os métodos descritos anteriormente são baseados no primeiro caso, o que pode ser pouco realista do ponto de vista biológico. Frisse et al. (2001) e Ptak et al. (2004) desenvolveram abordagem para estimar recombinação baseada tanto em *crossing-overs* quanto conversão de genes. De forma mais ampla, Wall (2004) propôs a estimação de recombinação tanto com base em *crossing-overs*, para comparar com abordagens anteriores, quanto com base em ambos, *crossing-overs* e conversão gênica.

2.2.5 Máxima Verossimilhança e Inferência Bayesiana

Trabalhos recentes apresentam comparações entre os métodos de máxima verossimilhança e inferência bayesiana na estimação de parâmetros via coalescência. Beerli (2006), ao utilizar o método proposto por Beerli and Felsenstein (2001) e sua versão bayesiana desenvolvida posteriormente, detectou diferenças importantes na es-

timação de fluxo gênico entre populações. Em contrapartida, Kuhner and Smith (2007) não detectaram diferenças substanciais entre tais métodos na estimação de crescimento populacional e recombinação.

Beerli (2006) menciona que a abordagem de máxima verossimilhança apresenta dificuldades para cenários complexos, com os quais grande quantidade de parâmetros é normalmente estimada. Considera que tais dificuldades também podem ocorrer quando altas taxas de fluxo gênico e menor variabilidade são esperadas entre populações. Isso porque o processo de simulação via MCMC não se comporta de forma apropriada em função da grande quantidade de eventos, dificultando a busca e a convergência para estimativas que maximizam a verossimilhança. Essas dificuldades devem ter estimulado o desenvolvimento de abordagem aproximada em trabalho recente (Palczewski and Beerli, 2013), muito embora a inferência bayesiana possa fornecer resultados confiáveis para situações complexas (Beerli, 2006).

2.3 Desequilíbrio de Ligação e Teoria da Coalescência

A associação não-aleatória, ou preferencial, entre alelos de diferentes locos em uma população fornece um profundo entendimento sobre a sua história evolutiva (McVean, 2002). Apesar dos estudos relacionados ao DL serem de longa data na genética de populações, sua relação com a teoria da coalescência apenas foi claramente apresentada por trabalhos recentes (Zöllner and von Haeseler, 2000; Pritchard and Przeworski, 2001; Nordborg and Tavaré, 2002; McVean, 2002; McVean and Cardin, 2005; McVean, 2007a,b).

Zöllner and von Haeseler (2000) utilizaram os conceitos da teoria da coalescência para estudar o DL em populações de tamanho constante e com crescimento exponencial, a partir de dados simulados de SNPs. Pritchard and Przeworski (2001) também utilizaram dados simulados para investigar DL considerando diferentes cenários demográficos via coalescência, tais como tamanho populacional constante, crescimento exponencial e estrutura de população. Pritchard and Przeworski (2001) e Nordborg and Tavaré (2002) apresentaram as perspectivas do DL via coalescência para estudos de mapeamento associativo, em dois artigos que esclareceram a importância de uma visão genealógica dos padrões de associação para tal objetivo.

Apesar de esses artigos terem mostrado ideias e resultados interessantes, a relação formal do DL com a teoria da coalescência foi recentemente apresentada por McVean (2002). Utilizando-se a tradicional medida r^2 do DL (Hill and Robertson, 1968), este autor sugeriu uma interpretação genealógica dos padrões de associação a partir do cálculo do seu valor esperado, que, embora não possua expressão analítica possível, pode ser aproximado por uma razão de esperanças, como mostrado na equação 2.23 (Ohta and Kimura, 1969a,b, 1971):

$$E[r^2_{AB}] \approx \frac{E[D^2_{AB}]}{E[f_A(1-f_A)f_B(1-f_B)]} \tag{2.23}$$

sendo A e B dois locos bi-alélicos fisicamente ligados em determinado cromossomo.

Desenvolvendo o numerador da equação 2.23, é possível mostrar que:

$$
\begin{aligned}
E[D_{AB}^2] &= E[(f_{AB} - f_A f_B)(f_{AB} - f_A f_B)] \\
&= E[f_{AB}^2 - 2f_{AB}f_A f_B + f_A f_B f_A f_B] \\
&= E[f_{AB}^2] - 2E[f_{AB}f_A f_B] + E[f_A f_B f_A f_B] \\
&= F_{A(ij)B(ij)} - 2F_{A(ij)B(ik)} + F_{A(ij)B(kl)} \\
&\vdots \\
\mathrm{Var}[D_{AB}] &= \frac{\mathrm{Cov}[t_{A(ij)},t_{B(ij)}] - 2\mathrm{Cov}[t_{A(ij)},t_{B(ik)}] + \mathrm{Cov}[t_{A(ij)},t_{B(kl)}]}{E[T_A T_B]}
\end{aligned}
$$

em que $F_{A(ij)B(ij)}$, $F_{A(ij)B(ik)}$ e $F_{A(ij)B(kl)}$ são os coeficientes de identidade para dados de sequências. Esses coeficientes correspondem às probabilidades de duas sequências (i) i e j serem idênticas para os locos A e B; (ii) i e j serem idênticas para o loco A, e i e k serem idênticas para o loco B; e (iii) i e j serem idênticas para o loco A, e k e l serem idênticas para o loco B. Além disso, esses coeficientes podem ser expressos na forma de covariância dos tempos de coalescência envolvendo os locos A e B ($\mathrm{Cov}[t_A,t_B]$), conforme descrito na última expressão da equação. T_A e T_B correspondem aos tempos dos locos A e B para alcançar o MRCA de toda a amostra.

Da mesma forma, desenvolvendo o denominador da equação 2.23, tem-se que:

$$
E[f_A(1-f_A)f_B(1-f_B)] = \frac{E[t]^2 + \mathrm{Cov}[t_{A(ij)},t_{B(kl)}]}{E[T_A T_B]}
$$

em que $E[t]$ é o tempo de coalescência esperado para um par de cromossomos.

Assim, reescrevendo a equação 2.23 a partir das derivações do seu numerador e denominador, é possível mostrar que:

$$
E[r_{AB}^2] \approx \sigma_d^2 = \frac{\mathrm{Cov}[t_{A(ij)},t_{B(ij)}] - 2\mathrm{Cov}[t_{A(ij)},t_{B(ik)}] + \mathrm{Cov}[t_{A(ij)},t_{B(kl)}]}{E[t]^2 + \mathrm{Cov}[t_{A(ij)},t_{B(kl)}]} \tag{2.24}
$$

Para uma amostra de tamanho finito, algumas modificações devem ser incorporadas de forma a considerar que as sequências i, j, k e l podem não ser totalmente distintas. Dessa forma, $E[r_{AB}^2]$, ou simplesmente σ_d^2 (Ohta and Kimura, 1969a), é mostrado de forma modificada pela equação 2.25 (Hudson, 1985; McVean, 2002):

$$
\sigma_d^2 = \frac{[n^2 - 2(n-1)]C_{ij,ij} - 2(n-2)^2 C_{ij,ik} + (n-2)(n-3)C_{ij,kl} + nE[t]^2}{n(n-1)E[t]^2 + 2C_{ij,ij} + 4(n-2)C_{ij,ik} + (n-2)(n-3)C_{ij,kl}} \tag{2.25}
$$

em que os $C's$ correspondem às abreviações das covariâncias dos tempos de coalescência entre os locos A e B. É possível notar o tamanho da amostra (n) nos termos da equação 2.25.

McVean (2002) mostrou que as covariâncias dos tempos de coalescência são funções diretas de processos evolutivos importantes, tais como recombinação e fluxo gênico. No caso da recombinação, essas ideias

já haviam sido propostas por trabalhos anteriores de coalescência (Griffiths, 1981; Hudson, 1983; Kaplan and Hudson, 1985), mostrando a motivação de se estudar tal processo do ponto de vista genealógico (equações 2.26, 2.27 e 2.28).

$$\mathrm{Cov}[t_{A(ij)},t_{B(ij)}] = \frac{18+\rho}{18+13\rho+\rho^2} \tag{2.26}$$

$$\mathrm{Cov}[t_{A(ij)},t_{B(ik)}] = \frac{6}{18+13\rho+\rho^2} \tag{2.27}$$

$$\mathrm{Cov}[t_{A(ij)},t_{B(kl)}] = \frac{4}{18+13\rho+\rho^2} \tag{2.28}$$

Utilizando-se tais covariâncias na equação 2.24, é possível expressar σ_d^2 em função de ρ (Ohta and Kimura, 1971; Weir and Hill, 1986; Weir et al., 1990; McVean, 2002):

$$\sigma_d^2 = \frac{10+\rho}{22+13\rho+\rho^2} = \frac{10+\rho}{(2+\rho)(11+\rho)} \tag{2.29}$$

assumindo que a amostra populacional é finita e muito grande (McVean, 2002). A expressão 2.29 foi desenvolvida a partir de uma população que apresenta recombinação, deriva genética e baixa mutação (Weir et al., 1990). No equilíbrio entre recombinação e deriva genética, σ_d^2, ou $E[r^2]$ como é comumente utilizado, torna-se (Sved, 1968, 1971; Weir et al., 1990; Nielsen and Slatkin, 2013):

$$E[r^2] \approx \sigma_d^2 = \frac{1}{1+\rho} = \frac{1}{1+4N_e r} \tag{2.30}$$

também assumindo que a amostra populacional é finita e muito grande (McVean, 2002).

É importante dizer que o DL das equações 2.29 e 2.30 não foi inicialmente proposto utilizando os princípios da teoria da coalescência (Wakeley, 2009). Tais equações foram obtidas a partir dos conceitos da clássica genética de populações. Porém, é possível investigar o DL a partir da recombinação populacional estimada via coalescência.

No caso do modelo de coalescência com fluxo gênico, McVean (2002) mostrou que apenas uma das cova-riâncias dos tempos de coalescência (equação 2.24) é função da taxa de migração com fluxo gênico ($\mathrm{Cov}[t_{A(ij)},t_{B(ij)}] = \frac{1}{4M^2}$, sendo M igual à $4N_em$ ou $2N_em$), sendo que as demais covariâncias ($\mathrm{Cov}[t_{A(ij)},t_{B(ik)}]$ e $\mathrm{Cov}[t_{A(ij)},t_{B(kl)}]$) são iguais a zero (McVean, 2002).

Até o momento, nenhuma abordagem do DL ainda foi proposta para considerar vários processos simultaneamente. Acredita-se que tal abordagem poderá ser desenvolvida em breve no contexto de coalescência, já que estudos recentes têm sido propostos para modelar processos conjuntamente (Lohse et al., 2011; Mailund et al., 2012).

Estudos do DL baseados na teoria da coalescência podem ser uma estratégia interessante para compreender o contexto histórico-evolutivo de gerações e os seus padrões de associação. Em última análise, tais estudos podem proporcionar um mapeamento genético mais refinado para espécies de interesse (Nordborg and Tavaré, 2002; Rosenberg and Nordborg, 2002; Zöllner and Pritchard, 2005; McVean, 2007a; Kimmel et al., 2008).

2.4 Sorgo

2.4.1 Classificação e Taxonomia

O gênero *Sorghum* é pertencente a tribo Andropogoneae, subfamília Panicoidae, família Poaceae (Gramineae). Dentre as várias classificações propostas para este gênero, as mais compreensíveis e interessantes (Singh and Lohithaswa, 2006) foram apresentadas por Snowden (1936), Garber et al. (1950), Harlan and De Wet (1972a) e De Wet (1978).

Snowden (1936) classificou o gênero *Sorghum* nas categorias de Eusorghum e Para-sorghum, sendo a primeira dividida em Arundinaceae, Spontanea (espécies ou raças selvagens), Sativa (raças cultivadas) e Halepensia. Garber et al. (1950) dividiram o gênero *Sorghum* em seis subgêneros com base em dados citotaxonômicos, a saber: Eusorghum, Chaetosorghum, Heterosorghum, Sorghastrum, Parasorghum e Stiposorghum. Harlan and De Wet (1972a) elaboraram uma classificação simplificada com base em características morfológicas, com as quais os indivíduos selvagens foram agrupados em seis raças espontâneas (arundinaceum, aethiopicum, virgatum, verticilliflorum, propinquum e shattercane) e os indivíduos cultivados foram agrupados em cinco raças básicas (bicolor, guinea, caudatum, kafir e durra) e dez raças híbridas (guinea-bicolor, caudatum-bicolor, kafir-bicolor, durra-bicolor, guinea-caudatum, guinea-kafir, guinea-durra, kafir-caudatum, durra-caudatum e kafir-durra). De Wet (1978) desenvolveu uma classificação dentro do subgênero Eusorghum (ou sorghum) com a qual reconheceu a existência de três diferentes espécies, a saber: *Sorghum halepense* (L.) Pers, *Sorghum propinquum* (Kunth) Hitchc. e *Sorghum bicolor* (L.) Moench.

O sorgo [*Sorghum bicolor* (L.) Moench] é uma espécie autógama e diploide (2n = 20), sendo subdividida em três subespécies: *S. bicolor* ssp. *bicolor*, *S. bicolor* ssp. *drummondii* e *S. bicolor* ssp. *verticilliflorum*. Todos os grupos comerciais de sorgo cultivado, tais como sorgo granífero, forrageiro e vassoura, são classificados dentro da subespécie bicolor (Singh and Lohithaswa, 2006; De Wet and Harlan, 1971). A classificação das cinco raças básicas e das dez raças híbridas intermediárias para o sorgo cultivado é bastante utilizada nos dias de hoje. Detalhes históricos e morfológicos sobre as espécies relatadas, a espécie cultivada *S. bicolor* ssp. *bicolor* e as raças descritas podem ser encontrados em Doggett et al. (1988).

2.4.2 Centro de Origem e Domesticação

A região da Etiópia, localizada na parte leste da África, é considerada como o centro de origem do sorgo (Singh and Lohithaswa, 2006). Esta região possui enorme quantidade de espécies e variedades da raça durra, que está entre as raças básicas cultivadas do sorgo (Singh and Lohithaswa, 2006). Segundo Smith and Frederiksen (2000) e Morris et al. (2013), a domesticação inicial do sorgo aconteceu no Sudão, África, por volta de 10.000 anos atrás, a partir dos indivíduos selvagens e cultivados provenientes da Etiópia, com difusão para outras regiões da África e países da Ásia entre 8.000 e 1.500 anos atrás. Essas regiões compreendem países como Nigéria, Niger e Mali, na parte oeste da África, Uganda e Kênia, na parte leste da África, e India e China, no continente asiático. Da parte oeste da África, o sorgo foi distribuído para os Estados Unidos e outras regiões do mundo por meio do tráfico de escravos há mais de 200 anos (Quinby, 1974, 1975; Rosenow et al., 1983; Rooney and Smith, 2000; Smith and Frederiksen, 2000), em meados do século 19 (Singh and Lohithaswa, 2006). Mesmo antes de 1900, o cultivo do sorgo nos Estados Unidos já apresentava-se de forma bastante extensiva (Rosenow et al., 1983; Singh

and Lohithaswa, 2006).

Em função da origem e difusão antigas, a adaptação do sorgo para climas e práticas culturais específicas foi refletida em termos de variações morfológicas e fisiológicas envolvendo as cinco raças básicas de sorgo domesticado (Harlan and De Wet, 1972b). Essas variações foram o resultado de eventos demográficos importantes, tais como redução do tamanho populacional, estrutura de população, seleção e migração (Hamblin et al., 2006). Mais especificamente, as raças cultivadas de sorgo foram originadas de processos de seleção disruptiva e domesticação ocorridos na região centro-leste da África, a partir de indivíduos selvagens representantes da espécie ancestral *Sorghum arundinaceum*. A seleção natural em espécies selvagens e a seleção artificial praticada pelo homem, para características relacionadas ao sorgo cultivado, resultaram em populações diversificadas com considerável fluxo gênico entre si, explicando muito das variações raciais e do processo evolutivo da espécie (Singh and Lohithaswa, 2006).

2.4.3 Importância e Melhoramento Genético

O sorgo é o quinto cereal mais importante do mundo, permanecendo atrás do trigo, arroz, milho e cevada. Representa grande importância para alimentação humana em diversos países da África, Ásia e América Central e para alimentação animal nos Estados Unidos, Austrália e América do Sul. O sorgo granífero possui enorme capacidade de tolerar condições ambientais adversas, produzindo de forma satisfatória em locais de grande estresse hídrico. Normalmente, esta condição ambiental é desfavorável para o crescimento e a produção de outras culturas, fazendo com que o sorgo mostre o seu potencial e adquira o seu espaço enquanto cultura (Santos et al., 2005; Singh and Lohithaswa, 2006).

Atualmente, o maior produtor mundial de sorgo é os Estados Unidos, com produção anual superior a 10 milhões de toneladas (15% da produção mundial) na safra 2014/2015 (USDA, 2015). O Brasil é o décimo maior produtor mundial, com produção anual de 2 milhões de toneladas, permanecendo atrás de México, Nigéria, Índia, Sudão, Argentina, Etiópia, China e Austrália. A importância do sorgo é recente à agricultura brasileira, sendo explorado de forma intensiva a partir de 1970, com a utilização de quatro tipos cultivados de sorgo: granífero, forrageiro, sacarino e vassoura. Os dois primeiros se destacam tanto pela área cultivada quanto pelo nível tecnológico adotado (Santos et al., 2005).

Em geral, o melhoramento de sorgo tem sido realizado para promover adaptação às condições específicas dos locais de interesse. Isso reflete na obtenção de cultivares com níveis aceitáveis de produtividade, estabilidade e qualidade dos grãos de acordo com o ambiente desejado. Além disso, essas cultivares precisam mostrar resistência ou tolerância a certos fatores bióticos, tais como doenças e pragas, e fatores abióticos, tais como altas temperaturas e estresse hídrico (condições de seca). Ainda, é preciso mostrar qualidade nutricional para atender às exigências das alimentações humana e animal (Singh and Lohithaswa, 2006).

Nos Estados Unidos, o melhoramento do sorgo passou por uma série de fases importantes, a saber (Klein et al., 2008): (i) introdução de um número pequeno de cultivares oriundos da África entre 1874 e 1908; (ii) seleção de plantas de maturação precoce e baixa estatura (insensíveis ao fotoperíodo) a partir de populações heterogêneas entre 1904 e 1936; (iii) melhoramento de plantas visando a colheita mecanizada entre 1930 e 1940; (iv) implementação da produção de sementes híbridas de 1946 até os dias de hoje; e (v) conversão de acessos tropicais exóticos de maturação tardia e alta estatura (sensíveis ao fotoperíodo) em acessos temperados de ma-

turação precoce e baixa estatura a partir de 1963 (Stephens et al., 1967; Quinby, 1974; Rosenow et al., 1983; Klein et al., 2008), de forma a introduzir variabilidade exótica de genótipos tropicais produtivos no germoplasma temperado (Stephens et al., 1967; Casa et al., 2008). O uso deste germoplasma é utilizado até os dias de hoje no melhoramento, e cada uma dessas etapas exerceu grande influência na produção de sorgo nas regiões semi-áridas do mundo (Klein et al., 2008).

A conversão de acessos tropicais em acessos temperados, conhecida como SCP (*Sorghum Conversion Program*) (Stephens et al., 1967), foi uma estratégia importante estabelecida pelo USDA (*United States Department of Agriculture*) em colaboração com a *Texas University* (Casa et al., 2008). Em linhas gerais, vários cruzamentos envolvendo cultivares tropicais (não-adaptadas) e temperadas (adaptadas) foram realizados, e os indivíduos das populações obtidas foram submetidos a vários ciclos de retrocruzamentos com as cultivares tropicais (Rosenow et al., 1997). Dessa forma, grande parte do genoma ($\sim 90\%$) dos indivíduos gerados correspondeu ao genoma da cultivar tropical, com a presença de alelos desejáveis de cultivares temperadas. Em torno de $840 - 850$ linhagens convertidas ou parcialmente convertidas foram desenvolvidas para constituir um novo germoplasma e, assim, uma nova fonte de variabilidade para o melhoramento do sorgo (Casa et al., 2008; Klein et al., 2008).

O melhoramento genético de sorgo no Brasil inclui a obtenção de híbridos visando obter resistência a fatores bióticos (doenças e pragas) e abióticos (tolerância ao estresse hídrico e à toxidez por alumínio, aumento da eficiência de uso de fósforo no solo e acamamento), além de características particulares tanto do local como do tipo de sorgo cultivado (Santos et al., 2005). A obtenção de híbridos resistentes a fatores bióticos e abióticos também tem sido objetivo do melhoramento de países da África e da Ásia (Morris et al., 2013).

2.4.4 Alguns Genes de Importância

Nos últimos 20 anos, diversos estudos de mapeamento genético têm sido realizados em sorgo (Mace and Jordan, 2010). Esses estudos tiveram como objetivo a investigação de características quantitativas importantes, tais como altura (Lin et al., 1995; Pereira and Lee, 1995; Brown et al., 2008), maturidade (Crasta et al., 1999; Chantereau et al., 2001; Hart et al., 2001; Feltus et al., 2006), tolerância ao estresse hídrico (Tuinstra et al., 1996; Crasta et al., 1999; Subudhi et al., 2000; Tao et al., 2000; Xu et al., 2000; Kebede et al., 2001; Harris et al., 2007), tolerância ao alumínio (Magalhaes et al., 2004), restauração da fertilidade (Klein et al., 2001; Jordan et al., 2010) e resistência a doenças e pragas (Tao et al., 2003). Com base nesses e outros trabalhos, Mace and Jordan (2010) reuniram um grupo de 35 genes de maior efeito com o objetivo de posicioná-los conjuntamente em um mapa genético consenso. Esses autores tiveram como proposta a integração das informações disponíveis para promover direcionamentos importantes em programas de melhoramento. Dentre os genes descritos e estudados, Mace and Jordan (2010) consideraram alguns importantes relacionados à maturidade e altura de plantas, que são características correlacionadas e amplamente relatadas em sorgo.

Devido à esta importância, os genes de maturidade e altura de plantas foram escolhidos para a realização do presente estudo, e portanto serão descritos a seguir.

2.4.4.1 Exemplos Relevantes Para o Estudo

Os genes relacionados à maturidade e altura de plantas são conhecidos de longa data em sorgo (Quinby, 1974, 1975). Estudos genéticos clássicos dessas características identificaram quatro locos relacionados à matu-

ridade (*Ma1*, *Ma2*, *Ma3* e *Ma4*) e quatro locos relacionados à altura de plantas (*Dw1*, *Dw2*, *Dw3* e *Dw4*), algo que certamente gerou importante progresso e estimulou o desenvolvimento de trabalhos subsequentes. Os três primeiros locos para maturidade foram identificados por Quinby and Karper (1945), e o quarto loco foi identificado por Quinby (1966). Todos os locos relacionados à altura de plantas foram inicialmente descritos por Quinby and Karper (1954), embora um estudo mais antigo já havia abordado com propriedade sobre o controle genético dessa característica (Karper, 1932).

Para ambas as características, a interação de dominância entre os alelos resulta na obtenção de plantas com maturação tardia e estatura elevada, como é o caso das cultivares tropicais exóticas. Nos Estados Unidos, pesquisadores e produtores selecionaram, ao longo de anos, plantas com maturação precoce e baixa estatura (alelos mutantes), como forma de obter cultivares adaptadas ao clima temperado e à colheita mecanizada (Klein et al., 2008).

Em geral, os alelos mutantes dos locos de maturidade e altura de plantas surgiram após a entrada do sorgo nos Estados Unidos. Para ambas as características, três locos foram inicialmente identificados como tendo alelos recessivos, a saber: (i) *ma1*, *ma2* e *ma3*; e (ii) *dw1*, *dw2* e *dw3*. Em geral, as cultivares convertidas para maturação precoce e baixa estatura apresentam as três mutações para os genes *Ma* e *Dw*. O alelo recessivo *ma4* foi identificado posteriormente com a chegada de uma variedade específica da África, que mostrou adaptação temperada com apenas uma mutação. O alelo recessivo *dw4* já estava presente em certos indivíduos mesmo antes da sua entrada nos Estados Unidos, e o alelo dominante *Dw4* foi verificado em apenas um indivíduo que apresentava características típicas de cultivares tropicais (Quinby, 1975).

Dentre os locos de maturidade, *Ma1* é o que tem maior impacto no período de florescimento e crescimento (Klein et al., 2008; Murphy et al., 2011). Grande proporção da variação no florescimento foi explicada por este loco em estudo anterior (Lin et al., 1995), mostrando que, possivelmente, o surgimento do alelo recessivo *ma1* foi o principal fator na conversão de cultivares tropicais em temperadas. Em estudo posterior, dois locos adicionais para maturidade (*Ma5* e *Ma6*) foram identificados por Rooney and Aydin (1999). A motivação surgiu pelo fato de que outros locos poderiam estar envolvidos no controle do fotoperíodo além dos quatro locos iniciais, algo que já havia sido sugerido por trabalhos anteriores (Quinby et al., 1973; Quinby, 1974; Lin et al., 1995).

Trabalhos recentes mostraram que os genes *Ma1*, *Ma3* e *Ma6* estão situados nos cromossomos 6 (40,30 Mb), 1 (60,92 Mb) e 6 (0,70 Mb) do genoma do sorgo, respectivamente. Os dois primeiros foram anteriormente clonados e possuem localização bem conhecida (Lin et al., 1995; Childs et al., 1997; Murphy et al., 2011), enquanto que a localização do último ainda pode ser incerta (Murphy et al., 2014). Um estudo recente relatou que o gene *Ma5* está localizado no cromossomo 1 (Murphy et al., 2014), muito embora não haja evidências suficientes para suportar tal afirmação. Mace and Jordan (2010) relatam que o gene *Ma5* está localizado no cromossomo 2, tendo-se, portanto, evidências pouco conclusivas. Também não há evidências claras das posições genômicas dos genes *Ma2* e *Ma4*, embora haja informações de que o último está localizado no cromossomo 10 (Mace and Jordan, 2010).

Um estudo recente mostrou que o loco *Ma1* é responsável pela codificação da proteína regulatória de falsa resposta 37 (PRR37, do inglês *Pseudoresponse Regulator Protein 37*) (Murphy et al., 2011). Esses autores avaliaram um germoplasma de sorgo granífero contendo alelos recessivos para *Ma1*, e detectaram variações significativas de PRR37. Essas variações podem estar envolvidas com a expressão e inibição de mecanismos

relacionados ao fotoperíodo em dias curtos e longos, o que é decisivo para a adaptação da espécie em temperaturas diferentes.

O loco *Ma3* foi investigado e clonado em estudo realizado por Childs et al. (1997). Esses autores genotiparam dois marcadores moleculares fisicamente ligados a *Ma3*, em população de linhagens recombinantes, e detectaram fortes associações com o gene *PHYB*. Este gene é responsável pela codificação do fitocromo 123-kD em plantas, e os indivíduos mutantes (*ma3*) apresentaram fenótipos parecidos com plantas que não produziram fitocromo 123-kD. Esses resultados foram decisivos para a conclusão de que o loco *Ma3* é um tipo de *PHYB*.

Com relação aos locos relacionados à altura de plantas, acredita-se que todos eles estejam em cromossomos diferentes no genoma do sorgo (Klein et al., 2008). Os locos *Dw1*, *Dw2* e *Dw3* estão localizados nos cromossomos 9 (57,20 Mb), 6 (42,20 Mb) e 7 (58,61 Mb), respectivamente (Lin et al., 1995; Brown et al., 2008). A localização do loco *Dw4* ainda parece ser incerta, mas acredita-se que ele esteja situado no cromossomo 6 (6,60 Mb), com base em estudo recente de mapeamento associativo (Morris et al., 2013). O loco *Dw2* encontra-se fisicamente ligado ao loco *Ma1* no cromossomo 6 (Klein et al., 2008), e o loco *Dw3* é o único que foi clonado em trabalho anterior (Multani et al., 2003).

2.4.5 Marcadores Moleculares e Genotipagem Por Sequenciamento

Os marcadores moleculares têm sido amplamente utilizados em sorgo para diversas finalidades. Dentre elas, tem-se: (i) construção de mapas genéticos e mapeamento de QTLs para características quantitativas importantes - Mace and Jordan (2010) apresentam uma ampla revisão sobre esses estudos; (ii) análise do desequilíbrio de ligação (Hamblin et al., 2004, 2005, 2007; Bouchet et al., 2012; Morris et al., 2013) e mapeamento associativo (Casa et al., 2008; Sukumaran et al., 2012; Morris et al., 2013); (iii) análise da diversidade genética e do fluxo gênico entre populações (Morden et al., 1989, 1990; Aldrich et al., 1992; Aldrich and Doebley, 1992; Agrama and Tuinstra, 2004; Casa et al., 2005; Folkertsma et al., 2005; Deu et al., 2006, 2008; de Alencar Figueiredo et al., 2008; Mutegi et al., 2011; Sagnard et al., 2011; Bouchet et al., 2012); e (iv) estimação de taxas de recombinação (Hamblin et al., 2005; Morris et al., 2013) e seleção (Hamblin et al., 2005, 2006; Casa et al., 2006; Bouchet et al., 2012). A realização desses estudos contemplou o uso de diversos tipos de marcadores moleculares, tais como aloenzima, RFLP (*Restricted Fragment Length Polymorphism*), RAPD (*Random Amplified Polymorphic DNA*), SSR (*Simple Sequence Repeat*), DArT (*Diversity Arrays Technology*) e SNP (*Single Nucleotide Polymorphism*).

Nos últimos anos, a descoberta de marcadores moleculares e a genotipagem de centenas de indivíduos têm sido realizadas a partir das novas plataformas de sequenciamento para muitos estudos (Varshney et al., 2009; Davey et al., 2011). Com base em um único procedimento de sequenciamento (Davey et al., 2011), essas plataformas possibilitam a geração de grandes quantidades de dados a uma velocidade bastante superior em comparação com as metodologias tradicionais (Varshney et al., 2009), mostrando eficiência, rapidez e custo reduzido. Dentre às diferentes técnicas desenvolvidas a partir desta abordagem, tem-se a genotipagem por sequenciamento (GBS, do inglês *Genotyping-By-Sequencing*), proposta por Elshire et al. (2011). Trata-se de uma técnica simples, rápida, eficiente, específica e de baixo custo, que se baseia na redução da complexidade do genoma através do uso de enzimas de restrição específicas e sensíveis à metilação (Elshire et al., 2011). Muitos fragmentos são gerados a partir dos cortes realizados por essas enzimas, possibilitando o sequenciamento e a geração de dezenas de milhões

de *reads* (sequências curtas de DNA), nas quais polimorfismos podem ser detectados posteriormente (Sansaloni, 2012).

Em função dessas e outras vantagens, a técnica de GBS tem sido utilizada em estudos de diversas espécies de plantas e animais. No entanto, o GBS apresenta desvantagens importantes, tais como a distribuição não uniforme das leituras realizadas e a alta quantidade de dados perdidos (Beissinger et al., 2013). Segundo Davey et al. (2011), Poland et al. (2012) e Poland and Rife (2012), esta última é decorrência da baixa cobertura do genoma exercida pela técnica, que pode ocorrer tanto por questões de amostragem quanto por motivos biológicos. Contudo, diversas estratégias de imputação têm sido desenvolvidas e podem ser utilizadas para recuperar informações perdidas (Marchini and Howie, 2010). No caso do mapeamento e seleção genômica, isso não chega a ser um problema, mas, para alguns tipos de estudo, a baixa cobertura pode ter implicações na qualidade dos resultados.

3 MATERIAL E MÉTODOS

Como foi dito na Introdução (seção 1), o objetivo deste trabalho foi analisar, com base na teoria da coalescência, a estrutura populacional e o desequilíbrio de ligação em um painel mundial de acessos de sorgo. Para alcançar tal objetivo, análises de mutação, migração com fluxo gênico e recombinação foram realizadas para diferentes regiões genômicas e populações que constituem este painel.

Nesta seção, descreveremos inicialmente o painel de acessos e os dados moleculares gerados a partir do procedimento de genotipagem, incluindo a estratégia adotada para a seleção das regiões genômicas de interesse. Na sequência, mostraremos as estatísticas de diversidade e os testes de neutralidade utilizados, que são métodos clássicos inicialmente investigados para se ter uma visão global da variação dos dados. Por fim, apresentaremos os modelos de coalescência, incluindo os modelos de estrutura populacional (migração com fluxo gênico) e recombinação (abordagens completa e aproximada), e o desequilíbrio de ligação via coalescência.

3.1 Material Biológico e Dados Moleculares

3.1.1 Painel de Acessos

O painel de acessos do presente estudo é constituído por 356 indivíduos de sorgo oriundos de diferentes continentes e regiões geográficas do mundo, tais como África, Ásia, América do Norte, América Central e América do Sul. Países historicamente importantes para a origem e o melhoramento desta espécie estão presentes, tais como Etiópia, Índia, Nigéria, África do Sul, Sudão, Uganda e Estados Unidos. Outros países importantes também estão presentes, mas com menor quantidade de acessos. É o caso de Kênia, Chad, Congo, Egito, Mali, Niger, Senegal, Burkina Faso, Tanzânia, entre outros. Este painel é parte de uma coleção elaborada para estudos de mapeamento associativo e diversidade genética, que foram recentemente publicados por Casa et al. (2008) e Morris et al. (2013). As localidades dos acessos foram coletadas no *Germplasm Resources Information Network (GRIN)*, pertencente ao *Agricultural Research Service (ARS)* do *United States Department of Agriculture (USDA)*.

Do total de 356 indivíduos, 114 correspondem a linhagens tropicais exóticas e 242 correspondem a linhagens convertidas. Em geral, as linhagens tropicais exóticas (mais antigas) foram selecionadas devido à sua importância histórica para os programas de melhoramento por serem fontes de resistência ou tolerância a estresses bióticos e abióticos, além da diversidade genética e geográfica presentes nos seus centros de origem (África). As linhagens convertidas (mais recentes) foram selecionadas de acordo com a insensibilidade ao fotoperíodo (SCP) e o desempenho como cultivares em diferentes localidades do mundo (Casa et al., 2008).

3.1.2 Genotipagem Por Sequenciamento e Imputação de Dados

A genotipagem de marcadores por sequenciamento foi realizada para os 356 indivíduos do painel de acessos. As etapas foram realizadas pelo *Institute for Genomic Diversity* pertencente à *Cornell University* (Ithaca, NY, EUA), seguindo o método proposto por Elshire et al. (2011) em plataforma HiSeq 2000 (Illumina$^{\delta}$ Inc., San Diego, CA, EUA). Inicialmente, o DNA genômico dos indivíduos foi digerido pela enzima *Ape*KI, que reconhece sítio contendo cinco bases (uma delas é degenerada) e possui sensibilidade à metilação. A partir desta digestão, bibliotecas de 96 e 384-plex foram construídas para os indivíduos, e as sequências obtidas via Illumina foram mapeadas contra o genoma de referência do sorgo (Paterson et al., 2009). Tal mapeamento foi realizado com base

no método proposto por LI ET AL. (2009).

Para a descoberta dos SNPs, utilizou-se o *pipeline* `TASSEL-GBS` implementado na versão 3.0 do software `TASSEL`. Este *pipeline* foi descrito detalhadamente por GLAUBITZ ET AL. (2014) com base em versão mais recente do software. *Tags* de GBS contendo 64 pares de bases (pb) foram alinhadas contra o genoma de referência do sorgo (linhagem *BTx623*), e aquelas que estiveram presentes, no mínimo, dez vezes ($10\times$) nas bibliotecas foram mantidas para o mapeamento. Com base no arquivo *HapMap* gerado ao final do processo, um total de 265,487 SNPs foram descobertos para os 356 indíviduos (e as repetições utilizadas como controle). A imputação de dados perdidos foi realizada via software `NPUTE` (ROBERTS ET AL., 2007), com janelas de SNPs variando de 5 a 150 pb para cada cromossomo. As janelas que mostraram as maiores acurácias foram selecionadas e utilizadas para imputação.

Os dados moleculares gerados com os procedimentos anteriores foram gentilmente disponibilizados pelo Prof. Dr. Geoff Morris, do *Department of Agronomy*, *Crop Genetics* & *Genomics* pertencente à *Kansas State University* (Manhattan, KS, EUA).

3.1.3 Análises Prévias no Painel

Para reduzir a complexidade das análises de coalescência devido à grande quantidade de locais, decidiu-se utilizar apenas os países Etiópia (ETI), Índia (IND), Nigéria (NIG), África do Sul (AFS), Sudão (SUD), Uganda (UGA) e Estados Unidos (EUA). No entanto, para não reduzir o painel inicialmente proposto, análises de componentes principais foram previamente realizadas na tentativa de agrupar os acessos dos demais países naqueles aqui mais representativos. Tais análises foram realizadas utilizando-se os dados totais de SNPs, obtidos via GBS e `NPUTE`, e o painel de acessos com os 356 indivíduos. Essas análises foram realizadas no pacote `pcaMethods` (STACKLIES ET AL., 2007), disponível no software `R` (R CORE TEAM, 2015).

3.1.4 Seleção das Regiões Genômicas

Sete locos genômicos importantes para a história de melhoramento do sorgo foram selecionados para investigação no presente estudo. Estes locos correspondem a genes relacionados com altura de plantas (*Dw1*, *Dw2*, *Dw3* e *Dw4*) e maturidade (*Ma1*, *Ma3* e *Ma6*). Vários estudos têm evidenciado suas posições estimadas no genoma do sorgo (LIN ET AL., 1995; CHILDS ET AL., 1997; MULTANI ET AL., 2003; BROWN ET AL., 2008; MURPHY ET AL., 2011; MORRIS ET AL., 2013; MURPHY ET AL., 2014). No entanto, as posições que delimitam o início e o fim desses genes são desconhecidas ou pouco precisas até o momento, sendo necessário a adoção de estratégias alternativas para a investigação.

Assim, a partir das posições estimadas desses locos no genoma, o DL baseado na medida r^2 (HILL and ROBERTSON, 1968) foi estimado entre todos os pares de SNPs a cada 5.000 pb, caminhando-se simultaneamente tanto para esqueda quanto para direita. Uma vez que SNPs com elevado DL tendem a mostrar uma mesma genealogia evolutiva, acredita-se que SNPs adjacentes possam fornecer evidências importantes para os genes de interesse. O número 5.000 foi utilizado como referência ao tamanho médio de um gene em sorgo, com base no comprimento de sequências preditas relacionadas à maturidade desta espécie (BHOSALE ET AL., 2012).

Variantes adjacentes às posições estimadas foram selecionadas enquanto houve DL entre os SNPs mais extremos, delimitando assim as regiões de interesse. A presença de DL foi detectada com base na correção de Bonferroni (WEIR ET AL., 1990) para múltiplos testes, a partir das probabilidades obtidas pelo teste exato de Fisher

(Fisher, 1935). Essas análises foram realizadas utilizando-se o software R (R CORE TEAM, 2015).

3.2 Diversidade Genética e Testes de Neutralidade

Estatísticas de diversidade genética e neutralidade foram estimadas para cada região genômica de interesse, a saber: (i) número de sítios polimórficos (S) considerando todos os acessos (Watterson, 1975); (ii) diferença nucleo-tídica entre pares de sequências (π) envolvendo todos os acessos (Tajima, 1983); (iii) taxa de mutação populacional (θ_W) para cada sítio e todos os sítios (Watterson, 1975); e (iv) estatísticas D de Tajima (Tajima, 1989) e R_2 de Ramos-Onsins-Rozas (Ramos-Onsins and Rozas, 2002) como testes de neutralidade para investigar desvios do modelo padrão baseado em neutralismo (Wakeley, 2009). A estatística R_2 foi especificamente utilizada para detectar desvios causados pela alteração do tamanho efetivo populacional (expansão ou redução). Essas análises foram realizadas utilizando-se o pacote `pegas` (Paradis, 2010), disponível no software R (R CORE TEAM, 2015).

3.3 Modelos de Coalescência

3.3.1 Estrutura Populacional e Fluxo Gênico

Análises de coalescência com fluxo gênico foram realizadas para sete populações de diferentes localidades utilizando-se a versão bayesiana da abordagem proposta por Beerli and Felsenstein (2001). Tais localidades foram apresentadas na seção 3.1.3. Esta abordagem baseia-se no modelo de isolamento geográfico com fluxo gênico (Wright, 1931), que é fundamentado nos pressupostos do modelo de Fisher (1930) e Wright (1931) estruturado. No total, 49 parâmetros populacionais foram simultaneamente estimados, correspondendo às sete taxas de mutação das populações e às 42 taxas de fluxo gênico assimétricas envolvendo todos os possíveis pares de populações. Uma matriz de migração com fluxo gênico (Nordborg, 2007) $\mathcal{P}$ foi obtida contendo todos os parâmetros populacionais estimados, conforme representação mostrada a seguir (Beerli and Felsenstein, 2001). Tal representação foi modificada em relação à matriz original de Beerli and Felsenstein (2001), que não manteve os índices (subescritos) das taxas de fluxo gênico de acordo com os conceitos da teoria de matrizes, dificultando o entendimento.

$$\mathcal{P} = \begin{pmatrix} \theta_1 & \mathcal{M}_{12} & \mathcal{M}_{13} & \dots & \mathcal{M}_{1n} \\ \mathcal{M}_{21} & \theta_2 & \mathcal{M}_{23} & \dots & \mathcal{M}_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ \mathcal{M}_{n1} & \mathcal{M}_{n2} & \dots & \mathcal{M}_{n,n-1} & \theta_n \end{pmatrix}$$

Na matriz $\mathcal{P}$, θ_1,θ_2,...,θ_n as taxas de mutação correspondentes às n populações e $\mathcal{M}_{12}$,$\mathcal{M}_{13}$,...,$\mathcal{M}_{n,n-1}$ as taxas de fluxo gênico assimétricas envolvendo pares de populações. O parâmetro $\mathcal{M}_{ij}$ é equivalente a razão $\frac{m_{ij}}{\mu}$, sendo m_{ij} a taxa de fluxo gênico por geração que ocorre da população i para a população j e μ a taxa de mutação por sítio e geração. O parâmetro θ_i é igual a $2N_e^{(i)}\mu$, sendo $N_e^{(i)}$ o tamanho efetivo da população i em um modelo de Fisher (1930) e Wright (1931). Oportunamente, o valor de θ do presente modelo será referido como θ_{LM}.

Para a estimação da matriz de parâmetros $\mathcal{P}$, genealogias ancestrais importantes e *correlacionadas* foram simuladas com base no procedimento de MCMC. A partir de uma distribuição a priori logarítmica para θ e

linear para $\mathcal{M}$ (KUHNER and SMITH, 2007), e do modelo evolutivo F84 apresentado por FELSENSTEIN and CHURCHILL (1996), o MCMC foi realizado por meio do algoritmo Metropolis-Hastings (METROPOLIS ET AL., 1953; HASTINGS, 1970). Assim, certa genealogia simulada no estado G_t foi obtida a partir de pequenas alterações nos ramos ancestrais da genealogia simulada no estado G_{t-1} (BEERLI and FELSENSTEIN, 2001), com base nos coalescentes condicionais (FELSENSTEIN ET AL., 1999). A transição de uma genealogia para outra ao longo do MCMC, indicando progresso no processo de simulação, ocorreu sempre que a probabilidade da genealogia G_t foi superior à probabilidade da genealogia G_{t-1}.

Inicialmente, 20 cadeias curtas (iniciais), cada qual com 20.000 genealogias, foram percorridas para obter estimativas iniciais e direcionadas dos parâmetros. Com base nessas estimativas, 10 cadeias longas (finais), cada qual com 400.000 genealogias, foram percorridas, e as estimativas da última cadeia longa foram utilizadas para as inferências populacionais. Adicionalmente, este mesmo procedimento foi realizado considerando quatro cadeias independentes e em paralelo (MCMCMC ou MC^3, do inglês *Metropolis-Coupled Markov Chain Monte Carlo*) (GEYER, 1991; KUHNER ET AL., 2000) com diferentes temperaturas (1,0; 1,2; 1,5 e 4,0) (MORRELL ET AL., 2003), sendo que, ao final de cada cadeia curta e longa, permutações foram realizadas com o objetivo de aumentar a busca e eventualmente obter estimativas mais precisas dos parâmetros. Para todos os processos, as primeiras 40.000 e 400.000 genealogias das cadeias curtas e longas, respectivamente, foram descartadas como período de aquecimento (*burnin*), correspondendo a 10% do total. O procedimento foi repetido por 10 vezes para o caso em que não houve MC^3 e 5 vezes para o caso em que houve MC^3, sendo realizado para cada região gênomica de interesse. Em função do tempo computacional, o MC^3 não foi realizado para a análise de todas as regiões simultaneamente. Um valor máximo de 10.000 eventos de migração foi utilizado para cada processo realizado.

A função de verossimilhança do presente modelo pode ser representada pela seguinte equação (BEERLI and FELSENSTEIN, 2001):

$$L(\mathcal{P}) = \int_{\mathcal{G}} P(\mathcal{G} \mid \mathcal{P}) P(D \mid \mathcal{G}) \tag{3.1}$$

sendo $\mathcal{G}$ o espaço de genealogias simuladas e $\mathcal{P}$ a matriz que contém os parâmetros correspondentes a θ e $\mathcal{M}$. Os parâmetros de $\mathcal{P}$ foram estimados para cada genealogia simulada, sendo estratégia utilizada pela abordagem bayesiana no presente contexto.

Na expressão de verossimilhança da equação 3.1, $P(\mathcal{G} \mid \mathcal{P})$ foi calculado com base no seguinte modelo (BEERLI and FELSENSTEIN, 2001):

$$P(\mathcal{G} \mid \mathcal{P}) = \prod_{j=1}^{T} \left[\exp\left(-u_j \left[\sum_{i=1}^{s} \frac{k_{ij}(k_{ij}-1)}{\theta_i} + k_{ij} \sum_{z \neq i}^{s} \mathcal{M}_{zi} \right] \right) \times \left(\delta_j \mathcal{M}_{w_j v_j} + (1-\delta_j) \frac{2}{\theta_v} \right) \right]$$

O termo exponencial é a probabilidade de que em cada tempo j e em seu comprimento de genealogia u_j não aconteçam nem eventos de fluxo gênico nem eventos de coalescência. O segundo termo da expressão indica que esses eventos podem supostamente acontecer. Tais eventos correspondem ao fluxo gênico ($\mathcal{M}$), que ocorre da população w_j para a população v_j, e ao coalescente, que ocorre para a população v_j. A variável indicadora δ_j apresenta o valor de 1 quando um evento de fluxo gênico acontece no tempo j ou o valor de 0 para o caso contrário. k_{ij} corresponde ao número de sequências presentes na população i durante o intervalo de tempo j.

As análises do modelo de coalescência com fluxo gênico foram realizadas utilizando a versão `Linux` do pacote `LAMARC` (KUHNER, 2006, 2009).

3.3.2 Recombinação

3.3.2.1 Abordagem Completa

Estimativas de recombinação constante via coalescência foram obtidas para cada população e região genômica de interesse. Tais estimativas consideram a ocorrência de uma única taxa de recombinação histórica entre todos os SNPs presentes em certa região. Para tanto, utilizou-se a abordagem de máxima verossimilhança proposta por KUHNER ET AL. (2000), que utiliza o procedimento do MCMC e o algoritmo Metropolis-Hastings (METROPOLIS ET AL., 1953; HASTINGS, 1970) para simular gráficos de recombinação ancestral ou ARGs (GRIFFITHS and MARJORAM, 1996), contemplando tanto eventos de coalescência quanto eventos de recombinação (HEIN ET AL., 2004). A partir de dois valores iniciais para a mutação ($\theta_{0(1)}$: valor de θ_W e $\theta_{0(2)}$: valor de 0,01 atribuído pelo software utilizado) e um único valor inicial para a recombinação (ρ_0: valor de 0,01 atribuído pelo software utilizado), ARGs foram simulados utilizando a estratégia dos coalescentes condicionais, baseada no modelo F84 proposto por FELSENSTEIN and CHURCHILL (1996). De maneira independente, eventos de mutação foram incorporados nos ARGs amostrados (HUDSON, 1990; MCVEAN and CARDIN, 2005), de modo que, ao final, estimativas de mutação e recombinação (únicas e constantes) foram obtidas a partir do conjunto de ARGs simulados. Essas estimativas representam os valores de mutação e recombinação que podem ser pontos de máxima verossimilhança. Desconsiderou-se a ocorrência de múltiplas recombinações e interferência entre intervalos adjacentes, bem como processos de conversão gênica.

No caso da mutação, o valor de $\theta_{L\rho} = 2N_e\mu$ por sítio foi estimado para cada região e população. No caso da recombinação, os valores de $\frac{r_L}{\mu_L}$ e ρ_L entre dois possíveis sítios também foram estimados para cada região e população. Utilizou-se a mesma quantidade de iterações estabelecida para o modelo com fluxo gênico, pois, embora a quantidade de parâmetros seja aqui menor, trata-se de modelo complexo e difícil de ser investigado (HEIN ET AL., 2004; MCVEAN and CARDIN, 2005; NORDBORG, 2007). Então, 20 cadeias curtas de 20.000 ARGs e 10 cadeias longas de 400.000 ARGs foram percorridas, com o descarte (*burnin*) das primeiras 40.000 e 400.000 cadeias, respectivamente. Quatro cadeias independentes e em paralelo, com diferentes temperaturas (1,0; 1,2; 1,5 e 4,0) (MORRELL ET AL., 2003), também foram adicionalmente utilizadas para aumentar o espaço de ARGs simulados.

Apenas duas repetições foram realizadas considerando ausência e presença de MC^3. Isso foi decidido basicamente por três motivos: (i) a abordagem de máxima verossimilhança é normalmente mais intensiva que a bayesiana do ponto de vista computacional; (ii) dois valores iniciais para θ foram considerados; e (iii) as estimativas foram obtidas para cada região e população. Um valor máximo de 1.000 recombinações foi utilizado para cada processo.

As análises de coalescência para taxa de recombinação constante também foram realizadas na versão `Linux` do pacote `LAMARC` (KUHNER, 2006, 2009).

3.3.2.2 Abordagem Aproximada

Estimativas de recombinação variável para cada região genômica e população também foram obtidas com base em abordagem aproximada de verossimilhança. Ao contrário de se estimar um único valor constante

de recombinação para toda a região, como foi realizado pela abordagem completa (seção 3.3.2.1), estimativas variáveis envolvendo pares de SNPs foram calculadas utilizando-se o método de Auton and McVean (2007). Tal método é baseado no rjMCMC (Green, 1995), que consiste em simular ARGs de alta probabilidade com base em diferentes dimensões, aumentando o espaço da busca.

O método de Auton and McVean (2007) foi utilizado por meio de três etapas. Na primeira delas, a taxa de mutação populacional foi estimada a partir de uma versão modificada do método de Watterson (1975), que é baseada no modelo de sítios finitos. Tal modelo considera que determinado sítio polimórfico pode ter ocorrido a partir de vários eventos históricos de mutação. Então, com base em uma amostra de n sequências de comprimento L e região contendo S sítios polimórficos, foi possível obter θ_W^* para cada sítio e geração (equação 3.2):

$$\theta_W^* = \left(\sum_{i=1}^{n-1} \frac{1}{i} \right)^{-1} \ln\left(\frac{L}{L-S}\right) \tag{3.2}$$

Na segunda etapa, a região em estudo foi dividida em conjuntos envolvendo pares de SNPs, reduzindo a complexidade das análises. Na terceira etapa, o método de verossimilhança completa proposto por Fearnhead and Donnelly (2001) foi utilizado para cada conjunto de SNPs, possibilitando a simulação de muitas histórias e, assim, a estimação das taxas de mutação e recombinação. Considerou-se a possibilidade de mutação reversa e utilizou-se um intervalo de diferentes taxas de recombinação constante. Ao final, um conjunto de verossimilhanças completas e pré-calculadas foi obtido, possibilitando o cálculo da verossimilhança composta ou aproximada (L_C) de acordo com a equação 3.3 (Auton and McVean, 2007):

$$L_C(\rho) = \sqrt[S-1]{\prod_{i,j} P(D_{ij} \mid \rho_{r(ij)})} \tag{3.3}$$

em que a raiz com índice dependente de S corresponde à uma correção para que a verossimilhança seja confiavelmente calculada. Esta correção foi sugerida por Auton and McVean (2007) por meio da análise de vários cenários com dados simulados. D_{ij} e $\rho_{r(ij)}$ correspondem aos dados e às taxas de recombinação populacional entre os SNPs i e j, respectivamente. $\rho_{r(ij)}$ é igual à $2N_e r_{r(ij)}$ (neste caso, os indivíduos são homozigotos para todos os locos), sendo $r_{r(ij)}$ a taxa de recombinação por indivíduo e geração entre os SNPs i e j.

Os parâmetros de simulação utilizados com o método de Fearnhead and Donnelly (2001) foram: (i) taxa de mutação inicial calculada para cada sítio com base no estimador de Watterson (1975); (ii) taxas de recombinação constante variando de 0 a 100 para cada conjunto de SNPs; e (iii) posições investigadas a cada 1 Kb de distância. Os parâmetros utilizados com o método de Auton and McVean (2007) foram: (i) taxas de mutação e recombinação obtidas a partir do método de Fearnhead and Donnelly (2001); (ii) 10.000.000 de iterações percorridas; (iii) penalizações variando de 0 a 50, conforme sugerido pelos desenvolvedores, e 10 repetições para cada uma das penalizações; (iv) reamostragem de ARGs a cada 100 processos de iteração; e (v) 1.000.000 de cadeias descartadas em cada processo referentes ao período de aquecimento (*burnin*). As estimativas médias de recombinação variável foram selecionadas a partir da penalização que apresentou a menor variação ao longo das 10 repetições.

As análises foram realizadas na versão 2.2 do pacote `LDhat` utilizando os programas `complete` (Fearnhead and Donnelly, 2001) e `rhomap` (Auton and McVean, 2007).

3.4 Desequilíbrio de Ligação Via Coalescência

Até o momento, nenhuma medida do DL foi proposta considerando múltiplos processos estimados via coa-lescência. Nesse sentido, as estimativas de recombinação variável obtidas via `rhomap` (AUTON and MCVEAN, 2007) foram utilizadas para o cálculo do valor esperado do DL ($E[r^2]$), considerando cada população e região genômica de interesse. A abordagem do $E[r^2]$ escolhida (equação 2.30) foi aquela em que a população encontra-se em equilíbrio entre recombinação e deriva genética (SVED, 1968, 1971; WEIR ET AL., 1990). Isso porque a outra abordagem disponível do $E[r^2]$ (equação 2.29) considera a presença adicional de (baixa) mutação, que, em essência, é importante apenas por revelar algo do processo evolutivo (NORDBORG and TAVARÉ, 2002). Então, com as atenções voltadas ao mapeamento associativo, onde a recombinação é o fator de importância para o DL (PRITCHARD and PRZEWORSKI, 2001; NORDBORG and TAVARÉ, 2002), decidiu-se utilizar o $E[r^2]$ da equação 2.30.

Assim, $E[r^2]$, ou especificamente $E[r^2_{\texttt{rhomap}}]$, pôde ser calculado pela seguinte equação:

$$E[r^2_{\texttt{rhomap}}] \approx \frac{1}{1+\rho_{r(ij)}} = \frac{1}{1+2N_e r_{r(ij)}} \tag{3.4}$$

em que $\rho_{r(ij)}$ e $r_{r(ij)}$ correspondem às taxas de recombinação populacional e por indivíduo entre os SNPs i e j, respectivamente. Note, portanto, que a equação 3.4 é equivalente à equação 2.30.

O DL via medida r^2 (HILL and ROBERTSON, 1968) também foi calculado para os mesmos pares de SNPs, e a correlação de Spearman entre r^2 e $E[r^2_{\texttt{rhomap}}]$ foi investigada utilizando o software `R` (R CORE TEAM, 2015). O DL via coalescência também foi calculado a partir da estimativa única e constante de recombinação obtida pelo método de KUHNER ET AL. (2000), obtendo-se $E[r^2_{\texttt{LAMARC}}]$. Esta quantidade foi adicionalmente calculada como forma de inferir sobre o DL de alta ordem envolvendo múltiplos locos simultaneamente dentro de certa região.

4 RESULTADOS

4.1 Seleção dos Acessos de Sorgo

Do total de 265.487 SNPs gerados com a técnica de GBS, 247.806 ($93{,}34\%$) foram selecionados e utilizados nas análises prévias de componentes principais. O processo de imputação via `NPUTE` identificou 17.681 ($6{,}66\%$) sítios como sendo monomórficos (único nucleotídeo) para o presente painel, os quais, portanto, foram excluídos das análises. As maiores acurácias obtidas estiveram entre $96{,}70\%$ e $97{,}38\%$ de acordo com o cromossomo, a partir de janelas variando de 26 a 34 SNPs.

O gráfico das análises de componentes principais é mostrado na Figura 4.1. É possível observar certo agrupamento dos indivíduos pertencentes a populações como ETI, IND, NIG e AFS. Contudo, os dados totais de SNPs não sugeriram a formação de grupos claros de acordo com as populações. Para os EUA, por exemplo, os indivíduos se agruparam com acessos de todas as demais populações, de forma que os componentes principais pouco sugeriram padrões que possibilitassem a inclusão dos indivíduos de populações menos representativas nas sete populações selecionadas. Portanto, 58 indivíduos pertencentes às populações não selecionadas foram excluídos do painel inicial, restando, assim, um subpainel de 298 para as análises.

4.2 Seleção das Regiões Genômicas

De um total de sete locos genômicos inicialmente selecionados, cinco foram mantidos para as análises de coalescência. Apenas 14 e 4 SNPs foram observados para os locos *Dw3* e *Ma6* (respectivamente) enquanto houve DL a partir das suas posições estimadas no genoma (resultados não mostrados). No nosso entendimento, essas quantidades são insuficientes para as inferências, e portanto ambos os locos foram desconsiderados do presente trabalho.

Os padrões do DL envolvendo pares de SNPs adjacentes às posições dos locos *Dw1*, *Dw2*, *Dw4*, *Ma1* e *Ma3* são mostrados na Figura 4.2. Os gráficos referentes a todos os locos são mostrados até o instante em que associações significativas estiveram presentes ou queda pronunciada do DL foi observada, com base na correção de Bonferroni para múltiplos testes.

Para o loco *Dw1*, uma região adjacente de 70 quilo bases (Kb) foi selecionada tanto para direita quanto para esquerda da sua posição estimada (cromossomo 9: $57{,}2$ Mb), totalizando 140 Kb. Nesta região, 89 SNPs foram observados e 900 ($22{,}98\%$) associações de pares foram estatisticamente significativas quanto ao desequilíbrio de um total de 3.916 associações. Para o loco *Dw2*, uma região de 75 Kb foi selecionada a partir da sua posição estimada (cromossomo 6: $42{,}2$ Mb), totalizando 150 Kb. Ao todo, 85 SNPs foram detectados e 682 ($19{,}10\%$) associações mostraram estar em DL a partir de 3.570 combinações possíveis envolvendo pares de SNPs. Para o loco *Dw4*, uma região total de 140 Kb foi selecionada ao redor da sua posição estimada (cromossomo 6: $6{,}6$ Mb), contendo 54 SNPs e 653 ($45{,}63\%$) associações em DL de um total de 1.431 pares de SNPs. Para o importante loco de maturidade *Ma1*, uma região total de 160 Kb foi selecionada adjacentemente à sua posição estimada (cromossomo 6: $40{,}3$ Mb), contendo 39 SNPs e 269 ($36{,}30\%$) associações preferenciais a partir de 741 associações totais. E, por fim, uma região total de 120 Kb foi selecionada em torno da posição estimada do loco *Ma3* (cromossomo 1: $60{,}92$ Mb), contendo 52 SNPs e 235 ($17{,}72\%$) associações estatisticamente significativas de um total de 1.326 combinações.

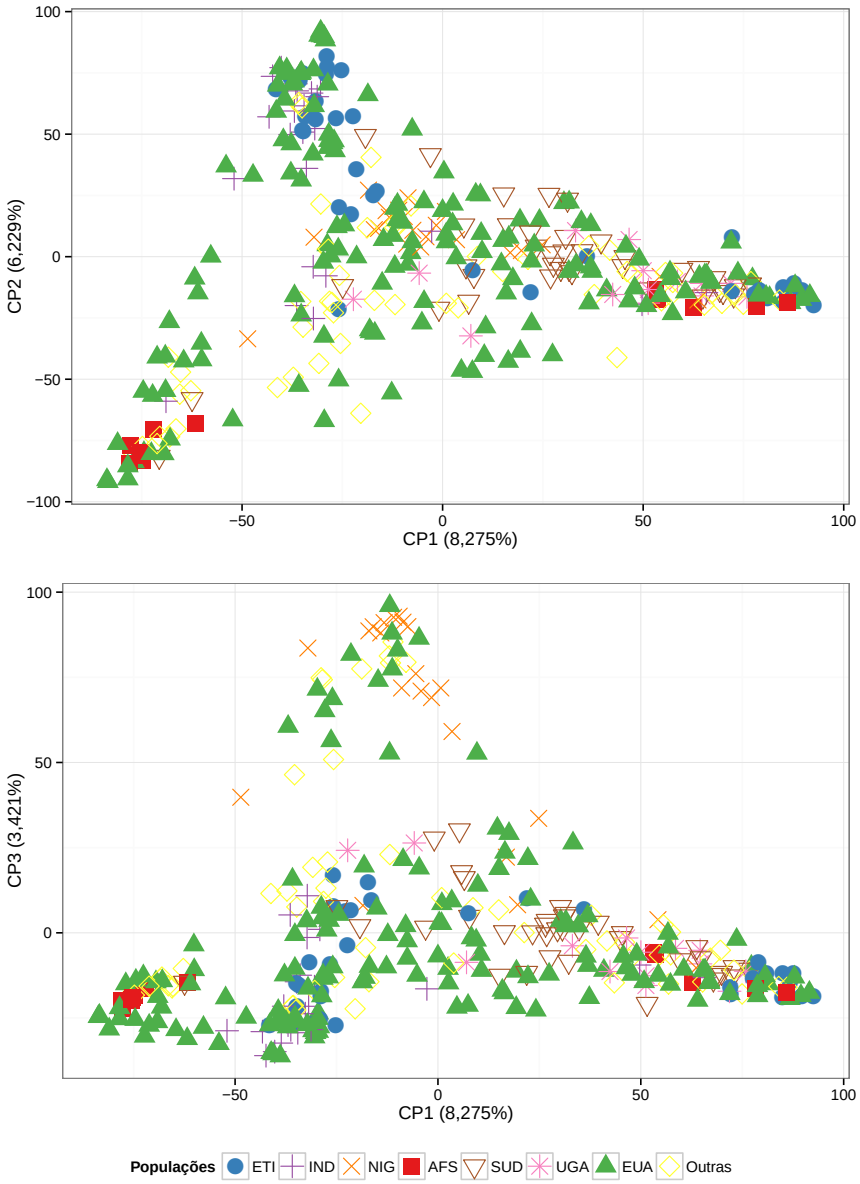

Figura 4.1: Gráfico das análises de componentes principais utilizando 247.806 SNPs e 356 indivíduos do painel mundial de acessos de sorgo. Os três primeiros componentes principais foram utilizados para inferir sobre os padrões de agrupamento. As populações referentes aos países Etiópia, Índia, Nigéria, África do Sul, Sudão, Uganda e Estados Unidos são indicadas na legenda como ETI, IND, NIG, AFS, SUD, UGA e EUA, respectivamente. Os indivíduos oriundos das demais populações estão representados na legenda por Outras

4.3 Estatísticas de Diversidade e Testes de Neutralidade

As estatísticas de diversidade genética e os testes de neutralidade são mostrados na Tabela 4.1. Tais quantidades são apresentadas para se ter uma visão inicial dos padrões de variação. É possível observar variabilidade genética entre as regiões genômicas de interesse, com base na quantidade de sítios polimórficos (S), nas estimativas das taxas de mutação para toda a região e por sítio ($\theta_{W(\text{Região})}$ e $\theta_{W(\text{Sítio})}$) e na diversidade nucleotídica (π). A região *Dw1*, localizada no cromossomo 9, foi a que apresentou o maior polimorfismo ($S = 89$) e os maiores valores das taxas de mutação ($\theta_{W(\text{Região})} = 14{,}189$ e $\theta_{W(\text{Sítio})} = 1{,}148 \times 10^{-4}$). Em contrapartida, a região que apresentou o menor polimorfismo ($S = 39$) e as menores estimativas de mutação foi a *Ma1* ($\theta_{W(\text{Região})} = 6{,}217$ e

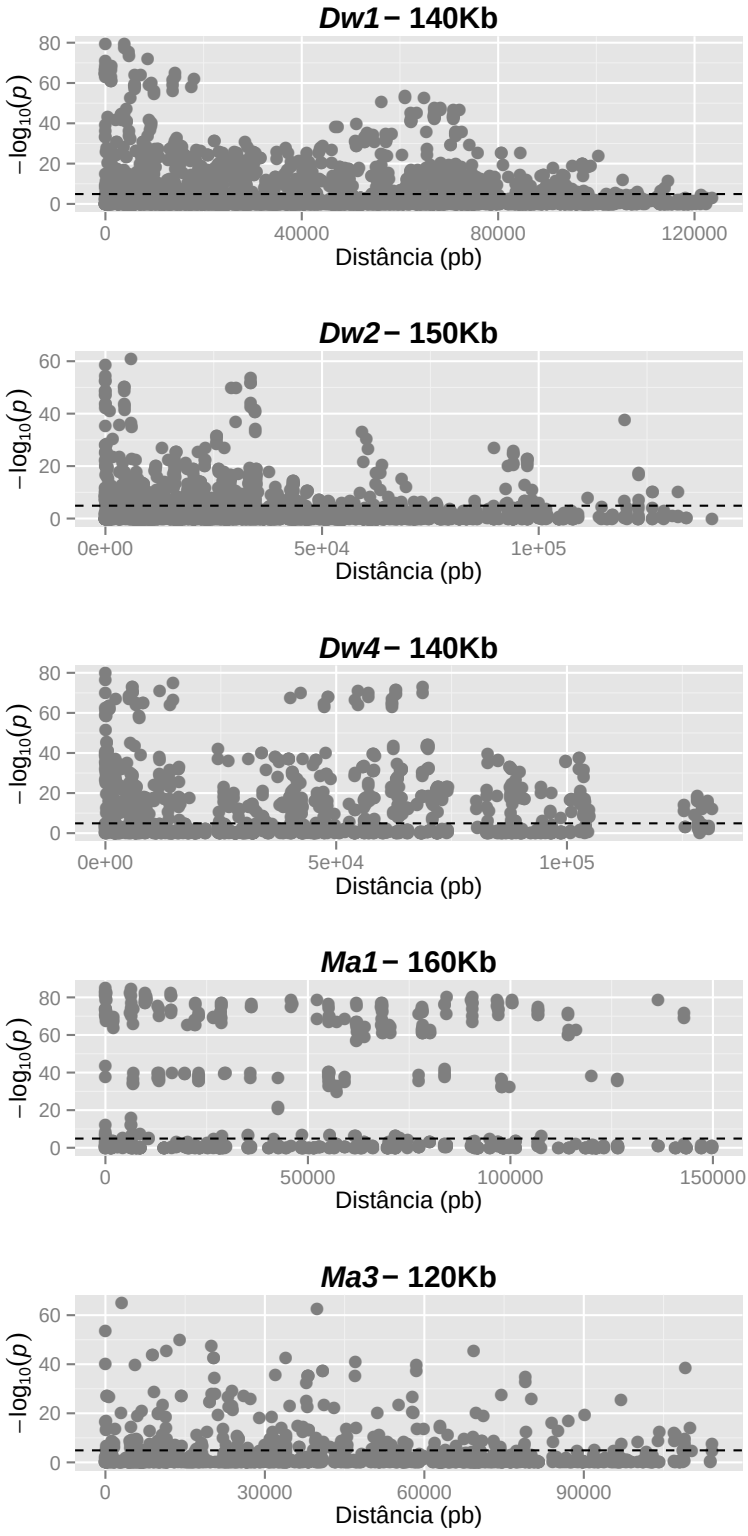

Figura 4.2: Queda do desequilíbrio de ligação, $-\log_{10}(p)$, em função da distância física entre SNPs em pares de bases (pb). A probabilidade obtida a partir do teste exato de Fisher é mostrada em escala logarítmica, e o limiar correspondente ao procedimento de Bonferroni para múltiplos testes é indicado pela linha horizontal pontilhada. Os pares de SNPs estatisticamente significativos (em DL) estão indicados acima da linha do Bonferroni. No título de cada gráfico, são mostradas as extensões em Kb determinadas para cada região genômica contendo o respectivo loco

$\theta_{W(\text{Sítio})} = 0{,}415 \times 10^{-4}$), situada no cromossomo 6. Um aspecto interessante é que os locos fisicamente ligados *Dw2* e *Ma1* apresentaram grandes variações de polimorfismo e das taxas de mutação por região e para cada sítio. Embora o tamanho definido para as regiões possa influenciar nos padrões de variação, *Ma1* foi a região que teve a maior quantidade de pares de bases e a menor variação. Esta influência pode ter sido aqui pouco expressiva.

Com base na comparação de pares envolvendo os 298 indivíduos do subpainel utilizado, as maiores diferenças entre nucleotídeos de um mesmo sítio polimórfico foram apresentadas pelas regiões *Ma1* ($\pi = 0{,}293$) e *Dw4* ($\pi = 0{,}238$). A região *Ma1* foi a única que apresentou valor significativo para a estatística *D*, indicando desvios do modelo padrão baseado em neutralismo. O valor positivo de *D* sugeriu uma menor proporção de polimorfismos de baixa e alta frequências para a região *Ma1*. Um valor positivo e elevado de *D* também foi observado para a região *Dw4*, embora este não tenha sido estatisticamente significativo. Nenhuma região genômica apresentou valor significativo para a estatística de crescimento populacional R_2.

Tabela 4.1: Estatísticas de diversidade genética e testes de neutralidade para as cinco regiões genômicas considerando o subpainel de 298 indivíduos de sorgo

Região	Cr.	*S*	Tamanho (pb)	$\theta_{W(\text{Região})}$	$\theta_{W(\text{Sítio})}$	π	*D*	R_2
Dw1	9	89	123.587	14,189	1,148	0,151	$-0{,}164^{ns}$	$0{,}075^{ns}$
Dw2	6	85	140.029	13,551	0,968	0,124	$-0{,}673^{ns}$	$0{,}062^{ns}$
Dw4	6	54	131.451	8,609	0,655	0,238	$1{,}432^{ns}$	$0{,}119^{ns}$
Ma1	6	39	149.808	6,217	0,415	0,293	$2{,}348^{**}$	$0{,}146^{ns}$
Ma3	1	52	114.138	8,290	0,726	0,185	$0{,}457^{ns}$	$0{,}092^{ns}$

Notas: *S* é o número de sítios polimórficos; θ_W é a taxa de mutação por geração (Watterson, 1975); π corresponde à diversidade nucleotídica, que é a proporção da soma das diferenças entre pares de indivíduos e o total de comparações de pares (Tajima, 1983); *D* (Tajima, 1989) e R_2 (Ramos-Onsins and Rozas, 2002) correspondem aos testes de neutralidade; ^{ns}: Não significativo; **: P-valor < 0,01; $\theta_{W(\text{Região})} = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}$; $\theta_{W(\text{Sítio})} = \frac{\theta_{W(\text{Região})}}{\text{Tamanho}}$; valores de $\theta_{W(\text{Sítio})}$ estão na escala de 10^{-4}

4.4 Estrutura Populacional e Fluxo Gênico

As estimativas de mutação ($\theta_{L\mathcal{M}}$) e fluxo gênico ($\mathcal{M}$) envolvendo as sete populações de sorgo são apresentadas na Figura 4.3. Uma inspeção desses resultados mostra que, em geral, as regiões genômicas contendo os locos *Dw1*, *Dw2*, *Dw4*, *Ma1* e *Ma3* sugeriram elevadas taxas de fluxo gênico para múltiplos pares de populações, algo que não foi observado com a mesma magnitude na análise com todas as regiões simultaneamente. Embora os intervalos de credibilidade tenham sido, em geral, menores e mais precisos na análise total, as estimativas de cada região mostraram evidências importantes para a história de melhoramento do sorgo. Tais evidências não se resumem apenas às elevadas taxas de fluxo de gênico, mas também aos padrões específicos obtidos para cada região, sugerindo histórias evolutivas distintas.

Um exemplo claro da importância da análise individual ocorreu para as populações da ETI e de UGA. O número médio de indivíduos que migraram da ETI para a UGA por geração, e que forneceram contribuição

genética-evolutiva, foi bastante elevado considerando as regiões *Dw1* ($\mathcal{M}_{16} = 66{,}55$) e *Dw4* ($\mathcal{M}_{16} = 91{,}90$). No entanto, tal número de migrantes foi menor e próximo para as regiões *Dw2* ($\mathcal{M}_{16} = 11{,}03$) e *Ma1* ($\mathcal{M}_{16} = 11{,}40$) e também para o *Total* ($\mathcal{M}_{16} = 7{,}11$), embora ainda tenha sido elevado. Para a região *Ma3*, o número médio de indivíduos que migraram da ETI para a UGA por geração foi bastante inferior ($\mathcal{M}_{16} = 0{,}55$).

O número médio de migrantes por geração que forneceu contribuição genética-evolutiva variou de $0{,}55$ (*Ma3*, ETI−UGA) a $99{,}10$ (*Dw2*, NIG−EUA). As maiores taxas médias de fluxo gênico foram observadas para as regiões *Dw4* ($52{,}07$), *Ma1* ($48{,}97$) e *Ma3* ($47{,}35$), enquanto que as menores taxas foram verificadas para *Dw1* ($44{,}56$), *Total* ($43{,}33$) e *Dw2* ($41{,}78$). O maior fluxo gênico observado foi da NIG para os EUA na região *Dw2* ($99{,}10$), com a primeira tendo a menor quantidade de polimorfismo (5 SNPs) entre todas as populações. Algo similar também foi observado para as populações da AFS e de UGA na região *Ma3*. No primeiro caso, elevado fluxo gênico foi verificado da AFS para as populações da ETI ($65{,}94$), da NIG ($88{,}33$) e do SUD ($79{,}23$), utilizando um total de 18 SNPs. No segundo caso, elevado fluxo gênico foi observado de UGA para a IND ($95{,}26$), o SUD ($93{,}89$) e os EUA ($98{,}19$), considerando 15 SNPs.

Todos os possíveis pares envolvendo as sete populações apresentaram taxas de fluxo gênico assimétricas, seguindo o modelo proposto. Isso significa que, por exemplo, o número médio de indivíduos que migraram da ETI para o SUD ($\mathcal{M}_{15} = 46{,}61$) por geração não foi o mesmo que do SUD para a ETI ($\mathcal{M}_{51} = 66{,}67$), com base nos intervalos de credibilidade não sobrepostos na análise total. Do total de 21 combinações envolvendo pares de populações, apenas 11 mostraram assimetria a partir da análise com todas as regiões. Tais combinações assimétricas são: (i) ETI-SUD ($\mathcal{M}_{15} = 46{,}61$ e $\mathcal{M}_{51} = 66{,}66$), (ii) ETI-EUA ($\mathcal{M}_{17} = 21{,}95$ e $\mathcal{M}_{71} = 41{,}33$), (iii) IND-AFS ($\mathcal{M}_{24} = 32{,}16$ e $\mathcal{M}_{42} = 58{,}82$), (iv) NIG-AFS ($\mathcal{M}_{34} = 67{,}83$ e $\mathcal{M}_{43} = 46{,}68$), (v) NIG-SUD ($\mathcal{M}_{35} = 73{,}51$ e $\mathcal{M}_{53} = 26{,}73$), (vi) NIG-EUA ($\mathcal{M}_{37} = 88{,}72$ e $\mathcal{M}_{73} = 58{,}68$), (vii) AFS-SUD ($\mathcal{M}_{45} = 93{,}64$ e $\mathcal{M}_{54} = 37{,}46$), (viii) AFS-UGA ($\mathcal{M}_{46} = 14{,}59$ e $\mathcal{M}_{64} = 47{,}04$), (ix) AFS-EUA ($\mathcal{M}_{47} = 49{,}87$ e $\mathcal{M}_{74} = 35{,}77$), (x) SUD-EUA ($\mathcal{M}_{57} = 42{,}42$ e $\mathcal{M}_{75} = 33{,}78$) e (xi) UGA-EUA ($\mathcal{M}_{67} = 57{,}09$ e $\mathcal{M}_{76} = 89{,}08$). As demais combinações mostraram assimetria com base nos padrões específicos de cada região, mostrando, novamente, a importância da análise individual no entendimento do processo evolutivo.

4.5 Recombinação e Desequilíbrio de Ligação

As taxas de mutação (por sítio) obtidas a partir do modelo de coalescência com recombinação ($\theta_{L\rho}$) mostraram estimativas superiores às taxas de mutação (por sítio) obtidas a partir da abordagem clássica (θ_W) (Tabela 4.2). Embora $\theta_{L\rho}$ e θ_W tenham apresentado uma elevada correlação de Spearman ($r = 0{,}779$), os resultados diferentes mostraram que a máxima verossimilhança foi importante para obter estimativas mais precisas da variabilidade genética. Isso porque θ_W, por pouco explorar a investigação genealógica a partir dos dados, não capitaliza toda a informação disponível nos mesmos. Além disso, maior precisão deve ter sido obtida pela estimação simultânea de $\theta_{L\rho}$ e da taxa de recombinação dependente do polimorfismo ($\frac{r_L}{\mu_L}$), mostrando um cenário evolutivo possivelmente mais real.

Ao comparar as taxas de mutação (por sítio) dos modelos de fluxo gênico ($\theta_{L\mathcal{M}}$) e recombinação ($\theta_{L\rho}$), observou-se grande diferença em suas estimativas (Figura 4.3 e Tabela 4.2). Sempre que $\theta_{L\mathcal{M}}$ mostrou estimativa inferior a 2×10^{-4} para certa região e população, $\theta_{L\rho}$ mostrou estimativa superior, atingindo um valor dez vezes maior ou mais para alguns casos (*Dw1*: NIG; *Dw2*: NIG; *Dw4*: NIG e AFS; *Ma1*: UGA). Em contrapartida,

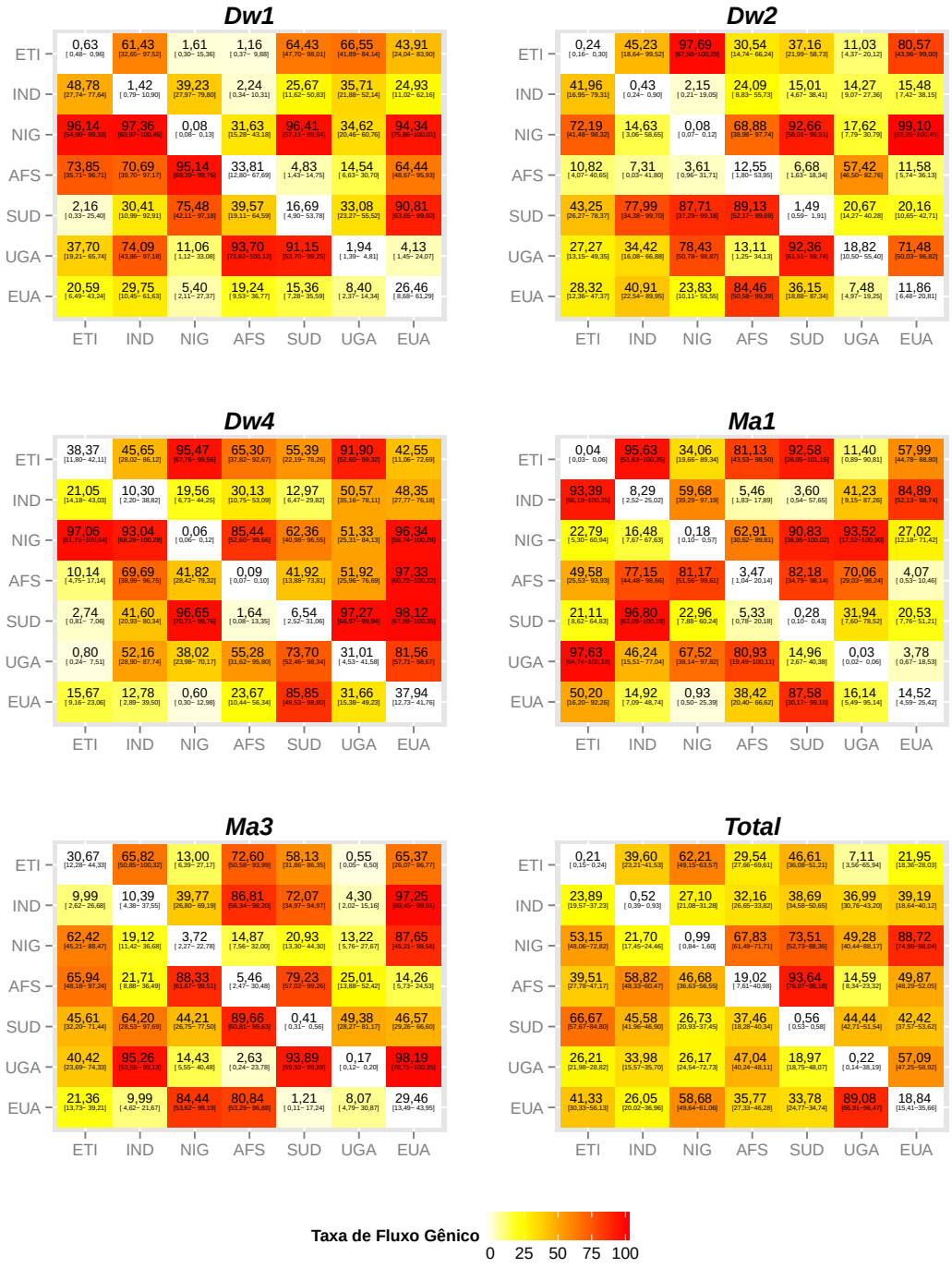

Figura 4.3: Taxas de mutação (θ_{LM}) e fluxo gênico ($\mathcal{M}$) estimadas via coalescência (LAMARC) para sete populações de sorgo (ETI, IND, NIG, AFS, SUD, UGA e EUA), considerando cada região genômica e todas elas simultaneamente. As taxas de mutação (escala de 10^{-4}) estão mostradas na diagonal principal das matrizes, correspondendo ao número médio de mutações por sítio entre dois indivíduos quaisquer até a ocorrência do evento de coalescência. As taxas de fluxo gênico encontram-se nos elementos fora da diagonal principal, correspondendo ao número médio de migrantes por geração. Os números entre colchetes correspondem aos intervalos de credibilidade. Neste modelo (BEERLI and FELSENSTEIN, 2001), as taxas de mutação e fluxo gênico são constantes ao longo das gerações. O primeiro e o segundo índices dos elementos das matrizes correspondem às populações que forneceram e receberam migrantes, respectivamente. Por exemplo, $\mathcal{M}_{12}$, de qualquer uma das matrizes mostradas, é a taxa de fluxo gênico por geração resultante da migração de indivíduos da ETI para a IND

sempre que $\theta_{L\mathcal{M}}$ foi superior a 2×10^{-4}, a estimativa de $\theta_{L\rho}$ foi inferior, com $\theta_{L\mathcal{M}}$ atingindo um valor dez vezes maior ou bem acima para alguns casos (*Dw1*: AFS, SUD e EUA; *Dw2*: AFS e UGA; *Dw4*: ETI, IND, UGA e EUA; *Ma1*: IND, AFS e EUA; *Ma3*: ETI, IND e EUA). A correlação de Spearman entre $\theta_{L\mathcal{M}}$ e $\theta_{L\rho}$ foi igual à $0{,}484$, mostrando uma estimativa inferior quando $\theta_{L\rho}$ foi correlacionado com θ_W da abordagem clássica ($r = 0{,}779$).

A região genômica que mostrou a maior taxa de mutação média (escala de 10^{-4}) foi a *Dw1* (2,27), seguida das regiões *Dw4* (1,36), *Dw2* (1,30), *Ma3* (1,17) e *Ma1* (0,73). As maiores variabilidades genéticas, em ordem decrescente e de acordo com a região e a população, foram: (i) 3,414 (*Dw1*, AFS); (ii) 2,780 (*Dw1*, EUA); (iii) 2,722 (*Dw1*, UGA); (iv) 2,659 (*Dw2*, EUA); (v) 2,189 (*Dw1*, NIG); (vi) 2,182 (*Ma3*, EUA); e (vii) 2,054 (*Dw4*, EUA).

As taxas de recombinação constante (`LAMARC`) que são dependentes da mutação ($\frac{r_L}{\mu_L}$) apresentaram variações entre as regiões e populações (Tabela 4.2). A maior taxa média de $\frac{r_L}{\mu_L}$ (escala de 10^{-3}) por Kb foi para a região *Dw2* (1,578), seguida das regiões *Ma3* (1,209), *Dw1* (0,347), *Ma1* (0,213) e *Dw4* (0,129). As médias elevadas de $\frac{r_L}{\mu_L}$ das regiões *Dw2* e *Ma3* ocorreram principalmente devido às altas estimativas das populações da NIG (10,159) e da IND (6,789), respectivamente. As taxas de recombinação independentes da mutação (ρ_L) mostraram menores variações, com o maior valor médio (escala de 10^{-3}) por Kb apresentado pela região *Ma3* (0,246), seguida das regiões *Dw2* (0,121), *Dw1* (0,115), *Dw4* (0,057) e *Ma1* (0,030). As taxas mais elevadas de ρ_L ocorreram para a região *Ma3*, por meio das populações da IND (0,882), da ETI (0,329) e o Subpainel (0,346). A taxa de ρ_L da NIG na região *Dw2* (0,224), embora elevada, mostrou ser bastante inferior ao valor de ρ_L da IND na região *Ma3* (0,882). Ambas tiveram valores parecidos de recombinação quando houve dependência da mutação, mas apresentaram valores bem distintos na sua independência.

A Tabela 4.3 apresenta as taxas de recombinação mínima ($\rho_{r(\mathrm{Min})}$), média ($\rho_{r(\mathrm{Med})}$) e máxima ($\rho_{r(\mathrm{Max})}$), estimadas via coalescência (`rhomap`), obtidas para cada região genômica e população. Ao contrário das taxas de recombinação mostradas na Tabela 4.2, que foram obtidas a partir do modelo de recombinação constante (Kuhner et al., 2000), tais recombinações foram obtidas a partir do modelo de recombinação variável ao longo da região (Auton and McVean, 2007). De forma geral, os padrões obtidos indicaram processos de recombinação específicos para cada região, e também para cada população dentro de certa região. Por exemplo, as menores e maiores taxas de recombinação detectadas para as regiões foram: (i) 0,053 (UGA) e 1,721 (ETI) para *Dw1*; (ii) 0,021 (UGA) e 76,320 (Subpainel) para *Dw2*; (iii) 0,010 (NIG) e 1,625 (ETI) para *Dw4*; (iv) 0,004 (AFS) e 1,934 (EUA) para *Ma1*; e (v) 0,058 (SUD) e 2,039 (IND) para *Ma3*. Considerando todas as populações, a maior taxa de recombinação média foi a da região *Dw2* (3,395), seguida das regiões *Ma3* (0,465), *Dw1* (0,403), *Dw4* (0,241) e *Ma1* (0,060). Além disso, algumas populações apresentaram recombinações máximas bastante diferentes da taxa de recombinação média para certa região, como foi o caso da NIG (6,729) e dos EUA (38,439) na região *Dw2*. As taxas de recombinação dos EUA e do Subpainel nesta região foram extremamente elevadas, estando muito além do valor médio observado.

Enquanto a Tabela 4.3 apresenta as taxas de recombinação mínima, média e máxima para todas as regiões genômicas e populações, o Gráfico 4.4 mostra as taxas de recombinação (ρ_r/Kb) entre todos os SNPs adjacentes para algumas regiões e populações. Tais regiões mostraram padrões bastante variáveis de recombinação, e portanto foram utilizadas como exemplo para discussão. Esses padrões podem ser claramente visualizados por meio da escala diferente do eixo y. O caso (A), correspondendo à população da ETI na região *Dw1*, e o caso (D),

Tabela 4.2: Taxas de mutação ($\theta_{L\rho}$) e recombinação ($\frac{r_L}{\mu_L}$ e ρ_L por Kb) estimadas via coalescência (LAMARC) para cada uma das cinco regiões genômicas e sete populações de sorgo, bem como para o subpainel de 298 indivíduos. A taxa de mutação estimada via abordagem clássica (θ_W) também é mostrada, e ambas θ_W e $\theta_{L\rho}$ (escala de 10^{-4}) correspondem ao número médio de mutações por sítio. $\frac{r_L}{\mu_L}$ (dependente da mutação) e ρ_L (independente da mutação) (escala de 10^{-3}) correspondem ao número médio de recombinações entre dois indivíduos quaisquer até a ocorrência de um evento de coalescência. Neste modelo (KUHNER ET AL., 1995), $\theta_{L\rho}$, $\frac{r_L}{\mu_L}$ e ρ_L são constantes ao longo das gerações. IC é o intervalo de confiança. $E[r^2_{\texttt{LAMARC}}]$ corresponde ao desequilíbrio de ligação de alta ordem calculado a partir do ρ_L de toda a região

Região	População	S[1]	$\theta_{W\,(\text{Sítio})}$	$\theta_{L\rho\,(\text{Sítio})}$	IC$_{\theta_{L\rho}}$ (95%)	$\frac{r_L}{\mu_L}$/Kb	IC $_{\frac{r_L}{\mu_L}}$ (95%)	ρ_L/Kb	$E[r^2_{\texttt{LAMARC}}]$
Dw1	ETI	47	0,924	1,275	0,771 – 2,153	0,237	0,119 – 0,451	0,042	0,996
	IND	50	1,083	1,635	0,943 – 2,427	0,194	0,093 – 0,368	0,044	0,996
	NIG	58	1,288	2,189	1,339 – 4,191	0,237	0,114 – 0,479	0,072	0,994
	AFS	60	1,565	3,414	2,477 – 5,466	0,237	0,119 – 0,451	0,042	0,996
	SUD	53	1,021	1,851	1,181 – 2,441	0,449	0,279 – 0,605	0,115	0,989
	UGA	47	1,196	2,722	1,412 – 5,336	0,306	0,128 – 0,491	0,116	0,989
	EUA	86	1,243	2,780	2,283 – 3,414	0,547	0,425 – 0,694	0,211	0,982
	Subpainel	89	1,148	3,572	3,025 – 4,105	0,565	0,455 – 0,726	0,281	0,976
Dw2	ETI	31	0,538	0,892	0,455 – 1,214	0,628	0,314 – 0,912	0,092	0,992
	IND	31	0,593	1,153	0,637 – 1,736	0,301	0,116 – 0,541	0,057	0,995
	NIG	05	0,098	0,134	0,097 – 0,170	10,159	5,202 – 14,19	0,224	0,982
	AFS	36	0,829	1,402	0,880 – 2,871	0,187	0,061 – 0,347	0,043	0,997
	SUD	68	1,156	1,718	1,238 – 2,477	0,217	0,148 – 0,365	0,062	0,995
	UGA	42	0,943	1,129	0,656 – 3,600	0,108	0,029 – 0,383	0,020	0,998
	EUA	81	1,033	2,659	2,125 – 2,981	0,394	0,339 – 0,524	0,172	0,986
	Subpainel	85	0,968	2,871	2,428 – 3,211	0,627	0,541 – 0,768	0,296	0,976
Dw4	ETI	46	0,850	1,282	0,756 – 2,029	0,130	0,022 – 0,146	0,040	0,998
	IND	42	0,856	1,343	0,653 – 2,469	0,040	0,010 – 0,093	0,013	0,999
	NIG	38	0,793	0,892	0,489 – 1,298	0,041	0,000 – 0,067	0,009	0,999
	AFS	33	0,809	0,896	0,468 – 5,525	0,006	0,000 – 0,068	0,001	0,999
	SUD	43	0,779	1,339	0,793 – 2,025	0,095	0,067 – 0,234	0,031	0,998
	UGA	38	0,909	1,693	0,781 – 2,867	0,096	0,038 – 0,179	0,040	0,998
	EUA	49	0,666	2,054	1,389 – 2,338	0,253	0,214 – 0,403	0,127	0,993
	Subpainel	54	0,655	2,112	1,721 – 2,329	0,374	0,301 – 0,530	0,192	0,989
Ma1	ETI	28	0,454	0,549	0,234 – 0,963	0,010	0,000 – 0,032	0,002	0,999
	IND	30	0,536	0,896	0,521 – 1,401	0,063	0,032 – 0,125	0,022	0,999
	NIG	25	0,458	0,555	0,331 – 1,054	0,005	0,000 – 0,025	0,001	0,999
	AFS	25	0,538	0,490	0,232 – 1,070	0,000	0,000 – 0,004	0,000	1,000
	SUD	26	0,413	0,607	0,336 – 0,836	0,036	0,007 – 0,081	0,008	0,999
	UGA	24	0,504	1,291	0,422 – 2,153	0,007	0,000 – 0,028	0,004	0,999
	EUA	39	0,465	0,687	0,620 – 0,765	0,283	0,244 – 0,349	0,075	0,997
	Subpainel	39	0,415	0,247	0,229 – 0,268	1,302	1,163 – 1,515	0,124	0,995
Ma3	ETI	35	0,745	1,144	0,943 – 1,326	1,311	1,139 – 1,621	0,329	0,983
	IND	34	0,798	0,592	0,529 – 0,642	6,789	6,215 – 7,717	0,882	0,957
	NIG	36	0,865	1,704	1,062 – 2,738	0,142	0,076 – 0,285	0,053	0,997
	AFS	18	0,508	0,957	0,451 – 1,986	0,072	0,000 – 0,190	0,015	0,999
	SUD	28	0,584	1,153	0,738 – 1,690	0,186	0,094 – 0,358	0,047	0,997
	UGA	15	0,413	0,478	0,219 – 1,075	0,000	0,000 – 0,031	0,000	1,000
	EUA	51	0,798	2,182	1,923 – 2,679	0,619	0,493 – 0,717	0,297	0,985
	Subpainel	52	0,726	2,848	2,556 – 3,422	0,554	0,429 – 0,622	0,346	0,983

[1] Número de sítios polimórficos

Tabela 4.3: Taxas de recombinação mínima ($\rho_{r(\mathrm{Min})}$), média ($\rho_{r(\mathrm{Med})}$) e máxima ($\rho_{r(\mathrm{Max})}$) estimadas via coalescência (`rhomap`) para cada uma das cinco regiões genômicas e sete populações de sorgo, bem como para o subpainel de 298 indivíduos. Neste modelo de coalescência (AUTON and MCVEAN, 2007), tais taxas correspondem ao número esperado de recombinações que acontece para determinado conjunto (par) de SNPs em uma distância de 1 Kb. Essas taxas são consideradas variáveis ao longo da região de interesse

Região	População	$\rho_{r(\mathrm{Min})}$	$\rho_{r(\mathrm{Med})}$	$\rho_{r(\mathrm{Max})}$
Dw1	ETI	0,344	0,680	1,721
	IND	0,073	0,234	1,169
	NIG	0,110	0,285	1,076
	AFS	0,115	0,364	1,252
	SUD	0,110	0,406	1,435
	UGA	0,053	0,223	0,698
	EUA	0,145	0,432	1,320
	Subpainel	0,186	0,599	1,712
Dw2	ETI	0,329	0,768	2,029
	IND	0,035	0,329	1,548
	NIG	0,302	2,420	6,729
	AFS	0,183	0,491	1,583
	SUD	0,285	0,888	1,954
	UGA	0,021	0,273	1,301
	EUA	0,210	7,125	38,439
	Subpainel	0,346	14,868	76,320
Dw4	ETI	0,060	0,288	1,625
	IND	0,036	0,061	0,181
	NIG	0,010	0,033	0,621
	AFS	0,100	0,278	0,905
	SUD	0,117	0,334	1,146
	UGA	0,079	0,381	1,131
	EUA	0,089	0,291	1,554
	Subpainel	0,148	0,265	0,660
Ma1	ETI	0,009	0,011	0,021
	IND	0,018	0,162	1,357
	NIG	0,009	0,022	0,134
	AFS	0,004	0,006	0,018
	SUD	0,013	0,016	0,025
	UGA	0,013	0,017	0,050
	EUA	0,039	0,159	1,934
	Subpainel	0,034	0,085	1,053
Ma3	ETI	0,279	0,619	1,132
	IND	0,558	0,838	2,039
	NIG	0,152	0,366	1,499
	AFS	0,171	0,340	0,551
	SUD	0,058	0,194	0,898
	UGA	0,599	0,695	0,938
	EUA	0,198	0,340	0,756
	Subpainel	0,181	0,327	1,241

correspondendo à população da IND na região *Ma3*, parecem se assemelhar na variação de ρ_r, que permanece entre $0{,}500$ e $2{,}000$ recombinações médias por Kb. O caso (B), que corresponde à população dos EUA na região *Dw2*, mostra um valor de $38{,}439$ recombinações médias por Kb, uma estimativa bastante elevada e diferente dos padrões das demais regiões e populações. Por fim, o caso (C) mostra os baixos padrões de recombinação da população do SUD na região *Ma1*, não ultrapassando o valor de $0{,}025$ recombinações médias por Kb.

O DL calculado via medida tradicional (r^2) e coalescência ($E[r^2_{\texttt{rhomap}}]$) é mostrado na Figura 4.5. É possível observar que os padrões de ambos foram muito próximos em alguns casos. Foi o que aconteceu, por exemplo, para as regiões e populações *Dw2* (IND), *Dw2* (UGA) e *Ma1* (IND, NIG, AFS, SUD e UGA), onde, em geral, forte DL foi observado em grandes blocos de associação. Contudo, é possível perceber que o r^2 não parece fornecer uma definição clara da extensão de possíveis blocos para casos em que os padrões são completamente elevados na região. Um exemplo claro acontece com a população da NIG na região *Dw4*. Embora ambas as medidas evidenciam que há forte DL nesta região, há certa descontinuidade nos padrões mostrados por r^2, algo que não acontece com o $E[r^2_{\texttt{rhomap}}]$. Possivelmente, esse comportamento está associado ao viés comumente verificado com o uso do r^2, de modo que o DL via coalescência ($E[r^2_{\texttt{rhomap}}]$) pode ser fundamental para uma melhor definição das associações.

O DL de alta ordem ($E[r^2_{\texttt{LAMARC}}]$, Tabela 4.2) sugeriu estimativas extremamente elevadas ($0{,}957 - 1{,}000$) para todas as regiões genômicas e populações. Tais estimativas foram obtidas a partir das baixas taxas de recombinação (ρ_L) sugeridas pelo `LAMARC` considerando o polimorfismo de toda a região genômica. Embora houve casos em que o r^2 e o $E[r^2_{\texttt{rhomap}}]$ foram bastante elevados e mostraram estar próximos de $E[r^2_{\texttt{LAMARC}}]$, este DL não mostrou estimativa representativa das associações para todas as regiões e populações. Tal evidência se torna ainda mais clara ao comparar as estimativas de $0{,}957 - 0{,}999$ do $E[r^2_{\texttt{LAMARC}}]$ com o DL médio a partir das estimativas do r^2 (Figura 4.5), que variaram da seguinte forma, de acordo com a região genômica: (i) $0{,}127 - 0{,}351$ para região *Dw1*; (ii) $0{,}106 - 0{,}486$ para a região *Dw2*; (iii) $0{,}212 - 0{,}567$ para a região *Dw4*; (iv) $0{,}259 - 0{,}713$ para a região *Ma1*; e (iv) $0{,}059 - 0{,}320$ para a região *Ma3*.

Os padrões observados na Figura 4.5 claramente evidenciaram uma estrutura específica do DL para as populações ao longo das regiões. Essa especificidade pode ser visualizada de duas diferentes maneiras: (i) DL entre regiões genômicas; e (ii) DL entre populações para uma dada região. No primeiro caso, é possível perceber que o DL variou consideravelmente de acordo com a região do genoma. As regiões que pareceram ter DL mais extensivo foram *Ma1*, *Dw4* e *Dw2*, em ordem decrescente de associação. As regiões *Dw1* e *Ma3* também apresentaram DL relativamente extensivo, só que em menor proporção e para uma ou outra população.

No segundo caso, o DL mostrou-se de forma bastante específica de acordo com a população. Com exceção da região *Ma1*, em que praticamente todas as populações apresentaram forte bloco de associação, todas as demais tiveram populações com padrões específicos de DL. Por exemplo, as populações da IND e de UGA mostraram maior DL em comparação com as demais para ambas as regiões *Dw1* e *Dw2*. Isso também ocorreu com as populações da ETI, IND e NIG na região *Dw4*, e do SUD na região *Ma3*.

Um resultado muito curioso foi verificado para o Subpainel de indivíduos. Para todas as regiões, o DL presente no Subpainel foi inferior ou próximo àquele verificado para as demais populações. Isso mostra que a mistura populacional de indivíduos oriundos de diferentes localidades não gerou acréscimos nos padrões do DL, algo que pode eventualmente acontecer.

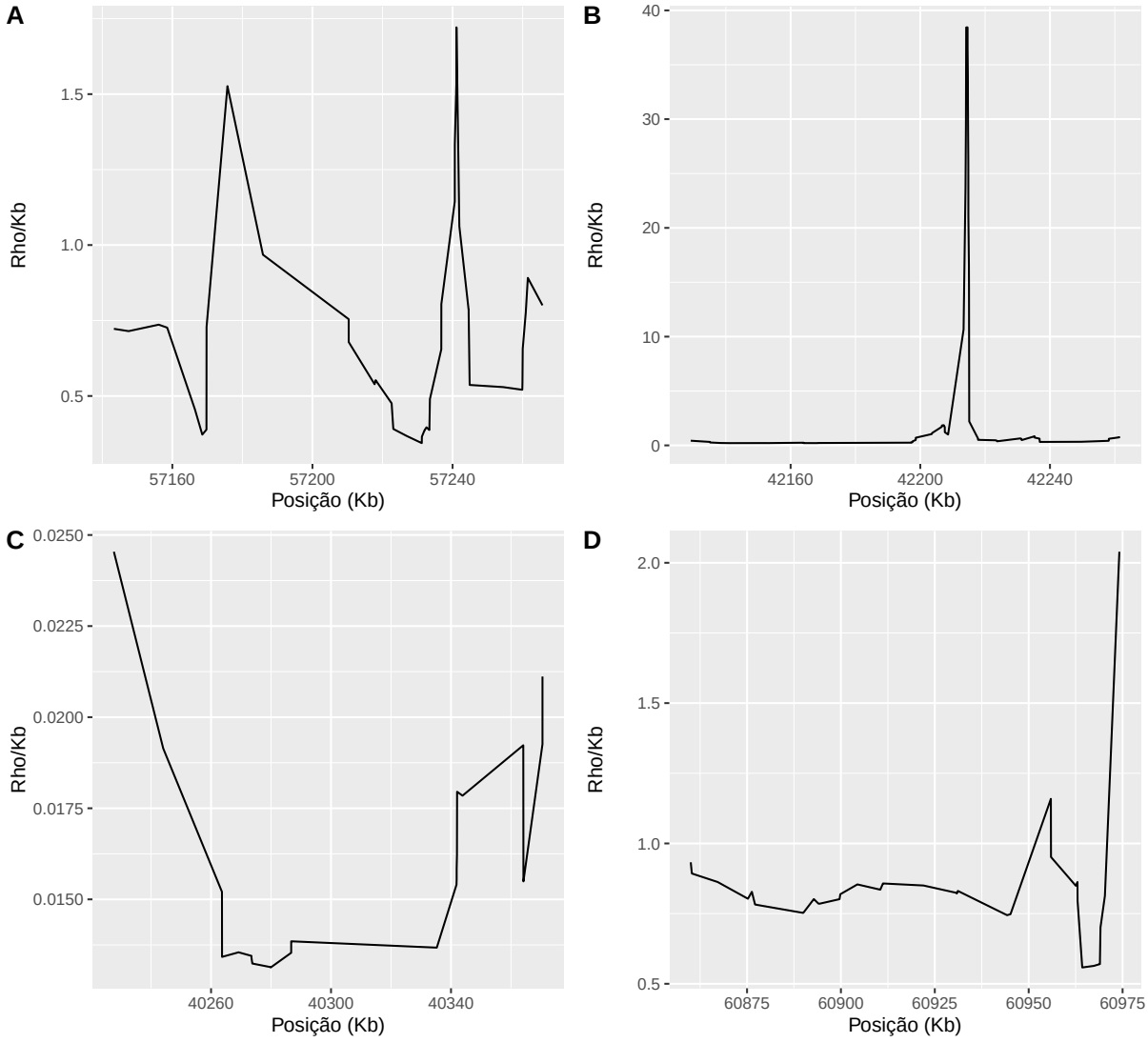

Figura 4.4: Taxas de recombinação (ρ_r/Kb) estimadas via coalescência (`rhomap`) entre todos os pares de SNPs adjacentes para algumas das regiões e populações mostradas na Tabela 4.3. No presente gráfico, todas as taxas de recombinação entre SNPs adjacentes são apresentadas, e não apenas as taxas mínima, média e máxima, como foi feito na referida Tabela. **(A)** Região *Dw1*, População da ETI; **(B)** Região *Dw2*, População dos EUA; **(C)** Região *Ma1*, População do SUD; e **(D)** Região *Ma3*, População da IND. Note a variação de escala mostrada no eixo y, indicando taxas de recombinação bastante distintas de acordo com a região genômica e a população. No caso **(B)**, uma taxa média de $38{,}439$ recombinações por Kb é mostrada entre dois SNPs adjacentes para a população dos EUA na região *Dw2*, sugerindo um possível ponto quente de recombinação (*hotspot*)

As correlações de Spearman entre r^2 e $E[r^2_{\texttt{rhomap}}]$ são mostradas na Tabela 4.4. É possível observar que a associação entre ambas as medidas é muito próxima de zero para todas as regiões genômicas e populações. Embora há vários casos em que r^2 e $E[r^2_{\texttt{rhomap}}]$ mostraram padrões de associação distintos (Figura 4.5), os resultados das correlações pouco refletem aquilo que, em geral, foi verificado nos *heatmaps* de DL. Isto ocorre porque as medidas de correlação são uma estatística de resumo.

Tabela 4.4: Correlações de Spearman entre o desequilíbrio de ligação via medida tradicional (r^2) e coalescência ($E[r^2_{\texttt{rhomap}}]$) para cinco regiões genômicas e sete populações de sorgo, bem como para o subpainel de 298 indivíduos

População	Regiões Genômicas				
	Dw1	*Dw2*	*Dw4*	*Ma1*	*Ma3*
ETI	−0,097	−0,147	0,024	−0,020	0,064
IND	−0,117	0,199	0,052	−0,055	−0,057
NIG	0,015	0,103	0,028	0,038	−0,052
AFS	−0,006	0,054	0,016	−0,077	−0,023
SUD	−0,063	−0,110	−0,122	−0,117	0,156
UGA	−0,023	−0,186	−0,067	−0,034	0,094
EUA	−0,027	0,001	−0,021	0,092	0,036
Subpainel	−0,024	−0,044	−0,013	−0,006	0,071

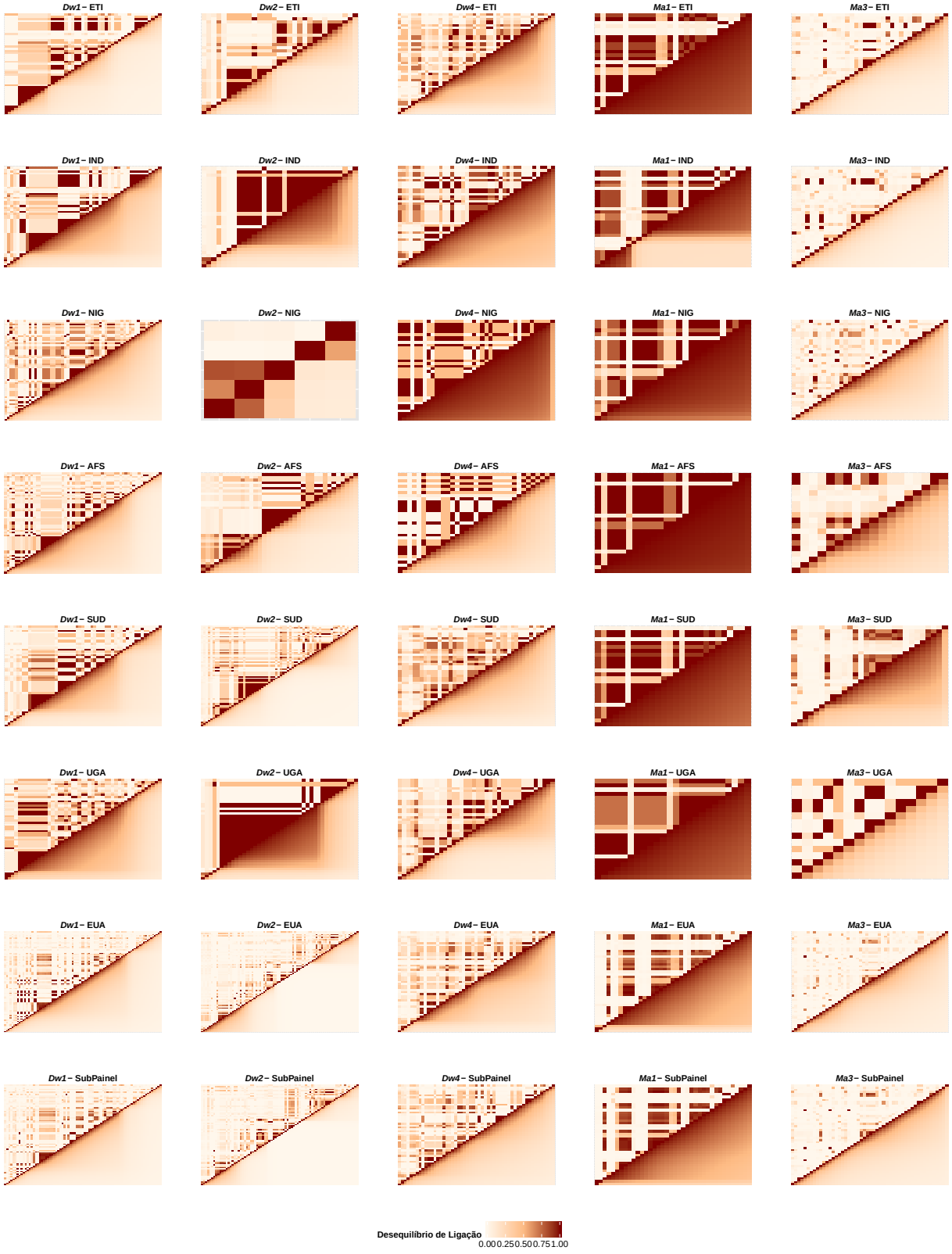

Figura 4.5: Gráfico do desequilíbrio de ligação via medida tradicional (r^2) e coalescência ($E[r^2_{\texttt{rhomap}}]$) para cada região genômica e população de sorgo. Os padrões do r^2 e $E[r^2_{\texttt{rhomap}}]$ são mostrados acima e abaixo da diagonal principal dos gráficos de *heatmap*, respectivamente. O $E[r^2_{\texttt{rhomap}}]$ foi calculado com base nas estimativas de recombinação variável obtidas via `rhomap` (ρ_r), considerando uma população em equilíbrio entre recombinação e deriva genética

5 DISCUSSÃO

No presente estudo, os princípios da teoria da coalescência foram utilizados para investigar a estrutura populacional e o desequilíbrio de ligação de um painel mundial de acessos de sorgo. Para tanto, análises de mutação, migração com fluxo gênico e recombinação foram realizadas para cinco regiões genômicas e sete populações que constituem este painel. Tais investigações foram realizadas com o propósito de fornecer informações relevantes para estudos futuros de mapeamento associativo em sorgo, bem como para um melhor entendimento do próprio painel. Além disso, como objetivo secundário, havia a expectativa de avaliar os modelos e pacotes disponíveis para análises baseadas em coalescência.

Uma das principais pressuposições assumida por grande parte dos modelos de coalescência é que as populações se comportam segundo o modelo proposto por Fisher (1930) e Wright (1931). Assume-se que uma amostra recente de indivíduos foi aleatoriamente tomada a partir de uma grande população panmítica, que é baseada em cruzamentos aleatórios entre os indivíduos. No entanto, o painel aqui utilizado foi construído de maneira a agrupar indivíduos historicamente importantes para o melhoramento do sorgo, não sendo necessariamente uma amostra aleatória de uma população panmítica. Assim, as estimativas aqui obtidas podem ter apresentado certo viés e variâncias elevadas (Hamblin et al., 2004), além do fato de que a taxa de autofecundação em populações cultivadas de sorgo pode ser superior a 90% (Hamblin et al., 2005). Isso significa que os cruzamentos não acontecem, em sua grande maioria, de forma cruzada e aleatória, tendo implicações na forma pela qual a teoria da coalescência é proposta (Fu, 1997; Nordborg and Donnelly, 1997; Hamblin et al., 2005). Portanto, estudos futuros de coalescência em sorgo deveriam levar essa realidade em consideração.

O número reduzido de indivíduos para certas populações pode ter contribuído para a obtenção de menor polimorfismo para as regiões genômicas de interesse. Consequentemente, as estimativas de mutação, fluxo gênico e recombinação podem ter apresentado certo viés em relação ao processo evolutivo da espécie, embora o interesse tenha permanecido especificamente no painel aqui utilizado. Normalmente, o aumento da amostra resulta na elevação logarítmica da acurácia a partir da estimação por máxima verossimilhança (Felsenstein, 2006). No entanto, esse aumento deve ocorrer de forma a elevar o polimorfismo, pois uma quantidade inferior a 20 locos pode resultar em estimativas pouco confiáveis dos parâmetros, principalmente quando o grau de diferenciação entre populações é pequeno (Wilson and Rannala, 2003). No presente estudo, poucas foram as populações que apresentaram quantidades inferiores a 20 SNPs, mas o polimorfismo foi considerado baixo por se tratar de variações de nucleotídeos únicos. Embora os SNPs sejam, em geral, abundantes ao longo do genoma, normalmente são menos informativos quando comparados com outros marcadores moleculares, tais como os microssatélites.

A utilização da técnica de GBS possibilitou a obtenção de dados de sequências para a realização deste trabalho. Muitos dados perdidos foram inicialmente obtidos, e o método de imputação via `NPUTE` foi utilizado para a recuperação dessas informações. Embora grande parte do genoma dos indivíduos deva estar em homozigose, o `NPUTE` desconsiderou a possibilidade de haver locos em heterozigose, o que poder ser algo pouco real. Informações de heterozigotos certamente seriam importantes para se ter maior variação genética, influenciando diretamente na estimação dos parâmetros populacionais. No entanto, acredita-se que, em geral, a técnica de GBS, aliada à imputação via `NPUTE`, forneceu genótipos bastante razoáveis.

Desvios do modelo padrão baseado em neutralismo apenas foram detectados para a região *Ma1*, com

base na estatística *D* (Tajima, 1989). Embora esta região tenha apresentado as menores taxas médias de recombinação populacional via `LAMARC` ($\rho_L = 0{,}030 \times 10^{-3}$/Kb) e `rhomap` (0,060 ρ_r/Kb), sua reduzida variabilidade pode ter sido consequência de um processo de seleção. Tal evidência também foi sugerida por Klein et al. (2008) ao detectarem menor variação na região que contém os locos *Ma1* e *Dw2*, sugerindo, especificamente, a ocorrência de seleção direcio-nal em decorrência das decisões do melhoramento no processo de conversão de linhagens. Contudo, o valor positivo de *D* aqui detectado, indicando a presença de ramos internos mais longos na genealogia, pode sugerir a ocorrência de seleção balanceadora ou até mesmo uma recente mistura entre diferentes populações (Hein et al., 2004).

No entanto, outros processos evolutivos podem ter contribuído para a significância da estatística *D* (Wakeley, 2009; Korneliussen et al., 2013) na região *Ma1*, tais como fluxo gênico e endogamia, mostrando que *D*, e outros testes clássicos de neutralidade, devem pouco esclarecer a história evolutiva ocorrida (Fu and Li, 1999; Wakeley, 2009; Kuhner, 2009). Mais do que isso, é difícil saber de que maneira *D* é afetado na presença de vários processos evolutivos (Hein et al., 2004; Kuhner, 2009). Portanto, deve-se considerar *D* apenas como evidência inicial de que processos mais complexos podem ter ocorrido ao longo das gerações, lembrando que valores não significativos de *D* podem desconsiderar evidências evolutivas importantes (Morrell et al., 2003) devido ao uso restrito da variação.

A estimação de fluxo gênico usando o método proposto por Beerli and Felsenstein (2001) mostrou evidências importantes da história de melhoramento do sorgo. Embora este método seja computacionalmente intensivo (Palczewski and Beerli, 2013) por ser baseado em abordagem completa, as regiões aqui investigadas não apresentaram grandes quantidades de SNPs a ponto de torná-lo inviável. Assim, o fluxo gênico pôde ser estimado a partir de diversas simulações, repetições e diferentes estratégias de busca no contexto bayesiano, na expectativa de que muitas histórias pudessem ser percorridas e supostamente consideradas.

Em geral, elevadas taxas de fluxo gênico foram obtidas para todas as regiões genômicas e todas elas simultaneamente. Isso significa que intenso intercâmbio de acessos deve ter historicamente ocorrido entre as sete populações aqui consideradas, muitas das quais estão situadas em regiões geográficas próximas. No caso dos EUA, que está localizado mais distantemente, o programa de conversão de linhagens deve ter favorecido a entrada de acessos provenientes da África, sendo, em geral, de maior magnitude em comparação à entrada de acessos dos EUA na África. Assim, acredita-se que as taxas de fluxo gênico aqui detectadas retratam eventos migratórios mais recentes, embora parte deles possa ter ocorrido em gerações mais antigas do período de domesticação, envolvendo populações africanas mais próximas. Eventos migratórios mais antigos podem representar a grande motivação do uso da teoria da coalescência em comparação aos métodos clássicos para inferir estrutura populacional. Além disso, o elevado fluxo gênico aqui detectado também pode indicar que as populações foram recentemente formadas a partir de uma mesma origem, a qual deve ter ocorrido na Etiópia e no Sudão.

Embora o fluxo gênico tenha sido considerado como elevado, é preciso ter cuidado com a sua magnitude. Originalmente, o método proposto por Beerli and Felsenstein (2001) utiliza o polimorfismo disponível na amostra para estimar, simultaneamente, as taxas de mutação ($\theta_{L\mathcal{M}}$) e fluxo gênico ($\mathcal{M}$). $\mathcal{M}$ corresponde à razão $\frac{m}{\mu}$ (Beerli and Felsenstein, 1999), mostrando que esta estimativa é dependente da mutação. O produto de $\mathcal{M}$ pelo valor de θ da população recipiente resulta no número médio de migrantes por geração (neste caso, $2N_em$), sendo este independente da mutação (Beerli, 1998). Esta quantidade é preferida porque altos valores de $\mathcal{M}$ podem indicar

tanto elevado fluxo gênico quanto baixa mutação, dificultando uma real interpretação (Beerli, 1998). Contudo, no cenário aqui investigado, as taxas de mutação apresentaram, em geral, grandes intervalos de credibilidade. Isso significa que $2N_em$ poderia ser sub ou superestimado, apresentando intervalos de credibilidade bastante elevados.

Grandes intervalos de credibilidade foram encontrados para as estimativas de fluxo gênico $\mathcal{M}$, sugerindo um cenário bastante complexo. Isso pode ter acontecido pelo fato de que poucos locos estiveram disponíveis para poucas regiões genômicas, reduzindo a variação genética entre populações (Beerli, 1998; Wilson and Rannala, 2003; Beerli, 2006). Com elevadas taxas de fluxo gênico, grande parte das mutações estará presente em todas as populações, de modo que poucos locos poderão ser insuficientes para se ter maior variação e confiança no processo de estimação (Beerli, 2006). Além disso, muitos eventos terão que ser investigados pelo algoritmo de MCMC caminhando-se do presente em direção ao passado, algo que se torna ainda mais complexo à medida que aumenta o número de populações, de indivíduos e, assim, a quantidade de parâmetros (Palczewski and Beerli, 2013). Portanto, a estimação confiável de fluxo gênico entre populações que estiveram em constante e intenso contato é um grande desafio (Faubet et al., 2007; Palczewski and Beerli, 2013).

Outro ponto importante é que a análise de fluxo gênico do presente estudo foi realizada no contexto bayesiano. Embora tal abordagem tem sido recomendada para cenários demográficos complexos (Beaumont and Rannala, 2004; Beerli, 2006), a estimação dos parâmetros pode ser sensivelmente dependente das informações estabelecidas a priori (Stephens, 2007). Mesmo que a distribuição a posteriori não apresente similaridade com a distribuição a priori, o que é indicativo de que variação suficiente está disponível nos dados (Beerli, 2006), a atribuição de informações iniciais é fundamental para a estimação confiável dos parâmetros. No presente estudo, distribuições típicas foram utilizadas como priori para θ (linear) e $\mathcal{M}$ (logarítimica) (Kuhner and Smith, 2007). Porém, o cenário aqui investigado é bastante complexo, sendo recomendado, em trabalhos futuros, a utilização de outras distribuições a priori consideradas importantes, tais como exponencial e uniforme (Beerli, 2006).

A análise de fluxo gênico para cada região individualmente mostrou evidências importantes da história evolutiva do sorgo. Por um lado, o uso simultâneo de múltiplas regiões não ligadas pode fornecer maior variação genética (Hey and Nielsen, 2004; Beerli, 2006) e possibilitar a investigação de processos que afetam todo o genoma (Sousa and Hey, 2013). Por outro, tais regiões podem evidenciar taxas de fluxo gênico variáveis entre as populações, mostrando histórias evolutivas específicas. Na presença de seleção, por exemplo, certas regiões podem experimentar completa ausência de fluxo gênico entre populações (Sousa et al., 2013; Sousa and Hey, 2013). No presente estudo, fluxo gênico assimétrico foi observado entre todos os pares de populações somente quando a análise individual das regiões foi considerada. Portanto, a análise de fluxo gênico para cada região genômica foi muito importante para detectar um cenário evolutivo mais complexo e, possivelmente, mais real.

O método proposto por Beerli and Felsenstein (2001) também foi utilizado em trabalhos anteriores envolvendo espécies domesticadas. Morrell et al. (2003) investigaram o fluxo gênico entre três populações selvagens de cevada (*Hordeum vulgare* ssp. *spontaneum*) utilizando dados de sequência de nove locos genômicos. Em média, um indivíduo migrante por geração foi observado entre todas as populações a partir da análise contendo todos os locos. Porém, taxas de fluxo gênico ($M = 4N_em$) elevadas (11,664, 17,844 e 21,272) foram detectadas na análise para cada loco separadamente, obtendo-se, muitas vezes, grandes intervalos de confiança. Embora os autores tenham concentrado suas conclusões na análise total, resultados específicos foram obtidos para cada loco, algo que também foi observado no presente estudo.

Também para cevada, HÜBNER ET AL. (2012) estimaram a taxa de fluxo gênico entre três populações selvagens e entre populações selvagem e cultivada utilizando marcadores SSRs. Os autores encontraram elevado fluxo gênico entre as três populações selvagens ($2{,}87-8{,}45$) e da população cultivada para a selvagem ($3{,}14$), como consequência de cruzamentos ocasionais e da dispersão de sementes realizada pelo homem e por outros animais.

As análises de coalescência mostraram que as taxas de recombinação variaram de acordo com a região genômica e a população em sorgo. Isso mostra que o valor médio de recombinação ($1{,}4$ ρ/Kb, `rhomap`) sugerido por MORRIS ET AL. (2013) é insuficiente para caracterizar os padrões de recombinação no genoma desta espécie. Os autores argumentam que diferentes taxas foram observadas para regiões centroméricas e teloméricas. No entanto, os resultados aqui obtidos via `rhomap` sugerem que as taxas de recombinação podem ser bastante variáveis para regiões contendo locos importantes para o melhoramento do sorgo, tais como *Dw2* ($3{,}395$ ρ_r/Kb), *Ma1* ($0{,}060$ ρ_r/Kb) e *Ma3* ($0{,}465$ ρ_r/Kb). Embora os dois primeiros estejam fisicamente ligados no cromossomo 6 (LIN ET AL., 1995; KLEIN ET AL., 2008), seus padrões de recombinação sugerem histórias distintas para cada população e o subpainel utilizado.

Em geral, as taxas médias de recombinação constantes e independentes da mutação (ρ_L/Kb), estimadas via `LAMARC` no presente estudo, mostraram estimativas superiores daquelas obtidas por HAMBLIN ET AL. (2005). Esses autores investigaram os padrões de recombinação em sorgo com base nos métodos aproximados de HUDSON (2001) e LI and STEPHENS (2003). Para tanto, seis regiões genômicas não ligadas foram analisadas para um painel constituído de 32 acessos (cultivados e selvagens) de diferentes localidades do mundo. Uma única e constante taxa de recombinação foi obtida para cada uma dessas seis regiões, assim como foi realizado no presente estudo utilizando o método de KUHNER ET AL. (2000). As taxas de recombinação obtidas por HAMBLIN ET AL. (2005) variaram de $2{,}073 - 15{,}660 \times 10^{-6}$/Kb e $0{,}997 - 9{,}957 \times 10^{-6}$/Kb utilizando os métodos de HUDSON (2001) e LI and STEPHENS (2003), respectivamente.

HAMBLIN ET AL. (2005) mostraram que a recombinação foi inferior para acessos cultivados em comparação aos acessos selvagens, considerando cinco das seis regiões estudadas. No entanto, uma comparação entre populações cultivadas também pode ser desejável. O nosso estudo mostrou que as taxas de recombinação entre populações cultivadas podem ser consideravelmente diferentes para certa região. Tal diferença foi aqui baseada no método de AUTON and MCVEAN (2007), que forneceu taxas de recombinação variáveis ao longo de certa região.

O caso mais extremo dessa diferença ocorreu para a população dos EUA ($38{,}439$ ρ/Kb) e o Subpainel ($76{,}320$ ρ/Kb) na região *Dw2*. Grande parte dos eventos de recombinação aqui detectados deve ter ocorrido recentemente no contexto de coalescência, já que não há indivíduos selvagens de períodos anteriores à domesticação. Assim, acredita-se que as taxas elevadas de tais populações correspondam a pontos quentes de recombinação (*hotspots*), com sítios adjacentes mostrando, em média, taxas próximas àquelas verificadas para as demais regiões. Taxa de recombinação elevada também foi observada para a população da NIG ($6{,}729$ ρ/Kb) na região *Dw2*, embora tal valor possa ter sido subestimado pela baixa quantidade de SNPs.

Estimativas de recombinação via coalescência também foram obtidas para outras espécies autógamas. Em *Arabidopsis thaliana*, NORDBORG ET AL. (2005) e KIM ET AL. (2007) encontraram taxas médias de $0{,}200$ e $0{,}800$ ρ/Kb para todo o genoma, respectivamente. KIM ET AL. (2007) observaram taxas mais elevadas ($1{,}500$ ρ/Kb) a curtas distâncias ($0 - 5$ Kb), obtendo-se valores próximos à média ($0{,}800$ ρ/Kb) a partir de 5 Kb. Os autores sugeriram a ocorrência de conversão gênica para explicar tais padrões, que evidenciam certa discrepância entre DL

de curtas e longas distâncias. Em sorgo, poucas evidências de recombinação foram observadas a curtas distâncias (Caniato et al., 2014), sugerindo que a conversão gênica não deve explicar discrepâncias de associação para este espécie (Hamblin et al., 2005).

Em cevada, Morrell et al. (2003) observaram taxas de recombinação variando de $0{,}094-0{,}511$/intersítio para nove genes de importância. Esses autores utilizaram o mesmo método de Kuhner et al. (2000) investigado no presente estudo, com o qual uma única e constante taxa de recombinação é estimada para certa região. Morrell et al. (2003) obtiveram as taxas de recombinação dependentes da mutação ($\frac{r}{\mu}$) para cada um dos nove genes, e os valores observados foram bastante elevados. As taxas médias de recombinação dependentes da mutação aqui detectadas variaram de acordo com a região, sendo muito inferiores às taxas obtidas por Morrell et al. (2003). A região que mostrou o maior valor médio de $\frac{r}{\mu}$/Kb foi a *Dw2* ($1{,}578\times10^{-3}$), seguida das regiões *Ma3* ($1{,}209\times10^{-3}$), *Dw1* ($0{,}347\times10^{-3}$), *Ma1* ($0{,}213\times10^{-3}$) e *Dw4* ($0{,}129\times10^{-3}$). Claramente, os genes estudados por Morrell et al. (2003) são bem menores que as regiões aqui investigadas, contendo, possivelmente, menor polimorfismo. Assim, como esses autores se basearam apenas nas taxas de recombinação dependentes da mutação, é possível que altas estimativas de recombinação tenham sido obtidas apenas em função de menores taxas de mutação.

As análises de fluxo gênico e recombinação não foram realizadas simultaneamente no presente estudo. Embora seja possível investigar tais processos em um único procedimento de análise no `LAMARC` (Morrell et al., 2003), as estimações não são realizadas com base em um modelo conjunto. Isso porque fluxo gênico e recombinação foram propostos por modelos de coalescência independentes, que consideram apenas a taxa de mutação como processo comum. Portanto, é esperado que as estimativas obtidas em ambos os casos apresentem certo viés, algo que pode ser maior ou menor à medida que outros processos evolutivos se tornam importantes para explicar a variação dos dados.

Os padrões do DL via r^2 e $E[r^2_{\texttt{rhomap}}]$ mostraram resultados muito próximos para algumas regiões e populações. No entanto, a medida tradicional r^2 mostrou certa descontinuidade de possíveis blocos de associação, principalmente quando houve forte DL entre os locos. Tal comportamento pode sugerir, muitas vezes, interpretações pouco precisas da estrutura do DL dentro de certa região. Hamblin et al. (2005) também observaram padrões pouco conclusivos do DL em sorgo com base em tal medida. Esses resultados reforçam a ideia de que o r^2 é fortemente influenciado pelo tamanho da amostra e/ou pela baixa frequência alélica (Flint-Garcia et al., 2003), fornecendo estimativas variáveis e pouco precisas de associação (Pritchard and Przeworski, 2001; Nordborg and Tavaré, 2002). Acredita-se que o r^2 pode ser útil apenas para indicar que certo DL está presente em determinada região, fornecendo padrões pouco refinados de associação (Evans and Cardon, 2005).

Em contrapartida, o DL obtido via coalescência sugeriu padrões que parecem fazer sentido para uma espécie autógama. Para muitos casos, grandes blocos de haplótipos, bem definidos e estruturados, foram observados, mostrando um extenso DL entre locos fisicamente ligados no genoma. Tais blocos não foram observados para todas as populações de certa região, mostrando que os padrões do DL foram específicos para cada população. Isso sugere que um único *pool* gênico pode ser insuficiente para a realização de mapeamento associativo em sorgo, algo que também foi sugerido em estudo anterior com humanos (Evans and Cardon, 2005).

No entanto, a abordagem do $E[r^2]$ aqui utilizada (equação 2.30) foi proposta considerando que a amostra populacional é muito grande. Claramente, algumas das populações investigadas não apresentaram grande quantidade de indivíduos, possivelmente por questões de disponibilidade ou recursos de captação. Isso ocorreu

principalmente para as populações da AFS e de UGA. Então, embora as menores amostras estejam próximas às quantidades que são normalmente utilizadas em estudos de coalescência, o $E[r^2]$ pode ter sido inflacionado por questões de amostragem. Tal ponto é muito importante, e deve ser considerado em estudos futuros.

Além disso, embora o DL via coalescência tenha fornecido padrões importantes, apenas a recombinação foi considerada em seu cálculo. Em geral, todas as regiões tiveram elevado fluxo gênico entre as populações. O número médio de indivíduos que entraram e saíram de certa população, considerando o cenário complexo aqui investigado, deve ter influenciado nos padrões históricos do DL obtidos. Da mesma forma, o processo de seleção balanceadora, sugerido para a região *Ma1*, também pode ter influenciado nos padrões elevados do DL. Portanto, é desejável que a interpretação genealógica do DL (McVean, 2002) seja estendida para contemplar vários processos evolutivos simultaneamente, tais como fluxo gênico, recombinação e seleção. Para uma espécie autógama tal como o sorgo, desvios do equilíbrio panmítico são esperados (Hamblin et al., 2005), e isso também deveria ser levado em consideração. Espera-se que uma estimação mais precisa do DL via coalescência possa fornecer informações mais refinadas para estudos de mapeamento associativo em sorgo.

No momento em que grandes quantidades de dados estão sendo geradas a partir do genoma de várias espécies, métodos de coalescência aproximados podem ser utilizados para investigar diferentes cenários evolutivos. Embora o uso eficiente de tais métodos, como o ABC (*Approximate Bayesian Computation*) e o PAC (*Product of Approximate Conditionals*), seja algo ainda discutível, acredita-se que abordagens aproximadas estarão cada vez mais presentes no contexto de coalescência para múltiplos locos (McVean and Cardin, 2005). Contudo, o grande desafio continuará sendo o desenvolvimento de abordagens completas mais eficientes (Sethuraman and Hey, 2016) e que consideram vários processos simultaneamente (Sousa and Hey, 2013), na busca de um entendimento mais profundo dos processos evolutivos.

6 CONSIDERAÇÕES FINAIS

No presente estudo, a estrutura populacional e o desequilíbrio de ligação foram investigados com base na teoria da coalescência. Para tanto, análises de mutação, migração com fluxo gênico e recombinação foram realizadas para cinco regiões genômicas e sete populações de sorgo, e resultados muito interessantes foram encontrados.

A experiência de ter tido que compreender os princípios da teoria da coalescência e utilizar pacotes especificamente desenvolvidos para tal abordagem foi incrível. As análises de fluxo gênico e recombinação constante ao longo da região foram realizadas com o pacote `LAMARC`. Trata-se de um software muito bem documentado e com informações suficientes para o uso adequado. Contudo, detalhes importantes para a correta interpretação dos resultados não são muito bem esclarecidos ou são apresentados de forma isolada em um ou outro arquivo, dificultando o entendimento. Além disso, por utilizar abordagem de verossimilhança completa, tal pacote demanda enorme tempo computacional, podendo se tornar inviável para casos em que muitos dados encontram-se disponíveis. O programa `rhomap`, disponível no pacote `LDhat`, foi utilizado para análises de recombinação variável ao longo da região. Trata-se de um pacote com documentação necessária para o uso correto, porém com poucas explicações para iniciantes. Foi especificamente desenvolvido para o uso da teoria da coalescência a partir de muitos dados, podendo ser uma alternativa interessante para estimar recombinação em cenários mais atuais.

Em geral, elevado fluxo gênico foi observado entre todas as populações de sorgo a partir das análises de cada região genômica e todas elas simultaneamente. Além disso, os resultados sugeriram que a história evolutiva de regiões específicas do genoma desta espécie pode ser distinta, até mesmo quando processos que influenciam todo o genoma estão sendo estudados. Ainda, o fluxo gênico mostrou ser assimétrico entre todos os possíveis pares de populações, sugerindo que a forma pela qual as populações se relacionaram e continuam se interagindo evolutivamente não é igual. Este resultado evidencia que os métodos clássicos para investigar a estrutura populacional podem ser insatisfatórios para explicar o processo histórico-evolutivo.

Utilizando o método proposto por KUHNER ET AL. (2000), que considera taxas de recombinação constantes para certa região (`LAMARC`), baixas estimativas foram observadas para todas as regiões genômicas e populações. No entanto, utilizando o método proposto por AUTON and MCVEAN (2007), que considera taxas de recombinação variáveis (`rhomap`), altas e baixas estimativas foram observadas de acordo com a região genômica e a população. A magnitude de recombinação para certas populações foi bastante elevada, sendo um resultado inesperado em sorgo.

Os padrões do DL via medida tradicional (r^2) e coalescência ($E[r^2_{\texttt{rhomap}}]$) mostraram resultados próximos em alguns casos. No entanto, o r^2 apresentou padrões descontínuos em várias ocasiões, dificultando o entendimento e a caracterização dos blocos de associação. Em contrapartida, o DL via coalescência ($E[r^2_{\texttt{rhomap}}]$) forneceu resultados que pareceram ter sido mais consistentes, sendo uma estratégia eventualmente importante para um refinamento dos padrões de associação. O DL de alta ordem via coalescência ($E[r^2_{\texttt{LAMARC}}]$) forneceu estimativas bastante elevadas para todas as regiões e populações, um resultado muito diferente dos padrões observados por r^2 e $E[r^2_{\texttt{rhomap}}]$.

Acredita-se que os resultados aqui encontrados possam fornecer informações relevantes para estudos futuros de mapeamento associativo em sorgo. Embora o contato entre as sete populações aqui estabelecidas tenha

sido historicamente intenso, os padrões do DL mostraram resultados específicos de acordo com a região genômica e a população. Tais resultados sugerem que o mapeamento genético a partir de um único *pool* gênico pode ser insuficiente para a detecção de associações causais importantes para características quantitativas em sorgo.

REFERÊNCIAS

Abdurakhmonov, I. Y. and A. Abdukarimov, 2008 Application of association mapping to understanding the genetic diversity of plant germplasm resources. International journal of plant genomics **2008**: 1–18, doi:10.1155/2008/574927.

Agrama, H. and M. Tuinstra, 2004 Phylogenetic diversity and relationships among sorghum accessions using SSRs and RAPDs. African Journal of Biotechnology **2**: 334–340.

Aldrich, P. and J. Doebley, 1992 Restriction fragment variation in the nuclear and chloroplast genomes of cultivated and wild *Sorghum bicolor*. Theoretical and Applied Genetics **85**: 293–302.

Aldrich, P., J. Doebley, K. Schertz, and A. Stec, 1992 Patterns of allozyme variation in cultivated and wild *Sorghum bicolor*. Theoretical and Applied Genetics **85**: 451–460.

Allendorf, F. W. and G. H. Luikart, 2007 Conservation and the genetics of populations. Wiley. com.

Ardlie, K. G., L. Kruglyak, and M. Seielstad, 2002 Patterns of linkage disequilibrium in the human genome. Nature Reviews Genetics **3**: 299–309.

Auton, A. and G. McVean, 2007 Recombination rate estimation in the presence of hotspots. Genome research **17**: 1219–1227.

Bahlo, M. and R. Griffiths, 2000 Inference from gene trees in a subdivided population. Theoretical population biology **57**: 79–95.

Beaumont, M. A. and B. Rannala, 2004 The Bayesian revolution in genetics. Nature Reviews Genetics **5**: 251–261.

Becquet, C. and M. Przeworski, 2007 A new approach to estimate parameters of speciation models with application to apes. Genome research **17**: 1505–1519.

Beerli, P., 1998 Estimation of migration rates and population sizes in geographically structured populations. NATO ASI SERIES A LIFE SCIENCES **306**: 39–54.

Beerli, P., 2006 Comparison of bayesian and maximum-likelihood inference of population genetic parameters. Bioinformatics **22**: 341–345.

Beerli, P. and J. Felsenstein, 1999 Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics **152**: 763–773.

Beerli, P. and J. Felsenstein, 2001 Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proceedings of the National Academy of Sciences **98**: 4563–4568.

Beerli, P. and M. Palczewski, 2010 Unified framework to evaluate panmixia and migration direction among multiple sampling locations. Genetics **185**: 313–326.

Beissinger, T. M., C. N. Hirsch, R. S. Sekhon, J. M. Foerster, J. M. Johnson, G. Muttoni, B. Vaillancourt, C. R. Buell, S. M. Kaeppler, and N. de Leon, 2013 Marker density and read depth for genotyping populations using genotyping-by-sequencing. Genetics **193**: 1073–1081.

Bhosale, S. U., B. Stich, H. F. W. Rattunde, E. Weltzien, B. I. Haussmann, C. T. Hash, P. Ramu, H. E. Cuevas, A. H. Paterson, A. E. Melchinger, et al., 2012 Association analysis of photoperiodic flowering time genes in west and central african sorghum [*Sorghum bicolor* (L.) Moench]. BMC plant biology **12**: 32.

Bouchet, S., D. Pot, M. Deu, J.-F. Rami, C. Billot, X. Perrier, R. Rivallan, L. Gardes, L. Xia, P. Wenzl, et al., 2012 Genetic structure, linkage disequilibrium and signature of selection in sorghum: lessons from physically anchored dart markers. PLoS One **7**: e33470.

Brown, P. J., W. L. Rooney, C. Franks, and S. Kresovich, 2008 Efficient mapping of plant height quantitative trait loci in a sorghum association population with introgressed dwarfing genes. Genetics **180**: 629–637.

Caniato, F. F., M. T. Hamblin, C. T. Guimaraes, Z. Zhang, R. E. Schaffert, L. V. Kochian, and J. V. Magalhaes, 2014 Association mapping provides insights into the origin and the fine structure of the sorghum aluminum tolerance locus, altsb. PloS One **9**: e87438.

Casa, A., S. Mitchell, M. Hamblin, H. Sun, J. Bowers, A. Paterson, C. Aquadro, and S. Kresovich, 2005 Diversity and selection in sorghum: simultaneous analyses using simple sequence repeats. Theoretical and Applied Genetics **111**: 23–30.

Casa, A. M., S. E. Mitchell, J. D. Jensen, M. T. Hamblin, A. H. Paterson, C. F. Aquadro, and S. Kresovich, 2006 Evidence for a selective sweep on chromosome 1 of cultivated sorghum. Crop science **46**: S–27.

Casa, A. M., G. Pressoir, P. J. Brown, S. E. Mitchell, W. L. Rooney, M. R. Tuinstra, C. D. Franks, and S. Kresovich, 2008 Community resources and strategies for association mapping in sorghum. Crop science **48**: 30–40.

Chantereau, J., G. Trouche, J. Rami, M. Deu, C. Barro, and L. Grivet, 2001 RFLP mapping of QTLs for photoperiod response in tropical sorghum. Euphytica **120**: 183–194.

Childs, K. L., F. R. Miller, M.-M. Cordonnier-Pratt, L. H. Pratt, P. W. Morgan, and J. E. Mullet, 1997 The sorghum photoperiod sensitivity gene, Ma3, encodes a phytochrome B. Plant Physiology **113**: 611–619.

Crasta, O., W. Xu, D. Rosenow, J. Mullet, and H. Nguyen, 1999 Mapping of post-flowering drought resistance traits in grain sorghum: association between QTLs influencing premature senescence and maturity. Molecular and General Genetics MGG **262**: 579–588.

Davey, J. W., P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen, and M. L. Blaxter, 2011 Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics **12**: 499–510.

de Alencar Figueiredo, L., C. Calatayud, C. Dupuits, C. Billot, J.-F. Rami, D. Brunel, X. Perrier, B. Courtois, M. Deu, and J.-C. Glaszmann, 2008 Phylogeographic evidence of crop neodiversity in sorghum. Genetics **179**: 997–1008.

De Wet, J., 1978 Systematics and evolution of Sorghum sect. Sorghum (Gramineae). American Journal of Botany pp. 477–484.

De Wet, J. M. and J. Harlan, 1971 The origin and domestication of *Sorghum bicolor*. Economic Botany **25**: 128–135.

Deu, M., F. Rattunde, and J. Chantereau, 2006 A global view of genetic diversity in cultivated sorghums using a core collection. Genome **49**: 168–180.

Deu, M., F. Sagnard, J. Chantereau, C. Calatayud, D. Hérault, C. Mariac, J.-L. Pham, Y. Vigouroux, I. Kapran, P. Traore, et al., 2008 Niger-wide assessment of in situ sorghum genetic diversity with microsatellite markers. Theoretical and Applied Genetics **116**: 903–913.

Doggett, H. et al., 1988 Sorghum.. Number 2. ed., Longman Scientific and Technical.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell, 2011 A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS One **6**: e19379.

Evans, D. M. and L. R. Cardon, 2005 A comparison of linkage disequilibrium patterns and estimated population recombination rates across multiple populations. The American Journal of Human Genetics **76**: 681–687.

Faubet, P., R. S. Waples, and O. E. Gaggiotti, 2007 Evaluating the performance of a multilocus bayesian method for the estimation of migration rates. Molecular Ecology **16**: 1149–1166.

Fearnhead, P. and P. Donnelly, 2001 Estimating recombination rates from population genetic data. Genetics **159**: 1299–1318.

Fearnhead, P. and P. Donnelly, 2002 Approximate likelihood methods for estimating local recombination rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **64**: 657–680.

Felsenstein, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of molecular evolution **17**: 368–376.

Felsenstein, J., 1988 Phylogenies from molecular sequences: inference and reliability. Annual review of genetics **22**: 521–565.

Felsenstein, J., 1992 Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genetical research **59**: 139–147.

Felsenstein, J., 2006 Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci? Molecular biology and evolution **23**: 691–700.

Felsenstein, J. and G. A. Churchill, 1996 A hidden Markov model approach to variation among sites in rate of evolution. Molecular Biology and Evolution **13**: 93–104.

Felsenstein, J., M. K. Kuhner, J. Yamato, and P. Beerli, 1999 Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data. Lecture Notes-Monograph Series pp. 163–185.

Feltus, F., G. Hart, K. Schertz, A. Casa, S. Kresovich, S. Abraham, P. Klein, P. Brown, and A. Paterson, 2006 Alignment of genetic maps and QTLs between inter-and intra-specific sorghum populations. Theoretical and applied genetics **112**: 1295–1305.

Fisher, R., 1930 The genetical theory of natural selection.

Fisher, R. A., 1935 The logic of inductive inference. Journal of the Royal Statistical Society pp. 39–82.

Fisher, R. A. et al., 1922 On the dominance ratio. Proceedings of the royal society of Edinburgh **42**: 321–341.

Flint-Garcia, S. A., J. M. Thornsberry, and E. S. IV Buckler, 2003 Structure of linkage disequilibrium in plants*. Annual Review of Plant Biology **54**: 357–374.

Folkertsma, R. T., H. F. W. Rattunde, S. Chandra, G. S. Raju, and C. T. Hash, 2005 The pattern of genetic diversity of guinea-race *Sorghum bicolor* (L.) Moench landraces as revealed with SSR markers. Theoretical and Applied Genetics **111**: 399–409.

Frisse, L., R. Hudson, A. Bartoszewicz, J. Wall, J. Donfack, and A. Di Rienzo, 2001 Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. The American Journal of Human Genetics **69**: 831–843.

Fu, Y.-X., 1997 Coalescent theory for a partially selfing population. Genetics **146**: 1489–1499.

Fu, Y.-X., 1998 Probability of a segregating pattern in a sample of DNA sequences. Theoretical population biology **54**: 1–10.

Fu, Y.-X. and W.-H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133**: 693–709.

Fu, Y.-X. and W.-H. Li, 1999 Coalescing into the 21st century: an overview and prospects of coalescent theory. Theoretical Population Biology **56**: 1–10.

Futuyma, D. J., 2005 Evolution. sunderland, ma. USA: Sinauer Associates. Gouy MAM, Guindon S, Gascuel O (2010). SeaView version **4**: 221–224.

Garber, E. D. et al., 1950 Cytotaxonomic studies in the genus Sorghum. University of California Publications in Botany **23**: 283–362.

Geyer, C. J., 1991 Markov chain monte carlo maximum likelihood. Interface Foundation of North America. Retrieved from the University of Minnesota Digital Conservancy, http://purl.umn.edu/58440.

Glaubitz, J. C., T. M. Casstevens, F. Lu, J. Harriman, R. J. Elshire, Q. Sun, and E. S. Buckler, 2014 TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS One **9**: E90346.

Green, P. J., 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82**: 711–732.

Griffiths, R. and S. Tavaré, 1994a Ancestral inference in population genetics. Statistical Science pp. 307–319.

GRIFFITHS, R. and S. TAVARÉ, 1994b Simulating probability distributions in the coalescent. Theoretical Population Biology **46**: 131–159.

GRIFFITHS, R. and S. TAVARÉ, 1995 Unrooted genealogical tree probabilities in the infinitely-many-sites model. Mathematical biosciences **127**: 77–98.

GRIFFITHS, R. C., 1981 Neutral two-locus multiple allele models with recombination. Theoretical Population Biology **19**: 169–186.

GRIFFITHS, R. C. and P. MARJORAM, 1996 Ancestral inference from samples of DNA sequences with recombination. Journal of Computational Biology **3**: 479–502.

GRIFFITHS, R. C. and S. TAVARÉ, 1994c Sampling theory for neutral alleles in a varying environment. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences **344**: 403–410.

GUPTA, P. K., S. RUSTGI, and P. L. KULWAL, 2005 Linkage disequilibrium and association studies in higher plants: present status and future prospects. Plant molecular biology **57**: 461–485.

HAHN, M. W., 2007 Detecting natural selection on cis-regulatory DNA. Genetica **129**: 7–18.

HAHN, M. W., 2008 Toward a selection theory of molecular evolution. Evolution **62**: 255–265.

HAMBLIN, M. T., A. M. CASA, H. SUN, S. C. MURRAY, A. H. PATERSON, C. F. AQUADRO, and S. KRESOVICH, 2006 Challenges of detecting directional selection after a bottleneck: lessons from *Sorghum bicolor*. Genetics **173**: 953–964.

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

HAMBLIN, M. T., S. E. MITCHELL, G. M. WHITE, J. GALLEGO, R. KUKATLA, R. A. WING, A. H. PATERSON, and S. KRESOVICH, 2004 Comparative population genetics of the panicoid grasses: sequence polymorphism, linkage disequilibrium and selection in a diverse sample of *Sorghum bicolor*. Genetics **167**: 471–483.

HAMBLIN, M. T., M. G. SALAS FERNANDEZ, M. R. TUINSTRA, W. L. ROONEY, and S. KRESOVICH, 2007 Sequence variation at candidate loci in the starch metabolism pathway in sorghum: prospects for linkage disequilibrium mapping. Crop Science **47**: S–125.

HAMILTON, M., 2009 Population genetics. Wiley. com.

HARLAN, J. and J. DE WET, 1972a A simplified classification of cultivated sorghum. Crop Science **12**: 172–176.

HARLAN, J. and J. DE WET, 1972b A simplified classification of cultivated sorghum. Crop Science **12**: 172–176.

HARRIS, K., P. SUBUDHI, A. BORRELL, D. JORDAN, D. ROSENOW, H. NGUYEN, P. KLEIN, R. KLEIN, and J. MULLET, 2007 Sorghum stay-green QTL individually reduce post-flowering drought-induced leaf senescence. Journal of experimental botany **58**: 327–338.

HART, G., K. SCHERTZ, Y. PENG, and N. SYED, 2001 Genetic mapping of *Sorghum bicolor* (L.) Moench QTLs that control variation in tillering and other morphological characters. Theoretical and Applied Genetics **103**: 1232–1242.

HARTL, D. L. and A. G. CLARK, 2007 Principles of population genetics, volume 116. Sinauer associates Sunderland.

HASTINGS, W. K., 1970 Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**: 97–109.

HEDRICK, P. W., 2010 Genetics of populations. Jones & Bartlett Publishers.

HEIN, J., M. SCHIERUP, and C. WIUF, 2004 Gene genealogies, variation and evolution: a primer in coalescent theory. Oxford university press.

HEY, J., 2010 Isolation with migration models for more than two populations. Molecular biology and evolution **27**: 905–920.

HEY, J. and R. NIELSEN, 2004 Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics **167**: 747–760.

HEY, J. and R. NIELSEN, 2007 Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proceedings of the National Academy of Sciences **104**: 2785–2790.

HILL, W. and A. ROBERTSON, 1968 Linkage disequilibrium in finite populations. Theoretical and Applied Genetics **38**: 226–231.

HOLSINGER, K. E. and B. S. WEIR, 2009 Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. Nature Reviews Genetics **10**: 639–650.

HÜBNER, S., T. GÜNTHER, A. FLAVELL, E. FRIDMAN, A. GRANER, A. KOROL, and K. J. SCHMID, 2012 Islands and streams: clusters and gene flow in wild barley populations from the levant. Molecular ecology **21**: 1115–1129.

HUDSON, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theoretical population biology **23**: 183–201.

HUDSON, R. R., 1985 The sampling distribution of linkage disequilibrium under an infinite allele model without selection. Genetics **109**: 611–631.

HUDSON, R. R., 1987 Estimating the recombination parameter of a finite population model without selection. Genetical research **50**: 245–250.

HUDSON, R. R., 1990 Gene genealogies and the coalescent process. Oxford surveys in evolutionary biology **7**: 44.

HUDSON, R. R., 2001 Two-locus sampling distributions and their application. Genetics **159**: 1805–1817.

JORDAN, D., E. S. MACE, R. HENZELL, P. KLEIN, and R. KLEIN, 2010 Molecular mapping and candidate gene identification of the Rf2 gene for pollen fertility restoration in sorghum [*Sorghum bicolor* (L.) Moench]. Theoretical and applied genetics **120**: 1279–1287.

Kaplan, N. and R. R. Hudson, 1985 The use of sample genealogies for studying a selectively neutral m-loci model with recombination. Theoretical population biology **28**: 382–396.

Karlin, S. and J. McGregor, 1967 The number of mutant forms maintained in a population. In Proceedings of the Fifth Berkeley Symposium on mathematics, Statistics and probability, volume 4, pp. 415–438.

Karper, R., 1932 A dominant mutation of frequent recurrence in sorghum. American Naturalist pp. 511–529.

Kebede, H., P. Subudhi, D. Rosenow, and H. Nguyen, 2001 Quantitative trait loci influencing drought tolerance in grain sorghum (*Sorghum bicolor* L. Moench). Theoretical and Applied Genetics **103**: 266–276.

Kim, S., V. Plagnol, T. T. Hu, C. Toomajian, R. M. Clark, S. Ossowski, J. R. Ecker, D. Weigel, and M. Nordborg, 2007 Recombination and linkage disequilibrium in *Arabidopsis thaliana*. Nature genetics **39**: 1151–1155.

Kim, Y., S. Feng, and Z.-B. Zeng, 2008 Measuring and partitioning the high-order linkage disequilibrium by multiple order Markov chains. Genetic epidemiology **32**: 301–312.

Kimmel, G., R. M. Karp, M. I. Jordan, and E. Halperin, 2008 Association mapping and significance estimation via the coalescent. The American Journal of Human Genetics **83**: 675–683.

Kimura, M., 1955 Solution of a process of random genetic drift with a continuous model. Proceedings of the National Academy of Sciences of the United States of America **41**: 144.

Kimura, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61**: 893.

Kingman, J., 1982a Exchangeability and the evolution of large populations.

Kingman, J. F., 1982b The coalescent. Stochastic processes and their applications **13**: 235–248.

Kingman, J. F., 1982c On the genealogy of large populations. Journal of Applied Probability pp. 27–43.

Kingman, J. F., 2000 Origins of the coalescent: 1974-1982. Genetics **156**: 1461–1463.

Klein, R., P. Klein, A. Chhabra, J. Dong, S. Pammi, K. Childs, J. Mullet, W. Rooney, and K. Schertz, 2001 Molecular mapping of the Rf1 gene for pollen fertility restoration in sorghum (*Sorghum bicolor* L.). Theoretical and applied genetics **102**: 1206–1212.

Klein, R. R., J. E. Mullet, D. R. Jordan, F. R. Miller, W. L. Rooney, M. A. Menz, C. D. Franks, and P. E. Klein, 2008 The effect of tropical sorghum conversion and inbred development on genome diversity as revealed by high-resolution genotyping. Crop science **48**: S–12.

Korneliussen, T. S., I. Moltke, A. Albrechtsen, and R. Nielsen, 2013 Calculation of Tajimas D and other neutrality test statistics from low depth next-generation sequencing data. BMC bioinformatics **14**: 289.

Kreitman, M., 2000 Methods to detect selection in populations with applications to the human. Annual review of genomics and human genetics **1**: 539–559.

Kuhner, M. K., 2006 LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. Bioinformatics **22**: 768–770.

Kuhner, M. K., 2009 Coalescent genealogy samplers: windows into population history. Trends in Ecology & Evolution **24**: 86–93.

Kuhner, M. K. and L. P. Smith, 2007 Comparing likelihood and bayesian coalescent estimation of population parameters. Genetics **175**: 155–165.

Kuhner, M. K., J. Yamato, and J. Felsenstein, 1995 Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics **140**: 1421–1430.

Kuhner, M. K., J. Yamato, and J. Felsenstein, 1998 Maximum likelihood estimation of population growth rates based on the coalescent. Genetics **149**: 429–434.

Kuhner, M. K., J. Yamato, and J. Felsenstein, 2000 Maximum likelihood estimation of recombination rates from population data. Genetics **156**: 1393–1401.

Lewontin, R., 1964 The interaction of selection and linkage. i. general considerations; heterotic models. Genetics **49**: 49.

Lewontin, R. and K.-i. Kojima, 1960 The evolutionary dynamics of complex polymorphisms. Evolution pp. 458–472.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, et al., 2009 The sequence alignment/map format and SAMtools. Bioinformatics **25**: 2078–2079.

Li, N. and M. Stephens, 2003 Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics **165**: 2213–2233.

Lin, Y.-R., K. F. Schertz, and A. H. Paterson, 1995 Comparative analysis of QTLs affecting plant height and maturity across the poaceae, in reference to an interspecific sorghum population. Genetics **141**: 391.

Lohse, K., R. Harrison, and N. H. Barton, 2011 A general method for calculating likelihoods under the coalescent process. Genetics **189**: 977–987.

Mace, E. and D. Jordan, 2010 Location of major effect genes in sorghum (*Sorghum bicolor* (L.) Moench). Theoretical and applied genetics **121**: 1339–1356.

Magalhaes, J. V., D. F. Garvin, Y. Wang, M. E. Sorrells, P. E. Klein, R. E. Schaffert, L. Li, and L. V. Kochian, 2004 Comparative mapping of a major aluminum tolerance gene in sorghum and other species in the poaceae. Genetics **167**: 1905–1914.

Mailund, T., A. E. Halager, M. Westergaard, J. Y. Dutheil, K. Munch, L. N. Andersen, G. Lunter, K. Prüfer, A. Scally, A. Hobolth, et al., 2012 A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. PLoS Genet **8**: e1003125.

Marchini, J. and B. Howie, 2010 Genotype imputation for genome-wide association studies. Nature Reviews Genetics **11**: 499–511.

Marjoram, P. and P. Joyce, 2011 Practical implications of coalescent theory. Problem Solving Handbook in Computational Biology and Bioinformatics **1**: 63.

Marjoram, P. and S. Tavaré, 2006 Modern computational approaches for analysing molecular genetic variation data. Nature Reviews Genetics **7**: 759–770.

McVean, G., 2007a Linkage disequilibrium, recombination and selection. Handbook of Statistical Genetics, Third Edition pp. 909–944.

McVean, G., 2007b The structure of linkage disequilibrium around a selective sweep. Genetics **175**: 1395–1406.

McVean, G., P. Awadalla, and P. Fearnhead, 2002 A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics **160**: 1231–1241.

McVean, G. A., 2002 A genealogical interpretation of linkage disequilibrium. Genetics **162**: 987–991.

McVean, G. A. and N. J. Cardin, 2005 Approximating the coalescent with recombination. Philosophical Transactions of the Royal Society B: Biological Sciences **360**: 1387–1393.

McVean, G. A., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley, and P. Donnelly, 2004 The fine-scale structure of recombination rate variation in the human genome. Science **304**: 581–584.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, 1953 Equation of state calculations by fast computing machines. The journal of chemical physics **21**: 1087.

Morden, C., J. Doebley, and K. Schertz, 1990 Allozyme variation among the spontaneous species of Sorghum section Sorghum (Poaceae). Theoretical and Applied Genetics **80**: 296–304.

Morden, C. W., J. F. Doebley, and K. F. Schertz, 1989 Allozyme variation in old world races of *Sorghum bicolor* (Poaceae). American journal of botany pp. 247–255.

Morrell, P. L., K. E. Lundy, and M. T. Clegg, 2003 Distinct geographic patterns of genetic diversity are maintained in wild barley (*Hordeum vulgare* ssp. *spontaneum*) despite migration. Proceedings of the National Academy of Sciences **100**: 10812–10817.

Morris, G. P., P. Ramu, S. P. Deshpande, C. T. Hash, T. Shah, H. D. Upadhyaya, O. Riera-Lizarazu, P. J. Brown, C. B. Acharya, S. E. Mitchell, et al., 2013 Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proceedings of the National Academy of Sciences **110**: 453–458.

Multani, D. S., S. P. Briggs, M. A. Chamberlin, J. J. Blakeslee, A. S. Murphy, and G. S. Johal, 2003 Loss of an MDR transporter in compact stalks of maize br2 and sorghum dw3 mutants. Science **302**: 81–84.

Murphy, R. L., R. R. Klein, D. T. Morishige, J. A. Brady, W. L. Rooney, F. R. Miller, D. V. Dugas, P. E. Klein, and J. E. Mullet, 2011 Coincident light and clock regulation of pseudoresponse regulator protein 37 (PRR37) controls photoperiodic flowering in sorghum. Proceedings of the National Academy of Sciences **108**: 16469–16474.

Murphy, R. L., D. T. Morishige, J. A. Brady, W. L. Rooney, S. Yang, P. E. Klein, and J. E. Mullet, 2014 Ghd7 (Ma6) represses sorghum flowering in long days: Alleles enhance biomass accumulation and grain production. The Plant Genome **7**.

Mutegi, E., F. Sagnard, K. Semagn, M. Deu, M. Muraya, B. Kanyenji, S. De Villiers, D. Kiambi, L. Herselman, and M. Labuschagne, 2011 Genetic structure and relationships within and between cultivated and wild sorghum (*Sorghum bicolor* (L.) Moench) in kenya as revealed by microsatellite markers. Theoretical and applied genetics **122**: 989–1004.

Myung, I. J., 2003 Tutorial on maximum likelihood estimation. Journal of mathematical Psychology **47**: 90–100.

Nath, H. and R. Griffiths, 1993 The coalescent in two colonies with symmetric migration. Journal of mathematical biology **31**: 841–851.

Nath, H. and R. Griffiths, 1996 Estimation in an island model using simulation. Theoretical population biology **50**: 227–253.

Nielsen, R., 2001 Statistical tests of selective neutrality in the age of genomics. Heredity **86**: 641–647.

Nielsen, R. and M. Slatkin, 2013 An introduction to population genetics: theory and applications. Sinauer Associates Sunderland, MA.

Nielsen, R. and J. Wakeley, 2001 Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics **158**: 885–896.

Nordborg, M., 2007 Coalescent theory. Hand of statistical genetics.

Nordborg, M. and P. Donnelly, 1997 The coalescent process with selfing. Genetics **146**: 1185–1195.

Nordborg, M., T. T. Hu, Y. Ishino, J. Jhaveri, C. Toomajian, H. Zheng, E. Bakker, P. Calabrese, J. Gladstone, R. Goyal, et al., 2005 The pattern of polymorphism in *Arabidopsis thaliana*. PLoS biology **3**: e196.

Nordborg, M. and S. Tavaré, 2002 Linkage disequilibrium: what history has to tell us. TRENDS in Genetics **18**: 83–90.

Notohara, M., 1990 The coalescent and the genealogical process in geographically structured population. Journal of mathematical biology **29**: 59–75.

Odong, T., J. Van Heerwaarden, J. Jansen, T. J. van Hintum, and F. Van Eeuwijk, 2011 Determination of genetic structure of germplasm collections: are traditional hierarchical clustering methods appropriate for molecular marker data? Theoretical and Applied Genetics **123**: 195–205.

Ohta, T. and M. Kimura, 1969a Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. Genetics **63**: 229.

Ohta, T. and M. Kimura, 1969b Linkage disequilibrium due to random genetic drift. Genetical Research **13**: 47–55.

Ohta, T. and M. Kimura, 1971 Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. Genetics **68**: 571.

Oraguzie, N. C., E. H. Rikkerink, S. E. Gardiner, H. d. Silva, et al., 2007 Association mapping in plants.. Springer-Verlag GmbH.

Palczewski, M. and P. Beerli, 2013 A continuous method for gene flow. Genetics **194**: 687–696.

Paradis, E., 2010 pegas: an R package for population genetics with an integrated–modular approach. Bioinformatics **26**: 419–420.

Paterson, A. H., J. E. Bowers, R. Bruggmann, I. Dubchak, J. Grimwood, H. Gundlach, G. Haberer, U. Hellsten, T. Mitros, A. Poliakov, et al., 2009 The *Sorghum bicolor* genome and the diversification of grasses. Nature **457**: 551–556.

Pearse, D. E. and K. A. Crandall, 2004 Beyond F_{ST}: analysis of population genetic data for conservation. Conservation Genetics **5**: 585–602.

Pereira, M. and M. Lee, 1995 Identification of genomic regions affecting plant height in sorghum and maize. Theoretical and Applied Genetics **90**: 380–388.

Poland, J. A., P. J. Brown, M. E. Sorrells, and J.-L. Jannink, 2012 Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PloS one **7**: e32253.

Poland, J. A. and T. W. Rife, 2012 Genotyping-by-sequencing for plant breeding and genetics. The Plant Genome **5**: 92–102.

Pritchard, J. K. and M. Przeworski, 2001 Linkage disequilibrium in humans: models and data. The American Journal of Human Genetics **69**: 1–14.

Pritchard, J. K., M. Stephens, N. A. Rosenberg, and P. Donnelly, 2000 Association mapping in structured populations. The American Journal of Human Genetics **67**: 170–181.

Ptak, S. E., K. Voelpel, and M. Przeworski, 2004 Insights into recombination from patterns of linkage disequilibrium in humans. Genetics **167**: 387–397.

Quinby, J., 1966 Fourth maturity gene locus in sorghum. Crop Science **6**: 516–518.

Quinby, J., 1975 The genetics of sorghum improvement. Journal of Heredity **66**: 56–62.

Quinby, J., J. Hesketh, and R. Voigt, 1973 Influence of temperature and photoperiod on floral initiation and leaf number in sorghum. Crop Science **13**: 243–246.

Quinby, J. and R. Karper, 1945 Inheritance of three genes that influence time of floral initiation and maturity date in milo. Journal of the American Society of Agronomy.

Quinby, J. and R. Karper, 1954 Inheritance of height in sorghum.

Quinby, J. R., 1974 Sorghum improvement and the genetics of growth.

R CORE TEAM, 2015 R: A Language and Environment for Statistical Computing.

Rafalski, A. and M. Morgante, 2004 Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. TRENDS in Genetics **20**: 103–111.

Ramírez-Soriano, A., S. E. Ramos-Onsins, J. Rozas, F. Calafell, and A. Navarro, 2008 Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. Genetics **179**: 555–567.

Ramos-Onsins, S. E. and J. Rozas, 2002 Statistical properties of new neutrality tests against population growth. Molecular biology and evolution **19**: 2092–2100.

Roberts, A., L. McMillan, W. Wang, J. Parker, I. Rusyn, and D. Threadgill, 2007 Inferring missing genotypes in large SNP panels using fast nearest-neighbor searches over sliding windows. Bioinformatics **23**: i401–i407.

Rooney, W. and C. W. Smith, 2000 Techniques for developing new cultivars. Sorghum: Origin, history, technology and production. John Wiley & Sons, New York pp. 329–347.

Rooney, W. L. and S. Aydin, 1999 Genetic control of a photoperiod-sensitive response in *Sorghum bicolor* (L.) Moench. Crop science **39**: 397–400.

Rosa, J. R. B. F., 2011 Análise do desequilíbrio de ligação e da estrutura populacional do germoplasma brasileiro de cana-de-açúcar. Ph.D. thesis, Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo.

Rosenberg, N. A. and M. Nordborg, 2002 Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nature Reviews Genetics **3**: 380–390.

Rosenow, D., J. Dahlberg, J. Stephens, F. Miller, D. Barnes, G. Peterson, J. Johnson, and K. Schertz, 1997 Registration of 63 converted sorghum germplasm lines from the sorghum conversion program. Crop Sci **37**: 1399–1400.

Rosenow, D., J. Quisenberry, C. Wendt, and L. Clark, 1983 Drought tolerant sorghum and cotton germplasm. Agricultural Water Management **7**: 207–222.

Sagnard, F., M. Deu, D. Dembélé, R. Leblois, L. Touré, M. Diakité, C. Calatayud, M. Vaksmann, S. Bouchet, Y. Mallé, et al., 2011 Genetic diversity, structure, gene flow and evolutionary relationships within the *Sorghum bicolor* wild–weedy–crop complex in a western african region. Theoretical and applied genetics **123**: 1231–1246.

Sansaloni, C., 2012 Desenvolvimento e aplicações de DArT (Diversity Arrays Technology) e genotipagem por sequenciamento (Genotyping-By-Sequencing) para análise genética em Eucalyptus, volume 1. Tese (Doutorado). Universidade de Brasília.

Santos, F., C. Casela, and J. Waquil, 2005 Melhoramento de sorgo, volume 2. UFV: Viçosa.

Sethuraman, A. and J. Hey, 2016 IMa2p–parallel MCMC and inference of ancient demography under the isolation with migration (IM) model. Molecular ecology resources **16**: 206–215.

Simonsen, K. L., G. A. Churchill, and C. F. Aquadro, 1995 Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141**: 413–429.

Singh, H. P. and H. Lohithaswa, 2006 Sorghum. In Cereals and Millets, pp. 257–302, Springer.

Slatkin, M., 2008 Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. Nature Reviews Genetics **9**: 477–485.

Slatkin, M., L. Excoffier, et al., 1996 Testing for linkage disequilibrium in genotypic data using the expectation-maximization algorithm. Heredity **76**: 377–383.

Smith, C. W. and R. A. Frederiksen, 2000 Sorghum: Origin, history, technology, and production, volume 2. John Wiley & Sons.

Snowden, J., 1936 Cultivated Races of Sorghum.. London; Trustees of Bentham-Moxon Fund.

Sousa, V. and J. Hey, 2013 Understanding the origin of species with genome-scale data: modelling gene flow. Nature Reviews Genetics **14**: 404–414.

Sousa, V. C., M. Carneiro, N. Ferrand, and J. Hey, 2013 Identifying loci under selection against gene flow in isolation-with-migration models. Genetics **194**: 211–233.

Stacklies, W., H. Redestig, M. Scholz, D. Walther, and J. Selbig, 2007 pcaMethods–a bioconductor package providing pca methods for incomplete data. Bioinformatics **23**: 1164–1167.

Stephens, J., F. Miller, and D. Rosenow, 1967 Conversion of alien sorghums to early combine genotypes. Crop Science **7**: 396–396.

Stephens, M., 2007 Inference under the coalescent. Handbook of statistical genetics.

Stumpf, M. P. and G. A. McVean, 2003 Estimating recombination rates from population-genetic data. Nature Reviews Genetics **4**: 959–968.

Subudhi, P., D. Rosenow, and H. Nguyen, 2000 Quantitative trait loci for the stay green trait in sorghum (*Sorghum bicolor* L. Moench): consistency across genetic backgrounds and environments. Theoretical and Applied Genetics **101**: 733–741.

Sukumaran, S., W. Xiang, S. R. Bean, J. F. Pedersen, S. Kresovich, M. R. Tuinstra, T. T. Tesso, M. T. Hamblin, and J. Yu, 2012 Association mapping for grain quality in a diverse sorghum collection. The Plant Genome **5**: 126–135.

Sved, J., 1968 The stability of linked systems of loci with a small population size. Genetics **59**: 543.

Sved, J., 1971 Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theoretical population biology **2**: 125–141.

Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105**: 437–460.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**: 585–595.

Takahata, N., 1988 The coalescent in two partially isolated diffusion populations. Genetical research **52**: 213–222.

TAKAHATA, N. and M. SLATKIN, 1990 Genealogy of neutral genes in two partially isolated populations. Theoretical Population Biology **38**: 331–350.

TAO, Y., A. HARDY, J. DRENTH, R. HENZELL, B. FRANZMANN, D. JORDAN, D. BUTLER, and C. MCINTYRE, 2003 Identifications of two different mechanisms for sorghum midge resistance through QTL mapping. Theoretical and Applied Genetics **107**: 116–122.

TAO, Y., R. HENZELL, D. JORDAN, D. BUTLER, A. KELLY, and C. MCINTYRE, 2000 Identification of genomic regions associated with stay green in sorghum by testing RILs in multiple environments. Theoretical and Applied Genetics **100**: 1225–1232.

TEMPLETON, A. R., 2006 Population genetics and microevolutionary theory. John Wiley & Sons.

TUINSTRA, M., E. GROTE, P. GOLDSBROUGH, and G. EJETA, 1996 Identification of quantitative trait loci associated with pre-flowering drought tolerance in sorghum. Crop Science **36**: 1337–1344.

VARSHNEY, R. K., S. N. NAYAK, G. D. MAY, and S. A. JACKSON, 2009 Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends in biotechnology **27**: 522–530.

WAKELEY, J., 1997 Using the variance of pairwise differences to estimate the recombination rate. Genetical research **69**: 45–48.

WAKELEY, J., 2009 Coalescent theory: an introduction, volume 1. Roberts & Company Publishers.

WALL, J. D., 2000 A comparison of estimators of the population recombination rate. Molecular Biology and Evolution **17**: 156–163.

WALL, J. D., 2004 Estimating recombination rates using three-site likelihoods. Genetics **167**: 1461–1473.

WATTERSON, G., 1975 On the number of segregating sites in genetical models without recombination. Theoretical population biology **7**: 256–276.

WEIR, B. and W. HILL, 1986 Nonuniform recombination within the human beta-globin gene cluster. American journal of human genetics **38**: 776.

WEIR, B. S. ET AL., 1990 Genetic data analysis. Methods for discrete population genetic data.. Sinauer Associates, Inc. Publishers.

WILSON, G. A. and B. RANNALA, 2003 Bayesian inference of recent migration rates using multilocus genotypes. Genetics **163**: 1177–1191.

WIUF, C. and J. HEIN, 1999 Recombination as a point process along sequences. Theoretical population biology **55**: 248–259.

WRIGHT, S., 1931 Evolution in mendelian populations. Genetics **16**: 97-159.

XU, W., P. K. SUBUDHI, O. R. CRASTA, D. T. ROSENOW, J. E. MULLET, and H. T. NGUYEN, 2000 Molecular mapping of QTLs conferring stay-green in grain sorghum (*Sorghum bicolor* L. Moench). Genome **43**: 461–469.

YU, J. and E. S. BUCKLER, 2006 Genetic association mapping and genome organization of maize. Current Opinion in Biotechnology **17**: 155–160.

ZHU, C., M. GORE, E. S. BUCKLER, and J. YU, 2008 Status and prospects of association mapping in plants. The plant genome **1**: 5–20.

ZÖLLNER, S. and J. K. PRITCHARD, 2005 Coalescent-based association mapping and fine mapping of complex trait loci. Genetics **169**: 1071–1092.

ZÖLLNER, S. and A. VON HAESELER, 2000 A coalescent approach to study linkage disequilibrium between single-nucleotide polymorphisms. The American Journal of Human Genetics **66**: 615–628.