

UNIVERSIDADE ESTADUAL PAULISTA - UNESP

CÂMPUS DE JABOTICABAL

**COMPARAÇÃO DE MÉTODOS DE CONSTRUÇÃO DE
HAPLÓTIPOS EM ESTUDO DE ASSOCIAÇÃO GENÔMICA
AMPLA COM DADOS SIMULADOS**

Cherlynn Daniela da Silva Arce

Estatística

2018

UNIVERSIDADE ESTADUAL PAULISTA - UNESP

CÂMPUS DE JABOTICABAL

**COMPARAÇÃO DE MÉTODOS DE CONSTRUÇÃO DE
HAPLÓTIPOS EM ESTUDO DE ASSOCIAÇÃO GENÔMICA
AMPLA COM DADOS SIMULADOS**

Cherlynn Daniela da Silva Arce

Orientador: Prof. Dr. Henrique Nunes de Oliveira

Coorientadora: Dra. Camila Urbano Braz

Dissertação apresentada à Faculdade de Ciências Agrárias e Veterinárias – Unesp, Câmpus de Jaboticabal, como parte das exigências para a obtenção do título de Mestre em Genética e Melhoramento Animal.

2018

Arce, Cherlynn Daniela da Silva
A668c Comparação de métodos de construção de haplótipos em estudos
de associação genômica ampla com dados simulados / Cherlynn
Daniela da Silva Arce. -- Jaboticabal, 2018
iii, 34 p. : il. ; 29 cm

Dissertação (mestrado) - Universidade Estadual Paulista,
Faculdade de Ciências Agrárias e Veterinárias, 2018
Orientador: Henrique Nunes de Oliveira
Coorientadora: Camila Urbano Braz
Banca examinadora: Lucia Galvão de Albuquerque, Maria Eugênia
Zerlotti Mercadante
Bibliografia

1. Estudos de Associação. 2. Haplótipos. 3. Intervalo de Confiança.
4. Janelas Sobrepostas. 5. Regra de Quatro Gametas. 6. Snp. I.
Título. II. Jaboticabal-Faculdade de Ciências Agrárias e Veterinárias.

CDU 636.082:636.2

Ficha catalográfica elaborada pela Seção Técnica de Aquisição e Tratamento da Informação –
Diretoria Técnica de Biblioteca e Documentação - UNESP, Câmpus de Jaboticabal

CERTIFICADO DE APROVAÇÃO

TÍTULO DA DISSERTAÇÃO: COMPARAÇÃO DE MÉTODOS DE CONSTRUÇÃO DE HAPLÓTIPOS EM ESTUDO DE ASSOCIAÇÃO GENÔMICA AMPLA COM DADOS SIMULADOS

AUTORA: CHERLYNN DANIELA DA SILVA ARCE

ORIENTADOR: HENRIQUE NUNES DE OLIVEIRA

COORDINADORA: CAMILA URBANO BRAZ

Aprovada como parte das exigências para obtenção do Título de Mestra em GENÉTICA E MELHORAMENTO ANIMAL, pela Comissão Examinadora:

Camila U. Braz
Pós-doutoranda CAMILA URBANO BRAZ
Departamento de Zootecnia / FCAV / UNESP - Jaboticabal

Maria Eugênia Z. Mercadante
Pesquisadora Dra. MARIA EUGÊNIA ZERLOTTI MERCADANTE
Instituto de Zootecnia / IZ - Sertãozinho/SP

Lucia Galvão de Albuquerque
Profa. Dra. LUCIA GALVÃO DE ALBUQUERQUE
Departamento de Zootecnia / FCAV / UNESP - Jaboticabal

Jaboticabal, 27 de fevereiro de 2018

DADOS CURRICULARES DO AUTOR

Cherlynn Daniela da Silva Arce, nascida em Curitiba – PR em 03 de Agosto de 1987, iniciou o curso de Bacharelado em Estatística na Universidade Federal do Paraná (UFPR), Curitiba – PR, em Agosto de 2009, tornando-se Estatística em Julho de 2013. Iniciou o curso de Pós-Graduação em Genética e Melhoramento Animal na Universidade Estadual Paulista “Júlio de Mesquita Filho” em março de 2016.

“Deficiências de capacidade inata podem ser compensadas através do trabalho duro persistente e da concentração. Pode-se dizer que o trabalho substitui o talento, ou, melhor ainda, que cria o talento.”

(Santiago de Ramón y Cajal)

Este trabalho é dedicado ao Thiago que sempre está do meu lado me apoiando, mesmo fisicamente distantes. À Letícia e Verônica pela amizade e apoio principalmente nos momentos mais difíceis.

AGRADECIMENTOS

Ao Prof. Dr. Henrique Nunes de Oliveira, pelas suas sábias palavras inspiradoras aliadas de orientações simpáticas e pacientes, e que sempre encontrou um tempo disponível para meu auxílio. O senhor é um grande exemplo de vida para ser seguido;

À Dra. Camila Urbano Braz por disponibilizar de seu tempo para meu auxílio, sempre com muita simpatia e cordialidade ao me orientar e me aconselhando para que eu seja uma pessoa e profissional melhor. Camila, você é um grande exemplo de vida para ser seguido;

À Coordenação de Aperfeiçoamento de pessoal de Nível Superior (CAPES) pela concessão da bolsa durante o período de mestrado;

À Universidade Estadual Paulista “Júlio de Mesquita Filho” pela oportunidade;

Aos Profs. Drs. Danísio Prado Munari e Salvador Ramos que me incentivaram a fazer o curso explanando dúvidas e me aconselhando sabiamente em ser uma pessoa e profissional melhor;

À Profa. Dra. Lúcia Galvão de Albuquerque e ao Prof. Dr. Fernando Sebastian Baldi Rey pelos seu ensinamentos, disponibilidade e exemplos de vida;

Às Profa. Dra. Lúcia Galvão de Albuquerque e à Dra. Maria Eugênia Z. Mercadante pela gentileza de ter aceitado o convite para serem membros da banca de defesa e pelas suas sugestões e críticas valiosas;

Ao Tiago Bresolin pelos conselhos e orientações além da sua disponibilidade e ajuda sempre muito cordial;

Aos meus amigos e colegas coorientados do Prof. Dr Henrique: André, Ândrea, Donicer, Ligia, Marina e Marisa pela amizade e convivência regados de cafés além dos seus sábios conselhos e apoios, me sinto muito afortunada de ter conhecido todos vocês;

Ao Thiago Roberto do Prado por sempre estar do meu lado me apoiando e ajudando nos momentos de maior necessidade.

Às minhas amigas do coração Letícia Inácio e Verônica Precybilovicz cujas pessoas eu sempre admirei e sempre vou admirar, seus apoios e incentivos sempre me foram muito preciosos. Contem sempre comigo!

Aos meus amigos André Mauric e Bianca Olivieri por todo apoio e auxílio quando mais precisei, sempre me lembrarei de vocês.

Aos meus amigos Anderson Alves e Rebeka Magalhães pela amizade e simpatia acompanhados de boas músicas, momentos de descontração e de ensinamentos além de todo o apoio, momentos os quais recordarei para toda a vida.

À minha amiga Anielly Freitas que gentilmente se dispôs a me ajudar e apoiar e que me ensinou muito em pouco tempo de convivência, mas são lições que levarei com muito carinho aonde quer que eu vá.

Aos meus sogros Edna e José do Prado por toda ajuda e apoio.

Ao Prof. Dr. Walter Maldonado Jr pela disponibilidade do pacote fcaTeX utilizado para a produção textual desse estudo.

Finalmente, a todos aqueles que diretamente ou indiretamente influenciaram e/ou apoiaram este trabalho.

SUMÁRIO

RESUMO.....	ii
ABSTRACT.....	iii
CAPÍTULO 1 – Considerações gerais.....	4
1. INTRODUÇÃO.....	4
2. Revisão de Literatura.....	5
2.1. Haplótipos.....	5
2.2. Estudos de associação genômica ampla utilizando haplótipos.....	6
2.3. Métodos de construção de haplótipos.....	7
3. Objetivo Geral.....	9
3.1. Objetivos específicos.....	9
4. MATERIAL E MÉTODOS.....	10
4.1. Simulação de dados Genômicos.....	10
4.2. Construção dos Blocos de Haplótipos.....	12
4.3. Matriz G.....	13
4.4. Modelo misto linear univariado.....	14
4.5. Testes de significância.....	15
4.9. Fórmula do desequilíbrio de ligação para marcadores multi-alélicos.....	16
5. RESULTADOS E DISCUSSÃO.....	17
5.1. Extensão do desequilíbrio de ligação das populações simuladas.....	17
5.2. Correlações entre as matrizes G.....	20
5.3. Análise de associação.....	21
6. CONCLUSÃO.....	28
7. REFERÊNCIAS.....	28

COMPARAÇÃO DE MÉTODOS DE CONSTRUÇÃO DE HAPLÓTIPOS EM ESTUDOS DE ASSOCIAÇÃO GENÔMICA AMPLA COM DADOS SIMULADOS

RESUMO – Com o avanço dos estudos em genética e da tecnologia aplicada à genotipagem de marcadores moleculares, a identificação de polimorfismos associados às características de interesse econômico se tornou mais acessível, possibilitando a sua utilização aumentar a acurácia de modelos de predição do mérito genético dos animais. Esse avanço também possibilitou aumentar a acurácia dos estudos para identificação de QTLs para características de interesse econômico. Entretanto, os marcadores comumente utilizados para tal fim são os SNPs, que por serem bi-alélicos podem não ser muito eficientes na identificação dos QTLs. Os haplótipos, multi-alélicos, apresentam maior possibilidade de estarem em desequilíbrios de ligação (DL) com os QTLs. Dessa forma, objetivou-se no presente trabalho identificar o melhor método de construção de haplótipos para utilização em estudos de detecção de QTLs, a partir da comparação dos três métodos mais comumente utilizados para este fim. Foram utilizadas três populações simuladas representando características com três diferentes valores de herdabilidade, para as quais foram armazenados os dados fenotípicos, genotípicos e de pedigree dos 6.000 animais da população mais recente: Pop1 com herdabilidade baixa (0,10); Pop2 com herdabilidade moderada (0,25); e, Pop3 com herdabilidade alta (0,35). Os genomas simulados consistiram de 750.000 marcadores do tipo SNP, e 750 QTLs, com dois a quatro alelos, dispostos aleatoriamente em 29 cromossomos com tamanho total de 2.333 centimorgans (cM). A partir da simulação foram eliminados os SNPs cuja frequência do menor alelo foi menor que 0,1, restando 576.027, 577.189 e 576.675 marcadores para as populações Pop1, Pop2 e Pop3, respectivamente. A variação fenotípica foi de 1,0 e a variação dos QTLs foi de 50% das herdabilidades, para cada população. As médias dos DL para cada cromossomo, medidas pela estatística D' , variaram de 0,20 até 0,30 para todas as populações, na última geração. Foram construídos haplótipos utilizando três métodos: Intervalo de Confiança (IC), Regra de Quatro Gametas (RQG) e Janelas Sobrepostas (JS). Para Pop1, no cromossomo 15, os métodos IC, RQG e JS identificaram cinco, oito e sete QTLs, respectivamente. Somente um QTL foi identificado nos cromossomos 19 e 29. Para a característica de herdabilidade alta, foi identificado um QTL no cromossomo 11. Em relação às análises de associação utilizando SNPs individuais, foram identificados quatro QTLs no cromossomo 15. Para a característica de herdabilidade moderada, não foram encontrados haplótipos ou SNPs isolados significativos. A metodologia de formação de haplótipos baseado na RQG foi considerada a mais eficiente para detecção de QTLs em relação aos métodos IC e JS, bem como ao uso dos SNPs isolados.

Palavras-chave: estudos de associação, haplótipos, intervalo de confiança, janelas sobrepostas, regra de quatro gametas, snp

COMPARISON OF HAPLOTYPES CONSTRUCTION METHODS IN GENOMIC ASSOCIATION STUDIES WITH SIMULATED DATA

ABSTRACT – With the advancement of genetic studies and the technology applied to the genotyping of molecular markers, the identification of polymorphisms associated with the characteristics of economic interest became more accessible, allowing its use to increase the accuracy of prediction models of the genetic merit of the animals. This advance also made it possible to increase the accuracy of studies to identify QTLs for characteristics of economic interest. However, the commonly used markers for this purpose are SNPs, which because they are bi-allelic may not be very efficient in identifying QTLs. The haplotypes, multi-allelic, are more likely to be in linkage disequilibrium (LD) with QTLs. Thus, the objective of this work was to identify the best haplotype construction method for use in QTLs detection studies, by comparing the three methods most commonly used for this purpose. Three simulated populations representing characteristics with three different heritability values were used for which the phenotypic, genotypic and pedigree data of the 6,000 animals were stored: Pop1 with low heritability (0.10); Pop2 with moderate heritability (0.25); and, Pop3 with high heritability (0.35). The simulated genomes consisted of 750,000 SNP-type markers, and 750 QTLs, with two to four alleles, arranged randomly on 29 chromosomes with a total size of 2,333 centimorgans (cM). From the simulation the SNPs whose frequency of the lowest allele was less than 0.1 were eliminated, leaving 576,027, 577,189 and 576,675 markers for Pop1, Pop2 and Pop3 populations, respectively. The phenotypic variation was 1.0 and the variation of QTLs was 50% of the heritabilities, for each population. The mean LD for each chromosome, measured by the D' statistic, ranged from 0.20 to 0.30 for all populations in the last generation. Haplotypes were constructed using three methods: Confidence Interval (CI), Four Gametes Rule (FGR) and Sliding-Window (SW). For Pop1, on chromosome 15, CI, FGR and SW methods identified five, eight and seven QTLs, respectively. Only one QTL was identified on chromosomes 19 and 29. For the high heritability characteristic, a QTL was identified on chromosome 11. Regarding the association analyzes using individual SNPs, four QTLs were identified on chromosome 15. For the moderate heritability characteristic, no significant isolated haplotypes or SNPs were found. The methodology of haplotype formation based on the FGR was considered the most efficient for the detection of QTLs in relation to CI and SW methods, as well as to the use of isolated SNPs.

Keywords: association study, confidence interval, haplotypes, four gamete rule, sliding-window, snp

CAPÍTULO 1 – Considerações gerais

1. INTRODUÇÃO

O sequenciamento do genoma bovino possibilitou o desenvolvimento de painéis de alta densidade com grande número de marcadores SNP distribuídos ao longo do genoma com potencial para capturar grande parte da variação genética de determinadas características (MATUKUMALLI et al., 2009). Esse avanço permitiu a detecção de QTL para diferentes características complexas por meio de estudos de associação genômica ampla (*Genome Wide-Association studies* - GWAS) (HAYES et al., 2007; CALUS et al., 2009; WU et al., 2014). Os estudos de GWAS são baseados no desequilíbrio de ligação (DL) entre os marcadores e os QTLs (HAMBLIN; JANNINK, 2011) sendo que quanto maior o DL, maior o poder de detectar os QTLs (LONG; LANGLEY, 1999).

A utilização de SNPs isolados em estudos de associação não tem sido muito eficiente na detecção de QTL (GUO et al., 2009; LORENZ; HAMBLIN; JANNINK, 2010; SERÃO et al., 2013; WU et al., 2014), o que pode ser explicado pelo fato dos SNPs capturarem pouca informação sobre a variabilidade genética presente em determinadas regiões do genoma (BALL et al., 2010). Além disso, de acordo com Hayes e Goddard (2001) muitos QTLs presentes no genoma bovino apresentam pequenos efeitos, o que dificulta sua detecção com a utilização de SNP isolados (WANG et al., 2016). Por outro lado, utilizar sequências de marcadores adjacentes, denominadas haplótipos, pode ser vantajoso em estudos de associação, visto que haplótipos são multi-alélicos e podem apresentar maior DL com o QTL (LONG; LANGLEY, 1999; CUYABANO; SU; LUND, 2014).

Existem diferentes metodologias para construção de haplótipos, as quais apresentam vantagens e desvantagens dependendo da população, da característica e da distribuição dos marcadores ao longo do genoma. O método de Intervalo de Confiança (IC) é baseado no DL entre marcadores adjacentes e pode ser mais eficiente para detectar os blocos de haplótipos em populações com maior decaimento do DL em função da distância entre os marcadores, mas que possuam grande diversidade haplotípica (GABRIEL et al., 2002). A Regra de Quatro Gametas (RQG) é um método baseado no processo de recombinação entre dois marcadores

e recomenda-se sua utilização em genomas com maior frequência de eventos de recombinação (WANG et al., 2002). No método de Janelas Sobrepostas (JS) todos os marcadores formam haplótipos de tamanhos constantes sem a necessidade de conhecimento prévio da estrutura genética da população (MATHIAS et al., 2006).

Como desvantagens, os métodos baseados no DL e em recombinação, como IC e RQG, desprezam regiões no genoma que não apresentam tais características, o que acarreta na não inclusão de SNPs “órfãos” na análise de associação (LI; SUNG; LIU, 2007). Por outro lado, o método JS não apresenta essa limitação, pois o mesmo cobre todo o genoma, aumentando a possibilidade de detectar o loco causal (GUO et al., 2009), no entanto, seu custo computacional é elevado.

Poucos estudos têm sido realizados com a finalidade de encontrar a melhor metodologia para construção de haplótipos em diferentes espécies (MEGENS et al., 2009; LORENZ; HAMBLIN; JANNINK, 2010; SODELAND, 2010; BRAZ, 2016;). Assim, não há consenso na literatura em relação ao melhor método de construção de haplótipos para utilização em estudos de associação, principalmente em bovinos de corte.

2. Revisão de Literatura

2.1. Haplótipos

Haplótipos são regiões genômicas de SNPs adjacentes com poucas evidências de recombinação histórica (GABRIEL et al., 2002) compostas por marcadores que apresentam alto desequilíbrio de ligação (DL) entre si (CARDON; ABECASIS, 2003).

Segundo Cuyabano et al. (2014) existe maior possibilidade de identificação de QTL com a utilização de haplótipos, em relação ao uso de SNPs isolados, em razão dos primeiros serem multi-alélicos. De acordo com os autores, as frequências alélicas dos marcadores bi-alélicos, como os SNPs, podem ficar inalteradas mesmo com a ocorrência de mutação e, dessa forma, essa mutação pode não ser detectada. Além disso, existe maior possibilidade do QTL estar em alto DL com um

haplótipo do que com um SNP que o compõe (CUYABANO et al., 2014; HAMBLIN, JANNINK, 2011).

Calus et al. (2009) demonstraram que o mapeamento de QTL foi considerado mais preciso com a utilização de haplótipos. Lorenz, Hamblin e Jannink (2010) relataram vários benefícios da utilização de haplótipos em análises de associação quando comparados com marcadores SNPs, tais como: habilidade de capturar interações epistáticas entre SNPs em um loco; mais informativos na identificação de pares de alelos herdados pelas progênes, sendo provenientes de um ancestral em comum (Idênticos por Descendência - ID); reduzem o número de testes de hipóteses (exceto no caso de JS), o que conseqüentemente, diminui o erro tipo I; capturam com maior eficiência informações de evolução histórica e fornecem maior poder de detecção devido sua composição ser uma sequência de marcadores em um loco.

Barendse (2011) estudou a característica de gordura corporal e sugeriu que com o uso haplótipos se pode extrair mais informações dos dados genômicos, pois eles possuem maior capacidade de detectar a variação fenotípica em comparação a SNPs isolados. Zhang et al. (2016) afirmam que estudos genômicos com haplótipos são ferramentas poderosas para busca de QTL. Salem et al. (2018) relataram que haplótipos foram eficientes para a identificação de vários genes associados com características de interesse econômico.

2.2. Estudos de associação genômica ampla utilizando haplótipos

O objetivo dos estudos de associação genômica ampla (GWAS) é identificar QTL para características de interesse (IKEGAWA, 2012; VISSCHER et al., 2012). Para tal análise, alguns princípios devem ser considerados para evitar resultados falsos-positivos e maximizar o poder de detecção dos QTL (HAMBLIN; JANNINK, 2011; ZENG et al., 2015). Para evitar variáveis de confundimento em estudos de GWAS é recomendável incluir a estrutura populacional para considerar o efeito poligênico. Isso pode ser feito com a inclusão da matriz de parentesco convencional, matriz de parentesco genômica, ou pelas suas componentes principais (ZENG et al., 2015). Além disso, o poder de detecção de QTL dos estudos de GWAS é diretamente proporcional ao tamanho da amostra (número de animais) e também da

densidade e distribuição dos marcadores utilizados no estudo, pois quanto maior a densidade do painel de marcadores maior a probabilidade de o QTL estar em DL com algum marcador (HAMBLIN; JANNINK, 2011; GODDARD; HAYES, 2012). Por outro lado, como os SNPs são analisados independentemente em estudos de GWAS (BUSH; MOORE, 2012; ZENG et al., 2015) um grande número de testes é gerado, causando redução do poder estatístico.

O uso de haplótipos em GWAS pode reduzir o número de testes, pois SNPs são analisados conjuntamente e os que não forem agrupados não são, comumente, incluídos na análise. Deste modo, o poder do teste é preservado e a taxa de falso-positivos pode ser controlada (SHIM et al., 2009), exceto para o caso das Janelas Sobrepostas. No entanto, não existe consenso sobre qual método é recomendável em estudos de associação em bovinos de corte, principalmente porque existe carência na literatura de estudos que comparem os desempenhos entre métodos de construção de haplótipos.

2.3. Métodos de construção de haplótipos

Diferentes metodologias para construção de haplótipos têm sido descritas na literatura (GABRIEL et al., 2002; WANG et al., 2002; MATHIAS et al., 2006), as quais apresentam vantagens e desvantagens. O método Intervalo de Confiança (IC), desenvolvido por Gabriel et al. (2002), é o método que tem sido mais utilizado em estudos de seleção e associação genômica (CALUS et al., 2008; SODELAND 2010; CUYABANO; SU; LUND, 2014; WU et al., 2014). Esse método utiliza como critério o DL existente entre marcadores adjacentes.

O DL ocorre quando dois alelos no mesmo cromossomo segregam de maneira dependente, sendo que quanto maior a correlação existente entre eles, mais ligados esses alelos estão (GUPTA; RUSTGI; KULWAL, 2005). Esse DL tende a decrescer de acordo com a distância entre os marcadores, entretanto o decaimento do DL difere entre espécies e raças (MCKAY et al., 2007; WANG et al., 2013; PORTO-NETO; KIJAS; REVERTER, 2014; BRITO et al., 2015).

Segundo Gabriel et al. (2002) populações que possuam grande extensão de DL ao longo do genoma ou pouca extensão, mas com maior diversidade haplotípica,

podem ser beneficiadas pelo método de IC. Entretanto, a utilização do método IC pode não ser vantajosa quando se utiliza painéis de baixa densidade (GABRIEL et al., 2002). Os valores da estatística D' irão variar de zero (0) até um (1), quanto mais próximo de zero menor evidência de desequilíbrio de ligação entre os pares de marcadores, e quanto mais próximo de um os pares de marcadores possuirão 'forte DL'. Os autores descrevem que para aplicação desse método em uma amostra, considera-se que um par de SNPs esteja em "forte DL" quando o limite superior do intervalo de confiança 95% da estimativa de D' para o par for igual ou maior que 0,98 e o limite inferior for igual ou maior que 0,70; quando o limite superior estiver abaixo de 0,90 existe "evidência de recombinação histórica" e nas outras situações o par de SNPs é considerado "não informativo". Quando 95% dos pares de SNPs informativos de um conjunto de SNPs adjacentes estiverem em "forte LD" esse conjunto é categorizado como um bloco de haplótipos. Marcadores com Frequência Alélica Mínima (MAF) inferior a 5% geralmente são considerados não informativos e, consequentemente, não são incluídos na análise.

O método Regra de Quatro Gametas (RQG) utiliza o evento de recombinação como critério para a formação dos haplótipos (WANG et al., 2002). De acordo com Hudson e Kaplan (1985) uma desvantagem é que o método RQG não é capaz de identificar todas as recombinações históricas existentes no genoma. O RQG considera que, para quaisquer dois locos adjacentes, um evento de recombinação ocorre se são observados quatro tipos de gametas, sendo que o quarto gameta possui frequência superior ou igual a 0,01 (WANG et al., 2002). A formação dos blocos é iniciada a partir do primeiro marcador até a primeira recombinação. Após a recombinação, é iniciado um novo bloco que irá terminar até a próxima recombinação ascendida de um evento não recombinante e, assim, seguirá até o último marcador do cromossomo, sendo que os haplótipos formados podem variar de tamanho (HUDSON; KAPLAN, 1985).

Outra metodologia para a construção de haplótipos é a baseada em Janelas Sobrepostas (JS), na qual haplótipos são construídos de tamanho fixo, começando com o primeiro SNP e movendo as janelas até cobrir todo o genoma, um SNP de cada vez (MATHIAS et al., 2006). Para a utilização deste método não é necessário conhecer previamente a estrutura genética da população, ou seja, o DL, a taxa de

recombinação, entre outros fatores que influenciam na construção dos blocos de haplótipos pelas metodologias apresentadas anteriormente (BRAZ, 2016). Além disso, nesse método todos os marcadores são incluídos em haplótipos, aumentando a possibilidade de detecção de QTL, pois todos SNPs sempre são considerados nas análises (GAO et al., 2012). Entretanto, existem algumas desvantagens na utilização do método JS como a redução do poder do teste de associação, visto que o número de testes analisados é muito grande (GUO et al., 2009); o tamanho ideal do haplótipo a ser utilizado precisa ser determinado (GUO et al., 2009; GAO et al., 2012; BRAZ, 2016), e há elevado custo computacional para a construção e análise dos haplótipos (CHENG et al., 2005).

A metodologia utilizada para a construção dos haplótipos pode afetar nos resultados obtidos da análise de associação (GABRIEL et al., 2002; WANG et al., 2002; HAMBLIN; JANNINK, 2011). No entanto, estudos com a finalidade de encontrar a melhor metodologia para construção de haplótipos em diferentes espécies são escassos (MEGENS et al., 2009; LORENZ; HAMBLIN; JANNINK, 2010; SODELAND, 2010; BRAZ, 2016), principalmente em bovinos de corte.

3. Objetivo Geral

Objetiva-se identificar o melhor método de construção de haplótipos para detecção de QTLs, a partir da comparação dos três métodos mais comumente utilizados para este fim, utilizando três populações simuladas representando características com três diferentes valores de herdabilidade.

3.1. Objetivos específicos

- Simular três populações, com herdabilidade diferente, baixa (0,10), moderada (0,25) e alta (0,35);
- Construir haplótipos para as três populações utilizando três métodos: Intervalo de Confiança, Regra de Quatro Gametas e Janelas Sobrepostas de tamanho cinco;

- Comparar o desempenho de cada método de construção de entre si e com as análises dos modelos construídos utilizando SNPs isolados em GWAS.

4. MATERIAL E MÉTODOS

4.1. Simulação de dados Genômicos

Para a obtenção de dados fenotípicos, genotípicos e de pedigree foi utilizado o simulador QMSim, versão 1.10 (SARGOLZAEI; SCHENKEL, 2009). Para cada simulação foi realizada uma réplica baseada nas simulações de Brito et al. (2011) e Melo (2015).

Para cada população foram simulados três passos que correspondem a população histórica, população de expansão e população recente, respectivamente. Inicialmente, na geração zero, a população histórica é composta de 1000 animais. A partir da geração 1.001 até a geração 2.020 o número de animais decai de 1000 para 200 com o objetivo de criar DL e equilíbrio de mutação e deriva.

Na população de expansão, foram utilizados como fundadores os 100 machos e as 100 fêmeas da última população histórica, o objetivo é aumentar o número de animais para que haja animais suficientes para serem os fundadores da população recente. Dessa forma, da última geração da população de expansão foram selecionados aleatoriamente 240 machos e 6.000 fêmeas, com o intuito de simular um rebanho bovino real, com o tamanho efetivo de 923 animais.

No genoma dos animais simulados, o número e a disposição dos marcadores por cromossomos foram baseados em chips de genotipagem encontrados atualmente no mercado. O diagrama de fluxo do processo de simulação das populações pode ser observado na Figura 4.1.

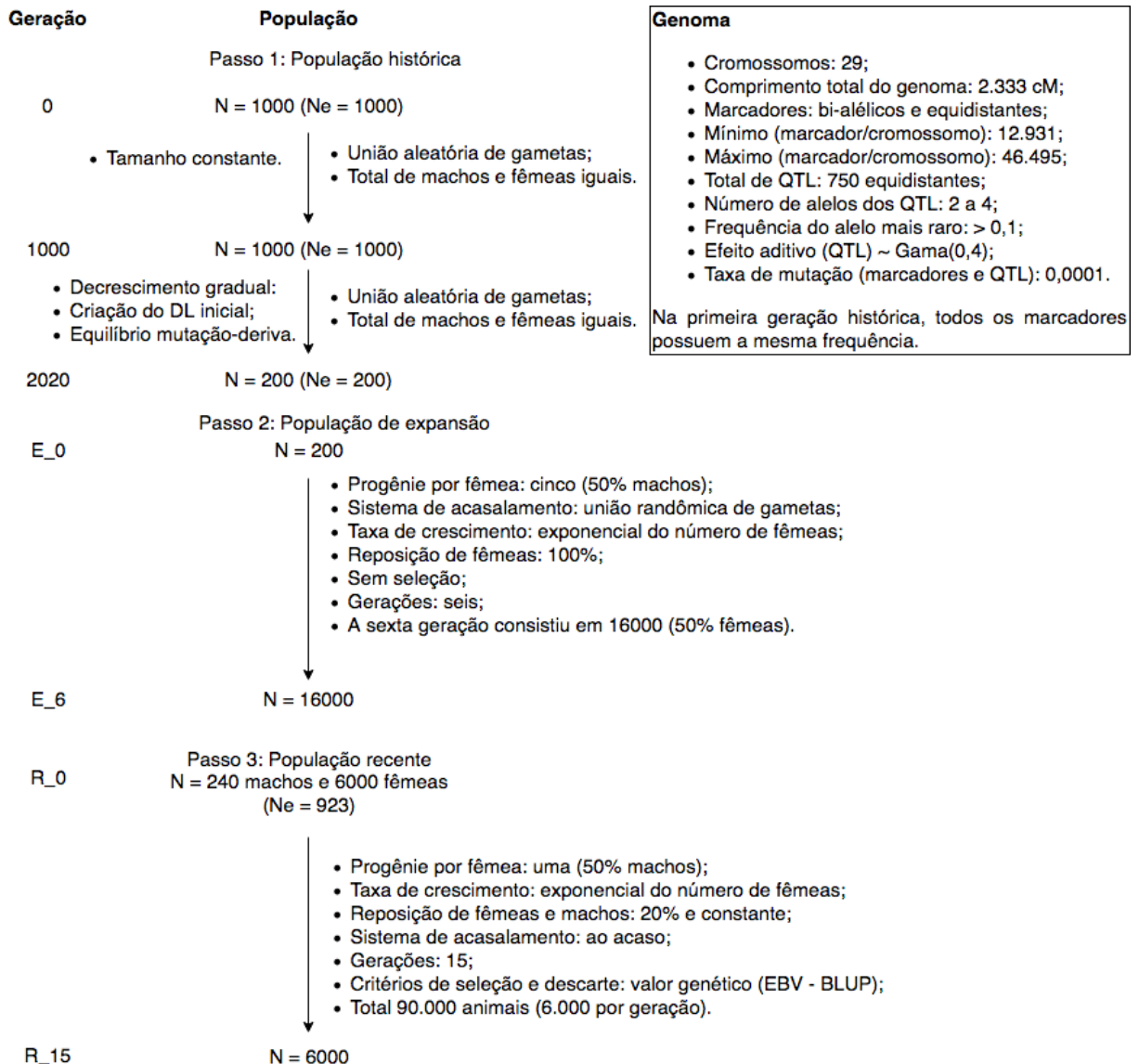


Figura 4.1. Diagrama de fluxo da simulação.

Após a simulação, o número de marcadores SNPs com $MAF > 0,1$ foi de 576.027, 577.189 e 576.675 para as populações Pop1 (herdabilidade baixa: 0,10), Pop2 (herdabilidade moderada: 0,25) e Pop3 (herdabilidade alta: 0,35), respectivamente, distribuídos em 29 cromossomos com comprimento total de 2.333 cM. A variação fenotípica foi de 100% e a variação atribuída aos QTLs foi de 50% dos valores das herdabilidades simuladas.

Os animais analisados no presente estudo pertenciam à última geração simulada para cada população (Pop1, Pop2 e Pop3), as quais foram compostas de

3.010, 2.980 e 3.058 fêmeas e 2.990, 3.020 e 2.942 machos, respectivamente. O sumário dos fenótipos para cada herdabilidade está disponível na Tabela 4.1.

Tabela 4.1. Estatísticas dos fenótipos correspondentes de cada população.

População	Mínimo	Máximo	Média	Desvio-Padrão
Pop1	-2,563	4,935	1,344	0,9948
Pop2	-1,297	6,676	2,723	0,1376
Pop3	-0,2276	7,1473	3,1843	0,9788

4.2. Construção dos Blocos de Haplótipos

Foram utilizados três métodos para a construção dos blocos de haplótipos: Intervalo de Confiança (IC), quando 95% dos pares de SNPs informativos de um conjunto de SNPs adjacentes estiverem em "forte LD" (o limite superior do IC de $D' \geq 0,98$ e o limite inferior $\geq 0,70$) esse conjunto é categorizado como um bloco de haplótipos e os marcadores que possuíam $MAF < 0,05$ foram considerados não informativos e ignorados (GABRIEL et al., 2002); Regra de Quatro Gametas (RQG), no qual foi considerado um evento de recombinação quando houve a identificação do quarto gameta com frequência observada maior que 0,01 (WANG et al., 2002); Janelas Sobrepostas (JS), em que os haplótipos foram constituídos por cinco marcadores adjacentes, representando o tamanho da janela de cada haplótipo, (MATHIAS et al., 2006). Após a formação do primeiro haplótipo, composto do primeiro marcador até o quinto, a janela será deslocada para formar o segundo haplótipo, do segundo marcador até o sexto, e continuará nesse processo iterativo até o final do genoma. Sendo assim, cada marcador poderá constituir de um ou mais haplótipos. Os números totais de haplótipos formados e a porcentagem do genoma coberto para cada método e população podem ser observados na Tabela 4.2.

Tabela 4.2. Números totais de haplótipos e alelos construídos e porcentagem de cobertura no genoma por IC, RQG e JS por população.

Pop	IC		RQG		JS*		Cobertura ¹
	Hap	Alelos	Hap	Alelos	Hap	Alelos	
Pop1	120.010	530.727	128.056	581.146	575.911	5.168.460	46% (IC) 48% (RQG)
Pop2	121.051	522.333	129.762	568.446	577.073	5.191.308	45% (IC) 46% (RQG)
Pop3	120.945	515.158	128.339	560.551	576.559	5.245.741	44% (IC) 45% (RQG)

Pop: População; IC: Intervalo de Confiança; RQG: Regra de Quatro Gametas; JS: Janelas Sobrepostas; Hap: haplótipos. * Porcentagem de cobertura para JS é sempre 100%; ¹Porcentagem de cobertura.

As construções dos haplótipos com base nas metodologias IC e RQG foram realizadas utilizando o software HaploView (BARRETT et al., 2004). O método JS foi aplicado utilizando o software R (R CORE TEAM, 2016).

4.3. Matriz G

As matrizes genômicas dos haplótipos e SNPs foram constituídas de acordo com o número de cópias de cada alelo que cada animal possuía, codificados da seguinte maneira: 0 para os animais que não possuíam cópia do alelo; 1 para os animais que possuíam apenas uma cópia do alelo proveniente de um dos pais; 2 para os animais que possuíam duas cópias do alelo, sendo que cada cópia foi proveniente de um dos pais. Dessa forma, cada matriz dos blocos de haplótipos e SNPs teve dimensão $N \times p$, onde N corresponde ao número total de alelos dos blocos de haplótipos ou SNPs e p corresponde ao número de animais no estudo.

Para as análises dos haplótipos formados pelos métodos IC e RQG foram utilizadas as matrizes G construídas com base nos respectivos haplótipos. Para a análise dos SNPs utilizou-se a matriz G construída utilizando os SNPs isolados.

Essa mesma matriz foi utilizada para a análise dos haplótipos formados pelo método JS, visto que neste método cada marcador forma mais de um haplótipo, e, dessa forma, a estrutura populacional no modelo poderia estar superestimada.

As matrizes G foram calculadas para cada população e método de haplótipo utilizando o software GEMMA (ZHOU; STEPHENS, 2012). Seja X a matriz de incidência que contém i alelos de n blocos de haplótipos de p animais, a estatística para o cálculo da matriz de parentesco genômica é definida como:

$$G = \frac{1}{p} \sum_{i=1}^p \frac{1}{v_{x_i}} (x_i - 1_n \bar{x}_i)(x_i - 1_n \bar{x}_i)^T.$$

Onde, x_i é a i -ésima coluna que corresponde ao i -ésimo alelo da matriz para os haplótipos e no caso dos SNPs corresponde ao i -ésimo marcador; $\bar{x}_i$ e v_{x_i} correspondem a média e variâncias amostrais do i -ésimo alelo ou SNP e 1_n é um vetor $n \times 1$ de uns.

Após a construção das matrizes foram estimadas as correlações, entre os elementos fora da diagonal principal, das matrizes G dos métodos IC (G_{IC}) e RQG (G_{RQG}) com as matrizes G convencionais, computadas utilizando os marcadores SNPs (G_{SNP}) com a finalidade de mensurar a associação linear dos graus de parentesco obtidos entre as mesmas. Caso as correlações entre as matrizes G_{IC} e G_{RQG} com as matrizes G_{SNP} forem altas, essas podem ser utilizadas nos modelos de associação.

4.4. Modelo misto linear univariado

As análises de associação utilizando haplótipos e SNPs isolados, para as populações Pop1, Pop2 e Pop3 foram realizadas utilizando o software GEMMA (ZHOU; STEPHENS, 2012) ajustando um modelo misto linear univariado, como segue:

$$y = \mu + x\beta + u + \epsilon.$$

Onde, y é vetor de dimensão n de características quantitativas de n -indivíduos, μ é uma constante, x é a matriz dos blocos de haplótipos, β é o efeito dos blocos de haplótipos ou SNP, u é o efeito aleatório dos animais, sendo $u \sim NMV_n(0, \lambda\tau^{-1}G)$, ϵ é o erro aleatório, $\epsilon \sim NMV_n(0, \tau^{-1}I_n)$, τ^{-1} é a variância residual

dos erros, λ é a razão entre dois componentes de variância, G é a matriz genética de parentesco, I_n é a matriz identidade com dimensão $n \times n$.

4.5. Testes de significância

Para a análise de associação utilizando os marcadores moleculares SNPs foram computados os valores de p dos efeitos dos marcadores pelo Teste de Wald (ZENG et al., 2015) após correção para múltiplos testes pelo método de Bonferroni.

Para os blocos de haplótipos, por serem multi-alélicos, o teste de Wald não se aplica e, dessa forma, houve necessidade de adaptação de um teste específico a partir da variância genética explicada por cada bloco de haplótipo. A variância genética aditiva dos haplótipos foi estimada utilizando as frequências e os efeitos de cada alelo para cada haplótipo e é definida como (LYNCH; WALSH, 1998):

$$\sigma_j^2 = \sum_i^I \sum_{l>i}^I (a_{ij} - a_{lj})^2 p_{ij} p_{lj}.$$

Onde, σ_j^2 é a variância genética aditiva do j -ésimo haplótipo; a_i é o efeito aditivo do i -ésimo alelo do j -ésimo haplótipo; a_l é o efeito aditivo do l -ésimo alelo do j -ésimo haplótipo, com $i \neq l$ e $l > i$; p_i é a frequência do i -ésimo alelo do j -ésimo haplótipo; p_l é a frequência do l -ésimo alelo do j -ésimo haplótipo, e I : é o total de alelos do j -ésimo haplótipo.

Como a variância esperada das estimativas dos efeitos dos alelos é dependente do número de alelos por bloco de haplótipos, uma vez calculada a variância para cada bloco de haplótipo foi ajustada inicialmente uma curva da distribuição de probabilidade gama para as variâncias dos blocos haplótipos de acordo com o número de alelos, sendo estimados os parâmetros de forma (α) e de taxa (β). Depois disto, foi ajustado um tipo de modelo para cada parâmetro. Para o parâmetro de forma α , utilizou-se o modelo de regressão linear:

$$y_{\alpha_{ij}} = a + b_1 x_i + b_2 x_i^2 + b_3 x_i^3 + \varepsilon_{ij}.$$

Onde, $y_{\alpha_{ij}}$ corresponde aos α ajustados; a corresponde ao intercepto; b_1 , b_2 e b_3 , correspondem aos coeficientes linear, quadrático e cúbico; x_i é o número

de alelos por haplótipo e ε_{ij} é o erro associado (SCHNEIDER; HOMMEL; BLETTNER, 2010).

As significâncias dos coeficientes de regressão foram avaliadas pelos seus valores de p e os modelos escolhidos foram os que possuíam os maiores valores para o R^2 ajustado.

Para o parâmetro de taxa β , utilizou-se um modelo de regressão não linear de Brody (BRODY, 1924), como segue:

$$y_{\beta_{ij}} = \theta_1(1 - \theta_2 \exp(-\theta_3 x_i)) + \varepsilon_{ij}.$$

Onde, $y_{\beta_{ij}}$ corresponde aos β ajustados; os coeficientes da regressão θ_1 é o limite assintótico que a curva poderá assumir, θ_2 é a proporção de valores iniciais que definem a angulação da curva e θ_3 representa a velocidade de crescimento da curva até atingir o limite assintótico; x_i é o número de alelos por haplótipo e ε_{ij} é o erro associado.

Os limiares baseados no resultado do Teste de Bonferroni foram estimados utilizando a função quantil de uma distribuição gama com os parâmetros $\hat{\alpha} = \frac{\hat{y}_{\alpha_i}}{\hat{y}_{\beta_i}^2}$ e $\hat{\beta} = \frac{\hat{y}_{\alpha_i}^2}{\hat{y}_{\beta_i}}$, onde $\hat{y}_{\alpha_i}$ são os valores preditos para o i-ésimo número de alelos obtidos do modelo selecionado referente ao parâmetro α e $\hat{y}_{\beta_i}^2$ são os valores preditos para o i-ésimo número de alelos obtidos do modelo selecionado referente ao parâmetro β .

4.9. Fórmula do desequilíbrio de ligação para marcadores multi-alelicos

Para verificar o desequilíbrio de ligação (DL) entre os haplótipos significativos com o QTL detectado e os marcadores que constituem os haplótipos significativos com o QTL detectado foi utilizada seguinte estatística para D' (ZHAO; NETTLETON; DEKKERS, 2007):

$$D' = \sum_{i=1}^k \sum_{j=1}^m p(A_i)p(B_j) \left| \frac{D_{ij}}{D_{ij}^{max}} \right|,$$

$$D_{ij} = p(A_i B_j) - p(A_i)p(B_j),$$

$$D_{ij}^{max} = \min \left[p(A_i)p(B_j), (1 - p(A_i))(1 - p(B_j)) \right] \text{ quando } D_{ij} < 0,$$

$$D_{ij}^{max} = \min \left[p(A_i) (1 - p(B_j)), (1 - p(A_i)) p(B_j) \right] \text{ quando } D_{ij} \geq 0.$$

E para o cálculo da correlação como coeficiente de DL (r^2) foi utilizada a seguinte estatística (ZHAO; NETTLETON; DEKKERS, 2007):

$$r^2 = \sum_{i=1}^k \sum_{j=1}^m p(A_i) p(B_j) r_{ij}^2,$$

$$r_{ij}^2 = \frac{D_{ij}^2}{p(A_i)(1 - p(A_i))p(B_j)(1 - p(B_j))}.$$

Onde, k e m são os números dos alelos alternados em um loco A e B , respectivamente, $p(A_i)$ é a frequência do alelo A_i no loco A , $p(B_j)$ é a frequência do alelo B_j no loco B e $p(A_i B_j)$ é a frequência do haplótipo $A_i B_j$.

5. RESULTADOS E DISCUSSÃO

5.1. Extensão do desequilíbrio de ligação das populações simuladas

O decaimento do DL dos dados simulados para as três herdabilidades foi suave e linear não diferindo entre si, e aproximando-se do que é encontrado na literatura para bovinos da raça Nelore (ESPIGOLAN et al., 2013; MELO, 2015). Em distâncias maiores do que 60 kb, o decaimento do LD (Figuras 5.1, 5.2 e 5.3) foi um pouco maior do que o documentado por Espigolan et al. (2013). Em dados reais, valores médios de DL, medido utilizando o r^2 , para bovinos Nelore variaram entre 0,17 (ESPIGOLAN et al. 2013) até 0,31 (FERNANDES JÚNIOR et al. 2016), no presente estudo, os valores médios do DL pertencem ao intervalo de 0,20 a 0,30. Portanto, pode-se considerar que as simulações de cada população se adequam ao que é observado em dados reais.

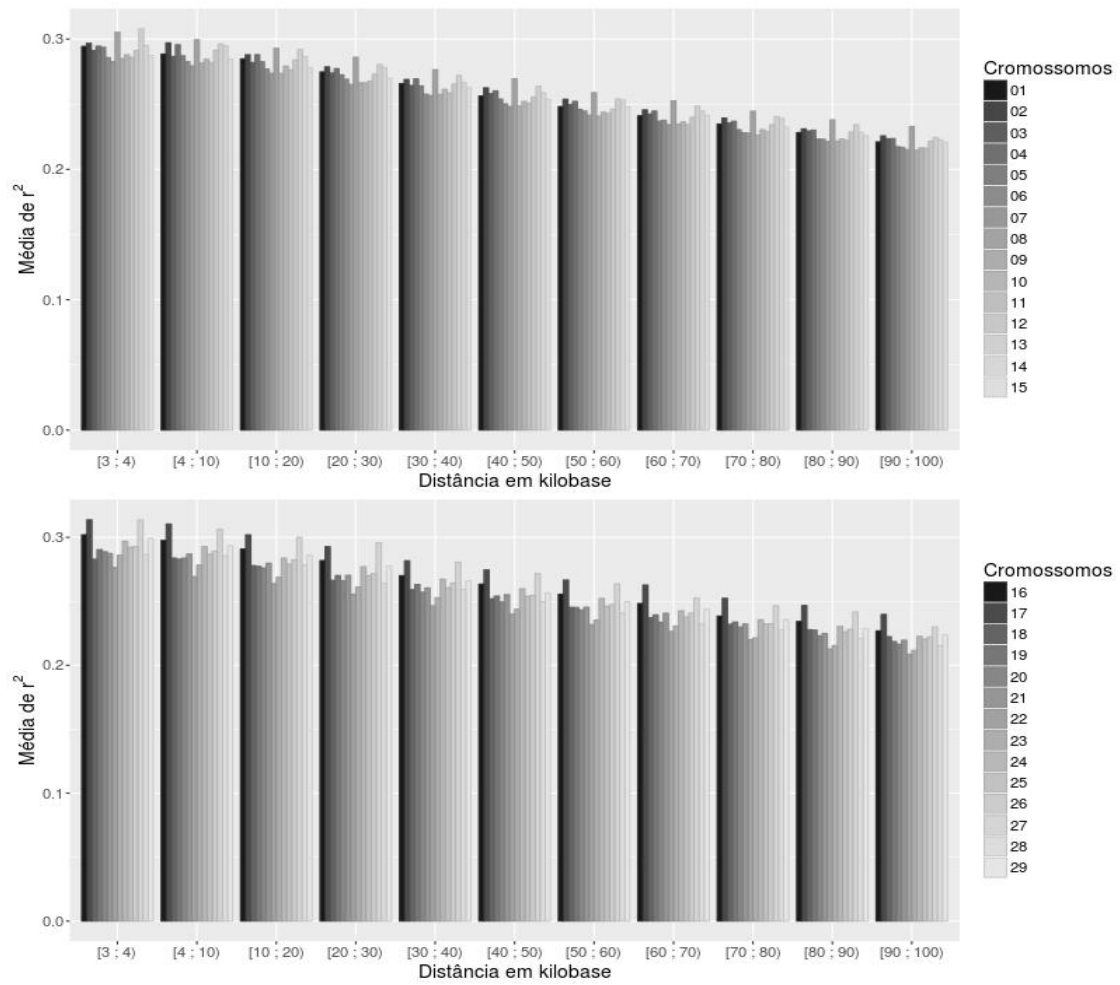


Figura 5.1. Gráfico do desequilíbrio de ligação (r^2) para a herdabilidade 0,10 referente aos cromossomos de 1 a 29 do genoma simulado.

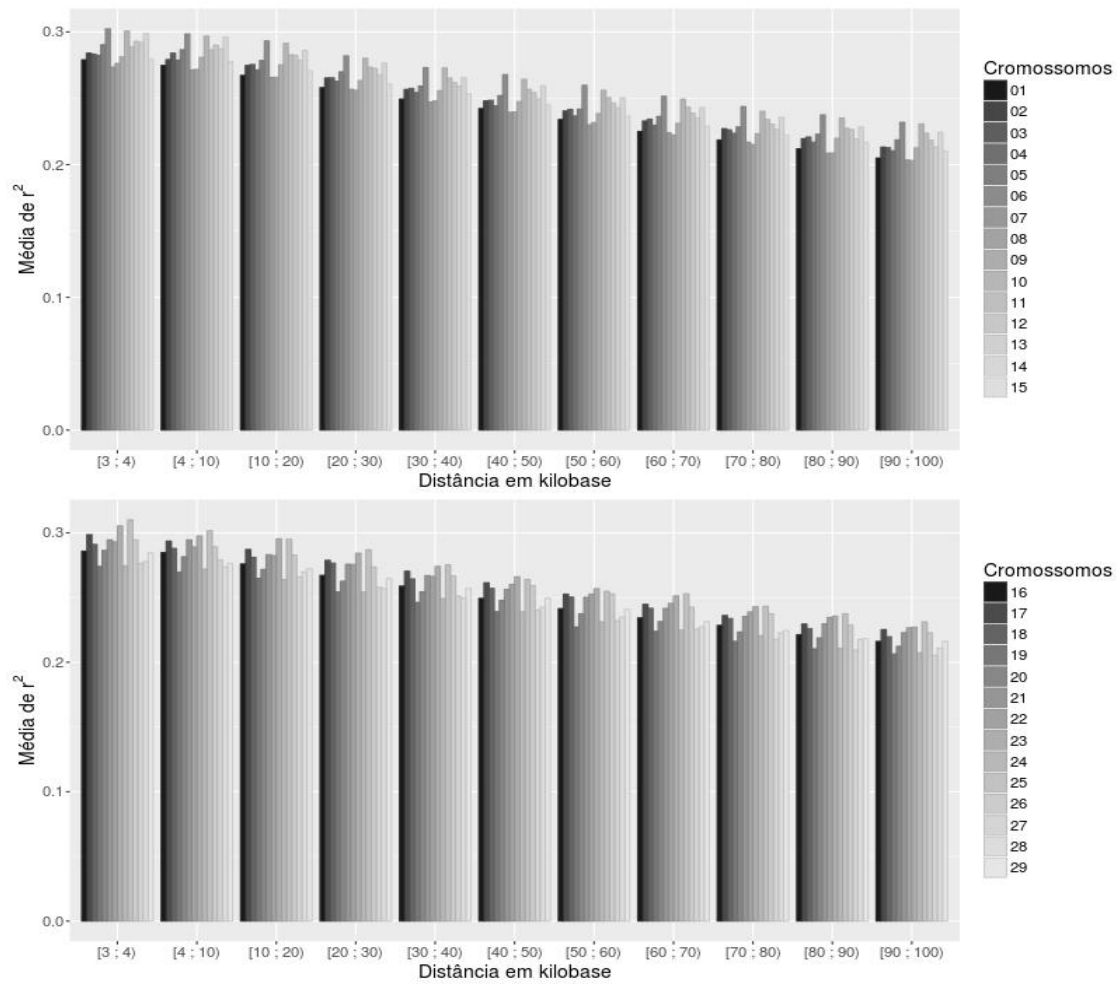


Figura 5.2. Gráfico do desequilíbrio de ligação (r^2) para a herdabilidade 0,25 referente aos cromossomos de 1 a 29 do genoma simulado.

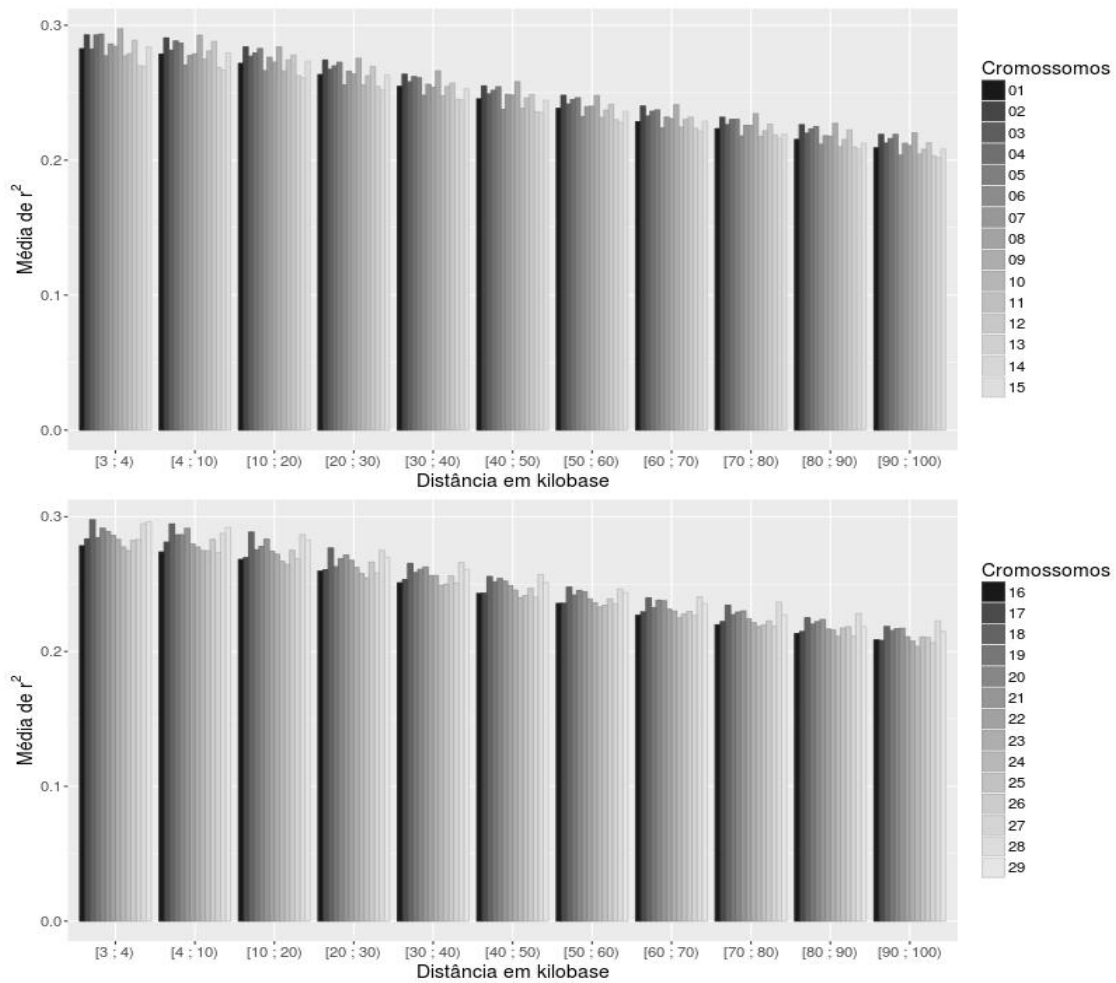


Figura 5.3. Gráfico do desequilíbrio de ligação (r^2) para a herdabilidade 0,35 referente aos cromossomos de 1 a 29 do genoma simulado.

5.2. Correlações entre as matrizes G

Na Tabela 5.1 observa-se que todas as correlações entre as matrizes G_{IC} e G_{RQG} com as matrizes G_{SNP} , para as três populações, possuem associação linear forte, indicando que as matrizes não são estatisticamente distintas.

Tabela 5.1. Correlação das matrizes genômicas construídas a partir dos métodos IC e RQG com as matrizes genômicas construídas a partir dos SNPs para as três populações simuladas.

Método	População		
	Pop1	Pop2	Pop3
IC	0,9974	0,9975	0,9977
RQG	0,9978	0,9979	0,9981

IC: Intervalo de Confiança; RQG: Regra de Quatro Gametas.

As correlações entre as matrizes G_{IC} e G_{RQG} foram estimadas em 0,9997. Este resultado era esperado, pois ambos os métodos possuem em comum a construção dos blocos de haplótipos baseada em eventos de recombinação que ocorrem no genoma (GABRIEL et al., 2002; WANG et al., 2002).

5.3. Análise de associação

Na análise de associação genômica ampla, foram encontrados haplótipos com efeitos significativos apenas nas populações Pop1 (herdabilidade de 0,10) e Pop3 (herdabilidade de 0,35). Para a população Pop1, todos os métodos de construção de haplótipos resultaram na identificação de haplótipos com efeitos significativos. Para a Pop3, apenas quando foi utilizado o método JS observou-se um bloco de haplótipo com efeito significativo, como pode ser visto na Tabela 5.2. Este resultado não era esperado uma vez que as três populações foram simuladas com estrutura genética muito semelhante e, como a variância fenotípica simulada era igual nas três populações, os maiores efeitos dos QTLs no caso das herdabilidades mais altas deveriam facilitar sua detecção.

Tabela 5.2. Número de haplótipos significativos por população, de acordo com o método de construção de haplótipo e cromossomo.

Populações	Método	Cromossomo	Total de haplótipos significativos
Pop1	IC	15	32
	RQG	15	52
		19	1
		29	1
	JS	15	200
Pop3	JS	11	1

IC: Intervalo de Confiança; RQG: Regra de Quatro Gametas; JS: Janelas Sobrepostas.

Na população Pop1, com o uso do método IC cada um dos 32 haplótipos significativos explicou entre 0,0070 a 0,0150 de variância aditiva, o que representou entre 0,14% e 30% da variância dos QTLs. Para os 52 haplótipos significativos do método RQG, o valor da variância genética explicada ficou entre 0,0061 e 0,0151, correspondendo entre 0,12% e 30% da variância dos QTLs. Com o uso das mesmas metodologias, Sodeland (2010) também detectou QTLs para características com herdabilidades baixas (0,02 a 0,13) e Oh et al. (2012) identificaram haplótipos significativos para características de baixa a moderada herdabilidades, em dados reais de bovinos. Com o uso do método JS os 200 haplótipos significativos explicaram variâncias aditivas de 0,0070 a 0,0150, equivalente a 0,14% e 30% da variância dos QTLs. Quando a população Pop3 foi analisada, apenas o método JS detectou haplótipo significativo, o qual explicou 0,0082 de variância genética aditiva, correspondendo a 0,05% da variância do QTL identificado.

Considerando que os haplótipos que detectaram QTLs são os mais próximos destes, pode-se observar que apesar do grande número de haplótipos significativos no cromossomo 15 da Pop1, foram detectados apenas cinco, oito e sete QTLs utilizando-se, os métodos IC, RQG e JS, respectivamente (Tabela 5.3). O método RQG foi o único a detectar o QTL Q386, com haplótipo constituído de dois SNPs e

três alelos, e também os QTLs Q511 e Q737, com haplótipos constituídos de três SNPs.

Na Tabela 5.3 observa-se que exceto o QTL Q387 no cromossomo 15 e o Q511 no cromossomo 19, cujas variâncias representavam pouco menos que 4,0% das variâncias simuladas dos QTLs na característica da Pop1, as variâncias dos QTLs detectados foram muito pequenas. Estes dois QTLs (Q387 e Q511) foram os que apresentaram maiores efeitos entre os QTLs simulados nesta população. O haplótipo significativo mais próximo do QTL Q737, cuja variância genética era essencialmente zero, estava também próximo do QTL Q738, com variância genética de 0,000217 (0,4% da variância simulada), ainda de pequena magnitude, mas suficiente para o haplótipo ser significativo. Observa-se ainda, que a maioria dos haplótipos significativos estavam em DL com o QTL Q387, bem como que para os três métodos de formação de haplótipos a máxima variância explicada pelo marcador foi dos haplótipos próximos ao QTL Q387. A região formada pelos QTLs Q384 a Q390, que compreende cerca de 11 cM do cromossomo 15 na Pop1, concentra mais de 99% dos haplótipos significativos. Os QTLs que estão mais próximos entre si são o Q387 e o Q388 com distância de 0,0325 cM e os mais distantes são os QTLs Q381 e Q390 com 12,5515 cM de distância. Deste conjunto de QTLs, apenas os pares Q384-Q385 e Q386-Q387 apresentam a estatística D' maior que 0,5 e r^2 maior que 0,10 indicando DL entre eles. Entretanto, pode-se considerar que a região que abrange o QTL Q384 até o QTL Q387 apresenta alguma ligação (Tabela 5.4), isto talvez explique o fato da região concentrar a maioria dos haplótipos significativos. Desta forma, supõe-se que apenas dois QTLs detectados foram de fato importantes (Q385 e Q387) e os haplótipos que detectaram outros QTLs são falso-positivos. Além disso, é importante ressaltar que o segundo QTL com maior variância simulada (Q511) foi detectado apenas por um haplótipo significativo formado pela RQG e, apesar disto, a variância do haplótipo não era particularmente grande.

Tabela 5.3. População, cromossomo, identificação dos QTL e variâncias genética aditiva simuladas dos QTLs detectados e, número de haplótipos significativos, valores máximos das medidas de desequilíbrio de ligação (r^2 e D') entre QTLs detectados e os haplótipos significativos e máxima variância do efeito estimado dos haplótipos significativos, de acordo com o método de formação dos haplótipos.

Pop	Cr	QTL	Var (QTL)	N hap	Maior r^2	Maior D'	Máx Var (hap)
Pop1	15	Q381	0,000000	IC(1)	IC(0,047)	IC(0,458)	IC(0,0084)
				RQG(1)	RQG(0,047)	RQG(0,475)	RQG(0,0085)
				JS(1)	JS(0,076)	JS(0,707)	JS(0,0068)
	15	Q384	0,000008	RQG(5)	RQG(0,228)	RQG(0,668)	RQG(0,0081)
				JS(1)	JS(0,286)	JS(0,886)	JS(0,0068)
	15	Q385	0,000016	RQG(4)	RQG(0,202)	RQG(0,702)	RQG(0,0070)
				JS(2)	JS(0,181)	JS(0,706)	JS(0,0069)
	15	Q386	0,000000	RQG(1)	RQG(0,208)	RQG(0,682)	RQG(0,0072)
	15	Q387	0,001937	IC(10)	IC(0,594)	IC(0,968)	IC(0,0150)
				RQG(13)	RQG(0,597)	RQG(0,995)	RQG(0,0151)
				JS(63)	JS(0,413)	JS(1,000)	JS(0,0131)
	15	Q388	0,000010	IC(11)	IC(0,406)	IC(0,931)	IC(0,0139)
				RQG(13)	RQG(0,226)	RQG(0,931)	RQG(0,0139)
				JS(77)	JS(0,212)	JS(0,937)	JS(0,0107)
	15	Q389	0,000000	IC(8)	IC(0,467)	IC(0,937)	IC(0,0122)
				RQG(12)	RQG(0,592)	RQG(0,979)	RQG(0,0108)
				JS(51)	JS(0,319)	JS(0,987)	JS(0,0121)
	15	Q390	0,000013	IC(2)	IC(0,310)	IC(0,812)	IC(0,0078)
				RQG(3)	RQG(0,249)	RQG(0,766)	RQG(0,0078)
				JS(2)	JS(0,275)	JS(0,844)	JS(0,0075)
	19	Q511	0,001896	RQG(1)	RQG(0,371)	RQG(0,885)	RQG(0,0068)
	29	Q737	0,000000	RQG(1)	RQG(0,133)	RQG(0,576)	RQG(0,0068)
Pop3	11	Q256	0,000071	JS(1)	JS(0,000)	JS(0,028)	JS(0,0081)

Pop: População; Cr: Cromossomo; Var (QTL): Variância do QTL; N hap: Número de haplótipos significativos; Máx Var (hap): Máxima variância dos haplótipos; QTL: *Quantitative Trait Loci*; IC: Intervalo de Confiança; RQG: Regra de Quatro Gametas; JS: Janelas Sobrepostas.

Estudos de associação genômica em que são utilizados haplótipos (WU et al., 2014; ZHANG et al., 2015; BRAZ, 2016) demonstram que é possível validar estudos de GWAS em que foram utilizando SNPs isolados e, inclusive, detectar QTLs que não foram detectados por SNPs isolados, como é observado no caso do QTL Q511. Tal observação também foi documentada em estudos realizados utilizando dados reais de frangos (ZHANG et al., 2015). Os autores estudaram a característica de peso em 396 frangos e utilizaram 46.665 SNPs nas análises. Os haplótipos construídos possuíam limiar de $r^2 > 0,8$ e o tamanho da janela foi de 200 kb. No total, foram formados 5.650 haplótipos sendo que 19 deles foram significativos e 7 deles foram constituídos com SNPs significativos. Na análise dos marcadores, 18 SNPs isolados também possuíam efeitos significativos. Em dados reais para ovelhas, Ren et al. (2016) estudaram o fenótipo “*polycerate*” em que foram utilizados dados de 50 animais, sendo 28 com dois chifres e 22 com quatro chifres, e o total de 491.507 SNPs. Foram identificados 10 SNPs e 20 haplótipos, construídos utilizando a metodologia RQG.

Resultados diferentes foram encontrados por Wu et al (2014). Os autores utilizando SNPs detectaram um número maior de variáveis significativas do que com o uso dos haplótipos. A suspeita para tal é que os SNPs adjacentes estejam na mesma região o que diminuiu o número de associações significativas. No presente estudo, vários SNPs significativos constituíam o mesmo haplótipo. No estudo de Wu et al (2014) alguns QTLs somente foram identificados utilizando SNPs isolados e outros utilizando os haplótipos. As características estudadas foram peso do corte de músculo dianteiro (*foreshank*) e níveis de triglicerídeos, em 942 bovinos da raça Simmental, sendo os haplótipos construídos pelo método de Intervalo de Confiança.

Para a população Pop3, o haplótipo significativo está localizado próximo do QTL Q257 que possui variância 0,000104 e corresponde a 0,2% do total da variância simulada, podendo ser o efeito combinado entre ambos a razão da significância do haplótipo.

Para os SNPs, ao nível de significância de 5%, a Pop1 foi a única população em que se detectou SNPs significativos, todos dispostos no cromossomo 15. No total, 42 marcadores SNPs foram significativos, e todos estavam localizados na mesma região, próximos dos QTLs Q384, Q387, Q388 e Q389. Dessa forma, no

presente trabalho, os modelos de associação utilizando haplótipos resultaram em um número maior de QTLs detectados do que os modelos de associação utilizando SNPs isolados. Em estudos de simulação, Hayes et al. (2007) também observaram em estudo de GWAS que com a utilização de blocos de haplótipos é possível obter melhor performance comparado ao uso de SNPs isolados. O resultado encontrado corrobora com os encontrados por Braz (2016), o qual observou que haplótipos baseados em Janelas Sobrepostas de cinco SNPs detectaram mais QTLs quando comparados com análises de associações de SNPs isolados, utilizando uma característica de baixa herdabilidade em Nelore.

Tabela 5.4. Desequilíbrio de ligação, D' e r^2 , dos QTL detectados no cromossomo 15 da população Pop1.

QTL	D'	r^2
Q381 e Q384	0,337	0,050
Q384 e Q385	0,523	0,156
Q385 e Q386	0,479	0,045
Q386 e Q387	0,805	0,177
Q387 e Q388	0,259	0,009
Q388 e Q389	0,001	0,000
Q389 e Q390	0,116	0,012

QTL: *Quantitative Trait Loci*.

Analisou-se a razão pela qual alguns métodos não identificaram QTLs que outros métodos identificaram. Para os QTLs identificados apenas pelo método RQG os seguintes casos foram observados: em relação ao QTL Q386, não há blocos de haplótipos construídos na posição do QTL pelo método IC e pelo método JS o bloco de haplótipo presente nesta região possui cinco SNPs e nove alelos, diferentemente do haplótipo significativo formado pelo método RQG. Em relação ao QTL Q511, o segundo em maior variância simulada, utilizando o método IC não foi formado bloco na posição do QTL e o bloco formado pelo método JS constituído de cinco SNPs e seis alelos não conseguiu detectar o efeito do QTL. Na posição do QTL Q737, foi

formado um bloco de haplótipo com seis SNPs e oito alelos pelo método IC, entretanto, este não foi significativo nos níveis de significância testados (5% e 10%).

Os QTLs Q384 e Q385 foram detectados pelos métodos RQG e JS, sendo que com o uso do método IC pode não ter sido possível identifica-los devido ao bloco formado na posição do QTL Q384 consistir de quatro SNPs e nove alelos e na posição do QTL Q385 não foi formado bloco de haplótipo.

Para a população Pop3, apenas o método JS conseguiu detectar um QTL (Q256) no cromossomo 11. Na mesma posição, foram construídos haplótipos pelos métodos IC e RQG, ambos formados por dois SNPs e três alelos, o que pode ter afetado na não identificação do QTL, visto que o haplótipo que o identificou pelo método JS foi composto de cinco SNP.

Em virtude do valor de p de Bonferroni ao nível de significância de 5% ser muito restritivo, os haplótipos de todos os métodos foram reanalisados utilizando o nível de significância de 10%, sendo que os resultados foram semelhantes. Para a Pop1, embora o número de haplótipos no cromossomo 15 tenha aumentado, apenas um novo QTL (Q402) foi identificado por um haplótipo formado pelo método RQG em região cerca de 30 cM distante do Q390 e com variância simulada de 0,000017. Este resultado indica que diminuir o nível de significância de 5% para 10% apenas aumentou o número de QTLs falso-positivos. Para os marcadores SNPs, não houve QTL identificado além dos documentados anteriormente, assim como para as outras duas populações.

A eficácia do uso de métodos de construção de haplótipos depende diretamente do número de animais presentes no estudo e da densidade de marcadores (HAMBLIN; JANNINK, 2011). No presente estudo, foram utilizados mais animais e maior densidade de marcadores do que é documentado na literatura (ZHANG et al., 2015; SALEM et al., 2018), os haplótipos também demonstraram ter desempenho melhor que SNPs isolados. Com alguns métodos de construção de haplótipos é possível obter maior eficiência em estudos de GWAS utilizando menos variáveis nos modelos de predição, o que resulta em maior poder de detecção.

6. CONCLUSÃO

Para a estrutura das populações aqui simuladas, apesar da baixa eficiência de todos os métodos para a detecção dos QTLs, os resultados indicam que o método de formação de haplótipos baseado na Regra de Quatro Gametas é o mais adequado para detecção de QTLs em relação aos métodos Intervalo de Confiança e Janelas Sobrepostas, e também ao uso dos SNPs isolados.

7. REFERÊNCIAS

BALL, A. D.; STAPLEY, J.; DAWSON, D. A.; BIRKHEAD, T. R.; BURKE, T.; SLATE, J. A comparison of SNPs and microsatellites as linkage mapping markers: lessons from the zebra finch (*Taeniopygia guttata*). **BMC Genomics**, v. 11, p. 218, 2010.

BARENDSE, W. Haplotype analysis improved evidence for candidate genes for intramuscular fat percentage from a genome wide association study in cattle. **Plos One**, v. 6, e29601, 2011.

BARRETT, J.; FRY, B.; MALLER, J.; DALY, M. Haploview: analysis and visualization of LD and haplotype maps. **Bioinformatics**, v. 21, p. 263–265, 2004.

BRAZ, C. U. **Detecção de QTL para Maciez da Carne em Bovinos da Raça Nelore**. 2016. 81 f. Tese (Doutorado em Melhoramento Genético Animal) – Faculdade de Ciências Agrárias e Veterinárias, Universidade Estadual Paulista “Júlio de Mesquita Filho”, Jaboticabal, 2016.

BRITO, F.; NETO, J.; SARGOLZAEI, M.; COBUCI, J.; SCHENKEL, F. S. Accuracy of genomic selection in simulated populations mimicking the extend of linkage disequilibrium in beef cattle. **BioMed Central Genetics**, v.12, p. 80, 2011.

BRITO, L. F.; JAFARIKIA, M.; GROSSI, D. A.; KIJAS, J. W.; PORTO-NETO, L. R.; VENTURA, R. V.; SARGOLZAEI, M.; SCHENKEL, F. S. C. Characterization of linkage disequilibrium, consistency of gametic phase and admixture in Australian and Canadian goats, **BMC Genetics**, v. 16, p. 67, 2015.

BRODY, S.; RAGSDALE, A. C.; TURNER, C. W. The relation between the initial rise and the subsequent decline of milk secretion following parturition. **The Journal of General Physiology**. V.6, p.541-545,1924.

BUSH, W.; MOORE, J. Genome-Wide Association Studies. **PLOS Computational Biology**, v. 8, 2012.

CALUS, M.; MEUWISSEN, T.; ROOS, A.; VEERKAMP, R. Accuracy of Genomic Selection Using Different Methods to Define Haplotypes. **Genetics Society of America**, v. 178, p. 553–561, 2008.

CALUS, M. MEUWISSEN, T.; WINDING, J.; KNOL, E. SCHROOTEN, C.; VEREIJKEN, A.; VEERKAMP, R. Effects of the number of markers per haplotype and clustering of haplotypes on the accuracy of QTL mapping and prediction of genomic breeding values. **Genetics Selection Evolution**, v. 41, n. 1, p. 11, 2009.

CARDON, L.; ABECASIS, G. Using haplotype blocks to map human complex trait loci. **Elsevier Science**, v. 19, p. 135–140, 2003.

CHENG, R.; MA, J.; ELSTON, R.; LI, M. Fine mapping functional sites or regions from case-control data using haplotypes of multiple linked SNPs. **Annals of Human Genetics**, v. 9, p. 102–112, 2005.

CUYABANO, B.; SU, G.; LUND, M. Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. **BioMed Central**, v. 15, p. 1171–1182, 2014.

ESPIGOLAN, R.; BALDI, F.; BOLIGON, A.; SOUZA, F.; GORDO, D.; TONUSSI, R.; CARDOSO, D. F.; OLIVEIRA, H.; TONHATI, H.; SARGOLZAEI, M.; SCHENKEL, F.; CARVALHEIRO, R.; FERRO, J.; ALBUQUERQUE, L. Study of whole genome linkage disequilibrium in Nellore cattle. **BMC Genomics**, v. 14, n. 1, p. 305, 2013.

FERNANDES JÚNIOR, G. A.; ROSA, G. J. M.; VALENTE, B. D.; CARVALHEIRO, R.; BALDI, F.; GARCIA, D. A.; GORDO, D. G. M.; ESPIGOLAN, R.; TAKADA, L.; TONUSSI, R.L.; ANDRADE, W. B. F.; MAGALHÃES, A. F. B.; CHARDULO, L. A. L.; TONHATI, H.; ALBUQUERQUE, L. G. Genomic prediction of breeding values for carcass traits in Nellore cattle. **Genetics Selection Evolution**, v.48, n.7, 2016.

GABRIEL, S.; SCHAFFNER, S.; NGUYEN, H.; MOORE, J.; JOY, J.; BLUMENSTIEL, B.; HIGGINS, J.; DEFELICE, M.; LOCHNER, A.; FAGGART, M.; LIU-CORDERO, S.; ROTIMI, C.; ADEYEMO, A.; COOPER, R.; WARD, R.; LANDER, E.; DAYLY, M.; ALTSHULER, D. The Structure of Haplotype Blocks in the Human Genome. **Science**, v. 296, p. 2225–2229, 2002.

GAO, Q.; YUAN, Z.; HE, Y.; ZHAO, J.; ZHANG, X.; LI, F.; ZHANG, B.; XUE, F. Exhaustive Sliding-Window Scan Strategy for Genome-Wide Association Study via PCA-Based Logistic Model. **Global Journal of Science Frontier Research Bio-Tech and Genetics**, v. 12, p. 4–15, 2012.

GODDARD, M. E.; HAYES, B. J. Genome-Wide Association Studies and Linkage Disequilibrium in Cattle. In: **Bovine Genomics**. Wiley-Blackwell, 2012. p. 192–210.

GUO, Y.; LI, J.; BONHAM, A.; WANG, Y.; DENG, H. Gain in power for exhaustive analyses of haplotypes using variable-sized sliding window strategy: a comparison of association-mapping strategies. **European Journal of Human Genetics**, v. 17, p. 785–792, 2009.

HAMBLIN, M.; JANNINK, J. Factors affecting the power of haplotype markers in association studies. **Nature Methods**, v. 4, p. 145–153, 2011.

HAYES, B.; CHAMBERLAIN, A.; MCPARTLAN, H.; MACLOEOD, I.; SETHURAMAN, L.; GODDARD, M. Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. **Genetics Research Cambridge**, v. 89, p. 215–220, 2007.

HAYES, B.; GODDARD, M. The distribution of the effects of genes affecting quantitative traits in livestock. **Genetics Selection Evolution**, v. 33, p. 209–229, 2001.

HUDSON, R.; KAPLAN, N. Statistical Properties of the Number of Recombination Events in the History of a Sample of DNA Sequences. **Genetics**, v. 111, p. 147–164, 1985.

IKEGAWA, S. A short history of the Genome-Wide Association Study: where we were and where we are going. **Genomics & Informatics**, v. 10, p. 220–225, 2012.

LI, Y.; SUNG, W.; LIU, J. L. Association Mapping Via Regularized Regression Analysis of Single-Nucleotide-Polymorphism Haplotypes in Variable-Sized Sliding Windows. **The American Journal of Human Genetics**, v. 80, p. 706-715, 2007.

LONG, A.; LANGLEY, C. The Power of Association Studies to Detect the Contribution of Candidate Genetic Loci to Variation in Complex Traits. **Genome Ressearch**, v. 9, p. 720–731, 1999.

LORENZ, A.; HAMBLIN, M.; JANNINK, J. Performance os Single Nucleotide Polymorphism versus Haplotypes for Genome-Wide Association Analysis in Barley. **Plos One**, v. 5, p. 1–11, 2010.

LYNCH, M.; WALSH, B. **Genetics and Analysis of Quantitativa Traits**. 1. ed. Sunderland: Sinauaer, 1998.

MATHIAS, E.; GAO, P.; GOLDSTEIN, J.; WILSON, A.; PUGH, E.; FURBERT-HARRIS, P.; DUNSTON, G.; MALVEAUX, F.; TOGIAS, A.; BARNES, K.; BEATY, T.; HUANG, S. A graphical assessment of p-values from sliding window haplotype tests of association to identify asthma susceptibility loci on chromosome 11q. **BMC Genetics**, v. 7, 2006.

MATUKUMALLI, L.K.; SCHOROEDER, S.; DENISE, S.K.; SONSTEGARD, T.; LAWLEY, C.T.; GEORGES, M.; COPPIETERS, W.; GIETZEN, K.; MEDRANO, J. F.; RINCON, G.; LINCE, D.; EGGEN, A.; GLASER, L.; CAM, G.; VAN TASSEL, C. V. Analyzing LD blocks and CNV segments in cattle: Novel genomic features identified using the BovineHD BeadChip. **Illumina Inc**, n. 370-2011-002, 2011.

MEGENS, H.; CROOJIMANS, R. P. A.; BASTIAASEN, J. W. M.; KERSTENS, H. H. D.; COSTER, A.; JALVING, R.; VEREIJKEN, A.; SILVA, P.; MUIR, W. M.; CHENG, H. H.; HANOTTE, O.; GROENEN, M. A. M. Comparison of linkage disequilibrium and haplotype diversity on macro- and microchromosomes in chicken. **BMC Genetics**, v. 10, p. 86, 2009.

MCKAY, S. D.; SCHANABEL, R. D.; MURDOCH, B. M.; MATUKUMALI, L. K.; AERTS, J.; COPPIETERS, W.; CREWS, D.; NETO, E. D.; GILL, C. A.; GAO, CH.; MANNEN, H.; STOTHARD, P.; WANG, Z.; VAN TASSEL, C. P.; WILLIAMS, J. L.; TAYLOR, J. F.; MOORE, S. S. Whole genome linkage disequilibrium maps in cattle. **BMC Genetics**, v. 8, p. 74, 2007.

MELO, T. P. DE. **Genome-Wide Association Study of Reproduction Traits in Nelore Cattle, Including Additional Phenotypic Information from Non-Genotyped Animals**. 2015. 73 f. Tese (Doutorado em Genética e Melhoramento Animal) - Faculdade de Ciências Agrárias e Veterinárias Universidade Estadual Paulista “Júlio de Mesquita Filho”, Jaboticabal, 2015.

OH, D.; LEE, Y.; LA, B.; YEO, J. Identification of the SNP (Single Nucleotide Polymorphism) for Fatty Acid Composition Associated with Beef Flavor-related FABP4 (Fatty Acid Binding Protein 4) in Korean Cattle. **Asian-Australasian Journal of Animal Sciences**, v. 25, p. 913-920, 2012.

PORTO-NETO, L. R.; KIJAS, J. W.; REVERTER, A. The extent of linkage disequilibrium in beef cattle breeds using high-density SNP genotypes. **Genetics Selection Evolution**, v. 46, p. 22, 2014.

R CORE TEAM (2016). **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

REN, X.; YANG, G. L.; PENG, W. F.; ZHAO, Y. X.; ZHANG, M.; CHEN, Z. H.; WU, F. A.; KANTANEN, J.; SHEN, M.; LI, M. H. A genome-wide association study identifies a genomic region for the polycerate phenotype in sheep (*Ovis aries*). **Scientific Reports**, v. 6, p. 21111, 2016.

SALEM, M. M. I.; THOMPSON, G.; CHEN, S.; BEJA-PEREIRA, A.; CARVALHEIRA, J. Linkage Disequilibrium and Haplotype Block Structure in Portuguese Holstein Cattle. **Czech Journal of Animal Science**, v. 2, p.61 – 69, 2018.

SARGOLZAEI, M.; SCHENKEL, F. QMSim: a large-scale genome simulator for livestock. **Bioinformatics**, v. 25, p. 680–681, 2009.

SCHNEIDER, A.; HOMMEL, G.; BLETTNER, M. Linear regression analysis—part 14 of a series on evaluation of scientific publications. **Pesquisa Agropecuária Brasileira**, v. 107, p. 76 – 82, 2010.

SERÃO, N.; GONZÁLEZ-PEÑA, D.; BEEVER, J. E.; FAULKNER, D. B.; SOUTHEY, B. R. RODRIGUEZ-ZAS, S. L. Single nucleotide polymorphisms and haplotypes associated with feed efficiency in beef cattle. **BMC Genetics**, v. 14, p. 94, 2013.

SHIM, H.; CHUN, H.; ENGELMAN, C. D.; PAYSEUR, B. A. Genome-wide association studies using single-nucleotide polymorphisms versus haplotypes: an empirical comparison with data from the North American Rheumatoid Arthritis Consortium. **BMC Proceedings**, v. 3, p. 35, 2009.

SODELAND, M. **Genomic architecture and complex traits in Norwegian Red cattle**. 2010. 100 f. Tese (Doutorado em Filosofia) - Norwegian University of Life Sciences, 2010.

VISSCHER, P. et al. Review: Five Years of GWAS Discovery. **The American Journal of Human Genetics**, v. 90, p. 7–24, 2012.

WANG, L.; SORENSEN, P.; JANSS, L.; OSTERSEN, T.; EDWARDS, D. Genome-wide and local pattern of linkage disequilibrium and persistence of phase for 3 Danish pig breeds. **BMC Genetics**, v. 14, p. 115, 2013.

WANG, N.; AKEY, J. M.; ZHANG, K.; CHAKRABORTY, R.; JIN, L. Distribution of Recombination Crossovers and the Origin of Haplotype Blocks: The Interplay of Population History, Recombination, and Mutation. **The American Society of Human Genetics**, v. 71, p. 1227–1234, 2002.

WANG, S.; WEN, Y.; REN, W.; NI, Y.; ZHANG, J.; FENG, J.; ZHANG, Y. Mapping small-effect and linked quantitative trait loci for complex traits in backcross or DH populations via a multi-locus GWAS methodology. **Scientific Reports**, v. 6, p. 29951, 2016.

WU, Y.; FAN, H.; WANG, Y.; ZHANG, L.; GAO, X.; CHEN, Y.; LI, J. Genome-Wide Association Studies Using Haplotypes and Individual SNPs in Simmental Cattle. **Plos One**, v. 9, p. 1–11, 2014.

ZENG, P. et al. Statistical analysis for genome-wide association study. **The Journal of Biomedical Research**, v. 4, p. 285–297, 2015.

ZHANG, C. X.; FANI, Q. C.; ZHANG, J. Y.; WANG, W. H.; XUE, Q.; WANG, Y. J. Genome-wide association study of growth traits in the Jinghai Yellow chicken. **Genetics and Molecular Search**, v. 4, p. 15331 – 15338, 2015.

ZHANG, W.; LI, J.; GUO, Y.; ZHANG, L.; XU, L.; GAO, X.; ZHU, B.; GAO, H.; NI, H.; CHEN, Y. Multi-strategy genome-wide association studies identify the DCAF16-NCAPG region as a susceptibility locus for average daily gain in cattle. **Scientific Reports**, v. 6, p. 38073, 2016.

ZHAO, H.; NETTLETON, D.; DEKKERS, J. C. M. Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between single nucleotide polymorphisms. **Genetics Research**, v. 89, p. 1 – 6, 2007.

ZHOU, X.; STEPHENS, M. Genome-wide efficient mixed-model analysis for association studies. **Nature Genetics**, v. 44, p. 821–824, 2012.