

**UNIVERSIDADE ESTADUAL PAULISTA - UNESP
CÂMPUS DE JABOTICABAL**

**EFEITO DA UTILIZAÇÃO DE DIFERENTES CRITÉRIOS
DE CONTROLE DE QUALIDADE DOS GENÓTIPOS EM
ESTUDOS DE ASSOCIAÇÃO E SELEÇÃO GENÔMICA
AMPLA**

Tiago Bresolin

Zootecnista

2015

**UNIVERSIDADE ESTADUAL PAULISTA - UNESP
CÂMPUS DE JABOTICABAL**

**EFEITO DA UTILIZAÇÃO DE DIFERENTES CRITÉRIOS
DE CONTROLE DE QUALIDADE DOS GENÓTIPOS EM
ESTUDOS DE ASSOCIAÇÃO E SELEÇÃO GENÔMICA
AMPLA**

Tiago Bresolin

Orientador: Profa. Dr. Lucia Galvão de Albuquerque

Co-orientador: Dr. Roberto Carneiro

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

2015

Bresolin, Tiago
B842e Efeito da utilização de diferentes critérios de controle de qualidade dos genótipos em estudos de associação e seleção genômica ampla / Tiago Bresolin. -- Jaboticabal, 2015
xix, 46 p.; 28cm

Dissertação (Mestrado) -- Universidade Estadual Paulista, Faculdade de Ciências Agrárias e Veterinárias, 2015.

Orientador: Lucia Galvão de Albuquerque
Banca examinadora: Fernando Sebastián Baldy Rey e Rusbel Raul Aspilcueta Borquis.
Bibliografia

1. Análise genômica. 2. Acurácia de predição. 3. Bovinos de corte. 4. Predições genômicas. 5. Marcadores moleculares. 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 – Serviço Técnico de Biblioteca e Documentação - UNESP, Câmpus de Jaboticabal.



UNIVERSIDADE ESTADUAL PAULISTA

CAMPUS DE JABOTICABAL

FACULDADE DE CIÊNCIAS AGRÁRIAS E VETERINÁRIAS DE JABOTICABAL

CERTIFICADO DE APROVAÇÃO

TÍTULO: EFEITO DA UTILIZAÇÃO DE DIFERENTES CRITÉRIOS DE CONTROLE DE QUALIDADE DOS GENÓTIPOS EM ESTUDOS DE ASSOCIAÇÃO E SELEÇÃO GENÔMICA AMPLA

AUTOR: TIAGO BRESOLIN

ORIENTADORA: Profa. Dra. LUCIA GALVAO DE ALBUQUERQUE

CO-ORIENTADOR: Prof. Dr. ROBERTO CARVALHEIRO

Aprovado como parte das exigências para obtenção do Título de MESTRE EM GENÉTICA E MELHORAMENTO ANIMAL , pela Comissão Examinadora:


Profa. Dra. LUCIA GALVAO DE ALBUQUERQUE

Departamento de Zootecnia / Faculdade de Ciências Agrárias e Veterinárias de Jaboticabal


Prof. Dr. FERNANDO SEBASTIÁN BALDI REY

Departamento de Zootecnia / Faculdade de Ciências Agrárias e Veterinárias de Jaboticabal


Prof. Dr. RUSBEL RAUL ASPILCUETA BORQUIS

Universidade Federal da Grande Dourados / Dourados/MS

Data da realização: 30 de julho de 2015.

DADOS CURRICULARES DO AUTOR

Tiago Bresolin, nascido em Caxambu do Sul - SC em 06 de janeiro de 1987, filho de Divaldino João Bresolin e Rilde Pagliari Bresolin, iniciou o curso de Zootecnia na Universidade Estadual de Santa Catarina (UDESC), Chapecó - SC em Fevereiro de 2009 e tornou-se Zootecnista em Julho de 2013 pela Universidade Federal de Santa Maria (UFSM), Santa Maria - RS. Durante a graduação foi bolsista CNPq e realizou iniciação científica sob orientação do Prof. Dr. Paulo Roberto Nogara Rorato. Em Agosto de 2013, ingressou no curso de Pós-Graduação em Genética e Melhoramento Animal na Faculdade de Ciências Agrárias e Veterinárias (UNESP), Jaboticabal - SP, sob orientação da Profa. Dra. Lucia Galvão de Albuquerque. De Outubro de 2014 a Março de 2015, realizou estágio de Mestrado no exterior na University of Wisconsin (UW), Madison - WI, sob orientação do Prof. Dr. Guilherme Jordão de Magalhães Rosa.

"Existem muitas hipóteses na ciência que estão erradas. Isso é perfeitamente aceitável, elas são a abertura para achar as que estão certas."

Carl Sagan

Aos meus queridos pais, Divaldino e Rilde.

A minha irmã Sinara.

A Camila, pelo companherismo e dedicação.

A todos aqueles que me ajudaram na caminhada até aqui.

Dedico e Ofereço.

AGRADECIMENTOS

À Deus, por iluminar a minha caminhada.

À Profa. Dra. Lucia Galvão de Albuquerque, pela orientação, amizade, confiança e pelas oportunidades de crescimento pessoal e profissional a mim oferecidas.

Aos membros da banca examinadora, Prof. Dr. Fernando Sebastián Baldi Rey e Prof. Dr. Rusbel Raul Aspilcueta Borquis, pelos importantes questionamentos e sugestões.

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pela concessão da bolsa de Mestrado (Processo nº 2013/26264-4) e da bolsa de estágio de pesquisa no exterior (Processo nº 2014/14431-6).

Ao Prof. Dr. Guilherme Jordão de Magalhães Rosa e Bruno Dourado Valente pela orientação e valiosas discussões e sugestões durante o estágio na University of Wisconsin.

À Dionéia Magda Everling pelo companherismo, ensinamentos e principalmente pelo incentivo durante os anos que trabalhamos juntos.

Ao Prof. Dr. Paulo Rorato pelos ensinamentos, confiança, amizade e oportunidades oferecidas.

Aos todos os amigos que passaram e que estão na salinha, pelos momentos de aprendizado, trabalho, descontração e comemorações.

A todos os meus amigos que encontrei na cidade de Madison e na University of Wisconsin. Em especial, deixo meu abraço de agradecimento à Tiago, Arthur, Gerson e Família, Mariana, Keiinch, Angelina e Fernando.

SUMÁRIO

SUMÁRIO	i
CAPÍTULO 1 CONSIDERAÇÕES GERAIS	7
1.1 Introdução	7
1.2 Revisão de Literatura	9
1.2.1 Critérios de Controle de Qualidade	9
1.2.2 Seleção Genômica Ampla	10
1.2.3 Associação Genômica Ampla	11
1.2.4 Modelos Estatísticos	12
1.3 Objetivos	15
CAPÍTULO 2 EFEITO DO CONTROLE DE QUALIDADE, DA DENSIDADE E DA FREQUÊNCIA ALÉLICA DOS MARCADORES NA ACURÁCIA DE PREDIÇÃO GENÔMICA DE CARACTERÍSTICAS COMPLEXAS NA RAÇA NELORE	17
2.1 Introdução	18
2.2 Material e Métodos	19
2.3 Resultados e Discussão	24
2.4 Conclusão	28
CAPÍTULO 3 EFEITO DO CONTROLE DE QUALIDADE DOS MARCADORES NOS ESTUDOS DE ASSOCIAÇÃO GENÔMICA AMPLA EM BOVINOS DA RAÇA NELORE	29
3.1 Introdução	30
3.2 Material e Métodos	31
3.3 Resultados e Discussões	35
3.4 Conclusões	39
REFERÊNCIAS	41

EFEITO DA UTILIZAÇÃO DE DIFERENTES CRITÉRIOS DE CONTROLE DE QUALIDADE DOS GENÓTIPOS EM ESTUDOS DE ASSOCIAÇÃO E SELEÇÃO GENÔMICA AMPLA

RESUMO - Na seleção genômica os estudos se concentram em estimar a habilidade ou acurácia de predição genômica. A acurácia de predição pode ser afetada por muitos fatores como a herdabilidade da característica, o tamanho efetivo da população, a densidade dos marcadores, o tamanho do genoma, o número de loci afetando a característica e distribuição dos seus efeitos, o desequilíbrio de ligação e o tipo e número de informações fenotípicas usadas nas análises estatísticas. Contudo, muitos desses fatores podem ser direta ou indiretamente afetados pela edição do conjunto de marcadores, podendo influenciar a acurácia de predição genômica. Para os estudos de associação genômica ampla a falta de qualidade dos dados genômicos pode levar a resultados falso-negativos ou positivos na fase de descoberta das regiões associadas a características de importância econômica. Além disso, o controle de qualidade pode levar a perdas de marcadores que capturam uma proporção significativa da variância genética. Sendo assim, este estudo teve como objetivos avaliar o efeito do controle de qualidade, da densidade e da frequência alélica dos marcadores sobre a seleção genômica e o efeito do controle de qualidade dos marcadores nos estudos de associação genômica ampla, em bovinos da raça Nelore. Para isso foram utilizadas informações fenotípicas, genotípicas e de pedigree provenientes de fazendas que integram os programas de melhoramento genético da DeltaGen e CRV Paint. Foram utilizadas as características idade ao primeiro parto, peso à desmama e peso ao sobrelano. Os animais foram genotipados com o painel Illumina Bovine HD, que contém, aproximadamente, 777.000 SNPs. Os fenótipos foram pré-corrigidos para os efeitos fixos, utilizando um modelo animal uni-característica. Para a seleção genômica foram criados 1) três cenários com diferentes limiares de exclusão dos SNPs de acordo com os critérios de controle de qualidade, 2) sete sub painéis com diferentes densidades de marcadores, variando de 10.000 a 734.271 SNPs e 3) cinco sub grupos de marcadores com base na menor frequência alélica. Na seleção genômica foram utilizadas as metodologias “Genomic Best Linear Unbiased Predictor” (GBLUP) e BayesC π . A acurácia de predição foi calculada pela correlação de Pearson entre os valores observados (pseudo-fenótipos) e os preditos (pseudo-fenótipos preditos), dividido pela raiz quadrada da herdabilidade da característica. Os diferentes cenários e situações foram comparadas por meio de um processo de validação cruzada com cinco grupos. Na associação genômica ampla foram criados três cenários com diferentes limiares

de exclusão dos SNPs de acordo com os critérios de controle de qualidade. Para estimar os efeitos dos marcadores foi utilizada a metodologia BayesC π . Diferenças na proporção da variância genética total explicada para cada janela de 1 Mb foram consideradas para comparar os diferentes cenários. Independente da metodologia e característica utilizada para seleção genômica, a diminuição do número de marcadores disponíveis para as análises não teve efeito significativo sobre as acurácias de predição genômicas entre os cenários do controle de qualidade. Os sub painéis, com diferentes densidades de marcadores, apresentaram acurácia de predição genômica semelhantes para todas as características estudadas. O sub grupo de marcadores com as menores frequências alélicas apresentaram menores acurácias de predição comparado com os demais sub grupos. Nos estudos de associação genômica ampla foi verificado efeito do controle de qualidade sobre a proporção da variância genética explicada por cada janela de 1 Mb. Na classificação das dez janelas que explicam uma maior proporção da variância genética foi verificado que as janelas alteram sua posição entre os cenários de maior e menor rigor no controle de qualidade, assim como entre os cenários de maior rigor e sem controle de qualidade. Entre os cenários de menor rigor e sem controle de qualidade as alterações na classificação das dez principais janelas foram menores comparada com a classificação nas demais janelas.

Palavras-chave: análise genômica, acurácia de predição, bovinos de corte, predições genômicas, marcadores moleculares

EFFECT OF THE DIFFERENT QUALITY CONTROL CRITERIA OF MARKERS ON GENOME-WIDE ASSOCIATION STUDIES AND GENOMIC SELECTION

ABSTRACT - In the genomic selection, the main objective of studies is to estimate the ability or accuracy of genomic prediction. The accuracy of genomic prediction can be influenced by the following factors: heritability of the trait, effective population size, marker density, genome size, number of loci affecting the trait and distribution of their effects, linkage disequilibrium and number and type of phenotype information used in the statistical analyses. However, many of these factors described above may be directly or indirectly influenced by the quality control of the markers and, as a consequence, the accuracy of genomic prediction may be affected. For genome-wide association studies, the low quality data can lead to false-negative or positive results in the discovery phase of the regions associated with economic traits. On the other hand, the use of a quality control may remove markers that capture a significant proportion of the genetic variance. Therefore, the objective of this study was to evaluate the effect of quality control, marker density and allele frequency of markers on the genomic selection and the effect of quality control of markers on genome-wide association studies in Nellore breed. Phenotype, genotype and pedigree information, belonging to two breeding programs – DeltaGen and CRV Paint, were used. Age at first calving, weaning weight and yearling weight traits were studied. The animals were genotyped using the Illumina Bovine HD panel which contains 777 thousand SNP markers, approximately. The phenotype was corrected by the fixed effects using a single trait animal model. For the genomic selection were created: 1) three scenarios with different exclusion threshold of SNP markers according to quality control criteria; 2) seven sub panels with different marker density ranging from 10,000 to 734,272 SNP markers; and 3) five bins of markers based on the minor allele frequency. In the genomic selection “Genomic Best Linear Unbiased Predictor” (GBLUP) and BayesC π methods were used. The accuracy of genomic prediction was calculated as the Pearson correlation between observed and predicted values, divided by the square root of the heritability of the trait. The different scenarios and situations were compared through 5-folds cross validation method. In the genome-wide association studies, three scenarios were constructed according to the quality control criteria with different exclusion thresholds of SNP markers. To estimate the marker effect BayesC π method was used. The proportion of genetic variance within each 1 Mb window was used to compare the different scenarios. Independent of the method and trait used for genomic selection, there was no significant effect on the accuracies of genomic prediction among the scenarios. The sub panels with different marker densities

showed similar accuracies of genomic prediction for the traits studied. The sub groups of markers with the lower allele frequency showed low accuracy of genomic prediction when compared with others sub groups. In the genome-wide association studies, it was verified the effect of quality control on the proportion of genetic variance explained by each 1 Mb window. At the top ten windows which explained the highest proportion of genetic variance, it was verified modifications on the windows rank among the different scenarios. The lowest change occurred between the scenario with lower quality control and the scenario without quality control.

Key words: genomic analysis, accuracy of prediction, beef cattle, genomic predictions, molecular markers

CAPÍTULO 1 - CONSIDERAÇÕES GERAIS

1.1 INTRODUÇÃO

Na última década o uso de marcadores SNPs (*single nucleotide polymorphisms* ou polimorfismos de base única) para predição do mérito genético de animais e para identificação de regiões no genoma responsáveis pela expressão de características complexas, tem crescido rapidamente. Esse crescimento foi acompanhado pelo desenvolvimento de tecnologias de genotipagem em larga escala e dos chips contendo um número significativo de marcadores SNP. Por exemplo, para bovinos, foi lançado em 2009 o HDBovineSNP BeadChip (Illumina Infinium[®]), que possui aproximadamente 777 mil marcadores distribuídos ao longo do genoma e, potencialmente, é capaz de capturar grande parte da variação genética para determinadas características.

A genotipagem de um animal, utilizando painéis de alta densidade, produz três grupos de sinais de acordo com o genótipo do indivíduo (ZIEGLER; KONIG; THOMPSON, 2008). Porém, muitos fatores podem afetar a intensidade dos sinais, entre eles: problemas de análises laboratoriais, problemas nas amostras de chips (efeito de lote), material biológico usado para extração de DNA (sangue, carne, sêmen, etc.), quantidade e concentração de DNA. Assim, mesmo que as tecnologias de genotipagem em larga escala sejam precisas, ainda não são perfeitas, e erros podem ocorrer na genotipagem de alguns SNPs e amostras (SCHEET; STEPHENS, 2006). Para remover ou minimizar os erros de genotipagem, realiza-se o um controle de qualidade dos SNPs e das amostras antes das análises estatísticas.

Existem diferentes critérios de controle de qualidade que podem ser usados, envolvendo cada SNP individualmente e para as amostras. Para os SNPs, normalmente, são utilizados como critérios o escore de genotipagem (GCscore), a taxa de genotipagem (*Call rate*), a menor frequência alélica (MAF), o desvio do equilíbrio de Hardy-Weinberg (HWE) e a correlação (r^2) entre SNPs (ZIEGLER; KONIG; THOMPSON, 2008; MOSER et al., 2009; KIM et al., 2011; MUJIBI et al., 2011; UTSUNOMIYA et al., 2013). Para o controle de qualidade da amostra, os critérios mais comumente empregados são o *Call rate* e o nível de heterozigose (BOUWMAN et al., 2011; KIM et al., 2011; SAATCHI et al., 2011). No entanto, não há um consenso entre os pesquisadores quanto aos valores de limiar de exclusão atribuído para cada critério de controle de qualidade que, em geral, são determinados de acordo com os objetivos do trabalho, permitindo maior ou menor rigor.

A Illumina recomenda que seja utilizado um GCscore acima de 0,15 para painéis Infinium, como é o caso do HDBovineSNP BeadChip (ILLUMINA, 2009). Para os demais critérios de controle de qualidade dos marcadores, são atribuídos limiares de

exclusão variando: 0,80 a 0,98 para *Call rate*; 0,01 a 0,05 para MAF; e p-valores entre 0,05 a 0,00001 para HWE (ZIEGLER; KONIG; THOMPSON, 2008; MOSER et al., 2009; KIM et al., 2011; MUJIBI et al., 2011; UTSUNOMIYA et al., 2013). No controle de qualidade das amostras os limiares de exclusão para *Call rate* podem variar de 0,90 a 0,98 (BOUWMAN et al., 2011; KIM et al., 2011; SAATCHI et al., 2011). A heterozigosidade é um bom indicador da qualidade das amostras e, normalmente, são excluídas aquelas fora da média mais ou menos determinado número de desvios-padrão (ZIEGLER; KONIG; THOMPSON, 2008).

Para seleção genômica, em diferentes espécies, os estudos se concentram, principalmente, em estimar a habilidade e a acurácia da predição genômica (MEUWISSEN; HAYES; GODDARD, 2001; EDRISS et al., 2013; ABDOLLAHI-ARPANAHI et al., 2014). No entanto, a habilidade e a acurácia de predição dependem da herdabilidade da característica, do tamanho efetivo da população, da densidade dos marcadores, do tamanho do genoma, do número de loci afetando a característica e distribuição dos seus efeitos, do desequilíbrio de ligação e do tipo e número de informações fenotípicas usadas nas análises estatísticas (HAYES; GODDARD, 2010; MAKOWSKY et al., 2011; de los CAMPOS et al., 2013b). Contudo, muitos desses fatores podem ser direta ou indiretamente afetados pela edição do conjunto de marcadores, podendo influenciar a acurácia de predição genômica.

Além da seleção genômica, os SNPs podem ainda ser utilizados para a detecção e localização de QTLs (*Quantitative Trait Loci*) que atuam na expressão de características quantitativas, via estudos de associação genômica ampla. Para esses estudos, também é feito um controle de qualidade dos marcadores e amostras, utilizando os mesmos critérios e limiares de exclusão descritos anteriormente. A falta de qualidade nos dados genômicos pode levar a resultados falso-negativos ou falso-positivos na fase de detecção das regiões associadas as características complexas. Por outro lado, o rigor atribuído durante o controle de qualidade pode ter efeito sobre a proporção da variância genética capturada pelos marcadores, diminuindo as chances de encontrar regiões associadas às características de interesse. Esses resultados são propagados em estudos posteriores, como por exemplo na seleção genômica, impactando de forma negativa em termos de tempo e capital investidos (TURNER et al., 2011). Portanto, a qualidade dos dados genômicos é muito importante para estudos de seleção e associação genômica ampla e o controle de qualidade deve ser feito de forma a não interferir nos resultados das análises estatísticas.

1.2 REVISÃO DE LITERATURA

1.2.1 Critérios de Controle de Qualidade

Para o controle de qualidade (CQ), existem diferentes critérios que podem ser usados envolvendo os SNPs individualmente como, por exemplo, GCscore, *Call rate*, MAF, HWE e r^2 , que elimina das análises SNPs redundantes. Para as amostras, o *Call rate* e o nível de heterozigose são os critérios de CQ que normalmente são utilizados. O GCscore é usado como uma medida de qualidade, indicando a confiabilidade da leitura de cada genótipo, podendo variar de 0 a 1 (OLIPHANT et al., 2002). O GCscore é calculado utilizando informações do agrupamento das amostras, sendo que cada SNP é avaliado com base nas seguintes características: ângulo, dispersão, sobreposição e intensidade. Não há interpretação global do GCscore, pois a pontuação depende do agrupamento das amostras para cada SNP, o que é afetado por muitas variáveis, incluindo a qualidade das amostras (ILLUMINA, 2009). Em geral, quando o GCscore é baixo, os genótipos estão localizados mais distantes do centro de um agrupamento e, portanto, tem uma confiabilidade mais baixa. A Illumina recomenda que seja utilizado um GCscore acima de 0,15 para painéis Infinium, como é o caso do HDBovineSNP BeadChip (ILLUMINA, 2009).

O *Call rate* dos SNPs e das amostras é um bom indicador da qualidade de leitura dos genótipos. SNPs que não são “lidos” em um grande número de amostras indicam que o processo de genotipagem foi falho. Assim, o *Call rate* determina o percentual de SNP que foram lidos eficientemente. Pelo critério do *Call rate* são excluídos aqueles SNPs que não foram genotipados em uma dada proporção das amostras (ex.: 97%), podendo assumir maiores ou menores valores, dependendo das condições de genotipagem (ZIEGLER; KONIG; THOMPSON, 2008). Para melhor exemplificar essa diferença de valores, foram encontrados na literatura trabalhos que consideram nas análises, apenas SNP com *Call rate* variando de 0,80 a 0,98 (ZIEGLER; KONIG; THOMPSON, 2008; MOSER et al., 2009; KIM et al., 2011; MUJIBI et al., 2011; UTSUNOMIYA et al., 2013).

Segundo Edriss et al. (2013), é difícil estimar de forma acurada os efeitos de SNPs raros na população e, portanto, os mesmos devem ser excluídos das análises estatísticas. Nestes casos, normalmente, a MAF é usada como filtro no controle de qualidade dos marcadores (ZIEGLER; KONIG; THOMPSON, 2008). Na literatura, o critério do filtro para a MAF varia de 0,01 a 0,05, ou seja, são mantidos nas análises, apenas SNPs com MAF acima desses valores descritos (ZIEGLER; KONIG; THOMPSON, 2008; MOSER et al., 2009; KIM et al., 2011; MUJIBI et al., 2011; UTSUNOMIYA et al., 2013).

A relação entre as frequências dos alelos e as genotípicas também pode ser utilizada como um passo final no CQ (TURNER et al., 2011). Assim, o teste de HWE é usado para buscar evidências de problemas com os dados genômicos no CQ. O princípio geral do HWE é que as frequências alélicas e genotípicas em uma população grande e com acasalamento aleatório, permanecem estáveis entre gerações, existindo relacionamento fixo entre as frequências alélicas e genotípicas (ZIEGLER; KONIG; THOMPSON, 2008). Desvios observados no HWE podem evidenciar erros no processo de genotipagem e, portanto, podem-se excluir aqueles marcadores que desviam do HWE (ZIEGLER; KONIG; THOMPSON, 2008; WIGGANS et al., 2009). Para cada marcador, o desvio a partir do HWE é calculado utilizando o teste de qui-quadrado e a exclusão dos SNPs é feita com base nos p-valores que podem variar de 0,05 a 0,00001 (ZIEGLER; KONIG; THOMPSON, 2008; MOSER et al., 2009; KIM et al., 2011; MUJIBI et al., 2011; UTSUNOMIYA et al., 2013).

Para as amostras, o fato de uma grande proporção de SNPs não serem “lidos” é um indicativo de que a amostra de DNA é de baixa qualidade ou concentração (TURNER et al., 2011). Nesse caso, quando um indivíduo não foi genotipado de forma eficiente (Call rate baixa) ele é excluído das análises estatísticas. São descritos na literatura valores para a Call rate das amostras de 0,90 a 0,98 (BOUWMAN et al., 2011; KIM et al., 2011; SAATCHI et al., 2011). Outro indicador da qualidade das amostras é a heterozigosidade, onde elevada heterozigosidade pode indicar contaminação do DNA, exceto em populações cruzadas. Normalmente, estima-se a média e o desvio-padrão da heterozigosidade em toda a população genotipada para excluir aqueles fora da média mais ou menos determinado número de desvios-padrão. Excluir “outliers” com base no histograma de heterozigosidade também é um procedimento comumente utilizado (ZIEGLER; KONIG; THOMPSON, 2008).

1.2.2 Seleção Genômica Ampla

Por muitos anos, o mérito genético dos animais candidatos à seleção foi predito utilizando apenas informações fenotípicas e de pedigree via BLUP (HENDERSON, 1973). Na última década, com o advento das tecnologias de marcadores de DNA, especialmente do tipo SNP, a seleção de animais pode ser realizada incluindo, além do fenótipo e do pedigree, informações de um conjunto denso de marcadores. Os primeiros a propor o uso de um grande conjunto de marcadores distribuídos por todo o genoma para prever o mérito genético de animais candidatos à seleção foram Meuwissen, Hayes e Goddard (2001), ao que os autores denominaram de seleção genômica (SG). Esta, segundo os autores, é um tipo de seleção assistida por marcadores (SAM) empregando um grande número de marcadores.

Na SG, são utilizadas informações de um conjunto denso de marcadores distribuídos por todo o genoma, explicando grande parte da variação genética para determinada característica, assumindo que os marcadores estão em desequilíbrio de ligação (DL) com os QTLs (MEUWISSEN; HAYES; GODDARD, 2001). De acordo com Dekkers (2007) e Meuwissen (2007), a SG comparada à tradicional (BLUP), fornece predições dos valores genéticos com maior acurácia, principalmente, para as características que se expressam em um único sexo e/ou que sejam de baixa herdabilidade. Além disso, com a SG é possível antecipar o processo de seleção, no caso de características mensuradas tardiamente na vida do animal ou de mensuração difícil e de alto custo.

Um dos principais objetivos nos estudos de SG tem sido estimar a acurácia das predições genômicas (MEUWISSEN; HAYES; GODDARD, 2001; EDRISS et al., 2013; ABDOLLAHI-ARPANAHI et al., 2014) que, no entanto, pode ser afetada por uma série de fatores. A acurácia de predição genômica depende da herdabilidade da característica (GODDARD, 2009; HAYES et al., 2009; PSZCZOLA et al., 2012), tamanho efetivo da população (GODDARD, 2009), desequilíbrio de ligação (MEUWISSEN; HAYES; GODDARD, 2001), distribuição das frequências e efeitos dos alelos (LETTRE, 2011; PARK et al., 2011), densidade de marcadores (MOSER et al., 2010; WEIGEL et al., 2009), pseudo-fenótipo utilizado nas análises estatísticas (DAETWYLER et al., 2013) e relação de parentesco entre população de treinamento e validação (HAYES et al., 2009; HABIER et al., 2010).

A acurácia de predição genômica ainda pode ser influenciada pela qualidade dos dados genômicos utilizados nas análises estatísticas. Mesmo as tecnologias de genotipagem sendo precisas, as mesmas não são perfeitas e, portanto, erros podem ocorrer durante o processo de genotipagem (SCHEET; STEPHENS, 2006). São muitos os fatores que podem contribuir para esses erros, como: problemas de análises laboratoriais, problemas nas amostras de chips (efeito de lote), material biológico usado para extração de DNA (sangue, carne, sêmen, etc.), quantidade e concentração de DNA. Esses erros de genotipagem podem ser minimizados ou removidos através do controle de qualidade dos SNPs e amostras antes das análises estatísticas. Por outro lado, a edição do conjunto de marcadores pode ter efeito direto ou indireto sobre os fatores descritos anteriormente, podendo influenciar à acurácia de predição genômica.

1.2.3 Associação Genômica Ampla

Os primeiros trabalhos de associação entre marcadores SNPs e características de importância econômica, se baseavam no sequenciamento de regiões de genes candidatos, identificação dos SNPs, genotipagem dos animais para validação dos

SNPs e condução de testes de associação. Essa abordagem durante muito tempo propiciou a detecção e localização de um grande número de associações entre SNPs e características complexas nas diferentes espécies animais (SCHENKEL et al., 2005; ALLAN et al., 2007; KHATIB et al., 2007; GILL et al., 2009). No entanto, essa abordagem é limitada, pois assume-se que poucos genes são responsáveis pela expressão da característica (de los CAMPOS; GIANOLA; ALLISON, 2010). Além disso, os autores comentam que para os testes de associação, é utilizada a regressão de um único marcador o que leva a altas taxas de associações falso-positivas.

Com o desenvolvimento de chips contendo um número significativo de marcadores, distribuídos ao longo do genoma, estudos de associação genômica ampla (GWAS) são conduzidos. Os estudos GWAS consideram todos os marcadores disponíveis de uma só vez nas análises estatísticas. Com os estudos de GWAS é possível conhecer a arquitetura genética das características de importância econômica em uma determinada população (CASAS et al., 2013). Além disso, pode-se detectar QTLs, construir bibliotecas de genes que, posteriormente, podem ser utilizadas na seleção genômica (PETERS et al., 2012). Os estudos de GWAS exploram o DL existente entre os marcadores e as variantes causais (BOLORMAA et al., 2013), sendo que, quanto maior a densidade dos painéis de marcadores, maiores tendem a ser os níveis de DL (RINCON et al., 2011).

Apesar de utilizar metodologias capazes de estimar o efeito de um grande número de marcadores ao mesmo tempo nos estudos de GWAS, a preocupação com o poder dos testes estatísticos continua. Assim o sucesso dos estudos de GWAS depende do poder do teste estatístico, que por sua vez pode ser afetado por uma série de fatores como: tamanho da amostra, tamanho efetivo da população, frequência do alelo causal, desequilíbrio de ligação, estrutura populacional e número de animais com informações utilizados nas análises estatísticas (GODDARD, 2009; LETTRE, 2011; STRANGER; STAHL; RAJ, 2011; HAYES, 2013). Além desses fatores, a qualidade dos dados genômicos utilizados nos estudos de GWAS pode levar a resultados falsos-negativos ou falsos-positivos, comprometendo as análises atuais e as futuras.

1.2.4 Modelos Estatísticos

Atualmente, são muitas as metodologias que podem ser utilizadas na estimação dos efeitos dos marcadores, sendo que, basicamente, a diferença entre esses métodos está nas suposições assumidas para os efeitos dos marcadores. Independente das pressuposições, todos os métodos baseiam-se na regressão dos fenótipos sobre os marcadores disponíveis, usando um modelo linear (de los CAMPOS et al., 2013a). Porém, para que um modelo estatístico seja utilizado na seleção ou associ-

ação genômica ampla é preciso que o mesmo permita estimar um número suficientemente grande de efeitos a partir de um limitado número de observações ($n < p$). Segundo Resende et al. (2012), estão disponíveis metodologias que propiciam os três atributos desejáveis de um método ideal, permitindo a acomodação da arquitetura genética do caráter, regularização da estimação e a seleção de covariáveis

Genomic Best Linear Unbiased Predictor

Para os animais que possuem genótipo e fenótipo uma metodologia utilizada para estimar os efeitos genéticos capturados pelos marcadores, segue um modelo tradicional de análises (BLUP) com a substituição da matriz de parentesco A (pedigree) pela matriz de parentesco genômica G (marcadores). Essa metodologia foi proposta por VanRaden (2008), chamada de "Genomic Best Linear Unbiased Predictor"(GBLUP).

Nas avaliações tradicionais por meio do BLUP (HENDERSON, 1973) o seguinte modelo animal pode ser utilizado para estimar os efeitos genéticos aditivos em uma população: $y = X\beta + Za + \epsilon$, em que y é o vetor de observações fenotípicas; β é o vetor de efeitos fixos; a é o vetor dos efeitos genéticos aditivos aleatórios; ϵ é o vetor de resíduos aleatórios; e X e Z são as matrizes de incidência para β e a , respectivamente. Utilizando a metodologia GBLUP (VANRADEN, 2008) o modelo é equivalente ao descrito anteriormente: $y = X\beta + Ma + \epsilon$ em que a é o vetor dos efeitos aleatórios de marcadores, M é a matriz de incidência para a . A matriz de incidência M contém os valores 0, 1 e 2 para o número de alelos do marcador em um indivíduo diploide. Assim, segundo VanRaden (2008) as equações dos modelos mistos para predição de a pode ser descrita como:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + G^{-1} \frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}. \quad (1.1)$$

Para o GBLUP assume-se que $a | \sigma_a^2 \sim N(0, \sigma_a^2 G)$, onde a variância σ_a^2 é a mesma para todos os marcadores e $e \sim N(0, I \sigma_e^2)$, onde I é uma matriz identidade e σ_e^2 é a variância residual (VANRADEN, 2008). Segundo os autores uma forma de se construir a matrix G pode ser a seguinte:

$$G = \frac{(M - P)(M - P)'}{2 \sum_{i=1}^m p_i(1 - p_i)} \quad (1.2)$$

em que M é uma matriz de marcadores ($m \times n$) em que m é o número total de marcadores e n é o total de indivíduos genotipados, e P é uma matriz contendo a frequência

do segundo alelo (p_i), expressa como $2(p_i - 0,5)$. Com o GBLUP é possível obter as estimativas dos valores genéticos genômicos $\hat{g}$ para cada animal. No entanto, pode-se obter os efeitos de cada marcador de forma indireta como segue: $\hat{a} = M\hat{a} \Rightarrow M'\hat{a} = M'M\hat{a} \Rightarrow \hat{m} = (M'M)^{-1} M'\hat{a}$.

BayesC π

Meuwissen, Hayes e Goddard (2001), propuseram alguns métodos bayesianos (BayesA e BayesB) que foram importantes para predição do mérito genético em animais, sendo que cada uma dessas abordagens, diferem na suposição sobre os efeitos dos marcadores. No entanto, Gianola et al. (2009) fizeram uma análise crítica dos métodos associados aos modelos hierárquicos bayesianos (BayesA e B), especificamente em relação às suas formulações em termos dos hiperparâmetros que propiciam variâncias específicas para cada marcador. Segundo os autores, nenhum dos métodos permite o aprendizado Bayesiano sobre essas variâncias específicas para cada marcador.

Para contornar os problemas levantados por Gianola et al. (2009), com relação aos hiperparâmetros que propiciam variâncias específicas para cada marcador, Habier et al. (2011) propuseram a metodologia BayesC π . Segundo os autores, essa metodologia permite o aprendizado Bayesiano e apresenta a vantagem de propiciar informação sobre a arquitetura genética do caráter quantitativo. Sendo assim, o seguinte modelo geral pode ser descrito:

$$y = \mu + \sum_{i=1}^m M_i a_i + e_i \quad (i = 1, \dots, m) \quad (1.3)$$

onde: y é o vetor de pseudo-fenótipo; μ é a média geral; m é o número de SNPs, M_i é o vetor ($N \times 1$) do genótipo do SNP i , a_i é o efeito aditivo do SNP i , e e_i é um vetor de efeitos residuais. Os genótipos podem ser codificados como 0, 1 e 2, para o homozigoto, heterozigoto e o segundo homozigoto, respectivamente.

Na metodologia BayesC π assume-se, *a priori*, que todos os SNPs possuem variância comum, a qual segue uma qui-quadrada escalonada invertida com v graus de liberdade (4,2) e um parâmetro de escala S_a^2 (HABIER et al., 2011). Além disso, *a priori* para a_i depende da variância de a_i ($\sigma_{a_i}^2$) e da probabilidade *a priori* para o parâmetro π do marcador i , ou seja, $a_i | \pi, \sigma_{a_i}^2 \sim \pi N(0, \sigma_{a_i}^2) + (1 - \pi) N(0, \sigma_{a_i}^2)$. Diferente dos métodos propostos por (MEUWISSEN; HAYES; GODDARD, 2001), no método BayesC π o parâmetro π é tratado como desconhecido e com distribuição uniforme (0, 1) *a priori*.

Como resultado, o efeito de um SNP, ajustado com probabilidade $(1 - \pi)$, vem a partir de uma mistura de distribuições t de Student multivariada $t(0, v_a, IS_a^2)$. O S_a^2 é derivado a partir do valor esperado de uma variável aleatória com distribuição qui-quadrada escalonada invertida como segue:

$$E(\sigma_{a_i}^2) = \frac{v_a S_a^2}{v_a - 2} = \tilde{\sigma}_a^2; \quad (1.4)$$

onde $S_a^2 = \frac{\sigma_a^2(\tilde{v}_a - 2)}{\tilde{v}_a}$, $v_a = 4,2$ como descrito em Meuwissen, Hayes e Goddard (2001) e $\tilde{\sigma}_a^2$ é a variância do efeito aditivo de um locus amostrado aleatoriamente, que pode ser considerado como a variância genética aditiva explicada pelos SNPs. Para os efeitos residuais assume-se *a priori* que $e_i \sim N(0, I \sigma_e^2)$, onde σ_e^2 segue uma qui-quadrada escalonada invertida com os parâmetros v e S , em que, arbitrariamente, v é menor que 4,2 (HABIER et al., 2011).

1.3 OBJETIVOS

Objetivo Geral

Estudar o efeito do controle de qualidade, da densidade de marcadores e das frequências alélicas dos marcadores na acurácia de predição genômica e o efeito do controle de qualidade dos genótipos nos estudos de associação genômica ampla para bovinos da raça Nelore, utilizando um painel de alta densidade, buscando definir critérios de edição adequados.

Objetivos Específicos

- Verificar o efeito do controle de qualidade, utilizando diferentes limiares para os critérios de exclusão dos marcadores, na acurácia de predição genômica;
- Utilizar sub painéis com diferentes densidades e verificar o impacto dos mesmos sobre a acurácia de predição genômica;
- Definir sub grupos de marcadores de acordo com a menor frequência alélica para observar o efeito da frequência alélica sobre a acurácia de predição genômica;
- Verificar o efeito do controle de qualidade, utilizando diferentes limiares para os critérios de exclusão dos marcadores, sobre a proporção da variância genética explicada por cada janela de 1 Mb nos estudos de associação genômica ampla.

CAPÍTULO 2 - EFEITO DO CONTROLE DE QUALIDADE, DA DENSIDADE E DA FREQUÊNCIA ALÉLICA DOS MARCADORES NA ACURÁCIA DE PREDIÇÃO GENÔMICA DE CARACTERÍSTICAS COMPLEXAS NA RAÇA NELORE

RESUMO - Na literatura são encontrados trabalhos de seleção genômica em diferentes espécies, dos quais muitos se concentram em estimar a acurácia de predição genômica em características complexas. Considerando a importância de ter estimativas acuradas, o presente trabalho foi conduzido com o objetivo de verificar o efeito do controle de qualidade, da densidade e frequência alélica dos marcadores sobre a acurácia de predição genômica utilizando três características com diferentes herdabilidades, além de duas metodologias, em uma população da raça Nelore, genotipada com o painel Illumina Bovine HD assay. As características estudadas foram: idade ao primeiro parto, peso à desmama e peso ao sobreano com 1.789, 3.228 e 2.803 informações respectivamente. Foram criados três cenários para os critérios de controle de qualidade, com diferentes níveis de restrição. Além disto, foram utilizados seis sub painéis com diferentes densidades de marcadores, obtidos a partir do painel de alta densidade. Por fim, o total de marcadores disponíveis foi subdividido em seis grupos de acordo com a menor frequência alélica. As metodologias Genomic Best Linear Unbiased Predictor (GBLUP) e BayesC π foram utilizadas para estimar os efeitos dos marcadores. Não foi verificado efeito dos critérios de controle de qualidade dos genótipos sobre a acurácias de predição genômicas. A maior densidade dos painéis não resultou em maiores acurácias de predição. Os sub grupos de SNPs com as menores frequências alélicas tiveram menores acurácias de predição do que sub grupos com maiores frequências alélicas.

Palavras-chave: bovinos de corte, edição dos marcadores, predições genômicas

ABSTRACT - The main objective of the genomic selection studies, in different species, is to estimate the accuracy of genomic prediction for complexes traits. Considering the importance of having accurate estimates, the objective of this study was to verify the effects of markers quality control, density and allele frequency of markers on accuracy of genomic prediction, using three traits with different heritability estimates and two methods of analyses in Nelore breed animals, genotyped with Illumina Bovine HD assay panel. The following traits were used: age at first calving, weaning weight and yearling weight with 1,789, 3,228 and 2,803 animals with genotypes and phenotypes, respectively. Three scenarios for the quality control criteria, with different restriction levels, were used. Moreover, six sub panels with different markers densities, derived from high density panel were used. The total number of markers available was divided in six groups according to the minor allele frequency. Genomic Best Linear Unbiased Predictor (GBLUP) and BayesC π were used to estimate the marker effects.

There was no effect of quality control criteria of genotypes on accuracy of genomic predictions. The high density panel does not result in higher genomic predictions compared to panels with 10,000 to 100,000 SNP markers. The sub groups of markers with low allele frequency, showed lower genomic prediction than groups with higher allele frequency.

Key words: beef cattle, markers edition, genomic predictions

2.1 INTRODUÇÃO

A seleção genômica (SG) proposta por Meuwissen, Hayes e Goddard (2001) é utilizada para prever o valor genético aditivo, utilizando milhares de polimorfismos de base única (SNP) distribuídos ao longo do genoma. Além do seu uso em programas de melhoramento animal, a SG é ainda empregada no melhoramento de plantas, na medicina humana e na biologia (ABDOLLAHI-ARPANAHI et al., 2014). Para a aplicação da SG, primeiramente, o efeito de cada SNP é estimado para um fenótipo de interesse, usando uma população de referência. Essas estimativas são usadas, posteriormente, para prever o mérito genético de indivíduos candidatos à seleção, que ainda não apresentam o fenótipo de interesse, mas foram genotipados. Para o sucesso da SG, independente da espécie considerada, é importante que as predições feitas na população de referência sejam acuradas.

Um dos principais objetivos nos estudos de SG tem sido estimar a acurácia de predição genômica (MEUWISSEN; HAYES; GODDARD, 2001; EDRISS et al., 2013; ABDOLLAHI-ARPANAHI et al., 2014), que no entanto, pode ser afetada por uma série de fatores. A acurácia de predição genômica depende da herdabilidade da característica (GODDARD, 2009; HAYES et al., 2009; PSZCZOLA et al., 2012), tamanho efetivo da população (GODDARD, 2009), desequilíbrio de ligação (MEUWISSEN; HAYES; GODDARD, 2001), distribuição das frequências e efeitos dos alelos (LETTRE, 2011; PARK et al., 2011), densidade de marcadores (MOSER et al., 2010; WEIGEL et al., 2009), pseudo-fenótipo utilizado nas análises estatísticas (DAETWYLER et al., 2013) e relação entre população de treinamento e validação (HAYES et al., 2009; HABIER et al., 2010).

A qualidade das informações genômicas utilizadas nas análises estatísticas, também pode ter efeito sobre a acurácia de predição genômica. Para evitar este problema, normalmente, se faz um controle de qualidade do conjunto de marcadores utilizando diferentes critérios. O escore de genotipagem (GCscore), a taxa de genotipagem (*Call rate*), a menor frequência alélica (MAF), o desvio do equilíbrio de Hardy-Weinberg (HWE), e a correlação entre SNPs (r^2) são os principais critérios utilizados

no controle de qualidade dos marcadores. Por outro lado, os critérios de edição podem ter efeito direto ou indireto sobre os fatores descritos anteriormente o que pode ter efeito sobre as predições genômicas. Não existe um consenso na literatura sobre os limites mínimos ou máximos a serem utilizados para cada critério de edição. Em bovinos de leite, Edriss et al. (2013) conduziram um estudo testando diferentes limites de exclusão dos marcadores, de acordo com alguns critérios de controle de qualidade (GCscore, MAF e HWE), e verificaram pequenas variações nas acurácias das predições genômicas.

Até o momento, são limitados os trabalhos que abordam sistematicamente os fatores utilizados para a consistência dos dados genômicos, e sua influência na acurácia de predição genômica. O presente trabalho foi conduzido com o objetivo de verificar o efeito do controle de qualidade, da densidade e da frequência alélica dos marcadores sobre a acurácia de predição genômica em uma população de animais da raça Nelore, buscando definir critérios de edição adequados.

2.2 MATERIAL E MÉTODOS

Dados

Foram utilizadas informações fenotípicas e genotípicas de uma população de animais pertencentes aos programas de melhoramento genético DeltaGen e CRV Paint. As características estudadas foram: idade ao primeiro parto (IPP), peso à desmama (PD) e peso ao sobreano (PS). Os fenótipos foram pré-corrigidos para os efeitos fixos de grupo de contemporâneos (IPP, PD e PS) e as covariáveis idade da mãe ao parto (PD) e idade do animal a pesagem (PD e PS), efeitos linear e quadrático. Para a característica PD foi considerado ainda o efeito materno como aleatório no modelo. A definição de grupo de contemporâneo para IPP foi: safra, fazenda e grupo de manejo ao nascimento, grupo de manejo à desmama e ao sobreano. Para a característica PD, o grupo de contemporâneo foi definido a partir da combinação de safra, fazenda, e grupo de manejo ao nascimento, fazenda e grupo de manejo à desmama e sexo. Na característica PS, além da definição para PD ainda foi incluído a fazenda e grupo de manejo ao sobreano. As soluções dos efeitos fixos foram obtidas utilizando um modelo animal uni-característica, empregando o software AIREMLF90 (MISZTAL et al., 2014). Todos os indivíduos considerados nas análises possuíam informações fenotípicas e genotípicas para cada característica. Os animais foram genotipados utilizando o painel de alta densidade Illumina Bovine HD assay (Illumina, San Diego, CA, USA) que possui 777.962 marcadores SNPs espalhados pelo genoma bovino. Foram considerados no presente estudo apenas marcadores dos cromossomos 1 ao 29 e amostras

com taxa de chamada (Call rate) maior que 0,90. O número de animais para cada característica está apresentado na Tabela 2.1.

Cenários

Para investigar o efeito do controle de qualidade (CQ) dos SNPs sobre a acurácia de predição genômica foram criados três cenários. Em cada cenário do CQ dos SNPs foram definidos valores e/ou limiares de exclusão diferentes para: taxa de genotipagem (*Call rate*), menor frequência alélica (MAF), equilíbrio de Hard-Weinberg (HWE) e correlação entre pares de SNPs (r^2). Dessa forma, os cenários foram definidos como:

a) Maior rigor (C1): *Call rate* ($< 0,98$), MAF ($< 0,05$), HWE ($p < 10^{-5}$) e r^2 ($> 0,999$). O r^2 foi avaliado dentro de uma janela de 100 marcadores, sendo excluído o SNP de menor MAF de cada par altamente correlacionado.

b) Menor rigor (C2): *Call rate* ($< 0,85$) e MAF ($< 0,01$);

c) Foram excluídos apenas marcadores duplicados (C3).

O controle de qualidade dos marcadores para os cenários C1 e C2 foi efetuado de forma iterativa até que nenhum SNP fosse descartado. Os diferentes cenários foram criados para cada característica em estudo (IPP, PD e PS). Todos os critérios descritos foram calculados utilizando o pacote snpStats do R (R Core Team, 2013). Um resumo do número de SNP utilizado em cada cenário está descrito na Tabela 2.1.

Tabela 2.1 – Descrição das amostras e cenários utilizados.

Números	IPP	PD	PS
Total de indivíduos	1.817	3.228	2.803
Indivíduos nas análises	1.756	3.165	2.780
Herdabilidades	0,13	0,20	0,35
Total de SNPs	777.962	777.962	777.962
SNPs no C1	316.252	321.988	321.992
SNPs no C2	569.720	576.332	576.487
SNPs no C3	734.272	734.272	734.272

IPP = idade ao primeiro parto; PD = peso à desmama; PS = peso ao sobreano; C1 = cenário de maior rigor; C2 = cenário de menor rigor; C3 = cenário sem controle de qualidade.

Para verificar o efeito da densidade dos marcadores na acurácia de predição genômica, seis sub painéis com diferentes números de SNPs foram utilizados: 10k, 50k, 100k, 300k, 500k, além do painel com todos os marcadores disponíveis (734.272 SNPs). Os sub painéis foram criados a partir do total de SNPs disponíveis (734.272) com genótipos faltantes e genótipos faltantes imputados. Os marcadores foram alo-

cados em cada painel de forma aleatória, incluindo os cromossomos de 1 a 29. A imputação dos genótipos faltantes foi realizada com o programa FImpute (SARGOLZAEI; CHESNAIS; SCHENKEL, 2014).

O efeito da frequência alélica sobre a acurácia de predição genômica foi verificada utilizando cinco sub grupos formados a partir da MAF de cada SNP (Tabela 2.2). Nos sub grupos os marcadores foram agrupados em intervalos de acordo com sua MAF. A definição dos intervalos foi realizada de forma que todos os sub grupos possuísem quantidade, aproximadamente, igual de marcadores. As análises estatísticas foram conduzidas para cada um dos sub grupos. Na definição dos sub grupos foram utilizados todos os marcadores disponíveis (734.272 SNPs) que tiveram os genótipos faltantes imputados. Os sub painéis e os sub grupos de MAF foram construídos para as três características (IPP, PD e PS) e analisados pelas duas metodologias descritas na seção Modelos.

Tabela 2.2 – Descrição dos sub grupos de marcadores com base na menor frequência alélica (MAF)

Grupos	IPP		PD		PS	
	MAF (N)	Média	MAF (N)	Média	MAF (N)	Média
G1	0,000-0,007 (147.297)	0,00199	0,000-0,009 (146.891)	0,00259	0,000-0,008 (146.823)	0,00244
G2	0,007-0,062 (146.900)	0,0297	0,009-0,064 (146.815)	0,0313	0,008-0,064 (146.823)	0,0308
G3	0,062-0,173 (146.726)	0,114	0,064-0,174 (146.832)	0,116	0,064-0,174 (146.855)	0,115
G4	0,173-0,324 (146.727)	0,247	0,174-0,325 (146.837)	0,247	0,174-0,325 (146.920)	0,247
G5	0,324-0,500 (146.621)	0,411	0,325-0,500 (146.896)	0,412	0,325-0,500 (146.850)	0,412
G6	0,000-0,500 (734.271)	0,161	0,000-0,500 (734.271)	0,162	0,000-0,500 (734.271)	0,161

IPP = idade ao primeiro parto; PD = peso à desmama; PS = peso ao sobreano; N = número de SNPs.

Modelos

As metodologias Genomic Best Linear Unbiased Predictor (GBLUP), proposta por VanRaden (2008) e BayesC π (HABIER et al., 2011) foram utilizadas para obter as soluções dos efeitos dos marcadores. O modelo geral pode ser descrito como:

$$y = \mu + \sum_{i=1}^m M_i a_i + e_i \quad (i = 1, \dots, m), \quad (2.1)$$

onde: y é o vetor de pseudo-fenótipo; μ é a média geral; m é o número de SNPs, M_i é o vetor ($N \times 1$) do genótipo do SNP i , a_i é o efeito aditivo do SNP i , e e_i é um vetor

de efeitos residuais. Os genótipos dos SNPs foram codificados como 0, 1 e 2, para o homozigoto, heterozigoto e o segundo homozigoto, respectivamente.

A diferença entre as metodologias está nas pressuposições assumidas para os efeitos dos marcadores a . Para o GBLUP assume-se que $a | \sigma_a^2 \sim N(0, \sigma_a^2 G)$, onde a variância σ_a^2 é a mesma para todos os marcadores e $e \sim N(0, I \sigma_e^2)$, onde I é uma matriz identidade e σ_e^2 é a variância residual (VANRADEN, 2008). O mesmo autor propõe que a matriz G seja construída da seguinte forma:

$$G = \frac{(M - P)(M - P)'}{2 \sum_{i=1}^m p_i(1 - p_i)}, \quad (2.2)$$

em que M é uma matriz de marcadores ($m \times n$) em que m é o número total de marcadores e n é o total de indivíduos genotipados, e P é uma matriz contendo a frequência do segundo alelo (p_i), expressa como $2(p_i - 0,5)$.

Com o método BayesC π (HABIER et al., 2011), assume-se *a priori* que todos os SNPs possuem variância comum (σ_a^2), seguindo uma distribuição qui-quadrada escalonada invertida com v graus de liberdade (4,2) e um parâmetro de escala S . Para os efeitos dos marcadores (a_i) assume-se uma informação *a priori* para o parâmetro π do marcador i . Como resultado, o efeito de um SNP, ajustado com probabilidade $(1 - \pi)$, vem a partir de uma mistura de distribuições t de Student, ou seja, $a_i | \pi, \sigma_{a_i}^2 \sim \pi N(0, \sigma_{a_i}^2) + (1 - \pi) N(0, \sigma_{a_i}^2)$. Diferente dos métodos propostos por Meuwissen, Hayes e Goddard (2001), no método BayesC π o parâmetro π é tratado como desconhecido e com distribuição uniforme (0, 1) *a priori*. Para os efeitos residuais, assume-se *a priori* que: $e_i \sim N(0, I \sigma_e^2)$, onde σ_e^2 , onde σ_e^2 segue uma qui-quadrada escalonada invertida com os parâmetros v e S , em que, arbitrariamente, v é menor que 4,2 (HABIER et al., 2011). Essas duas metodologias (GBLUP e BayesC π) foram implementadas no *software* GS3 desenvolvido por Legarra, Ricard e Filangi (2014), e utilizado para estimar os efeitos dos marcadores no presente estudo.

Na metodologia BayesC π foram realizadas 500 mil iterações MCMC, com período de burn-in de 100.000 e os resultados salvos a cada 20 iterações para obter as soluções dos marcadores. Foram realizados testes de convergência para a variância genética aditiva, com base nas cadeias MCMC, com o pacote estatístico Bayesian Output Analysis Program (BOA) for MCMC (SMITH, 2008) do R (R Core Team, 2013). Além disso, foi realizada uma inspeção visual das cadeias para verificar a convergência das mesmas. O critério de convergência para a metodologia GBLUP foi baseado no logaritmo da função de verossimilhança (10^{-8}).

Acurácia de predição genômica

Para verificar o impacto do controle de qualidade, da densidade e da frequência alélica dos marcadores sobre a acurácia de predição genômica, foi utilizado o método de validação cruzada. Para isto, o total de animais foi dividido, de forma aleatória, em duas populações: uma de treinamento (animais com fenótipos e genótipos) e outra de validação (animais apenas com genótipo). Primeiramente, a população de treinamento foi usada para estimar os efeitos dos SNPs. Posteriormente, a população de validação teve seus fenótipos preditos com base nos efeitos dos SNPs estimados na população de treinamento. No presente trabalho, o total de animais foi dividido em 5 conjuntos (aleatoriamente). Destes, 4 foram utilizados para estimar os efeitos dos marcadores e um para prever os fenótipos. O processo foi repetido 5 vezes de forma que cada grupo de animais fosse utilizado uma vez como população de validação até que todos os animais tivessem seus fenótipos preditos. O número médio de animais utilizados na população de treinamento para as características IPP, PD e PS foram 1.431, 2.532 e 2.213, respectivamente. Na população de validação, o número médio de animais foi 358 (IPP), 633 (PD) e 553 (PS). Os mesmos animais foram utilizados nas análises dos cenários de CQ, densidades de painéis e sub grupos de marcadores. A acurácia de predição foi calculada pela correlação de Pearson entre os valores observados (pseudo-fenótipos) e os preditos (pseudo-fenótipos preditos) dividido pela raiz quadrada da herdabilidade da característica (PRYCE et al., 2012). Foi realizado um teste de comparação de médias para verificar as diferenças significativas entre as acurácias de predição genômica nos cenários do CQ e nos sub painéis de marcadores.

Desequilíbrio de ligação

O desequilíbrio de ligação (LD) entre dois SNPs foi avaliado usando o r^2 (HILL; ROBERTSON, 1968) que é calculado da seguinte forma:

$$r^2 = \frac{(freq.AB * freq.ab - freq.Ab * freq.aB)^2}{(freq.A * freq.a * freq.B * freq.b)}, \quad (2.3)$$

onde freq.A, freq.a, freq.B e freq.b são as frequências dos alelos A, a, B e b, respectivamente e freq.AB, freq.ab, freq.aB e freq.Ab são as frequências dos haplótipos AB, ab, aB e Ab na população, respectivamente. Se dois *loci* são independentes, a frequência esperada do haplótipo AB é calculada como o produto entre freq.A e a freq.B. Assim, se freq.AB for maior ou menor que o valor esperado, isso indica que esses dois loci tendem a segregar juntos e estão em LD. O LD (r^2) foi calculado para todos os pares de SNPs de cada cromossomo usando o pacote snpStats (CLAYTON, 2014) do R (R Core Team, 2013).

2.3 RESULTADOS E DISCUSSÃO

Independente do método de predição, a diminuição no total de marcadores disponível para as análises de seleção genômica (SG), devido ao CQ, não teve impacto significativo sobre a acurácia de predição genômica nas características estudadas (Tabela 2.3). Em bovinos de leite Edriss et al. (2013) utilizando diferentes limiares de exclusão para alguns critérios do CQ dos SNPs (MAF, HWE e GC score), encontraram pequenas diferenças nas acurácias de predição. No entanto, os autores comentam que, mesmo apresentando pequeno efeito sobre a acurácia de predição, o processo de edição dos SNPs é fácil de ser realizado sem que marcadores informativos sejam perdidos. Segundo Wiggans et al. (2009) se um SNP não tem seu efeito estimado de forma acurada e não contribui para as predições genômicas, o mesmo pode ser eliminado reduzindo assim a demanda computacional e melhorando a precisão das estimativas dos efeitos dos SNPs que permanecem nas análises.

Tabela 2.3 – Estimativas da acurácia de predição genômica e do desequilíbrio de ligação (LD) nos diferentes cenários do controle de qualidade.

	C1	C2	C3
<i>Idade ao Primeiro Parto</i>			
GBLUP	0,449 ^a ± 0,128 ¹	0,461 ^a ± 0,127	0,465 ^a ± 0,127
BAYES	0,433 ^a ± 0,105	0,453 ^a ± 0,107	0,459 ^a ± 0,110
LD	0,312 ± 0,202	0,393 ± 0,231	0,313 ± 0,123
<i>Peso a Desmama</i>			
GBLUP	0,500 ^a ± 0,118	0,496 ^a ± 0,117	0,498 ^a ± 0,119
BAYES	0,491 ^a ± 0,116	0,489 ^a ± 0,117	0,488 ^a ± 0,117
LD	0,310 ± 0,210	0,391 ± 0,228	0,312 ± 0,128
<i>Peso ao Sobreano</i>			
GBLUP	0,437 ^a ± 0,071	0,440 ^a ± 0,069	0,440 ^a ± 0,068
BAYES	0,438 ^a ± 0,071	0,442 ^a ± 0,069	0,442 ^a ± 0,071
LD	0,313 ± 0,207	0,390 ± 0,234	0,314 ± 0,126

GBLUP = acurácia de predição para a metodologia Genomic Best Linear Unbiased Prediction; BAYES = acurácia de predição para a metodologia BayesC π ; ¹ = desvio-padrão; C1 = maior rigor no controle de qualidade; C2 = menor rigor no controle de qualidade; C3 = sem controle de qualidade; letras sobescritas representam as diferenças significativas (p < 0,05).

As acurácias de predição genômica encontradas no presente trabalho diferem entre as características estudadas (Tabela 2.3). De forma geral o maior valor da acurácia de predição foi encontrado para a característica peso à desmama (PD), porém, semelhante entre as características idade ao primeiro parto (IPP) e peso ao sobreano (PS). Este resultado não seria esperado uma vez que as acurácias foram maiores para a característica de menor herdabilidade (IPP e PD). No entanto, resultados semelhantes aos encontrados no presente trabalho são reportados na literatura em bovinos de

corte (WEBER et al., 2012; ZHANG et al., 2014) e bovinos de leite (LI et al., 2014; ZHANG et al., 2014). As acurácias de predição no presente trabalho foram de baixas a moderadas. Maiores valores de acurácia de predição são reportados na literatura quando o valor genético ou valor genético desregredido foram utilizados como pseudo-fenótipos na população de treinamento (SAATCHI et al., 2011; BODDHIREDDY et al., 2014a; BODDHIREDDY et al., 2014b). No entanto, Daetwyler et al. (2013) comenta que diferenças nas acurácias de predição utilizando diferentes pseudo-fenótipos podem acontecer, pois os mesmos apresentam diferentes sinais de natureza genética. Além disso, os autores comentam que a acurácia de predição depende do número de QTLs que são responsáveis pela variação genética da característica.

As acurácias de predição genômicas considerando os sub painéis foram semelhantes, tanto para aqueles que tiveram os SNPs faltantes imputados como para os não imputados. Dessa forma, optou-se por apresentar apenas os resultados dos sub painéis que tiveram os SNPs faltantes imputados (Tabela 2.4). Os valores de acurácia de predição encontradas para os sub painéis foram próximos aos apresentados nos diferentes cenários do CQ. O maior valor de acurácia de predição para a característica IPP foi encontrada no sub painel 10 K. Para as características PD e PS os subs painéis 50 K e 300 k, respectivamente, apresentaram as maiores acurácias de predição. Entretanto, as pequenas diferenças nas acurácias de predição observadas entre os sub painéis não foram significativas ($p < 0,05$) para as três características estudadas e nas metodologias utilizadas.

Resultado semelhante a do presente trabalho é reportado por Abdollahi-Arpanahi et al. (2014) em aves. Os autores encontraram um pequeno aumento na habilidade de predição quando aumentaram a densidade dos marcadores de 5.000 para 10.000, comparado com todos os marcadores. Em humanos, Makowsky et al. (2011) observaram que, com o aumento da densidade de marcadores até 400 mil havia um aumento nas habilidades de predições genômicas. Zhang et al. (2011) comparando a acurácia de predição entre um painel de 10.000 SNPs e painéis de baixa densidade (50 a 5000 SNPs), simulados, observaram que 95 % da acurácia de predição do painel de 10.000 SNPs pode ser obtida com poucos marcadores. Em bovinos de leite, genotipados com 50 K, Weigel et al. (2009) e Vazquez et al. (2010) relatam que o uso de poucos marcadores (aproximadamente 1.000 SNPs) é suficiente para obter resultados significativos na acurácia de predição genômica.

Os resultados do presente trabalho mostram que o uso de painéis mais densos (> 100 K) pode não refletir em maiores ganhos na acurácia de predição genômica. Porém, os resultados encontrados para os sub painéis não podem ser replicados em outras raças, pois segundo Vazquez et al. (2010) os sub painéis são raça específicos.

Tabela 2.4 – Estimativas da acurácia de predição genômica e desequilíbrio de ligação (LD) e seus respectivos desvio-padrões, para as diferentes densidades de SNPs.

Paineis	GBLUP	BAYES	LD
<i>Idade ao Primeiro Parto</i>			
10k	0,497 ^a ± 0,107 ¹	0,495 ^a ± 0,080	0,072 ± 0,161
50k	0,447 ^a ± 0,123	0,436 ^a ± 0,105	0,154 ± 0,260
100k	0,449 ^a ± 0,132	0,435 ^a ± 0,117	0,197 ± 0,297
300k	0,464 ^a ± 0,130	0,453 ^a ± 0,109	0,262 ± 0,343
500k	0,466 ^a ± 0,125	0,462 ^a ± 0,105	0,294 ± 0,362
700k	0,465 ^a ± 0,127	0,459 ^a ± 0,110	0,315 ± 0,373
<i>Peso a Desmama</i>			
10k	0,447 ^a ± 0,159	0,453 ^a ± 0,163	0,067 ± 0,154
50k	0,507 ^a ± 0,124	0,501 ^a ± 0,122	0,146 ± 0,252
100k	0,494 ^a ± 0,114	0,485 ^a ± 0,112	0,188 ± 0,290
300k	0,496 ^a ± 0,115	0,488 ^a ± 0,114	0,252 ± 0,337
500k	0,496 ^a ± 0,119	0,485 ^a ± 0,116	0,283 ± 0,357
700k	0,498 ^a ± 0,119	0,488 ^a ± 0,117	0,303 ± 0,368
<i>Peso ao Sobreano</i>			
10k	0,393 ^a ± 0,056	0,395 ^a ± 0,057	0,068 ± 0,155
50k	0,435 ^a ± 0,080	0,434 ^a ± 0,082	0,147 ± 0,253
100k	0,434 ^a ± 0,072	0,436 ^a ± 0,073	0,190 ± 0,291
300k	0,440 ^a ± 0,068	0,441 ^a ± 0,070	0,254 ± 0,338
500k	0,442 ^a ± 0,069	0,444 ^a ± 0,071	0,285 ± 0,358
700k	0,440 ^a ± 0,068	0,442 ^a ± 0,071	0,305 ± 0,369

GBLUP = Genomic Best Linear Unbiased Prediction; BayesC π (BAYES);

¹ = desvio-padrão; letras sobescritas diferentes representam as diferenças significativas (p < 0,05).

Ainda, segundo os autores, para outras espécies o número de marcadores necessários pode mudar dependendo do LD. Como reportado na literatura (WEIGEL et al., 2009; VAZQUEZ et al., 2010; ZHANG et al., 2011) o uso de painéis de baixa densidade, capazes de capturar grande parte da variação genética, são economicamente viáveis para seleção de animais. Contudo, é importante que animais de maior contribuição genética (machos e fêmeas) sejam genotipados com painel de alta densidade (WEIGEL et al., 2009). Talvez, até o momento, um dos entraves da utilização da seleção genômica na maioria dos programas de melhoramento genético no Brasil sejam os altos custos da genotipagem dos animais. Assim, o menor custo dos chips de baixa densidade tornaria mais atrativo o uso de marcadores para seleção de animais.

As acurácias de predições genômicas para os sub grupos de marcadores formados a partir da menor frequência alélica (MAF) foram semelhantes entre as metodologias utilizadas. Assim, os resultados apresentados são do método BayesC π (Figura 2.1). Para as características IPP, PD e PS as menores acurácias de predição

foram encontradas para os sub grupos de menor MAF (G1). Porém, entre os grupos G2 a G5, não foram observadas grandes diferenças nas acurácias de predição para essas características. A característica PS apresentou a menor acurácia de predição no grupo G1, com aumento na acurácia nos demais grupos (G2 a G5). De forma geral, os grupos com maior MAF (G3, G4 e G5) foram os que apresentaram as maiores acurácias de predição, mas não maior que considerando todos os marcadores (734.272).

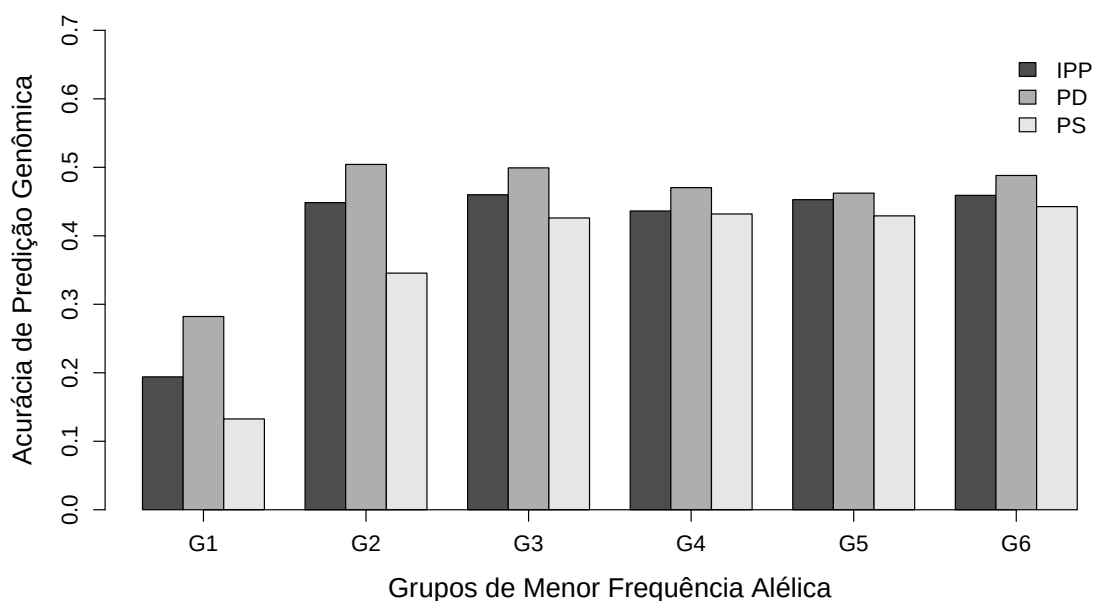


Figura 2.1 – Acurácia de predição genômica utilizando grupos de MAF para as características idade ao primeiro parto (IPP), peso à desmama (PD) e peso ao sobreano (PS).

Os resultados encontrados no presente estudo, para os sub grupos de MAF, diferem dos reportados por Abdollahi-Arpanahi et al. (2014) em aves. Os autores observaram que SNPs com menor MAF apresentaram melhor habilidade de predição, mas não foram superiores ao painel incluindo todos os marcadores. No entanto, a magnitude das habilidades de predição encontradas por esses autores foi maior que as do presente trabalho. Edriss et al. (2013) em bovinos de leite atribuindo diferente limites de exclusão para a MAF encontraram pequenas diferenças nas acurácias de predição entre as estratégias utilizadas. Os autores observaram que ao excluir SNPs com MAF acima de 0,10 houve uma redução na acurácia de predição. As menores acurácias encontradas no sub grupo G1 do presente trabalho podem estar relacionadas com a proporção de SNPs monomórficos ou raros que esses grupos apresentam. A proporção de marcadores monomórficos observados no sub grupo G1 para as características IPP, PD e PS foram 27 %, 14 % e 16 %, respectivamente. Segundo, Edriss et al. (2013), é difícil estimar o efeito quando um alelo é raro na população e possíveis

erros de estimação dos efeitos dos marcadores podem ocorrer. Por outro lado, quando todos os marcadores são considerados nas análises os alelos raros ou monomórficos são “mascarados” pelos demais marcadores.

O LD foi calculado para os cenários do CQ e sub painéis para as características estudadas (Tabela 2.3, Tabela 2.4). Os resultados encontrados para o LD nos diferentes cenários do CQ foram semelhantes entre si e concordam com os reportados por Espigolan et al. (2013) em animais da raça Nelore genotipados com o chip de 770 K. Por outro lado, menores valores de LD foram encontrados por Ogawa et al. (2014) na raça Japanese Black, genotipados com o chip de 50 K. Nos sub painéis, independente da característica, o LD aumentou com o aumento na densidade do painel. Esse comportamento também foi verificado por Ogawa et al. (2014), porém considerando uma menor quantidade de SNPs (100 a 38.000 SNPs). Como verificado por Espigolan et al. (2013) e Ogawa et al. (2014), com o aumento da distância entre SNPs adjacentes há uma diminuição no LD. Assim, as diferenças no LD encontradas nos sub painéis, está relacionado com a maior distância entre SNP adjacentes nos sub painéis de menor densidade e vice-versa. Além da distância entre marcadores, a MAF também pode ter efeito sobre o LD (KHATKAR et al., 2008; ESPIGOLAN et al., 2013).

Embora a metodologia BayesC π faça um controle maior daqueles marcadores que não possuem grande efeito, quando comparada com a metodologia GBLUP, não foram observadas diferenças nas acurácias de predição entre as duas metodologias. Resultado semelhante ao do presente trabalho foi reportado por Daetwyler et al. (2013) utilizando oito metodologias distintas. Por outro lado, Meuwissen, Hayes e Goddard (2001), Habier et al. (2011) e Neves et al. (2014), relatam acurácias de predição diferentes entre as metodologias testadas. Segundo Daetwyler et al. (2013) a acurácia de predição depende da arquitetura genética da característica e da população. Além disso, o fato de haver um grande número de parâmetros a ser estimado e um pequeno número de amostras pode levar a resultados similares entre as metodologias (de los CAMPOS et al., 2013a).

2.4 CONCLUSÃO

Os critérios de controle de qualidade dos genótipos não afetaram a acurácia de predição genômica para a população e características estudadas. O aumento no número de marcadores nos sub painéis, não refletiram em maiores acurácias de predições genômicas. Os sub grupos de marcadores com baixa frequência alélica resultaram em menores estimativas de acurácia de predição do que sub grupos de marcadores com maiores frequências alélicas.

CAPÍTULO 3 - EFEITO DO CONTROLE DE QUALIDADE DOS MARCADORES NOS ESTUDOS DE ASSOCIAÇÃO GENÔMICA AMPLA EM BOVINOS DA RAÇA NELORE

RESUMO - Alguns pesquisadores têm ressaltado a importância do poder dos testes estatísticos durante a fase de descoberta das associações entre marcadores e o fenótipo de interesse. No entanto, o poder do teste estatístico é uma função do tamanho da amostra, frequência do alelo causal e magnitude do seu efeito, desequilíbrio de ligação, estrutura populacional e do número de animais com informações utilizados nas análises estatísticas. Outro fator que pode ter efeito sobre o poder do teste estatístico é a falta de qualidade das informações genômica utilizadas. Contudo, é possível remover marcadores com qualidade questionável por meio de um controle de qualidade dos marcadores. Por outro lado, o controle de qualidade pode remover marcadores que capturam uma quantidade significativa da variância genética aditiva. Sendo assim, o objetivo do presente trabalho foi verificar o efeito do controle de qualidade dos marcadores sobre a proporção de variância genética explicada por cada janela de 1 Mb nos estudos de associação genômica ampla na raça Nelore. Foram utilizadas informações fenotípicas e genotípicas de animais da raça Nelore, pertencentes aos programas de melhoramento genético DeltaGen e CRV Paint. As características consideradas no presente trabalho foram idade ao primeiro parto, peso à desmama e peso ao sobreano. Para os marcadores foram definidos três cenários com diferentes limites de exclusão para os critérios utilizados no controle de qualidade. Os efeitos dos marcadores foram estimados com a metodologia BayesC π . A proporção da variância genética aditiva capturada pelos marcadores em cada janela de 1 Mb foi utilizada para comparar os diferentes cenários do controle de qualidade. O controle de qualidade dos marcadores teve efeito sobre a distância entre marcadores adjacentes, a menor frequência alélica e o desequilíbrio de ligação em cada cromossomo. A proporção da variância genética explicada por cada janela de 1 Mb foi diferente nos três cenários de controle de qualidade. Alteração na classificação das janelas entre os cenários mostra que existe efeito do controle de qualidade sobre os estudos de associação genômica ampla.

Palavras-chaves: desequilíbrio de ligação, menor frequência alélica, variância genética aditiva

ABSTRACT - Some researchers have emphasized the importance of the power of statistical tests during the discovery phase of the associations between markers and phenotype. However, the statistical test is a function of sample size, frequency of causal allele and their effect size, linkage disequilibrium, population structure and number of animals with information used in the statistical analysis. Another factor that can

influence the power of statistical test is the low quality of genotypic information used. Nevertheless, it is possible to remove markers with questionable quality through quality control of makers. On the other hand, the quality control can remove markers that capture a significant amount of genetic variance. The objective of this study was to verify the effect of the quality control of makers on proportion of genetic variance explained by each 1 Mb window in genome-wide association studies in Nellore breed. Phenotype and genotype information of animals belonging to two different breeding programs – DeltaGen and CRV Paint, were used. Age at first calving, weaning weight and yearling weight traits were studied. For the quality control of the markers, three scenarios with different thresholds of exclusion criteria were defined. The BayesC π method was used to estimate the SNP markers effect. The proportion of genetic variance explained within each 1 Mb window was used to compare the different quality control scenarios. The quality control of markers affected on the distance between adjacent markers, minor allele frequency and linkage disequilibrium in each chromosome. The proportion of genetic variance explained for each 1 Mb window was different in the three quality control scenarios. Changes at the window ranking between scenarios have shown that there is effect on the quality control of markers on genome-wide association studies.

Key words: linkage desiquilibrium, minor allele frequency, aditive genetic variance

3.1 INTRODUÇÃO

Nos últimos anos o uso de painéis densos de marcadores SNP (*single nucleotide polymorphism* ou polimorfismos de base única) para estudos de associação genômica ampla (GWAS) cresceu consideravelmente. Com o GWAS, é possível conhecer a arquitetura genética de características complexas de uma determinada população (CASAS et al., 2013). Além disso, segundo Peters et al. (2012), pode-se detectar QTLs e construir bibliotecas de genes que, posteriormente, podem ser utilizadas na seleção genômica. Independente da finalidade, os estudos de GWAS se concentram em localizar marcadores que estejam associados às características de interesse, explorando o desequilíbrio de ligação (LD) entre marcador e variante causal (BOLORMAA et al., 2013).

Mesmo utilizando métodos estatísticos robustos para encontrar associações entre marcadores e o fenótipo de interesse, o poder dos testes estatísticos é uma preocupação constante das pesquisas (MCCARTHY et al., 2008; de los CAMPOS; GIANOLA; ALLISON, 2010; LI et al., 2014). A precisão dos testes estatísticos na fase de descoberta das associações é determinante para o sucesso desses estudos, bem como de pesquisas posteriores como, por exemplo, a seleção genômica, em termos de tempo e capital investido (MCCARTHY et al., 2008; TURNER et al., 2011). No entanto,

o poder do teste estatístico é uma função do tamanho da amostra, frequência do alelo causal e magnitude do seu efeito, desequilíbrio de ligação, estrutura populacional e o número de animais com informações utilizados nas análises estatísticas (GODDARD, 2009; LETTRE, 2011; STRANGER; STAHL; RAJ, 2011; HAYES, 2013).

A qualidade das informações genômicas utilizadas nas análises estatísticas também pode influenciar o poder dos testes estatísticos na associação genômica (TURNER et al., 2011). A falta de qualidade das informações genômicas, pode resultar em associações falso-negativas ou falsas-positivas entre marcadores e variantes causais (ZIEGLER; KONIG; THOMPSON, 2008). No entanto, é possível remover aqueles marcadores com qualidade questionável que possam produzir estimativas viesadas. Por outro lado, o rigor atribuído ao controle de qualidade dos marcadores pode remover SNPs que capturam uma quantidade significativa da variância genética aditiva.

Para os critérios de controle de qualidade são atribuídos limiares que, normalmente, variam de acordo com os objetivos da pesquisa e do pesquisador. Na literatura, são encontrados limiares de exclusão dos marcadores que podem variar de 0,80 a 0,98 para Call rate; 0,01 a 0,05 para MAF; e p-valores entre 0,05 a 0,00001 para HWE (ZIEGLER; KONIG; THOMPSON, 2008; MOSER et al., 2009; KIM et al., 2011; MUJIBI et al., 2011; UTSUNOMIYA et al., 2013). Não há um consenso sobre quais são os limiares de exclusão dos SNPs que devem ser utilizados, bem como, quais são seus impactos sobre os estudos de GWAS. Assim, o objetivo do presente trabalho foi verificar o efeito do controle de qualidade dos marcadores sobre a proporção de variância genética explicada por cada janela de 1 Mb nos estudos de GWAS na raça Nelore, buscando definir critérios de edição adequados. Adicionalmente foi estimado o desequilíbrio de ligação na população estudada.

3.2 MATERIAL E MÉTODOS

Dados

No presente trabalho foram utilizados dados fenotípicos e genotípicos de uma população de animais da raça Nelore, provenientes dos programas de melhoramento genético DeltaGen e CRV Paint. Foram estudadas as características idade ao primeiro parto (IPP), peso à desmama (PD) e peso ao sobreano (PS). Uma descrição das características é apresentada na Tabela 3.1. Os fenótipos foram pré-corrigidos para os efeitos fixos de acordo com a característica. Para obter as soluções dos efeitos fixos foi utilizado um modelo animal uni-característico considerando os efeitos fixos de grupo de contemporâneo (IPP, PD e PS) e as covariáveis idade da mãe ao parto (PD) e idade do animal a pesagem (PD e PS), além dos efeitos aleatórios direto (IPP, PD e

PS) e materno (PD). Os grupos de contemporâneos foram definidos como: fazenda, estação e ano de nascimento, fazenda e grupo de manejo à desmama e ao sobreano para IPP; fazenda, estação, sexo, ano e grupo de manejo ao nascimento e à desmama para PD; e fazenda, estação, sexo, ano e grupo de manejo ao nascimento, à desmama e ao sobreano para PS. Para obter as soluções foram considerados todos os animais que possuíam informação fenotípica. O software AIREMLF90 (MISZTAL et al., 2014) foi utilizado para obter as soluções. Os fenótipos pré-corrigidos foram obtidos através da soma do valor genético e o resíduo. Após a correção dos fenótipos foram mantidos apenas animais que possuíam informações fenotípicas e genotípicas para cada característica. Os animais foram genotipados utilizando o painel de alta densidade Illumina Bovine HD assay (Illumina, San Diego, CA, USA) que possui 777.962 marcadores SNPs espalhados pelo genoma bovino.

Tabela 3.1 – Descrição das características utilizadas

Características	N	Média	SD	Mínimo	Máximo
Idade ao Primeiro Parto	1.817	1.051	136,17	698	1.200
Peso à Desmama	3.228	174,80	28,54	92,00	268,00
Peso ao Sobreano	2.803	275,90	37,00	182,00	385,00

N = número de informações; SD = Desvio Padrão

Edição dos marcadores

Foram considerados apenas marcadores dos cromossomos 1 ao 29 e amostras com taxa de chamada (*Call rate*) maior que 0,90. O efeito do controle de qualidade (CQ) sobre os estudos de associação genômica ampla (GWAS) foi verificado utilizando três cenários. Os cenários foram definidos atribuindo diferentes valores e/ou limiares de exclusão para os critérios: taxa de genotipagem (*Call rate*), menor frequência alélica (MAF), equilíbrio de Hard-Weinberg (HWE) e correlação entre pares de SNPs (r^2). Dessa forma, os cenários foram:

a) Maior rigor (C1): *Call rate* ($< 0,98$), MAF ($< 0,05$), HWE ($p < 10^{-5}$) e r^2 ($> 0,999$). O r^2 foi avaliado dentro de uma janela de 100 marcadores, sendo excluído o SNP de menor MAF de cada par altamente correlacionado.

b) Menor rigor (C2): *Call rate* ($< 0,85$) e MAF ($< 0,01$);

c) Foram excluídos apenas marcadores duplicados (C3).

Para os cenários C1 e C2 o CQ dos marcadores foi efetuado de forma iterativa até que nenhum SNP fosse descartado. Os diferentes cenários foram criados para cada característica em estudo (IPP, PD e PS). Todos os critérios descritos foram calculados utilizando o pacote *snpStats* (CLAYTON, 2014) do R (R Core Team, 2013). Um

resumo do número de amostras e SNPs utilizados em cada cenário e característica está descrito na Tabela 3.2.

Tabela 3.2 – Descrição dos cenários utilizados

Números	IPP	PD	PS
Amostras usadas	1.756	3.165	2.780
Total de SNPs	777.962	777.962	777.962
SNPs usados no C1	316.252	321.988	321.792
SNPs usados no C2	569.720	576.332	576.487
SNps usados no C3	734.272	734.272	734.272

IPP = idade ao primeiro parto; PD = peso à desmama; PS = peso ao sobreano; C1 = cenário de maior rigor; C2 = cenário de menor rigor; C3 = cenário sem controle de qualidade.

Modelo

As soluções para os efeitos dos SNPs foram estimadas utilizando a metodologia BayesC π (HABIER et al., 2011) com o seguinte modelo:

$$y = \mu + \sum_{i=1}^m M_i a_i + e_i \quad (i = 1, \dots, m), \quad (3.1)$$

onde: y é o vetor de pseudo-fenótipo; μ é a média geral; m é o número de SNPs, M_i é o vetor ($N \times 1$) do genótipo do SNP i , a_i é o efeito aditivo do SNP i , e e_i é um vetor de efeitos residuais. Os genótipos dos SNPs foram codificados como 0, 1 e 2, para o homozigoto, heterozigoto e o segundo homozigoto, respectivamente.

Para os efeitos dos marcadores (a_i) o método BayesC π assume *a priori* que todos os SNPs possuem variância comum (σ_a^2), seguindo uma distribuição qui-quadrada escalonada invertida com ν graus de liberdade (4,2) e um parâmetro de escala S . Além disso, com esse método para (a_i) é assumida uma informação *a priori* para o parâmetro π do marcador i . Dessa forma, o efeito de um SNP, ajustado com probabilidade $(1 - \pi)$, vem a partir de uma mistura de distribuições t de Student, ou seja, $a_i | \pi, \sigma_{a_i}^2 \sim \pi N(0, \sigma_{a_i}^2) + (1 - \pi) N(0, \sigma_{a_i}^2)$. Diferente dos métodos propostos por Meuwissen, Hayes e Goddard (2001), no método BayesC π o parâmetro π é tratado como desconhecido e com distribuição uniforme (0, 1) *a priori*. Para os efeitos residuais assume-se *a priori* que: $e_i \sim N(0, I \sigma_e^2)$, onde σ_e^2 , onde σ_e^2 segue uma qui-quadrada escalonada invertida com os parâmetros ν e S , em que, arbitrariamente, ν é menor que 4,2 (HABIER et al., 2011).

Essa metodologia (BayesC π) foi implementada no *software* GS3 (LEGARRA; RICARD; FILANGI, 2014), o qual foi utilizado para estimar os efeitos dos marcadores

no presente estudo. Foram utilizadas 500 mil iterações MCMC, com período de burn-in de 100.000 e os resultados salvos a cada 20 iterações para obter as soluções dos marcadores. Para a variância genética aditiva, foi realizado teste convergência utilizando o pacote estatístico Bayesian Output Analysis Program (BOA) for MCMC (SMITH, 2008) do R (R Core Team, 2013). Além disso, foi realizada uma inspeção visual das cadeias para verificar a convergência das mesmas.

Desequilíbrio de ligação

O desequilíbrio de ligação (LD) entre dois SNPs foi avaliado usando o r^2 (HILL; ROBERTSON, 1968) que é calculado da seguinte forma:

$$r^2 = \frac{D^2}{f(A)f(a)f(B)f(b)}, \quad (3.2)$$

onde $D = f(AB) - f(A)f(B)$, e $f(AB)$, $f(A)$, $f(a)$, $f(B)$ e $f(b)$ são as frequências observadas dos haplótipos AB e dos alelos A, a, B e b, respectivamente, na população. Se dois loci são independentes, a frequência esperada do haplótipo AB é calculada como o produto entre $f(A)$ e $f(B)$. Assim, se $f(AB)$ for maior ou menor que o valor esperado, isso indica que esses dois loci tendem a segregar juntos e estão em LD. Segundo Espigolan et al. (2013) e Khatkar et al. (2008) o r^2 é a medida mais apropriada para estimar o LD em marcadores bialélicos, tais como os SNPs, pois o r^2 é menos sensível a frequências alélicas e o tamanho da amostra. O LD foi calculado para todos os pares de SNPs de cada cromossomo usando o pacote snpStats (CLAYTON, 2014) do R (R Core Team, 2013). Para comparar as diferenças do LD entre os cenários foi realizado um teste de comparação de médias a 5% de significância, utilizando o R (R Core Team, 2013).

Variância dos marcadores

A proporção da variância genética total explicada para cada janela de 1 Mb foi usada para comparar os diferentes cenários do controle de qualidade (PETERS et al., 2012). Segundo LEE et al. (2013) a proporção da variância genética explicada para cada SNP ($\%V_{g_i}$) pode ser calculada da seguinte forma:

$$\%V_{g_i} = \frac{100 * 2p_i q_i \sigma_{a_i}^2}{\sigma_g^2} \quad (3.3)$$

em que p_i e q_i são as frequências alélicas para o i^{esimo} SNP, a_i é o efeito aditivo estimado do i^{esimo} SNP e $\sigma_{a_i}^2$ variância genética explicada pelo SNP.

3.3 RESULTADOS E DISCUSSÕES

A extensão do desequilíbrio de ligação (LD), a distância (d) entre marcadores adjacentes e a menor frequência alélica (MAF) em cada cromossomo variaram entre os três cenários (Figura 3.1). Como estes parâmetros apresentaram comportamento semelhante para as três características estudadas, optou-se em apresentar apenas os resultados da característica IPP. Os valores de LD observados no presente trabalho são semelhantes aos reportados por Espigolan et al. (2013) na raça Nelore e por Mokry et al. (2014) para animais da raça Canchim, ambos utilizando um painel de 777 K. A comparação, entre os resultados do presente estudo e os reportados na literatura, foi feita nas mesmas d média (1 kb) entre marcadores adjacentes.

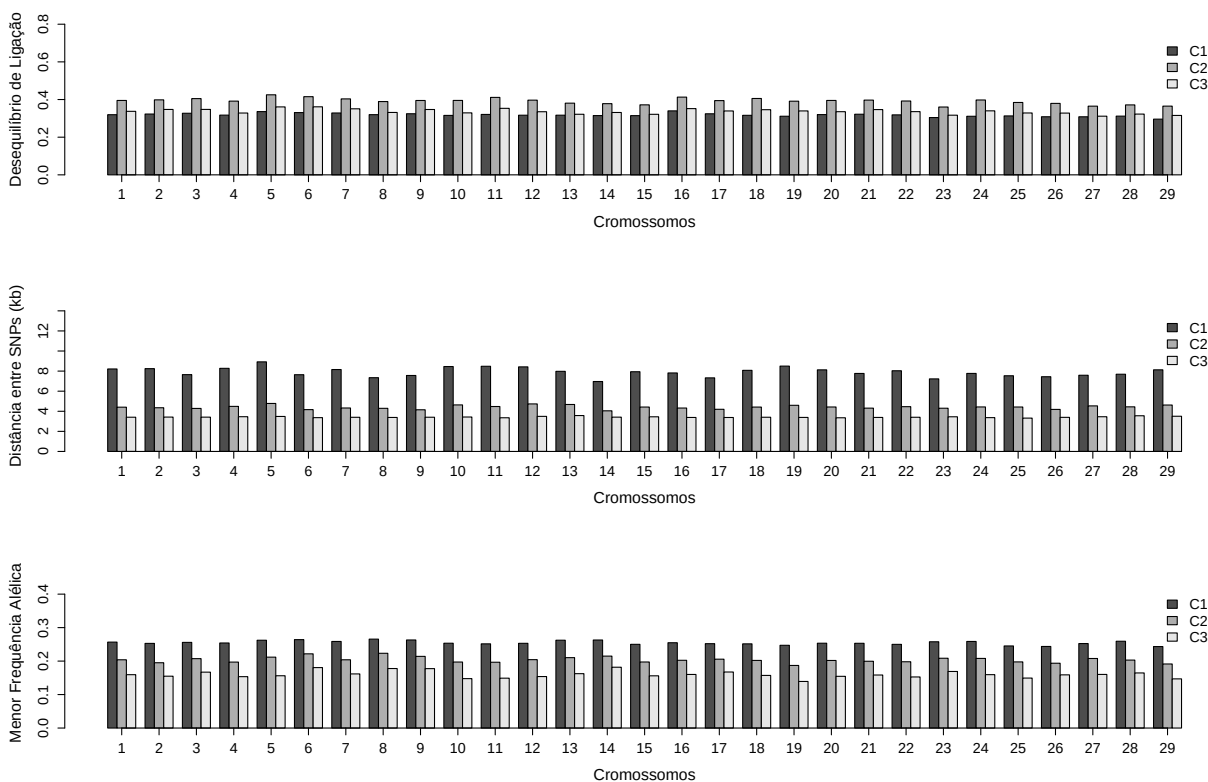


Figura 3.1 – Média do desequilíbrio de ligação (superior), da distância entre marcadores adjacentes (centro) e da menor frequência alélica (inferior) por cromossomo, nos cenários de maior rigor (C1), menor rigor (C2) e sem controle de qualidade (C3) para a característica idade ao primeiro parto.

Os cenários de maior rigor (C1), menor rigor (C2) e sem controle de qualidade (C3) produziram, respectivamente, 39.627.025, 123.921.424 e 215.884.249 combinações possíveis de pares de SNPs para estimar o LD entre os 29 cromossomos. O LD, para cada cromossomo, foi significativamente maior para o cenário C2, seguido pelos cenários C3 e C1. As médias (desvio-padrão) do LD para os cenários C1, C2 e C3

Tabela 3.3 – Classificação das dez principais janelas (1 Mb) com maior proporção da variância genética aditiva capturada pelos marcadores na raça Nelore.

Rank	C1				C2				C3			
	JN	CR	PVAR	N	JN	CR	PVAR	N	JN	CR	PVAR	N
<i>Idade ao Primeiro Parto</i>												
1º	005	18	0,240	185	050	17	0,263	339	050	17	0,260	366
2º	050	17	0,232	206	005	18	0,235	306	005	18	0,236	371
3º	107	08	0,204	202	107	08	0,203	316	107	08	0,202	383
4º	025	23	0,186	333	004	07	0,192	272	004	07	0,189	304
5º	004	07	0,171	184	028	03	0,188	331	028	03	0,189	393
6º	029	23	0,169	247	063	14	0,160	283	063	14	0,159	318
7º	028	03	0,167	205	025	23	0,159	451	025	23	0,154	468
8º	027	01	0,162	220	092	10	0,156	392	031	03	0,153	289
9º	063	14	0,156	180	031	03	0,155	267	092	10	0,152	483
10º	013	05	0,147	213	029	23	0,154	347	027	01	0,151	361
<i>Peso à Desmama</i>												
1º	025	14	0,281	156	003	15	0,270	172	003	15	0,275	222
2º	008	14	0,278	215	008	14	0,255	335	008	14	0,263	396
3º	008	04	0,232	194	025	14	0,249	231	025	14	0,253	281
4º	026	14	0,216	116	026	14	0,236	218	026	14	0,247	273
5º	022	14	0,187	179	022	14	0,212	316	022	14	0,209	335
6º	002	06	0,181	158	008	04	0,204	297	008	04	0,204	339
7º	110	07	0,174	183	002	06	0,183	257	002	06	0,184	304
8º	003	15	0,170	074	110	07	0,173	328	110	07	0,173	337
9º	120	05	0,163	179	069	08	0,156	281	069	08	0,164	343
10º	024	29	0,160	171	120	05	0,151	269	120	05	0,149	359
<i>Peso ao Sobreano</i>												
1º	041	08	0,224	169	036	06	0,255	263	036	06	0,257	299
2º	035	29	0,186	233	041	08	0,234	285	041	08	0,234	398
3º	143	01	0,185	207	003	01	0,192	352	003	01	0,195	399
4º	003	01	0,178	221	035	29	0,185	358	035	29	0,183	416
5º	030	08	0,174	173	003	12	0,181	386	003	12	0,182	410
6º	028	03	0,166	209	028	03	0,180	334	028	03	0,180	393
7º	005	18	0,162	186	030	08	0,174	270	030	08	0,177	349
8º	006	18	0,160	261	143	01	0,173	357	143	01	0,174	396
9º	130	01	0,155	184	005	02	0,156	307	005	02	0,161	357
10º	032	29	0,152	159	005	18	0,154	307	101	04	0,154	339

C1 = cenário de maior rigor; C2 = cenário de menor rigor; C3 = cenário sem controle de qualidade; JN = posição da janela de 1 Mb em cada cromossomo; CR = cromossomo; PVAR = proporção da variância genética explicada pelos SNPs na janela; N = número de SNPs por janela.

foram 0,31 (0,20), 0,39 (0,23) e 0,33 (0,12), respectivamente. Fatores como a d entre marcadores adjacentes e a MAF podem ter efeito sobre o LD (KHATKAR et al., 2008). No presente trabalho foi observado que quanto maior a MAF e a d entre marcadores, por cromossomo, menores foram as estimativas de LD. Resultados semelhantes ao do

presente estudo são reportados por Khatkar et al. (2008) e por Espigolan et al. (2013) em animais da raça Holandesa e Nelore, respectivamente. Khatkar et al. (2008) comentam que com o aumento na MAF, há um aumento no número de pares de SNPs com frequência alélica similar e, conseqüentemente, um aumento no LD, uma vez que as frequências alélicas são consideradas para estimar o LD. Segundo Espigolan et al. (2013), o uso de pares de marcadores com baixa frequência alélica pode subestimar o LD.

O LD têm impacto sobre as estimativas dos efeitos dos marcadores e, conseqüentemente, nos estudos de GWAS e seleção genômica (BARENDSE et al., 2007; BOLORMAA et al., 2013; PORTO-NETO; KIJAS; REVERTER, 2014). Com um menor LD é possível que uma quantidade significativa de QTLs, que contribuem para a variação genética do caracter, não tenham seus efeitos estimados (BARENDSE et al., 2007). Além disso, segundo os autores, perdas de QTLs pode acontecer quando uma menor proporção do genoma está sendo coberta. Dessa forma, mesmo utilizando chip de alta densidade com o objetivo de cobrir todo o genoma e incrementar o LD entre marcador e QTL, fatores como, por exemplo, o CQ pode ter efeito sobre o LD e as estimativas dos efeitos dos marcadores.

A classificação das dez principais janelas de 1 Mb, nos diferentes cenários, mostra que pode existir efeito do CQ sobre os estudos de GWAS para as três características estudadas (Tabela 3.3). Essa classificação tem como base a proporção da variância genética aditiva explicada pelos marcadores dentro de cada janela. Assim, diferenças na variância capturada determinam as mudanças na classificação das janelas observadas nos diferentes cenários do CQ. As maiores alterações na classificação das janelas foram encontradas quando comparado o cenário C1 com os demais cenários (C2 e C3). Este resultado ocorreu, provavelmente, pela exclusão de marcadores, devido ao CQ, que capturavam uma quantidade significativa da variância genética aditiva. Algumas alterações na classificação das dez principais janelas de 1 Mb podem ser destacadas. Para a característica idade ao primeiro parto, por exemplo, a janela 063 no cromossomo 14 que foi classificada em 9º no cenário C1 e 6º nos cenários C2 e C3. É importante observar que o número de marcadores nessa janela é maior nos cenários C2 e C3 que no C1. Outra situação, é a observada com a janela 025 do cromossomo 23 (333 SNP) que foi classificada em 4º no cenário C1, e em 7º nos cenários C2 e C3 (451 e 468 SNP, respectivamente). Neste caso, o aumento no número de marcadores não resultou em maior proporção da variância genética aditiva explicada. Algumas janelas, como por exemplo, a 107 do cromossomo 8, apresenta a mesma classificação nos três cenários do CQ (C1, C2 e C3), mesmo variando o número de SNP. Na característica peso à desmama a janela 003 do cromossomo 15

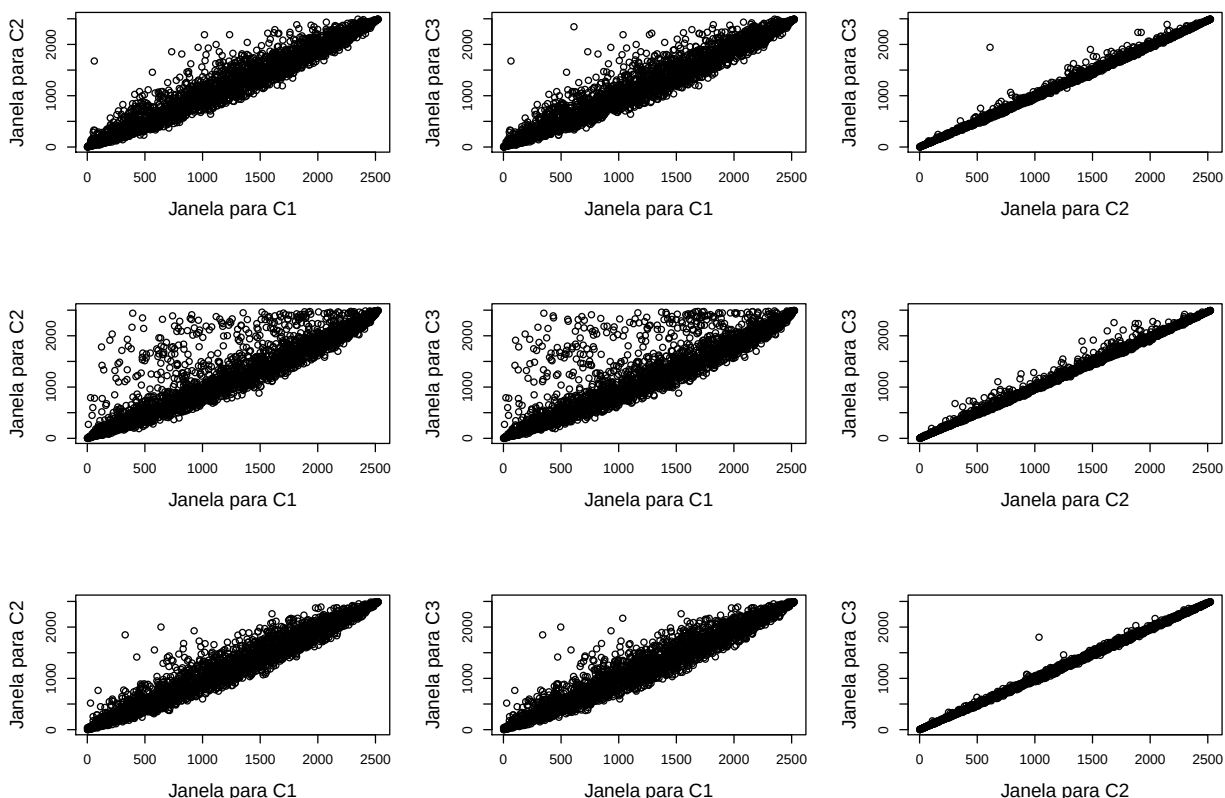


Figura 3.2 – Diagrama de dispersão da classificação das janelas de 1 Mb de acordo com a proporção da variância genética aditiva explicada por cada janela para as características idade ao primeiro parto (superior), peso à desmama (centro) e peso ao sobreano (inferior) nos cenários de maior rigor (C1), menor rigor (C2) e sem controle de qualidade (C3).

foi classificada em 8º para o cenário C1 e como 1º nos cenários C2 e C3. O maior número de marcadores nos cenários C2 e C3 contribuíram para a maior proporção da variância genética aditiva explicada para essa janela. Por outro lado, o maior número de marcadores na janela 022 do cromossomo 14 não aumentou a proporção da variância genética aditiva, pois foi classificada em 5º para os três cenários. Para a característica peso ao sobreano a janela 143 do cromossomo 1 foi classificada em 3º para o cenário C1, porém, em 8º para os cenários C2 e C3. Nesse exemplo o maior número de marcadores nos cenários C2 e C3 não resultou em aumento na proporção de variância genética aditiva.

Alteração na classificação das janelas de 1 Mb não aconteceu apenas nas dez principais, mas também nas demais janelas criadas ao longo do genoma (Figura 3.2). A comparação de todas as janelas entre os cenários mostra as diferenças na classificação das mesmas para as características estudadas. De forma geral, para as três características, as maiores dispersões foram encontradas ao comparar os cenários

C1 com os cenários C2 e C3. As correlações entre as classificações das janelas nos cenários C1 – C2, C1 – C3 e C2 – C3 foram, respectivamente, 0,96, 0,97 e 0,99 para IPP; 0,91, 0,89 e 0,99 para PD e 0,97, 0,97 e 0,99 para PS. Apesar das altas correlações nas classificações das janelas entre os três cenários, as mudanças na classificação das janelas que explicam a maior proporção da variância genética, podem levar a conclusões diversas quanto às regiões genômicas de maior efeito sobre a característica estudada.

3.4 CONCLUSÕES

A distância entre marcadores adjacentes, a menor frequência alélica e o desequilíbrio de ligação foram diferentes nos três cenários do controle de qualidade. O rigor no controle de qualidade dos marcadores teve efeito sobre a proporção da variância genética capturada por cada marcador dentro das janelas de 1 Mb, na população estudada. Os limiares de exclusão atribuídos aos critérios de controle de qualidade dos marcadores devem ser definidos de forma que marcadores importantes não sejam excluídos das análises estatísticas.

REFERÊNCIAS

- ABDOLLAHI-ARPANAHI, R.; NEJATI-JAVAREMI, A.; PAKDEL, A.; MORADI-SHAHRBABA, M.; MOROTA, G.; VALENTE, B.; KRANIS, A.; ROSA, G.; GIANOLA, D. Effect of allele frequencies, effect sizes and number of markers on prediction of quantitative traits in chickens. **Journal of Animal Breeding and Genetics**, v. 131, p. 123–133, 2014.
- ALLAN, M. F.; THALLMAN, R. M.; CUSHMAN, R. A.; WHITE, S. E. E. and S. N.; KUEHN, L. A.; CASAS, E.; SMITH, T. P. L. Association of a single nucleotide polymorphism in SPP1 with growth traits and twinning in a cattle population selected for twinning rate. **Journal of Animal Science**, v. 85, p. 341–347, 2007.
- BARENDSE, W.; REVERTER, A.; BUNCH, R. J.; HARRISON, B. E.; BARRIS, W.; THOMAS, M. B. A validated whole-genome association study of efficient food conversion in cattle. **Genetics**, v. 176, p. 1893–1905, 2007.
- BODDHIREDDY, P.; KELLY, M. J.; NORTHCUTT, S.; PRAYAGA, K. C.; RUMPH, J.; DENISE, S. Genomic predictions in angus cattle: Comparisons of sample size, response variables, and clustering methods for cross-validation. **Journal of Animal Science**, v. 92, p. 485–497, 2014.
- BODDHIREDDY, P.; PRAYAGA, K.; BARROS, P.; DENISE, R. L. S. Genomic predictions of economically important traits in nelore cattle of brazil. In: 10th WORLD CONGRESS OF GENETICS APPLIED TO LIVESTOCK PRODUCTION, 2014, Vancouver, Canadá. **Proceeding...** Vancouver, Canadá, 2014.
- BOLORMAA, S.; PRYCE, J. E.; KEMPER, K. E.; HAYES, B. J.; ZHANG, Y.; TIER, B.; BARENDSE, W.; REVERTER, A.; GODDARD, M. E. Detection of quantitative trait loci in *Bos indicus* and *Bos taurus* cattle using genome-wide association studies. **Genetics Selection Evolution**, v. 45, p. 1–12, 2013.
- BOUWMAN, A. C.; BOVENHUIS, H.; VISKER, M. H. P. W.; ARENDONK, J. A. M. van. Genome-wide association of milk fatty acids in dutch dairy cattle. **Genetics**, v. 12, n. 43, p. 1–12, 2011.
- CASAS, E.; LEACH, R. L.; REINHARDT, T.; THALLMAN, R. M.; LIPPOLIS, J. D.; BENNETT, L. G.; KUEHN, A. L. A genomewide association study identified CYP2J2 as a gene controlling serum vitamin d status in beef cattle. **Journal of Animal Science**, v. 91, p. 3549–3556, 2013.
- CLAYTON, D. **SnpMatrix and XSnpmatrix classes and methods**. [S.l.], 2014. 78 p.
- DAETWYLER, H. D.; CALUS, M. P. L.; PONG-WONG, R.; CAMPOR, G. de los; HICKKEY, J. M. Genomic prediction in animals and plants: Simulation of data, validation, reporting, and benchmarking. **Genetics**, v. 193, p. 347–365, 2013.
- de los CAMPOS, G.; GIANOLA, D.; ALLISON, D. B. Predicting genetic predisposition in humans: the promise of whole-genome markers. **Nature Reviews Genetics**, v. 11, p. 880–886, 2010.

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

de los CAMPOS, G.; VAZQUEZ, A. I.; FERNANDO, R.; KLIMENTIDIS, Y. C.; SORENSEN, D. Prediction of complex human traits using the genomic best linear unbiased predictor. **PLOS Genetics**, v. 9, p. 1–15, 2013.

DEKKERS, J. C. M. Marker-assisted selection for commercial crossbred performance. **Journal of Animal Science**, v. 85, p. 2104–2114, 2007.

EDRISS, V.; GULDBRANDTSEN, B.; LUND, M. S.; SU, G. Effect of marker-data editing on the accuracy of genomic prediction. **Journal of Animal Breeding and Genetics**, v. 130, p. 128–135, 2013.

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

GIANOLA, D.; de los CAMPOS, G.; HILL, W. G.; MANFREDI, E.; FERNANDO, R. Additive genetic variability and the bayesian alphabet. **Genetics**, v. 183, p. 347–363, 2009.

GILL, J. L.; BISHOP, S. C.; MCCORQUODALE, C.; WILLIAMS, J. L.; WIENER, P. Association of selected SNP with carcass and taste panel assessed meat quality traits in a commercial population of Aberdeen Angus-sired beef cattle. **Genetics Selection Evolution**, v. 41, p. 1–12, 2009.

GODDARD, M. Genomic selection: prediction of accuracy and maximisation of long term response. **Genetica**, v. 136, p. 245–257, 2009.

HABIER, D.; FERNANDO, R. L.; KIZILKAYA, K.; GARRICK, D. J. Extension of the bayesian alphabet for genomic selection. **BMC Bioinformatics**, v. 12, n. 186, p. 1–12, 2011.

HABIER, D.; TETENS, J.; SEEFRIED, F. R.; LICHTNER, P.; THALLER, G. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. **Genetics Selection Evolution**, v. 42, n. 5, p. 1–12, 2010.

HAYES, B. **Genome-Wide Association Studies and Genomic Prediction, Methods in Molecular Biology**: Overview of statistical methods for genome-wide association studies (gwas). Hatfield, Hertfordshire, UK: Humana Press, 2013.

HAYES, B.; GODDARD, M. Genome-wide association and genomic selection in animal breeding. **Genome**, v. 53, p. 876–883, 2010.

HAYES, B. J.; BOWMAN, P. J.; CHAMBERLAIN, A. J.; GODDARD, M. E. Genomic selection in dairy cattle: Progress and challenges. **Journal of Dairy Science**, v. 92, p. 433–443, 2009.

HENDERSON, C. R. Sire evaluation and genetic trends. In: ANIMAL BREEDING AND GENETICS SYMPOSIUM IN HONOUR OF J. LUSH, 1973, Champaign. **Proceeding...** Champaign, 1973.

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

ILLUMINA. **GenomeStudio: An integrated platform for data visualization and analysis**. [S.I.], 2009. 184 p. Disponível em: <<http://support.illumina.com/>>.

KHATIB, H.; ZAITOUN, I.; WIEBELHAUS-FINGER, J.; CHANG, Y. M.; ROSA, G. J. M. The association of bovine PPARGC1A and OPN genes with milk composition in two independent Holstein cattle populations. **Journal of Dairy Science**, v. 90, p. 2966–2970, 2007.

KHATKAR, M. S.; NICHOLAS, F. W.; COLLINS, A. R.; ZENGER, K. R.; CAVANAGH, J. A. L.; BARRIS, W.; SCHNABEL, R. D.; TAYLOR, J. F.; RAADSMA, H. W. Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density snp panel. **BMC Genomics**, v. 9, n. 187, p. 1–18, 2008.

KIM, Y.; RYU, J.; WOO, J.; KIM, J. B.; KIM, C. Y.; LEE, C. Genome-wide association study reveals five nucleotide sequence variants for carcass traits in beef cattle. **Animal Genetics**, v. 42, p. 361–365, 2011.

LEE, S. H.; CHOI, B. H.; LIM, D.; GONDRO, C.; CHO, Y. M.; DANG, C. G.; S., A.; JANG, G. W.; LEE, K. T.; YOON, D.; LEE, H. K.; YEON, S. H.; YANG, B. S.; KANG, H. S.; K, H. S. Genome-wide association study identifies major loci for carcass weight on BTA14 in Hanwoo (Korean Cattle). **PLOS ONE**, v. 8, p. 1–9, 2013.

LEGARRA, A.; RICARD, A.; FILANGI, O. **Manual for GS3 program**. [S.I.], 2014. 25 p.

LETTRE, G. Recent progress in the study of the genetics of height. **Human Genetics**, v. 129, p. 465–472, 2011.

LI, X.; WANG, S.; HUANG, J.; LI, L.; ZHANG, Q.; DING, X. Improving the accuracy of genomic prediction in Chinese Holstein cattle by using one-step blending. **Genetics Selection Evolution**, v. 46, n. 66, p. 2014, 2014.

MAKOWSKY, R.; PAJEWSKI, N. M.; KLIMENTIDIS, Y. C.; VAZQUEZ, A. I.; DUARTE, C. W.; ALLISON, D. B.; de los CAMPOS, G. Beyond missing heritability: Prediction of complex traits. **PLOS Genetics**, v. 7, p. 1–9, 2011.

MCCARTHY, M. I.; , G. R. A. ; CARDON, L. R.; GOLDSTEIN, D. B.; LITTLE, J.; IO-ANNIDIS, J. P. A.; HIRSCHHORN, J. N. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. **Nature Publishing Group**, v. 9, p. 356–369, 2008.

MEUWISSEN, T. H. E. Genomic selection: marker assisted selection on genome-wide scale. **Journal of Animal Breeding and Genetics**, v. 124, p. 321–322, 2007.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v. 157, p. 1819–1829, 2001.

MISZTAL, I.; TSURUTA, S.; LOURENCO, D.; AGUILAR, I.; LEGARRA, A.; VITEZICA, Z. **Manual for BLUPF90 family of programs**. [S.l.], 2014. 125 p.

MOKRY, F. B.; BUZANSKAS, M. E.; MUDADU, M. A.; GROSSI, D. A.; HIGA, R. H.; VENTURA, R. V.; LIMA, A. O.; SARGOLZAEI, M.; MEIRELLES, S. L. C.; SCHENKEL, F. S.; SILVA, M. V. G. B.; NICIURA, S. C. M.; ALENCAR, M. M.; MUNARI, D. P.; REGITANO, L. C. A. Linkage disequilibrium and haplotype block structure in a composite beef cattle breed. **BMC Genomics**, v. 15, p. 1–9, 2014.

MOSER, G.; KHATKAR, M. S.; HAYES, B. J.; RAADSMA, H. W. Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. **Genetics Selection Evolution**, v. 42, n. 37, p. 1–15, 2010.

MOSER, G.; TIER, B.; CRUMP, R. E.; KHATKAR, M. S.; RAADSMA, H. W. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. **Genetics Selection Evolution**, v. 41, n. 56, p. 1–16, 2009.

MUJIBI, F. D. N.; NKURUMAH, J. D.; DURUNNA, O. N.; STOTHARD, P.; MAH, J.; WANG, Z.; BASARAB, J.; PLASTOW, G.; CREWS, D. H.; MOORE, J.; MOORE, S. S. Accuracy of genomic breeding values for residual feed intake in crossbred beef cattle. **Journal of Animal Science**, v. 89, p. 3353–3361, 2011.

NEVES, H. H. R.; CARVALHEIRO, R.; O'BRIEN, A. M. P.; UTSUNOMIYA, Y. T.; CARMO, A. S.; SCHENKEL, F. S.; SÖLKNER, J.; MCEWAN, J. C.; TASSELL, C. P. V.; COLE, J. B.; SILVA, M. V. G. B.; QUEIROZ, S. A.; SONSTEGARD, T. S.; GARCIA, J. F. Accuracy of genomic predictions in *Bos indicus* (Nellore) cattle. **Genetics Selection Evolution**, v. 46, n. 17, p. 1–13, 2014.

OGAWA, S.; MATSUDA, H.; TANIGUCHI, Y.; WATANABE, T.; NISHIMURA, S.; SUGIMOTO, Y.; IWAISAKI, H. Effects of single nucleotide polymorphism marker density on degree of genetic variance explained and genomic evaluation for carcass traits in Japanese Black beef cattle. **BMC Genomics**, v. 15, n. 15, p. 1–13, 2014.

OLIPHANT, A.; BARKER, D. L.; STUELPNAGEL, J. R.; CHEE, M. S. Beadarray technology: Enabling an accurate, cost-effective approach to high-throughput genotyping. **BioTechniques**, v. 32, p. S56–S61, 2002.

PARK, J. H.; GAIL, M. H.; WEINBERG, C. R.; CARROLL, R. J.; CHUNG, C. C.; WANG, Z.; CHANOCK, S. J.; FRAUMENI, J. F.; CHATTERJEE, N. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. **Proceedings of the National Academy of Sciences**, v. 108, p. 18026–18031, 2011.

PETERS, S. O.; KIZILKAYA, K.; GARRICK, D. J.; FERNANDO, R. L.; REECY, J. M.; WEABER, R. L.; SILVER, G. A.; THOMAS, M. G. Bayesian genome wide association analyses of growth and yearling ultrasound measures of carcass traits in Brangus heifers. **Journal of Animal Science**, v. 90, p. 3398–3409, 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, n. 22, p. 1–5, 2014.

PRYCE, J. E.; ARIAS, J.; BOWMAN, P. J.; DAVIS, S. R.; MACDONALD, K. A.; WAGHORN, G. C.; WALES, W. J.; WILLIAMS, Y. J.; SPELMAN, R. J.; HAYES, B. J. Accuracy of genomic predictions of residual feed intake and 250-day body weight in growing heifers using 625,000 single nucleotide polymorphism markers. **Journal of Dairy Science**, v. 95, p. 2108–2119, 2012.

PSZCZOLA, M.; STRABEL, T.; MULDER, H. A.; CALUS, M. P. L. Reliability of direct genomic values for animals with different relationships within and to the reference population. **Journal of Dairy Science**, v. 95, p. 389–400, 2012.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2013. Disponível em: <<http://www.R-project.org/>>.

RESENDE, M. D. V.; SILVA, F. F.; LOPES, P. S.; AZEVEDO, C. F. **Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada (RRM) e Estatística Espacial**. 21. ed. Viçosa: Universidade Federal de Viçosa/Departamento de Estatística, 2012. 291 p.

RINCON, G.; WEBER, K. L.; EENENNAAM, A. L. V. V.; GOLDEN, B. L.; MEDRANO, J. F. Hot topic: Performance of bovine high-density genotyping platforms in Holsteins and Jerseys. **American Dairy Science Association**, v. 94, n. 12, p. 6116–6121, 2011.

SAATCHI, M.; MCCLURE, M. C.; MCKAY, S. D.; ROLF, M. M.; KIM, J.; DECKER, J. E.; TAXIS, T. M.; CHAPPLE, R. H.; RAMEY, H. R.; NORTHCUTT, S. L.; BAUCK, S.; WOODWARD, B.; DEKKERS, J. C. M.; FERNANDO, R. L.; SCHNABEL, R. D.; GARRICK, D. J.; TAYLOR, J. F. Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. **Genetics Selection Evolution**, v. 43, n. 40, 2011.

SARGOLZAEI, M.; CHESNAIS, J. P.; SCHENKEL, F. S. A new approach for efficient genotype imputation using information from relatives. **BMC Genomics**, v. 15, n. 478, 2014.

SCHEET, P.; STEPHENS, M. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. **The American Journal of Human Genetics**, v. 78, p. 629–644, 2006.

SCHENKEL, F. S.; MILLER, S. P.; YE, X.; MOORE, S. S.; NKRUMAH, J. D.; LI, C.; YU, J.; MANDELL, I. B.; WILTON, J. W.; WILLIAMS, J. L. Association of single nucleotide polymorphisms in the leptin gene with carcass and meat quality traits of beef cattle. **Journal of Animal Science**, v. 83, p. 2009–2020, 2005.

SMITH, B. J. **Bayesian Output Analysis Program (BOA) for MCMC**. [S.l.], 2008. 40 p. Disponível em: <<http://www.public-health.uiowa.edu/boa>>.

STRANGER, B. E.; STAHL, E. A.; RAJ, T. Progress and promise of genome-wide association studies for human complex trait genetics. **Genetics**, v. 187, p. 367–383, 2011.

TURNER, S.; ARMSTRONG, L. L.; BRADFORD, Y.; CARLSON, C. S.; CRAWFORD, D. C.; CRENSHAW, A. T.; ANDRADE, M.; DOHENY, K. F.; HAINES, J. L.; HAYES, G.;

JARVIK, G.; JIANG, L.; KULLO, I. J.; LI, R.; LING, H.; MANOLIO, T. A.; MATSUMOTO, M.; MCCARTY, C. A.; MCDAVID, A. N.; MIREL, D. B.; PASCHALL, J. E.; PUGH, E. W.; RASMUSSEN, L. V.; WILKE, R. A.; ZUVICH, R. L.; RITCHIE, M. D. **Quality control procedures for genome wide association studies**. [S.l.], 2011. 1-29 p. Disponível em: <doi:10.1002/0471142905.hg0119s68>.

UTSUNOMIYA, Y. T.; CARMO, A. S.; CARVALHEIRO, R.; NEVES, H. H. R.; MATOS, M. C.; ZAVAREZ, L. B.; O'BRIEN, A. M. P.; SÖLKNER, J.; MCEWAN, J. C.; COLE, J. B.; TASSELL, C. P. V.; SCHENKEL, F. S.; SILVA, M. V. G. B.; NETO, L. R. P.; SONSTEGARD, T. S.; GARCIA, J. F. Genome-wide association study for birth weight in Nellore cattle points to previously described orthologous genes affecting human and bovine height. **BMC Genetics**, v. 14, n. 52, p. 1–12, 2013.

VANRADEN, P. M. Efficient methods to compute genomic predictions. **Journal of Dairy Science**, v. 91, p. 4414–4423, 2008.

VAZQUEZ, A. I.; ROSA, G. J. M.; WEIGEL, K. A.; de los CAMPOS, G.; GIANOLA, D.; ALLISON, D. B. Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. **Journal of Dairy Science**, v. 93, p. 5942–5949, 2010.

WEBER, K. L.; THALLMAN, R. M.; KEELE, J. W.; SNELLING, W. M.; BENNETT, G. L.; SMITH, T. P. L.; MCDANELD, T. G.; ALLAN, M. F.; EENENNAAM, A. L. V.; KUEHN, L. A. Accuracy of genomic breeding values in multibreed beef cattle populations derived from deregressed breeding values and phenotypes. **Journal of Animal Science**, v. 90, p. 4177–4190, 2012.

WEIGEL, K. A.; de los CAMPOS, G.; GONZÁLEZ-RECIO, O.; NAYA, H.; WU, X. L.; LONG, N.; ROSA, G. J. M.; GIANOLA, D. Predictive ability of direct genomic values for lifetime net merit of Holstein. **Journal of Dairy Science**, v. 92, p. 5248–5257, 2009.

WIGGANS, G. R.; SONSTEGARD, T. S.; VANRADEN, P. M.; MATUKUMALLI, L. K.; SCHNABEL, R. D.; TAYLOR, J. F.; SCHENKEL, F. S.; TASSELL, C. P. V. Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. **Journal of Dairy Science**, v. 92, p. 3431–3436, 2009.

ZHANG, Y. D.; JOHNSTON, D. J.; BOLORMAA, S.; HAWKEN, R. J.; TIER, B. Genomic selection for female reproduction in Australian tropically adapted beef cattle. **Animal Production Science**, v. 54, p. 16–24, 2014.

ZHANG, Z.; DING, X.; LIU, J.; ZHANG, Q.; KONING, D. de. Accuracy of genomic prediction using low-density marker panels. **Journal of Dairy Science**, v. 94, p. 3642–3650, 2011.

ZHANG, Z.; OBER, U.; ERBE, M.; ZHANG, H.; GAO, N.; HE, J.; LI, J.; SIMIANER, H. Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. **PLOS One**, v. 9, p. 1–12, 2014.

ZIEGLER, A.; KONIG, I. R.; THOMPSON, J. R. Biostatistical aspects of genome-wide association studies. **Biometrical Journal**, v. 50, p. 8–28, 2008.