

UNIVERSIDADE ESTADUAL PAULISTA - UNESP
CÂMPUS DE JABOTICABAL

**IMPUTAÇÃO E ESTUDOS GENÔMICOS DE BOVINOS
NELORE**

Priscila Arrigucci Bernardes
Médica veterinária

2018

UNIVERSIDADE ESTADUAL PAULISTA - UNESP
CÂMPUS DE JABOTICABAL

**IMPUTAÇÃO E ESTUDOS GENÔMICOS DE BOVINOS
NELORE**

Priscila Arrigucci Bernardes

Orientador: Prof. Dr. Danísio Prado Munari

Coorientador: Prof. Dr. Ricardo Vieira Ventura

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

2018

Bernardes, Priscila Arrigucci
B521i Imputação e estudos genômicos de bovinos Nelore / Priscila Arrigucci Bernardes. -- Jaboticabal, 2018
vii, 99 p. : il. ; 29 cm

Tese (doutorado) - Universidade Estadual Paulista, Faculdade de Ciências Agrárias e Veterinárias, 2018

Orientador: Danísio Prado Munari

Coorientador: Ricardo Vieira Ventura

Banca examinadora: Lenira El Faro Zadra, Ana Fabrícia Braga Magalhães, Rodrigo Pelicioni Savegnago, João Ademir de Oliveira

Bibliografia

1. Affymetrix. 2. Bovinos de corte. 3. Illumina. 4. Informação genômica. 5. Segmentos de homozigose. I. Título. II. Jaboticabal-Faculdade de Ciências Agrárias e Veterinárias.

CDU 636.082:636.2

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

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: IMPUTAÇÃO E ESTUDOS GENÔMICOS DE BOVINOS NELORE

AUTORA: PRISCILA ARRIGUCCI BERNARDES

ORIENTADOR: DANISIO PRADO MUNARI

COORIENTADOR: RICARDO VIEIRA VENTURA

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



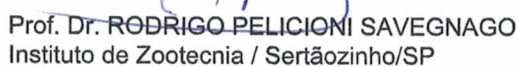
Prof. Dr. DANISIO PRADO MUNARI
Departamento de Ciências Exatas / FCAV / UNESP - Jaboticabal



Pesquisadora Dra. LENIRA EL FARO ZADRA
Instituto de Zootecnia / Sertãozinho/SP



Pós-doutoranda ANA FABRÍCIA BRAGA MAGALHÃES
Departamento de Zootecnia / FCAV / UNESP - Jaboticabal



Prof. Dr. RODRIGO PELICIONI SAVEGNAGO
Instituto de Zootecnia / Sertãozinho/SP



Prof. Dr. JOÃO ADEMIR DE OLIVEIRA
Departamento de Ciências Exatas / FCAV / UNESP - Jaboticabal

Jaboticabal, 29 de junho de 2018

DADOS CURRICULARES DO AUTOR

PRISCILA ARRIGUCCI BERNARDES – nascida em 22 de julho de 1987, na cidade de São João da Boa Vista, São Paulo, filha de Gilberto de Azevedo Bernardes e Luciana de Carvalho Arrigucci Bernardes. Iniciou o curso de Medicina Veterinária na Faculdade de Ciência Agrárias e Veterinárias, UNESP, Campus de Jaboticabal em março de 2006. Foi bolsista FAPESP de iniciação científica em 2009 e obteve o título de Médica Veterinária em 2011. Em agosto de 2012, ingressou no curso de mestrado pelo Programa de Pós-Graduação em Genética e Melhoramento Animal na Faculdade de Ciência Agrárias e Veterinárias, UNESP, Campus de Jaboticabal sob orientação do Prof. Dr. Danísio Prado Munari e coorientação da Dr. Daniela do Amaral Grossi e do Dr. Rodrigo Pelicioni Savegnago. Foi bolsista FAPESP de mestrado pelo período de agosto de 2013 a agosto de 2014. Em agosto de 2014, ingressou no curso de doutorado pelo Programa de Pós-Graduação em Genética e Melhoramento Animal na Faculdade de Ciência Agrárias e Veterinárias, UNESP, Campus de Jaboticabal sob orientação do Prof. Dr. Danísio Prado Munari e coorientação do Dr. Ricardo Vieira Ventura. Realizou estágio sanduíche com bolsa estágio de pesquisa no exterior (BEPE) na University of New England no período de março de 2017 a fevereiro de 2018 sob supervisão do Dr. Cedric Gondro. Foi bolsista FAPESP de março de 2016 a junho de 2018.

O dinheiro faz homens ricos,
o conhecimento faz homens sábios
e a humildade faz grandes homens.

(Mahatma Gandhi)

Aos meus amados pais Luciana e Gilberto,
Ao meu namorado que tanto amo,
Aos meus queridos irmãos e avós,
Exemplos de paciência, amor e dedicação.

AGRADECIMENTOS

Gostaria de agradecer a Deus por sempre estar ao meu lado e me dar forças para continuar firme e sempre seguir em frente.

Aos meus pais por me proporcionarem tudo o que tenho até hoje e por não ter medido esforços para me fornecer a melhor educação possível. Por sempre me fazer olhar para o lado positivo de tudo e aprender a amar o que faz. Aos meus irmãos Bruno e Thaís, por participarem do meu crescimento e aprendizagem da vida e me apoiarem, estando sempre dispostos a ajudar. Aos meus avós Márcia e Antenor (*in memoriam*) por toda a dedicação em minha criação e o apoio ao longo dos meus estudos. À minha querida avó Marly por todos os mimos, por ser meu exemplo de vida e minha inspiração no amor aos estudos. Ao meu avô Tabajara (*in memoriam*) por toda paciência que teve e por me despertar a curiosidade pelas coisas da vida, essência importante para um bom pesquisador. A minha prima Carolina, que participou de todos os momentos bons e ruins pelo qual passei, estando sempre ao meu lado, fisicamente ou não, para me apoiar nas decisões mais difíceis. Enfim, a toda a minha família, que colaborou cada um a sua maneira para hoje eu ser quem sou.

Ao meu namorado, parceiro e companheiro de todas as horas Eliéder. Obrigada por estar ao meu lado sempre, me inspirando, ensinando e ajudando em tudo o que pode. Agradeço por ser meu suporte de todos os momentos e por me fazer muito feliz. Aos meus sogros Lucélia e Cidinho, por me tratar como filha, sempre ajudando e se disponibilizando para o que for necessário, assim como toda família.

Gostaria de agradecer a todos meus amigos de São João da Boa Vista, Ana Lia, Ana Luiza, Bruna, Cacainho, Du, Gabi, Gui, Jorge, Kaká, Neto, Má, Marcilio, Sy, Talitha, Vê, Victor, que me suportaram por tanto tempo e me ajudaram a formar a pessoa que hoje sou, com certeza são as pessoas que considero da minha família e são as que eu escolhi para ter ao meu lado sempre.

Às minhas queridas amigas Lary e Tetxu, por me divertirem e chorarem comigo quando preciso. Obrigada por serem sempre tão maravilhosas e essenciais na minha vida. Às demais amigas que a faculdade me trouxe Franga, Charlie, Barraka, Disput's, Nariz, Fórfi, Tandela e Pírim, por todos os momentos de estudos, choros, diversão e muita comilança. Também aos meninos Espirro, Pogo, Ieth, Roska, Pastel, Mentor, Bitcha, Piri e todos da VET06 que muito me divertiram nos melhores anos da minha vida.

Gostaria de agradecer a família que escolhi aqui em Jaboticabal, República As Coyotes, desde as moradoras mais antigas que convivi até as mais novas, as quais não conviveram comigo e ainda me consideram da família e me recebem sempre com sorrisos e braços abertos.

À todo o pessoal do EAGMA Tati, Jaque, Ismael, Thiago, Salvador, Mi, Sabrina, Diego, Nicole, Guilherme Venturini, Val, Natalia, Bia, Jorge, Marcos, Leticia, Ana Paula, Alejandro, Rebeka, Samla, Záfia, Lucas e Rafael, obrigada por fazer os dias melhores e por toda ajuda que me deram. Sem vocês este trabalho não seria o mesmo.

Aos funcionários do Departamento de Exatas, Adriana, Zezé e Shirley e aos funcionários da pós-graduação por proporcionar o ambiente de trabalho o melhor possível.

À EMBRAPA pecuária sudeste, Dra. Luciana Regitano e toda sua equipe pela colaboração e por fornecer os dados para a realização deste trabalho.

Aos membros da banca de qualificação Dr. Rodrigo, Dra. Ana Fabrícia, Dra. Nedenia e Dr. Guilherme, pela amizade, pelo tempo dispendido e considerações realizadas para melhoria do presente trabalho. Aos membros da banca de defesa da tese Dra. Lenira, pela paciência, colaboração e sugestões e Dr. João Ademir pela paciência, consideração e por todos os ensinamentos desde a graduação.

Ao meu coorientador Dr. Ricardo pelo apoio, ensino, orientação e sugestões que foram muito importantes para o desenvolvimento deste trabalho. A todos os professores que tive desde o início de minha educação até minha formação profissional, por terem dedicado tempo para minha formação. Ao Prof. Dr. José Jurandir Fagliari por ter me aceitado como estagiária e estudante de iniciação, me apresentando a área acadêmica, ensinando, orientando e me incentivando sempre.

Em especial ao Prof. Dr. Danísio Prado Munari por me receber como aluna e orientada, tendo muita paciência e dedicação, por todos os ensinamentos e conselhos e por sempre se preocupar não somente com minha formação profissional, mas também com minha formação pessoal.

Ao Dr. Cedric Gondro e sua esposa Dra. Simone, por todo o aprendizado e por me receberem muito bem, fazendo com que eu me sentisse a vontade mesmo não estando no meu país. À University of New England pela oportunidade única. À inspiradora e guerreira Sara com sua linda filha Ana, assim como aos grandes amigos que fiz em Armidale, Netsanet, Antonio, José, Peter, Naomi, Bank, Gian, Matias, Rodrigo, Paulo, Rodrigo gaúcho, Tharcilla e Flávio e muitos outros que aqui não citei, obrigada por fazer dessa experiência uma das mais incríveis e por todos os momentos divertidos que passamos juntos.

À FCAV-UNESP Jaboticabal por proporcionar excelente ambiente para formação profissional. À CAPES pela bolsa de estudos concedida de agosto de 2014 a março de 2016, à FAPESP por conceder bolsa de estudos no país (Processo 2015/25096-6) no período de março de 2016 a junho de 2018 e de estágio de pesquisa no exterior (Processo: 2016/22940-3) no período de março de 2017 a fevereiro de 2018, as quais foram fundamentais para a realização do presente trabalho.

Muito obrigada!

SUMÁRIO

	Página
RESUMO.....	IV
ABSTRACT.....	V
LISTA DE ABREVIATURAS.....	VI
CAPÍTULO 1 – CONSIDERAÇÕES GERAIS.....	1
1. INTRODUÇÃO	1
2. REVISÃO DE LITERATURA	3
2.1. <i>Desequilíbrio de ligação.....</i>	<i>3</i>
2.2. <i>Blocos de haplótipos.....</i>	<i>5</i>
2.3. <i>Imputação</i>	<i>7</i>
2.4. <i>Redes Neurais Artificiais</i>	<i>11</i>
2.5. <i>Segmentos de homozigose (ROH).....</i>	<i>13</i>
3. REFERÊNCIAS BIBLIOGRÁFICAS.....	16
CAPÍTULO 2 - ESTUDO DE IMPUTAÇÃO EM BOVINOS DA RAÇA NELORE UTILIZANDO A COMBINAÇÃO DE PAINÉIS DE ALTA DENSIDADE	22
1. INTRODUÇÃO	23
2. MATERIAL E MÉTODOS	24
2.1. <i>Descrição dos dados e controle de qualidade.....</i>	<i>24</i>
2.2. <i>Painéis considerados para a imputação.....</i>	<i>26</i>
2.3. <i>Imputação</i>	<i>30</i>
2.4. <i>Desequilíbrio de ligação e bloco de haplótipos</i>	<i>31</i>
3. RESULTADOS E DISCUSSÃO	32
3.1. <i>Imputação</i>	<i>32</i>

3.2. Desequilíbrio de ligação (DL) e bloco de haplótipos	40
4. CONCLUSÃO	47
5. REFERÊNCIAS	47
CAPÍTULO 3. PREDIÇÃO DA ACURÁCIA DE IMPUTAÇÃO EM BOVINOS DA RAÇA NELORE UTILIZANDO REDES NEURAS ARTIFICIAIS	51
1. INTRODUÇÃO	52
2. MATERIAL E MÉTODOS	53
2.1. Descrição dos dados.....	53
2.2. Controle de qualidade	54
2.3. Determinação da composição genética dos animais	55
2.4. Descrição dos cenários de imputação.....	55
2.5. Redes neurais artificiais: estrutura e funcionamento.....	57
2.6. Redes neurais artificiais e regressão linear múltipla na predição de acurácia de imputação	58
2.7. Modelos utilizados para predição de acurácia de imputação	60
2.8. Análise hierárquica e não hierárquica	62
3. RESULTADOS E DISCUSSÃO	63
3.1. Predição da acurácia de imputação	63
3.2. Análise hierárquica e não hierárquica	70
4. CONCLUSÃO	72
5. RERÊNCIAS.....	73
CAPÍTULO 4 - ESTUDO DE ENDOGAMIA UTILIZANDO SEGMENTOS DE HOMOZIGOSE EM BOVINOS NELORE	76
1. INTRODUÇÃO	77
2. MATERIAL E MÉTODOS	78
2.1. Descrição dos dados e controle de qualidade.....	78

2.2. <i>Análise de segmentos de homozigose</i>	79
3. RESULTADOS E DISCUSSÃO	80
4. CONCLUSÃO	94
5. REFERÊNCIAS	94
APÊNDICE	98
APÊNDICE A.	99

IMPUTAÇÃO E ESTUDOS GENÔMICOS DE BOVINOS NELORE

RESUMO – Dentre as informações fornecidas pelas metodologias que utilizam marcadores do tipo polimorfismo de nucleotídeo único (SNPs), as de segmentos de homozigose (ROH) e de desequilíbrio de ligação têm colaborado para estudos de aplicação direta da informação genômica em populações de bovinos de corte, como em estudos de associação com cobertura ampla do genoma, de seleção genômica e de estrutura da população, dentre outros. Atualmente a imputação vem sendo utilizada principalmente para reduzir custos com a genotipagem dos animais e pode ser utilizada combinando informações genômicas de diferentes painéis. Para que dados imputados sejam utilizados de forma eficiente, é necessário que a imputação tenha sido implementada de forma que todos os animais tenham seus genótipos inferidos com elevada acurácia. No entanto, esta é verificada apenas se houver o genótipo real para avaliar a confiabilidade do genótipo imputado. Dessa maneira, os objetivos deste trabalho foram: (1) estudar a imputação de painéis comercial e customizados de baixa densidade para painéis de alta densidade (Illumina e Affymetrix), assim como para um painel combinado (Illumina + Affymetrix) para bovinos da raça Nelore, e estudar o desequilíbrio de ligação e conformação de blocos de haplótipos antes e após a imputação; (2) estudar estratégias para predição da acurácia de imputação, utilizando redes neurais artificiais e regressão linear múltipla; (3) estudar os segmentos de homozigose e, com isso, a endogamia presente em uma população de bovinos da raça Nelore, assim como identificar os genes presentes nos segmentos de homozigose mais frequentes na população. Os estudos de ROH foram realizados com utilização de informações de 34 touros de diferentes linhagens e suas progênes, totalizando 809 animais genotipados da raça Nelore com informação de 509.107 SNPs (Illumina). Para as análises de imputação e de predição da acurácia de imputação foram utilizados os mesmos animais, sendo que 93 destes também foram genotipados com o painel “Axion Genome-Wide BOS 1 Array Plate”. A partir dos resultados das análises de imputação demonstrou-se que o uso combinado de painéis pode ser uma alternativa para aumentar a densidade e o número de bloco de haplótipos, aumentando a probabilidade de obter um marcador próximo a um QTL de interesse. Além disso, essa estratégia indica que a escolha de SNPs em comum entre os painéis de alta densidade (Illumina e Affymetrix) pode ser utilizada para customizar um painel de menor densidade, permitindo elevar a acurácia de imputação do painel da Illumina e Affymetrix. Na análise de predição da acurácia de imputação, a utilização de redes neurais artificiais foi mais eficiente comparada ao modelo de regressão linear múltipla, podendo ser utilizada com esta finalidade. A partir dos resultados das análises de ROH observou-se que a população encontra-se com baixo nível de endogamia, no entanto os reprodutores apresentaram maior valor de endogamia comparado a progênie, o qual somado a presença de segmentos de homozigose mais longos nestes animais podem indicar que tenha ocorrido intensa utilização de poucos reprodutores nas gerações mais recentes em algumas famílias.

Palavras-chave: Affymetrix, bovinos de corte, Illumina, informação genômica, segmentos de homozigose

IMPUTATION AND GENOMIC STUDIES IN BOVINE NELORE

ABSTRACT – Among all the information provided by methodologies that use single nucleotide polymorphism (SNPs), the runs of homozygosity (ROH) and linkage disequilibrium have been used for studies that explore genomic information in beef cattle population, as the genome-wide association, genomic selection, the structure of population and others. Nowadays, the imputation is used in these studies to reduce genomic costs and this also can be used combining genomic information from different panels. The animals used to be imputed should present genotypes inferred with high accuracy to allow the use imputed genotypes in other studies. However, the accuracy is verified only if there is a real genotype to evaluate the imputed genotype. Therefore, this study aimed: (1) Evaluate imputation of commercial and customized low density panels to high density panels (Illumina and Affymetrix), as well as to a combined panel (Illumina + Affymetrix) in Nelore beef cattle, and estimating linkage disequilibrium and haplotype blocks conformation to high density panels individually and after imputation; (2) Study a strategy to predict imputation accuracy using artificial neural network and linear regression; (3) Study runs of homozygosity and inbreeding in a populations from Nelore beef cattle, as well as identify genes present in ROH with high frequency in population. For ROH studies were used 34 bulls from different lines and the progeny, totalizing 809 Nelore animals genotyped with information of 509.107 SNPs (Illumina). The imputation analysis and imputation accuracy prediction used the same animals, wherein 93 were also genotyped with Axion Genome-Wide BOS 1 Array Plate. The imputation analysis demonstrates that the combined panels used from different panels can be considered due to increasing the density and number of haplotype blocks, increasing the probability to find a marker close to an important QTL. Furthermore, this strategy indicates that the choice for common SNPs between high-density panels Illumina and Affymetrix to customize a lower density panel can increase the imputation accuracy to Illumina and Affymetrix. The prediction of imputation accuracy analysis showed that the neural network is more efficient compared to linear regression, and could be used for this purpose. The results from ROH analysis showed low population inbreeding, however the sires presented higher inbreeding compared to progenies and longer runs of homozygosity, which suggest that has occurred intense use of few sires in recent generations in some families.

Keywords: beef cattle, genomic information, runs of homozygosity, Illumina, Affymetrix

LISTA DE ABREVIATURAS

SNP: Polimorfismo de nucleotídeo único

ROH: Segmentos de homozigose

DL: Desequilíbrio de ligação

GWAS: Associação com cobertura ampla do genoma

MAF: Frequência alélica do Alelo de menor frequência

MPL: “Multilayer perceptron”

QTL: *Loci* de características quantitativas

IllumHD: Painel BovineHD BeadChip

AffyHD: Painel Axion Genome-Wide BOS 1 Array Plate

PC: Painel combinado

20kCom: Painel GeneSeek Genomic Profiler LD v2

50kCom: Painel Illumina BovineSNP50 v2 BeadChip

20kCust1: Painel customizado da 1° forma com densidade 20 mil SNPs

20kCust2: Painel customizado da 2° forma com densidade de 20 mil SNPs

20kCust3: Painel customizado da 3° forma com densidade 20 mil SNPs

50kCust1: Painel customizado da 1° forma com densidade 50 mil SNPs

50kCust2: Painel customizado da 2° forma com densidade 50 mil SNPs

50kCust3: Painel customizado da 3° forma com densidade 50 mil SNPs

PROP: Proporção de genótipos imputados corretamente

COR: Correlação simples de Pearson entre o genótipo imputado e observado

Novo_DL: Pares de SNPs que estavam ausentes nos outros painéis de alta densidade

DP: Desvio-padrão

BTA: Cromossomo

20k: Painel GeneSeek Genomic Profiler LD v2

50k: Painel Illumina BovineSNP50 v2 BeadChip

700k: Painel BovineHD BeadChip

600k: Painel Axion Genome-Wide BOS 1 Array Plate

800k: Painel combinado

QME: Quadrado médio dos erros

NN: Redes neurais artificiais

LM: Regressão linear múltipla

BLUP: Melhor predição linear não-viesada

CAPÍTULO 1 – Considerações gerais

1. INTRODUÇÃO

As tecnologias que utilizam informações de marcadores do tipo polimorfismo de nucleotídeo único (SNPs) permitiram grandes avanços na área de melhoramento genético animal, uma vez que permitem o estudo aprofundado de características com baixa resposta ao processo de seleção tradicional. Dentre as metodologias que utilizam a informação de SNPs, as informações fornecidas por segmentos de homozigose (ROH – “Runs of homozygosity”) e de desequilíbrio de ligação (DL) têm colaborado para estudos de associação com cobertura ampla do genoma (GWAS), de seleção genômica e de estrutura da população (MEUWISSEN; HAYES; GODDARD, 2001; BUZANSKAS et al., 2014).

O método de ROH permite identificar o tamanho e a quantidade dos segmentos em homozigose, os quais fornecem informações importantes sobre o histórico populacional e relações genéticas entre os animais (FERENČAKOVIĆ et al., 2013). O DL é a associação não aleatória entre alelos de dois ou mais *loci*, e esta associação pode diferir ao longo do genoma, pois na herança genética os alelos são passados para a próxima geração em blocos denominados haplótipos, os quais possibilitam que alelos herdados juntos apresentem maior associação (O'BRIEN et al., 2014). Estas informações podem auxiliar na compreensão de diferenças genômicas observadas entre populações de animais taurinos (*Bos taurus taurus*) e zebuínos (*Bos taurus indicus*).

O painel disponível comercialmente com maior densidade de SNPs para bovinos é o BovineHD BeadChip (Illumina), contendo cerca de 777 mil SNPs.

Estudos comparando o DL em taurinos e zebuínos, utilizando este painel, revelaram que mesmo em curtas distâncias (até 100kb), os animais zebuínos apresentam menor DL comparado aos animais taurinos e cruzados (PORTO-NETO; KIJAS; REVERTER, 2014). Esta diferença foi justificada devido às características das estruturas populacionais durante a formação das diversas raças. Dessa maneira, a partir de estudos realizados com zebuínos sugere-se a utilização de painéis de maiores densidades de SNPs para melhor desempenho ao utilizar abordagens genômicas, como estudos de GWAS e seleção genômica (O'BRIEN et al., 2014).

A imputação é amplamente estudada para inferir marcadores não definidos em animais que foram genotipados com baixa densidade, com a finalidade de reduzir custos da aplicação de seleção genômica (VANRADEN et al., 2011; HUANG et al., 2012) e pode ser utilizada também com a finalidade de obter genótipo de animais não genotipados (BERRY et al., 2014). Esta é também importante ferramenta para combinar dados genotipados com diferentes painéis e de diferentes densidades (SARGOLZAEI; CHESNAIS; SCHENKEL, 2014), podendo fornecer maior densidade de SNPs com alto DL para animais zebuínos. Esta abordagem colabora para melhor identificação dos blocos de haplótipos e maior acurácia nos resultados de metodologias que explorem o DL.

Para que dados imputados sejam utilizados em análises genômicas, é necessário que a imputação tenha sido implementada de forma que todos os animais tenham seus genótipos inferidos com elevada acurácia (BADKE et al., 2014). No entanto, a dificuldade na obtenção da acurácia de imputação devido à necessidade da presença do genótipo real para verificar a confiabilidade do genótipo imputado pode reduzir a abrangência da utilização destes genótipos. Dessa

maneira, os objetivos deste trabalho foram: (1) realizar estudos de imputação utilizando painéis de alta densidade (Illumina e Affymetrix) e painel customizado em uma população de bovinos da raça Nelore, estimar o desequilíbrio de ligação (DL) e a conformação de blocos de haplótipos para estes painéis de alta densidade individualmente e após a imputação; (2) estudar estratégias para predição da acurácia de imputação antes que a imputação seja efetuada, utilizando redes neurais artificiais e regressão linear múltipla e (3) estudar os segmentos de homozigose (ROH) em bovinos da raça Nelore, assim como identificar os genes presentes nos segmentos de homozigose mais frequentes na população.

2. REVISÃO DE LITERATURA

2.1. Desequilíbrio de ligação

O DL pode ser definido como a associação não aleatória entre alelos de dois ou mais *loci*, não necessariamente no mesmo cromossomo, dentro de uma mesma população. Dois cálculos estatísticos de desequilíbrio de ligação são os mais utilizados: o coeficiente de correlação entre alelos de dois *loci* (r^2) (HILL; ROBERTSON, 1968) e o valor padronizado de D, $|D'|$ (LEWONTIN, 1964). A primeira medida (r^2) varia de 0 a 1, em que 0 significa a ausência de correlação entre pares de alelos dos diferentes *loci*, ou seja, total equilíbrio de ligação e 1 significa correlação perfeita entre estes pares, ou seja, total desequilíbrio de ligação entre esse alelos. A verificação do valor de $|D'|$ menor que 1 indica que houve recombinação gênica entre os pares de alelos dos diferentes *loci* e, quando este é igual a 1, indica ausência de recombinação e portanto total desequilíbrio de ligação.

Alguns fatores podem afetar o DL, tais como: a estrutura da população, a migração, a seleção, a deriva genética, a mutação e principalmente a recombinação gênica (HEDRICK, 2010). Eventos históricos que promovem redução do tamanho efetivo da população tendem a aumentar a extensão de alelos com maior DL, enquanto que após rápida expansão populacional, eventos de recombinação gênica recorrentes tendem a reduzir o DL entre alelos mais distantes (REICH et al., 2001). O processo de seleção também promove alterações no DL em determinadas regiões do genoma, uma vez que há aumento na frequência dos alelos favoráveis à característica selecionada, levando aos alelos mais próximos a estes possuírem maior frequência, ampliando a correlação entre estes e consequentemente o DL (THE INTERNATIONAL HAPMAP CONSORTIUM, 2007).

Um dos primeiros estudos de DL em bovinos utilizou 284 microssatélites em uma população de bovinos de leite (FARNIR et al., 2000). Com a automação da genotipagem dos SNPs, estes passaram a ser aplicados em estudos de DL em diferentes populações. McKay et al. (2007), utilizando aproximadamente 2670 SNPs, revelaram que em distâncias de até 5 kb no genoma, animais zebuínos apresentaram níveis de desequilíbrio de ligação inferiores comparados aos taurinos. Os mesmos autores observaram que animais da raça Nelore apresentaram os menores valores de DL (r^2), em curtas distâncias, sendo o grande tamanho efetivo desta população uma das justificativas para este resultado observado na raça. Estudos em que os marcadores utilizados eram espaçados, de maneira geral, indicaram a utilização de painéis com maior densidade de marcadores para melhor detecção do DL (KHATKAR et al., 2007; LU et al., 2012).

Ao utilizarem painéis com cerca de 777 mil SNPs, Porto-Neto, Kijas e Reverter (2014) também observaram que os animais zebuínos apresentaram menor DL em curtas distâncias (até 100kb), quando comparados aos animais taurinos. Estes autores identificaram que a utilização de maior número de SNPs aumenta a cobertura sobre o genoma, sendo suficiente para identificar sinais de associações que não são observados com painéis de baixa densidade.

Espigolan et al. (2013), estudando DL em bovinos da raça Nelore com painéis de alta densidade, concluíram que as estimativas de DL para SNPs dentro de uma distância física de 30 kb ($r^2 \approx 0,20$) justificam o uso de painéis de alta densidade para a implementação de seleção genômica para esta raça, visto que a literatura reporta que valores médios de $r^2 \approx 0,20$ são considerados em suficiente DL para atingir acurácia de 0,85 ao obter estimativas de valores genômicos (MEUWISSEN; HAYES; GODDARD, 2001). No entanto, Neves et al. (2014) obtiveram média de r^2 igual a 0,29 para bovinos da raça Nelore e ao utilizar diferentes metodologias para predição genômica, observaram máximo de 0,74 de acurácia empírica para predição genômica. Dessa forma, o aumento na densidade dos marcadores pode auxiliar na identificação do DL, o que colabora com estudos de seleção e associação genômica na raça Nelore.

2.2. *Blocos de haplótipos*

Os blocos de haplótipos são regiões dos cromossomos que estão em alto DL (CHEN et al., 2013) e que são herdados em conjunto na próxima geração. Estes blocos geralmente representam regiões com baixa taxa de recombinação gênica, cercados por regiões com pontos de alta taxa de recombinação gênica (KHATKAR et al., 2007). Diversos algoritmos foram desenvolvidos para a identificação de blocos

de haplótipos (GABRIEL et al., 2002; ZHANG; JIN, 2003). Grande parte desses algoritmos utiliza critérios com base nos valores de DL, na diversidade de haplótipos e na localização de regiões conhecidas como pontos de alta taxa de recombinação gênica (KHATKAR et al., 2007).

A elucidação da estrutura de blocos de haplótipos pode trazer importantes considerações para estudos de associação e seleção genômica, devido a possibilidade de selecionar conjuntos de SNPs e diminuir a redundância causada pela informação de vários SNPs, resumindo essa informação em blocos de haplótipos (ZHANG et al., 2002). Segundo Cuyabano et al. (2015), a utilização de informação de blocos de haplótipos para seleção genômica possui a vantagem de que cada haplótipo pode apresentar um desequilíbrio de ligação maior com uma mutação causal do que quando utilizada a informação de SNPs individualmente. Além disso, os alelos dentro de blocos de haplótipos podem ser considerados para detectar maior variação dos efeitos.

Diferentes formas de definição de blocos de haplótipos têm sido utilizadas para predição genômica, sendo determinadas janelas com SNPs no mesmo centimorgan (BOICHARD et al., 2012); janelas com número fixo de SNPs (VILLUMSEN; JANSS; LUND, 2009) e informação de desequilíbrio de ligação (CUYABANO; SU; LUND, 2015), que pode ser medido por três medidas: r^2 , D e D', sendo esta última a mais utilizada pois esta depende menos da frequência alélica. A informação de blocos de haplótipos também podem ser utilizada para detecção de regiões genômicas sob seleção durante o processo evolutivo e identificação de recentes assinaturas de seleção (SABETI et al., 2002).

Villa-Angulo et al. (2009), estudando o genoma de bovinos de diferentes raças, relataram que, na faixa de 1 a 100 kb de extensão, a estrutura de blocos de haplótipos possui similaridade à de humanos. Khatkar et al. (2006), estudando o cromossomo 6 de bovinos, identificaram que em regiões mais distantes, o DL em bovinos manteve maiores magnitudes quando comparado ao DL em humanos, devido às diferenças no tamanho efetivo das populações.

Similaridades na estrutura de blocos de haplótipos em bovinos de leite e de corte taurinos foram relatadas por Villa-Angulo et al. (2009). Neste estudo, ao avaliar bovinos de leite taurinos e animais zebuínos, os autores detectaram elevado grau de dissimilaridade na estrutura de blocos de haplótipos destes animais, concluindo que estas diferenças podem ser devido ao histórico ancestral desses subgrupos. Outros estudos que identificaram blocos de haplótipos utilizando painéis baixa densidade de SNPs em bovinos revelaram que os blocos cobriam pequenas regiões do genoma, sugerindo a utilização de maior densidade de SNPs para a maior cobertura do genoma, facilitando a identificação dos blocos e determinação dos SNPs mais representativos para características de importância econômica (KHATKAR et al., 2007; VILLA-ANGULO et al., 2009).

2.3. *Imputação*

A aplicação da seleção genômica com a utilização de todos os animais genotipados com painéis de alta densidade exige elevado custo. Dessa maneira, genotipar animais jovens com painéis de baixa densidade, que é de menor custo e então realizar a imputação dos marcadores não conhecidos utilizando a informação de uma população referência genotipada com painéis de alta densidade tornou-se

uma alternativa para redução de custos para a aplicação desta tecnologia (HAYES et al., 2011).

A imputação também pode ser utilizada para predizer os genótipos faltantes e melhorar a taxa de leitura de genótipo dos animais (MARCHINI et al., 2007). Outra abordagem da imputação é a combinação de dados genotipados com diferentes painéis, o que pode contribuir para a melhoria dos resultados nos estudos genômicos (SARGOLZAEI; CHESNAIS; SCHENKEL, 2014).

Diversos programas computacionais foram desenvolvidos para a imputação (SARGOLZAEI; CHESNAIS; SCHENKEL, 2014; BROWNING; BROWNING, 2013). A maioria destes realiza a imputação utilizando informação de genealogia ou informação populacional, sendo que esta última considera que os indivíduos não são aparentados e determina os haplótipos utilizando o desequilíbrio de ligação, existindo também aqueles programas computacionais que utilizam as duas abordagens.

Após a imputação é recomendada a verificação da eficiência desta imputação, a qual é medida pela acurácia. A acurácia de imputação pode ser influenciada pela composição da população referência, em que o maior número de animais e a relação próxima de parentesco entre estes e os animais da população a ser imputada, melhoram a acurácia de imputação (KHATKAR et al., 2012).

Ventura et al. (2014), ao estudarem o impacto da população referência na imputação de raças puras e cruzadas, identificaram melhores acurácias quando o grupo racial dos animais que precisavam ser imputados estava bem representado na população referência. A frequência alélica também pode afetar a acurácia de imputação, uma vez que para obter a informação de fase de DL de um alelo, este

precisa estar em maior frequência em sua estrutura de haplótipo, sendo dessa maneira mais difícil a imputação de alelos raros (BROWNING; BROWNING, 2011). Outro fator que influencia a acurácia de imputação é a densidade de marcadores. Os painéis com alta densidade de marcadores permitem a identificação de mais níveis de DL, constatando maior DL entre os marcadores, principalmente para zebuínos (O'BRIEN et al., 2014), podendo reduzir os erros de imputação e elevar a acurácia desta prática.

Diversos estudos de imputação têm sido realizados com o objetivo de redução de custos para a aplicação de seleção genômica, sendo observadas altas acurácias de imputação (HUANG et al., 2012; CHUD et al., 2015). No entanto, ainda há poucos estudos de imputação em bovinos combinando dados genotipados com diferentes painéis de diferentes densidades, com o objetivo de ampliar a densidade dos marcadores e melhorar a acurácia das demais aplicações genômicas, principalmente em animais zebuínos.

Dois painéis de alta densidade de SNPs estão disponíveis comercialmente para bovinos: o BovineHD BeadChip (Illumina) e o Axion Genome-Wide BOS 1 Array Plate (Affymetrix). O painel da Illumina possui aproximadamente 777 mil SNPs, distribuídos homogeneamente ao longo do genoma com distância média de 3,4 Kb entre SNPs adjacentes. O painel da Affymetrix possui aproximadamente 640 mil SNPs, que foram selecionados para minimizar a redundância de cobertura de grandes grupos de SNPs em DL e maximizar a cobertura de pequenos grupos de SNPs em DL.

Ventura et al. (2013) realizaram a imputação de dois painéis de alta densidade (Illumina + Affymetrix) em animais taurinos e observaram acurácia igual a 98,03%,

resultando em um painel combinado de marcadores contendo 1.261.128 SNPs. Estes mesmos autores relataram que é possível de se obter a acurácia de imputação relativamente alta entre dois painéis de alta densidade, mesmo com uma pequena população referência (80 animais).

Em um estudo de desempenho entre os dois painéis de alta densidade em bovinos das raças Holandesa e Jersey (RINCON et al., 2011) foi identificado que, embora o painel da Affymetrix apresente número menor de SNPs, 19% a mais de SNPs restaram após a remoção de SNPs redundantes calculados pelo valor de DL (*prunning*), resultando em menor distância física entre os SNPs restantes (5,2 vs. 6,9 kb), em relação ao painel da Illumina. No mesmo estudo, também foi constatado que o uso combinado de ambos os painéis aumentou a cobertura do genoma, se comparada ao uso de cada painel de genotipagem separadamente, pois diminuiu a distância física entre os SNPs adjacentes, o que pode ser importante para demais estudos genômicos de raças puras e cruzadas. Isso também pode ser de interesse para a pecuária brasileira, visto que grande parte do rebanho comercial nacional possui alguma forma de cruzamento com animais da raça Nelore.

Outra forma de aumentar a densidade de SNPs e obter maior quantidade de informação genômica dos animais, além da utilização de painéis de alta densidade e a combinação dos mesmos, é realizar o sequenciamento de alguns animais e a imputação dos demais para toda a sequência genômica. No entanto, torna-se difícil identificar a eficiência da imputação quando não se tem a informação dos SNPs obtidos pela genotipagem de alta densidade ou do sequenciamento completo dos animais imputados, uma vez que tanto a genotipagem em alta densidade como o sequenciamento ainda possui elevado custo para ser aplicado em grande número de

animais. Desta forma, tornou-se necessário estudar a estimativa do valor da acurácia de imputação, para que genótipos imputados possam ser usados em demais análises genômicas com a garantia da qualidade da imputação. O programa computacional minimac3 (DAS et al., 2016) fornece uma opção de estimativa de acurácia por SNP, o que torna possível observar e monitorar regiões de baixa acurácia de imputação.

Van Binsbergen et al. (2014) propuseram uma forma de prever a confiabilidade da imputação para os SNPs utilizando informações de desequilíbrio de ligação, frequência alélica do alelo de menor frequência (MAF – “minor allele frequency”) e o número de indivíduos na população referência. Estes autores concluíram que funções que estimam o desequilíbrio de ligação com base apenas em distância física ou na diferença de MAF entre o SNP imputado e o SNP mais próximo presente no painel de baixa densidade não fornecem adequada confiabilidade de imputação. No entanto, quando estas são consideradas juntas, a predição da confiabilidade de imputação aumenta. Dessa maneira, estudos comparando diferentes alternativas na predição da acurácia de imputação para SNPs e para os animais podem ser interessantes para obter uma estimativa mais próxima da acurácia real, podendo esta ser utilizada para apoiar o uso de genótipos imputados em demais estudos que abordam informação genômica.

2.4. Redes Neurais Artificiais

As redes neurais artificiais são utilizadas em estudos de modelagem de dados, predição, classificação, identificações de padrões, controle de processos, otimização e suporte à decisão em diversas áreas do conhecimento (SARLE, 1994). A predição por redes neurais artificiais tem sido explorada por identificar relações

não-lineares entre os preditores e variáveis resposta, apresentando a característica de aprendizagem de máquina, em que as informações podem ser inferidas a partir dos dados previamente inseridos, por meio do potencial da rede em detectar interações complexas entre variáveis predictoras, ou seja, esta explora a construção de algoritmos que podem aprender com seus erros e fazer previsões sobre dados. Essa característica faz com que as redes neurais não necessitem das pressuposições de independência e normalidade das variáveis exigidas nas relações lineares. Além disso, não é preciso pré-especificar a função matemática a ser utilizada para modelar o conjunto de dados (TAM, 1991; GAUDART et al., 2004; PAO, 2008).

De maneira geral, as diferentes estruturas de redes neurais artificiais vêm sendo utilizadas para obter informações na agropecuária com resultados satisfatórios, como os observados para a predição de fenótipos de características complexas para bovinos de leite e cultivares de trigo (GIANOLA et al., 2011), predição de diferença esperada de progênie em score de marmoreio (OKUT et al., 2013) e para predição de valores genéticos genômicos em uma linhagem de milho (GONZÁLEZ-CAMACHO et al.; 2012).

Dentre as diferentes estruturas de redes neurais artificiais existentes, a “multilayer perceptron” (MLP) é composta por camadas alinhadas de neurônios, em que os neurônios de uma camada são interligados com os da camada adjacente (CHENG; TITTERINGTON, 1994; PALIWAL; KUMAR, 2009). As MLP são redes neurais com aprendizado supervisionado, ou seja, a extração do conhecimento do banco de dados e o mapeamento das variáveis de entrada com as de saída são feitos por meio do algoritmo “backpropagation”, que indica para a MLP o quão perto

ou distante as predições estão dos valores reais. Isso permite que a MLP possa melhorá-las, por meio de mudanças nos pesos sinápticos (WARNER; MISRA, 1996). O desenvolvimento matemático do algoritmo “backpropagation” foi apresentado por Rumelhart et al. (1986). Essa característica de aprendizado supervisionado permite que a rede MLP seja flexível com relação aos dados utilizados, característica fundamental na análise de dados no campo da agropecuária.

2.5. Segmentos de homozigose (ROH)

O estudo de parâmetros populacionais possui relevância devido à informação que este fornece sobre a população estudada, o que permite entender a plasticidade das populações e o monitoramento do desempenho de características de interesse econômico em raças comercialmente importantes (CURIK; FERENČAKOVIĆ; SÖLKNER, 2014). A análise da estrutura da população utilizando dados de pedigree tem sido realizada principalmente com a finalidade de identificar aumento na endogamia da população (SANTANA JR et al., 2010).

A endogamia modifica as frequências genotípicas, aumentando a homozigose e, conseqüentemente, reduzindo a heterozigose, o que implica na redução da variabilidade genética da população. Elevadas taxas de endogamia podem reduzir a viabilidade e a fertilidade dos animais, além de reduzir a média de desempenho fenotípico de algumas características economicamente importantes, definida como depressão endogâmica (FALCONER; MACKAY, 1996).

Dessa maneira, o monitoramento deste parâmetro populacional torna-se importante na condução de programas de melhoramento genético. No entanto, a ausência de registros de pedigree é frequente e esta prejudica a avaliação dos parâmetros populacionais que, muitas vezes, podem ser super ou subestimados.

Para solucionar este problema, metodologias que utilizam informações de marcadores moleculares têm sido utilizadas (FERENČAKOVIĆ et al., 2013).

Os segmentos de homozigose (ROH) identificam regiões longas de segmentos cromossômicos em homozigose que estão presentes nos indivíduos devido à transmissão de haplótipos idênticos de pais para a progênie. Esta informação tem sido utilizada em estudos relacionados a doenças (GHANI et al., 2013), recombinação gênica (BOSSE et al. 2012), histórico e endogamia populacional (PURFIELD et al., 2012).

Bjelland et al. (2013) obtiveram diferentes medidas de endogamia utilizando informações de SNPs e identificaram alta correlação entre estas medidas. Keller, Visscher e Goddard (2011), estudaram estimativas de endogamia com base no uso individual de SNP, de ROH e de pedigree. Estes autores, ao correlacionar as estimativas de endogamia obtidas pelas informações genômicas e a obtida pelo pedigree, identificaram que a endogamia estimada utilizando ROH apresentou maior magnitude, porque a endogamia calculada por ROH tende a detectar endogamia mais recente quando comparada às demais estimativas genômicas.

Estudos que utilizam a metodologia de ROH variam na definição de limites mínimos para um segmento de homozigose, os quais podem ser considerados pela quantidade de SNPs homozigotos consecutivos e/ou uma determinada distância. Ferenčaković et al. (2011) consideraram diferentes tamanhos mínimos de ROH de 1, 2, 4, 8 e 16 Mb com pelo menos 15 SNPs homozigotos consecutivos. Bjelland et al. (2013) estudaram ROH em bovinos de leite e utilizaram 30 SNPs como tamanho mínimo para ser considerado um segmento em homozigose. Howrigan, Simonson e Keller (2011) estudaram limites mínimos com relação a quantidades de SNPs e, ao

estabelecer diferentes quantidades de SNPs para ser considerado um ROH, os autores concluíram que o tamanho mínimo depende do período que iniciou a autozigose. Assim, valores de limites mínimos de alta magnitude detectam autozigose recente e valores de limites mínimos de baixa magnitude detectam autozigoses mais antigas. De maneira geral, segmentos muito longos de ROH representam endogamia recente e segmentos curtos de ROH representam endogamia antiga.

Ferenčaković et al. (2013) estudaram quatro raças bovinas da Áustria, utilizando informações de ROH e identificaram que a raça Brown Swiss foi a que apresentou maior endogamia. Os autores observaram que esta raça apresentava poucos segmentos de ROH, porém de grande extensão e relacionaram este fato com o histórico populacional de importação de pequeno número de animais da raça. Em um estudo de ROH com diversas raças, zebuínos africanos apresentaram menor quantidade de ROH quando comparados a zebuínos da América e Madagascar, nos quais foram observados segmentos curtos, indicando que estas raças foram estabelecidas por uma população pequena de fundadores, ou seja, apresentaram endogamia antiga e não foram afetadas por endogamia recente (PURFIELD et al., 2012). Zavarez et al. (2015) estudaram uma população de bovinos da raça Nelore e observaram médias baixas de autozigose (4,79%) e ausência de uniformidade nos ROH ao longo do genoma destes animais, indicando que esta era uma população não endogâmica.

3. REFERÊNCIAS BIBLIOGRÁFICAS

BADKE, Y. M.; BATES, R. O.; ERNST, C. W.; FIX, J.; STEIBEL, J. P. Accuracy of estimation of genomic breeding values in pigs using low-density genotypes and imputation. **G3**, Toronto, v.4, p. 623-631, 2014.

BERRY, D. P.; MCPARLAND, S.; KEARNEY, J. F.; SARGOLZAEI, M.; MULLEN, M. P. Imputation of ungenotyped parental genotypes in dairy and beef cattle from progeny genotypes. **Animal**, Cambridge, v. 8, n. 6, p. 895-903, 2014.

BJELLAND, D. W.; WEIGEL, K. A.; VUKASINOVIC, N.; NKRUMAH, J. D. Evaluation of inbreeding depression in Holstein cattle using whole-genome SNP markers and alternative measures of genomic inbreeding. **Journal of Dairy Science**, Champaign, v. 96, p. 4697-4706, 2013.

CHEN, W.; HUNG, C.; LIN, Y. Efficient haplotype block partitioning and tag SNP selection algorithms under various constraints. **BioMed research international**, London, v. 2013, 2013

DAS, S.; FORER, L.; SCHÖNHERR, S.; SIDORE, C.; LOCKE, A. E.; KWONG, A.; SCHLESSINGER, D. Next-generation genotype imputation service and methods. **Nature genetics**, v. 48, n. 10, p. 1284, 2016.

BOICHARD, D.; GUILLAUME, F.; BAUR, A.; CROISEAU, P.; ROSSIGNOL, M. N.; BOSCHER, M. Y.; DRUET, T.; GENESTOUT, L.; COLLEAU, J. J.; JOURNAUX, L.; DUCROCQ, V.; FRITZ, S. Genomic selection in French dairy cattle. **Animal Production Science**, Clayton, v. 52, n. 3, p. 115-120, 2012.

BOSSE, M.; MEGENS, H.; MADSEN, O.; PAUDEL, Y.; FRANTZ, L. A. F.; SCHOOK, L. B.; CROOIJMANS, R. P. M. A.; GROENEN, M. A. M. Regions of Homozygosity in the Porcine Genome: Consequence of Demography and the Recombination Landscape. **Plos Genetics**, Cambridge, v. 8, n. 11, 2012.

BROWNING, B.L.; BROWNING, S.R. A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. **The American Journal of Human Genetics**, Houston, v. 84, p. 210-223, 2009.

BROWNING, B.L.; BROWNING, S.R. Improving the accuracy and efficiency of identity-by-descent detection in population data. **Genetics**, New York, v. 194, n. 2, p. 459–471, 2013.

BROWNING, S. R.; BROWNING, B. L. Haplotype phasing: existing methods and new developments. **Nature Reviews Genetics**, London, v. 12, p. 703-714, 2011.

BUZANSKAS, M.E.; GROSSI, D.A.; VENTURA, R.V.; SCHENKEL, F.S.; SARGOLZAEI, M.; MEIRELLES, S.L.C.; MOKRY, F.B.; HIGA, R.H.; MUDADU, M.A.; DA SILVA, M.V.G.B.; NICIURA, S.C.M.; JÚNIOR, R.A.A.T.; ALENCAR, M.M.; REGITANO, L.C.A.; MUNARI, D.P. Genome-Wide Association for Growth Traits in Canchim Beef Cattle. **PLoS One**, São Francisco, v. 9, n. 4, 2014.

CHENG, B.; TITTERINGTON, D. M. Neural networks: a review from a statistical perspective. **Statistical Science**, Piscataway, v. 9, p. 2-54, 1994.

CHUD, T. C. S.; VENTURA, R. V.; SCHENKEL, F. S.; CARVALHEIRO, R.; BUZANSKAS, M. E.; ROSA, J. O.; MUDADU, M. A.; SILVA, M. V. G. B.; MOKRY, F. B.; MARCONDES, C. R.; REGITANO, L. C. A.; MUNARI, D. P. Strategies for genotype imputation in composite beef cattle. **BMC Genetics**, London, v. 16, n. 99, 2015.

CURIK, I.; FERENČAKOVIĆ, M.; SÖLKNER, J. Inbreeding and runs of homozygosity: A possible solution to an old problem. **Livestock Science**, Oxford, v. 166, p. 26-34, 2014.

CUYABANO, B. C.; SU, G.; LUND, M. Selection of haplotype variables from a high-density marker map for genomic prediction. *Genetics Selection Evolution*, v. 47, n. 1, p. 61, 2015.

CUYABANO, B. C. D.; SU, G.; ROSA, G. J. M.; LUND, M. S.; GIANOLA, D. Bootstrap study of genome-enabled prediction reliabilities using haplotype blocks across Nordic Red cattle breeds. **Journal of dairy science**, Champaign, v. 98, n. 10, p. 7351-7363, 2015.

DAS, S.; FORER, L.; SCHÖNHERR, S.; SIDORE, C.; LOCKE, A. E.; KWONG, A.; SCHLESSINGER, D. Next-generation genotype imputation service and methods. **Nature Genetics**, New York, v. 48, n. 10, p. 1284-1287, 2016.

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

FALCONER, D. S.; MACKAY, T. F. C. Introduction to quantitative genetics. 4. ed. Harlow: Longman House, 1996. p. 245 – 253

FARNIR, F.; COPPIETERS, W.; ARRANZ, J.; BERZI, P.; CAMBISANO, N.; GRISART, B.; KARIM, L.; MARCQ, F.; MOREAU, L.; MNI, M.; NEZER, C.; SIMON, P.; VANMANSHOVEN, P.; WAGENAAR, D.; GEORGES, M. Extensive Genome-wide Linkage Disequilibrium in Cattle. **Genome Research**, San Francisco, v. 10, n. 2, p. 220–227, 2000.

FERENČAKOVIĆ, M.; HAMZIĆ, E.; GREDLER, B.; CURIK, I.; SÖLKNER, J. Runs of Homozygosity Reveal Genomewide Autozygosity in the Austrian Fleckvieh Cattle. **Agriculturae Conspectus Scientificus**, Zagreb, v. 76, n. 4, p. 325-328, 2011.

FERENČAKOVIĆ, M.; HAMZIĆ, E.; GREDLER, B.; SOLBERG, T. R.; KLEMETSDAL, G.; CURIK, I.; SÖLKNER, J. Estimates of autozygosity derived from runs of homozygosity: empirical evidence from selected cattle populations. **Journal of Animal Breeding and Genetics**, Malden, v.130, p. 286-293, 2013.

GABRIEL, S.B.; SCHAFFNER, S.F.; NGUYEN, H.; MOORE, J.M.; ROY, J.; BLUMENSTIEL, B.; HIGGINS, J.; DEFELICE, M.; LOCHNER, A.; FAGGART, M.; LIU-CORDERO, S.N.; ROTIMI, C.; ADEYEMO, A.; COOPER, R.; WARD, R.; LANDER, E.S.; DALY, M.J.; ALTSHULER, D. The structure of haplotype blocks in the human genome. **Science**, Washington, v. 296, n. 5576, p. 2225–2229, 2002.

- GAUDART, J.; GIUSIANO, B.; HUIART, L. Comparison of the performance of multi-layer perceptron and linear regression for epidemiological data. **Computational Statistics & Data Analysis**, California, v. 44, p. 547-570, 2004.
- GHANI, M.; SATO, C.; LEE, J. H.; REITZ, C.; MORENO, D.; MAYEUX, R.; ST GEORGE-HYSLOP, P.; ROGAEVA, E. Evidence of Recessive Alzheimer Disease Loci in a Caribbean Hispanic Data Set: Genome-wide Survey of Runs of Homozygosity. **JAMA Neurology**, Dallas, v. 70, n. 10, p. 1261-1267, 2013.
- GIANOLA, D.; OKUT, H.; WEIGEL, K. A.; ROSA, G. J. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. **BMC Genetics**, London, v. 12, n. 87, 2011.
- GONZÁLEZ-CAMACHO, J. M.; DE LOS CAMPOS, G.; PÉREZ, P.; GIANOLA, D.; CAIRNS, J. E.; MAHUKU, G.; CROSSA, J. Genome-enabled prediction of genetic values using radial basis function neural networks. **Theoretical and Applied Genetics**, Heidelberg, v. 125, n. 4, p. 759-771, 2012.
- HAYES, B. J.; BOWMAN, P. J.; DAETWYLER, H. D.; KIJAS, J. W.; VAN DER WERF, J. H. J. Accuracy of genotype imputation in sheep breeds. **Animal Genetics**, Malden, v. 43, p. 72-80, 2011.
- HEDRICK, P.W. **Genetics of Populations**. Jones and Bartlett Publishers: Sudbury, 2010. p. 523-598.
- HILL, W.; ROBERTSON, A. Linkage disequilibrium in finite populations. **Theoretical and Applied Genetics**, Heidelberg, v. 38, n. 6, p. 226 - 231, 1968.
- HOWRIGAN, D. P.; SIMONSON, M. A.; KELLER, M. C. Detecting autozygosity through runs of homozygosity: A comparison of three autozygosity detection algorithms. **BMC Genomics**, London, v. 12, n. 460, 2011.
- HUANG, Y.; HICKEY, J. M.; CLEVELAND, M. A.; MALTECCA, C. Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. **Genetics Selection Evolution**, London, v. 44, n. 25, 2012.
- KELLER, M. C.; VISSCHER, P. M.; GODDARD, M. E. Quantification of Inbreeding Due to Distant Ancestors and Its Detection Using Dense Single Nucleotide Polymorphism Data. **Genetics**, Bethesda, v. 189, p. 237-249, 2011.
- KHATKAR, M. S.; COLLINS, A.; CAVANAGH, J. A. L.; HAWKEN, R. J.; HOBBS, M.; ZENGER, K. R.; BARRIS, W.; MCCLINTOCK, A. E.; THOMSON, P. C.; NICHOLAS, F. W.; RAADSMA, H. W. A First-Generation Metric Linkage Disequilibrium Map of Bovine Chromosome 6. **Genetics**, Bethesda, v. 174, p. 79 - 85, 2006.
- KHATKAR, M.S.; ZENGER, K.R.; HOBBS, M.; HAWKEN, R.J.; CAVANAGH, J.A.L.; BARRIS, W.; MCCLINTOCK, A.E.; MCCLINTOCK, S.; THOMSON, P.C.; TIER, B.; NICHOLAS, F.W.; RAADSMA, H.W. A primary assembly of a bovine haplotype block map based on a 15,036-single-nucleotide polymorphism panel genotyped in holstein-friesian cattle. **Genetics**, Bethesda, v. 176, n. 2, p. 763 - 72, 2007.
- KHATKAR, M. S.; MOSER, G.; HAYES, B. J.; RAADSMA, H. W. Strategies and utility of imputed SNP genotypes for genomic analysis in dairy cattle. **BMC Genomics**, London, v. 13, n. 538, 2012.

LEWONTIN, R.C. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. **Genetics**, Bethesda, v. 49, n. 1, p. 49 - 67, 1964.

LU, D.; SARGOLZAEI, M.; KELLY, M.; LI, C.; VOORT, G.V.; WANG, Z.; PLASTOW, G.; MOORE, S.; MILLER, S.P. Linkage disequilibrium in Angus, Charolais, and Crossbred beef cattle. **Frontiers in Genetics**, Lausanne, v. 3, n. 152, 2012.

MARCHINI, J.; HOWIE, B.; MYERS, S.; MCVEAN, G.; DONNELLY, P. A new multipoint method for genome-wide association studies by imputation of genotypes. **Nature Genetics**, New York, v. 39, p. 906-913, 2007.

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

MEUWISSEN, T.H.; HAYES, B.J.; GODDARD, M.E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, Bethesda, v. 157, n. 4, p. 1819 - 1829, 2001.

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.; VAN TASSELL, C. P.; 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**, London, v. 46, n. 17, 2014.

O'BRIEN, A. M. P.; MÉSZÁROS, G.; UTSUNOMIYA, Y. T.; SONSTEGARD, T. S.; GARCIA, J. F.; TASSEL, C. P. V.; CARVALHEIRO, R.; SILVA, M. V. B.; SOLKNER, J. Linkage disequilibrium levels in *Bos indicus* and *Bos taurus* cattle using medium and high density SNP chip data and different minor allele frequency distributions. **Livestock Science**, Oxford, v. 166, p. 121 -132, 2014.

OKUT, H.; WU, X. L.; ROSA, G. J.; BAUCK, S.; WOODWARD, B. W.; SCHNABEL, R. D.; GIANOLA, D. Predicting expected progeny difference for marbling score in Angus cattle using artificial neural networks and Bayesian regression models. **Genetics Selection Evolution**, London, v. 45, n. 34, 2013.

PALIWAL, M.; KUMAR, U. A. Neural networks and statistical techniques: a review of applications. **Expert Systems with Applications**, Amsterdam, v. 36, p. 2-17, 2009.

PAO, H. T. A comparison of neural networks and multiple regression analysis in modeling capital structure. **Expert Systems with Applications**, Amsterdam, v. 35, p. 720-727, 2008.

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**, London, v. 46, p. 22 - 26, 2014.

PURFIELD, D. C.; BERRY, D. P.; MCPARLAND, S.; BRADLEY, D. G. Runs of homozygosity and population history in cattle. **BMC Genetics**, London, v. 13, n. 70, 2012.

REICH, D. E.; CARGILL, M.; BOLK, S.; IRELAND, J.; SABETI, P. C.; RICHTER, D. J.; LAVERY, T.; KOUYOUMJIAN, R.; FARHADIAN, S. F.; WARD, R.; LANDER, E. S.

Linkage disequilibrium in the human genome. **Nature**, New York, v. 411, p. 199 - 204, 2001.

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

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning Representations by Back-Propagating Errors. **Nature**, New York, v. 323, n. 6088, p. 533-536, 1986.

SABETI, P.C.; REICH, D.E.; HIGGINS, J.M.; LEVINE, H.Z.P.; RICHTER, D.J.; SCHAFFNER, S.F.; GABRIEL, S.B.; PLATKO, J.V.; PATTERSON, N.J.; MCDONALD, G.J.; ACKERMAN, H.C.; CAMPBELL, S.J.; ALTSHULER, D.; COOPER, R.; KWIATKOWSKI, D.; WARD, R.; LANDER, E.S. Detecting recent positive selection in the human genome from haplotype structure. **Nature**, New York, v. 419, n. 6909, p. 832–837, 2002.

SANTANA JR., M. L.; OLIVEIRA, P. S.; PEDROSA, V. B.; ELER, J. P.; GROENEVELD, E.; FERRAZ, J. B. S. Effect of inbreeding on growth and reproductive traits of Nellore cattle in Brazil. **Livestock Science**, Oxford, v. 131, p. 212-217, 2010.

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

SARLE, W. S. Neural networks and statistical models. In: Proceedings of the Nineteenth Annual SAS Users Group International Conference. 1994, Cary, NC. Anais... Cary, NC.: SAS Institute Inc., 1994, 1538-1550.

TAM, K. Y. Neural network models and the prediction of bank bankruptcy. **Omega**, Philadelphia, v. 19, n. 5, p. 429-445, 1991.

THE INTERNATIONAL HAPMAP CONSORTIUM A second generation human haplotype map of over 3.1 million SNPs. **Nature**, New York, v.449, p. 851–861, 2007.

VAN BINSBERGEN, R.; BINK, M. C.; CALUS, M. P.; VAN EEUWIJK, F. A.; HAYES, B. J.; HULSEGGE, I.; VEERKAMP, R. F. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. **Genetics Selection Evolution**, London, v. 46, n. 41, 2014.

VANRADEN, P. M.; O'CONNELL, J. R.; WIGGANS, G. R.; WEIGEL, K. A. Genomic evaluations with many more genotypes. **Genetics Selection Evolution**, London, v. 43, p. 10 – 20, 2011.

VENTURA, R.; SCHENKEL, F.; SARGOLZAEI, M.; MILLER, S. P. Accuracy of Imputation to High Density SNP Data in Multibreed Beef Cattle. In: International Plant & Animal Genome XXI Conference, 21., 2013, San Diego. Disponível em: < <https://pag.confex.com/pag/xxi/webprogram/Paper7920.html>>. Acesso em: 23 de fevereiro de 2013.

VENTURA, R. V.; LU, D.; SCHENKEL, F. S.; WANG, Z.; LI, C.; MILLER, S. P. Impact of reference population on accuracy of imputation from 6K to 50K single nucleotide

polymorphism chips in purebred and crossbreed beef cattle. **Journal of Animal Science**, Champaign, v. 92, p. 1433 - 1444, 2014.

VILLA-ANGULO, R.; MATUKUMALLI, L. K.; GILL, C. A.; CHOI, J.; TASSELL, C. P. V.; GREFENSTETTE, J. J. High-resolution haplotype block structure in the cattle genome. **BMC Genetics**, London, v. 10, p. 19 - 31, 2009.

VILLUMSEN, T. M.; JANSS, L.; LUND, M. S. The importance of haplotype length and heritability using genomic selection in dairy cattle. **Journal of Animal Breeding and Genetics**, Malden, v. 126, p.3-13, 2009.

WARNER, B.; MISRA, M. Understanding neural networks as statistical tools. **The American Statistician**, New York, v. 50, n. 4, p. 284-293, 1996.

ZAVAREZ, L. B.; UTSUNOMIYA, Y. T.; CARMO, A. S.; NEVES, H. H. R.; CARVALHEIRO, R.; FERENCAKOVIC, M.; O'BRIEN, A. M. P.; CURIK, I.; COLE, J. B.; VAN TASSELL, C. P.; SILVA, M. V. G. B.; SONSTEGARD, T. S.; SÖLKNER, J.; GARCIA, J. F. Assessment of autozygosity in Nellore cows (*Bos indicus*) through high-density SNP genotypes. **Frontiers in Genetics**, Lausanne, v. 6, n. 5, 2015.

ZHANG, K.; DENG, M.; CHEN, T.; WATERMAN, M. S.; SUN, F. A dynamic programming algorithm for haplotype block partitioning. **PNAS**, Washington, v. 99, n. 11, p. 7335 - 7339, 2002.

ZHANG, K.; JIN, L. HaploBlockFinder: haplotype block analyses. **Bioinformatics**, Oxford, v. 19, n. 10, p. 1300 - 1301, 2003.

CAPÍTULO 2 - Estudo de imputação em bovinos da raça Nelore utilizando a combinação de painéis de alta densidade

RESUMO – A imputação pode ser utilizada para combinar informações genômicas de diferentes painéis, o que pode fornecer maior densidade de marcadores e favorecer a identificação de maior desequilíbrio de ligação. Dessa forma, este estudo teve como objetivo avaliar a imputação de painéis de baixa densidade comercial e customizados para painéis de alta densidade (Illumina e Affymetrix), assim como para um painel combinado (Illumina + Affymetrix - PC) para bovinos da raça Nelore, e estudar o desequilíbrio de ligação (DL) e conformação de blocos de haplótipos antes e após a imputação. Os 814 animais utilizados foram genotipados com BovineHD BeadChip (IllumHD), em que 93 animais (23 touros e 70 progênie) também foram genotipados com Axion Genome-Wide BOS 1 Array Plate (AffyHD). As imputações foram realizadas com 23 touros compondo a população referência. De maneira geral, a customização considerando DL e MAF (frequência do alelo de menor frequência) apresentaram as maiores acurácias. O painel IllumHD apresentou maiores estimativas de DL comparado ao AffyHD e PC para curtas distâncias entre os marcadores SNP. O PC apresentou maior quantidade de bloco de haplótipos com pequenos tamanhos. O uso combinado de painéis é recomendado por aumentar a densidade e número de blocos de haplótipos, aumentando a probabilidade de obter um marcador próximo a um QTL de interesse. Deve-se considerar SNPs em comum para o painel IllumHD e AffyHD para customizar um painel de menor densidade, pois permite elevar a acurácia de imputação para o painel IllumHD, AffyHD e PC.

Palavras-chave: Affymetrix, Illumina, painéis comerciais, painel combinado, painéis customizados

1. INTRODUÇÃO

A seleção genômica já tem sido aplicada para bovinocultura de leite. No entanto, sua aplicação na bovinocultura de corte tem sido mais lenta devido aos desafios de se obter uma população referência grande o suficiente a qual permita alcançar níveis desejáveis de acurácia de valores genéticos (HAYES et al., 2013). Além do aumento no número de animais genotipados, o maior número de marcadores moleculares pode elevar a acurácia nos valores genéticos (MEUWISSEN, 2009), uma vez que o maior número de marcadores aumenta a possibilidade de identificar *loci* de características quantitativas (QTL) em elevado desequilíbrio de ligação com um marcador.

Dois painéis comerciais de alta densidade são utilizados para bovinos (RINCON et al., 2011): o BovineHD BeadChip (Illumina) e o Axion Genome-Wide BOS 1 Array Plate (Affymetrix). O painel Illumina contém, aproximadamente, 777 mil SNPs distribuídos homogeneamente pelo genoma. O painel Affymetrix tem cerca de 640 mil SNPs, os quais foram selecionados para diminuir a possível redundância na cobertura de SNPs que estão em alto desequilíbrio de ligação. Devido ao alto custo para genotipar muitos animais com estes painéis, o uso de genótipos imputados em painéis de baixa densidade para os de alta densidade tem sido estudado na seleção genômica (WEIGEL et al., 2010; BERRY; KEARNEY, 2011).

Existem muitos estudos de imputação de painéis de baixa densidade para alta densidade utilizando o painel Illumina (HOZÉ et al., 2013; BERRY et al., 2014; CARVALHEIRO et al., 2014; CHUD et al., 2015). Entretanto, existem poucos estudos de imputação considerando alta densidade do painel Affymetrix para bovinos (VANRADEN et al., 2013; SCHENKEL, 2014). Genótipos imputados são

importantes para estudos de associação com cobertura ampla do genoma (GWAS) e seleção genômica, quando apresentam elevada acurácia de imputação (BADKE et al., 2014). Assim, a obtenção de um painel de baixa densidade que apresente elevada acurácia de imputação para ambos os painéis de alta densidade pode trazer benefícios para a aplicação da seleção genômica.

Além da utilização da imputação para inferir marcadores não definidos em animais que foram genotipados com painéis de baixa densidade, com a finalidade de reduzir custos da aplicação de seleção genômica (VANRADEN et al., 2011; HUANG et al., 2012), esta também pode ser utilizada para combinar dados de genótipos de diferentes painéis (SARGOLZAEI et al., 2014). Isso pode fornecer maior densidade de marcadores e elevar o nível do desequilíbrio de ligação entre os SNPs. Essa abordagem pode contribuir para melhor identificação de blocos de haplótipos e favorecer metodologias que utilizam informação de desequilíbrio de ligação. Portanto, este estudo teve como objetivo avaliar a imputação de painéis de baixa densidade comercial e customizados para painéis de alta densidade e um painel combinado formado por painéis de alta densidade (Illumina e Affymetrix) em bovinos da raça Nelore, assim como estimar o desequilíbrio de ligação (DL) e a conformação de blocos de haplótipos para painéis de alta densidade individualmente e após a imputação.

2. MATERIAL E MÉTODOS

2.1. Descrição dos dados e controle de qualidade

Dados genômicos foram obtidos a partir de 34 touros registrados da raça Nelore e de suas progênies, totalizando 780 novilhos machos. Os touros

constituíram famílias de meio-irmãos que foram geradas por inseminação artificial. Os novilhos foram produzidos em três estações de monta, nascidos em 2007, 2008 e 2009. Os animais foram criados em fazendas distintas da Embrapa Pecuária Sudeste, localizada na cidade de São Carlos (SP), da Embrapa Gado de Corte, situada no município de Campo Grande (MS) e de propriedades particulares dos estados de Mato Grosso e Mato Grosso do Sul.

A escolha dos touros para genotipagem foi realizada a partir de consultas aos catálogos das centrais de inseminação artificial. Para a composição desta amostra, 34 touros foram selecionados na população, de maneira que representassem as principais linhagens e genealogias que compõem a raça Nelore. Ao mesmo tempo, buscou-se minimizar o grau de parentesco entre os progenitores.

Os 814 animais da raça Nelore utilizados nesse estudo foram genotipados com o BovineHD BeadChip (Illumina), dos quais 93 animais (23 touros e 70 progênies) também foram genotipados com Axiom Genome-Wide BOS 1 Array Plate (Affymetrix), no Laboratório Multiusuários Centralizado de Genômica Funcional Aplicada à Agropecuária e Agroenergia, Piracicaba, São Paulo.

O controle de qualidade foi feito com a finalidade de remover erros de genotipagem excluindo-se SNPs localizados em cromossomos não autossômicos, com posições desconhecidas, com um p-valor no z-teste para o equilíbrio de Hardy-Weinberg menor que 10^{-5} , taxa de leitura ("call rate") menor que 0,98 e amostras que apresentaram taxa de leitura menor que 0,90. Após o controle de qualidade, restaram 809 animais com informação de 509.107 SNPs para o painel Illumina (IllumHD) e 93 animais com 427.875 SNPs para o painel Affymetrix (AffyHD).

2.2. Painéis considerados para a imputação

Um painel combinado (PC) foi feito utilizando IllumHD e AffyHD para formar um painel com maior densidade, com o objetivo de aumentar a probabilidade de obter SNPs com elevado DL com QTLs (“Quantitative Trait Loci”). Os painéis de alta densidade (IllumHD e AffyHD) apresentaram 56.646 SNPs em comum após o controle de qualidade e o painel combinado apresentou 880.336 SNPs.

Para estudar a acurácia de imputação de painéis de menor densidade para os painéis de alta densidade (IllumHD, AffyHD e PC), SNPs do genótipo IllumHD foram omitidos para imitar que os animais foram genotipados com GeneSeek Genomic Profiler LD v2 (20kCom) e com Illumina BovineSNP50 v2 BeadChip (50kCom). O painel 20kCom totalizou 15.575 SNPs e o 50kCom totalizou 27.946 SNPs.

Os SNPs em comum entre os painéis Illumina e Affymetrix (56.646 SNPs) foram utilizados para customizar painéis de 20k e 50k SNPs de três formas diferentes. Pelas diferentes formas de customização buscou-se selecionar a mesma quantidade de SNPs presentes nos painéis comerciais (20kCom e 50kCom) com a intenção de comparar a performance de imputação nas mesmas condições para todos os painéis. Para a primeira maneira de customizar os painéis (20kCust1 e 50kCust1), considerou-se os SNPs em comum com maior DL com os SNPs presentes no painel comercial (20kCom e 50kCom). Para a segunda maneira (20kCust2 e 50kCust2), considerou-se a mesma situação da primeira, porém com o adicional de que os SNPs que apresentavam frequência do alelo de menor frequência (MAF) menor que 0,09 foram descartados. Na terceira maneira de customizar os painéis, em uma janela de três SNPs, a MAF de cada SNP foi multiplicada pelo DL calculado entre este SNP e os outros presentes na janela.

Então, selecionou-se o SNP da janela que apresentou maior valor para soma dos resultados, formando o 20kCust3. Uma janela de dois SNPs foi considerada para obter o painel customizado 50kCust3. Customização semelhante foi realizada no estudo de Carvalheiro et al. (2014). Um esquema representando a terceira maneira de customização está apresentado na Figura 1.

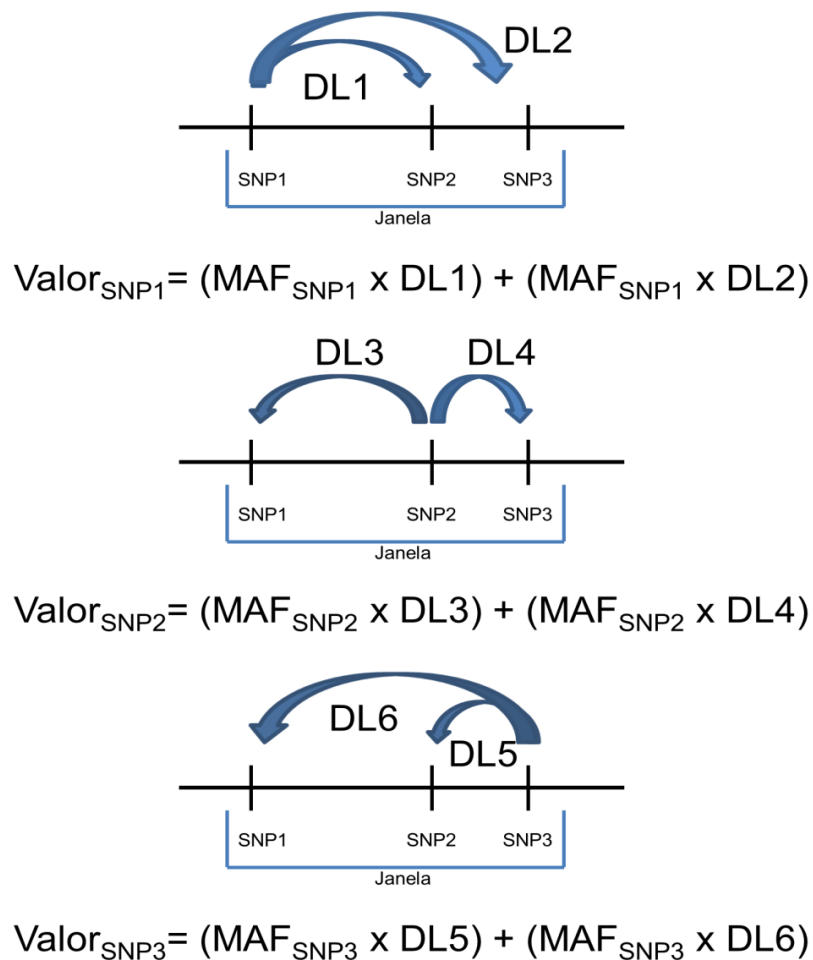


Figura 1. Esquema representando como foi realizada a terceira maneira de customização. Dentro da janela considerada foi calculado o desequilíbrio de ligação (DL1) entre o SNP1 e o SNP2 e este valor de DL1 foi multiplicado pela MAF do SNP1. Também foi calculado o desequilíbrio de ligação (DL2) entre o SNP1 e o SNP3, este valor de DL2 foi multiplicado pela MAF do SNP1. O resultado dos dois cálculos foi então somado. O mesmo cálculo foi repetido para cada SNP dentro da janela. O SNP que apresentou maior valor foi selecionado.

O DL para a customização foi mensurado por r^2 proposto por Hill and Robertson (1968). Assim, foram formados seis painéis customizados. Os painéis utilizados e as imputações realizadas estão demonstrados na Figura 2 e o número de SNPs em cada painel utilizado e o número de SNPs em comum entre cada painel está na Tabela 1.

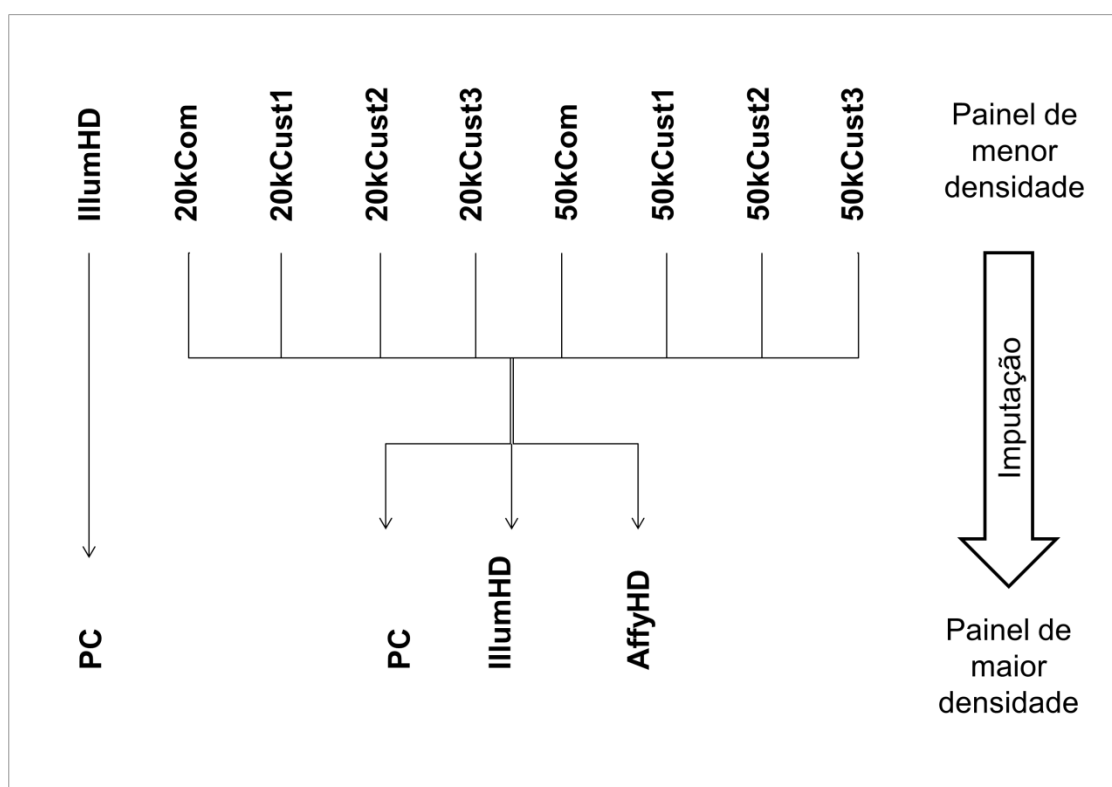


Figura 2. Organograma das imputações realizadas no presente estudo envolvendo o painel BovineHD BeadChip (IllumHD), um painel reduzido para GeneSeek Genomic Profiler LD v2 (20kCom) e um para BovineSNP50 v2 BeadChip, seis painéis customizados nas densidades de 20k e 50k (Cust1, Cust2 e Cust3), o painel Axion Genome-Wide BOS 1 Array Plate (AffyHD) e um painel combinado formado pelos painéis IllumHD e AffyHD.

Tabela 1. Número de SNPs em cada painel (diagonal) e o número de SNPs em comum (acima da diagonal) entre os painéis utilizados.

	PC	IllumHD	AffyHD	20kCom	50kCom	20kCust1	20kCust2	20kCust3	50kCust1	50kCust2	50kCust3
PC	880.336	509.107	427.875	15.575	27.946	15.575	15.575	15.575	27.946	27.946	27.946
IllumHD	-	509.107	56.646	15.575	27.946	15.575	15.575	15.575	27.946	27.946	27.946
AffyHD	-	-	427.875	15.575	27.946	15.575	15.575	15.575	27.946	27.946	27.946
20kCom	-	-	-	15.575	5.908	0	0	0	4.532	3.617	2.813
50kCom	-	-	-	-	27.946	0	0	0	18.753	12.832	8.765
20kCust1	-	-	-	-	-	15.575	12.680	5.647	10.813	10.239	9.075
20kCust2	-	-	-	-	-	-	15.575	6.654	9.681	12.207	10.065
20kCust3	-	-	-	-	-	-	-	15.575	7.783	10.579	14.291
50kCust1	-	-	-	-	-	-	-	-	27.946	19.580	13.880
50kCust2	-	-	-	-	-	-	-	-	-	27.946	16.867
50kCust3	-	-	-	-	-	-	-	-	-	-	27.946

IllumHD – Painel de alta densidade da Illumina; AffyHD – Painel de alta densidade da Affymetrix; PC – Painel combinado usando IllumHD e AffyHD; 20kCom – Painel comercial com cerca de 20 mil SNPs; 20kCust1, 20kCust2, 20kCust3- Primeiro, segundo e terceiro painel customizado com cerca de 20 mil SNPs, respectivamente; 50kCom - Painel comercial com cerca de 50 mil SNPs; 50kCust1, 50kCust2, 50kCust3- Primeiro, segundo e terceiro painel customizado com cerca de 50 mil SNPs, respectivamente.

2.3. *Imputação*

As análises de imputação foram conduzidas utilizando o programa FImpute v.2.2 (SARGOLZAEI et al., 2014). Em geral, os animais mais velhos são genotipados com painéis de alta densidade. Portanto, o cenário utilizado para todas as imputações foram com os 23 touros na população referência e os outros genótipos (786 animais) inseridos na população a ser imputada. O painel IllumHD foi imputado para o PC. Todos os oito painéis de baixa densidade (2 comerciais e 6 customizados) foram imputados para IllumHD, para AffyHD e para o PC (Figura 1). Para as imputações para IllumHD, a acurácia de imputação foi calculada para todos os 786 animais. Para as imputações para AffyHD e PC, a acurácia foi calculada para apenas os animais que possuíam o genótipo do AffyHD (70 progênies). Devido ao baixo número de animais para calcular a acurácia de imputação destes painéis, a acurácia de imputação para IllumHD foi feita também utilizando os mesmos 70 animais e as médias de acurácia foram comparadas com aquelas obtidas quando utilizados os 786 animais.

As acurácias foram calculadas pelo erro obtido comparando os marcadores imputados e observados, calculando a razão da proporção de genótipos imputados corretamente sobre todos os genótipos imputados (PROP). Também foi utilizada a correlação simples de Pearson (COR) entre o genótipo imputado e o observado. A relação entre a acurácia de imputação e MAF foi estudada estabelecendo 10 classes com intervalo de MAF de 0,05. A acurácia observada ao longo dos cromossomos também foi avaliada para todas as imputações.

2.4. Desequilíbrio de ligação e bloco de haplótipos

Duas medidas de DL foram consideradas neste estudo: o coeficiente de correlação entre alelos de dois loci (r^2) (HILL; ROBERTSON, 1968) e o $|D'|$ (LEWONTIN, 1964). As duas medidas foram calculadas utilizando o programa PLINK (PURCELL et al., 2007). O DL foi calculado entre todos os pares de SNP dentro de uma janela de 500kb e o decaimento do DL foi feito utilizando o programa R (R Core Team, 2018), pelo cálculo das médias de DL de 100 bins de 5 kb. Devido à combinação de painéis, quando calculado o DL para PC observou-se pares de SNPs que já tinham sido observados para os painéis IllumHD e AffyHD. Entretanto, para o painel PC também foi calculado o DL entre pares de SNPs que estavam ausentes nestes outros painéis de alta densidade, ou seja, pares de SNPs que foram formados com a junção dos painéis, assim, o decaimento do DL para esses novos pares de SNPs foi chamado de Novo_DL.

Médias de r^2 foram calculadas também entre SNPs que estavam presentes no painel de baixa densidade e SNPs presentes apenas nos painéis de alta densidade. Esta média foi calculada apenas para pares de SNPs com distância igual ou menor do que a média da distância entre dois SNPs no painel de baixa densidade, em que para a densidade de 20k foi de 173kb e para o de 50k foi de 96kb.

O estudo de bloco de haplótipos foi feito nos painéis IllumHD, AffyHD e PC, podendo assim comparar a quantidade e o tamanho dos blocos nestes painéis. A reconstrução dos blocos de haplótipos foi realizada após a imputação. Portanto, as considerações sobre os blocos de haplótipos foram feitas para um painel de alta densidade. As fases dos haplótipos foram reconstruídas utilizando o programa FImpute v.2.2 (SARGOLZAEI et al., 2014). Após isso, foi aplicada a definição de

Gabriel et al. (2002), que utiliza a informação de D' , para blocos de haplótipos, utilizando o programa Haploview (BARRETT et al., 2005) em cada cromossomo. Os resultados foram utilizados para construir um cariógrama com quatro classes de tamanho de haplótipos para IllumHD, AffyHD e PC utilizando o pacote “ggbio” (YIN; COOK; LAWRENCE, 2012) do programa R (R Core Team, 2018).

3. RESULTADOS E DISCUSSÃO

3.1. Imputação

De maneira geral, as acurácias de imputação foram elevadas considerando o tamanho da população referência (Tabela 2). Estes animais foram estudados por Mudadu et al. (2016), que avaliaram a estrutura desta população com informações genômicas e a descreveram com baixa diversidade, o que pode justificar a elevada acurácia observada para as análises. Embora a população de referência seja pequena, a presença dos touros na população auxiliou na reconstrução dos blocos de haplótipos, o que reduziu o erro de imputação. Khatkar et al. (2012), estudando bovinos de leite, observaram que a inclusão dos reprodutores na população referência diminuiu a taxa de erro da imputação quando comparado a população referência sem os reprodutores. Estes mesmos autores relataram que a média de taxa de erro decresce com o aumento da relação de parentesco entre a população referência e a população teste, o que pode ser mais pronunciado quando a população referência é pequena.

Tabela 2. Número de SNPs para serem imputados em bovinos da raça Nelore, acurácias medidas pela correlação (COR) e proporção de genótipos imputados corretamente (PROP) e desvios-padrão (DP) para imputações de painéis de baixa densidade para painéis de alta densidade.

Análise	Painel de menor densidade	Painel de maior densidade	Nº SNPs imputado	COR (DP)	PROP (DP)
1	IllumHD	PC	371.229	0,84(0,24)	84,11(18,39)
2	20kCom	IllumHD	493.532	0,88(0,09)	84,96(5,80)
3	20kCust1	IllumHD	493.532	0,87(0,09)	83,89(5,85)
4	20kCust2	IllumHD	493.532	0,87(0,09)	84,34(5,80)
5	20kCust3	IllumHD	493.532	0,89(0,09)	86,22(5,64)
6	20kCom	AffyHD	412.300	0,70(0,18)	68,08(10,00)
7	20kCust1	AffyHD	412.300	0,74(0,20)	73,47(12,23)
8	20kCust2	AffyHD	412.300	0,75(0,20)	74,14(12,54)
9	20kCust3	AffyHD	412.300	0,77(0,20)	75,49(13,22)
10	20kCom	PC	864.761	0,84(0,09)	81,24(6,22)
11	20kCust1	PC	864.761	0,83(0,09)	80,22(6,05)
12	20kCust2	PC	864.761	0,83(0,09)	80,67(6,10)
13	20kCust3	PC	864.761	0,84(0,09)	82,16(6,26)
14	50kCom	IllumHD	481.161	0,88(0,10)	85,94(6,05)
15	50kCust1	IllumHD	481.161	0,88(0,09)	85,99(5,85)
16	50kCust2	IllumHD	481.161	0,89(0,09)	86,96(5,70)
17	50kCust3	IllumHD	481.161	0,90(0,09)	88,21(5,52)
18	50kCom	AffyHD	399.929	0,73(0,19)	72,65(11,93)
19	50kCust1	AffyHD	399.929	0,74(0,20)	73,95(12,61)
20	50kCust2	AffyHD	399.929	0,77(0,21)	75,93(13,40)
21	50kCust3	AffyHD	399.929	0,78(0,21)	76,91(13,95)
22	50kCom	PC	852.390	0,84(0,09)	81,93(6,31)
23	50kCust1	PC	852.390	0,84(0,09)	82,00(6,33)
24	50kCust2	PC	852.390	0,85(0,09)	82,87(6,43)
25	50kCust3	PC	852.390	0,86(0,09)	83,87(6,55)

IllumHD – Painel de alta densidade da Illumina; AffyHD – Painel de alta densidade da Affymetrix; PC – Painel combinado usando IllumHD e AffyHD; 20kCom – Painel comercial com cerca de 20 mil SNPs; 20kCust1, 20kCust2, 20kCust3- Primeiro, segundo e terceiro painel customizado com cerca de 20 mil SNPs, respectivamente; 50kCom - Painel comercial com cerca de 50 mil SNPs; 50kCust1, 50kCust2, 50kCust3- Primeiro, segundo e terceiro painel customizado com cerca de 50 mil SNPs, respectivamente.

A estimativa de COR foi em geral maior que o valor de PROP para as mesmas análises (Tabela 2). A maior acurácia observada para ambas as medidas foi a de 50kCust3 para IllumHD (análise 17) e as menores acurácias foram de 20kCom para AffyHD (análise 6). Valores menores para a medida PROP quando comparado a COR também foi observado por Carvalheiro et al. (2014). Segundo estes autores, a maior penalidade dada para um alelo imputado incorretamente para a primeira medida pode justificar o resultado obtido.

A alta acurácia observada para a análise 17 (Tabela 2) foi provavelmente devido à alta densidade somada ao elevado DL observado entre SNPs que estavam presentes no painel de baixa densidade e SNPs presentes apenas no painel de alta densidade (Tabela 3). Embora a estimativa de DL entre os SNPs presentes no painel de baixa densidade e os SNPs presentes apenas no painel de alta densidade (Tabela 3) seja maior de 20kCust1 para PC, de 20kCust3 para IllumHD e de 20kCust3 para PC do que de 50kCust3 para IllumHD, o fato de haver menor quantidade de SNPs a serem imputados quando utilizada densidade de 50k quando comparado com 20k, favorece a construção de haplótipos pelo programa, o que pode reduzir os erros de imputação.

Segundo Pei et al. (2008), muitos fatores podem influenciar na acurácia de imputação. No entanto, o DL apresentou papel central dentre os métodos testados pelos autores. Os mesmos fatores (desequilíbrio de ligação e densidade) podem contribuir para as baixas estimativas de acurácia de imputação observados para a análise 6 (Tabela 2), em que dentre as menores estimativas de DL (Tabela 3), o 20kCom é o painel com menor densidade.

Tabela 3. Médias dos valores de r^2 de desequilíbrio de ligação e desvios-padrão (DP) entre SNPs presentes no painel de baixa densidade e SNPs presentes apenas no painel de maior densidade.

Menor densidade	Maior densidade		
	IllumHD	AffyHD	PC
20kCom	0,25±0,29	0,21±0,26	0,28±0,31
20kCust1	0,29±0,32	0,22±0,26	0,31±0,33
20kCust2	0,29±0,31	0,22±0,25	0,30±0,31
20kCust3	0,33±0,33	0,24±0,26	0,33±0,32
50kCom	0,26±0,32	0,20±0,26	0,28±0,33
50kCust1	0,28±0,32	0,20±0,26	0,29±0,33
50kCust2	0,27±0,31	0,21±0,25	0,29±0,31
50kCust3	0,30±0,32	0,22±0,26	0,30±0,32

A imputação de IllumHD para PC (análise 1) apresentou elevada estimativa de acurácia, porém seguida do maior erro-padrão observado no presente estudo (Tabela 2). A seleção dos SNPs para compor os painéis da IllumHD e da AffyHD, foram feitas de formas diferentes, o que pode promover dificuldades para formar os haplótipos quando a imputação é realizada entre painéis de diferentes empresas. Assim, animais com menor nível de parentesco com a população de referência podem apresentar maior dificuldade em identificar os haplótipos e consequentemente podem apresentar menor acurácia de imputação.

Embora possua maior número de SNPs a serem imputados para o PC, as menores densidades de 20k e 50k apresentaram estimativas maiores ou iguais à acurácias de imputação observada para a análise 1 (IllumHD para PC - Tabela 2). Este resultado pode ser explicado pelo decaimento de DL, em que o PC, mesmo em curtas distâncias entre os marcadores SNP apresenta estimativas menores que o IllumHD (resultados não mostrados), o que implica que SNPs a serem imputados do

IllumHD para PC tem menor DL com SNPs presentes no painel IllumHD, dificultando a imputação.

As acurácias obtidas quando o painel de maior densidade era o IllumHD (análises de 2-5 e 4-17) utilizando apenas 70 animais apresentaram estimativas semelhantes às obtidas utilizando os 786 animais, em que a diferença foi de 0,01 para COR e de uma unidade (1,0) para PROP. Esta semelhança pode ter sido observada devido à estrutura desta população, em que os animais apresentam elevado grau de parentesco, assim a média das estimativas de acurácia observada para poucos animais reflete na média das estimativas obtidas para a população. De maneira geral, as menores acurácias de imputação foram para AffyHD, seguida de imputações para PC e então imputações para o IllumHD. A densidade e o DL são fatores que podem explicar a razão para elevadas estimativas de acurácia de imputação quando o painel de alta densidade é o IllumHD, o qual, embora apresente menor DL entre SNPs presentes no painel de baixa densidade e os presentes apenas no painel de alta densidade (Tabela 3) comparado ao PC, o IllumHD possui menor número de SNPs a serem imputados.

A primeira e a segunda maneira de customizar os painéis de baixa densidade (Cust1 e Cust2) apresentaram acurácia de imputação menor que o painel comercial para a densidade de 20k para PC e IllumHD (Tabela 2). Para imputações da densidade de 50k para PC e IllumHD, foram observadas estimativas de acurácia muito semelhantes entre 50kCom e 50kCust1 e estimativa de acurácia um pouco superior para 50kCust2 (Tabela 2). Entretanto, foi possível observar que a acurácia para estes dois painéis customizados quando imputados para o painel AffyHD foi maior que aquela para o painel comercial. Em geral, quando comparados os painéis

comerciais com os painéis customizados, a terceira forma de customização (Cust3) foi a que aumentou a acurácia de imputação para quase todas as situações estudadas (Tabela 2). Esta forma de customização considera dois fatores que afetam a acurácia de imputação: DL e MAF. Escolher SNP com maior DL com outro SNP em uma janela contribui para a reconstrução de haplótipos. Este fato somado a exclusão de SNPs quase fixados na população pode ser considerado para customizar um painel. Isso porque a ausência de variabilidade de genótipos para determinado SNPs na população pode dificultar estudos de associação com cobertura ampla do genoma, exigindo uma população muito grande para poder encontrar alguma associação.

O aumento na acurácia de imputação ao utilizar a terceira forma de customização (Cust3) foi mais visível quando a imputação foi para o painel AffyHD (Tabela 1). Isto pode ter ocorrido pois os painéis customizados da terceira maneira removeu os SNPs com baixa MAF, conseqüentemente, estes SNPs precisam ser imputados quando realizada a imputação. Além disso, segundo Pei et al. (2008), a influência da MAF pode ser maior em regiões do genoma de baixo DL, pois nestas regiões, marcadores com baixa MAF provavelmente apresentaram maior DL local com marcadores próximos, mesmo que a região como um todo apresente baixo DL. O fato de o AffyHD apresentar menor média de DL (Tabela 3) e assim, provavelmente conter mais regiões em baixo DL, pode fazer com que os SNPs com baixa MAF que precisam ser imputados quando utilizada a terceira forma de customização (Cust3) tenham maior influência, fazendo com que o aumento na acurácia seja maior para o painel AffyHD.

As relações entre a acurácia de imputação e a MAF observadas para imputação de IllumHD para PC (Figura 3) foram semelhantes para todos os painéis (resultados não mostrados). A relação observada entre a acurácia medida pela COR e MAF foi oposta aquela verificada entre PROP e MAF, em que ao aumentar a MAF, a COR também aumentou, mas a PROP reduziu. Estes resultados também foram observados por Ma et al. (2013), que justificaram o resultado por ser PROP uma medida que não considera que a correta imputação tenha ocorrido de forma aleatória, o que favorece SNPs com baixa MAF. Os autores reportaram que por esta razão, a medida de correlação para calcular a acurácia de imputação pode ser melhor para SNPs com baixa MAF.

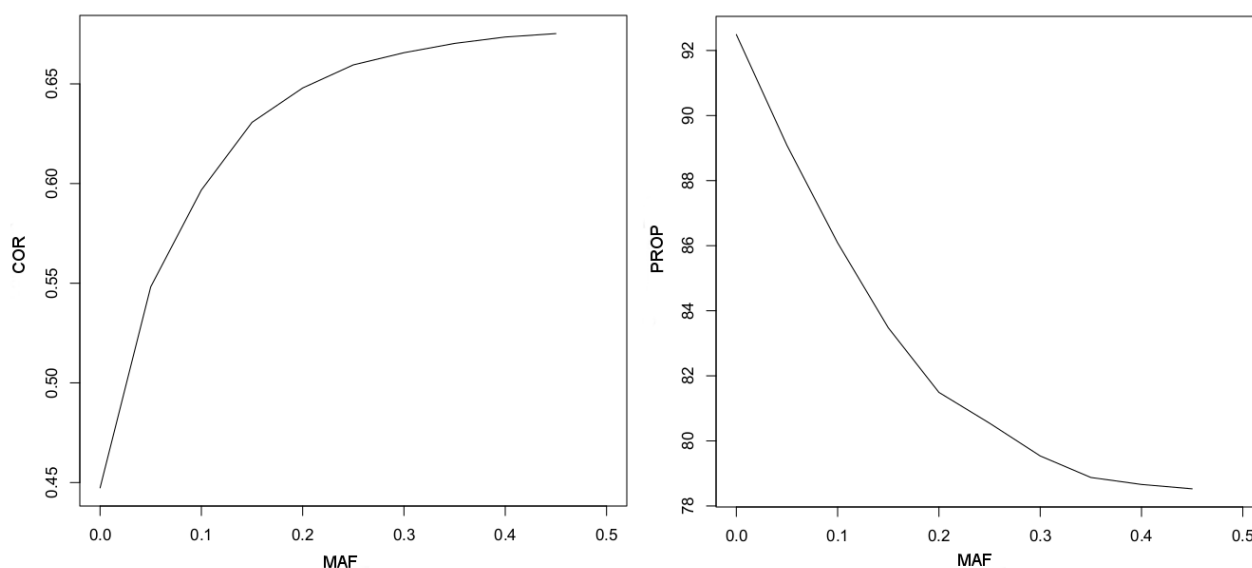


Figura 3. Relação observada entre as acurácias medidas pela correlação (COR) e medida pela proporção de genótipos imputados corretamente (PROP) em relação à MAF.

Mesmo com pequenas diferenças, a acurácia de imputação foi diferente entre os cromossomos. Nas imputações para IllumHD, as maiores acurácias foram observadas para os cromossomos (BTA) 6 e 8, enquanto que as menores foram

observadas para o BTA 25. Quando a imputação foi para AffyHD, os cromossomos com maiores acurácias foram os BTA 5 e BTA 6 e os com menores foram os BTA 19, BTA 25 e BTA 29. Devido ao PC representar a soma dos painéis, os cromossomos que apresentaram maiores estimativas para acurácia para IllumHD e AffyHD, também apresentaram elevada acurácia para PC (BTA 5, BTA 6 e BTA 8). O mesmo foi observado para as menores estimativas de acurácias de imputação, em que os menores valores foram nos BTA 19, BTA 25 e BTA 29.

Piccoli et al. (2014) sugeriram que o tamanho do cromossomo está relacionado com a acurácia de imputação, sendo que maiores cromossomos apresentam maiores acurácias de imputação. No presente estudo, foi possível identificar pequena diferença na performance da acurácia de imputação entre cromossomos quando o painel de maior densidade era IllumHD, AffyHD ou PC. Este resultado sugere que não apenas o tamanho dos cromossomos, mas também a estrutura de bloco de haplótipos pode influenciar a performance de imputação para os diferentes cromossomos. Quando a imputação foi para o IllumHD, o BTA 25 (o menor dos autossomos) apresentou menor estimativa de acurácia. No entanto, mesmo não sendo os cromossomos mais longos, o BTA 6 e o BTA 8 apresentaram maiores estimativas para acurácia. Estes cromossomos estão entre os cinco com maior número de haplótipos com mais de 10 SNPs para IllumHD (Figura 4A) e estes são os dois primeiros quando considerado a razão do número de haplótipos dividido pelo tamanho do cromossomo (densidade). Em adição, estes haplótipos são concentrados ao longo dos cromossomos. O número de haplótipos parece também influenciar a imputação para AffyHD, em que as maiores estimativas de acurácia foram observadas para cromossomos com maior número de haplótipos (BTA 5 e

BTA 6 – Figura 4B). O PC também apresentou muitos blocos de haplótipos nos cromossomos que apresentaram as maiores estimativas de acurácia de imputação (Figura 4C).

3.2. *Desequilíbrio de ligação (DL) e bloco de haplótipos*

O decaimento do DL para IllumHD, AffyHD e Novo_DL está representado na Figura 5. A medida de r^2 apresenta diferença de decaimento quando comparado com D' , o qual para IllumHD e AffyHD observou-se estimativas próximas em curtas distâncias e para longas distâncias o painel AffyHD manteve elevadas estimativas. O Novo_DL para esta medida seguiu o painel IllumHD. Estimativas maiores para D' foram observadas e, segundo Espigolan et al. (2013), esta medida pode superestimar o DL. Estes autores relataram que outra desvantagem desta medida é que a mesma pode ser fortemente superestimada quando utilizada em pequenas amostras ou SNPs com baixa MAF. Este fato pode explicar o decaimento de D' para o painel AffyHD o qual, embora possua menor proporção de SNPs com baixa MAF, deve-se considerar que a amostra utilizada para o DL foi menor (70 animais) do que para os demais painéis.

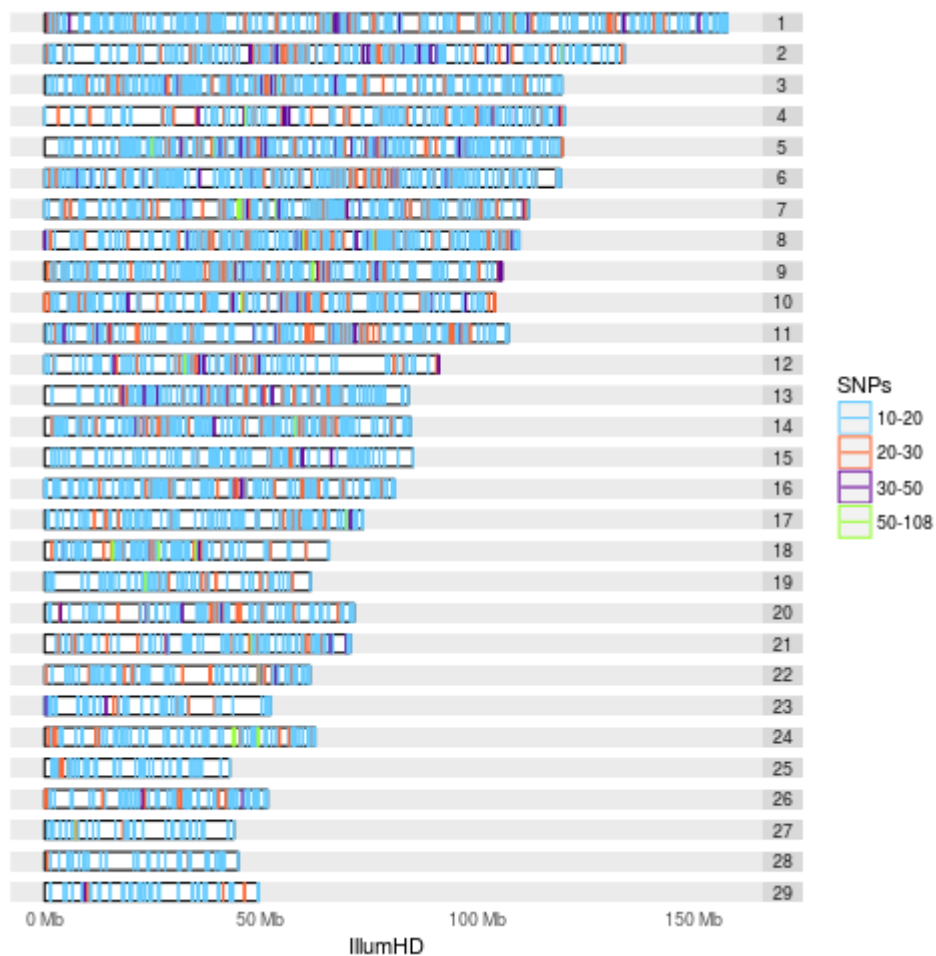


Figura 4A. Blocos de haplótipos compostos por mais de 10 SNPs, distribuídos ao longo dos cromossomos detectados quando utilizado o painel de alta densidade Illumina (IllumHD) em bovinos Nelore.

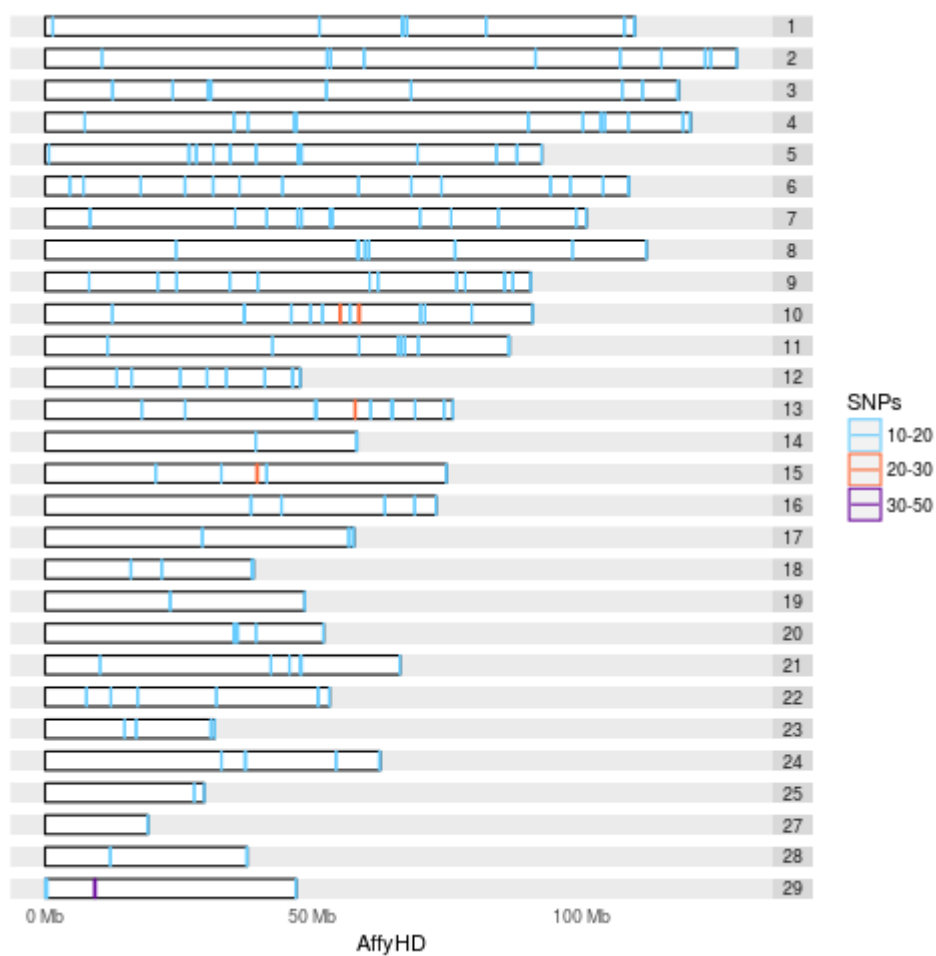


Figura 4B. Blocos de haplótipos compostos por mais de 10 SNPs, distribuídos ao longo dos cromossomos detectados quando utilizado o painel de alta densidade Affymetrix (AffyHD) em bovinos Nelore.

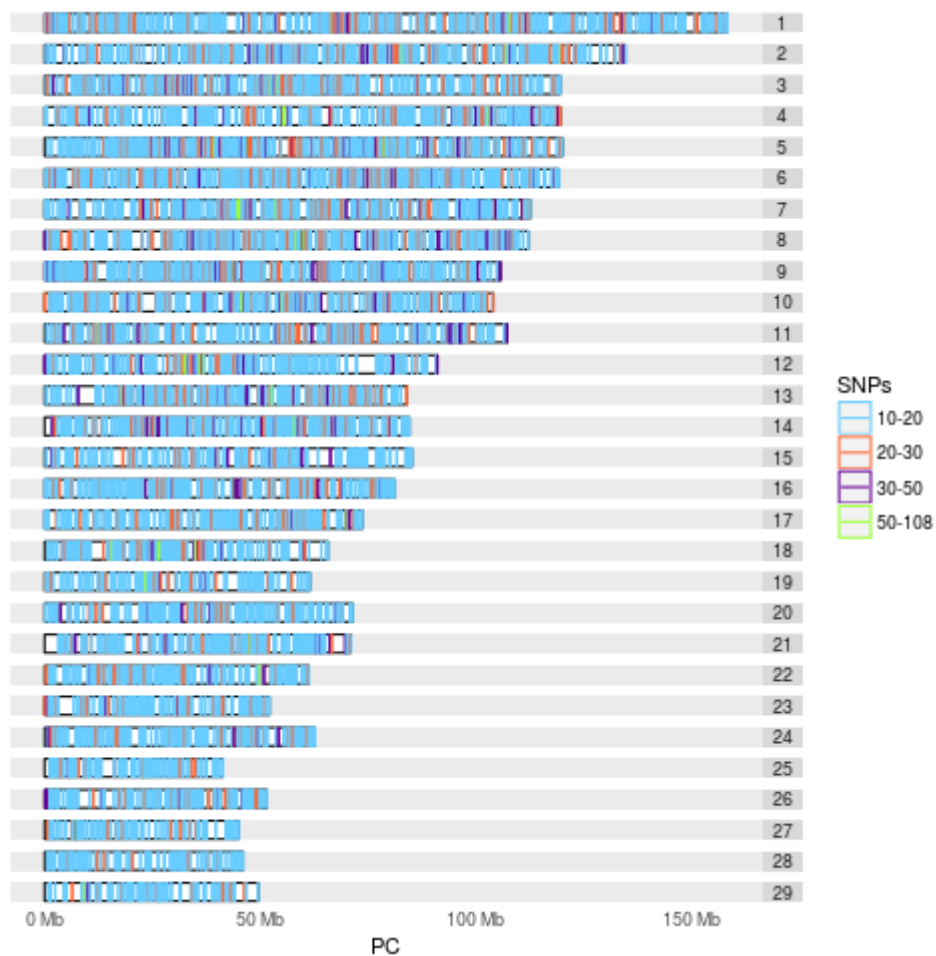


Figura 4C. Blocos de haplótipos compostos por mais de 10 SNPs, distribuídos ao longo dos cromossomos detectados quando utilizado o painel combinado (PC) utilizando IllumHD e AffyHD em bovinos Nelore.

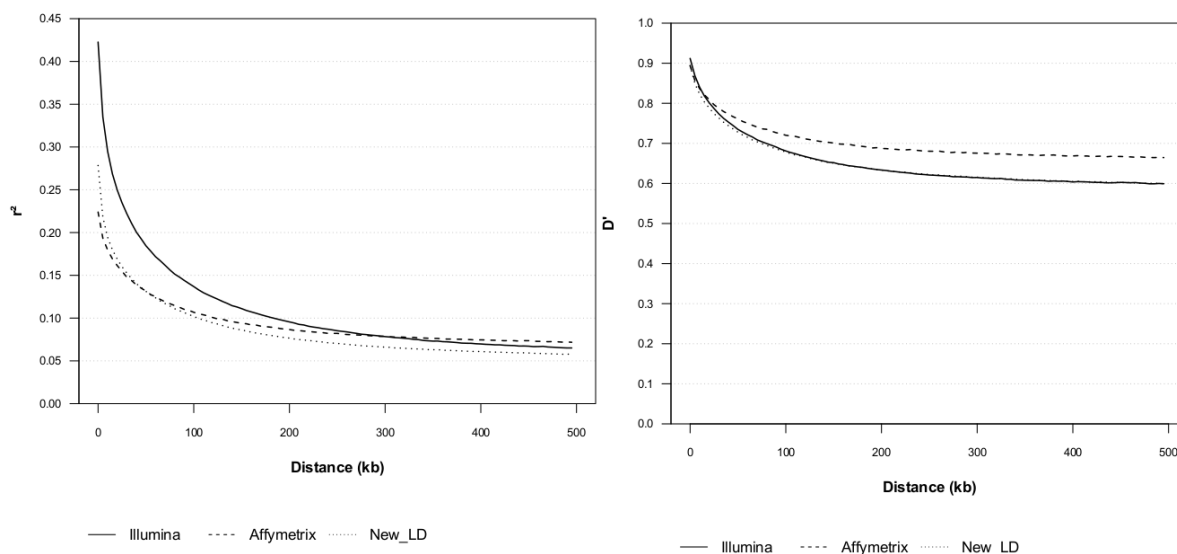


Figura 5. Decaimento de desequilíbrio de ligação em função do painel de alta densidade Illumina (IllumHD), Affymetrix (AffyHD) e de par de alelos que estão presentes no painel combinado (Illumina + Affymetrix) e ausentes nos demais painéis de alta densidade individualmente (Novo_DL).

Para r^2 , IllumHD apresentou maiores estimativas de DL quando comparado a AffyHD para curtas distâncias e o Novo_DL apresentou estimativas pouco maiores que o AffyHD, mas com o aumento da distância entre marcadores de SNP, o Novo_DL apresentou estimativas menores. As estimativas de r^2 observadas no presente estudo foram menores que as reportadas por O'Brien et al. (2014) para bovinos da raça Nelore. Esta diferença pode ser atribuído ao fato que neste estudo as médias de DL foram obtidas a cada 1Kb, enquanto no presente trabalho foi utilizado médias a cada 5kb.

Mesmo com estimativas frágeis de DL devido ao tamanho da amostra, o painel AffyHD apresentou menores estimativas de r^2 para curtas distâncias. A medida de r^2 é influenciada pela frequência dos SNPs e, segundo Van Binsbergen et al. (2014), em extremas diferenças de MAF entre um par de SNPs, mesmo que a distância entre os mesmos seja pequena, a máxima estimativa de DL que este par

de SNPs pode ter, é baixa. Para os painéis IllumHD e AffyHD, foi observado que 47% e 39%, respectivamente, dos pares de SNPs apresentaram diferenças de MAF menor que 0,1, o que pode justificar o menor DL para o painel AffyHD em curtas distâncias.

Como resultado da soma dos SNPs presente nos dois painéis de alta densidade, o PC apresentou decaimento do DL equivalente à média desta combinação (resultado não apresentado). Embora o PC não tenha superado o decaimento de DL do IllumHD, a maior densidade para o PC aumenta a probabilidade de obter um SNP com alto DL com um QTL.

O painel IllumHD apresentou 84.529 blocos de haplótipos (Figura 4A), o que foi mais do que observado para AffyHD, com 63.967 blocos de haplótipos (Figura 4B) e menos do que o observado para o painel combinado (PC), com 140.336 blocos de haplótipos (Figura 4C). Foi possível observar também variabilidade no tamanho dos blocos quando foram utilizados diferentes painéis, em que o IllumHD foi composto por blocos com média igual a $137,70 \pm 219,05$ kb e mediana de 69,35 kb, o AffyHD com média de $102,10 \text{ kb} \pm 155,47$ kb e mediana de 57,55 kb e o PC com média igual a $107,10 \text{ kb} \pm 169,14$ kb e mediana de 55,39 kb.

As médias observadas neste estudo para os tamanhos dos blocos de haplótipos dos três painéis foram maiores que as relatadas para bovinos Simmental Chinês e Wagyu (NIU et al., 2016) genotipados com alta densidade Illumina e utilizando a mesma metodologia que o presente estudo, o que pode ser explicado pelas diferenças observadas para a estrutura populacional. A maior parte dos blocos de haplótipos nos três painéis é composta por menos de 10 SNPs, em que apenas 3.882, 193 e 8.462 blocos de haplótipos são compostos por 10 SNPs ou mais para

os painéis IllumHD (Figura 4A), AffyHD (Figura 4B) e PC (Figura 4C), respectivamente. No painel IllumHD foi possível identificar 27 blocos de haplótipos com 50 SNPs ou mais, sendo que o maior bloco era composto por 95 SNPs em 326,5 kb. Para este painel, os BTA 7 e BTA 18 foram os que apresentaram maiores quantidades destes haplótipos maiores, com três longos blocos de haplótipos. Para o painel AffyHD não foram detectados blocos de haplótipos maiores que 50 SNPs. O painel AffyHD foi o que apresentou menor quantidade de blocos de haplótipos em menores distâncias, comparado com os demais painéis. O fato deste painel apresentar menor DL entre SNPs contribui para a menor identificação de blocos de haplótipos.

O painel combinado (PC) apresentou 38 haplótipos com mais de 50 SNPs, em que os BTA 2, BTA 5, BTA 7, BTA 12 e BTA 18 apresentaram três destes longos blocos. Para o PC, o maior bloco de haplótipo possui 108 SNPs em 310,5kb. Segundo Cuyabano et al. (2015), a principal vantagem da utilização da informação de blocos de haplótipos na seleção genômica, em vez da informação de SNPs individualmente, é que cada haplótipo pode estar em maior DL com uma mutação causal do que qualquer SNP individualmente estaria. Assim, ao estudar a predição genômica em bovinos de leite, utilizando informação de blocos de haplótipos, Cuyabano, Su e Lund (2015) observaram aumento na acurácia de predição do valor genômico do que ao obtidos usando-se SNP individualmente.

A combinação dos dois painéis parece colaborar com o aumento da quantidade de blocos de haplótipos, o que pode ser importante para demais estudos de abordagem genômica, mesmo considerando que a média de tamanho dos blocos de haplótipos para PC foi menor que para o IllumHD. A mediana menor que a média

para o PC evidencia que a maioria dos haplótipos são pequenos. No entanto, o aumento na “cobertura” dos cromossomos com longos haplótipos, quando utilizado o PC, comparado ao IllumHD e o aumento na cobertura para os cromossomos curtos (BTA 23 a BTA29 – Figura 4) podem contribuir em estudos que exploram blocos de haplótipos.

4. CONCLUSÃO

O uso do painel combinado (PC) é recomendado devido ao aumento da densidade e o número de blocos de haplótipos, o que eleva a probabilidade de obter um marcador próximo a um QTL de interesse. Deve-se considerar SNPs em comum para o painel IllumHD e AffyHD para customizar um painel de menor densidade, pois permite elevar a acurácia de imputação para o painel IllumHD, AffyHD e PC.

5. REFERÊNCIAS

- BARRETT, J. C.; FRY, B.; MALLER, J.; DALY, M. J. Haploview: analysis and visualization of LD and haplotype maps. **Bioinformatics**, Oxford, v. 21, n. 2, p. 263 – 265, 2005.
- BERRY, D. P.; KEARNEY, J.F. Imputation of genotypes from low- to high-density genotyping platforms and implications for genomic selection. **Animal**, Cambridge, v. 5, n. 8, p. 1162 – 1169, 2011.
- BERRY, D. P.; MCPARLAND, S.; KEARNEY, J. F.; SARGOLZAEI, M.; MULLEN, M. P. Imputation of ungenotyped parental genotypes in dairy and beef cattle from progeny genotypes. **Animal**, Cambridge, v. 8, n. 6, p. 895 – 903, 2014.
- BROWNING, B. L.; BROWNING, S. R. A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. **The American Journal of Human Genetics**, Houston, v. 84, p. 210 – 223, 2009.
- CARVALHEIRO, R.; BOISON, S.; NEVES, H. H. R.; SARGOLZAEI, M.; SCHENKEL, F. S.; UTSUNOMIYA, Y. T.; O'BRIEN, A. M. P.; SÖLKNER, J.; MCEWAN, J. C.; VAN TASSELL, C. P.; SONSTEGARD, T. S.; GARCIA, J. F. Accuracy of genotype imputation in Nelore cattle. **Genetics Selection Evolution**, London, v. 46, p. 69, 2014.

CHUD, T. C. S.; VENTURA, R. V.; SCHENKEL, F. S.; CARVALHEIRO, R.; BUZANSKAS, M. E.; ROSA, J. O.; MUDADU, M. A.; SILVA, M. V. G. B.; MOKRY, F. B.; MARCONDES, C. R.; REGITANO, L. C. A.; MUNARI, D. P. Strategies for genotype imputation in composite beef cattle. **BMC Genetics**, London, v. 16, n. 99, 2015.

CUYABANO, B. C.; SU, G.; LUND, M. Selection of haplotype variables from a high-density marker map for genomic prediction. *Genetics Selection Evolution*, v. 47, n. 1, p. 61, 2015.

CUYABANO, B. C. D.; SU, G.; ROSA, G. J. M.; LUND, M. S.; GIANOLA, D. Bootstrap study of genome-enabled prediction reliabilities using haplotype blocks across Nordic Red cattle breeds. *Journal of dairy science*, Champaign, v. 98, n. 10, p. 7351-7363, 2015.

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

GABRIEL, S. B.; SCHAFFNER, S. F.; NGUYEN, H.; MOORE, J. M.; ROY, J.; BLUMENSTIEL, B.; HIGGINS, J.; DEFELICE, M.; LOCHNER, A.; FAGGART, M.; LIU-CORDERO, S. N.; ROTIMI, C.; ADEYEMO, A.; COOPER, R.; WARD, R.; LANDER, E. S.; DALY, M. J.; ALTSHULER, D. The structure of haplotype blocks in the human genome. **Science**, Washington, v. 296, n. 5576, p. 2225 – 2229, 2002.

HAYES, B. J.; LEWIN, H. A.; GODDARD, M. E. The future of livestock breeding: genomic selection for efficiency, reduced emissions intensity, and adaptation. **Trends in Genetics**, Philadelphia, v. 29, n. 4, p. 206 – 214, 2013.

HILL, W.; ROBERTSON, A. Linkage disequilibrium in finite populations. **Theoretical and Applied Genetics**, Heidelberg, v. 38, n. 6, p. 226 – 231, 1968.

HOZÉ, C.; FOUILLOUX, M. N.; VENOT, E.; GUILLAUME, F.; DASSONNEVILLE, R.; FRITZ, S.; DUCROCQ, V.; PHOCAS, F.; BOICHARD, D.; CROISEAU, P. High-density marker imputation accuracy in sixteen French cattle breeds. **Genetics Selection Evolution**, London, v. 45, n. 33, 2013.

HUANG, Y.; HICKEY, J. M.; CLEVELAND, M. A.; MALTECCA, C. Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. **Genetics Selection Evolution**, London, v. 44, n. 25, 2012.

KHATKAR, M. S.; MOSER, G.; HAYES, B. J.; RAADSMA, H. W. Strategies and utility of imputed SNP genotypes for genomic analysis in dairy cattle. **BMC Genomics**, London, v. 13, p. 538, 2012.

LEWONTIN, R. C. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. **Genetics**, Bethesda, v. 49, n. 1, p.49 – 67, 1964.

MA, P.; BRØNDUM, R. F.; ZHANG, Q.; LUND, M. S.; SU, G. Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish red Cattle. **Journal of Dairy Science**, Champaign, v. 96, p. 4666–4677, 2013.

MEUWISSEN, T. H. Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. **Genetics Selection Evolution**, London, v. 41, n. 35, 2009.

MUDADU, M. A.; PORTO-NETO, L. R.; MOKRY, F. B.; TIZIOTO, P. C.; OLIVEIRA, P. S. N.; TULLIO, R. R.; NASSU, R. T.; NICIURA, S. C. M.; THOLON, P.; ALENCAR, M. M.; HIGA, R. H.; ROSA, A. N.; FEIJÓ, G. L. D.; FERRAZ, A. L. J.; SILVA, L. O. C.; MEDEIROS, S. R.; LANNA, D. P.; NASCIMENTO, M. L.; CHAVES, A. S.; SOUZA, A. R. D. L.; PACKER, I. U.; TORRES JR, R. A. A.; SIQUEIRA, F.; MOURÃO, G. B.; COUTINHO, L. L.; REVERTER, A.; REGITANO, L. C. A. Genomic structure and marker-derived gene networks for growth and meat quality traits of Brazilian Nelore beef cattle. **BMC Genomics**, London, v. 17, p. 235, 2016.

NIU, H.; ZHU, B.; GUO, P.; ZHANG, W.; XUE, J.; CHEN, J.; ZHANG, L.; GAO, H.; GAO, X.; XU, L.; LI, J. Estimation of linkage disequilibrium levels and haplotype block structure in Chinese Simmental and Wagyu beef cattle using high-density genotypes. **Livestock Science**, Oxford, v. 190, p. 1-9, 2016.

O'BRIEN, A. M. P.; MÉSZÁROS, G.; UTSUNOMIYA, Y. T.; SONSTEGARD, T. S.; GARCIA, J. F.; VAN TASSELL, C. P.; CARVALHEIRO, R.; DA SILVA, M. V. B.; SÖLKNER, J. Linkage disequilibrium levels in *Bos indicus* and *Bos Taurus* cattle using medium and high density SNP chip data and different minor allele frequency distributions. **Livestock Science**, Oxford, v. 166, p. 121-132, 2014.

PEI Y, LI J, ZHANG L, PAPASIAN CJ, DENG H. Analyses and Comparison of Accuracy of Different Genotype Imputation Methods. **Plos One**. 2008, 3:10.

PICCOLI, M.; BRACCINI, J.; CARDOSO, F.; SARGOLZAEI, M.; LARMER, S.G.; SCHENKEL, F. Accuracy of genome-wide imputation in Braford and Hereford beef cattle. **BMC Genetics**, London, v. 15, n. 157, 2014.

PURCELL, S.; NEALE, B.; TODD-BROWN, K.; THOMAS, L.; FERREIRA, M. A. R.; BENDER, D.; MALLER, J.; SKLAR, P.; BAKKER, P. I. W.; DALY, M. J.; SHAM, P. C. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. **The American Journal of Human Genetics**, Houston, v. 81, p. 559 – 575, 2007.

R Core Team. R: A language and environment for statistical computing. Viena: R Foundation for Statistical Computing. Available: <http://www.R-project.org/>. Acessado: 12 Janeiro 2018.

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

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

SCHENKEL, F. Genome Wide Imputation in Canadian Beef Cattle. In: 10th World Congress of Genetics Applied to Livestock Production, 2014, Vancouver. Available in: https://www.asas.org/docs/default-source/wcgalp-proceedings-oral/259_paper_10342_manuscript_1326_0b.pdf?sfvrsn=2.

SUN, C.; WU, X-L.; WEIGEL, K. A.; ROSA, G. J. M.; BAUCK, S.; WOODWARD, B. W.; SCHNABEL, R. D.; TAYLOR, J. F.; GIANOLA, D. An ensemble-based approach to imputation of moderate-density genotypes for genomic selection with application to Angus cattle. **Genetical Research**, Cambridge, v. 94, p. 133–150, 2012.

VAN BINSBERGEN, R.; BINK, M. C.; CALUS, M. P.; VAN EEUWIJK, F. A.; HAYES, B. J.; HULSEGGE, I.; VEERKAMP, R. F. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. **Genetics Selection Evolution**, London, v. 46, n. 1, n. 41, 2014.

VANRADEN, P. M.; O'CONNELL, J. R.; WIGGANS, G. R.; WEIGEL, K. A. Genomic evaluations with many more genotypes. **Genetics Selection Evolution**, London, v. 43, p. 10-20, 2011.

VANRADEN, P. M.; NULL, D. J.; SARGOLZAEI, M.; WIGGANS, G. R.; TOOKER, M. E.; COLE, J. B.; SONSTEGARD, T. S.; CONNOR, E. E.; WINTERS, M.; VAN KAAM, J. B. C. H. M.; VALENTINI, A.; VAN DOORMAAL, B. J.; FAUST, M. A.; DOAK, G. A. Genomic imputation and evaluation using high-density Holstein genotypes. **Journal of Dairy Science**, Champaign, v. 96, p. 668 – 678, 2013.

WEIGEL, K. A.; DE LOS CAMPOS, G.; VAZQUEZ, A. I.; ROSA, G. J. M.; GIANOLA, D.; VAN TASSELL, C. P. Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. **Journal of Dairy Science**, Champaign, v. 93, p. 5423 – 5435, 2010.

YIN, T.; COOK, D.; LAWRENCE, M. “ggbio: an R package for extending the grammar of graphics for genomic data.” **Genome Biology**, London, v. 13, n. 8, 2012.

CAPÍTULO 3. Predição da acurácia de imputação em bovinos da raça Nelore utilizando redes neurais artificiais

RESUMO – Para que dados genômicos imputados sejam utilizados em subsequentes estudos de associação ou de seleção genômica, é necessário que a imputação tenha sido implementada de forma que todos os animais envolvidos tenham seus genótipos inferidos com elevada acurácia. No entanto, a acurácia de imputação somente pode ser verificada quando os marcadores contidos nos painéis de alta densidade são omitidos da análise de imputação para obtenção do painel de baixa densidade de interesse. Dessa maneira, este estudo teve como objetivo estudar estratégias para predição da acurácia de imputação em dados genômicos de bovinos da raça Nelore, antes que a imputação seja efetuada, utilizando redes neurais artificiais e regressão linear múltipla. Foram utilizados dados de 814 bovinos da raça Nelore genotipados com o BovineHD BeadChip, sendo que destes, 93 animais também foram genotipados com o Axion Genome-Wide BOS 1 Array Plate. Diferentes cenários de imputação foram realizados para estes animais utilizando o programa FImpute v.2.2 e foram calculadas as acurácias de imputação. Foram escolhidos sete preditores para serem incluídos nas análises de regressão linear múltipla e de redes neurais: composição genética, tamanho da população referência, as duas densidades dos painéis utilizados, duas medidas de desequilíbrios de ligação (DL) dos painéis envolvidos e relacionamento genético entre os animais. A população de treinamento foi feita com 60% dos animais para ambos os modelos. A habilidade de predição dos modelos foi calculada por meio da correlação entre os valores de acurácia de imputação observados com os valores de acurácia preditos pelos dois métodos além do cálculo do erro padrão médio. As redes neurais mostraram-se mais eficientes para predizer a acurácia de imputação comparado ao modelo de regressão linear múltipla, podendo ser utilizadas com esta finalidade. Os preditores que envolveram DL forneceram informações similares à informação de densidade dos painéis para predição da acurácia de imputação, podendo ser desconsiderados quando utilizado as redes neurais artificiais.

Palavras-chave Affymetrix, Illumina, preditores, regressão linear múltipla, relacionamento genético

1. INTRODUÇÃO

A seleção genômica vem sendo utilizada em bovinos de leite e em algumas raças de bovinos de corte, o que permite que empresas de melhoramento genético possam obter vantagens econômicas com a utilização do valor genético genômico na seleção, por reduzir o tempo de obtenção de ganho genético (HAYES et al., 2009; MEUWISSEN et al., 2016). No entanto, a utilização de valores genéticos genômicos para a seleção de animais superiores geneticamente depende da confiabilidade de predição dos mesmos.

A acurácia da predição do valor genético genômico pode ser maior com o aumento do número de animais com registros fenotípicos e genotípicos, bem como com a utilização de painéis de alta densidade na população referência (MEUWISSEN, 2009). Entretanto, a obtenção de muitos animais genotipados em alta densidade pode ter elevado custo (CLARK et al., 2012). Uma alternativa seria a imputação, ferramenta utilizada para inferir marcadores não definidos em animais genotipados com painéis de baixa densidade, com a finalidade de reduzir custos da implementação de seleção genômica (VANRADEN et al., 2011; HUANG et al., 2012).

Para que dados genômicos imputados sejam utilizados em subsequentes estudos de associação e seleção genômica, é necessário que a imputação tenha sido implementada de forma que todos os animais envolvidos tenham seus genótipos inferidos com elevada acurácia (BADKE et al., 2014). A acurácia de imputação é afetada por diversos fatores, como o relacionamento genético entre os animais imputados e a população referência, o tamanho da população referência, a densidade dos painéis, o desequilíbrio de ligação e a frequência do alelo de menor

ocorrência (PEI et al. 2008; ZHANG et al., 2010). No entanto, a acurácia de imputação somente pode ser verificada quando marcadores contidos nos painéis de alta densidade são omitidos da análise de imputação para a obtenção do painel de baixa densidade de interesse.

As redes neurais têm sido utilizadas para estudos de modelagem, predição e classificação de dados, identificações de padrões, controle de processos, otimização e suporte à decisão em diversas áreas do conhecimento (SARLE, 1994). Estas são alternativas de predição, pelo fato de que as mesmas podem identificar relações não-lineares entre os preditores e as variáveis respostas, aprendendo de forma adaptativa as funções matemáticas que ligam esses dois conjuntos de variáveis devido a uma série de transformações por meio das “funções de ativação”. Essa característica faz com que as redes neurais não necessitem das pressuposições de independência e normalidade das variáveis exigidas nas relações lineares e não é preciso pré-especificar a função matemática a ser utilizada para modelar o conjunto de dados (TAM, 1991; GAUDART et al., 2004; PAO, 2008). Assim, este estudo teve como objetivo estudar estratégias para predição da acurácia de imputação em dados genômicos de bovinos da raça Nelore, utilizando redes neurais artificiais e regressão linear múltipla, de forma que animais que obtiverem baixos valores de acurácia predita sejam descartados e não façam parte do processo de seleção genômica.

2. MATERIAL E MÉTODOS

2.1. Descrição dos dados

Para o estudo da melhor estratégia de predição da acurácia de imputação, antes da implementação do mesmo, foram utilizados dados de 814 bovinos da raça

Nelore, em que 34 eram touros e 780 progênies desses touros. Esses dados foram cedidos pela Embrapa Pecuária Sudeste e foram genotipados com o BovineHD BeadChip (Illumina -700k), sendo que destes, 93 animais (23 touros e 70 progênies) também foram genotipados com o Axion Genome-Wide BOS 1 Array Plate (Affymetrix – 600k). O organograma das análises realizadas no presente estudo está descrito na Figura 1.

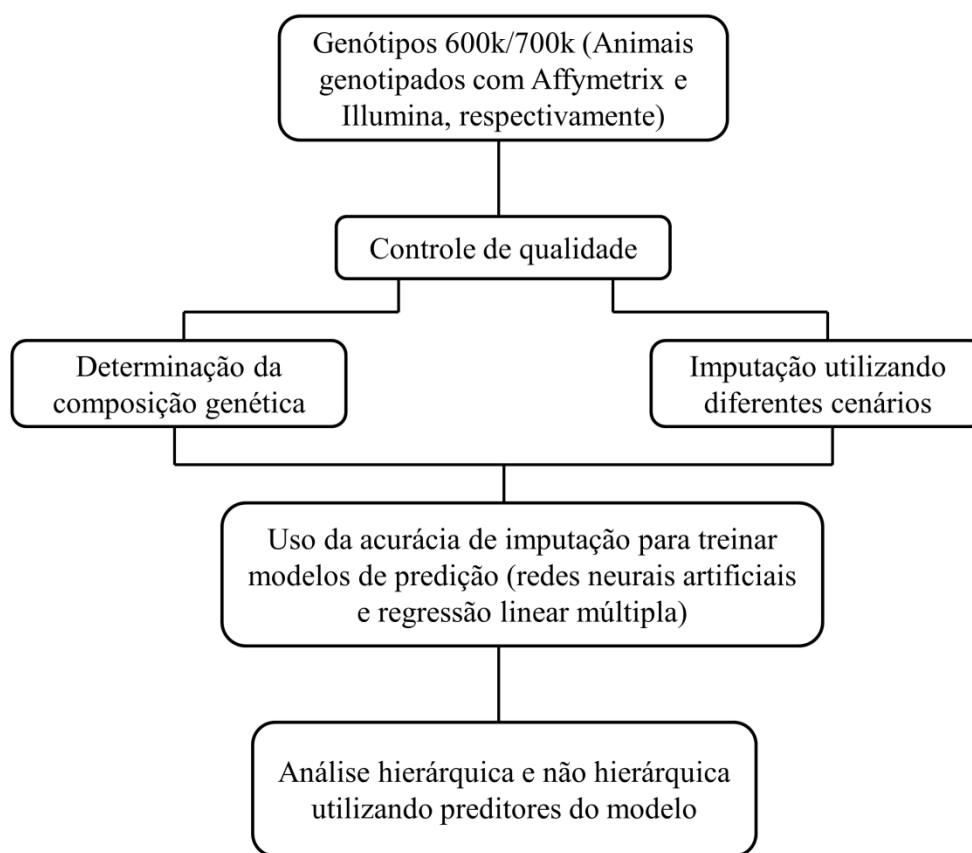


Figura 1. Organograma com a descrição da sequência das análises realizadas no presente estudo.

2.2. Controle de qualidade

Para todos os animais, o controle de qualidade dos genótipos foi feito com a finalidade de remover erros de genotipagem excluindo SNPs localizados em regiões não autossômicas, com posições desconhecidas, com desvios significativos

($p < 0,00001$) do equilíbrio de Hardy-Weinberg, com taxa de leitura (“call rate”) menor que 0,98 e amostras que apresentaram “call rate” menor que 0,90. Após o controle de qualidade, restaram 809 animais com 509.107 SNPs para o painel Illumina e, 93 animais com 427.875 SNPs para o painel Affymetrix. Assim, após aplicado o controle de qualidade, ambos os painéis apresentaram 56.646 SNPs em comum.

2.3. Determinação da composição genética dos animais

A identificação da composição do grupo genético dos animais foi realizada com o auxílio do programa computacional Admixture (ALEXANDER et al., 2009). Para essa análise foram determinados diferentes grupos, extraindo do programa um número “K” de colunas para cada animal, sendo que em cada linha apresentava-se a proporção daquele animal para aquele grupo genético (representado pela coluna). O número de grupos genéticos (K) foi escolhido para este estudo variando de 2 a 5 grupos, uma vez que quando foram considerados valores maiores de K, as proporções de cada grupo genético nos animais passaram a ser muito pequenas, dificultando a caracterização dos grupos.

2.4. Descrição dos cenários de imputação

As imputações foram realizadas utilizando o programa FImpute v.2.2 (SARGOLZAEI et al., 2014). Para os diferentes cenários de imputação foi realizada a redução dos painéis de maior densidade para painéis de menor densidade comerciais, omitindo os SNPs para a população a ser imputada, obtendo assim painéis com 3k, 8k, 20k e 50k. Os SNPs em comum entre os painéis de alta densidade Illumina e Affymetrix (56.646 SNPs) foram utilizados para customizar de duas formas diferentes painéis de 20k e 50k SNPs.

Na primeira maneira de customizar os painéis (20kCust1 e 50kCust1), considerou-se os SNPs em comum com maior desequilíbrio de ligação com os SNPs presentes no painel comercial (20k e 50k). Para a segunda maneira (20kCust2 e 50kCust2), considerou-se a mesma situação da primeira, porém com o adicional de que os SNPs que apresentavam MAF (“Minor Allele Frequency”) menor que 0,09 foram descartados. Além disso, foi realizada uma combinação entre os SNPs presentes no painel Illumina e Affymetrix para os 23 touros que foram genotipados com ambos os painéis, formando um painel contendo 880.336 SNPs (800k).

As imputações foram realizadas considerando quatro cenários na população referência: (a) 23 touros, (b) cinco touros, (c) 100 animais e (d) 40 animais, escolhidos aleatoriamente (Tabela 1). Assim, foi possível obter tamanhos diferentes da população referência e diferentes relacionamentos genéticos entre a população de imputação e referência. A acurácia de imputação por animal foi estimada pela correlação simples de Pearson entre o genótipo imputado e o observado (COR).

Tabela 1. Número de animais utilizados na população referência para as imputações a partir dos painéis de menor densidade para os de maior densidade em seus respectivos cenários e que foram incluídas nas análises de redes neurais artificiais.

Menor densidade	Maior densidade					
	50k	600k	700k	700k	700k	800k
3k	40	-	-	-	-	23
8k	100	-	40	-	-	23
20k	-	23	23	5	100	23
20kCust1	-	23	23	5	-	23
20kCust2	-	23	23	5	-	23
50k	-	23	23	5	100	23
50kCust1	-	23	23	5	-	-
50kCust2	-	23	23	5	-	-

2.5. Redes neurais artificiais: estrutura e funcionamento

As estruturas das redes neurais “multilayer perceptron” (MLP) são compostas por camadas alinhadas de neurônios, em que os neurônios de uma camada são interligados com os da camada adjacente (CHENG; TITTERINGTON, 1994; PALIWAL; KUMAR, 2009). As MLP são redes neurais com aprendizado supervisionado, ou seja, a extração do conhecimento do banco de dados e o mapeamento das variáveis de entrada com as de saída são feitos por meio do algoritmo “backpropagation”, que indica para a MLP o quão perto ou distante as previsões estão dos valores reais, permitindo que a MLP possa melhorá-las por meio de mudanças nos pesos sinápticos (WARNER; MISRA, 1996). O desenvolvimento matemático do algoritmo “backpropagation” foi descrito por Rumelhart et al. (1986).

Segundo Warner e Misra (1996), cada neurônio de entrada conecta-se a cada neurônio da camada intermediária adjacente via pesos sinápticos. A entrada em um neurônio é uma soma ponderada das saídas dos neurônios conectados ao mesmo. Nos neurônios da camada intermediária, cada unidade transforma o valor gerado na função sináptica por meio da função de ativação. Uma função de ativação comumente usada é a função logística. A resposta ativada no neurônio da camada escondida é propagada para o neurônio da camada de saída da MLP como:

$$a_l = \sum_{k=1}^m w_k g_k \left(b_k + \sum_{j=1}^n w_{kj} x_j \right) + b_l$$

em que a_l é o valor de saída no neurônio da camada de saída x_j são as variáveis normalizadas ponderadas pelos pesos sinápticos w_{kj} , entre o neurônio j da camada de entrada e k da camada intermediária, e somadas ao intercepto b_k , g_k é a

função de ativação e b_1 é o neurônio intercepto da camada de saída (HAYKIN, 1999). Após isso, a_1 é ativado pela função de ativação $g(a_1)$, que é uma função de ativação linear.

Segundo Cheng e Titteringto (1994), o nível de desempenho de predição de uma rede neural para um conjunto de dados amplo, ou seja, não só para aqueles que foram usados em sua construção, é conhecido como capacidade de generalização da resposta. Essa capacidade tem desempenho reduzido quando há “overfitting”, que ocorre quando a rede neural possui alto desempenho de predição em sua construção, mas não possui o mesmo desempenho para prever futuras respostas. Isso pode ser causado por má escolha da estrutura da rede neural. Para evitar o “overfitting”, a rede neural é construída em duas fases: fase de treinamento e de teste.

2.6. Redes neurais artificiais e regressão linear múltipla na predição de acurácia de imputação

Foram utilizadas redes neurais MLP, as quais são aquelas com aprendizado supervisionado, com duas camadas intermediárias de cinco e dois neurônios, respectivamente. Com isso, foi possível obter a predição da acurácia de imputação. Diferentes combinações de quantidades de neurônios para cada camada foram previamente testadas.

No presente estudo, a supervisão da rede ocorreu por meio da inclusão da acurácia de imputação calculada (COR) para cada cenário. Na camada de entrada das redes neurais foram utilizadas sete preditores (Figura 2): (P1) proporção do animal para cada grupo genético; (P2) número de animais da população referência; (P3) número de SNPs do painel de baixa densidade; (P4) número de SNPs do painel

de alta densidade; (P5) média do desequilíbrio de ligação (DL) entre os SNPs presentes no painel de baixa densidade; (P6) média do DL entre os SNPs presentes no painel de baixa densidade com os SNPs presentes no de alta densidade e ausentes no de baixa densidade; (P7) média de relacionamento genético, calculado pela matriz genômica, entre cada animal a ser imputado com os cinco animais presentes na população referência que fossem mais relacionados com este animal.

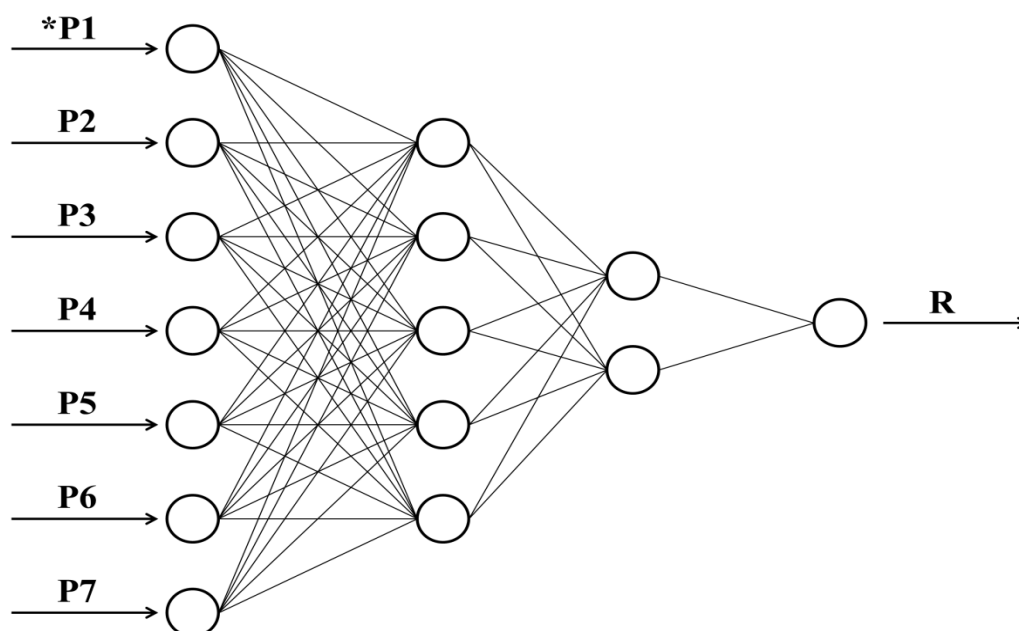


Figura 2. Estrutura da rede neural quando utilizadas todas as informações na camada de entrada em que os preditores foram (P1) proporção do animal para cada grupo genético, *variando de zero e de dois a cinco grupos; (P2) número de animais da população referência; (P3) número de SNPs do painel de baixa densidade; (P4) número de SNPs do painel de alta densidade; (P5) média do desequilíbrio de ligação do painel de baixa densidade e (P6) média do DL entre os SNPs presentes no painel de baixa densidade com os SNPs presentes no de alta densidade e ausentes no de baixa densidade; (P7) média de relacionamento genético entre cada animal a ser imputado com os cinco animais presentes na população referência que fossem mais relacionados com este animal; (R) acurácia de imputação predita.

As análises de DL foram realizadas utilizando o programa computacional PLINK (PURCELL et al., 2007), em que foram consideradas as medidas de r^2 , proposta por Hill e Robertson (1968). As análises com redes neurais artificiais foram realizadas por meio do pacote “neuralnet” (FRITSCH et al., 2012), do programa computacional R (R Core Team, 2018).

Os mesmos preditores e variável-resposta foram utilizados no modelo de regressão linear múltipla, a fim de prever a acurácia de imputação. Assim, foi possível comparar se, em um modelo linear, obteve-se a mesma habilidade de predição que em um modelo não linear. As análises com modelos de regressão múltipla foram realizadas por meio do programa computacional R (R Core Team, 2018).

A habilidade de predição dos modelos de regressão múltipla e das redes neurais artificiais foi calculada por meio da correlação entre os valores de acurácia de imputação observados nas análises de imputação (valores considerados verdadeiros), com os valores preditos pelos dois modelos. O quadrado médio do erro também foi utilizado para avaliar a habilidade de predição dos dois métodos. As análises de redes neurais artificiais foram repetidas por três vezes. Assim, as correlações e quadrado médio dos erros apresentados são médias das três repetições.

2.7. Modelos utilizados para predição de acurácia de imputação

A fim de identificar o conjunto de preditores que melhor descreveria a variável acurácia de imputação, foram consideradas diferentes situações, variando os

preditores utilizados e o número e a composição dos animais escolhidos para compor a população de treinamento.

Inicialmente, modificou-se o número de grupos genéticos considerados nas análises, sendo zero (sem considerar o preditor) e variando de dois até cinco grupos genéticos, em que os demais preditores foram considerados. Em seguida, foram realizadas análises em que cada preditor foi excluído unicamente de cada análise. Outras duas análises foram realizadas com a exclusão simultânea das informações de DL (P5 e P6) e apenas considerando as informações de número de animais na população referência (P2) e as densidades dos painéis (P3 e P4). Estas análises foram realizadas utilizando 60% dos animais (8.817 observações) escolhidos aleatoriamente para compor a população de treinamento.

Adicionalmente, também foram consideradas diferentes formas de população de treinamento para os modelos de redes neurais artificiais e regressão linear múltipla. Na primeira, consideraram-se animais cujo preditor de relacionamento genético (P7) era menor que 0,11147 como população de treinamento (9.234 observações). Na segunda, considerou-se aleatoriamente a mesma quantidade de observações (9.234 observações) para a população de treinamento. Na terceira, considerou-se para a população de treinamento 9.234 observações de animais que possuíam em sua composição genética a proporção maior do que 0,4 para um de dois grupos genéticos quando K foi igual a 2.

A escolha do valor de relacionamento genético menor que 0,11147 e composição genética com proporção maior do que 0,4 foi determinada devido a necessidade de manter o mesmo número de observações na população treinamento, para que a comparação entre diferentes formas de escolha de

população treinamento fosse justa. Após obter as predições, foi realizado um teste Tukey (significância de 5%) para comparar as correlações entre os valores de acurácia de imputação observados nas análises com os valores preditos pelos modelos utilizando diferentes combinações entre preditores e população de treinamento.

2.8. Análise hierárquica e não hierárquica

Para identificar a influência de cada preditor na eficiência da predição, separou-se, por análise hierárquica e não hierárquica, as observações utilizadas na população de teste em grupos. Então foi calculada a média do quadrado médio dos erros para cada grupo e assim foi possível comparar os grupos, identificando se o perfil dos preditores observado para cada grupo prejudicou ou favoreceu a predição. Assim, após verificar qual situação promoveu a predição mais adequada da acurácia de imputação, as informações de cada preditor foram padronizadas e utilizadas para realizar análise hierárquica e identificar a possível quantidade de grupos homogêneos. A divisão em grupos foi realizada utilizando o número de clusters sugerido pelo pacote NbCust do programa computacional R (R Core Team, 2018).

Depois do estabelecimento destes grupos, realizou-se a análise de agrupamento não-hierárquico pelo método *K-means*. A dissimilaridade entre os indivíduos foi medida pela distância euclidiana. Então, a média do quadrado médio do erro foi calculada para cada grupo, a fim de realizar comparação com os demais possíveis grupos identificados na análise. As análises hierárquicas e não-hierárquicas e a comparação entre as médias do quadrado médio dos erros por grupo foram realizadas para compreender o efeito dos preditores em cada grupo.

3. RESULTADOS E DISCUSSÃO

3.1. *Predição da acurácia de imputação*

O valor da correlação para os dois modelos com o quadrado médio dos erros, considerando seis preditores (P2 a P7) e excluindo a informação de proporção de grupo genético (P1), assim como quando considerado os sete preditores em que K (número de grupos genéticos) variou de dois a cinco, está descrito na Tabela 2.

O maior valor observado para a correlação entre valores preditos e os reais pela rede neural artificial foi igual a 0,919 e pelo modelo de regressão linear múltipla foi de 0,813. De maneira geral, as correlações observadas entre os valores preditos e os reais pelas redes neurais artificiais foram maiores que as correlações observadas utilizando o modelo de regressão linear múltipla. O mesmo observa-se para o quadrado médio dos erros, em que o modelo de rede neural artificial proporcionou menores valores quando comparado ao modelo de regressão linear múltipla. Não foi observada grande variação nos resultados das três repetições para as redes neurais artificiais.

A ausência da informação de proporção do animal para cada grupo genético (P1) promoveu redução na correlação (Tabela 2), afetando a predição para ambos os modelos. Conforme aumentou o número de grupos genéticos (K) em que a população de animais foi dividida, houve pequena alteração da correlação (Tabela 2), este fato pode ter ocorrido porque os animais são de apenas uma raça. Assim, a informação de grupo genético mostrou-se importante para a predição da acurácia de imputação; no entanto, a divisão da população em apenas dois grupos genéticos (K=2) não apresentou diferença significativa quando utilizado mais de dois grupos

genéticos para o modelo de redes neurais artificiais, o que indica ser eficiente para a predição da acurácia de imputação.

Tabela 2. Estimativas das correlações (Cor), desvios-padrão (entre parênteses) e quadrado médio dos erros (QME) dos valores de acurácia de imputação observados nas análises com os valores preditos pelos modelos de redes neurais artificiais (NN) e de regressão linear múltipla (LM) utilizando todos os preditores, variando o número de grupos genéticos (K) e com ausência da informação de grupo genético.

Número de grupos genéticos (P1)	NN*		LM	
	Cor	QME	Cor	QME
2	0,908 (0,008) ^{Aa}	0,003	0,654 (0,000) ^B	0,010
3	0,919 (0,000) ^{Aa}	0,001	0,806 (0,000) ^B	0,006
4	0,900 (0,014) ^{Aa}	0,003	0,639 (0,000) ^B	0,010
5	0,916 (0,000) ^{Aa}	0,003	0,813 (0,000) ^B	0,006
Ausência de grupo genético	0,750 (0,003) ^{Ab}	0,008	0,627 (0,000) ^B	0,011

*Os resultados mostrados são a média de três repetições e entre parênteses o desvio-padrão observado para as três repetições. Letras maiúsculas diferentes nas colunas indicam diferença estatística pelo teste de Tukey ($P < 0,05$); A, B: Letras minúsculas diferentes nas linhas indicam diferença estatística pelo teste Tukey ($P < 0,05$); a, b Letras iguais não diferem entre si significativamente.

A informação de proporção do grupo genético pode ter contribuído para a predição da acurácia de imputação porque o programa utilizado para realizar as imputações (Fimpute), além de utilizar informações de parentesco, utiliza informações de haplótipos e, animais do mesmo grupo genético, apresentam haplótipos semelhantes e, conseqüentemente, tendem a apresentar acurácia de imputação semelhante. Estudos realizados por Xiang et al. (2015) com raças puras e cruzadas de suínos, evidenciaram que a proporção de haplótipos compartilhados entre as populações de referência e de validação pode fornecer performance apropriada da imputação. Assim, se os haplótipos de um animal pertencente a um

grupo genético estão bem representados na população referência, implica que os haplótipos de todo grupo genético estão representados na população referência, proporcionando maior acurácia de imputação para este animal e todo o seu grupo genético. Este fato pode ser observado quando a imputação é estudada em animais cruzados, pois estudos observaram que a maior representação racial na população referência melhora a acurácia de imputação para estes animais (VENTURA et al., 2014; MOGHADDAR et al., 2015).

Devido à análise considerando a proporção de grupo genético com apenas $K=2$ mostrar-se eficiente para prever a acurácia de imputação pelo modelo de redes neurais artificiais, ter menor custo computacional e não apresentar diferença significativa quando utilizado mais que dois grupos genéticos, o modelo considerando P1 com $K=2$ foi considerado como o modelo completo. Então, foi retirado cada um dos outros seis preditores (P2 a P7) e foi verificado o quadrado médio do erro e o valor da correlação para o modelo de redes neurais artificiais e o modelo linear múltipla (Tabela 3).

Comparado ao modelo completo que apresentou correlação igual a 0,908, a remoção do preditor que indicava o número de animais da população referência (P2) promoveu redução da correlação para o modelo de redes neurais artificiais e para o modelo de regressão linear múltipla (Tabela 3). Isto pode ter ocorrido porque para as imputações foram utilizados poucos animais na população referência. Hozé et al. (2013), estudando diversas raças de bovinos de leite e de corte, relataram que para bovinos de corte a taxa de erro de imputação decresceu com o aumento do número de animais na população referência até aproximadamente 400 animais e que a partir deste número a taxa de erro se estabilizou.

Embora o relacionamento genético dos animais seja elevado nessa população (dados não demonstrados), os cenários de imputação utilizaram número inferior de 400 animais, o que pode implicar em maior variabilidade nos resultados da acurácia de imputação para os diferentes cenários. Sendo assim, o número de animais na população referência pode ser uma informação importante para a aprendizagem da rede, o que auxilia para a correta predição para a acurácia de imputação.

A remoção do preditor que indicava o relacionamento genético entre cada animal a ser imputado com os cinco animais mais relacionados com o mesmo da população referência (P7) reduziu a correlação entre o valor predito de acurácia e a calculada (Tabela 3), quando comparado ao modelo completo ($Cor=0,908$). Este fato pode ter ocorrido porque o aumento do relacionamento genético entre animais da população a ser imputada e da população referência pode ocasionar aumento da acurácia de imputação. Assim, este preditor fornece importante informação para a predição da acurácia de imputação.

A redução da correlação obtida com a remoção do número de animais da população referência (P2) e com a remoção do relacionamento genético (P7 - Tabela 3) pode ser explicada pelas diferenças da acurácia de imputação de acordo com a variação no tamanho populacional e parentesco entre os animais da população referência e imputação. Zhang e Druet (2010) identificaram que a taxa de erro de imputação reduziu quando parentes próximos (pais e avós) eram genotipados em maior densidade e incluídos na população referência para bovinos de leite. A influência do número de animais na população referência e a relação entre animais da população referência e da imputação sobre a acurácia de

imputação, também foram relatadas para demais espécies, como em ovinos (HAYES et al., 2011) e aves (HEIDARITABAR et al., 2015).

Tabela 3. Estimativas das correlações (Cor), desvios-padrão (entre parênteses) e quadrado médio dos erros (QME) dos valores de acurácia de imputação observados nas análises com os valores preditos pelos modelos de redes neurais artificiais (NN) e de regressão linear múltipla (LM), com situações em que os preditores (de P1 a P7) foram excluídos de cada análise.

Preditores ausentes	NN*		LM	
	Cor	QME	Cor	QME
P2	0,883 (0,004) ^{Aab}	0,004	0,600 (0,000) ^B	0,010
P3 e P4	0,910 (0,002) ^{Aa}	0,003	0,639 (0,000) ^B	0,010
P5	0,910 (0,002) ^{Aa}	0,003	0,641 (0,000) ^B	0,010
P6	0,901 (0,007) ^{Aab}	0,003	0,654 (0,000) ^B	0,010
P5 e P6	0,904 (0,011) ^{Aab}	0,003	0,637 (0,000) ^B	0,010
P7	0,867 (0,026) ^{Ab}	0,005	0,547 (0,000) ^B	0,012
P1, P5, P6 e P7	0,685 (0,008) ^{Ac}	0,009	0,494 (0,000) ^B	0,013
Modelo Completo (sem preditores ausentes e K=2)	0,908 (0,008) ^{Aa}	0,003	0,654 (0,000) ^B	0,010

Preditores do modelo completo – (P1) proporção do animal para cada grupo genético; (P2) número de animais da população referência; (P3) número de SNPs presentes no painel de alta densidade e (P4) de baixa densidade; (P5) média do desequilíbrio de ligação do painel de baixa densidade e (P6) média do DL entre os SNPs presentes no painel de baixa densidade com os SNPs presentes no de alta densidade e ausentes no de baixa densidade; (P7) média de relacionamento genético entre cada animal a ser imputado com os cinco animais presentes na população referência que fossem mais relacionados com este animal. *Os resultados mostrados são a média de três repetições e entre parênteses o desvio-padrão observado para as três repetições. Letras maiúsculas diferentes nas colunas indicam diferença estatística pelo teste de Tukey ($P < 0,05$); A, B: Letras minúsculas diferentes nas linhas indicam diferença estatística pelo teste Tukey ($P < 0,05$); a, b Letras iguais não diferem entre si significativamente.

A retirada dos preditores que continham informação de painel de alta e baixa densidade (P3 e P4) foi realizada simultaneamente por serem informações

complementares. Pei et al. (2008) afirmaram que o DL destaca-se dentre os diversos fatores que afetam a acurácia de imputação. Adicionalmente, estes mesmos autores ressaltaram a relação entre a densidade de marcadores e o DL, em que normalmente painéis mais densos proporcionam maiores DL. Este fato pode estar relacionado com a resposta observada pelas redes neurais artificiais, em que a correlação removendo P3 e P4 apresentou pequena alteração comparada à correlação com a remoção da informação de DL (Tabela 3). Possivelmente, este resultado pode ter ocorrido, pois apenas umas dessas informações já é o suficiente para a aprendizagem das redes neurais artificiais. Embora ainda pequena, houve uma redução quando a média do DL entre os SNPs presentes no painel de baixa densidade com os SNPs ausentes foi removida, quando comparado ao modelo completo, modelos removendo P3 e P4 conjuntamente e removendo apenas P5. Isto pode ter ocorrido devido à utilização de painéis customizados nas imputações destes animais.

Na situação em que foram removidos quatro dos sete preditores (P1, P5, P6 e P7), mantendo apenas o número de animais da população referência (P2), número aproximado de SNPs do painel de baixa densidade (P3) e número aproximado de SNPs do painel de alta densidade (P4) promoveu importante redução na correlação entre o valor de acurácia de imputação calculado e predito (Tabela 3), possivelmente pela remoção de preditores que mostraram-se importantes na aprendizagem das redes, como proporção de grupo genético e relacionamento genético. Mais estudos acrescentando ou removendo preditores ou incluindo diferentes formas de imputação são importantes para estabelecer a melhor forma de utilização desta predição.

A escolha de animais para a população de treinamento diferente da forma aleatória mostrou-se ineficiente na predição dos dois modelos (Tabela 4). Este resultado pode ser observado porque os preditores escolhidos para selecionar a população de treinamento (raça e relacionamento genético) mostraram-se importantes para melhorar a predição. Dessa forma, ao utilizar observações com semelhança dentro de algum destes preditores, dificulta-se o estabelecimento de um modelo de regressão linear múltipla que possa ser aplicado de forma eficiente para as demais observações. Da mesma maneira, populações de treinamento formadas utilizando informações de raça ou relacionamento genético, promovem o aprendizado de um padrão pelas redes neurais artificiais, que não se aplica adequadamente quando é testada em outros padrões.

Tabela 4. Estimativas das correlações (Cor), desvios-padrão (entre parênteses) e quadrado médio dos erros (QME) dos valores de acurácia de imputação observados nas análises com os valores preditos pelos modelos de redes neurais artificiais (NN) e de regressão linear múltipla (LM), utilizando diferentes formas para a escolha dos animais que iriam compor a população de treinamento.

Forma de escolha dos animais	NN*		LM	
	Cor	QME	Cor	QME
Relacionamento < 0,11147	0,631 (0,068) ^{Ab}	0,006	0,380 (0,000) ^B	0,023
Aleatoriamente	0,938 (0,007) ^{Aa}	0,002	0,700 (0,000) ^B	0,008
Proporção do grupo genético 1 maior que 0,4	0,671 (0,005) ^{Ab}	0,013	0,580 (0,000) ^B	0,016

*Os resultados mostrados são a média de três repetições e entre parênteses o desvio-padrão observado para as três repetições. Letras maiúsculas diferentes nas colunas indicam diferença estatística pelo teste de Tukey ($P < 0,05$); A, B: Letras minúsculas diferentes nas linhas indicam diferença estatística pelo teste Tukey ($P < 0,05$); a, b Letras iguais não diferem entre si significativamente.

3.2. Análise hierárquica e não hierárquica

Em função dos resultados, as análises hierárquica e não hierárquica foram realizadas utilizando a predição considerando a proporção de cada animal nos grupos genéticos (P1) com $K=2$ e incluindo os demais preditores estudados (modelo completo), com a população de treinamento de 60% dos animais. Realizou-se a análise hierárquica e então com o auxílio do pacote NbClust do programa computacional R (R Core Team, 2018), foi determinado que a população deveria ser dividida em dois grupos de similaridade (Figura 3). Então realizou-se a análise não-hierárquica, considerando os dois grupos (Figura 4).

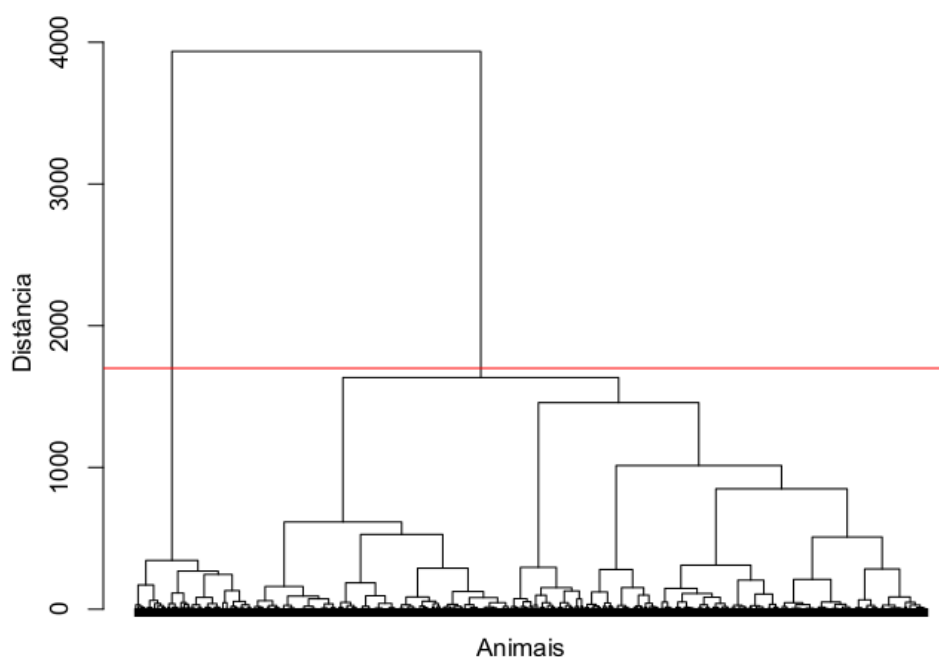


Figura 3. Dendrograma obtido pela análise hierárquica das observações inseridas nas redes neurais artificiais para a análise utilizando o modelo completo que inclui informação padronizada de proporção de grupo genético; número de animais considerado na população referência para imputação; número de SNPs no painel de alta densidade e de baixa densidade dos painéis; média do desequilíbrio de ligação do painel de baixa densidade e média do DL entre os SNPs presentes no painel de baixa densidade com os SNPs presentes no de alta densidade e ausentes no de baixa densidade e média de relacionamento genético, entre cada animal a ser imputado com os cinco animais presentes na população referência que fossem mais relacionados com este animal.

Observou-se que a média do Grupo 2 foi baixa para número de animais na população referência (P2) e para o relacionamento genético entre os animais da população de imputação e os cinco mais aparentados da população referência (P7 - Figura 4). Esses preditores (P2 e P7) mostraram-se importantes para predição satisfatória da acurácia de imputação (Tabela 3). No entanto, o Grupo 1 apresentou média de quadrado médio do erro igual a 0,0038, valor semelhante ao Grupo 2, que apresentou média de 0,0034, indicando que a elevada média para os painéis de alta e baixa densidade e o elevado DL pode ter colaborado para melhorar a predição da acurácia de imputação para as observações que compunham este grupo. É importante ressaltar que, como observado nas diferentes análises para as redes neurais artificiais, as informações de densidade e DL fornecem para o modelo informações semelhantes.

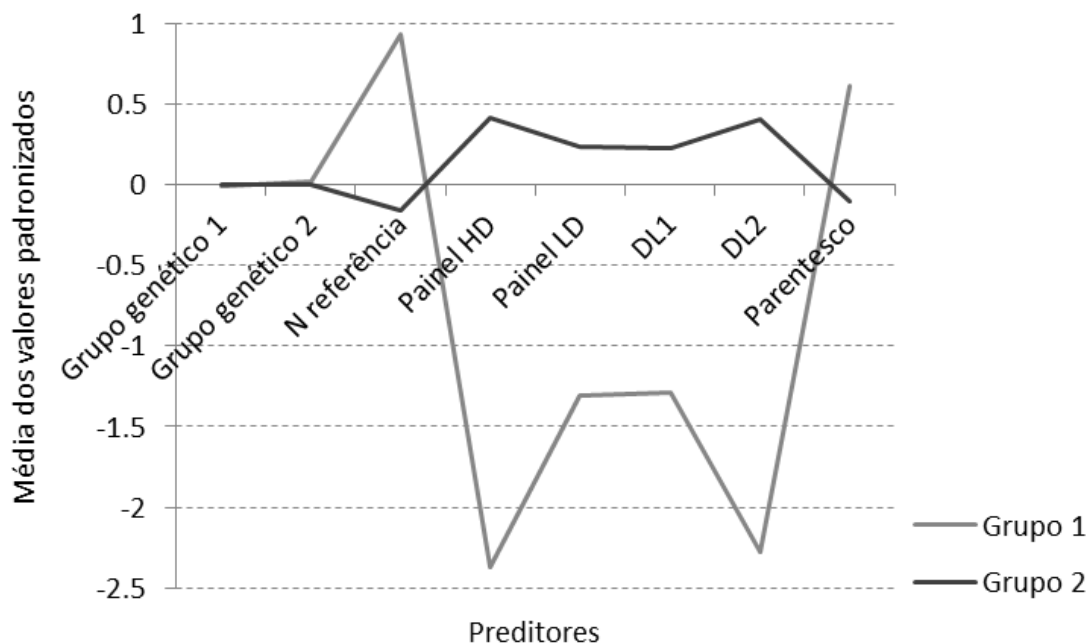


Figura 4. Distribuição pela análise não-hierárquica das médias dos valores padronizados para cada preditor do modelo completo para os dois grupos estabelecidos. (Grupo genético 1) proporção do animal para um grupo quando a população foi dividida em dois grupos genéticos; (Grupo genético 2) proporção do animal para o outro grupo quando a população foi dividida em dois grupos genéticos; (N referência) número de animais considerado na população referência para imputação; (Painel HD) número de SNPs no painel de alta densidade e (Painel LD) de baixa densidade dos painéis; (DL1) média do desequilíbrio de ligação do painel de baixa densidade e (DL2) média do DL entre os SNPs presentes no painel de baixa densidade com os SNPs presentes no de alta densidade e ausentes no de baixa densidade; (Parentesco) média de relacionamento genético, entre cada animal a ser imputado com os cinco animais presentes na população referência que fossem mais relacionados com este animal.

4. CONCLUSÃO

As redes neurais mostraram-se mais eficientes para prever a acurácia de imputação comparado ao modelo de regressão linear múltipla, podendo ser utilizadas com esta finalidade. Os preditores que envolveram DL forneceram informações similares à informação de densidade dos painéis para predição da

acurácia de imputação, podendo ser desconsiderados quando utilizado as redes neurais artificiais.

5. RERÊNCIAS

ALEXANDER, D. H.; NOVENBRE, J.; LANGE, K. Fast model-based estimation of ancestry in unrelated Individuals. **Genome Research**, San Francisco, v. 19, p. 1655–1664, 2009.

BADKE, Y. M.; BATES, R. O.; ERNST, C. W.; FIX, J.; STEIBEL, J. P. Accuracy of Estimation of Genomic Breeding Values in Pigs Using Low-Density Genotypes and Imputation. **G3**, Toronto, v.4, p. 623-631, 2014.

BROWNING, B.L.; BROWNING, S.R. A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. **The American Journal of Human Genetics**, Houston, v. 84, p. 210-223, 2009.

CHENG, B.; TITTERINGTON, D. M. Neural networks: a review from a statistical perspective. **Statistical Science**, Piscataway, v. 9, p. 2-54, 1994.

CLARK, S. A.; HICKEY, J. M.; DAETWYLER, H. D.; VAN DER WERF, J. H. J. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. **Genetics Selection Evolution**, London, v. 44, n. 4, 2012.

DAETWYLER, H. D.; CALUS, M. P. L.; PONG-WONG, R.; DE LOS CAMPOS, G.; HICKEY, J. M. Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting, and Benchmarking. **Genetics**, Bethesda, v. 193, p. 347-365, 2013.

FRITSCH, S.; GUENTHER, F.; SULING, M. Neuralnet: Training of neural networks. R package version 1.32; 2012. Available: <http://CRAN.R-project.org/package=neuralnet>. Accessed 18 August 2016.

GAUDART, J.; GIUSIANO, B.; HUIART, L. Comparison of the performance of multi-layer perceptron and linear regression for epidemiological data. **Computational Statistics & Data Analysis**, California, v. 44, p. 547-570, 2004.

GODDARD, M. E.; HAYES, B. J. Mapping genes for complex traits in domestic animals and their use in breeding programmes. **Nature Reviews Genetics**, New York, v. 10, p. 381-391, 2009.

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

HAYES, B. J.; BOWMAN, P. J.; DAETWYLER, H. D.; KIJAS, J. W.; VAN DER WERF, J. H. J. Accuracy of genotype imputation in sheep breeds. **Animal Genetics**, Malden, v. 43, p. 72-80, 2011.

HAYKIN S. **Neural Networks a comprehensive foundation**. 2 ed. UpperSaddle River, NJ: Prentice Hall International, 1999.

HEIDARITABAR, M.; CALUS, M. P. L.; VEREIJKEN, A.; GROENEN, M. A. M.; BASTIAANSEN, J. W. M. Accuracy of imputation using the most common sires as reference population in layer chickens. **BMC Genetics**, London, v. 16, p. 101, 2015.

HILL, W.; ROBERTSON, A. Linkage disequilibrium in finite populations. **Theoretical and Applied Genetics**, Heidelberg, v. 38, n. 6, p. 226 - 231, 1968.

HOZÉ, C.; FOUILLOUX, M. N.; VENOT, E.; GUILLAUME, F.; DASSONNEVILLE, R.; FRITZ, S.; DUCROCQ, V.; PHOCAS, F.; BOICHARD, D.; CROISEAU, P. High-density marker imputation accuracy in sixteen French cattle breeds. **Genetics Selection Evolution**, London, v. 45, n. 33, 2013.

HUANG, Y.; HICKEY, J. M.; CLEVELAND, M. A.; MALTECCA, C. Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. **Genetics Selection Evolution**, London, v. 44, n. 25, 2012.

MEUWISSEN, T. H. Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. **Genetics Selection Evolution**, London, v. 41, n. 35, 2009.

MEUWISSEN, T.; HAYES, B.; GODDARD, M. Genomic selection: A paradigm shift in animal breeding. **Animal Frontiers**, Champaign, v. 6, n. 1, p. 6-14, 2016.

MOGHADDAR, N.; GORE, K. P.; DAETWYLER, H. D.; HAYES, B. J.; VAN DER WERF, J. H. J. Accuracy of genotype imputation based on random and selected reference sets in purebred and crossbred sheep populations and its effect on accuracy of genomic prediction. **Genetics Selection Evolution**, London, v. 47, n. 97, 2015.

PALIWAL, M.; KUMAR, U. A. Neural networks and statistical techniques: a review of applications. **Expert Systems with Applications**, Amsterdam, v. 36, p. 2-17, 2009.

PAO, H. T. A comparison of neural networks and multiple regression analysis in modeling capital structure. **Expert Systems with Applications**, Amsterdam, v. 35, p. 720-727, 2008.

PEI, Y.; LI, J.; ZHANG, L.; PAPASIAN, C. J.; DENG, H. Analyses and Comparison of Accuracy of Different Genotype Imputation Methods. **Plos One**. São Francisco, v. 3, n. 10, 2008.

PURCELL, S.; NEALE, B.; TODD-BROWN, K.; THOMAS, L.; FERREIRA, M. A. R.; BENDER, D.; MALLER, J.; SKLAR, P.; BAKKER, P. I. W.; DALY, M. J.; SHAM, P. C. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. **The American Journal of Human Genetics**, Houston, v. 81, p. 559 – 575, 2007.

R Core Team. R: A language and environment for statistical computing. Viena: R Foundation for Statistical Computing; 2016. Available: <http://www.R-project.org/>. Acessado: 10 Agosto 2018.

RIGGS, E. R.; JACKSON, L.; MILLER, D. T.; VAN VOOREN, S. Phenotypic Information in Genomic Variant Databases Enhances Clinical Care and Research:

The ISCA Consortium Experience. **Human Mutation**, Hoboken, v. 33, n. 5, p. 787-796, 2012.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning Representations by Back-Propagating Errors. **Nature**, New York, v. 323, n. 6088, p. 533-536, 1986.

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

SARLE, W. S. Neural networks and statistical models. In: Proceedings of the Nineteenth Annual SAS Users Group International Conference. 1994, Cary, NC. Anais... Cary, NC.: SAS Institute Inc., 1994, 1538-1550.

TAM, K. Y. Neural network models and the prediction of bank bankruptcy. **Omega**, Philadelphia, v. 19, n. 5, p. 429-445, 1991.

VANRADEN, P. M.; O'CONNELL, J. R.; WIGGANS, G. R.; WEIGEL, K. A. Genomic evaluations with many more genotypes. **Genetics Selection Evolution**, London, v. 43, p. 10-20, 2011.

VENTURA, R. V.; LU, D.; SCHENKEL, F. S.; WANG, Z.; LI, C.; MILLER, S. P. Impact of reference population on accuracy of imputation from 6K to 50K single nucleotide polymorphism chips in purebred and crossbred beef cattle. **Journal of Animal Science**, Champaign, v. 92, p. 1433-1444, 2014.

WARNER, B.; MISRA, M. Understanding neural networks as statistical tools. **The American Statistician**, New York, v. 50, n. 4, p. 284-293, 1996.

XIANG, T.; MA, P.; OSTERSEN, T.; LEGARRA, A.; CHRISTENSEN, O. F. Imputation of genotypes in Danish purebred and two-way crossbred pigs using low-density panels. **Genetics Selection Evolution**, London, v. 47, n. 54, 2015.

ZHANG, Z.; DRUET, T. Marker imputation with low-density marker panels in Dutch Holstein cattle. **Journal of Dairy Science**, Champaign, v. 93, p. 5487-5494, 2010.

CAPÍTULO 4 - Estudo de endogamia utilizando segmentos de homozigose em bovinos Nelore

RESUMO – Elevado nível de endogamia pode reduzir a viabilidade e fertilidade dos animais, resultante da depressão endogâmica. A endogamia pode ser monitorada utilizando o registro de pedigree. No entanto, a falta de informações e erros de registros podem prejudicar a acurácia dos resultados obtidos. A utilização dos marcadores de polimorfismo de nucleotídeo único (SNPs) pode auxiliar na obtenção do coeficiente de endogamia com base nas informações genômicas. Dessa maneira, o objetivo deste trabalho foi estudar os segmentos de homozigose (ROH) e, com isso, o coeficiente de endogamia presente em uma população de bovinos da raça Nelore, assim como identificar os genes presentes nos segmentos de homozigose mais frequentes na população. Foram utilizadas informações de 34 touros de diferentes linhagens e sua progênie, totalizando 809 registros de animais genotipados da raça Nelore com informação de 509.107 SNPs. Para verificar a quantidade e tamanho dos segmentos de homozigose presentes na população deste estudo foram consideradas janelas contendo 50 SNPs e 500 Kb de distância, em que foi permitida a presença de um heterozigoto e nenhum genótipo desconhecido nesta janela. Para ser considerado segmento de homozigose, este deveria conter no mínimo 50 SNPs e 1 Mb de distância com SNPs em homozigose. Os resultados de ROH foram utilizados para calcular o coeficiente de endogamia. A endogamia da população apresentou valores para média, mediana, mínimo e máximo iguais a 5,84%, 5,40%, 0% e 24,88%, respectivamente. A média de endogamia dos reprodutores foi maior que a média da progênie, sugerindo que a escolha das fêmeas para os acasalamentos foi adequada, refletindo na redução da média de endogamia para a geração seguinte. Também foram observados longos segmentos de homozigose para alguns reprodutores. Os cromossomos 1, 5, 7, 12 e 21 apresentaram segmentos com frequência maior que 20% na população, nos quais foram observados genes relacionados a características de adaptação, sobrevivência e crescimento. A partir dos resultados apresentados, demonstra-se a baixa endogamia para a população, porém sugere-se que tenha ocorrido intensa utilização de poucos reprodutores nas gerações mais recentes em algumas linhagens.

Palavras-chave: bovinos de corte, linhagens, progênies, regiões de homozigose, reprodutores

1. INTRODUÇÃO

Os programas de melhoramento animal atualmente no Brasil realizam as avaliações genéticas utilizando modelo animal e melhor predição linear não-viesada (*best linear unbiased prediction* - BLUP). Nesta metodologia, quando há informação de animais de parentesco próximo, há maior ponderação para estas informações, enquanto que informações de parentescos distantes tornam-se menos importantes (CLARK et al., 2012). Esta abordagem resulta em taxas mais altas de co-seleção de parentes colaterais e pode promover grande aumento na taxa de endogamia ao longo das gerações (WOOLLIAMS et al., 2015).

A elevada taxa de endogamia em uma população pode reduzir a viabilidade e fertilidade dos animais (HARTL; CLARK, 2007), assim como promover a depressão endogâmica (FALCONER; MACKAY, 1996), em que é possível observar redução no desempenho de algumas características economicamente importantes. Dessa maneira, este parâmetro populacional deve ser constantemente monitorado nos programas de melhoramento genético. Este monitoramento pode ser realizado utilizando registros de pedigree. No entanto, a falta de informações e erros de registros podem prejudicar a acurácia dos resultados obtidos. Santana Jr. et al. (2010), sugeriram possível subestimação do coeficiente de endogamia observado em bovinos da raça Nelore devido a falta de informações no pedigree.

O avanço da tecnologia e a propagação da utilização dos polimorfismos de nucleotídeo único (SNPs) podem auxiliar na obtenção de um coeficiente de endogamia com base nas informações genômicas, o qual pode fornecer maiores informações para animais que apresentam deficiências no registro de pedigree (SAURA et al., 2013). Dentre as diferentes maneiras de se obter o coeficiente de

endogamia com informações genômicas, os segmentos de homozigose (ROH – “Runs of Homozigosity”) tem sido mais utilizados por permitir a detecção de endogamia recente ou antiga, e assim, têm sido amplamente utilizados em estudos para diversas raças bovinas (FERENČAKOVIĆ et al., 2011; BJELLAND et al., 2013; ZAVAREZ et al., 2015). Desse modo, os objetivos deste estudo foram estudar os segmentos de homozigose (ROH) e, com isso, a endogamia presente em uma população de bovinos da raça Nelore, assim como identificar os genes presentes nos segmentos de homozigose mais frequentes na população.

2. MATERIAL E MÉTODOS

2.1. Descrição dos dados e controle de qualidade

Dados genômicos foram obtidos a partir de 34 touros registrados da raça Nelore e de suas progênes, totalizando 780 novilhos machos. Os touros constituíram famílias de meios-irmãos que foram geradas por inseminação artificial. Os novilhos foram produzidos em três estações de monta, nascidos em 2007, 2008 e 2009. Os animais foram criados nas fazendas da Embrapa Pecuária Sudeste, localizada na cidade de São Carlos (SP), da Embrapa Gado de Corte, situada no município de Campo Grande (MS) e de propriedades particulares dos estados de Mato Grosso e Mato Grosso do Sul.

A escolha dos touros para genotipagem foi realizada a partir de consultas aos catálogos das centrais de inseminação artificial disponíveis no país. Para a composição desta amostra, 34 touros foram selecionados na população, de maneira que representassem as principais linhagens e genealogias que compõem a raça

Nelore, ao mesmo tempo em que se buscou minimizar o grau de parentesco entre os progenitores.

Os 814 animais da raça Nelore envolvidos nesse estudo foram genotipados com o BovineHD BeadChip (Illumina) no Laboratório Multiusuários Centralizado de Genômica Funcional Aplicada à Agropecuária e Agroenergia, Piracicaba, São Paulo. O controle de qualidade foi feito com a finalidade de remover erros de genotipagem excluindo-se SNPs localizados em regiões não autossômicas, com posições desconhecidas, com um p-valor no z-teste para os desvios significativos do equilíbrio de Hardy-Weinberg menor que 10^{-5} , taxa de leitura (“call rate”) menor que 0,98 e animais que apresentaram “call rate” menor que 0,90. Após o controle de qualidade, restaram 809 animais com informação de 509.107 SNPs.

2.2. Análise de segmentos de homozigose

Para verificar a quantidade e o tamanho dos segmentos de homozigose presentes na população utilizou-se o programa PLINK (PURCELL et al., 2007). Este programa computacional utiliza uma janela com número mínimo estabelecido de SNPs (10 SNPs por padrão) que percorre os segmentos cromossômicos paternos e maternos, de SNP a SNP, determinando se estes são homozigotos. O número de janelas completamente homozigotas, assim como o número total de janelas, é somado para cada SNP. Se os 10 SNPs consecutivos tiverem mais de 5% de todas as janelas que os cobrem, sendo homozigotos, o segmento é chamado de ROH (BJELLAND et al., 2013). Neste estudo, as janelas consideradas continham 50 SNPs e 500 Kb de distância no genoma, em que foi permitido a presença de 1 heterozigoto e nenhum SNP desconhecido nesta janela. Para ser considerado

segmento de homozigose, este deveria conter no mínimo 50 SNPs e 1 Mb de distância com SNPs em homozigose.

Os resultados de ROH foram utilizados para calcular o coeficiente de endogamia, sendo: $F_{ROH} = \frac{\sum_{j=1}^n L_{ROHj}}{L_{total}}$, em que L_{ROHj} é o tamanho do ROH e L_{total} é o tamanho total do genoma utilizado (2,548,724 kb). O cálculo para a endogamia foi realizado utilizando o programa R (R CORE TEAM, 2018).

Após a identificação dos segmentos de homozigose, foi observada a frequência dos mesmos na população e realizado um cariógrama com o auxílio do pacote “OmicCircos” (HU et al., 2014) do programa R (R Core Team, 2018) e para aqueles segmentos mais frequentes investigou-se a presença de genes utilizando a plataforma Ensembl (<https://www.ensembl.org/index.html>) e montagem do genoma versão UMD3.1. Após a identificação dos genes, os mesmos foram inseridos na plataforma DAVID Bioinformatic Resources 6.8 Beta (HUANG, D. W.; SHERMAN, B.T.; LEMPICKI, R. A., 2009a;b) para anotação gênica funcional.

3. RESULTADOS E DISCUSSÃO

Foram observados 42.753 segmentos de homozigose em todos os animais em que 31.644 segmentos, embora com possíveis sobreposições, foram diferentes nos animais estudados. Os segmentos apresentaram tamanho médio de 3,34 Mb com mediana de 2 Mb, mínimo de 1 Mb e máximo de 80,8 Mb, incluindo de 50 a 14.954 SNPs. O animal que apresentou a maior extensão total de homozigose é um touro com a soma da extensão de segmentos igual a 634 Mb (24,88% do genoma) distribuídos em 92 segmentos. Dentre os 10 primeiros animais com maior extensão total de homozigose estão 4 touros, os quais apresentaram mais de 11% de

segmentos maiores que 10 Mb. Segundo Howrigan et al. (2011), segmentos com tamanho maior que 10 Mb podem ser relacionados a endogamia que ocorreu dentro das cinco últimas gerações. Dessa maneira, embora estes animais sejam provenientes de diferentes linhagens (MUDADU et al., 2016), estes segmentos podem indicar recente uso intenso de touros dentro de suas linhagens.

A distribuição entre número de animais e extensão de segmentos de homozigose está mostrada na Figura 1. O animal que apresentou o segmento mais extenso (80,8 Mb) é progênie de um touro da linhagem Lemgruber, a qual, em estudo avaliando-se o pedigree, relatou-se redução na variabilidade genética e aumento do coeficiente de endogamia (OLIVEIRA et al., 2011). No entanto, a ausência de outros descendentes desta linhagem entre os animais com segmentos de homozigose mais extensos pode indicar que a escolha de progenitores com menor grau de parentesco foi eficiente para animais desta linhagem.

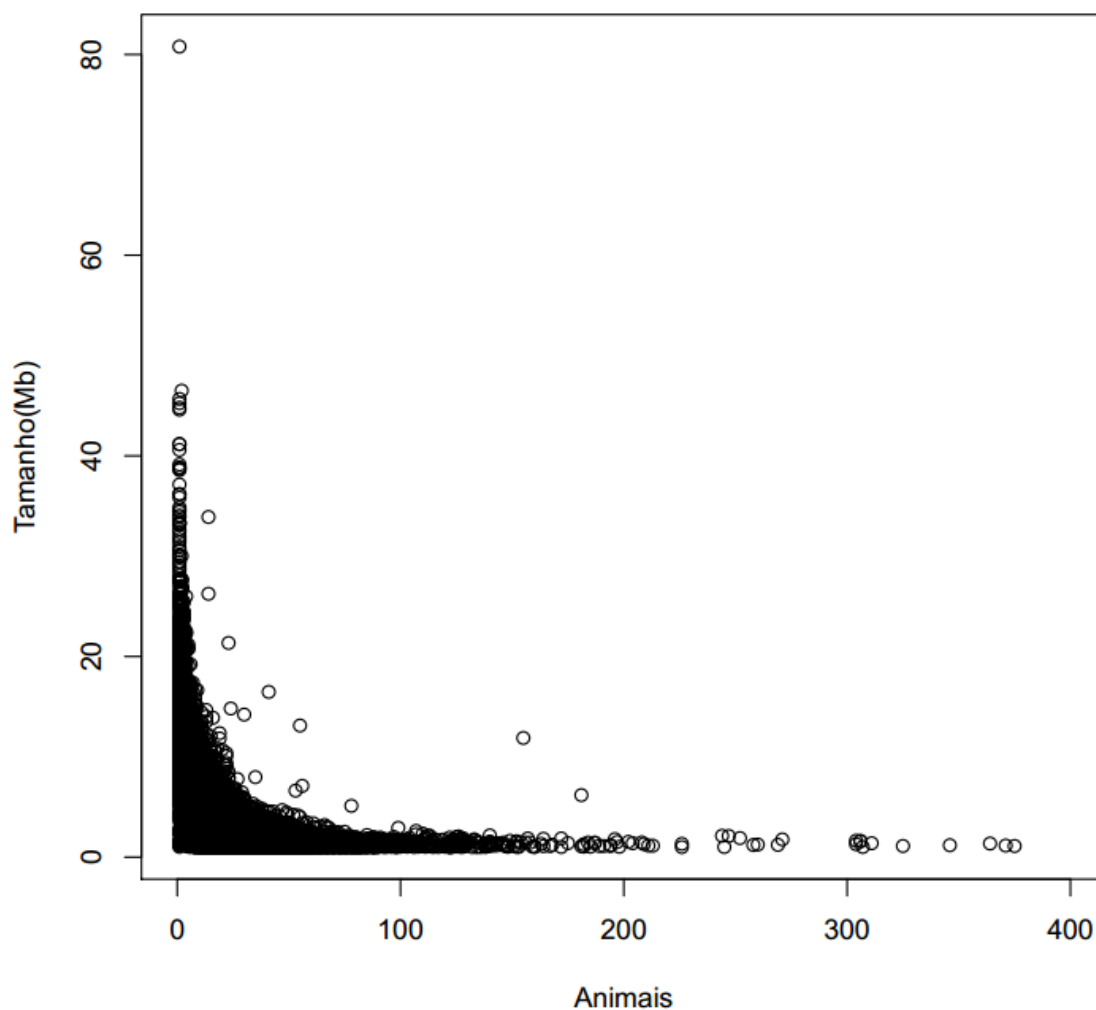


Figura 1. Distribuição entre número de animais e extensão de segmentos de homozigose em uma população de animais da raça Nelore.

A distribuição do número de segmentos de homozigose entre os cromossomos está mostrada na Figura 2. O cromossomo (BTA) 5 foi o que apresentou maior quantidade de segmentos de homozigose. Os segmentos de homozigose mais frequentes neste cromossomo apresentaram tamanho entre 2000 Kb e 5000 Kb (Figura 3). Em estudo de outra população de animais Nelore, o BTA 5 também foi o que apresentou maior média da porcentagem de “cobertura” por segmentos de homozigose dentre os cromossomos autossomos (ZAVAREZ et al.,

2015). Neste cromossomo também já foi relatado a presença de cinco diferentes genes em regiões de homozigose em população de Brahman e Tropical Composite (REVERTER et al., 2017). Este fato, somado a presença do maior segmento de homozigose observado neste estudo estar presente neste cromossomo, indica que este pode ser importante para estudos de endogamia.

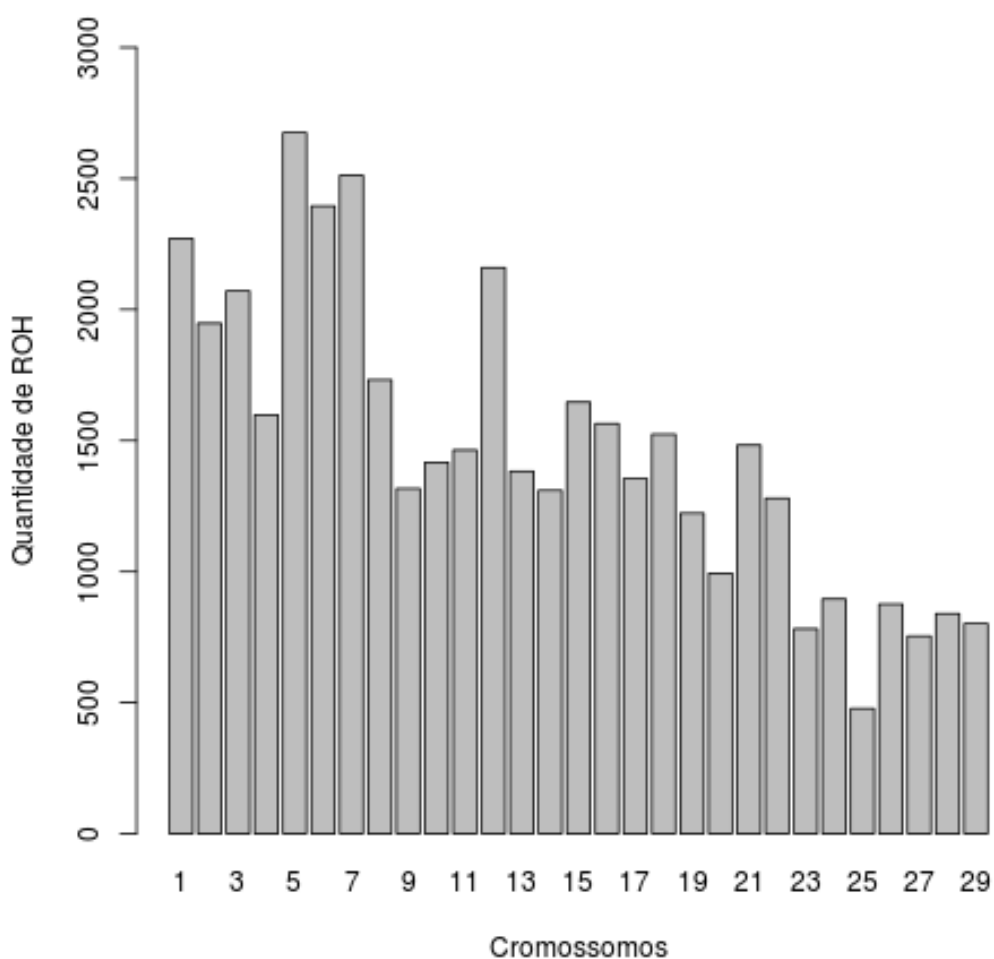


Figura 2. Distribuição entre os cromossomos dos 42.753 segmentos de homozigose observados em uma população de bovinos Nelore

Na Figura 3 está a distribuição dos tamanhos dos segmentos de homozigose observado ao longo dos cromossomos. O cromossomo 3 foi o que apresentou maior

quantidade de segmentos mais extensos (maiores que 10 Mb) dentre todos os autossomos estudados. No entanto, estes segmentos foram observados em baixa frequência na população, sendo representados em poucos animais e em diferentes regiões para cada animal.

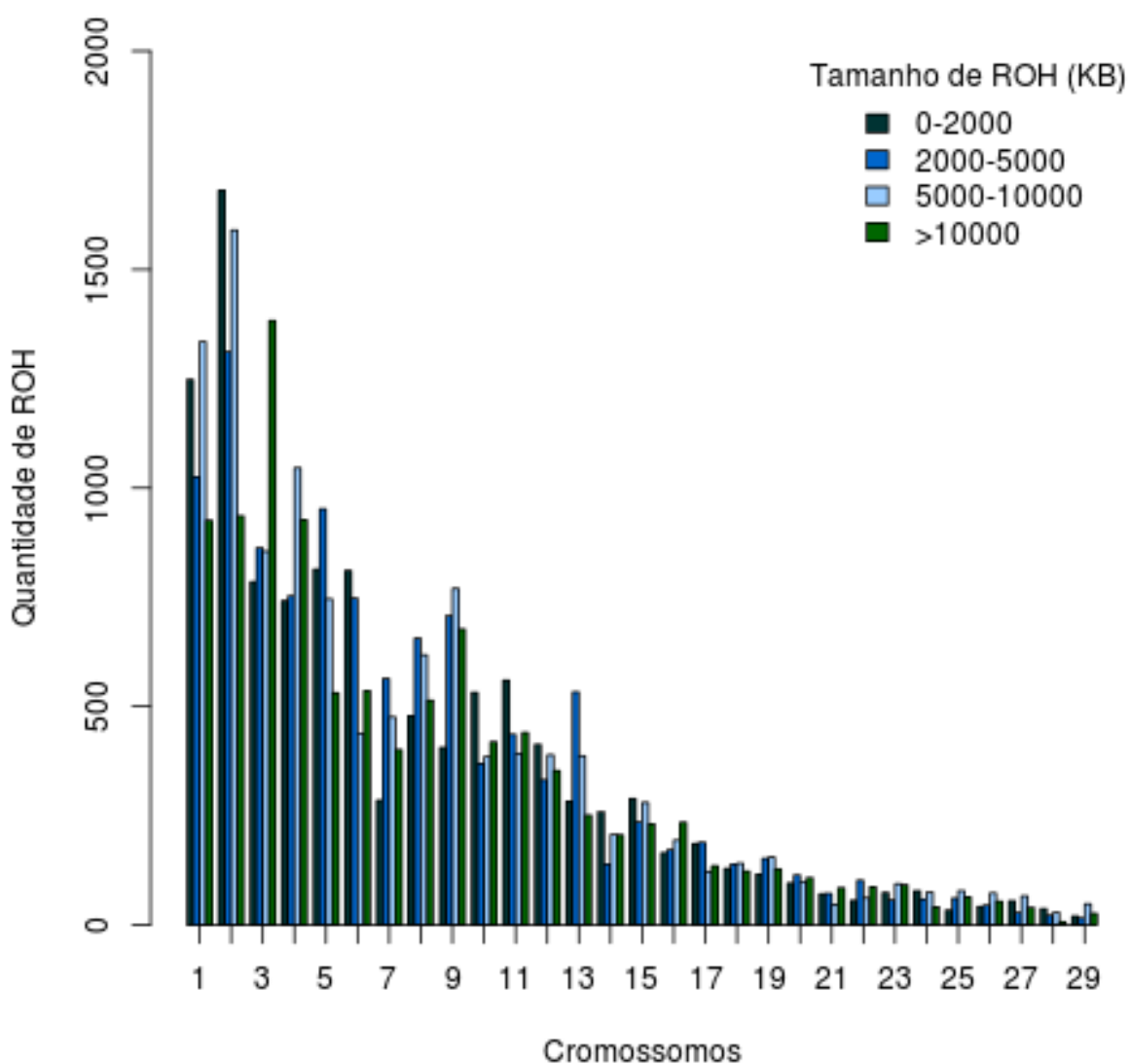


Figura 3. Distribuição entre os cromossomos dos 42.753 segmentos de homozigose por tamanho em uma população de bovinos Nelore

As regiões em que foram observados os segmentos de homozigose mais longos estão mostradas na Figura 4. Cada segmento de homozigose apresentado na Figura 4 foi observado em pelo menos um animal e apresenta tamanho maior que 10 Mb, o que pode indicar endogamia que ocorreu dentro das cinco últimas gerações na população destes animais. O cromossomo 5 foi o que apresentou o maior segmento e apesar deste estar presente em apenas um animal, a presença de segmentos em todas as categorias de extensão (de 10-13 Mb, 13-20 Mb, 20-40 Mb, 40-50 Mb e maior que 50 Mb) comprova que este pode ser um cromossomo a ser estudado ao investigar os efeitos da homozigose na raça Nelore.

A proporção dos segmentos de homozigose presentes na população variou de 0,12% até 46,35%, com média de 3,03%. Os segmentos mais frequentes na população apresentaram menor extensão (Figura 1), o que indica que grande parte da população apresenta endogamia antiga. No entanto, 63,53% dos animais apresentaram pelo menos um segmento maior que 10 Mb, o que sugere possíveis regiões de “autozigose” recente, principalmente para os reprodutores.

A endogamia da população apresentou média, mediana, mínimo e máximos iguais a 5,84%, 5,40%, 0% e 24,88%, respectivamente. Estes valores foram maiores do que os valores observados por Zavarez et al., (2015) em uma população Nelore. A média observada para a população foi semelhante à observada para 12 gerações anteriores em uma população de Guzerá, sendo que esta, em gerações atuais, apresenta médias de F_{ROH} inferiores (DE SOUZA FONSECA et al., 2016).

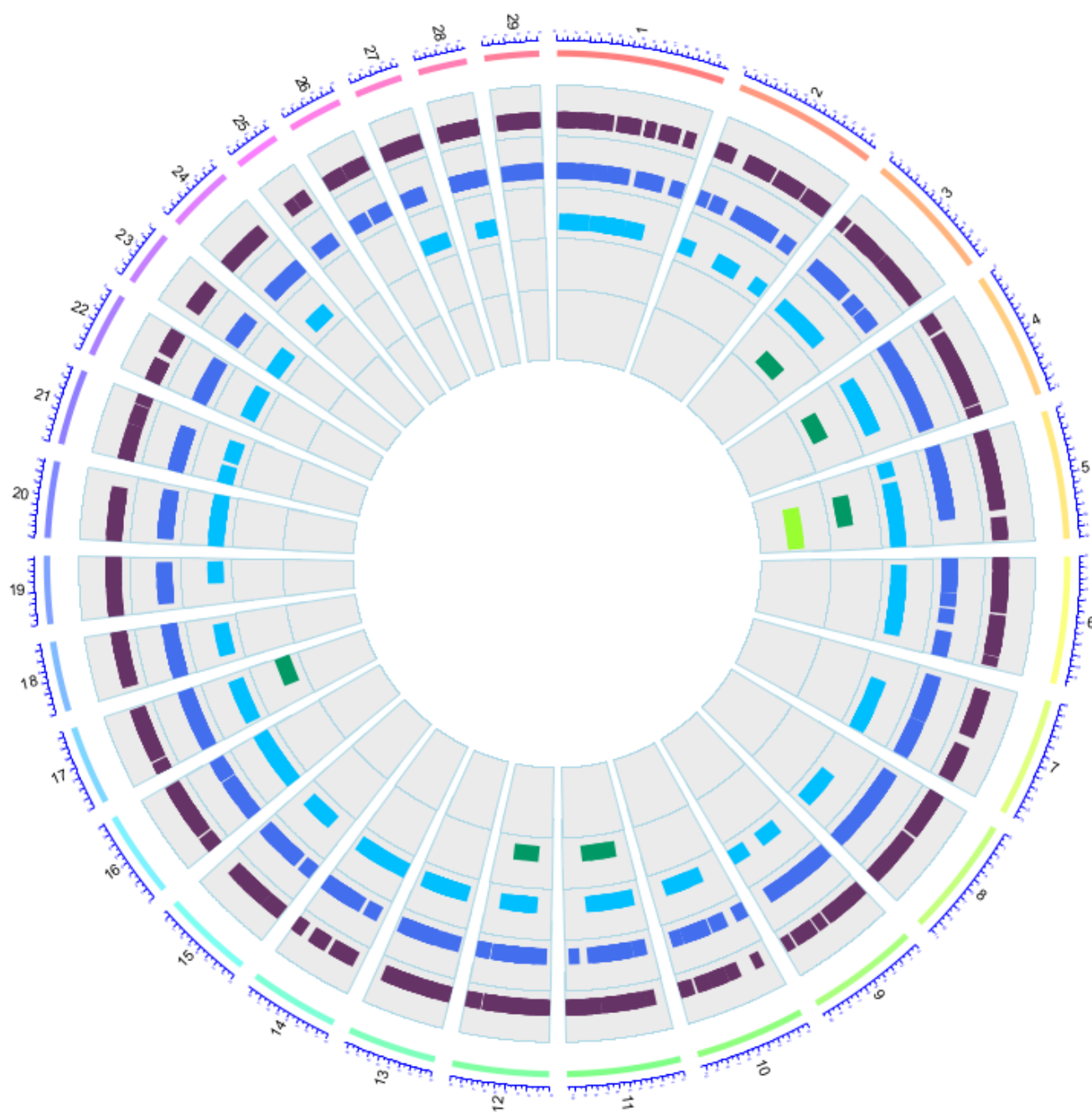


Figura 4. Distribuição nos cromossomos de segmentos de homozigose com tamanho entre 10 a 13 Mb (roxo), 13 a 20 Mb (azul escuro), 20 a 40 Mb (azul claro), 40 a 50 Mb (verde escuro) e maior que 50 Mb (verde claro) observados em uma população de bovinos Nelore.

Na Figura 5 é possível observar que a média de endogamia dos reprodutores (9,25%) foi maior que a média da progênie (5,69%), sugerindo que a escolha das fêmeas para os acasalamentos foi adequada, refletindo na redução da média de endogamia para a geração seguinte. Ferenčaković et al. (2011), ao estudarem touros da raça Simental, relataram média de F_{ROH} igual a 9%. A média de endogamia elevada para os reprodutores, juntamente com o fato da presença de segmentos de homozigose maiores que 10 Mb nos mesmos, sugerem intensa utilização de poucos reprodutores nas gerações mais recentes para algumas linhagens.

A localização dos segmentos de homozigose mais frequentes nos cromossomos (BTA) está representada na Figura 6. Os BTA 5, BTA 7, BTA 12 e BTA 21, foram os que apresentaram maior quantidade de segmentos de homozigose com maior frequência nesta população. Regiões próximas às observadas neste estudo para os BTA 7, BTA 12 e BTA 12 também foram relatadas com frequência de segmentos de homozigose maior que 40% em um estudo com raças taurinas e zebuínas (SÖLKNER et al., 2014).

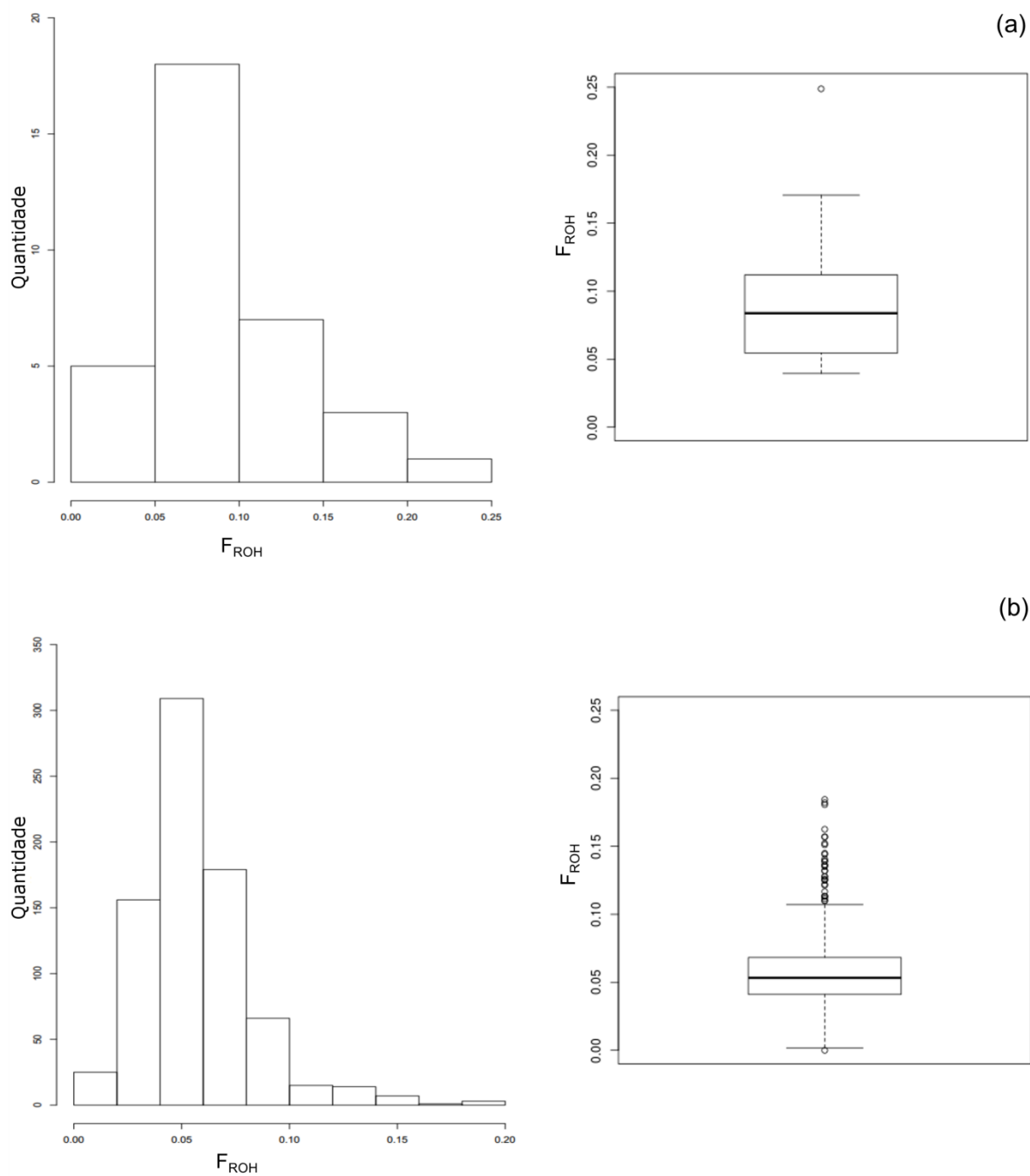


Figura 5. Histograma e boxplot da F_{ROH} para os reprodutores (a) e progênies (b) na população de bovinos Nelore estudada.

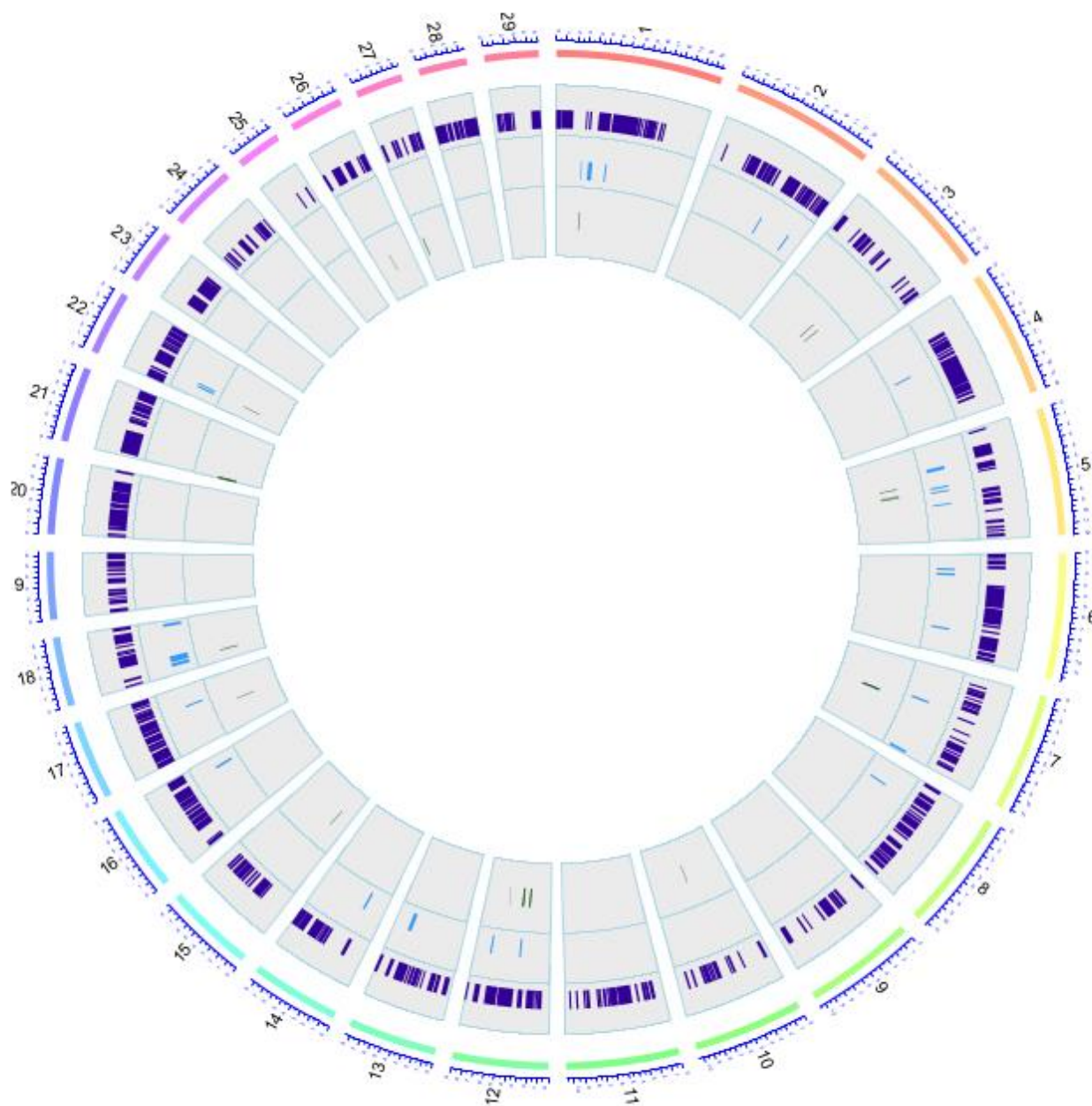


Figura 6. Distribuição nos cromossomos de segmentos de homozigose com frequência entre 4% a 10% (vermelho), entre 10% a 20% (azul) e maiores que 20% (roxo) em uma população de bovinos Nelore.

O tamanho dos segmentos com frequência maior que 20% na população variou de 1,27 Kb até 842 Kb. Ao todo foram identificados 323 genes presentes

nestes segmentos de homozigose. Os genes com função biológica anotada (Apêndice A) foram alocados em 12 grupos de acordo com processos celulares, componentes celulares e função molecular e foram enriquecidos em 1 via metabólica (Tabela 1).

A maioria dos genes identificados está relacionada com características adaptativas ou de crescimento. Dentre as características adaptativas, as que são relacionadas à defesa e imunidade do organismo são de importância para a propagação da espécie. Os genes *DEFBs* são pertencentes ao grupo da β -defensinas, o qual é uma das três subfamílias em que se dividem as defensinas. As defensinas são pequenos peptídeos que atuam como primeira linha de defesa do hospedeiro. Sendo assim, são importantes mediadores na resposta imune inata e sua atuação ocorre principalmente nos epitélios da pele e mucosas, rompendo a membrana microbiana (CHEN et al., 2013). Estudando as β -defensinas em diferentes espécies (humano, chimpanzé, camundongo, rato e cachorro), Patil et al. (2005) identificaram que a maioria destas defensinas mostraram-se preferencialmente expressas no sistema reprodutor masculino. Os autores afirmaram que as principais funções biológicas das β -defensinas podem estar relacionadas com fertilidade e reprodução, além da defesa do hospedeiro. Também foi observado o gene *TMEM173* (“transmembrane protein 173” – BTA 7), que foi relacionado com resposta imune em bovinos da raça Holandês (ZHENG et al., 2014).

O *NACA* (“nascent polypeptide-associated complex alpha subunit”) é um gene responsável por codificar uma proteína que compõe a subunidade alfa do complexo nascente, o qual se liga a polipeptídios recentemente sintetizados que emergem do

ribossomo (WIEDMANN et al., 1994; LI, RANDALL; DU., 2009). Schwerin et al. (2006) estudaram genes expressos que poderiam estar potencialmente envolvidos na transformação de alimentos em bovinos e identificaram níveis de transcrição hepática aumentados em Charolês quando comparados com a raça Holandesa, o que pode indicar que este gene pode influenciar o crescimento e desenvolvimento de bovinos de corte.

O gene *HMG2* (“high mobility group AT-hook 2”), que pode ser relacionado ao crescimento, é um fator de transcrição que impede que as células-tronco se diferenciem e, ao estudar regiões genômicas com grande efeito genético em bovinos de corte, Bolormaa et al. (2014) relataram a presença deste gene, afirmando ser o mesmo responsável por afetar o tamanho na maturidade e a distribuição de gordura destes animais.

Genes que favorecem a reprodução podem ser naturalmente selecionados, assim podendo ocorrer em regiões de homozigose nas espécies. A função do gene *SNRPN* (“small nuclear ribonucleoprotein polypeptide N” - BTA 21) em bovinos pode ainda ser explorada. Estudos identificaram relação deste gene com desempenho reprodutivo, em que foram observadas modificações no “imprinting” do gene *SNRPN* nos estágios pré e pós implantação embrionária quando utilizadas culturas *in vitro* ou transferência nuclear de células somáticas (SUZUKI et al., 2009). Este gene juntamente com o *NDN* (“necdin, MAGE family member” – BTA 21) e *MAGEL2* (“MAGE family member L2” – BTA 21) foram relatados por Irano et al. (2016) dentro de uma das 10 janelas que mais explicaram a variabilidade genética para a ocorrência de precocidade na prenhes de bovinos da raça Nelore.

Tabela 1. Descrição da anotação gênica funcional (Processo biológico, Componente celular ou Função molecular), enriquecimento de termos do “Gene Ontology” (GO) e principal KEGG (“Kyoto Encyclopedia of Genes and Genomes”) vias biológicas para os genes observados nos segmentos de homozigose com frequência maior que 20% em uma população de bovinos Nelore.

Número de acesso	Termo	N*	p-valor	Bonferroni	Benjamini
Processo Biológico					
GO:0042742	Resposta de defesa a bactéria	9	1,2E-06	6,8E-04	6,8E-04
GO:0015031	Transporte de proteínas	6	0,034	1,000	1,000
GO:0034613	Localização de proteína celular	3	0,034	1,000	1,000
GO:0045892	Regulação negativa da transcrição	7	0,040	1,000	1,000
GO:0031052	Quebra de cromossomo	2	0,048	1,000	1,000
GO:0008543	Via de sinalização do receptor do fator de crescimento fibroblástico	3	0,048	1,000	0,990
Componente Celular					
GO:0005886	Membrana plasmática	47	0,000	0,004	0,004
GO:0005576	Região extracelular	13	0,012	0,810	0,560
GO:0016021	Componente integral da membrana	60	0,027	0,980	0,720
Função Molecular					
GO:0004984	Atividade do receptor olfatório	34	0,000	0,000	0,000
GO:0004930	Atividade do receptor acoplado a proteína G	36	0,000	0,000	0,000
GO:0004745	Atividade da retinol-deidrogenase	3	0,009	0,820	0,440
Vias Biológicas					
-	Transdução olfatória	34	1,6E-8	1,5E-6	1,5E-6

*N – número de genes com significância de p-valor <0,05.

Genes relacionados à reprodução de fêmeas bovinas também foram observados nestas regiões de segmentos de homozigose com frequência maior que

20% na população, sendo o *KIF20A* (“kinesin family member 20A” – BTA 7) e *MZB1* (“marginal zone B and B1 cell specific protein” – BTA 7) relacionado com o processo de ovulação em bovinos (ROBKER et al., 2000; HAYASHI et al., 2010; HATZIRODOS et al., 2014). Outro gene associado ao desenvolvimento embrionário é o *ETF1* (“eukaryotic translation termination factor 1” – BTA 7) o qual, em estudo de criotolerância de embriões em zebuínos (*Bos taurus indicus*) e taurinos (*Bos taurus taurus*), foi identificado diferencial de expressão neste gene comparando blastocistos de animais da raça Simmental e Nelore (SUDANO, 2013).

Os genes *IRAK3* (“interleukin 1 receptor associated kinase 3” – BTA 5), *TMBIM4* (“transmembrane BAX inhibitor motif containing 4” – BTA 5), *LLPH* (“LLP homolog, long-term synaptic facilitation” – BTA 5) e *HMGA2* (“high mobility group AT-hook 2” – BTA 5) foram observados nas janelas que mais explicaram a variância genética para a característica habilidade de permanência no rebanho em estudo realizado por Teixeira et al. (2017) em bovinos Nelore, podendo também estar relacionados a características reprodutivas.

O gene *SLC23A1* (“solute carrier family 23 member 1” – BTA 7) parece se relacionar à características de adaptação bem como à assinatura de seleção para termotolerância em bovinos africanos (TAYE et al., 2017). A via metabólica identificada está relacionada com genes do sistema olfatório. Nos mamíferos, os genes sensoriais sofreram grandes mudanças ao longo da evolução e os animais dependem de adequada performance olfativa para comunicação social (LIM et al., 2013) e obtenção de alimentos. De fato, em estudo realizado com bovinos Nelore, genes associados com taxa de conversão alimentar também enriqueceram esta mesma via metabólica, relacionando a eficiência alimentar aos genes do sistema

olfatório (DE ALMEIDA SANTANA et al., 2016). De maneira geral, os genes observados nas regiões de segmentos de homozigose com maior frequência na população estudada estão relacionados com características reprodutivas, de defesa, imunidade, adaptação ao ambiente e desenvolvimento do animal.

4. CONCLUSÃO

Embora a média geral de endogamia para a população estudada seja baixa, foi possível observar que os reprodutores apresentaram maior valor de F_{ROH} , o qual somado a presença de segmentos de homozigose mais longos nestes animais, pode indicar possível uso intenso recente de poucos animais em algumas linhagens. Nos segmentos de homozigose mais frequentes na referida população foram identificados genes relacionados a características de adaptação, sobrevivência e crescimento.

5. REFERÊNCIAS

- BJELLAND, D. W.; WEIGEL, K. A.; VUKASINOVIC, N.; NKRUMAH, J. D. Evaluation of inbreeding depression in Holstein cattle using whole-genome SNP markers and alternative measures of genomic inbreeding. **Journal of Dairy Science**, Champaign, v. 96, p. 4697-4706, 2013.
- CHEN, Y.; ZHAO, H.; ZHANG, X.; LUO, H.; XUE, X.; LI, Z.; YAO, B. Identification, expression and bioactivity of *Paramisgurnus dabryanus* β -defensin that might be involved in immune defense against bacterial infection. **Fish & shellfish immunology**, Champaign, v. 35, n. 2, p. 399-406, 2013.
- BOLORMAA, S.; PRYCE, J. E.; REVERTER, A.; ZHANG, Y.; BARENDSE, W.; KEMPER, K.; GODDARD, M. E. A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle. **PLoS Genetics**, Cambridge, v. 10, n. 3, p. e1004198, 2014.
- CLARK, S. A.; HICKEY, J. M.; DAETWYLER, H. D.; VAN DER WERF, J. H. J. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. **Genomic Selection Evolution**, London, v. 44, n. 4, 2012.

DE ALMEIDA SANTANA, M. H.; JUNIOR, G. A. O.; CESAR, A. S. M.; FREUA, M. C.; DA COSTA GOMES, R.; E SILVA, S. D. L.; COUTINHO, L. L. Copy number variations and genome-wide associations reveal putative genes and metabolic pathways involved with the feed conversion ratio in beef cattle. **Journal of Applied Genetics**, Poznan, v. 57, n. 4, p. 495-504, 2016.

DE SOUZA FONSECA, P. A.; DOS SANTOS, F. C.; ROSSE, I. C.; VENTURA, R. V.; BRUNELLI, F. Â. T.; PENNA, V. M.; PEIXOTO, M. G. C. D. Retelling the recent evolution of genetic diversity for Guzerá: Inferences from LD decay, runs of homozygosity and N_e over the generations. **Livestock Science**, Oxford, v. 193, p. 110-117, 2016.

FALCONER, D. S.; MACKAY, T. F. C. **Introduction to quantitative genetics**. 4. ed. Harlow: Longman House, 1996. p. 245 – 253.

FRIEDRICH, J.; BRAND, B.; GRAUNKE, K. L.; LANGBEIN, J.; SCHWERIN, M.; PONSUKSILI, S. Adrenocortical Expression Profiling of Cattle with Distinct Juvenile Temperament Types. **Behavior Genetics**, New York, v. 47, n. 1, p. 102-113, 2017.

HARTL, D. L.; CLARK, A. G. **Principles of population genetics**. 4. ed. Massachusetts: Sinauer Associates, 2007. p. 265 - 268.

HATZIRODOS, N.; HUMMITZSCH, K.; IRVING-RODGERS, H. F.; HARLAND, M. L.; MORRIS, S. E.; RODGERS, R. J. Transcriptome profiling of granulosa cells from bovine ovarian follicles during atresia. **BMC Genomics**, London, v. 15, n. 40, 2014.

HAYASHI, K. G.; USHIZAWA, K.; HOSOE, M.; TAKAHASHI, T. Differential genome-wide gene expression profiling of bovine largest and second-largest follicles: identification of genes associated with growth of dominant follicles. **Reproductive Biology and Endocrinology**, London, v. 8, n. 11, 2010.

HOWRIGAN, D. P., SIMONSON, M. A., AND KELLER, M. C. Detecting autozygosity through runs of homozygosity: a comparison of three autozygosity detection algorithms. **BMC Genomics**, London, v.12, n. 460, 2011.

HUANG D. W.; SHERMAN, B.T.; LEMPICKI, R. A. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. **Nature Protocols**, New York, v. 4, n. 1, p. 44-57, 2009a.

HUANG, D.W.; SHERMAN, B.T.; LEMPICKI, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. **Nucleic Acids Research**, Oxford, v. 37, n. 1, p. 1-13, 2009b.

HU, Y.; YAN, C.; HSU, C. H.; CHEN, Q. R.; NIU, K.; KOMATSOULIS, G. A.; MEERZAMAN, D. OmicCircos: a simple-to-use R package for the circular visualization of multidimensional omics data. **Cancer Informatics**, v. 13, n. CIN. S13495, 2014.

IRANO, N.; DE CAMARGO, G. M. F.; COSTA, R. B.; TERAKADO, A. P. N.; MAGALHÃES, A. F. B.; DE OLIVEIRA SILVA, R. M.; DE OLIVEIRA, H. N. Genome-wide association study for indicator traits of sexual precocity in nellore cattle. **PLoS one**, Cambridge, v. 11, n. 8, e0159502, 2016.

LI, H.; RANDALL, W. R.; DU, S. J. skNAC (skeletal Naca), a muscle-specific isoform of Naca (nascent polypeptide-associated complex alpha), is required for myofibril organization. **The FASEB Journal**, Bethesda, v. 23, n. 6, p. 1988-2000, 2009.

LIM, D.; GONDRO, C.; PARK, H. S.; CHO, Y. M.; CHAI, H. H.; SEONG, H. H.; LEE, S. H. Identification of recently selected mutations driven by artificial selection in Hanwoo (Korean cattle). **Asian-Australasian journal of animal sciences**, Gwanak-gu, v. 26, n. 5, p. 603, 2013.

MUDADU, M. A.; PORTO-NETO, L. R.; MOKRY, F. B.; TIZIOTO, P. C.; OLIVEIRA, P. S. N.; TULLIO, R. R.; NASSU, R. T.; NICIURA, S. C. M.; THOLON, P.; ALENCAR, M. M.; HIGA, R. H.; ROSA, A. N.; FEIJÓ, G. L. D.; FERRAZ, A. L. J.; SILVA, L. O. C.; MEDEIROS, S. R.; LANNA, D. P.; NASCIMENTO, M. L.; CHAVES, A. S.; SOUZA, A. R. D. L.; PACKER, I. U. TORRES JR., R. A. A.; SIQUEIRA, F.; MOURÃO, G. B.; COUTINHO, L. L.; REVERTER, A.; REGITANO, L. C. A. Genomic structure and marker-derived gene networks for growth and meat quality traits of Brazilian Nelore beef cattle. **BMC Genomics**, London, v.17, n.235, 2016.

OLIVEIRA, P. S.; SANTANA JÚNIOR, M. L.; PEDROSA, V. B.; OLIVEIRA, E. C. M.; ELER, J. P.; FERRAZ, J. B. S. Estrutura populacional de rebanho fechado da raça Nelore da linhagem Lemgruber. **Pesquisa Agropecuária Brasileira, Brasília**, v. 46, n. 6, p. 639 – 647, 2011.

PATIL, A. A.; CAI, Y.; SANG, Y.; BLECHA, F.; ZHANG, G. Cross-species analysis of the mammalian β -defensin gene family: presence of syntenic gene clusters and preferential expression in the male reproductive tract. **Physiological genomics**, Bethesda, v. 23, n. 1, p. 5-17, 2005.

PEREIRA, J. C. C. **Melhoramento genético aplicado à produção animal**. Belo Horizonte: FEPMVZ, 2012. p.140 - 157.

PURCELL, S.; NEALE, B.; TODD-BROWN, K.; THOMAS, L.; FERREIRA, M. A. R.; BENDER, D.; MALLER, J.; SKLAR, P.; BAKKER, P. I. W.; DALY, M. J.; SHAM, P. C. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. **The American Journal of Human Genetics**, Houston, v. 81, p. 559-575, 2007.

R CORE TEAM. R: A language and environment for statistical computing. Viena: R Foundation for Statistical Computing; 2016. Available: <http://www.R-project.org/>. Acessado: 10 Agosto 2018.

REVERTER, A.; PORTO-NETO, L. R.; FORTES, M. R. S.; KASARAPU, P.; DE CARA, M. A. R.; BURROW, H. M.; LEHNERT, S. A. Genomic inbreeding depression for climatic adaptation of tropical beef cattle. **Journal of Animal Science**, Champaign, v. 95, n.9 , p. 3809, 2017.

ROBKER, R. L.; RUSSELL, D. L.; ESPEY, L. L.; LYDON, J. P.; O'MALLEY, B. W.; RICHARDS, J. S. Progesterone-regulated genes in the ovulation process: ADAMTS-1 and cathepsin L proteases. **Proceedings of the National Academy of Sciences**, Washington, v. 97, n. 9, p. 4689-4694, 2000.

SAURA, M.; FERNÁNDEZ, A.; RODRÍGUEZ, M. C.; TORO, M. A.; BARRAGÁN, C.; FERNÁNDEZ, A. I.; VILLANUEVA, B. Genome-wide estimates of coancestry and

inbreeding in a closed herd of ancient Iberian pigs. **PLoS One**, v. 8, n. 10, e78314, 2013.

SCHWERIN, M.; KUEHN, C.; WIMMERS, S.; WALZ, C.; GOLDAMMER, T. Trait-associated expressed hepatic and intestine genes in cattle of different metabolic type—putative functional candidates for nutrient utilization. **Journal of Animal Breeding and Genetics**, Malden, v. 123, n. 5, p. 307-314, 2006.

SÖLKNER, J.; FERENČAKOVIĆ, M.; KARIMI, Z.; PEREZ O'BRIEN, A.M.; MÉSZÁROS, G.; EAGLEN, S.; BOISON, S. A.; CURIK, I. Extremely Non-uniform: Patterns of Runs of Homozygosity in Bovine Populations. In: Proceedings 10th World Congress of Genetics Applied to Livestock Production. Vancouver, Canada, 2014. Disponível em: <<https://asas.confex.com/asas/WCGALP14/webprogram/Paper9877.html>>. Acessado em: 25 de Outubro de 2017.

SUDANO, M. J. **Criotolerância de Embriões Bos taurus indicus e Bos taurus taurus Produzidos In Vitro e In Vivo**. 2013. 117 f. Tese (Doutorado em Medicina Veterinária) – Faculdade de Medicina Veterinária e Zootecnia, Universidade Estadual Paulista, Botucatu, 2013.

SUZUKI, J.; THERRIEN, J.; FILION, F.; LEFEBVRE, R.; GOFF, A. K.; SMITH, L. C. In vitro culture and somatic cell nuclear transfer affect imprinting of SNRPN gene in pre-and post-implantation stages of development in cattle. **BMC Developmental Biology**, London, v. 9, n. 1, n. 9, 2009.

TAYE, M.; LEE, W.; CAETANO-ANOLLES, K.; DESSIE, T.; HANOTTE, O.; MWAI, O. A.; KIM, H. Whole genome detection of signature of positive selection in African cattle reveals selection for thermotolerance. **Animal Science Journal**, Medford, v. 88, n. 12, p. 1889-1901, 2017.

TEIXEIRA, D. B. A.; JÚNIOR, G. A. F.; DOS SANTOS SILVA, D. B.; COSTA, R. B.; TAKADA, L.; GORDO, D. G. M.; ALBUQUERQUE, L. G. Genomic analysis of stayability in Nelore cattle. **PloS one**, São Francisco, v. 12, n. 6, p. e0179076, 2017.

WIEDMANN, B.; SAKAI, H.; DAVIS, T. A.; WIEDMANN, M. A protein complex required for signal-sequence-specific sorting and translocation. **Nature**, New York, v. 370, n. 6489, p. 434-440, 1994.

WOOLLIAMS, J. A.; BERG, P.; DAGNACHEW, B. S.; MEUWISSEN, T. H. E. Genetic contributions and their optimization. **Journal of Animal Breeding and Genetics**, Malden, v. 132, p. 89 - 99, 2015.

ZAVAREZ, L. B.; UTSUNOMIYA, Y. T.; CARMO, A. S.; NEVES, H. H. R.; CARVALHEIRO, R.; FERENČAKOVIC, M.; O'BRIEN, A. M. P.; CURIK, I.; COLE, J. B.; VAN TASSELL, C. P.; SILVA, M. V. G. B.; SONSTEGARD, T. S.; SÖLKNER, J.; GARCIA, J. F. Assessment of autozygosity in Nelore cows (*Bos indicus*) through high-density SNP genotypes. **Frontiers in Genetics**, Lausanne, v. 6, n. 5, 2015.

ZHENG, Y.; CHEN, K. L.; ZHENG, X. M.; LI, H. X.; WANG, G. L. Identification and bioinformatics analysis of microRNAs associated with stress and immune response in serum of heat-stressed and normal Holstein cows. **Cell Stress and Chaperones**, Storrs, v. 19, n. 6, p. 973-981, 2014.

APÊNDICE

Apêndice A. Cromossomo (BTA), posição e os genes presentes nos segmentos de homozigose com frequência maior que 20% na população.

BTA	Segmento	Genes
1	30941859 - 32090411	RF00026
3	66296159- 66300051	IFI44
3	66303205 - 67444200	IFI44L,RF00568,PTGFR,GIPC2,RF00026,DNAJB4,RF00026,FUBP1,NEXN,MIGA1,RF00026,USP33,ZZZ3,AK5
5	47007950 - 48167401	GRIP1,RF00003,HELB,IRAK3,TMBIM4,LLPH,RF00001,HMGA2,bta-mir-763
5	56757198 - 57311731	ZBTB39,GPR182,RDH16,SDR9C7,RF00494,RF00493,HS D17B6,PRIM1,NACA,PTGES3,ATP5F1B,RF00273,RF00273,BAZ2A,RBMS2,GLS2,SPRYD4,MIP,TIMELESS,ApoN,APOF
5	57398518 - 57444406	ANKRD52,bta-mir-2433,SLC39A5,NABP2,RNF41
5	57751697 - 58372509	MMP19,DNAJC14,ORMDL2,SARNP,GDF11,CD63,RDH5,BLOC1S1,ITGA7,METTTL7B,RF01241,OR10P1,OR2AP1,RF00026
5	58580333 - 59946266	OR6C76,OR6C75,OR6C1,OR10A7
7	51172404 - 53433352	NME5,BRD8,KIF20A,CDC23,RF01225,GFRA3,CDC25C,SLBP2,FAM53C,bta-mir-2459,KDM3B,REEP2,EGR1,ETF1,HSPA9,RF00154,RF00154,CTNNA1,RF00001,LRRTM2,SIL1,RF00090,RF00090,MATR3,PAIP2,SLC23A1,MZB1,PROB1,SPATA24,DNAJC18,ECSCR,SMIM33,TMEM173,UBE2D2,CXXC5,PSD2,NRG2,RF00026,PURA,IGIP,CYSTM1,PFDN1,HBEGF,SLC4A9,EIF4EBP3,RF00026,SRA1,APBB3,SLC35A4
10	53080510 - 53923664	TCF12,RF00026
12	28035606 - 29794535	KL,PDS5B,N4BP2L1,BRCA2,ZAR1L,FRY,RXFP2,bta-mir-2299,B3GLCT
15	19120353 - 20105690	C15H11orf87
17	35391726 - 36349534	IL21,IL2,ADAD1,RF00026
18	13730814 - 14991573	ZNF469,ZFPM1,ZC3H18,IL17C,CYBA,MVD,SNAI3,RNF166,CTU2,PIEZO1,bta-mir-2327,CDT1,APRT,GALNS,TRAPPC2L,PABPN1L,CBFA2T3,ACSF3,CDH15,SLC22A31,ANKRD11,SPG7,RPL13,RF00324,CPNE7,DP EP1,CHMP1A,CDK10,SPATA2L,VPS9D1,ZNF276,FANCA,SPIRE2,TCF25,MC1R,TUBB3,DEF8,DBNDD1,GAS8,RF00003,SHCBP1
18	14999189 - 15057726	VPS35
21	5191 - 2267906	SNRPN,NDN,MAGEL2,RF00026,RF00001,RF00108,RF00105,RF01278
22	15310501 - 16219978	VIPR1,SEC22C,NKTR,ZBTB47,KLHL40,HHATL,CCDC13,HIGD1A,ACKR2,CYP8B1,POMGNT2,SNRK,bta-mir-2368,ANO10
26	21976305 - 22009768	BTRC
26	22032363 - 22653239	POLL,DPCD,FBXW4,FGF8,NPM3,MGEA5,RF00001,KCNIP2,C10orf76,HPS6,LDB1
27	5036606 - 6324087	LAP,DEFB7,DEFB6,DEFB13,DEFB4A,DEFB,EBD,DEFB5,DEFB1,DEFB10,GPM6A