

UNIVERSIDADE ESTADUAL PAULISTA – UNESP

CÂMPUS JABOTICABAL

**Estratégias de imputação e associação genômica com
dados de sequenciamento para características de
produção de leite na raça Gir**

Guilherme Batista do Nascimento

Biólogo

2018

UNIVERSIDADE ESTADUAL PAULISTA – UNESP
CÂMPUS JABOTICABAL

**Estratégias de imputação e associação genômica com
dados de sequenciamento para características de
produção de leite na raça Gir**

Guilherme Batista do Nascimento

Orientador: Prof. Dr. Danísio Prado Munari

Coorientadores: Dr. Rodrigo Pelicioni Savegnago

Dr. Marcos Vinicius Gualberto Barbosa da Silva

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

N124e Nascimento, Guilherme Batista
Estratégias de imputação e associação genômica com dados de sequenciamento para características de produção de leite na raça Gir / Guilherme Batista do Nascimento. – – Jaboticabal, 2018
iv, 60 p. : il. ; 29 cm

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

Orientador: Danísio Prado Munari

Coorientadores: Rodrigo Pelicioni Savegnago, Marcos Vinicius Gualberto Barbosa da Silva

Banca examinadora: Marcos Eli Buzanskas, Ana Fabricia Braga Magalhães, Adriana Santana do Carmo, Fernando Sebastian Baldi Rey

Bibliografia

1. Acurácia de imputação. 2. False Discovery Rate. 3. Next-generation sequence. 4. *Bos taurus indicus*. 5. Bovinocultura de leite 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.



UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Jaboticabal



CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: ESTRATÉGIAS DE IMPUTAÇÃO E ASSOCIAÇÃO GENÔMICA COM DADOS DE SEQUENCIAMENTO PARA CARACTERÍSTICAS DE PRODUÇÃO DE LEITE NA RAÇA GIR

AUTOR: GUILHERME BATISTA DO NASCIMENTO

ORIENTADOR: DANISIO PRADO MUNARI

COORIENTADOR: RODRIGO PELICIONI SAVEGNAGO

COORIENTADOR: MARCOS VINÍCIUS GUALBERTO BARBOSA DA SILVA

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



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



Prof. Dr. MARCOS ELI BUZANSKAS
Departamento de Zootecnia / Universidade Federal da Paraíba - Areia/PB



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



Profa. Dra. ADRIANA SANTANA DO CARMO
Departamento de Zootecnia / Universidade Federal de Goiás - Goiânia/GO



Prof. Dr. FERNANDO SEBASTIAN BALDI REY
Departamento de Zootecnia / FCAV / UNESP - Jaboticabal

Jaboticabal, 22 de fevereiro de 2018

DADOS CURRICULARES DO AUTOR

Guilherme Batista do Nascimento – Filho de Deise Mara Batista e Gildo Divino Soares do Nascimento, nasceu em Tupã – SP, no dia 12 de outubro de 1988. Iniciou a graduação em Ciências Biológicas em março de 2008, na Faculdade de Ciências Agrárias e Veterinárias – UNESP - Câmpus de Jaboticabal e obteve o Título de Bacharel em Ciências Biológicas em 2012. Durante os anos de 2010 e 2011, desenvolveu dois trabalhos de Iniciação Científica na mesma instituição de ensino, sob orientação do Prof. Dr. Danísio Prado Munari e Dr. Rodrigo Pelicioni Savegnago, o qual lhe rendeu o prêmio Ciências Biológicas, pelo melhor trabalho de conclusão de curso. Em fevereiro de 2014, concluiu o Curso de Mestrado do Programa de Genética e Melhoramento Animal na Faculdade de Ciências Agrárias e Veterinárias – UNESP - Câmpus de Jaboticabal, sob a orientação do Prof. Dr. Danísio Prado Munari e co-orientação da Pesquisadora Dra. Mônica Corrêa Ledur e Prof. Dr. Ricardo Vieira Ventura. Ingressou no Curso de Doutorado do mesmo programa e orientação do Prof. Dr. Danísio Prado Munari em março de 2014, sendo coorientado por Dr. Rodrigo Pelicioni Savegnago e Dr. Marcos Vinicius G. B. da Silva. Durante o período de um ano realizou estágio sanduíche na *University of Guelph*, sendo supervisionado pelo Prof. Dr. Flavio Schenkel.

"A maturidade nos permite olhar com menos ilusões, aceitar com menos sofrimento,
entender com mais tranquilidade e querer com mais doçura".
(Lya Luft)

AGRADECIMENTOS

Agradeço primeiramente aos ausentes nesses agradecimentos. A todos que de alguma forma passaram em minha vida ao longo desses quatros anos de elaboração deste trabalho. A todos esses marcadores de pequeno efeito, meu muito obrigado!

Ao meu orientador, Prof. Danísio, pelos 9 anos de aprendizados. O convívio ao longo desses anos deixa um eterno respeito e consideração. O qual estendo a Rose. Meu amigo e eterno coorientador Rodrigo, você foi essencial para que tudo chegasse até aqui.

Minha única contemporânea, Chud. Nossa irmandade vai além da academia, você é a garantia de uma boa discussão! Obrigado por tudo isso. Meus caros Lucas e Rafa, sócios em uma sociedade secreta que pisca. Sem vocês os dias certamente teriam sido mais chatos. Aos amigos Lê, Ana, Eli, Alejandro, Jaque, Lu, Marcos, Pri, Salvs, Rebeka, Eliéder, vocês são os melhores bebedores de café que existe no mundo. A todos do departamento de Ciências Exatas, que tenho o prazer de ter ao menos uma boa recordação.

Meu porto seguro em Jaboticabal, minha família que a vida permitiu ter. Parli, Alê e Ma. Foi incrível cada momento, cada almoço de domingo. Parli, meu irmão, que me deu uma família toda Ana, Isa e Carol e emprestou amigos como Marcio, Fer, Marcelinho, Leandro e tantos outros. Que junto com Rullian e Dou fizeram mesas de bar se tornarem lugares incríveis. Aos queridos Naomi e Faminta, amigos incríveis que me proporcionaram ótimos dias em um lar provisório.

Meu orientador na *University of Guelph* Prof. Flávio, um exemplo de profissional que agradeço por todo auxílio e dados cedidos, sem os quais não haveria esse trabalho. Minha querida amiga Virgulinha, o sorriso ao escrever seu nome descreve meu carinho por você e nossos queridos Ala (Hey Dude!) e Hans. E eterno carinho pelos amigos em Guelph, Val, Mari, Luiz, Flávia, Pedro, Doug e Norma, Kirsten, Isis.

As pessoas mais importantes em meu mundo, minha família. Minha mãe Deise, você não só me permitiu sonhar como me ajudou a realizar cada sonho. Você é a pessoa mais forte e incrível que terei a oportunidade de conhecer. Meu pai Gildo, você me fez curioso, a virtude mais importante que tive para chegar até aqui. Minha irmã Natália, por me fazer entender mais sobre o que é amar incondicionalmente.

Minha querida avó Yvone, com seu humor cativante faz uma tarde parecer um minuto. Meu avô Celso (*In memoriam*), pela lição sobre o tempo. Além do meu churrasqueiro predileto Marcel e ao melhor peixe do Clóvis e meus tios Rogério,

Denise e Celso (*In memoriam*) e primo Lucas por serem essa pequena, mas querida família.

A Universidade Estadual Paulista (UNESP), pelo orgulho de fazer parte dessa prestigiada Universidade por 10 anos. A todos da Bio08, por serem ornitorrincos bizarros. A agência de fomento Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela cessão da bolsa PPGGMA-DS e bolsa do convênio Capes/Embrapa (edital 15/2014). Ao programa Ciência Sem Fronteiras da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CNPq) pela concessão da bolsa processo 200968/2015-7, mesmo em tempos tão difíceis. A Empresa Brasileira de Pesquisa Agropecuária (Embrapa) Gado de Leite em Juiz de Fora (MG) e ao Laboratório Multiusuário de Bioinformática - Embrapa Informática Agropecuária, pela cessão de parte dos dados e infraestrutura computacional para análises de dados.

Minha velha amiga, Thata e meus tios por adoção Dora e Ferrugem. Meus bons amigos Kilombolas, cujo o copo amigo sempre se fez presente. A minha querida amiga Ju, por partilhar sonhos e surpresas. Meus velhos amigos Rao e Hércio que os anos só fortaleceram a amizade.

Ao acaso, Deus ou mesmo o resíduo aleatório desse modelo. Que mesmo sem poder ser explicado sempre se faz presente.

SUMÁRIO

Página

RESUMO	III
CAPÍTULO 1 - CONSIDERAÇÕES GERAIS	1
1 INTRODUÇÃO	1
2 OBJETIVOS.....	2
2.1 OBJETIVO GERAL	2
2.2 OBJETIVOS ESPECÍFICOS.....	3
3 REVISÃO DE LITERATURA	3
3.1 SEQUENCIAMENTO DE NOVA GERAÇÃO.....	3
3.2 IMPUTAÇÃO DOS GENÓTIPOS	5
3.3 RAÇA GIR E CARACTERÍSTICAS AVALIADAS	8
4 REFERÊNCIAS.....	11
CAPÍTULO 2 - IMPUTAÇÃO DE PAINÉIS DE GENOTIPAGEM PARA DADOS DE SEQUENCIAMENTO DE ANIMAIS ZEBUÍNOS.....	15
RESUMO	15
1 INTRODUÇÃO	16
2 MATERIAL E MÉTODOS	17
2.1 DADOS E CONTROLE DE QUALIDADE	17
2.2 PAINÉIS DE ALTA DENSIDADE	18
2.3 IMPUTAÇÃO	19
2.4 ANÁLISE DE COMPONENTES PRINCIPAIS (PCA)	20
3 RESULTADOS E DISCUSSÃO	20
3.1 DADOS DE SEQUENCIAMENTO E PAINÉIS DE GENOTIPAGEM.....	20
3.2 ESTRUTURA DA POPULAÇÃO	22
3.3 ACURÁCIA DE IMPUTAÇÃO.....	24
3.4 ACURÁCIA DE IMPUTAÇÃO AO LONGO DO GENOMA	25
4 CONCLUSÃO	29
REFERÊNCIAS.....	30
CAPÍTULO 3 - ASSOCIAÇÃO GENÔMICA AMPLA COM DADOS DE SEQUENCIAMENTO EM CARACTERÍSTICAS DE PRODUÇÃO E QUALIDADE DE LEITE NA RAÇA GIR	33
RESUMO	33
1 INTRODUÇÃO	34
2 OBJETIVOS.....	37
3 MATERIAL E MÉTODOS	37

3.1 DADOS DE GENÓTIPOS E DE SEQUENCIAMENTO	37
3.2 DADOS FENOTÍPICOS	38
3.3 ANÁLISE DE COMPONENTES PRINCIPAIS (PCA)	39
3.4 IMPUTAÇÃO	39
3.5 ESTUDO DE ASSOCIAÇÃO GENÔMICA AMPLA.....	40
3.6 ANÁLISE DE ENRIQUECIMENTO	41
4. RESULTADOS E DISCUSSÃO	41
4.1 IMPUTAÇÃO	41
4.2 ANÁLISE DE ASSOCIAÇÃO GENÔMICA (GWAS).....	44
4.3 ANÁLISE DE ENRIQUECIMENTO	54
5. CONCLUSÕES	56

ESTRATÉGIAS DE IMPUTAÇÃO E ASSOCIAÇÃO GENÔMICA COM DADOS DE SEQUENCIAMENTO PARA CARACTERÍSTICAS DE PRODUÇÃO DE LEITE NA RAÇA GIR

RESUMO – A implementação de dados de sequenciamento de nova geração - “next-generation sequence” (NGS) em programas de melhoramento genético animal representa a mais recente ferramenta na utilização de dados genotípicos nos modelos de associação genômica, tendo em vista que todo polimorfismo é considerado nas associações entre registros fenotípicos e dados de sequenciamento. Como em toda nova tecnologia, a prospecção das variantes ainda representa um desafio no sentido computacional e de viabilidade dos custos para sua implementação em larga escala. Diante desses desafios, neste trabalho buscou-se meios de explorar os benefícios na utilização da NGS nas predições genômicas e superar as limitações inerentes a esse processo. Registros fenotípicos e genotípicos (Illumina Bovine HD BeadChip) de 2.279 animais da raça Gir (*Bos taurus indicus*) foram disponibilizados pela Embrapa Gado de Leite (MG) e utilizados para as análises de associação genômica. Além disso, dados de sequenciamento de 53 animais do 1000 “Bulls Project” deram origem à população de referência de imputação. Visando verificar a eficiência de imputação, foram testados diferentes cenários quanto a sua acurácia de imputação por meio da análise “leave-one-out”, utilizando apenas os dados de sequenciamento, que apresentaram eficiências de até 84%, no cenário com todos os 51 animais disponíveis após o controle de qualidade. Também foram verificadas as influências das variantes em baixa frequência na acurácia de imputação em diferentes regiões do genoma. Com a escolha da melhor estrutura da população de referência de imputação e aplicação dos controles de qualidade nos dados de NGS e genômicos, foi possível imputar os 2.237 animais genotipados, que passaram pelo controle de qualidade para dados de sequenciamento e realizar análise de associação genômica para as características produção de leite (PL305), teor de gordura (PG305), proteína (PP305) e sólidos totais (PS305), mensuradas aos 305 dias em animais da raça Gir leiteiro. Para tal, foram utilizados os valores genéticos desregredidos (dEBV) como variável resposta no modelo de regressão múltipla. Regiões de 1Mb que contivessem 100 ou mais variantes com “False Discovery Rate” (FDR) inferior a 0,05, foram consideradas significativas e submetidas a análise de enriquecimento por meio dos termos MeSh (“Medical Subject Headings”). As três regiões significativas (FDR<0,05) para PS305 foram observadas nos cromossomos 11, 12 e 28 e a única região significativa em PG305 foi no cromossomo 6. Tais regiões apresentaram variantes associadas com vias metabólicas da produção de leite, ausentes nos painéis comerciais de genotipagem, podendo representar genes candidatos a seleção.

Palavras-chave: Acurácia de imputação, False Discovery Rate, Next-generation sequence, *Bos taurus indicus*, Bovinocultura de leite

IMPUTATION STRATEGIES AND GENOME-WIDE ASSOCIATION WITH SEQUENCE DATA FOR MILK PRODUCTION TRAITS IN GYR CATTLE

ABSTRACT - Implementing "next-generation sequence" (NGS) data in animal breeding programs represents the latest tool in the use of genotypic data in genomic association models, since all polymorphisms are considered in the associations between phenotypic records and sequencing data. As with any new technology, variant prospecting still represents a computational and cost-effective challenge for large-scale implementation. Front to these challenges, this work sought ways to explore the benefits of using NGS in genomic predictions and overcome the inherent limitations of this process. Phenotypic and genotypic (Illumina Bovine HD BeadChip) records of 2,279 Gir animals (*Bos taurus indicus*) were made available by Embrapa Gado de Leite (MG) and used for genomic association analysis. In addition, sequence data of 53 animals from the 1000 Bulls Project gave rise to the imputation reference population. In order to verify the imputation efficiency, different scenarios were tested for their imputation accuracy through the leave-one-out analysis, using only the sequencing data, which presented efficiencies of up to 84%, in the scenario with all the 51 animals available after quality control. Influences from the low-frequency variants on the accuracy of imputation in different regions of the genome were also verified. After identifying the best reference population structure of imputation and applying the quality controls in the NGS and genomic data, it was possible to impute the 2 237 genotyped animals that passed in the quality control to sequencing data and perform genomic association analysis for (PL305), fat content (PG305), protein (PP305) and total solids (PS305), measured at 305 days in dairy Gir animals. For this, unregulated genetic values (dEBV) were used as response variable in the multiple regression model. Regions of 1Mb containing 100 or more variants with a False Discovery Rate (FDR) lower than 0.05 were considered statistically significant and submitted to pathways enrichment analysis using the MeSh (Medical Subject Headings) terms. The three significant regions (FDR <0.05) for PS305 were observed on chromosomes 11, 12 and 28 and only one significant region in PG305, was on chromosome 6. These regions presented variants associated with metabolic pathways of milk production, absent in the panels genotyping, and may represent genes that are candidates for selection

Keywords: Accuracy of imputation, False Discovery Rate, Next-generation sequence, *Bos taurus indicus*, Dairy Cattle

CAPÍTULO 1 - CONSIDERAÇÕES GERAIS

1 INTRODUÇÃO

Com a disponibilidade dos painéis de genotipagem, os programas de melhoramento genético animal passaram a incorporar informações do genoma nas análises de predição genética. Com isso foi possível mensurar o parentesco realizado entre os animais utilizando a matriz de relacionamento genômico (Matrix G), por meio de marcadores moleculares. Isso permitiu que as estimativas dos valores genéticos dos animais fossem obtidas com maior acurácia e que fosse possível realizar análises de associação genômica, capazes de determinar regiões do genoma responsáveis pela variação genética aditiva de características fenotípicas de interesse econômico (HAYES; GODDARD, 2010).

A introdução da genômica no melhoramento genético animal também foi responsável pela diminuição do intervalo de gerações, uma das razões fundamentais para sua implementação na bovinocultura de leite (SCHAEFFER, 2006). No entanto, limitações como a necessidade de desequilíbrio de ligação (DL) entre os marcadores dificultam a identificação de variações causais no genoma (RIVAS et al., 2011). O mesmo ocorre com alelos em baixa frequência, pouco presentes em painéis de marcadores “single nucleotide polymorphisms” (SNP) e responsáveis por parte da variância genética aditiva em características quantitativas ou até mesmo causadores de doenças que atingem os rebanhos (LARMER, 2016).

Com o advento dos sequenciadores de nova geração - “next-generation sequence” (NGS), foi possível transpor essas limitações e aprimorar o estudo do genoma considerando toda variação genotípica. Dados de NGS ainda são pouco utilizados em programas de melhoramento genético animal, mas há grande potencial a ser explorado tendo em vista os resultados obtidos principalmente em humanos (NAIDOO et al., 2011; SUN et al., 2016). Os principais fatores limitantes na utilização da NGS são o alto custo com

sequenciamento dos animais e a dependência da taxa de cobertura na qualidade de leitura das variantes (JOHNSON; SLATKIN, 2008).

O desenvolvimento de algoritmos de imputação mais rápidos e precisos e a aplicação do controle de qualidade permitiu a utilização em larga escala de dados de NGS, ampliando o entendimento da estrutura genética das características fenotípicas (SAHANA et al., 2014). Aliado à otimização computacional dos “softwares” os dados de sequenciamento vêm possibilitando o contínuo progresso genético e o aumento nas acurácias das predições genômicas, especialmente na bovinocultura de leite (ZHANG et al., 2016). Para características de resistência a parasitas, controle de doenças e de difícil mensuração a NGS expressa todo seu potencial, uma vez que essas características tendem a ser influenciadas por alelos raros na população ou causadas por mutações recentes, variantes que ainda não estão presentes nos painéis de marcadores SNP. Para as demais características são esperados ganhos com maior captura da variabilidade genética e maior persistência nas estimativas do efeito dos marcadores, além de análises de associação genômica que identifiquem as mutações causais e maior compreensão nos mecanismos de herança (HAYES; GODDARD, 2010).

2 OBJETIVOS

2.1 Objetivo geral

Os objetivos do presente trabalho foram verificar estratégias de imputação em dados de sequenciamento e conduzir análise de associação genômica para investigar genes e regiões genômicas associadas às características de produção de leite, teor de gordura, proteína e sólidos totais aos 305 dias de produção, utilizando dados imputados para sequenciamento em animais da raça Gir (*Bos taurus indicus*).

2.2 Objetivos específicos

- (1) Avaliar se a estrutura e a inclusão de animais de outras raças na população de referência influenciam a acurácia de imputação, principalmente em variantes com baixa frequência no genoma.
- (2) Identificar regiões no genoma associadas à variação do valor genético das características estudadas, por meio de análises de associação ampla do genoma.
- (3) Investigar possíveis genes funcionais e vias metabólicas associadas às regiões do genoma em que as variantes foram significativas na análise de associação ampla do genoma.

3 REVISÃO DE LITERATURA

3.1 Sequenciamento de nova geração

O primeiro genoma sequenciado foi a bactéria *Haemophilus influenzae* (FLEISCHMANN et al., 1995), seguido pela *Saccharomyces cerevisiae*, primeiro organismo eucarioto com todo genoma sequenciado (GOFFEAU et al., 1996). Esses genomas foram sequenciados pela metodologia de Sanger, a mesma utilizada na obtenção do genoma humano (VENTER et al., 2001), demandando grande esforço laboratorial e financeiro. Com o advento dos sequenciadores de nova geração - “next-generation sequence” (NGS), houve um avanço na obtenção de dados de sequenciamento, uma vez que foi possível sequenciar as mais de três bilhões de pares de bases do genoma humano em apenas 24 horas.

Em dados de NGS, as sequências são obtidas por milhões de pequenos e redundantes fragmentos de DNA obtidos de forma automatizada pelos sequenciadores (NG; KIRKNESS, 2010). Esses pequenos fragmentos são alinhados de duas maneiras. Se há um genoma de referência para a espécie em questão, usa-se essa referência como mapa na montagem dos fragmentos (GNERRE et al., 2009). Quando essa referência não existe, as sequências devem ser construídas “de novo”. Os fragmentos devem ser alinhados por

meio da sobreposição dos fragmentos sequenciados, como um quebra-cabeça (BAKER, 2012).

Sequenciadores de NGS são capazes de gerar grande quantidade de fragmentos de forma automática, permitindo que regiões sejam sequenciadas diversas vezes (taxa de cobertura), o que é fundamental para a detecção de *loci* polimórficos. Portanto, em análises genômicas utilizando dados de NGS é possível selecionar diretamente os *loci* de características quantitativas (QTL), o que pode gerar previsões genéticas mais acuradas e maiores persistências das estimativas do efeito dos marcadores. Ou seja, a utilização de dados de NGS na estimativa do efeito de um marcador genético será mais consistente ao longo das gerações do que quando se utiliza painéis de genotipagem, diminuindo a perda de acurácia das estimativas, causada pela recombinação genética (MEUWISSEN; GODDARD, 2010).

O primeiro genoma sequenciado em bovinos foi oriundo do Bovine Genome Sequencing and Analysis Consortium et al. (2009), que sequenciaram uma vaca da raça Hereford a uma cobertura de aproximadamente 7X, por meio da metodologia de Sanger. Os autores encontraram ao menos 22.000 genes, sendo 14.345 deles ortólogos entre outras sete espécies de mamíferos. Esses genes, em sua maioria, estavam associados a vias metabólicas comuns às espécies. No entanto, os outros 7.655 genes foram associados a características específicas em bovinos, como produção de leite e carne.

O projeto 1000 “Bulls Genomes” propôs reunir sequências de animais do mundo todo, criando assim um banco de dados que permitisse diversos estudos genômicos (DAETWYLER et al., 2014). Isso vem resultando em estudos que contornam as dificuldades em se trabalhar com dados de NGS, como a adoção de critérios de controle de qualidade padronizados, que permitam maior confiabilidade nas bases lidas pelos sequenciadores. Além disso, há grande esforço na otimização de análises genômicas, uma vez que o volume de dados gerado é muitas vezes superior ao dos painéis de genotipagem, o que demanda maior poder computacional. Com isso será possível realizar análises de associação ampla do genoma (GWAS), que permitam investigar mutações causais, e elevar a acurácia da previsão dos valores genéticos de características de produção.

Binsbergen et al. (2014), em animais com genótipos imputados para NGS da raça Holstein Friesian, não obtiveram ganhos nas predições genômicas quando se comparou aos resultados obtidos no painel de alta densidade de marcadores SNP (HD). Os autores relataram ainda que dados de NGS dependem de ajustes nas metodologias para sua melhor utilização e que o processo de imputação demanda maior número de animais na população de referência de imputação do que em análises com dados de painéis de genotipagem. Estudos demonstram as vantagens a serem exploradas nos dados de NGS, como Eck et al. (2009), que encontraram 2,44 milhões de variantes em animais da raça Fleckvieh, sendo que 82% dessas variantes eram desconhecidas até então. Para se analisar toda essa variabilidade genética, análises com dados de NGS necessitam de um grande conjunto de animais sequenciados, para que as análises não percam o poder de predição.

Em animais zebuínos (*Bos taurus indicus*), poucos estudos com NGS foram realizados até o momento. Liao et al. (2013), em um “pool” gamético com 14 animais sendo 3 da raça Gir, identificaram 9.990.733 variantes, 604 inserções e 308 deleções. Entre as variantes, 62,34% representaram novos polimorfismos no genoma e 83,62% das inserções e deleções ainda não haviam sido reportadas. Esse resultado indicou que há muitas variantes que podem explicar a parcela genética das diferenças fenotípicas entre os animais. Além disso, a busca por alternativas metodológicas para a implementação de dados de NGS em análises genômicas é primordial para sua maior utilização e, consequentemente, diminuição nos custos com o sequenciamento dos animais. Esse processo poderá levar a maior utilização de dados de NGS em modelos de predição genômica, seguindo a evolução das análises computacionais, assim como foi no início da utilização de painéis de genotipagem.

3.2 Imputação dos genótipos

Utilizando dados simulados, Meuwissen et al. (2001) apresentaram os ganhos em acurácia ao introduzir as informações do DNA na predição dos valores genéticos em programas de melhoramento animal. No entanto, em dados reais, o custo com a genotipagem em alta densidade dos animais dificultou a implementação da seleção genômica em rebanhos comerciais. Com a imputação de genótipos foi possível contornar esse problema, uma vez que a

maioria dos animais do rebanho poderia ser genotipada com painéis de menor densidade de marcadores SNP, a um menor custo, sendo posteriormente imputados para painéis HD. Com o genótipo dos animais imputados para painéis em HD, foi possível implementar a seleção genômica de forma mais eficiente e aumentar o progresso genético, especialmente na bovinocultura de leite (HAYES et al., 2009).

Os métodos de imputação buscam por padrões de haplótipos em uma população de referência, genotipada com painéis de maior densidade, para estimar os genótipos não observados em um conjunto de dados genotipados com painéis de menor densidade de marcadores SNP. Várias abordagens foram propostas para determinar os subconjuntos de haplótipos da população de referência, visando maximizar a acurácia de imputação da população a ser imputada (HOWIE et al., 2011). As abordagens mais difundidas nos programas que realizam análises de imputação baseiam-se em informações familiares, usando as regras da segregação Mendeliana por meio dos registros do pedigree e/ou informações populacionais; em que os genótipos imputados são estimados por meio do desequilíbrio de ligação mensurado entre marcadores na população de referência (HAYES et al., 2012).

Como todo processo de estimação, a imputação de dados é sujeita a erros de predição que são influenciados por diversos fatores, entre os quais a estrutura da população, o tamanho efetivo da população de referência, o número de SNP no painel de baixa densidade, a frequência alélica dos marcadores, o nível de parentesco entre as populações de referência e a que será imputada e as metodologias empregadas no procedimento de imputação (HAYES et al., 2012; PAUSCH et al., 2013).

A eficiência de imputação pode ser medida pela acurácia, que quantifica a taxa de acerto na predição das regiões imputadas. Chud et al., (2015), ao imputar diferentes densidades de painéis de genotipagem em bovinos, relataram acurácias de até 0,96 no cenário com maior nível de parentesco entre os indivíduos da população de referência e a população que foi imputada. Os resultados foram superiores aos observados por Ventura et al. (2014), que ao imputar um painel de 6k para HD em bovinos, obtiveram acurácia de 0,80. Os mesmos autores, ao imputarem o painel 50k para HD conseguiram acurácias de 0,91.

Há grande variação nos resultados das imputações devido à quantidade de fatores que influenciam a sua eficiência, sendo a diferença entre as densidades dos painéis de genotipagem das populações de referência e da população a ser imputada um dos fatores que mais influenciam na acurácia de imputação. Isso pode dificultar a imputação de painéis em HD para dados de NGS, à medida que uma grande quantidade de haplótipos deverá ser estimada entre as densidades de genotipagem e de sequenciamento. No entanto, a imputação entre painéis em HD e NGS tende a ser mais eficiente com a utilização de populações multirraciais, ao contrário do observado em imputações entre painéis de genotipagem. Isso porque a população de referência sequenciada é capaz de capturar maior número de haplótipos ancestrais na população de referência do que os painéis de genotipagem (BAES, 2015). Além disso, a inclusão de animais de diferentes raças na população de referência eleva as chances de haver múltiplas cópias de determinados alelos, o que favorece na construção dos haplótipos a serem imputados (LI et al., 2011).

Análises de GWAS possibilitam que variantes comuns expliquem parte da variância genética das características quantitativas. Portanto, com a inclusão de variantes raras presentes nos dados de NGS são esperados ganhos na quantidade da variância genética aditiva capturada pelos modelos de regressão múltipla.

Variantes comuns podem ser definidas com “minor allele frequency” (MAF) superior a 0,05, enquanto as variantes raras são aquelas com MAF abaixo de 0,05. A eficiência de imputação com dados de NGS é fortemente influenciada pela acurácia de imputação dessas variantes raras no genoma (HOWIE et al., 2011). Li et al. (2011), em estudo com humanos, relataram que para se atingir acurácia de imputação de 0,70 em variantes com MAF inferior a 0,05 são necessários 1.200 indivíduos na população de referência de imputação. Portanto, o tamanho da população de referência também se mostrou extremamente influente na imputação dos dados de NGS, principalmente entre as variantes raras.

Em bovinos, a redução de 88 para 44 animais na população de referência ocasionou a redução de até 9% na acurácia de imputação entre o painel de 50k e dados de NGS (BINSBERGEN et al., 2014). Devido as

populações de bovinos terem menor tamanho efetivo da população (N_e) em relação às de humanos, é possível que o tamanho da população de referência exerça menor influência na acurácia de imputação.

3.3 Raça Gir e características avaliadas

O Brasil é detentor do segundo maior rebanho bovino do mundo com 212,3 milhões de cabeças, sendo considerado o maior rebanho no aspecto comercial. Com esse rebanho, o país ocupa a posição de maior exportador de carne bovina, segundo maior na produção de carne (USDA, 2014) e quarto no ranking mundial de produtores de leite de vaca, com produção de 34 milhões de toneladas de leite em 2013 (FAO, 2015). Essa posição de destaque da pecuária brasileira no cenário mundial teve início com a importação de diferentes raças de animais zebuínos (*Bos taurus indicus*), entre os anos de 1870 e 1962. Foram importados 6.262 animais, sendo que destes 700 animais eram da raça Gir, que devido a sua rusticidade, longevidade produtiva e boa produção leiteira o tornaram importante animal na cadeia produtiva do Brasil (CURI, 2002).

A raça Gir é amplamente utilizada para a produção de leite e pode ser cruzada com raças taurinas na busca de animais que reúnam a adaptabilidade ao clima tropical do Gir com os melhores desempenhos produtivos das raças taurinas (LEDIC et al., 2002). Presente em diferentes estados desde a década de 60, o Gir teve seu teste de progênie iniciado em 1985, sob responsabilidade da Associação Brasileira dos Criadores de Gir Leiteiro (ABCGIL) que posteriormente associou-se a Empresa Brasileira de Pesquisa Agropecuária (Embrapa) Gado de Leite em Juiz de Fora (MG). Juntas, estas coordenam o Programa Nacional de Melhoramento do Gir Leiteiro (PNMGL) que tem como principal objetivo o melhoramento genético da raça por meio da identificação e seleção de animais geneticamente superiores para as características de produção, conformação e manejo (VERNEQUE et al., 2011).

O programa de melhoramento genético dos animais da raça Gir abrange diversos rebanhos existentes no país. Com base nas informações de progênies e de pedigree são realizadas as avaliações genéticas, que incluem dados de produção e qualidade do leite (PANETTO et al., 2015). A raça vem mostrando o quanto está inserida no mercado nacional de gado de leite, com mais de 400 mil animais registrados pela Associação Brasileira dos Criadores de Zebu

(ABCZ), com crescimento de aproximadamente 70% no registro de novos animais até 2013 (ABCZ, 2013). Essa expansão corrobora os dados de comercialização de sêmen, uma vez que 47,8% de todo o sêmen de gado leiteiro comercializado no Brasil no ano de 2010 foi da raça Gir. Países como Angola, Canadá, Colômbia e Uruguai tem investido no potencial produtivo da raça, utilizando o material genético em cruzamentos com raças taurinas. Somente no ano de 2010, 37.491 doses de sêmen foram vendidas a esses países, que por meio de inseminação artificial, vêm incorporando a raça Gir aos seus plantéis (ASBIA, 2013).

Os programas de melhoramento genético realizam avaliações genéticas para várias características de importância econômica. Na bovinocultura de leite, o valor genético da produção de leite, teor de gordura, proteína e sólidos totais são, geralmente, os principais critérios de seleção. Esses critérios são usados para indicar animais superiores candidatos à seleção, visando promover o progresso genético do rebanho. Cerca de 23 características são avaliadas pelo PNMGL, sendo a produção de leite a de maior expressão econômica.

Tradicionalmente a produção de leite é mensurada pela quantidade de leite produzido em 305 dias (PL305), ou seja, a produção de leite da vaca em quilos durante todo seu período de lactação. Essa é a principal medida no critério de seleção em programas de avaliação genética de bovinos leiteiros no Brasil (LEDIC et al., 2002; PEREIRA et al., 2012). A razão para se usar PL305 está associada ao período ideal de lactação em raças leiteiras, algo em torno de 10 meses. Por sua vez, a escolha desse período de tempo permite a vaca um intervalo entre as lactações de 60 dias, garantindo a correta regeneração dos tecidos da glândula mamária do animal, além de garantir novo acúmulo de reservas nutricionais para as lactações seguintes. Esse intervalo também permite que a vaca tenha um parto ao ano, o que é desejado no sistema de produção atual. No entanto, animais da raça Gir têm em média lactações menores que 305 dias e intervalo de parto maiores, quando comparado a outras raças selecionadas para produção de leite (GONZÁLEZ-HERRERA et al., 2008).

Quando a avaliação genética é feita antes do animal chegar aos 305 dias de produção de leite, chamada lactação acumulada até os 305 dias, uma

alternativa é estimar PL305. Essa estimativa é baseada no somatório das aproximações diárias da produção de leite, valor proveniente do controle leiteiro, os quais são realizados periodicamente em intervalos diários, semanais ou mensais, dependendo do sistema produtivo. Uma alternativa é considerar apenas a lactação encerrada, ou seja, apenas animais com medidas até 305 dias. A adoção da lactação encerrada diminui possíveis erros associados à necessidade de se estimar os valores fenotípicos de produção em vacas com a lactação ainda em curso, evitando viés na avaliação genética causado por fatores de correção.

Albuquerque et al. (1999), em estudo que contou com 1.402 animais da raça Gir, observaram média geral de PL305 de 2.457 ± 141 kg com herdabilidade e repetibilidade de 0,15 e 0,44, respectivamente. As estimativas foram obtidas considerando as três primeiras lactações em modelo animal. Essas estimativas foram inferiores aquelas obtidas por Ledic et al. (2002), que com rebanho de 3.185 animais da raça Gir obtiveram médias para PL305 de 2.746 ± 1.299 kg e herdabilidade de 0,19, usando as três primeiras lactações e modelo animal. González-Herrera et al. (2008), em rebanho de animais Gir com 2.440 vacas e considerando apenas a primeira lactação, estimaram herdabilidades que variaram de 0,30 e 0,29 em análises uni e bicaracterísticas, respectivamente. Os mesmos autores relataram produção de PL305 inferior aos anteriores, com média de $2.170,24 \pm 927,11$ kg. Prata et al. (2015) e Panetto et al. (2015), usando modelo animal, obtiveram herdabilidades de 0,33 e 0,25, e 2.965 ± 1.558 e 3.042 ± 1.641 kg de média de produção de PL305, respectivamente. Essas estimativas indicaram contínuo progresso genético para a principal característica do critério de seleção da raça Gir.

Visando atender ao mercado consumidor, que cada vez mais exige produtos de qualidade, características relacionadas à qualidade do leite vêm sendo gradativamente incluídas como critérios de seleção, como os teores de proteína e gordura que possuem alto valor econômico para empresas que produzem queijo e derivados do leite. Trabalhos recentes que incluíram o estudo dessas características apresentaram valores médios para produção de gordura (PG) variando entre $112,38 \pm 58,40$ e $124 \pm 69,00$ kg. Para a produção de proteína (PP), foram obtidas médias em torno de $90,14 \pm 48,81$ e $104 \pm 57,00$ kg por lactação. As estimativas de herdabilidade variaram entre 0,20 e

0,24 para PG e 0,26 e 0,21 para PP, próximos aos obtidos para PL305 (PRATA et al., 2015; PANETTO et al., 2015). Quanto às estimativas de correlação genética, os mesmos autores relataram resultados que variaram entre 0,92 entre PL305 e PG e 0,98 para PL305 e PP, considerados de alta magnitude e sentido favorável para atender aos objetivos de seleção, de produzir maior quantidade de leite com maior quantidade de proteína e gordura.

4 REFERÊNCIAS

ALBUQUERQUE, M. S. M.; FREITAS, M. A. R.; TEODORO, R. L. Genetic and phenotypic parameters of productivity traits on the first three lactations in gyr cattle herds. **Genetics and Molecular Biology**, n. 22, v. 2, p. 177-181, 1999.

ASSOCIAÇÃO BRASILEIRA DE INSEMINAÇÃO ARTIFICIAL – ASBIA. Relatório estatístico de importação, exportação e comercialização de sêmen. [2013]. Disponível em: <<http://www.asbia.org.br>> Acesso em: 08/07/2016.

ASSOCIAÇÃO BRASILEIRA DOS CRIADORES DE ZEBU – ABCZ. [2013]. Disponível em: <<http://www.abcz.org.br>> Acesso em: 08/07/2016.

BAES, C. F. Across-breed imputation with whole genome sequence data in dairy cattle. In: PLANT AND ANIMAL GENOME XXIII CONFERENCE, 2015, San Diego - USA.

BAKER, M. De novo genome assembly: What every biologist should know. **Nature Methods**, v. 9, n. 4, p. 333-337, 2012.

BINSBERGEN R.; BINK, M. C.; CALUS, M. P. L.; 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**, p. 46:41, 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**, v. 16, p. 99, 2015.

CURI, R. A.; LOPES, C. R. Evaluation of nine microsatellite loci and misidentification paternity frequency in a population of Gyr breed bovines. **Brazilian Journal of Veterinary Research and Animal Science**, v. 39 p. 129-135, 2002.

DAETWYLER, H. D.; CAPITAN, A.; PAUSCH, H.; STOTHARD, P.; VAN BINSBERGEN, R.; BRONDUM, R. F.; et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. **Nature Genetics**, v. 46, n. 8, p. 858-65, 2014.

ECK, S. H.; BENET-PAGÈS, A.; FLISIKOWSKI, K.; MEITINGER, T.; FRIES, R.; STROM, T. M. Whole genome sequencing of a single *bos taurus* animal for single nucleotide polymorphism discovery. **Genome Biology**, v. 10, n. 8, p. R82, 2009.

FLEISCHMANN, R.; ADAMS, M.; WHITE, O.; CLAYTON, R.; KIRKNESS, E.; KERLAVAGE, A.; BULT, C.; TOMB, J.; DOUGHERTY, B.; MERRICK, J. Whole-genome random sequencing and assembly of *Haemophilus influenzae*. **Science**, v. 269, n. 5223, p. 496–512, 1995.

FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS – FAO. **FAO/Statistical databases - FAOSTAT**. [2015]. Disponível em: <<http://faostat3.fao.org/>> Acesso em: 08/07/2016.

GNERRE, S.; LANDER, E. S.; LINDBLAD-TOH, K.; JAFFE, D. B. Assisted assembly: How to improve a de novo genome assembly by using related species. **Genome Biology**, v. 10, n. 8, p. R88, 2009.

GOFFEAU, A.; BARRELL, B. G.; BUSSEY, H.; DAVIS, R. W.; DUJON, B.; FELDMANN, H.; GALIBERT, F.; HOHEISEL, J. D.; JACQ, C.; JOHNSTON, M.; LOUIS, E. J.; MEWES, H. W.; MURAKAMI, Y.; PHILIPPSSEN, P.; TETTELIN, H.; OLIVER, S. G. Life with 6000. Genes. **Science**, v. 274, n. 5287, p. 546–567, 1996.

GONZÁLEZ-HERRERA, L. G.; EL FARO, L.; ALBUQUERQUE, L. G.; TONHATI, H.; MACHADO, C. H. C. Parâmetros genéticos para produção de leite no dia do controle e para produção de leite até 305 dias nas primeiras lactações de vacas da raça Gir. **Revista Brasileira de Zootecnia**, v. 37, n. 10, p. 1774-1780, 2008.

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**, v. 92, p. 433- 43, 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**, v. 80, p. 43-72, 2012.

HAYES, B.; GODDARD, M. Genome-wide association and genomic selection in animal breeding. **Genome**, v. 53, n. 11, p. 876-883, 2010.

HOWIE, B.; MARCHINI, J.; STEPHENS, M. Genotype Imputation with Thousands of Genomes. **G3: GENES, GENOMES, GENETICS**, v. 1, n. 6, p. 457-470, 2011.

JOHNSON, P. L.; SLATKIN, M. Accounting for bias from sequencing error in population genetic estimates. **Molecular biology and evolution**, v. 25, n.1, p. 199-206, 2008.

LARMER, S. G. Next generation sequencing data in bovine: quality control, imputation, and application. 2016. 142 f. Tese (Doctor of Philosophy) – University of Guelph, Canadá, 2016.

LEDIC, I. L.; TONHATI, H.; VERNEQUE, R. S.; EL FARO, L.; MARTINEZ, M. L.; COSTA, C.N.; PEREIRA, J. C. C.; FERNANDES, L. O.; ALBUQUERQUE, L. G. Estimativa de Parâmetros Genéticos, Fenotípicos e Ambientes para as Produções de Leite no Dia do Controle e em 305 Dias de Lactação de Vacas da Raça Gir. **Revista Brasileira de Zootecnia**, v. 31, n. 5, p. 1953-1963, 2002.

LI, L.; LI, Y.; BROWNING, S. R.; BROWNING, B. L.; SLATER, A. J.; KONG, X.; APONTE, J. L.; MOOSER, V. E.; CHISSOE, S. L.; WHITTAKER, J. C.; NELSON, M. R.; EHM, M. G. Performance of genotype imputation for rare variants identified in exons and flanking regions of genes. **PLoS One**, v. 6, n. 9, p. e24945, 2011.

LIAO, X.; PENG, F.; FORNI, S.; MCLAREN, D.; PLASTOW, G.; STOTHARD, P. Whole genome sequencing of Gir cattle for identifying polymorphisms and loci under selection. **Genome**, v. 56, p. 592–598, 2013.

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

MEUWISSEN, T.; GODDARD, M. Accurate prediction of genetic values for complex traits by whole-genome resequencing. **Genetics**, v. 185, n. 2, p. 623-631, 2010.

NAIDOO, N.; PAWITAN, Y.; SOONG, R.; COOPER, D. N.; KU, C.-S. Human genetics and genomics a decade after the release of the draft sequence of the human genome. **Human Genomics**, v. 5 n. 6, p. 577–622, 2011.

NG, P. C.; KIRKNESS, E. F. Whole genome sequencing. **Methods in Molecular Biology**, v. 628, p. 215-26, 2010.

PANETTO, J. C. do C.; VERNEQUE, R. da S.; PEIXOTO, M. G. C. D.; BRUNELI, F. A. T.; MACHADO, M. A.; MARTINS, M. F.; SILVA, M. V. G. B.; ARBEX, W. A.; REIS, D. R. de L.; GERALDO, C. C.; MACHADO, C. H. C.; PEREIRA, M. A.; HORTOLANI, B.; VERCESI FILHO, A. E.; MACIEL, R. DA S.; FERNANDES, A. R. (2015). Programa Nacional de Melhoramento do Gir Leiteiro. Sumário Brasileiro de Touros. Resultado do Teste de Progênie. 6a Prova de Pré-Seleção de Touros - Maio 2015. Embrapa Gado de Leite (Documentos 177), Juiz de Fora.

PAUSCH, H.; AIGNER, B.; EMMERLING, R.; EDEL, C., GÖTZ, K. U., FRIES, R. Imputation of high-density genotypes in the Fleckvieh cattle population. **Genetics Selection Evolution**, v. 13, p. 30-45, 2013.

PEREIRA, R. J.; VERNEQUE, R. S.; LOPES, P.S.; et al. Milk yield pesistency in brasilian Gil cattle on a random regression model. **Genetics a Molecular Research**, v. 2, n. 11, p. 1599-1609, 2012.

PRATA, M. A.; EL FARO, L. E.; MOREIRA, H. L. et al. Genetic parameters for milk production traits and breeding goals for Gir dairy cattle in Brazil. **Genetics and Molecular Research**, v. 4, n. 14, p. 12585-12594, 2015.

RIVAS, M. A.; BEAUDOIN, M.; GARDET, A.; STEVENS, C.; SHARMA, Y.; ZHANG, C. K. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. **Nature Genetics**, v.43, n. 11, p. 1066-1073, 2011.

SAHANA, G.; GULDBRANDTSEN, B.; THOMSEN, B.; HOLM, L.; PANITZ, F.; BRØNDUM, R. F.; BENDIXEN, C.; LUND, M. S. Genome-wide association study using high-density single nucleotide polymorphism arrays and whole-genome sequences for clinical mastitis traits in dairy cattle. **Journal of Dairy Science**, v. 97, n. 11, p. 7258-7275, 2014.

SCHAEFFER, L. R. Strategy for applying genome-wide selection in dairy cattle. **Journal of Animal Breeding and Genetics**, v.123, n.4, p.218-23, 2006.

SUN, X.; SUI, W.; WANG, X.; HOU, X.; OU, M.; DAI, Y.; XIANG, Y. Whole-genome re-sequencing for the identification of high contribution susceptibility gene variants in patients with type 2 diabetes. **Molecular Medicine Reports**, v. 13, p. 3735-3746, 2016.

The Bovine Genome Sequencing and Analysis Consortium. The genome sequence of taurine cattle: A window to ruminant biology and evolution. **Science**, v. 324, n. 5926, p. 522-528, 2009.

UNITED STATES DEPARTMENT OF AGRICULTURE – USDA. **National Agricultural Statistics Service (NASS)**. [2014]. Disponível em: <<https://www.nass.usda.gov/>> Acesso em: 08 jul. 2016

VENTER, J. C.; ADAMS, M. D.; MYERS, E. D. et. al. The Sequence of the Human Genome. **Science**, v. 291, n. 5507, p. 1304-1351, 2001.

VENTURA, R.; LU, D.; SCHENKEL, F.; WANG, Z.; LI, C.; MILLER, S. 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**, v. 92, n. 4, p. 1433-1444, 2014.

VERNEQUE, R. da S.; PANETTO, J. C. do C.; BRUNELI, F. A. T.; GERALDO, C. C.; MACHADO, M. A.; FERNANDES, A. R.; MACIEL, R. da S.; MACHADO, C. H. C.; VERCESI FILHO, A. E.; SILVA, M. V. G. B.; SANTOS, G. G. dos; PEIXOTO, M. G. C. D. (2011). Programa nacional de melhoramento do Gir Leiteiro: Sumário brasileiro de touros: resultado do teste de progênie. Embrapa Gado de Leite (Documentos 145), Juiz de Fora.

ZHANG, Q.; GULDBRANDTSEN, B.; THOMSEN, J. R.; LUND, M. S.; SAHANA, G. Genome-wide association study for longevity with whole-genome sequencing in 3 cattle breeds. **Journal of Dairy Science**, v. 99, n. 9, p. 7289-7298, 2016.

CAPÍTULO 2 - Imputação de painéis de genotipagem para dados de sequenciamento de animais zebuínos

RESUMO – A viabilidade na utilização de dados de sequenciamento (NGS) passa pelo processo de imputação dos dados, uma vez que este permite aumentar o número de animais com informações genotípicas sem elevar demasiadamente os custos. A eficiência da imputação pode ser medida pela acurácia, que sofre entre outros fatores, com a influência das frequências alélicas e da estrutura da população de referência. Os objetivos desse estudo foram avaliar a influência da estrutura genética da população de referência e o efeito das variantes em baixa frequência na acurácia de imputação. Além disso, verificar se há diferenças nos níveis de acurácia de imputação ao longo do genoma de animais de diferentes raças zebuínas. A fim de avaliar essas influências em raças zebuínas, foram escolhidos 53 animais das raças Brahman (BRM), Gir, Nelore (NEL), Uganda Admixed (UGA) e Irã Admixed (IRN), oriundos do “1000 Bull Genomes Project (Run 5.0)” para formar as populações de referência e validação do tipo “leave-one-out”. Dois cenários para imputar painéis de genotipagem em alta densidade para dados de sequenciamento foram propostos. O cenário 51ind, composto por todos os 51 animais disponíveis após o controle de qualidade e o cenário 16ind, formado pelos 16 animais das raças Brahman, Gir e Nelore, definidas como as raças geneticamente mais próximas por análise de componentes principais. As acurácias de imputação desses cenários foram determinadas pela taxa de correlação (COR) e taxa de concordância (TC) entre os dados originais e imputados. Em média, a diferença de acurácia entre as raças foi superior a observada entre os cenários. Para as raças IRN e UGA as acurácias foram em média 15 e 16 pontos percentuais superiores as das raças BRM, GIR e NEL em TC e COR, respectivamente. A partir dos resultados, observou-se que a inclusão de dados de sequenciamento de animais de outras raças na população de referência aumentou as estimativas de TC e COR. Entre os cenários, os ganhos médios em acurácia do cenário 51ind em relação à 16ind foram de 6 e 9 pontos percentuais para TC e COR, respectivamente. A estrutura de uma população de referência multirracial favoreceu a correta imputação dos blocos de haplótipos se comparada a estrutura da população de referência formada apenas por animais geneticamente mais próximos. Além disso, as regiões centrais dos cromossomos apresentaram, em média, maior acurácia de imputação do que as suas extremidades e regiões com maior variabilidade genética e consequentemente maior índice de variantes em baixa frequência tendem a ter menor eficiência de imputação.

1 INTRODUÇÃO

O uso de técnicas de imputação em conjunto de dados genômicos contribuiu para viabilizar a predição genômica em programas de melhoramento genético animal. As metodologias de imputação possibilitaram que marcadores do tipo “single nucleotide polymorphism” (SNP) fossem utilizados de forma ampla, por diminuir os custos com a genotipagem dos animais, pelo fato de estimar o genótipo de determinada população, genotipada em menor densidade, com base em uma população referência, genotipada em maior densidade (CHUD et al., 2015).

Painéis de genotipagem apresentam certas limitações, pois dependem dos níveis de desequilíbrio de ligação (DL) entre marcadores e “quantitative trait loci” (QTL) e do número de polimorfismos disponíveis (IMELFORT et al., 2009; PORTO-NETO et al., 2014). Com a introdução da “Next generation sequencing” (NGS) foi possível transpor essas limitações, pois utilizando as variantes do genoma não é mais necessário depender dos níveis de LD entre marcadores, além de ser possível capturar maior número de variações ao longo do genoma, aumentando a possibilidade de encontrar regiões associadas a características de interesse econômico. No entanto, como toda nova tecnologia, os dados de sequenciamento ainda dependem de estudos para serem utilizados em larga escala de forma eficiente e economicamente viável (LIAO et al., 2013; BINSBERGEN et al., 2014). Um dos desafios é elevar a confiabilidade das variantes lidas no processo de sequenciamento, que sofre grande influência do número de vezes que cada “read” é sequenciada (taxa de cobertura). Para superar essa limitação, critérios de controle de qualidade são adotados visando retirar as variantes com baixa confiabilidade de leitura (LARMER, 2016).

Assim como na implementação dos painéis de genotipagem em programas de melhoramento animal, os custos são um dos fatores que mais limitam a maior utilização dos dados de NGS. Portanto, análises de imputação têm sido utilizadas para aumentar o número de animais com informações em NGS, sem elevar demasiadamente os custos. Devido a necessidade de um certo grau de acurácia para se utilizar os genótipos imputados de forma

eficiente, medidas da acurácia de imputação permitem quantificar a taxa de acerto na predição das regiões imputadas, sendo muita dessa acurácia devida a estrutura genética da população de referência, que deve ser capaz de representar geneticamente a população a ser imputada.

Estudos de imputação com dados de NGS que utilizaram população de referência multirracial, ou seja, diferentes raças na população de referência, apresentaram ganhos de até 5% na eficiência de imputação do cromossomo 1 em variantes comuns e raras (BOUWMAN; VEERKAMP, 2014; BERG et al., 2016). Brøndum et al. (2014), utilizando população de referência multirracial com 242 animais sequenciados, relataram estimativas de acurácia variando de $0,84 \pm 0,22$ a $0,90 \pm 0,20$. Esses resultados diferem dos observados em estudos com painéis de genotipagem, em que é desejável que cada raça possua sua própria população de referência (BERRY; MCCLURE; MULLEN, 2014). Essa maior eficiência das populações de referência multirraciais na imputação pode representar importante estratégia para superar os altos custos com o sequenciamento de animais e viabilizar economicamente sua utilização. Os objetivos desse estudo foram avaliar a influência da estrutura genética da população de referência e o efeito das variantes em baixa frequência na acurácia de imputação e verificar se existem diferenças nos níveis de acurácia de imputação ao longo do genoma de animais de diferentes raças zebuínas.

2 MATERIAL E MÉTODOS

2.1 Dados e controle de qualidade

O conjunto de dados foi composto por dados de sequenciamento de 53 animais, de 5 raças zebuínas: Brahman (BRM), Gir, Nelore (NEL), Uganda Admixed (UGA) e Irã Admixed (IRN), oriundos do “1000 Bull Genomes Project (Run 5.0)” (DAETWYLER et al., 2014). Esses animais foram usados para formar as populações de referência e validação das análises de imputação. Os dados de sequenciamento foram mapeados segundo o genoma de referência UMD 3.1 e os alinhamentos foram feitos com auxílio do “software” BWA 0.5.9, usando as configurações padrão (LI et al., 2010).

O procedimento para determinar as variantes (“variant calling”) foi feito para os 29 cromossomos autossômicos por meio do “Samtools-0.1.18 mpileup” (LI et al., 2008). Com objetivo de diminuir o viés na acurácia das análises de imputação, os dados foram submetidos a controle de qualidade visando garantir que variantes com baixa qualidade de sequenciamento fossem removidas (LI et al., 2009; LIAO et al., 2013; LARMER, 2016). Foram aplicados filtros que removeram as variantes seguindo os critérios: (I) variantes com escore de qualidade média menor que 20, ou seja, 99% das bases lidas corretamente segundo a estimativa de probabilidade de erro de leitura do sequenciador; (II) cobertura mínima menor que 10 ou maior que a média da cobertura mais três vezes o desvio-padrão, excluindo variantes discrepantes quanto a quantidade média de vezes que a variante foi lida; (III) Qualidade de mapeamento menor que 30, critério baseado nas estimativas de erro de posicionamento das bases lidas; (IV) Variantes com distância menor que 10 pb uma da outra, evitando que a mesma variante fosse considerada mais de uma vez; (V) Variantes que não estivessem presentes em ao menos 4 animais, tendo em vista que essas variantes tendem a ter baixa acurácia de imputação (LARMER, 2016). Os filtros de controle de qualidade foram submetidos a todos os animais e aos 29 cromossomos autossômicos por meio dos “softwares” VCFtools v0.1.13 (DANECEK et al., 2011) e pacotes em R snpStats (CLAYTON, 2015) e GWASTools (GOGARTEN et al., 2012).

2.2 Painéis de alta densidade

Os painéis de alta densidade (HD), utilizados no grupo de validação, foram criados com base nas posições compartilhadas entre os dados de sequenciamento e o mapa de genotipagem Illumina Bovine HD (Illumina, Inc., San Diego, CA).

Para cada animal do grupo de referência foi atribuído um painel de alta densidade, dando origem à população de validação. A fim de quantificar a confiabilidade da criação do painel HD, foi calculada a taxa de detecção de SNP (TDS). Essa taxa é influenciada pelos critérios do controle de qualidade e cobertura das sequências e foi calculada como a porcentagem de SNP no mapa de genotipagem Illumina Bovine HD detectada nos dados de sequência (ZHAN et al., 2011).

2.3 Imputação

As análises de imputação foram conduzidas do painel HD para dados de sequência por meio do programa FImpute v2.2 (SARGOLZAEI et al., 2014), com as configurações padrão e ausência de pedigree. A inclusão de registros do pedigree pode aumentar a acurácia de alelos em baixa frequência (BRØNDUM et al., 2013), mas não tem proporcionado ganhos quando o pedigree foi considerado em imputações em que o painel de baixa densidade possui mais de 50 mil marcadores (SARGOLZAEI et al., 2014).

A acurácia de imputação pode ser influenciada por diversos fatores, como a estrutura genética da população, número de animais na população de referência e a frequência alélica das variantes. A fim de verificar esses efeitos neste conjunto de dados, foram criados dois cenários de imputação (51ind e 16ind). O cenário 51ind foi formado por 51 animais disponíveis após o controle de qualidade, permitindo que todas as variantes disponíveis estivessem presentes na população de referência. Para verificar a influência da estrutura da população, foi criado o cenário 16ind, formado por 16 animais das raças Brahman, Gir e Nelore, definidas como as raças geneticamente mais próximas por análise de componentes principais (seção 2.4).

Após a imputação, a validação “leave-one-out” foi realizada em ambos os cenários, sendo cada animal retirado da população de referência, formado por dados de sequência e avaliado no grupo de validação, formado por dados em HD. A acurácia de imputação foi verificada pela correlação média entre os genótipos verdadeiros e os imputados (COR) e pela Taxa de Concordância (TC), definida como a porcentagem média de genótipos imputados corretamente. Essas estatísticas foram calculadas com base nos dados genotípicos, sendo codificados como 0 o homozigoto dominante, 2 o homozigoto recessivo e 1 o heterozigoto em ambos os cenários e dentro de cada raça avaliada. Hickey et al. (2014) relataram que a TC apresenta várias limitações metodológicas, sendo fortemente influenciada pela frequência alélica. Os autores relataram que a correlação média supera esses problemas, sendo assim a medida mais confiável na comparação das eficiências de imputação. A utilização de TC ficou apenas para fins de comparação de resultados com outros trabalhos disponíveis na literatura e comparação com

COR, visando verificar a influência das variantes em baixa frequência na acurácia de imputação.

2.4 Análise de componentes principais (PCA)

Os animais utilizados nas análises de imputação são provenientes de diferentes raças zebuínas. A fim de determinar o distanciamento genético entre os animais, foi efetuada a análise de componentes principais a partir da matriz de relacionamento genômico. No entanto, o grande número de variantes presentes em dados de sequenciamento favorece a formação de “clusters” de variantes em alto desequilíbrio de ligação, o que pode representar viés na análise de PCA (LAURIE et al., 2010). Para diminuir esse viés foi realizado o “pruning” das variantes, retirando as variantes com LD maior que 0,2. O “pruning” e PCA foram realizados no pacote em R SNPRelate (ZHENG et al., 2012).

3 RESULTADOS E DISCUSSÃO

3.1 Dados de sequenciamento e painéis de genotipagem

Dados de sequências foram analisados ao longo de 2.512 Mb, o que correspondeu aos 29 cromossomos autossômicos de 51 animais em 5 raças zebuínas (Tabela 1). A média de cobertura das “reads” considerando todos os animais foi de $10,10 \pm 1,88$, variando de $6,94 \pm 1,23$ nos animais Brahman até $10,32 \pm 0,53$ em animais Uganda Admixed. A precisão na leitura das variantes desempenha importante papel na acurácia de imputação em dados de NGS (LARMER, 2016). Portanto, com a aplicação de controle de qualidade antes das análises de imputação espera-se que variantes e amostras de animais com baixa confiabilidade tenham sido retiradas do conjunto de dados. Assim, dois animais da raça Brahman foram removidos devido à baixa cobertura de leitura em suas sequências, o que resultou na exclusão de mais de 80% das variantes pelo controle de qualidade, restando 51 animais disponíveis para os cenários de imputação. Os controles de qualidade por variantes foram aplicados de forma independente em cada cenário, resultando em diferentes quantidades de variantes e SNP nos grupos de referência e validação, respectivamente (Tabela 2). Devido a diferença no número de animais entre os cenários,

existiram variantes que se tornaram monomórficas e, por isso, foram retiradas do conjunto de dados.

Tabela 1. Número de animais por raça e cobertura de leitura das sequências dos animais após o controle de qualidade.

Raça	N	Cobertura		
		Média (DP)	Min	Max
BRM	5	6,94 (1,23)	5,06	8,79
GIR	6	8,44 (3,41)	4,54	14,06
NEL	5	7,31 (0,69)	6,04	8,11
IRN	9	10,17 (0,89)	9,12	12,15
UGA	26	10,32 (0,53)	9,10	11,15

N: Número de animais; **DP:** Desvio-padrão; **Max;** **Min:** Valores máximos e mínimos; **BRM:** Brahman; **GIR:** Gir **NEL:** Nelore; **UGA:** Uganda Admixed; **IRN:** Irã Admixed.

Em uma população, o nível da variabilidade genética e a escolha dos animais que serão sequenciados influenciam no número de variantes de um conjunto de dados. Ao todo, 13.171.566 variantes estavam presentes na população contendo todos os 51 animais. Essa quantidade de variantes foi superior à relatada por Liao et al. (2013), que usando dados de NGS em bovinos da raça Gir (*Bos taurus indicus*), observaram 9.990.733 variantes polimórficas oriundas de um “pool” gamético composto por 3 animais. Portanto, espera-se que o maior número de variantes encontradas neste trabalho tenha auxiliado na construção dos haplótipos de forma acurada na análise de imputação.

Tabela 2. Número de variantes na população de referência de imputação, quantidade de SNP na população de validação da imputação e taxa de detecção de SNP dos cenários analisados.

Cenários	Variantes na população de referência	SNP na população de validação	TDS (%)
51 animais	13.171.566	355.697	48,38
16 animais	12.819.520	304.797	41,45

SNP: Single Nucleotide Polymorphism; **TDS:** Taxa de Detecção de SNP.

O número de SNP nos painéis HD corresponderam a 2,43% e 2,37% das variantes presentes nas populações de referência de imputação dos cenários 51ind e 16ind, respectivamente. Esses valores foram próximos aos encontrados por Binsbergen et al. (2014) que, ao avaliarem o cromossomo 1 de 114 animais da raça Holstein Friesian, encontraram em média 2,33% das variantes compartilhadas com o painel em HD. A quantidade de SNP nos

painéis HD das populações de validação (Tabela 2) foram próximas aos 290.620 e 320.230 SNP observados em outros trabalhos com animais zebuínos, em que foi utilizado painel HD (SANTANA et al., 2014; NEVES et al., 2014). Liao et al. (2013), a partir de dados de 3 animais sequenciados da raça Gir, relataram valores de TDS de 30,60%. Essa estimativa foi inferior aos 48,38% do cenário com todos os 51 animais e aos 41,45% do cenário com 16 animais. Isso sugere que a estratégia para formar os painéis de baixa densidade (população de validação) foi eficiente, por representar valores próximos aos observados em estudos com painel de genotipagem em animais de raças zebuínas.

3.2 Estrutura da população

A estrutura populacional dos animais é apresentada na análise de componentes principais (Figura 1), sendo possível notar maior proximidade genética entre os animais das raças Brahman, Nelore e Gir. Essa proximidade era esperada, tendo em vista que a raça Brahman tem origem em cruzamentos absorventes de diferentes raças zebuínas, dentre as quais Gir e Nelore (FARIA et al., 2010). Nessas três raças foi possível notar maior distanciamento entre os animais da raça Gir e isso pode ser devido à estratégia de escolha dos animais a serem sequenciados, que tende a buscar animais geneticamente mais distantes dentro do rebanho, a fim de elevar o número de variantes capturadas e melhor representar o grupo de animais (DRUET et al., 2014).

Os animais das raças Uganda Admixed e Irã Admixed representam os animais zebuínos com maior número de sequências depositadas no “1000 Bull Genomes Project (Run 5.0)” e apresentaram grupos dispersos e distantes dos demais animais. Possivelmente, a dispersão dentro do grupo é devido ao fato dos animais serem de raças compostas, o que eleva a distância genética entre os animais. A distância entre esses dois grupos pode estar relacionada à própria distância geográfica na origem desses animais, oriundos respectivamente da África e Oriente Médio (MASON, 1988). Após a aplicação do “pruning”, foram mantidas apenas as variantes com LD menor que 0,2, restando assim 91.604 variantes não redundantes na análise de PCA. Esta estratégia garantiu maior confiabilidade na porcentagem da variância explicada pelos autovalores na PCA, sendo o primeiro componente principal capaz de explicar 4,38% da variação dos dados.

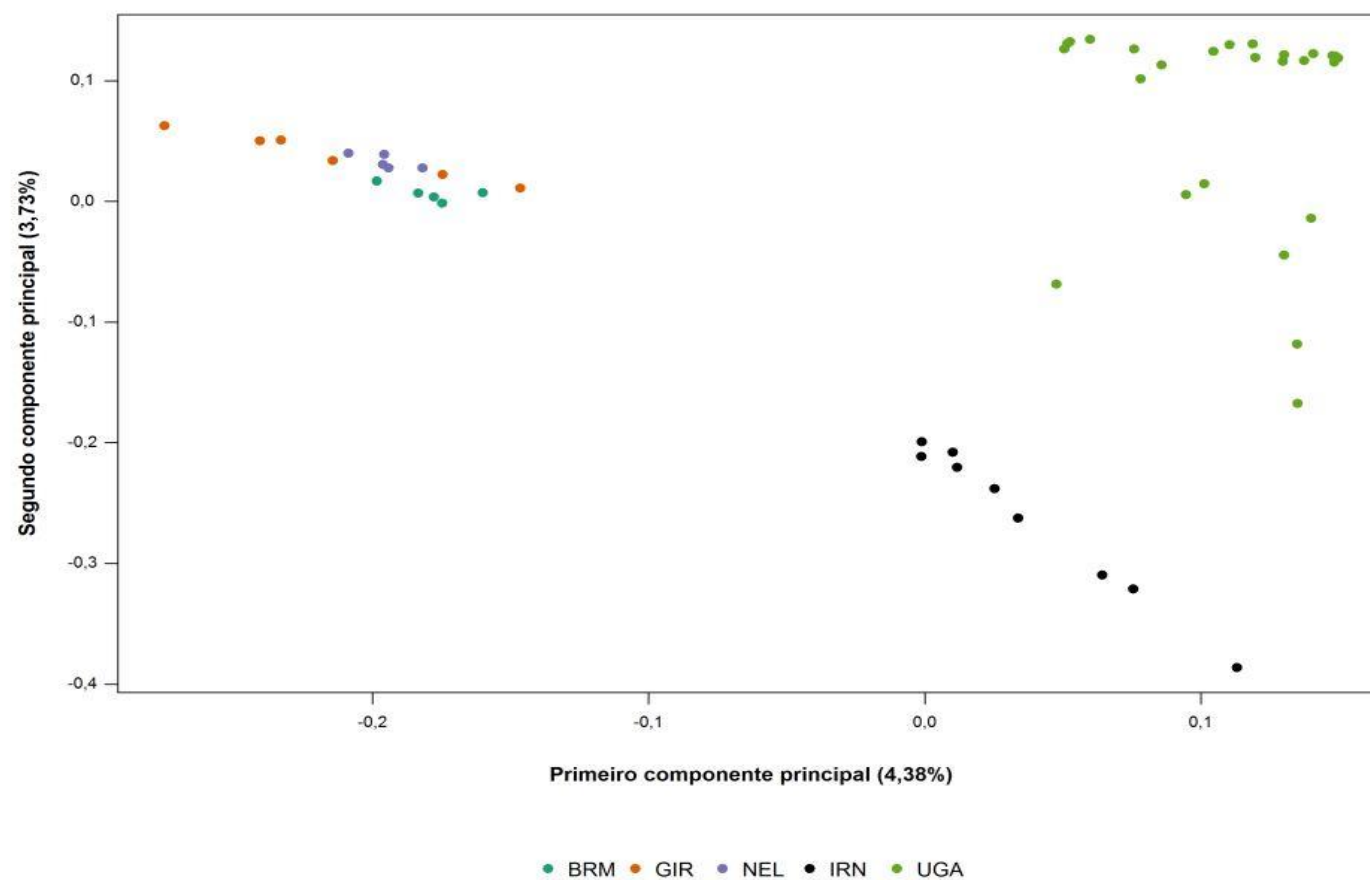


Figura 1. Análise de componentes principais dos 51 animais sequenciados. **BRM:** Brahman; **GIR:** Gir **NEL:** Nelore; **UGA:** Uganda Admixed; **IRN:** Irã Admixed.

3.3 Acurácia de imputação

Em média, a diferença de acurácia de imputação entre as raças foi superior a observada entre os cenários (Tabela 3). Isso porque o número de animais sequenciados na população de referência de imputação foi o fator de maior influência nas acurácias de imputação. Os valores de acurácia de imputação nas raças IRN e UGA foram superiores, em média, 15 e 16 pontos percentuais em relação às das raças BRM, GIR e NEL para a taxa de concordância e correlação, respectivamente. A partir desses resultados, foi possível observar que raças com maior número de animais na população de referência apresentaram as maiores estimativas de acurácia de imputação tanto em TC como para COR.

Comparando os resultados das raças UGA e IRN, foi possível observar também que a diferença entre as raças em TC foi de apenas 2 pontos percentuais. No entanto, os 17 animais a mais da raça UGA em relação a IRN foram suficientes para elevar a magnitude de COR de UGA em relação à IRN em 8 pontos percentuais. Ao comparar as estimativas médias de acurácia nos animais das raças BRM, GIR e NEL, que tinham quase o mesmo número de animais sequenciados, não foram observadas diferenças superiores a 2 pontos percentuais nos valores de TC e COR em ambos os cenários. Portanto, para elevar os valores de acurácia nos animais das raças BRM, GIR e NEL, é necessário que mais animais dessas raças componham a população de referência.

Entre os cenários, os ganhos médios em acurácia do cenário 51ind em relação à 16ind foram de 6 e 9 pontos percentuais para TC e COR, respectivamente (Tabela 3). Essas diferenças são observadas quando são comparadas as acurácias médias entre os animais das raças BRM, GIR e NEL nos dois cenários. Os resultados indicaram que a inclusão dos 35 animais das raças IRN e UGA no cenário 51ind possibilitou o aumento da eficiência de imputação. A presença desses animais auxiliou principalmente a correta imputação de alelos menos frequentes da população, tendo em vista que COR foi superior a TC para todas as raças no cenário 51ind. Ao adotar uma população de referência multirracial em dados de NGS, Bouwman; Veerkamp (2014) relataram ganhos em acurácia de até 13% em relação as imputações que utilizavam populações de referência compostas por apenas uma raça.

Tabela 3. Média e desvio-padrão (parênteses) da taxa de concordância e correlação por raças e cenários com 51 e 16 animais.

Raça	51 animais		16 animais	
	TC	Cor	TC	Cor
BRM	0,68 (0,01)	0,58 (0,03)	0,62 (0,01)	0,48 (0,03)
GIR	0,69 (0,02)	0,60 (0,04)	0,62 (0,01)	0,51 (0,03)
NEL	0,68 (0,01)	0,58 (0,02)	0,62 (0,01)	0,50 (0,02)
Média	0,68 (0,01)	0,59 (0,01)	0,62 (0,00)	0,50 (0,02)
IRN	0,82 (0,04)	0,71 (0,04)	-	-
UGA	0,84 (0,03)	0,79 (0,03)	-	-
Média	0,83 (0,01)	0,75 (0,06)	-	-

TC: Taxa de concordância; Cor: Correlação; BRM: Brahman; NEL: Nelore; UGA: Uganda Admixed; IRN: Irã Admixed.

O processo de validação “leave-one-out” foi escolhido para validar o processo de imputação devido ao número de animais disponíveis para análise e a estrutura da população ser formada por 5 raças distintas. No cenário 51ind, a população de referência composta por 50 animais em cada processo de validação, mais que dobrou o tamanho daquelas do cenário 16ind, em que apenas 15 animais formavam cada população de referência. Com os resultados obtidos foi possível verificar a influência do número de animais na população de referência de imputação, assim como os ganhos em utilizar uma população multirracial na imputação de dados de NGS, principalmente nas variantes em baixa frequência na população.

3.4 Acurácia de imputação ao longo do genoma

A acurácia média dos 29 cromossomos autossômicos no cenário com 51 animais foi de $0,50 \pm 0,29$, variando de $0,55 \pm 0,28$ a $0,42 \pm 0,29$ nos cromossomos 6 e 28 respectivamente (Figura 2). Essa diferença nas estimativas de acurácias entre e ao longo dos cromossomos indicam a relação da eficiência de imputação com regiões específicas do genoma e a influência da taxa de cobertura de sequenciamento, que foi relativamente baixa nesse conjunto de dados. As metodologias de imputação tendem a apresentar menor eficiência na construção dos blocos de haplótipos nas extremidades do genoma, o que diminui a acurácia de imputação nessas regiões. Além disso, essas regiões teloméricas dos cromossomos tendem a ser degradadas e com grande quantidade de nucleotídeos repetidos, o que também dificulta o processo de imputação.

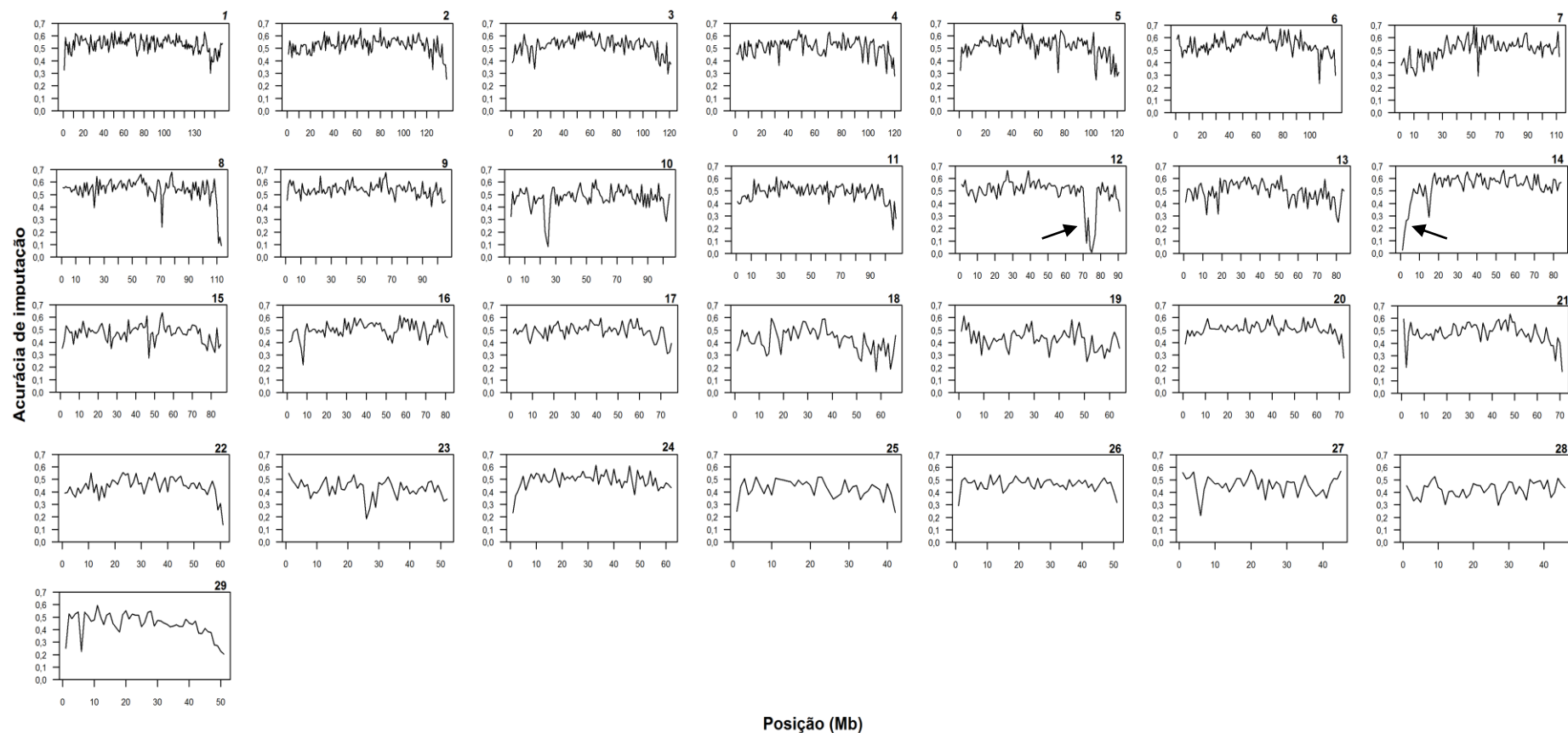


Figura 2. Média da acurácia mensurada pela correlação entre as variantes sequenciadas e imputadas em janelas de 1 Mb ao longo dos 29 cromossomos autossômicos no cenário com 51 animais. Setas nos cromossomos 12 e 14 indicam redução acentuada da acurácia de imputação.

Em geral, as posições centrais dos cromossomos tiveram melhores resultados de acurácia, sendo possível observar perda de acurácia nas extremidades em vários cromossomos. Ressalta-se que o cromossomo 14, mesmo tendo uma das maiores médias de acurácia, apresentou estimativas inferiores a 0,40 nos primeiros 5 Mb do comprimento do cromossomo. A exceção foi o cromossomo 9, em que foi obtida a segunda melhor média de acurácia de imputação, mantendo a acurácia média superior a 0,40 ao longo de todo o cromossomo.

Dentro de cada cromossomo foi possível observar também variações na eficiência de imputação, como no cromossomo 12, que apresentou desvio-padrão de 0,31, a maior variação entre os 29 cromossomos. No entanto, muito dessa variação foi observada pela perda abrupta de acurácia na região entre 70 e 80 Mb. Nessa região do genoma de referência UMD 3.1 do “1000 Bull Genomes Project”, foram observadas regiões codificadoras de proteína de resistência a drogas (MRA) por meio da ferramenta Ensembl, ou seja, regiões com alta variabilidade e consequentemente com alta concentração de alelos em baixa frequência. De maneira similar, Brondum et al. (2013) relataram diferenças na acurácia de imputação entre raças de animais taurinos ao longo do cromossomo 29. Os autores associaram a perda de acurácia em determinadas regiões com maior incidência de variantes com “minor allele frequency” (MAF) menor que 0,3 e a presença de poucos animais na população de referência.

O uso da NGS eleva o número de variantes disponíveis nas análises genômicas. No entanto, grande parte destas variantes está presente em baixa frequência na população, o que se torna um desafio no processo de imputação de dados. Na Figura 3, apresenta-se a média da distribuição da MAF após a aplicação do controle de qualidade no cenário com 51 animais. Das 13.171.566 variantes disponíveis no cenário 51ind, 1.813.475 apresentaram MAF menor que 0,05, ou seja, 12,42% do total de variantes.

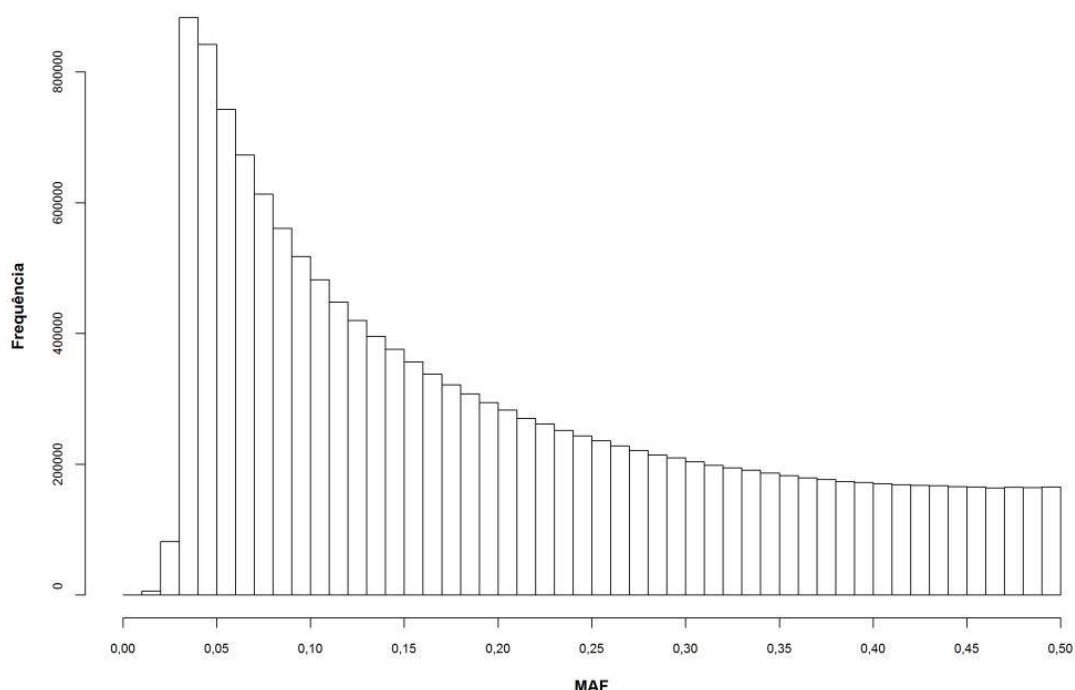


Figura 3. Distribuição das “minor allele frequency” (MAF) após a aplicação do controle de qualidade no cenário com 51 animais.

As medianas dos valores de acurácias de imputação ao longo dos 29 cromossomos em ambos os cenários tiveram suas eficiências influenciadas pela MAF (Figura 4). No cenário 51ind, variantes com MAF inferiores a 5% não apresentaram acurácias maiores que 0,30 e 45,24% das variantes não tiveram acurácias superiores a 0,5. Esse resultado indicou que os 51 animais desse cenário não foram suficientes para imputar variantes com baixa frequência de forma acurada. No entanto, o aumento na frequência das variantes gerou imputações eficientes, chegando a 70% em variantes com MAF superior a 0,25. As médias do cenário 16ind foram fortemente influenciados pelos altos valores dos desvios-padrão. Portanto, não é possível mensurar o real efeito da MAF nesse cenário. Entretanto, a diferença entre os cenários indicou que a inclusão de animais sequenciados de outras raças na população de referência foi capaz de capturar mais da variabilidade genética da população e, conseqüentemente, elevar as acurácias de imputação.

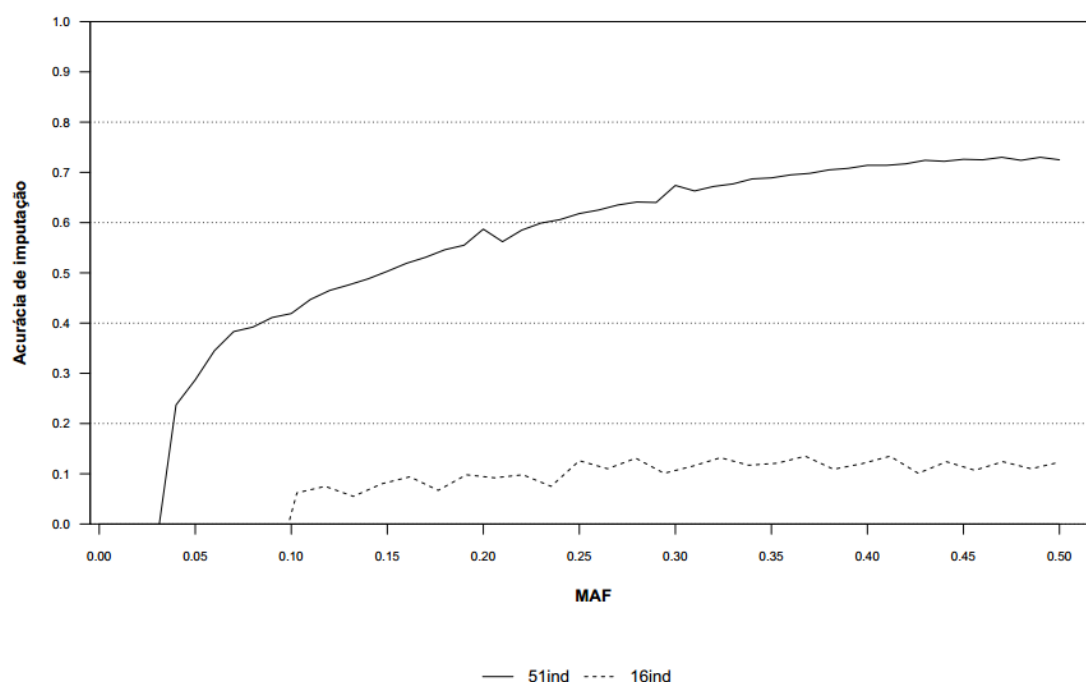


Figura 4. Medianas dos valores de acurácias de imputação mensuradas pela correlação entre as variantes sequenciadas e imputadas em função das “minor allele frequency” (MAF) considerando os 29 cromossomos nos cenários com 51 animais (51ind) e 16 animais (16ind).

4 CONCLUSÃO

A inclusão de animais de diferentes raças na população de referência pode elevar a eficiência de imputação, se comparado a estrutura da população de referência formada apenas por animais geneticamente mais próximos.

Para a condução de um estudo de associação genômica com dados de NGS é necessário maior número de animais na população de referência de imputação.

Existem diferenças na acurácia de imputação ao longo do genoma, dado que regiões centrais dos cromossomos apresentam em média maior acurácia de imputação do que as extremidades. Regiões com maior variabilidade genética e consequentemente maior índice de variantes em baixa frequência tendem a ter menor eficiência de imputação.

REFERÊNCIAS

BERG, I. D.; BOICHARD, D.; LUND, M. S. Comparing power and precision of within-breed and multibreed genome-wide association studies of production traits using whole-genome sequence data for 5 French and Danish dairy cattle breeds. *Journal of Dairy Science*, v. 99, n. 11, p. 8932-8945, 2016.

BERRY, D. P.; MCCLURE, M. C.; MULLEN M. P. Within- and across-breed imputation of high-density genotypes in dairy and beef cattle from medium- and low-density genotypes. **Journal of Animal Breeding and Genetics**, v. 131, n. 3, p. 165-172, 2014.

BINSBERGEN R.; BINK, M. C.; CALUS, M. P. L.; 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**, p. 46-41, 2014.

BOUWMAN, A. C.; VEERKAMP, R. F. Consequences of splitting whole-genome sequencing effort over multiple breeds on imputation accuracy. **BMC Genetics**, v. 15, n. 105, 2014.

BRONDUM M. P.; ZHANG, R. F.; LUND, Q.; S. U.; M.S. Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle. **Journal of Dairy Science**, v. 96, p. 4666-4677, 2013.

BRONDUM, R. F.; GULDBRANDTSEN, B.; SAHANA, G.; LUND, M. S.; SU, G. Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. **BMC Genomics**, v. 15, n. 1, p. 728, 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**, v. 16, p. 99, 2015.

CLAYTON, D. (2015). snpStats: SnpMatrix and XSnpmatrix classes and methods. R package version 1.24.0.

DAETWYLER, H. D.; CAPITAN, A.; PAUSCH, H., STOTHARD, P.; VAN BINSBERGEN, R.; BRONDUM, R. F.; et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. **Nature Genetics**, v. 46, n. 8, p. 858-65, 2014.

DANECEK, P.; AUTON, A.; ABECASIS, G.; ALBERS, C. A.; BANKS, E.; DEPRISTO, M. A. The variant call format and VCFtools. **Bioinformatics**, v. 27, n. 15, p.2156-2158, 2011.

DRUET, T.; MACLEOD, I. M.; HAYES, B. J. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. **Heredity**, v. 112, p. 39-47, 2014.

FARIA L. C.; QUEIROZ, S. A.; VOZZI, P. A.; LÔBO, R. B.; MAGNABOSCO, C.U.; OLIVEIRA, J. A. Variabilidade genética da raça Brahman no Brasil detectada por meio de análise de pedigree. **Pesquisa agropecuária brasileira**, v. 45, n. 10, 2010.

GOGARTEN, S. M.; BHANGALE, T.; CONOMOS, M.P.; LAURIE, C. A.; MCHUGH, C. P.; PAINTER, I.; ZHENG, X.; CROSSLIN, D. R.; LEVINE, D.; LUMLEY, T.; NELSON, S. C.; RICE, K.; SHEN, J.; SWARNKAR, R.; WEIR, B. S.; LAURIE, C. C. GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. **Bioinformatics**, v.28, n. 24, p. 3329-3331, 2012.

HICKEY, J. M.; CROSSAB, J.; BABUC, R.; DE LOS CAMPOS, G. Factors Affecting the Accuracy of Genotype Imputation in Populations from Several Maize Breeding Programs. **American Society of Agronomy**, v. 52, n. 2, p. 654-663, 2014.

IMELFORT, M.; DURAN, C.; BATLEY, J.; EDWARDS, D. Discovering genetic polymorphisms in next-generation sequencing data. **Plant Biotechnology Journal**, v. 7, p. 312-317, 2009.

LARMER, S. G. NEXT GENERATION SEQUENCING DATA IN BOVINE: QUALITY CONTROL, IMPUTATION, AND APPLICATION. 2016. 142 f.. Tese (Doutorado em Filosofia) - University of Guelph (UofG), Guelph, 2016.

LAURIE, C. C.; DOHENY, K. F.; MIREL, D.B. Quality control and quality assurance in genotypic data for genome-wide association studies. **Genetic Epidemiology**, v. 34, p. 591–602, 2010.

LI, H.; DURBIN, R. Fast and accurate long-read alignment with Burrows Wheeler transform. **Bioinformatics**, v. 26, p. 589–595, 2010.

LI, H.; HANDSAKER B.; WYSOKER, A.; FENNELL, T.; RUAN, J.; HOMER, N.; MARTH, G.; ABECASIS, G.; DURBIN, R. 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. **Bioinformatics**, v. 15, n. 16, p. 2078-2079, 2009.

LI, H.; RUAN, J.; DURBIN, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. **Genome Research**, v. 18, p. 1851-8, 2008.

LIAO, X.; PENG, F.; FORNI, S.; MCLAREN, D.; PLASTOW, G.; STOTHARD, P. Whole genome sequencing of Gir cattle for identifying polymorphisms and loci under selection. **Genome**, v. 56, p. 592–598, 2013.

MASON, I. L. World dictionary of livestock breeds (N). C.A.B. International, 1988. p. 31-54, 83, 113.

NEVES, H. R.; CARVALHEIRO, R.; PÉREZ O'BRIEN, A. M.; 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**, p. 46-17, 2014.

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

SANTANA, M. H. A.; UTSUNOMIYA, Y. T.; NEVES, H. H. R.; GOMES, R. C.; GARCIA, J. F.; FUKUMASU, H.; SILVA, S. L.; LEME, P. R.; COUTINHO, L. L.; ELER, J. P.; FERRAZ, J. B. S. Genome-wide association study for feedlot average daily gain in Nellore cattle (*Bos indicus*). **Journal of Animal Breeding and Genetics**, p. 210–216, 2014.

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

ZHAN, B.; FADISTA, J.; THOMSEN, B.; HEDEGAARD, J.; PANITZ, F.; BENDIXEN, C. Global assessment of genomic variation in cattle by genome resequencing and high-throughput genotyping. **BMC Genomics**, v. 12, p. 557, 2011.

ZHENG, X.; LEVINE, D.; SHEN, J.; GOGARTEN, S. M.; LAURIE, C.; WEIR, B. S. A high-performance computing toolset for relatedness and principal component analysis of SNP data. **Bioinformatics**, v. 28, n. 24, p. 3326-3328, 2012.

Capítulo 3 - ASSOCIAÇÃO GENÔMICA AMPLA COM DADOS DE SEQUENCIAMENTO EM CARACTERÍSTICAS DE PRODUÇÃO E QUALIDADE DE LEITE NA RAÇA GIR

RESUMO – Estudo de associação genômica ampla (GWAS) utilizando dados de “next-generation sequence” (NGS) permite que toda a variação do genoma seja verificada quanto ao seu efeito na variação fenotípica. A acurácia dessas associações depende da estrutura genética das características avaliadas, do número de variantes disponíveis e a distribuição de seus efeitos ao longo do genoma. Análises de GWAS foram realizadas para as características produção de leite (PL305), teor de gordura (PG305), proteína (PP305) e sólidos totais (PS305), com base em dados de sequenciamento de todo o genoma de 2.237 animais da raça Gir (*Bos taurus indicus*), imputados a partir de dados genotípicos. A análise de imputação desses genótipos foi feita em duas etapas. Primeiramente os genótipos do painel em 50k foram imputados para o painel de alta densidade (HD), com animais genotipados em HD formando a população de referência de imputação. Posteriormente, esses genótipos foram imputados para dados de NGS, sendo a população de referência de imputação formada por 51 animais sequenciados pelo “1000 Bull Genomes Project (Run 5.0)”. Para a GWAS utilizou-se os valores genéticos desregredidos (dEBV) de cada característica estudada como variável respostas no modelo de regressão múltipla e considerou-se significativas as regiões de 1Mb que contivessem 100 ou mais variantes com “False Discovery Rate” (FDR) inferior a 0,05. Essas regiões foram então submetidas a análise de enriquecimento por meio dos termos MeSh (“Medical Subject Headings”) em que se verificou a relação das variantes com vias biológicas conhecidas. Em PS305, três regiões nos cromossomos 11, 12 e 28 foram declaradas como significativas e foram associadas com vias metabólicas ligadas à produção de leite e sólidos totais. Para PG305, uma região foi significativa no cromossomo 6 e associada ao gene SLIT2, responsável por interações metabólicas entre o sistema imune e endócrino em vacas leiteiras. Grande parte das variantes significativas apresentaram “minor allele frequency” (MAF) inferior a 0,05 e não estavam presentes no painel 50k e HD. Essas ausências indicam que introduzir dados de NGS nas avaliações genéticas dessa população pode representar aumento na captura de regiões que influenciam a variabilidade genética aditiva das características avaliadas e que não são consideradas nos painéis de genotipagem.

1 INTRODUÇÃO

O estudo de associação ampla do genoma (GWAS) se tornou a principal abordagem na identificação de genes a partir de variantes comuns, que podem influenciar a variabilidade genética de características complexas (BUZANSKAS et al., 2014). Variantes comuns, presentes nos painéis comerciais de genotipagem permitiram que características de produção, reprodução e resistência a doenças de diferentes espécies fossem associadas com regiões específicas do genoma. No entanto, grande parte da variabilidade genética destas características ainda não foi atribuída a nenhuma região específica do genoma. Muito disso se deve ao caráter poligênico que a maioria das características quantitativas de interesse econômico possuem. Essas apresentam estrutura genética ainda desconhecida, sendo muitas vezes controladas por uma grande quantidade de genes de pequeno efeito e presentes ao longo de todo o genoma, tornando difícil a associação entre as variações fenotípicas e as regiões do genoma utilizando painéis de genotipagem.

Em bovinos, estudos de GWAS têm sido conduzidos desde o desenvolvimento dos painéis de genotipagem (MAI et al., 2010; SAHANA et al., 2010). Trabalhos com painéis de moderada (50k) a alta densidade (HD) de marcadores do tipo “single nucleotide polymorphism” (SNP) permitiram maior compreensão da estrutura genética de características produtivas. Além disso, esses painéis capturaram com maior acurácia a variabilidade genética de características tradicionais em programas de melhoramento genético (ZHANG et al., 2014). As associações entre regiões do genoma e características fenotípicas são dependentes dos níveis de desequilíbrio de ligação (DL) entre marcadores e regiões de *loci* de características quantitativas (QTL) (BERG; BOICHARD; LUND, 2016). Quando essas regiões apresentam marcadores em baixa frequência, as associações tendem a ser mais difíceis de serem obtidas, devido a baixa contribuição que dessas regiões na variabilidade genética (KIM et al., 2015). A necessidade de aumentar a eficiência em detectar regiões que afetam a variância genética aditiva das características de importância econômica, principalmente em regiões com alto índice de alelos em baixa

frequência, levou ao desenvolvimento de tecnologias de sequenciamento de nova geração – “next generation sequence” (NGS) (LU et al., 2011).

Com a implementação da NGS foi possível associar variantes raras a doenças em humanos, permitindo que doenças fossem melhor compreendidas e, conseqüentemente, fosse aprimorado o seu tratamento (CHEN et al., 2014). Há muito potencial a ser explorado com a utilização de dados de NGS, que permite que milhões de novas variáveis sejam incluídas nos modelos de associação. Entretanto, o elevado custo de cada animal sequenciado e as taxas de erro de leitura ainda são desafios a serem superados para sua implementação em larga escala (LU et al., 2011). Para que a associação entre variantes e o fenótipo avaliado seja eficiente, são necessários: (I) indivíduos sequenciados que representem geneticamente a população estudada; (II) dados fenotípicos com alta acurácia individual e (III) metodologias estatísticas apropriadas para o grande volume de dados gerados (PRICE et al., 2010).

Dados de animais sequenciados são escassos, se compararmos ao volume de dados genotipados disponíveis, mas com o auxílio de metodologias de imputação, dados de sequenciamento têm sido obtidos com a imputação dos genótipos já disponíveis. Essa alternativa permite que grandes conjuntos de dados em NGS estejam disponíveis para serem utilizados em GWAS, sem a necessidade do alto custo de sequenciar rebanhos inteiros. Daetwyler et al. (2014), utilizando animais sequenciados de diferentes raças, verificaram que a imputação de dados genotípicos para dados de sequência poderia ser uma alternativa para detectar novas variantes, aumentando assim a compreensão da estrutura de características de interesse econômico em gado de leite.

As estratégias para escolha dos animais que irão compor a população de referência visam capturar o maior número de variantes presentes na população em estudo. Druet et al. (2014) propuseram estratégias na escolha dos animais que irão compor a população de referência, sendo a melhor estratégia a combinação de animais sequenciados e geneticamente mais distantes uns dos outros.

Em dados de NGS, variantes comuns tendem a ter elevada eficiência de imputação (BINSBERGEN, et al., 2014), mas os mesmos resultados não são observados para variantes raras, ou seja, “minor allele frequency” (MAF) inferior a 5% (BAES, 2015). Em busca de maiores valores de acurácia, animais

selecionados de diferentes raças tem dado origem a populações de referência multirraciais, que tem elevado a eficiência de imputação nas variantes comuns e raras (BOUWMAN; VEERKAMP, 2014; BERG et al., 2016). A inclusão de várias raças na população de referência de imputação representa importante estratégia para superar os altos custos com o sequenciamento de animais.

Os resultados das análises de GWAS são fortemente influenciados pela correta imputação dos dados, que deve fornecer alta acurácia nas variantes imputadas (RUTKOSKI, et al., 2013). Parte dos erros no processo de imputação é devido à erros de posição de determinado marcador no mapa de genotipagem. Esse tipo de erro se torna mais grave quando o marcador é, ou está próximo a uma região de QTL. Erros de imputação fazem as associações perderem o sentido biológico, levando a associações falso-positivas. No que se refere à imputação de genótipos para sequência, esses erros tendem a ser frequentes, devido à presença de maior número de variantes raras. Isso reforça a necessidade da aplicação de rigoroso controle de qualidade, mas sem retirar excessivamente variantes em baixa frequência, que podem estar associadas com as características avaliadas.

A população avaliada nesse estudo foi composta por animais da raça Gir (*Bos taurus indicus*), raça leiteira de grande importância no Brasil. (REIS FILHO et al., 2010). A intensa utilização de animais Gir para produção de leite se deu por sua robustez e resistência a parasitas, características ideais em rebanhos no Brasil. Várias características fenotípicas são avaliadas no programa de melhoramento genético da raça, entre as quais a produção de leite, teores de gordura, proteína e sólidos totais, sendo consideradas as produções acumuladas até os 305 dias de lactação nas avaliações. Tais características são de fácil mensuração e boa resposta à seleção genômica. No entanto, pode haver um percentual da variabilidade genética não capturada pelas metodologias utilizando painéis de genotipagem, permitindo que novas abordagens sejam empregadas e que o progresso genético de tais características seja otimizado.

2 OBJETIVOS

O objetivo deste estudo foi associar regiões do genoma com características de interesse econômico para a raça Gir e investigar possíveis genes funcionais associados às regiões determinadas como significativas.

3 MATERIAL E MÉTODOS

3.1 Dados de Genótipos e de Sequenciamento

Os animais avaliados nesse estudo são provenientes do Programa de Melhoramento Genético de bovinos de Leite, que atualmente é conduzido pela Empresa Brasileira de Pesquisa Agropecuária (Embrapa) Gado de Leite em Juiz de Fora (MG), em parceria com a Associação Brasileira dos Criadores de Gir Leiteiro (ABCGIL). Foram disponibilizados para esse estudo registros de 2.279 animais da raça Gir, sendo 601 animais (585 machos e 16 fêmeas) genotipados com painel de alta densidade de marcadores (HD) e 1.678 fêmeas com painel de 50 mil marcadores (50k), ambos utilizando a plataforma de genotipagem Illumina (Illumina Inc., San Diego, USA).

Os animais e os dados de genótipo foram submetidos a critérios de controle de qualidade, visando reduzir a quantidade de amostras e marcadores com baixa qualidade. Registros de animais e marcadores com taxa de leitura (“call rate”) inferior a 0,90 foram excluídos, além de marcadores monomórficos e com desvios significativos ($p < 10^{-5}$) do equilíbrio de Hardy-Weinberg. Após a aplicação do controle de qualidade, restaram 583 e 1.654 animais com 594.974 e 33.382 marcadores nos painéis HD e 50k, respectivamente.

Os animais sequenciados foram obtidos do “1000 Bull Genomes Project (Run 5.0)” (DAETWYLER et al., 2014). Foram selecionados 53 animais de 5 raças zebuínas: 7 Brahman (BRM), 6 Gir, 5 Nelore (NEL), 26 Uganda Admixed (UGA) e 9 Irã Admixed (IRN). Dois animais da raça Brahman foram removidos devido a baixa cobertura de leitura das sequências, que resultou na retirada de mais de 80% das variantes pelo controle de qualidade. Portanto, 51 animais compuseram a população de referência de imputação.

Visando a qualidade das sequências, as 44.949.598 variantes polimórficas presentes no conjunto de dados foram submetidas a um rigoroso critério de qualidade. Os critérios seguiram as indicações do “1000 Bull Genomes Project” e permitiram manter apenas variantes com alta qualidade de leitura, mas mantendo a maior quantidade possível de variantes para análise de associação genômica (DAETWYLER et al., 2014; LARMER, 2016). Os critérios de exclusão adotados foram: (I) variantes com qualidade média menor que 20, ou seja, 99% das bases lidas corretamente segundo a estimativa de probabilidade de erro de leitura do sequenciador; (II) cobertura mínima menor que 10 ou maior que a média da cobertura mais três vezes o desvio-padrão, excluindo variantes discrepantes quanto a quantidade média de vezes que a variante foi lida; (III) qualidade de mapeamento menor que 30, critério baseado nas estimativas de erro de posicionamento das bases lidas; (IV) Variantes com distância menor que 10 pb uma da outra, evitando que a mesma variante seja considerada mais de uma vez; (V) Variantes que não estivessem presentes em ao menos 4 animais, tendo em vista que essas variantes tendem a ter baixa acurácia de imputação (LARMER, 2016).

Após o controle de qualidade, 13.171.566 variantes restaram no conjunto de dados. Os filtros de controle de qualidade foram submetidos a todos os animais e aos 29 cromossomos autossômicos por meio dos “softwares” VCFtools v0.1.13 (DANECEK et al., 2011) e pacotes em R snpStats (CLAYTON, 2015) e GWASTools (GOGARTEN et al., 2012).

Para determinar a proporção de variantes compartilhadas entre dados de NGS e o painel HD foi obtida a taxa de detecção de SNP (TDS). A TDS indica quantos marcadores do painel HD estão presentes no conjunto de dados avaliado, ou seja, quantas das posições polimórficas nos genomas sequenciados são marcadores do painel HD (ZHAN et al., 2011).

3.2 Dados fenotípicos

Os 2.279 registros fenotípicos utilizados, assim como os genótipos, fazem parte da avaliação genética de animais da raça Gir, realizada pela Embrapa Gado de Leite em Juiz de Fora (MG). As características avaliadas foram: produção de leite (PL305), teor de gordura (PG305), proteína (PP305) e

sólidos totais (PS305) e foram obtidos com base na lactação encerrada, ou seja, as produções em quilos de leite acumuladas até os 305 dias.

As estimativas dos valores genéticos das características estudadas foram obtidas a partir das observações fenotípicas e do pedigree de 89.849 animais, utilizando o método da máxima verossimilhança restrita, via “best linear unbiased prediction” (BLUP), por meio das equações de modelos mistos (HENDERSON, 1975) em modelo animal:

$$y = \mu + X\beta + Z\alpha + \varepsilon$$

em que y é o vetor de características fenotípicas; μ é a média geral; X a matriz de incidência dos efeitos fixos; β o vetor dos efeitos fixos; Z é a matriz de incidência dos efeitos aleatórios; α é o vetor de efeitos genéticos aditivos e ε é o vetor de erros aleatórios. O modelo considerou os efeitos fixos de grupos de contemporâneos (rebanho e ano no parto), estação de parto e a covariável linear e quadrática da idade da vaca ao parto. A predição dos valores genéticos, assim como dos erros aleatórios, foi obtida pelo programa Wombat (MEYER, 2007). Com base nas estimativas dos valores genéticos, foram obtidos os valores genéticos desregredidos (dEBV), assim como suas estimativas de confiabilidade, como proposto por Garrick et al. (2009).

3.3 Análise de Componentes Principais (PCA)

Os animais sequenciados da população de referência de imputação eram todos de raças zebuínas. No entanto, apenas seis desses eram da raça Gir. Assim, para determinar o grau de relacionamento genético entre os animais, foi efetuada a análise de componentes principais a partir da matriz de relacionamento genômico. Foram considerados na análise os 51 animais sequenciados e os 2.237 animais genotipados, mantendo apenas os marcadores em comum com o painel de 50k. Os marcadores e animais selecionados foram submetidos ao mesmo controle de qualidade dos dados genotípicos, sendo analisados 15.010 marcadores e 2.282 animais na PCA, que foi realizada por meio do pacote em R SNPRelate (ZHENG et al., 2012).

3.4 Imputação

Para se obter os dados de sequenciamento dos 2.237 animais genotipados que passaram pelo controle de qualidade, foram realizadas duas

análises de imputação. Primeiramente foram imputadas as 1.654 fêmeas em 50k para HD, utilizando os 583 animais em HD na população de referência. Posteriormente, as 1.654 fêmeas imputadas para HD, juntamente com os animais genotipados em HD, foram imputados para dados de sequenciamento. Para isso foram utilizados os 51 animais do “1000 Bull Genomes Project (Run 5.0)” na população de referência.

Todas as análises de imputação foram realizadas pelo programa FImpute v2.2 (SARGOLZAEI et al., 2014), com as configurações padrão e ausência de pedigree. A inclusão de registros do pedigree poderia aumentar a acurácia de alelos em baixa frequência (BRØNDUM et al., 2014), mas não tem apresentado ganhos quando incluído em imputações em que o painel de baixa densidade possui mais de 50 mil marcadores (SARGOLZAEI et al., 2014).

3.5 Estudo de associação genômica ampla

Os GWAS foram realizados entre as variantes imputadas para NGS e as características de produção e qualidade de leite, por meio do programa SNP1101 (SARGOLZAEI, 2014). A matriz de relacionamento genômico foi calculada segundo VanRaden (2008), sendo consideradas apenas as variantes com MAF superior a 0,01. Com o objetivo de associar as variantes às variações fenotípicas, a análise de regressão linear múltipla foi realizada regredindo as variantes disponíveis com os dEBV com valor de acurácia superior a 0,10. Esses critérios visaram garantir maior confiabilidade as variáveis utilizadas no modelo de regressão, garantindo maior acurácia nas associações entre variantes e variações fenotípicas.

Para serem consideradas significativas, as variantes deveriam apresentar taxa de “genome-wise False Discovery Rate” (FDR) de, no mínimo, 0,05. Mas apenas aquelas que formaram um pico, com ao menos 100 variantes dentro de 1 Mb, foram consideradas na análise de enriquecimento, que buscou associar tais regiões do genoma à vias biológicas conhecidas (LARMER, 2016).

3.6 Análise de enriquecimento

A definição das vias biológicas funcionais dos genes foi realizada por meio do conjunto de dados do “Medical Subject Headings” (MeSH) (MOROTA et al., 2015), para a busca de aspectos biológicos coerentes com as características estudadas dentro dos diferentes termos descritos no MeSH. Visando possíveis genes associados às variantes significativas, foram consideradas na análise regiões 500 kb acima (“upstream”) e abaixo (“downstream”) de cada variante.

O pacote biomaRt (DURINCK et al., 2005), juntamente com o genoma de referência bovino UMD 3.1 (ZIMIN et al., 2009), foi utilizado no mapeamento das variantes com base nas informações da plataforma Ensembl. A significância dos termos enriquecidos foi assegurada pelo teste hipergeométrico (ADAMS; SKOPEK 1987), que calcula a probabilidade de determinada anotação aparecer mais do que apenas por acaso no conjunto de genes selecionados, quando comparada com os genes no genoma de referência.

4. RESULTADOS E DISCUSSÃO

4.1 Imputação

Ao imputar os genótipos dos animais em HD e 50k para dados de NGS assume-se que parte das variantes não serão imputadas corretamente, o que pode acarretar em associações falso-positivas nas análises de GWAS. A fim de minimizar tais erros inerentes ao processo de imputação, foi realizada a imputação em duas etapas. Em estudo com animais de raças taurinas, a adoção do processo de imputação de genótipos para NGS em duas etapas apresentou ganhos na acurácia de imputação dos cromossomos 1 e 29 de até 65% em relação a imputação direta de animais em 50k para NGS (BINSBERGEN et al. 2014).

Na primeira etapa de imputação, os genótipos em 50k foram imputados para HD, permitindo que as 1.654 fêmeas em 50k tivessem seus genótipos imputados para os 594.974 marcadores do painel em HD. Devido ao reduzido número de vacas genotipadas em alta densidade, não foi possível verificar claramente os valores de acurácia de imputação por meio de validação

cruzada. Assim, para indicar a eficiência com que os dados foram imputados usou-se as distâncias genéticas entre os animais. Isso porque a acurácia de imputação tende a ser maior quanto maior for o grau de parentesco entre os indivíduos a serem imputados e sua população de referência (CARVALHEIRO et al., 2014; BAES, 2015).

O grau de parentesco entre os indivíduos foi obtido pela matriz de parentesco genômico, como apresentado na análise de componentes principais (Figura 1). A PCA evidenciou elevado grau de parentesco entre os animais que foram imputados (50k) com os da população de referência (HD), o que conferiu confiabilidade aos genótipos imputados nessa etapa da imputação.

Na segunda etapa de imputação, os animais em HD juntamente com os imputados na primeira etapa foram novamente imputados para dados de sequenciamento. Nessa etapa, eram esperados valores de acurácia inferiores aos da primeira, devido ao menor grau de parentesco entre os animais genotipados (HD e 50k) e os sequenciados (Figura 1), além do número reduzido de animais na população de referência. Porém, populações de referência compostas por diferentes raças têm favorecido a eficiência com que variantes são imputadas, uma vez que aumentam as chances de que as variantes presentes na população de referência representem a variabilidade genética da população a ser imputada (BAES, 2015). Brøndum et al. (2014), utilizando população de referência multirracial com 242 animais sequenciados, relataram valores de acurácia variando de $0,84 \pm 0,22$ a $0,90 \pm 0,20$.

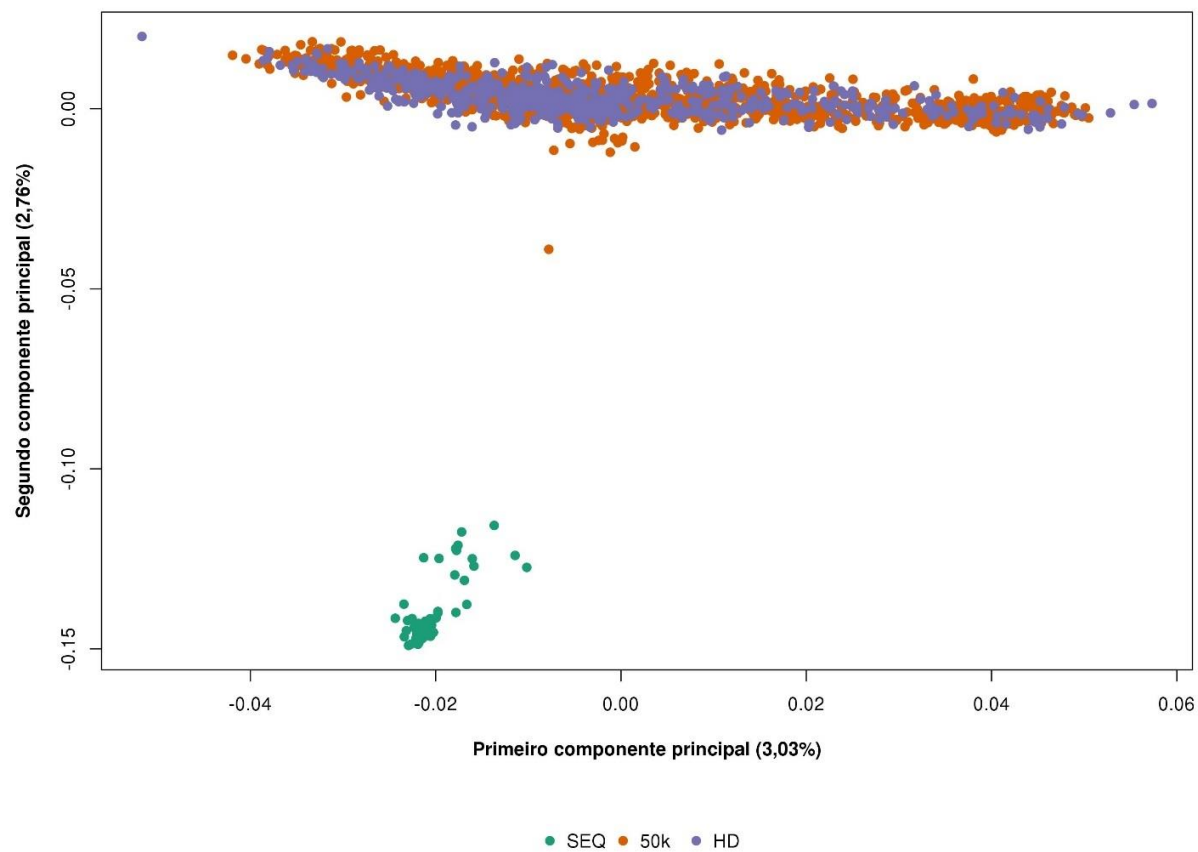


Figura 1. Análise de componentes principais com animais sequenciados e genotipados. **SEQ:** Animais com dados de sequenciamento; **50k:** Animais genotipados com painel 50k; **HD:** Animais genotipados com painel HD.

Estudos testando diferentes cenários na formação da população de referência de imputação revelaram ganhos de até 5% na eficiência de imputação do cromossomo 1, quando animais da mesma raça que a população a ser imputada estão presentes na população de referência (BOUWMAN; VEERKAMP, 2014). Visando atingir os mesmos ganhos, a população de referência na segunda etapa da imputação contou com seis animais da raça Gir, o que auxiliou na formação dos haplótipos durante o processo de imputação, além de todos os animais sequenciados serem de raças zebuínas.

Outro fator que influenciou a acurácia de imputação foi a presença de variantes raras. Foram consideradas raras todas as variantes com MAF entre 1 e 5%. A imputação dessas variantes representa um desafio nos processos de imputação, seja entre painéis de genotipagem com diferentes densidades ou imputações para dados de NGS (BARRETT et al., 2014). Pausch et al. (2017) relataram que variantes comuns podem ter valores médios de acurácia até 19% superiores às médias observadas nas variantes raras. No entanto, ao utilizarem uma população de referência multirracial, os autores obtiveram aumento de 5% na acurácia de imputação dessas variantes menos frequentes. Os 51 animais sequenciados utilizados na segunda etapa de imputação são oriundos de cinco raças distintas. Essa escolha permitiu aumentar o número de variantes presentes na população de referência, representando melhor a diversidade genética encontrada nas raças zebuínas e consequentemente auxiliando a correta imputação das variantes raras.

4.2 Análise de associação genômica (GWAS)

Os avanços nas tecnologias de sequenciamento vêm garantindo maior confiabilidade a cada nucleotídeo lido. Isso, aliado ao desenvolvimento de algoritmos de imputação mais eficientes, viabilizou a utilização dos dados de NGS em análises de associação, principalmente em gado de leite (DAETWYLER et al., 2014; MEUWISSEN; GODDARD, 2010; ZHANG et al., 2016).

Em busca de maior confiabilidade nas análises de associação, o dEBV foi utilizado como pseudo-fenótipo por ser a variável resposta mais acurada para o modelo de regressão múltipla (GARRICK et al., 2009). A quantidade de pseudo-fenótipos disponíveis variou de 2.080 a 2.138 para PG305 e PL305,

respectivamente (Tabela 1). Essa variação era esperada, uma vez que PL305 é a característica de maior valor econômico e a mais utilizada nos critérios de seleção de animais da raça Gir, o que a tornou a característica com maior média de acurácia entre as avaliadas.

Das mais de 13 milhões de variantes disponíveis para as análises de associação genômica, foram removidas em média $2.543.568 \pm 3.754$ com MAF inferior a 1% (Tabela 1). Esse critério visou retirar do conjunto de dados variantes com baixa confiabilidade e reduzir as chances de associações espúrias. Isso porque o valor da acurácia de imputação em variantes com MAF inferior a 1% tende a ser extremamente baixo (BOUWMAN et al. 2014).

Tabela 1. Número de animais avaliados e número de variantes em características de produção e qualidade do leite na população estudada.

Característica	N	Acurácia média dos dEBV (DP)	Variantes	
			Disponíveis	Excluídas
PL305	2.138	0,25 (0,21)	10.627.280	2.537.147
PG305	2.080	0,21 (0,18)	10.623.675	2.540.755
PP305	2.090	0,19 (0,17)	10.628.241	2.536.273
PS305	2.099	0,22 (0,19)	10.632.795	2.531.603

N: Número de animais; **DP:** Desvio-padrão; **dEBV:** valores genéticos desregredidos; **PL305:** produção de leite aos 305 dias; **PG305:** produção de gordura aos 305 dias; **PP305:** produção de proteína aos 305 dias; **PS305:** produção de sólidos totais aos 305 dias.

Após o controle de qualidade, as análises de associação foram conduzidas com mais de 10 milhões de variantes. Esse valor foi 15 vezes superior ao número de SNP normalmente disponíveis no painel HD em animais zebuínos.

Para todas as características, os valores médios de TDS foram próximos a 38% para as variantes comuns. Quando observado esse índice entre as variantes raras ($MAF < 0,05$), essa proporção se torna ainda menor, sendo próxima a 7% (Tabela 2). Portanto, análises de GWAS com dados de NGS permitem que muitas regiões ausentes nas análises genômicas com painéis de genotipagem sejam associadas a variação fenotípica, principalmente regiões com SNP em baixa frequência na população. Mesmo que a porcentagem da variação fenotípica explicada por variantes em baixa frequência seja menor que

a de variantes comuns, é possível que a somatória dos seus efeitos forneça maior compreensão da estrutura genética das características avaliadas.

Considerando todas as características estudadas (Tabela 3), foram observadas 950.686 variantes significativas ($FDR < 0,05$), mas apenas 627 dessas formaram picos significativos, com ao menos 100 variantes dentro de 1 Mb (Figuras 2 e 3). Essas variantes tiveram em média MAF de 0,03 e foram associadas a 25 genes utilizados nas análises de enriquecimento (Tabela 3). Para PS305, 500 variantes deram origem a picos, observados nos cromossomos 11, 12 e 28 (Figura 3B). O cromossomo 11, com 239 variantes, foi o que apresentou a maior proporção de variantes significativas que se localizaram entre as regiões de 58.063.850 a 59.996.090 pb. Nessa mesma região foram encontrados 221 SNP no mapa do painel HD, sendo 24 desses SNP compartilhados com os dados de sequenciamento. O valor médio de MAF das 239 variantes foi de $0,03 \pm 0,007$. No cromossomo 28, 151 variantes foram encontradas entre as posições 29.211.676 e 29.999.044 pb. Essa região contém 199 SNP no painel HD, 7 deles compartilhados com dados de NGS. A MAF dessas variantes variou de 0,01 a 0,06, sendo a maior amplitude observada entre as variantes dentro de um pico significativo.

No cromossomo 12, 110 variantes foram consideradas como significativas, localizadas entre 75.255.615 e 75.788.411 pb. A MAF média foi de $0,04 \pm 0,02$, sendo as variantes com maiores frequências dentre as 500 significativas para PS305. Nessa região não foram observados SNP compartilhados entre os dados de sequência e o painel HD, que continha apenas 55 SNP nessa região. A baixa quantidade de SNP nessa região pode explicar a ausência de compartilhamento. Devido a baixa cobertura do painel HD nessa região, o que reduz os valores de DL, essas variantes significativas são um indicativo de áreas do genoma de zebuínos que afetam a variabilidade genética e não são representadas pelas variantes comuns do painel comercial de genotipagem. O valor de MAF das variantes nos picos significativos é outro fator de interesse que, por serem inferiores a 5%, são comumente retiradas durante o controle de qualidade de estudos genômicos, ignorando seus efeitos na variabilidade genética aditiva de características de interesse econômico.

Tabela 2. Tamanho dos cromossomos com os respectivos números de variantes comuns e raras e taxa de detecção de SNP (TDS) da característica produção de leite aos 305 dias (PL305).

Cromossomo	Tamanho (Mb)	Variantes comuns	TDS¹ (%)	Variantes raras	TDS² (%)
1	158,33	558.740	44,70	185.396	9,52
2	137,06	442.009	42,17	150.479	7,93
3	121,43	364.724	41,39	122.439	7,85
4	120,83	419.279	41,40	139.480	7,84
5	121,17	373.753	40,21	128.144	8,64
6	119,45	417.355	47,89	139.804	7,64
7	112,64	360.120	42,72	122.136	7,52
8	113,38	379.629	43,73	118.483	7,32
9	105,71	359.979	46,61	119.961	7,76
10	104,30	320.469	36,28	114.196	8,74
11	107,31	293.750	35,61	107.638	7,81
12	91,16	340.473	44,65	115.582	8,48
13	84,24	232.115	32,78	75.047	6,83
14	84,65	280.531	44,13	87.752	6,51
15	85,29	290.593	37,11	109.017	7,01
16	81,70	257.665	38,94	88.314	6,72
17	75,15	225.946	38,09	81.102	7,86
18	66,00	161.442	32,10	60.066	4,99
19	64,05	139.838	27,02	48.467	5,26
20	72,03	231.961	40,33	83.158	8,38
21	71,60	209.511	35,44	73.385	7,48
22	61,43	155.990	33,18	60.080	7,20
23	52,53	160.769	36,04	62.693	5,78
24	62,71	208.551	39,33	70.160	7,66
25	42,90	88.256	27,90	32.698	5,45
26	51,66	153.957	35,88	56.806	6,90
27	45,41	154.510	40,54	56.437	6,80
28	46,31	154.446	34,53	60.584	6,70
29	51,50	161.786	34,08	59.629	6,26
Total	2.511,93	7.898.147	-	2.729.133	-

Variantes raras: Variantes com “minor allele frequency” entre 1% e 5%; **TDS^{1,2}:** taxa de detecção de SNP (porcentagem de variantes compartilhadas entre dados de NGS e painel HD) das variantes comuns e raras, respectivamente.

Em PG305, das 4.866 variantes com $FDR < 0,05$, apenas 127 deram origem a um pico significativo (Figura 2B). A região em que as 127 variantes foram encontradas está situada entre as posições 40.319.629 e 40.872.143 pb do cromossomo 6. Nessa mesma região foram encontrados 146 SNP no painel

HD, mas nenhum deles na mesma posição que as variantes. A MAF média das variantes nessa região foi de $0,011 \pm 0,005$, evidenciando que tais variantes apresentaram incidências muito baixas no genoma dessa população. No entanto, a partir da análise de enriquecimento, encontrou-se genes de efeitos condizentes com a característica.

Tabela 3. Número de variantes significativas, valor médio da “minor allele frequency” e número de genes identificados nas características de produção e qualidade de leite.

Características	Variantes significativas		Média MAF (DP)	Número de Genes identificados
	FDR<0,05	FDR<0,01		
PL305	718	120	0,025±0,042	-
PG305	2.923	1.943	0,029±0,036	1
PP305	15.160	9.197	0,027±0,029	-
PS305	32.922	50.541	0,042±0,037	24

MAF: “minor allele frequency”; **PL305:** produção de leite aos 305 dias; **PG305:** produção de gordura aos 305 dias; **PP305:** produção de proteína aos 305 dias; **PS305:** produção de sólidos totais aos 305 dias; **FDR:** *False Discovery Rate*.

Com o elevado número de variantes consideradas no estudo, era esperado que mais picos claros e significativos fossem encontrados, tendo em vista que as características avaliadas normalmente apresentam resposta a seleção. Diferentes fatores podem influenciar os resultados da GWAS, entre os quais estão o número de animais com informações genotípicas e a baixa acurácia da informação utilizada como variável resposta no modelo de predição genômica (PASANIUC et al., 2012). Ambos os fatores foram considerados no presente estudo. Além disso, ao se utilizar dados de NGS com baixa cobertura há um maior índice de erro de leitura, quando se compara aos índices dos painéis de genotipagem. Esses erros são introduzidos ao resíduo dos modelos de predição, diminuindo as acurácias das estimativas dos efeitos dos marcadores. Tais erros podem ser transmitidos no processo de imputação, principalmente se há baixa relação de parentesco entre as populações de referência e validação dos dados. Assim, o critério utilizado para considerar os picos significativos visou dar maior confiabilidade a análise de associação o que resultou em apenas 3 e 1 picos nas características PS305 e PG305, respetivamente.

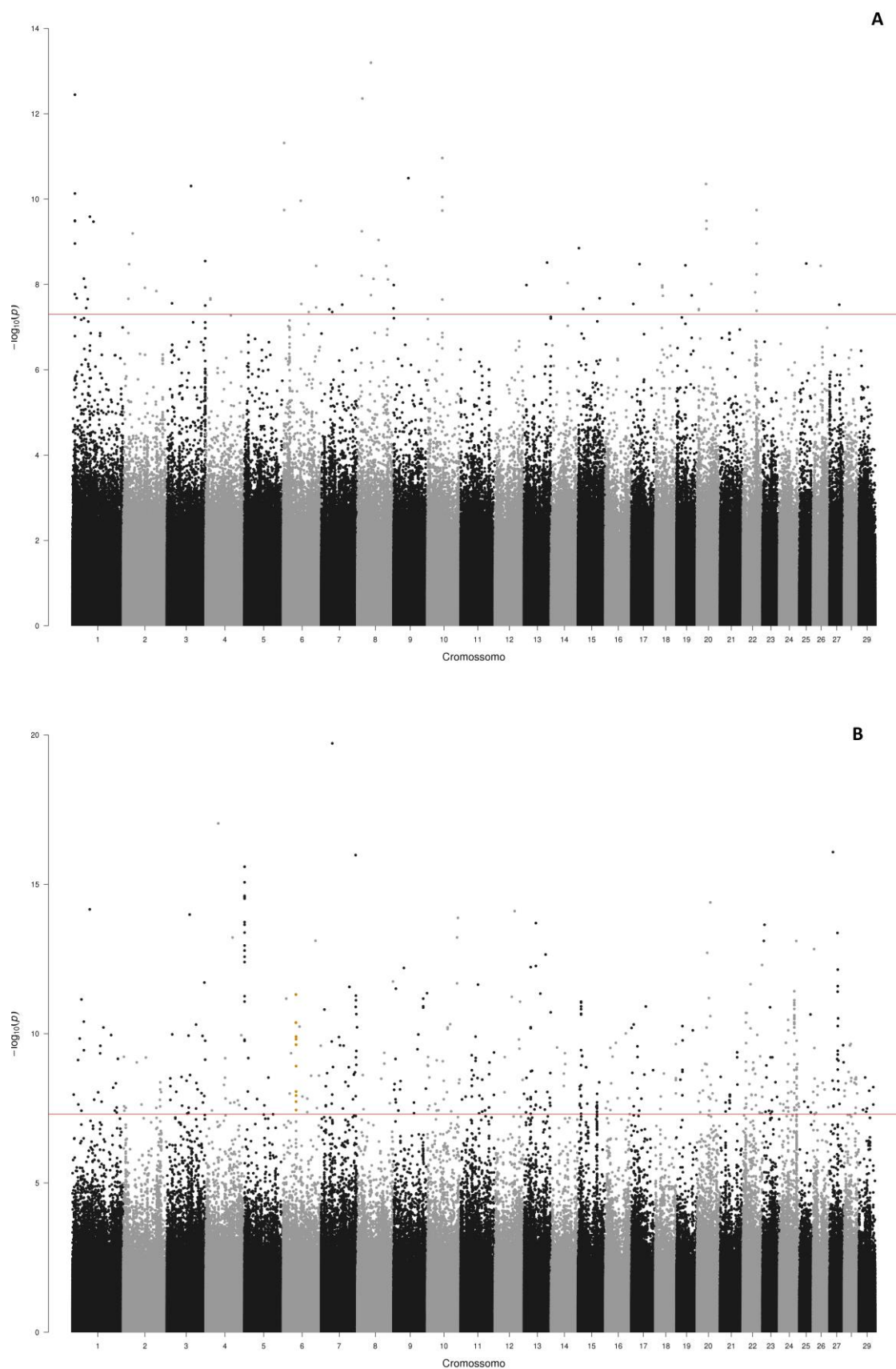


Figura 2. Manhattan Plot das características de produção e qualidade. **PL305 (A):** produção de leite aos 305 dias; **PG305 (B):** produção de gordura aos 305 dias.

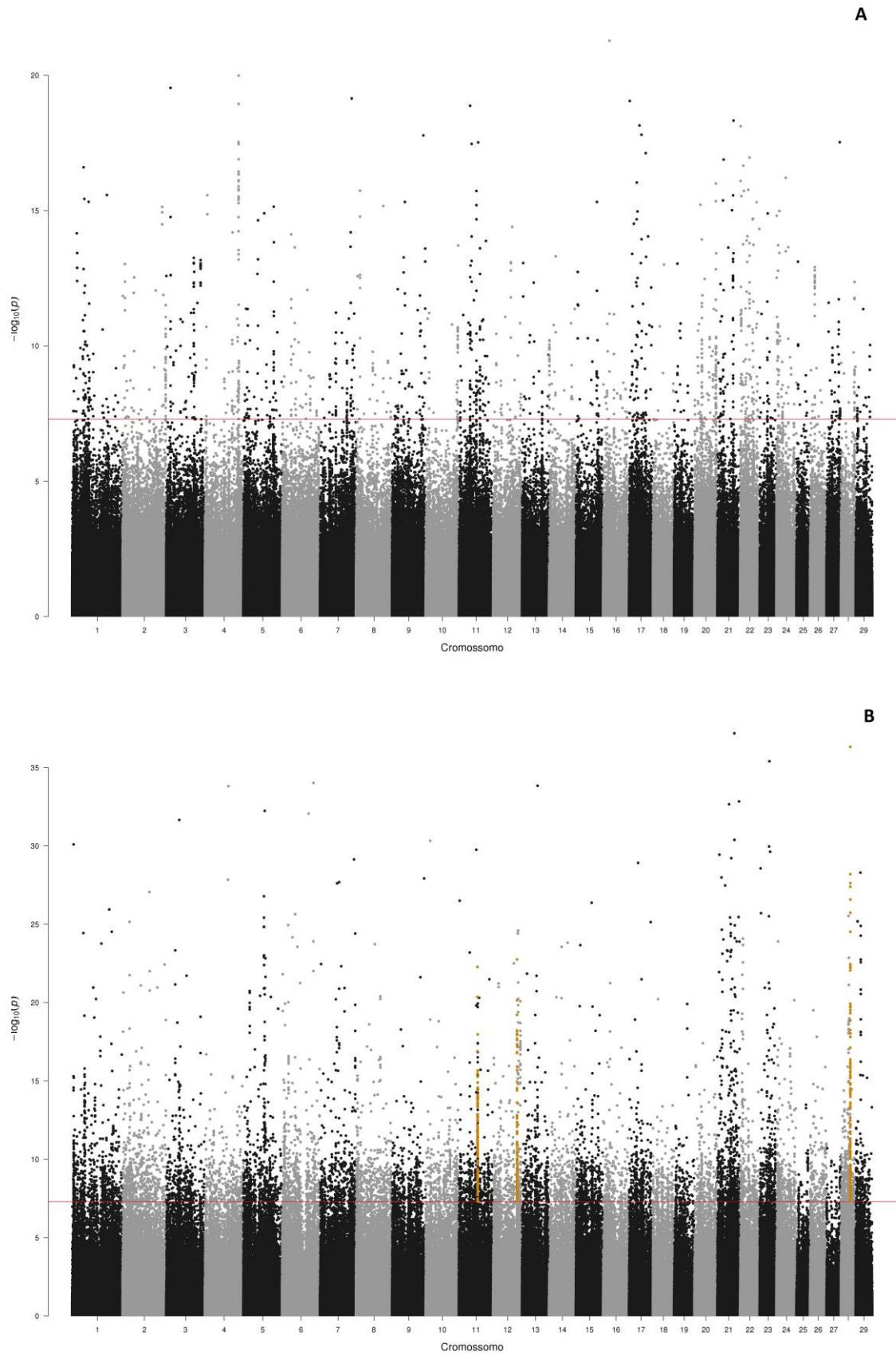


Figura 3. Manhattan Plot das características de qualidade. **PP305 (A):** produção de proteína aos 305 dias; **PS305 (B):** produção de sólidos totais aos 305 dias.

As demais características avaliadas tiveram variantes significativas ($FDR < 0,05$), mas sem a formação de nenhum pico significativo. Deve-se ressaltar que para PL305, característica com maior valor médio de acurácia dos dEBV entre as características avaliadas, obteve o menor número de variantes significativas, além dos maiores p-valores das análises de GWAS. Isso pode indicar que o seu efeito é controlado por várias regiões ao longo do genoma, não tendo regiões que expliquem grande parte da variação fenotípica.

Para PP305 24.357 variantes foram significativas ($FDR < 0,05$), sendo 9.197 variantes com taxa de FDR inferior a 0,01. Mesmo tendo a segunda maior proporção de variantes significativas entre as características avaliadas, nenhuma janela de 1Mb apresentou as 100 variantes para ser considerada como um pico de associação.

Utsunomiya (2015), em estudo de associação com painel comercial de genotipagem e utilizando parte do conjunto de dados desse estudo, não encontrou nenhuma região que causasse grande efeito na variância genética aditiva das características PL305, PG305, PP305 e PS305. O maior efeito observado na variação genética foi de 1,96% para PG305, o que ressalta o forte caráter poligênico nos mecanismos de herança dessas características. Os mesmos autores relataram maior número de picos, quando comparado ao presente estudo, o que era esperado tendo em vista que dados de NGS demandam mais animais com dados genômicos para obter os mesmos resultados de trabalhos com painel de genotipagem (FAYE et al., 2013).

Variantes significativas podem ser incluídas em painéis de genotipagem, elevando o seu poder de predição dos efeitos dos marcadores. No caso de características em que não há QTL de grande efeito, essa inclusão pode representar melhora significativa nas predições dos valores genéticos genômicos (GODDARD et al., 2015). No entanto, Larmer (2016), com 4.000 animais da raça Holstein, ao comparar o número de regiões significativas entre dados de NGS e painel HD relatou grande sobreposição de regiões significativas e quase nenhum ganho em acurácia de predição ao incluir as novas variantes no painel de genotipagem.

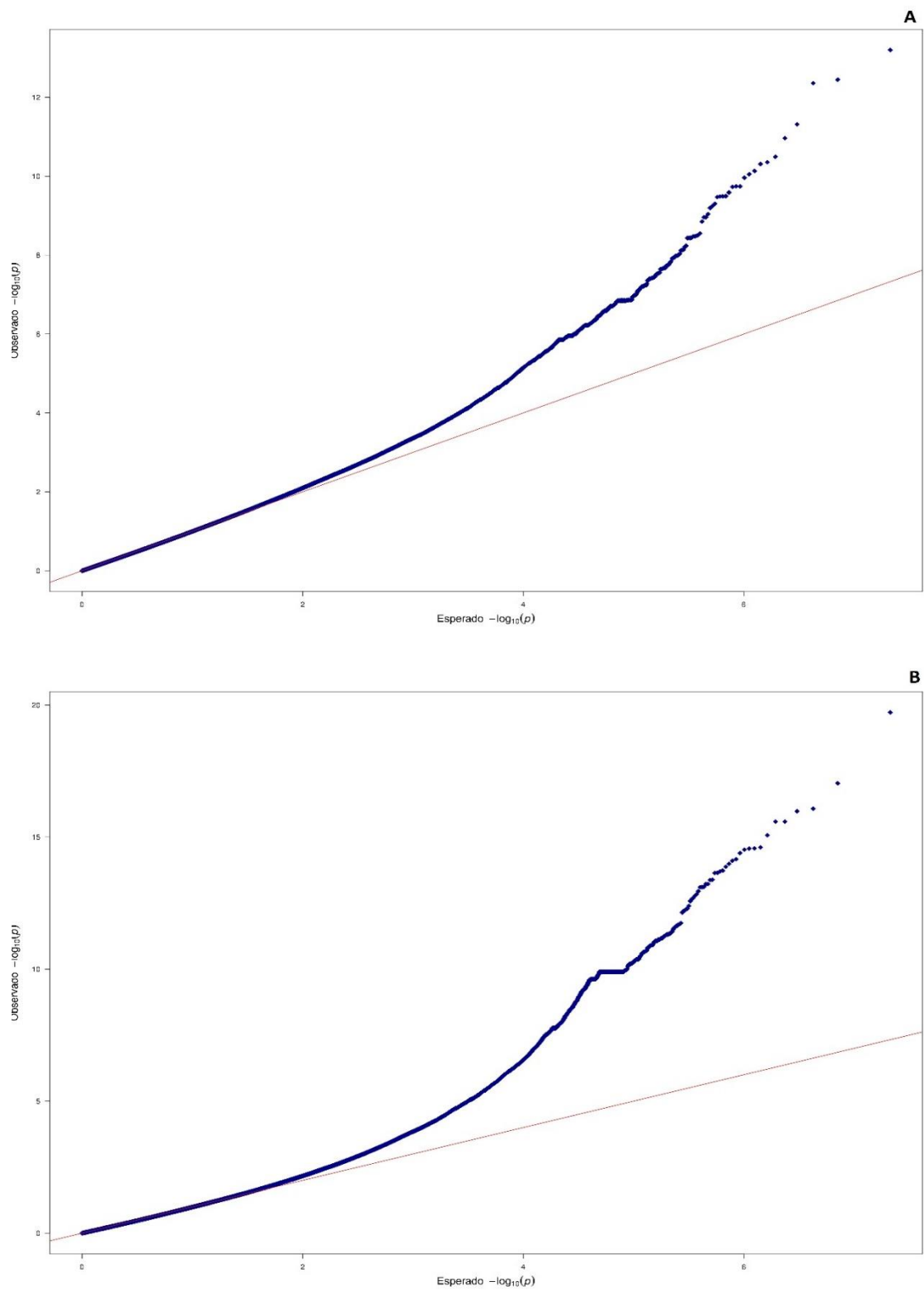


Figura 4. Q-Q Plot dos p-valores ($-\log_{10}$) estimados pela análise de regressão para as características de produção e qualidade de leite. **PL305 (A):** produção de leite aos 305 dias; **PG305 (B):** produção de gordura aos 305 dias.

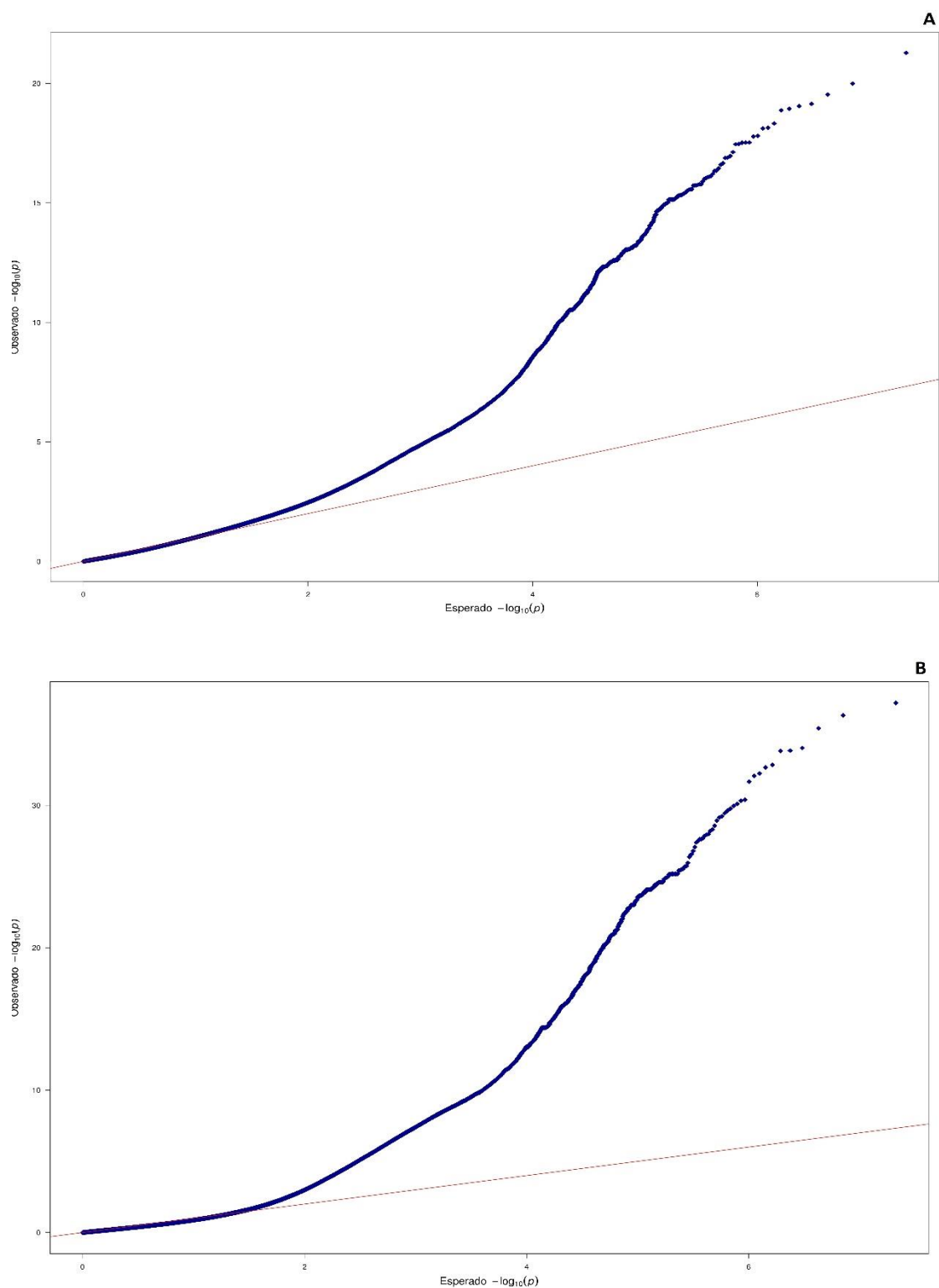


Figura 5. Q-Q Plot dos p-valores ($-\log_{10}$) estimados pela análise de regressão para as características de qualidade de leite. **PP305 (A):** produção de proteína aos 305 dias; **PS305 (B):** produção de sólidos totais aos 305 dias.

Para PL305 e PP305 não foram observados picos significativos na GWAS. Todavia, o desvio da distribuição dos efeitos dos marcadores sob a hipótese nula, indicou que as variantes influenciaram na variância genética aditiva das características (Figuras 4 e 5). No entanto, parte desse desvio pode ter sido causado apenas por um viés nas estimativas. As mesmas características demonstraram ser fortemente controladas por várias regiões do genoma, como observado pela distribuição dos efeitos das variantes ao longo do genoma nos gráficos Manhattan Plot (Figuras 2 e 3). Para capturar esse caráter poligênico, serão necessários mais animais genotipados e sequenciados, o que elevaria a acurácia de imputação e o poder de predição das análises de associação. Além disso, são necessários registros fenotípicos mais acurados, o que pode favorecer o aumento da confiabilidade dos valores genéticos e consequentemente reduzir as chances de associações falso-positivas.

4.3 Análise de enriquecimento

Todas as características fenotípicas avaliadas tiveram variantes significativas ($FDR < 0,05$), mas apenas quatro regiões foram consideradas na análise de enriquecimento. Isso porque o elevado número de variantes consideradas na análise de GWAS, frente ao número de animais utilizados na análise, favorece o surgimento de associações espúrias.

O pico significativo de PG305 foi formado por 127 variantes e situado entre 40.319.629 e 40.872.143 pb do cromossomo 6. Nessa região foi encontrado o gene SLIT2, associado a processos inflamatórios (LIM et al., 2014). Esposito et al. (2014) relataram interações metabólicas entre o sistema imune e endócrino em vacas leiteiras durante o período de transição. Os autores observaram que, durante o período de transição, há grande quantidade de metabolismo celular, pois o corpo da vaca se prepara para o parto e para a síntese do leite. Cohen-Zinder et al. (2005) relataram a presença desse gene dentro do intervalo de confiança de uma região de QTL. O mesmo faz parte de uma série de genes que afetam a produção de leite, gordura e proteína, presentes no cromossomo 6 de animais da raça Holandesa Israelense.

Raven et al. (2014), em estudo com mais de 8 mil marcadores e 16 mil animais das raças Holandesa -Frisia e Jersey, relataram que o gene SLIT2 foi

responsável por 0,09% da variação da produção de gordura e 0,46% em porcentagem de proteína, sendo sua atuação no genoma associada ao desenvolvimento das glândulas mamárias. No entanto, quando esse gene foi submetido a análise de enriquecimento pelos termos MeSH, não foram encontradas vias enriquecidas.

Para PS305, nos três picos declarados como significativos foram encontrados 23 genes e 14 vias enriquecidas que podem estar relacionadas à produção de sólidos totais (Tabela 4). A via mais enriquecida foi a do cálcio, mineral essencial e principal sólido presente no leite. Vias como Glicólise, Ácido pirúvico e Transferência de energia foram associadas a 15, 16 e 5 genes, respectivamente. O enriquecimento dessas vias indica a forte associação entre os picos significativos do GWAS com vias metabólicas conhecidas e associadas a características de produção e qualidade do leite. Esses resultados dão credibilidade as regiões consideradas como significativas, além de indicar possíveis genes candidatos a serem incluídos nas análises de estimação dos valores genéticos em animais da raça Gir e outros zebuínos geneticamente próximos.

Tabela 4. Termos relacionados a genes que apresentaram expressão diferencial em produção de sólidos totais aos 305 dias (PS305) na análise MeSH (“Medical Subject Headings”).

Categoria	MeSH ID¹	Termo MeSH
Químicos e drogas	D017310 (2)	Annexin A7
	D020747 (9)	Canais de cálcio, T-Type
	D058255 (3)	Protease Nexins
	D019289 (5)	Ácido pirúvico
	D058258 (3)	Serpin E2
	D002118 (505)	Cálcio
	D003847 (8)	Deoxiglucose
Fenômenos e processos	D060833 (3)	Microambiente celular
	D004735 (16)	Transferência de energia
	D006019 (15)	Glicólise
	D006706 (27)	Homeostase
	D015640 (62)	Canais de íons
Anatomia	D004730 (374)	Endotélio, Vascular
	D005187 (36)	Trompas de Falópio

¹ Identificação da via nos termos MeSH, em parênteses o número de genes associados

5. Conclusões

As características avaliadas apresentaram forte caráter poligênico, havendo grande número de variantes significativas na análise GWAS. Isso indica que a utilização de dados de NGS nas avaliações genéticas dessa população pode representar aumento na captura de regiões que influenciam a variabilidade genética das características avaliadas.

Regiões do genoma determinadas como significativas na GWAS, apresentam genes e vias metabólicas enriquecidas associadas as características de produção e qualidade de leite.

REFERÊNCIAS

- ADAMS, W. T.; SKOPEK, T. R. Statistical test for the comparison of samples from mutational spectra. **Journal of Molecular Biology**, v. 194, p. 391-6, 1987.
- BAES, C. F. Across-breed imputation with whole genome sequence data in dairy cattle. In: PLANT AND ANIMAL GENOME XXIII CONFERENCE, 2015, San Diego - USA.
- BARRETT, J. H.; TAYLOR, J. C.; ILES, M. M. Statistical Perspectives for GenomeWide Association Studies (GWAS). **Clinical Bioinformatics**, v. 1168, p. 47-61, 2014.
- BERG, I. D.; BOICHARD, D.; LUND, M. S. Comparing power and precision of within-breed and multibreed genome-wide association studies of production traits using whole-genome sequence data for 5 French and Danish dairy cattle breeds. **Journal of Dairy Science**, v. 99, n. 11, p. 8932-8945, 2016.
- BERRY, D. P.; MCCLURE, M. C.; MULLEN M. P. Within- and across-breed imputation of high-density genotypes in dairy and beef cattle from medium- and low-density genotypes. **Journal of Animal Breeding and Genetics**, v. 131, n. 3, p. 165-172, 2014.
- BINSBERGEN, R. V.; 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**, v. 46, n. 1, p. 41, 2014.
- BOUWMAN, A. C.; VEERKAMP, R. F. Consequences of splitting whole-genome sequencing effort over multiple breeds on imputation accuracy. **BMC Genetics**, v. 15, n. 105, 2014.

BRØNDUM, R. F.; GULDBRANDTSEN, B.; SAHANA, G.; LUND, M. S.; SU, G. Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. **BMC Genomics**, v. 15, n. 1, p. 728, 2014.

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.; SILVA, M. V. G. B. da; NICIURA, S. C. M.; TORRES JÚNIOR, R. A. A.; ALENCAR, M. M.; REGITANO, L. C. A.; MUNARI, D. P. Genome-Wide Association for Growth Traits in Canchim Beef Cattle. **Plos One**, v. 9, n. 4, p. e94802, 2014.

CARVALHEIRO, R.; BOISON, S. A.; 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**, v. 46, n. 1, p. 69, 2014.

CHEN, P. C.; YIN, J.; YU, H. W.; YUAN, T.; FERNANDEZ, M.; YUNG, C. K.; TWOOG-DUBE, E. Next-generation sequencing identifies rare variants associated with Noonan syndrome. **Proceedings of the National Academy of Sciences**, v. 111, n. 31, p. 11473-11478, 2014.

CLAYTON, D. snpStats: SnpMatrix and XSnpmatrix classes and methods. R package version 1.24.0, 2015.

COHEN-ZINDER, M.; SEROUSSI, E.; LARKIN, D. M.; LOOR, J. J.; WIND, A. E.; LEE, J. H.; DRACKLEY, J. K.; BAND, M. R.; HERNANDEZ, A. G.; SHANI, M.; LEWIN, H. A.; WELLER, J. L.; RON, M. Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. **Genome Research**, v. 15, n. 7, p. 936-944, 2005.

DAETWYLER, H. D.; CAPITAN, A.; PAUSCH, H.; STOTHARD, P.; BINSBERGEN, R. V.; BRØNDUM, R. F.; et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. **Nature Genetics**, v. 46, n. 8, p. 858-65, 2014.

DANECEK, P.; AUTON, A.; ABECASIS, G.; ALBERS, C. A.; BANKS, E.; DEPRISTO, M. A. The variant call format and VCFtools. **Bioinformatics**, v.27, n.15, p. 2156-2158, 2011.

DRUET, T.; MACLEOD, I.; M.; HAYES, B. J. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. **Heredity**, v. 112, p. 39-47, 2014.

DURINCK, S.; MOREAU, Y.; KASPRZYK, A.; DAVIS, S.; DE MOOR B.; BRAZMA A.; HUBER W. BIOMART and Bioconductor: a powerful link between biological databases and microarray data analysis. **Bioinformatics**, v. 21, p. 3439-40, 2005.

DURINCK, S.; SPELLMAN, P. T.; BIRNEY, E.; HUBER, W. Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package BIOMART. **Nature Protocols**, v. 4, p. 1184-91, 2009.

ESPOSITO, G.; IRONS, P.C.; WEBB, E.C.; CHAPWANYA, A. Interactions between negative energy balance, metabolic diseases, uterine health and immune response in transition dairy cows. **Animal Reproduction Science**, v. 30, n. 144, p. 60-71, 2014.

FAYE, L. L.; MACHIELA, M. J.; KRAFT, P.; BULL, S. B.; SUN, L. Re-ranking sequencing variants in the post-GWAS era for accurate causal variant identification. **PLoS Genetics**, v. 9, n. 8, p. e1003609, 2013.

GARRICK, D. J.; TAYLOR, J. F.; FERNANDO, R. L. Deregressing estimated breeding values and weighting information for genomic regression analyses. **Genetics Selection Evolution**, v. 41, p. 55, 2009.

GODDARD, M. E.; MACLEOD, I. M.; BOLORMAA, S.; HAYES, B. J.; KEMPER, K. E. Improving the accuracy of across breed genomic predictions. **Association for the Advancement of Animal Breeding and Genetics**, v. 21, p. 149-152, 2015.

GOGARTEN, S. M.; BHANGALE, T.; CONOMOS, M.P.; LAURIE, C. A.; MCHUGH, C. P.; PAINTER, I.; ZHENG, X.; CROSSLIN, D. R.; LEVINE, D.; LUMLEY, T.; NELSON, S. C.; RICE, K.; SHEN, J.; SWARNKAR, R.; WEIR, B. S.; LAURIE, C. C. GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. **Bioinformatics**, v. 28, n. 24, p. 3329-3331, 2012.

HENDERSON, C. R. Best linear unbiased estimation and prediction under a selection model. **Biometrics**, v. 31, p. 423-447, 1975.

KIM, Y. J.; LEE, J.; KIM, B.; T2D-Genes Consortium; PARK, T. A new strategy for enhancing imputation quality of rare variants from next-generation sequencing data via combining SNP and exome chip data. **BMC Genomics**, v. 16, n. 1109, 2015.

LARMER, S. G **Next generation sequencing data in bovine: quality control, imputation, and application**. 2016. 142 f. Tese (Doctor of Philosophy) – University of Guelph, Canadá, 2016.

LIM, R.; LIONG, S.; BARKER, G.; LAPPAS, M. Slit2 is decreased after spontaneous labour in myometrium and regulates pro-labour mediators. **Journal of Reproductive Immunology**, v. 106, p. 76-88, 2014.

LU, L.; BOERWINKLE, E.; XIONG, M. Association studies for next-generation sequencing. **Genome Research**, v. 21, p. 1099-1108, 2011.

MAI, M. D.; SAHANA, G.; CHRISTIANSEN, F. B.; GULDBRANDTSEN, B. A genome-wide association study for milk production traits in Danish Jersey cattle using a 50K single nucleotide polymorphism chip. **Journal of Animal Science**, v. 88, n. 11, p. 3522-3528, 2010.

MEUWISSEN, T.; GODDARD, M. Accurate Prediction of Genetic Values for Complex Traits by Whole-Genome Resequencing. **Genetics**, v. 185, n. 2, p. 623-631, 2010.

MEYER, K. WOMBAT - A tool for mixed model analyses in quantitative genetics by REML. **Journal of Zhejiang University Science B**, v. 8, n. 11, p. 815-821, 2007.

MOROTA, G.; PEÑAGARICANO, F.; PETERSEN, J. L.; CIOBANU, D. C.; TSUYUZAKI, K.; NIKAI, I. An application of MeSH enrichment analysis in livestock. **Animal Genetics**, v. 46, n. 4, p. 381–387, 2015.

PASANIUC, B.; ROHLAND, N.; MCLAREN, P. J.; GARIMELLA, K.; ZAITLEN, N.; LI, H.; PRICE, A. L. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. **Nature genetics**, v. 44, n. 6, p. 631-635, 2012.

PAUSCH, H.; MACLEOD, I. M.; FRIES, R.; EMMERLING, R.; BOWMAN, P. J.; DAETWYLER, H. D.; GODDARD, M. E. Evaluation of the Accuracy of Imputed Sequence Variant Genotypes and Their Utility for Causal Variant Detection in Cattle. **Genetics Selection Evolution**, v. 49 p. 24, 2017.

PRICE, A. L.; ZAITLEN, N. A.; REICH, D.; PATTERSON, N. New approaches to population stratification in genome-wide association studies. **Nature Reviews Genetics**, v. 11, n. 7, p. 459-463, 2010.

RAVEN, L. A.; COCKS, B. G.; GODDARD, M. E.; PRYCE, J. E.; HAYES, B. J. Genetic variants in mammary development, prolactin signalling and involution pathways explain considerable variation in bovine milk production and milk composition. **Genetics Selection Evolution**, v. 46, p. 29, 2014

REIS FILHO, J. C. R.; LOPES, P. S.; VERNEQUE, R. D.; TORRES, R. D.; TEODORO, R. L.; CARNEIRO, P. L. S.; Population structure of Brazilian Gyr dairy cattle. **Revista Brasileira de Zootecnia**, v. 39, n. 12, p. 2640-2645, 2010.

RUTKOSKI, J. E.; POLAND, J.; JANNINK, J. L.; SORRELLS, M. E.; Imputation of Unordered Markers and the Impact on Genomic Selection Accuracy. **Genes Genomes Genetics**, v. 3, n. 3, p. 427–439, 2013.

SAHANA, G.; GULDBRANDTSEN, B.; BENDIXEN, C.; LUND, M. S. Genome-wide association mapping for female fertility traits in Danish and Swedish Holstein cattle. **Animal Genetics**, v. 41, n. 6, p. 579-588, 2010.

SARGOLZAEI, M. SNP1101 User's Guide. Version 1.0.0. Mehdi Sargolzaei, Semex Alliance, Ontario, Canadá, 2014.

SARGOLZAEI, M.; CHESNAIS, J.; SCHENKEL, F. A new approach for efficient genotype imputation using information from relatives. **BMC Genomics**, v. 15, n. 1, p. 478, 2014.

UTSUNOMIYA, A. T. H. **Análise ampla do genoma para detecção de erros de montagem no genoma de referência bovino e para detecção de locos relacionados a características de produção e reprodução da raça gir.** 2015. 64 f. Tese (Doutor em Genética e Melhoramento Animal) – Faculdade de Ciências Agrárias e Veterinárias, Brasil, 2015.

VANRADEN, P. Efficient methods to compute genomic predictions. **Journal of dairy science**, v. 91, n. 11, p. 4414-4423, 2008.

ZHAN, B.; FADISTA, J.; THOMSEN, B.; HEDEGAARD, J.; PANITZ, F.; BENDIXEN, C. Global assessment of genomic variation in cattle by genome resequencing and high-throughput genotyping. **BMC Genomics**, v.12, p.557, 2011.

ZHANG, Q.; GULDBRANDTSEN, B.; THOMSEN, J. R.; LUND, M. S.; SAHANA, G. Genome-wide association study for longevity with whole-genome sequencing in 3 cattle breeds. **Journal of Dairy Science**, v. 99, n. 9, p. 7289-7298, 2016.

ZHANG, Z.; OBER, U.; ERBE, M.; ZHANG, H.; GAO, N.; HE, J.; LI, J.; SIMIANER, H. Improving the Accuracy of Whole Genome Prediction for Complex Traits Using the Results of Genome Wide Association Studies. **Plos One**, v. 9, n. 3, p. e93017, 2014.

ZHENG, X.; LEVINE, D.; SHEN, J.; GOGARTEN, S.; LAURIE, C.; WEIR, B. A High-performance Computing Toolset for Relatedness and Principal Component Analysis of SNP Data. **Bioinformatics**, v. 28, n. 24, p. 3326-3328, 2012.

ZIMIN A. V.; DELCHER, A. L.; FLOREA, L.; KELLEY, D. R.; SCHATZ M. C.; PUIU, D.; HANRAHAN, F.; PERTEA, G.; VAN TASSELL, C. P.; SONSTEGARD, T. S. A whole-genome assembly of the domestic cow, *Bos taurus*. **Genome Biology**, v. 10, p. R42, 2009.