UNIVERSIDADE ESTADUAL PAULISTA – UNESP
CÂMPUS DE JABOTICABAL

IDENTIFICAÇÃO DE VARIAÇÕES ESTRUTURAIS DO GENOMA DE BOVINOS GIR LEITEIRO

Larissa Graciano Braga
Médica veterinária

2021

UNIVERSIDADE ESTADUAL PAULISTA – UNESP
CÂMPUS DE JABOTICABAL

IDENTIFICAÇÃO DE VARIAÇÕES ESTRUTURAIS DO GENOMA DE BOVINOS GIR LEITEIRO

Discente: Larissa Graciano Braga
Orientador: Prof. Dr. Danísio Prado Munari
Coorientadores: Dra. Tatiane Cristina Seleguim Chud
Dr. Marcos Vinicius Gualberto Barbosa da Silva

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

2021

B813i

Braga, Larissa Graciano
Identificação de variações estruturais do genoma de bovinos Gir Leiteiro / Larissa Graciano Braga. -- Jaboticabal, 2021
87 p.

Dissertação (mestrado) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal
Orientador: Danísio Prado Munari

1. Bovino de leite. 2. Genes. 3. Genética animal. 4. Variações do número de cópias de DNA. 5. Zebu. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Jaboticabal

CERTIFICADO DE APROVAÇÃO

TÍTULO DA DISSERTAÇÃO: IDENTIFICAÇÃO DE VARIAÇÕES ESTRUTURAIS DO GENOMA DE BOVINOS GIR LEITEIRO

AUTORA: LARISSA GRACIANO BRAGA
ORIENTADOR: DANISIO PRADO MUNARI
COORIENTADORA: TATIANE CRISTINA SELEGUIM CHUD
COORIENTADOR: MARCOS VINÍCIUS GUALBERTO BARBOSA DA SILVA

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

Prof. Dr. DANISIO PRADO MUNARI (Participaçao Virtual)
Departamento de Engenharia e Ciências Exatas (DECEx) / FCAV / Unesp - Jaboticabal

Profa. Dra. ADRIANA SANTANA DO CARMO (Participaçao Virtual)
Escola de Veterinária e Zootecnia / Universidade Federal de Goiás/UFG

Pesquisador Dr. ROBERTO CARVALHEIRO (Participaçao Virtual)
Departamento de Zootecnia / FCAV / UNESP - Jaboticabal

Jaboticabal, 23 de julho de 2021

DADOS CURRICULARES DA AUTORA

Larissa Graciano Braga nasceu em Porto Velho - RO, no dia 29 de maio de 1996, filha de Bárbara Braga Graciano e Sebastião Graciano de Souza. Iniciou sua graduação em Medicina Veterinária em março de 2014, na Escola de Veterinária e Zootecnia da Universidade Federal de Goiás (EVZ/UFG), em Goiânia e obteve o Título de Bacharel em Medicina Veterinária em julho de 2019. No período de 2016 a 2017 foi voluntária de iniciação científica (PIVIC) no Programa de Iniciação à Pesquisa PIP/UFG, sob a orientação do Prof. Dr. Paulo Henrique Jorge da Cunha. No ano de 2016 foi Diretora de Projetos na CONPAVet Jr, empresa Júnior situada na EVZ/UFG. Em 2017, obteve bolsa para programa de mobilidade acadêmica internacional (MARCA/MERCOSUR) e realizou intercâmbio de graduação *sanduíche* no segundo semestre na Facultad de Ciências Veterinárias, Universidad Nacional del Litoral, na cidade de Esperanza, Província de Santa Fé, Argentina. Em agosto de 2019 ingressou no 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 de Dra. Tatiane Cristina Seleguim Chud e Dr. Marcos Vinicius Gualberto Barbosa da Silva.

AGRADECIMENTOS

Pelas oportunidades e privilégios que me foram concedidos ao longo de toda minha vida acadêmica.

Gostaria de agradecer a Deus, por ter me permitido realizar grandes sonhos, agraciando-me com saúde e por ter me guiado durante toda a caminhada. A minha mãe Maria, que sempre passou na frente dos meus medos, dificuldades e desafios.

Agradeço aos meus pais, Sebastião e Bárbara, por serem minha fortaleza e por me amarem incondicionalmente. Sempre me deram força e acolhimento nos momentos mais difíceis, principalmente nos que eu era mais chata. Fizeram o que podiam para que eu tivesse uma boa educação e me ensinaram o significado de responsabilidade e honestidade. A distância tem sido muito pesada para vocês (e também para mim), mas mesmo assim, vocês me apoiaram e apoiam muito. Tenho muita sorte de tê-los de ser filha de vocês e sou muito grata por isso.

À Faculdade de Ciências Agrárias e Veterinárias da Universidade Estadual Paulista (FCAV/UNESP), por ter sido minha segunda casa durante dois anos e pela excelência em ensino. Aos professores e funcionários por terem dado o melhor de si para minha formação acadêmico-profissional e como pessoa, sendo exemplos de responsabilidade, dedicação e respeito. Agradeço ao Programa de Pós Graduação em Genética e Melhoramento Animal pelo mestrado e à CAPES pelo financiamento da minha bolsa de pesquisa.

A todos os professores que muito contribuíram com a minha formação acadêmica e profissional, meu muito obrigada pelos conhecimentos e ensinamentos passados. Especialmente agradeço:

Ao meu orientador, Prof. Dr. Danísio Prado Munari, sempre muito paciente e solicito comigo, agradeço por todos os ensinamentos, críticas, conversas, conselhos, oportunidades e pela confiança depositada em mim. Esse apoio e orientação tem sido fundamental para a minha formação.

A minha coorientadora, Dra. Tatiane Cristina Seleguim Chud, por toda paciência, dedicação (inclusive com reuniões no seu horário de descanso), direção, críticas, incentivo e ensinamentos pessoais e profissionais. Sua coorientação foi essencial para o meu desenvolvimento acadêmico-profissional e pessoal. Ao meu coorientador, Dr. Marcos Vinicius Gualberto da Silva, por ter cedido parte dos dados,

pelo direcionamento, pela confiança depositada em mim e por ter sempre sido muito solicito durante a realização deste trabalho.

Aos membros da banca do exame geral de qualificação, Prof. Dr. Fernando Baldi e Prof. Dr. Roberto Carvalheiro, e aos membros da banca de defesa da dissertação, Prof. Dr. Adriana Santa do Carmo e Prof. Dr. Roberto Carvalheiro pelas críticas, reflexões e sugestões a este trabalho.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Agradeço à Embrapa, CNPq e FAPEMIG pelo financiamento das amostras utilizadas neste trabalho. Agradeço também a Embrapa Gado de Leite pelos dados cedidos e ao Laboratório Multiusuário de Bioinformática da Embrapa Informática Agropecuária pela infraestrutura computacional e recurso de TI, especialmente a Adhemar Neto e Leandro Cintra.

Durante esses anos tive a oportunidade de trabalhar com alunos e pesquisadores incríveis, que me ensinaram muito. Sou grata por todas as oportunidades que tive e agradeço especialmente à Rafael Nakamura Watanabe pela amizade e paciência e por ter me ajudado com os problemas do servidor, dúvidas, scripts, e-mails, que me escutou nos meus momentos de lamentação e de estresse. Agradeço também a Rodrigo Pelicioni Savegnago, por todo ensinamento e discussão durante a realização deste projeto.

Agradeço aos meus demais colegas membros e ex-membros do grupo de pesquisa Estatística Aplicada a Genética e Melhoramento Animal (EAGMA) que me auxiliaram durante esse processo: Ana Paula Sbardella, Guilherme Batista, Priscila Bernardes, Samla Cunha, Tádia Emanuelle, Thomaz Sena e especialmente a Letícia Joaquim Borges por todas as conversas, companhia e aconselhamento. Sou extremamente grata à paciência e ao acolhimento que vocês dedicaram a mim

Aos companheiros de pós graduação pelo apoio durante toda a trajetória. Com vocês compartilhei muitas dúvidas, momentos de estudo, momentos de alegria e confraternização que fizeram meus dias mais leves e felizes. Em especial agradeço a Bruna Salatta, Caio Teixeira, Eliéder Romanzini, Fernando Ongaratto, Gabriela Bonfá, Gustavo Schettini, Hayala Gomes, Isabella Ferreira, Ivan Filho, Kétuly Ataíde, Leyde Emanuelle, Pablo Domingues, Patrícia Schmidt, Patrícia Perin,

William Gallena. À Adriana e Shirley, funcionárias do departamento de Engenharia e Ciências Exatas por toda atenção e disposição em me auxiliar.

À Lucas Gazolla Rodrigues Soares, que tem sido um companheiro incrível. Por vibrar a cada conquista minha, por acreditar em mim (por muitas vezes mais do que eu mesma), por me incentivar e apoiar em tudo e por ter sido um porto seguro nesses últimos meses. O seu apoio e suporte tem sido um acalento, especialmente nos dias mais tempestuosos. À Alzira e Sidney Rodrigues Soares, por todo cuidado e carinho.

Agradeço a Daniel de Paiva e Flávia Dorjó Pedra, meus mentores, por todo incentivo, conselho, questionamento e reflexão. Vocês estão contribuindo para a construção e visão da minha melhor versão.

Agradeço às minhas grandes amigas, Ana Caroline Pimenta e Larissa Lenz, por todas as escutas de qualidade, conversas, colos reais e virtuais, conselhos e incentivos ao longo desses anos. Poder contar com vocês, mesmo que a distância, é sensacional. Agradeço a Érico Caldeira e Isabella Messias pela amizade e por acreditar muito em mim e me apoiar. Vocês me fazem enxergar e acreditar que sou capaz e suficiente.

Às minhas colegas de casa, Fernanda Salti, Laís Teixeira, Maria Júlia Santiago por terem tornado a minha estadia em Jaboticabal mais leve e divertida. À Rosilene da Silva, amiga e vizinha, por vibrar por minhas vitórias e por todo conselho, cuidado, carinho, abraço e atenção.

Agradeço também a todos que, de alguma forma contribuíram com a minha formação e com a realização deste trabalho, mas que não estão citados.

“O que eu faço é uma gota no meio do oceano.
Mas sem ela, o oceano será menor”

Madre Teresa de Calcutá

Sumário

IDENTIFICAÇÃO DE VARIAÇÕES ESTRUTURAIS DO GENOMA DE BOVINOS GIR LEITEIRO

RESUMO – O Gir Leiteiro resulta da seleção fenotípica de animais da raça zebuína Gir com maior aptidão para produção de leite. Esses animais possuem tolerância ao calor, às doenças e aos parasitas tropicais. Essas características adaptativas tornam o Gir Leiteiro um recurso genético importante para a produção de leite nos trópicos. O estudo e caracterização de variantes estruturais (SV) do DNA desses indivíduos relacionadas a características adaptativas, de sanidade e produtivas são de interesse para o melhoramento genético da raça. Um exemplo de variações estruturais estudadas é a variação no número de cópias (CNV), que compreende deleção ou duplicação de um segmento de DNA em relação ao genoma referência. Esse tipo de SV envolve muitos pares de bases, sendo capaz de alterar a expressão gênica e de ocasionar alterações fenotípicas. Os objetivos deste trabalho foram: (1) detectar CNV no genoma de bovinos Gir Leiteiro; (2) identificar regiões de CNV (CNVR) de alta confiabilidade; (3) determinar as regiões genômicas em que ocorrem as CNV que coincidem com regiões de interesse, tais como genes e "Quantitative trait loci" (QTL) previamente relacionados às características de sanidade, reprodução e produção de leite; (4) categorizar funcionalmente os genes sobrepostos às CNV e identificar as vias biológicas enriquecidas. Após o controle de qualidade, dados de intensidade de sinal de genotipagem de SNP (polimorfismo de nucleotídeo único) provenientes do ensaio Bovine HD SNPChip de 545 indivíduos e dados oriundos de sequenciamento de genoma completo de 38 touros foram utilizados para a detecção de CNV. Ao total, 547 animais foram utilizados, dos quais 36 possuíam informação de painel de SNP e sequenciamento. Para a identificação das CNV nos dados de genotipagem foi utilizado o modelo oculto de Markov por meio do programa PennCNV e, nos dados de sequenciamento, foram utilizadas as metodologias de "Read Depth", por meio do programa CNVnator e, "Split-Read" e "Read Pair", pelo programa DELLY. O genoma referência ARS_UCD1.2 foi utilizado para o alinhamento das amostras de sequenciamento e para o mapa de SNP. Dois conjuntos de regiões de CNV (CNVR) de alta confiança foram definidos, contendo variantes encontradas nos dados de painéis de SNP e de sequenciamento. Esses conjuntos foram relativos às CNVR presentes em, no mínimo, 5% da população estudada (CNVR_POP) e aos indivíduos representativos da população Gir Leiteiro (CNVR_ANI). No conjunto CNVR_POP, foram encontradas dez CNVR, que representam 1,04 Mb do genoma bovino. No conjunto CNVR_ANI, foram encontradas 45 CNVR, que somadas representam mais de 4,4 Mb do genoma bovino. Os dois conjuntos de CNVR de alta confiança foram unidos para análise funcional resultando em 48 CNVR únicas e de alta confiança. Essas se sobrepuseram a 69 genes, incluindo *FILIP1*, *SENP6*, *CA5A*, *BANP*, *HERC2*, *RHOU*, *GBP2*, *GBP4*, *GBP6*, *BLA-DQB*, *ENSBTAG00000037605* e genes de receptores olfativos. Na análise de enriquecimento, dois termos de Gene Ontology (GO) (FDR<0,05) e 17 termos da plataforma Medical Subject Headings (MeSH) (p-ajustado<0,05) foram enriquecidos significativamente. Nas CNVR únicas e de alta confiança foram encontrados 44 "Quantitative trait loci" (QTL) significativamente associados (p<0,05) a produção (29,54%), reprodução (22,73%), conformação (18,18%), sanidade (13,64%), leite (13,63%), e carne e carcaça (2,27%). As CNVR

presentes no conjunto CNVR_POP podem ser consideradas como regiões de polimorfismos no número de cópia, pois estão presentes em mais de 1% da população estudada. Em adição, esse conjunto pode ser utilizado como critério de escolha de CNV a serem validadas, como por exemplo, por PCR. Este estudo poderá auxiliar na escolha de SNP que estejam sobrepostos às CNV para o desenvolvimento de painéis de genotipagem para o Gir leiteiro. Ainda, a inclusão de CNVR na avaliação genômica poderá trazer benefícios para o processo de seleção e deve ser verificada. Nossos resultados abrangem a identificação e caracterização de 48 CNVR de alta confiança no genoma de bovinos Gir Leiteiro, o que contribui para a elaboração de um mapa de SV na raça Gir e para o melhor entendimento do genoma dos zebuínos. As CNVR identificadas neste estudo podem afetar potencialmente genes que estão envolvidos no processo evolutivo e no controle fenotípico de característica de interesse para a cadeia produtiva leiteira, como imunidade, lactação, reprodução, reconhecimento de estímulos e sanidade.

Palavras-chave: bovinos de leite, genes, genética animal, variações do número de cópias de DNA, zebu

IDENTIFICATION OF STRUCTURAL VARIANTS OF THE DAIRY GIR CATTLE

ABSTRACT – The Dairy Gir cattle results from the selection of animals from the Zebu Gir breed with greater aptitude for milk production. These animals are tolerant of heat, disease, and tropical parasites. These adaptive traits make the Dairy Gir an important genetic resource for milk production in the tropics. The study and characterization of structural variants (SV) of the DNA of these animals related to adaptive, healthy, and productive traits are interesting to the animal breeding of the breed. An example of structural variants is copy number variation (CNV), which comprises deletion or duplication of a DNA segment compared to the reference genome. The CNV can alter the gene expression and consequently change phenotypes because it encompasses many base pairs. The objectives of the work were: (1) detect CNV in the genome of Dairy Gir cattle; (2) identify high-confidence CNVR; (3) determine the genomic regions overlapping CNVR that coincide with the interest regions, such as genes and “Quantitative Trait Loci” (QTL) previously related to health, reproduction, and milk production traits; (4) functionally categorize genes overlapped with CNVR and identify the enriched biological pathways. After the quality control, luminosity data from 545 individuals genotyped with Illumina BovineHD BeadChip and whole-genome sequencing data from 38 bulls were used to detect CNV. In total, 547 Dairy Gir animals were used, and 36 animals had both types of information. To identify the CNV in the SNP (single nucleotide polymorphism) array data, the hidden Markov model methodology was used by the PennCNV software and, in the sequencing data, the “Read Depth” methodology was used by the CNVnator software, and “Split-Read” and “Read Pair” methodologies by the DELLY software. The reference genome ARS_UCD1.2 was used for the alignment and the SNP map. Two sets of high confidence CNV regions (CNVR) were defined, containing variants found in the SNP array and sequencing data. These sets are related to the CNVR present in at least 5% of the studied population (CNVR_POP) and the representative animals of the Dairy Gir population (CNVR_ANI). In the CNVR_POP set, ten CNVR were found, covering 1.04 Mb of the bovine genome. In the CNVR_ANI set, 45 CNVR were found, which together cover more than 4.4 Mb of the bovine genome. The two high confidence CNVR sets were merged for functional analysis, resulting in 48 unique and high-confidence CNVR. These overlapped 69 genes, including FILIP1, SENP6, CA5A, BANP, HERC2, RHOU, GBP2, GBP4, GBP6, BLA-DQB, ENSBTAG00000037605, and odorant receptors genes. In the enrichment analysis, two Gene Ontology (GO) terms (FDR<0.05), and 17 terms (p-adjusted<0.05) from the Medical Subject Headings (MeSH) platform were significant. In the unique and high confidence CNVR, 44 Quantitative trait loci (QTL) were significantly associated (p<0,05) with production (29.54%), reproduction (22.73%), conformation (18.18%), health (13.64%), milk (13, 63%), and meat and carcass (2.27%). The CNVR in the CNVR_POP set may be considered regions of copy number polymorphism since they are present in more than 1% of the studied population. In addition, this set can be used to choose the CNV to be validated, such as by PCR. This study may support the selection of markers that flanks CNV to the development of SNP arrays for the Dairy Gir. Also, the inclusion of CNVR in the genomic evaluation could benefit the selection process and must be verified. Our results include the identification and characterization of 48 high-confidence CNVR in the genome of Dairy Gir cattle. These data contribute to elaborate an SV map in the

Gir breed and a better understanding of the genome of the zebu cattle. The CNVR identified in this study may potentially affect genes involved in the evolutionary process and the phenotypic control of traits of interest to the dairy products market, such as immunity, lactation, reproduction, stimulus recognition, and health.

Keyword: animal genetics, dairy cattle, DNA copy number variations, genes, zebu

Lista de Abreviaturas

BAF	Frequência do alelo B
bp	Pares de base
CGH	Hibridização genômica comparativa
CNV	Variações no número de cópias
CNVR	Região de variações no número de cópias
DS	Duplicações segmentares
FISH	Hibridização *in situ* fluorescente
FoSTeS	“Fork stalling and template switching”
GC	Guanina e citosina
GO	“Gene Ontology”
HMM	Modelo oculto de Markov
HTS	Sequenciamento de alto rendimento
Kb	Quilobases
LINE	“Long interspersed nuclear elements”
KEGG	“Kyoto encyclopedia of genes and genomes”
LRR	Log da razão de R
Mb	Megabases
MeSH	“Medical subject headings”
NAHR	Recombinação homóloga não alélica
NHEJ	Ligação das extremidades não homólogas
PCA	Análise de componentes principais
PNMGL	Programa Nacional de Melhoramento do Gir Leiteiro
qPCR	Reação em cadeia da polimerase em tempo real
QTL	“Quantitative trait loci”
RFLP	Polimorfismo no comprimento dos fragmentos de restrição
RP	“Read-pair”
RD	“Read depth”
SNP	Polimorfismo de nucleotídeo único
SR	“Split-read”
SV	Variações estruturais

Lista de Tabelas

Lista de Figuras

CAPÍTULO 01 – CONSIDERAÇÕES GERAIS

1 INTRODUÇÃO

Espera-se que a produção mundial leiteira cresça 1,6% ao ano, passando a 977 mil toneladas de leite em 2029. Esse crescimento deverá ser impulsionado pela otimização dos sistemas de produção de leite, melhoria da sanidade animal, maior eficiência na alimentação e melhoria genética dos rebanhos (OECD/FAO, 2020). Mesmo com a retração na produção leiteira nacional em 2020, devido à redução de demanda por causa da pandemia do COVID-19 (Agro em Dia, 2020) e à seca em parte do país, o Brasil ocupa a quinta posição no ranking de países produtores de leite com 33.954 mil toneladas de leite (FAO, 2020).

Devido ao clima nos trópicos e as condições de criação, o cenário da produção leiteira no país é composto principalmente por rebanhos zebuínos ou cruzados. Animais cruzados, principalmente aqueles resultantes do cruzamento entre a raça Holandês e Gir Leiteiro respondem por mais de 80% do leite produzido no Brasil (Cole e Silva, 2016). A raça Gir, oriunda da Índia, foi introduzida no país no século XIX e é um dos recursos genéticos mais importantes para a produção de leite na região tropical, uma vez que é tolerante aos parasitas e ao calor (Santana et al., 2014).

O Gir Leiteiro resulta da seleção fenotípica de animais da raça Gir com maior aptidão para produção de leite. A seleção artificial nesses animais tem sido direcionada para a produção de leite e seus constituintes, características reprodutivas, de conformação e manejo e, desde 2018, o Programa Nacional de Melhoramento do Gir Leiteiro (PNMGL) utiliza a informação genômica para a estimação de valores genéticos (Panetto et al., 2020). A informação genômica é utilizada pelos principais programas de melhoramento no mundo, incluindo o do Gir Leiteiro, sendo necessário conhecer e caracterizar o seu potencial genético continuamente.

Na raça Holandesa, a inclusão de informação genômica na avaliação genética nos Estados Unidos, além de ter aumentado a acurácia das predições, diminuiu o intervalo de geração e aumentou o ganho genético anual (García-Ruiz et al., 2016). Em adição, a informação genômica é capaz de tornar a seleção de touros para o

teste de progênie mais confiável e diminuir os custos do teste que, especialmente no gado de leite, é longo e dispendioso.

A proposição da estrutura do DNA (WATSON; CRICK, 1953) permitiu o início da realização de estudos genômicos e, a partir de então, houve avanço nesse tipo de estudo. Na década de 1980, iniciaram-se as pesquisas com marcadores genéticos para espécies de interesse zootécnico, sendo as primeiras publicações sobre a caracterização de polimorfismo no comprimento de fragmentos de restrição (RFLP) em suínos (Chardon et al., 1985) e bovinos (BECKMANN et al., 1986; GEORGES et al., 1987). O avanço das técnicas de biologia molecular e de bioinformática vem permitindo estudos genômicos de larga escala, por meio da genotipagem de marcadores e do sequenciamento do material genético.

Essas técnicas moleculares permitem o estudo e caracterização de variantes do DNA, tais como variações estruturais, pequenas inserções e deleções (“indels”) e o polimorfismo de nucleotídeo único (SNP). O SNP é definido como uma alteração em um único nucleotídeo na sequência de DNA, que ocorre em mais de 1% da população estudada (Collins et al., 1998).

Uma das variações estruturais estudadas é a variação no número de cópias (CNV), que compreende deleção ou duplicação de um fragmento de DNA em relação ao genoma referência (Feuk et al., 2006). As CNV são menos frequentes que SNP e “indels” e, ainda assim, essas variantes podem provocar impactos funcionais e na evolução, uma vez que, por afetar a expressão gênica, pode afetar fenótipos de interesse (Bickhart et al., 2012).

Alterações no número de cópias de genes podem conduzir a mudanças na dosagem gênica, que é um mecanismo pelo qual a alteração do número de cópias de um gene modifica o seu perfil de expressão. Por exemplo, CNV que se sobrepõe a um gene sensível à dosagem pode provocar redução da expressão ou, ainda, induzir a formação de novos transcritos por meio da interrupção gênica (Gamazon e Stranger, 2015). Em suínos da raça Landrace, Rubin et al. (2012) relataram que a duplicação de uma região de 450 quilobases (Kb) que envolve o gene *KIT* (*KIT Proto-Oncogene, Receptor Tyrosine Kinase*) e uma mutação de um sítio de “splicing” em pelo menos uma das cópias desse gene provoca o fenótipo de coloração totalmente branca.

A detecção do tipo e tamanho de CNV depende do tamanho da amostra da população, da variabilidade genética, da metodologia de detecção e da técnica molecular utilizada para obter a informação genômica (Yang et al., 2021). Dentre esta última, estão o sequenciamento de alto rendimento (HTS - “high-throughput sequencing”) e os painéis de SNP. A detecção de CNV está sujeita a ocorrência de falsos positivos e estratégias devem ser tomadas para aumentar a sua acurácia. Por exemplo, regiões de CNV (CNVR) identificadas por dois tipos de técnicas moleculares e por meio de duas metodologias podem ser consideradas como de alta confiança (Zhan et al., 2011). As CNVR são determinadas pelo agrupamento de CNV sobrepostas (Redon et al., 2006).

As CNV foram identificadas na variação fenotípica de características de interesse zootécnico, tais como: características reprodutivas em suínos (Zheng et al., 2020), adaptação térmica em peixes marinhos (Cayuela et al., 2021); resposta ao estresse em codornas (Khatri et al., 2019) e características de crescimento em ovinos (Feng et al., 2020). Estudo anteriores identificaram associação desse tipo de variante com características de interesse econômico, como, afecções do casco (Butty et al., 2021), produção de leite (Ben Sassi et al., 2016; Xu et al., 2014), sanidade (Szyda et al., 2019), reprodução e características de conformação (Ben Sassi et al., 2016). Dada a importância econômica do Gir Leiteiro e a relevância das variações estruturais, novos conhecimentos sobre CNV na raça poderão auxiliar na identificação de animais superiores geneticamente, na escolha de touros para participar do teste de progênie e no aumento do ganho genético da raça. Isso evidencia a necessidade de mais estudos sobre o assunto na raça Gir. Os objetivos deste trabalho foram: (1) detectar CNV no genoma de bovinos Gir Leiteiro; (2) identificar CNVR de alta confiabilidade; (3) determinar as regiões genômicas em que ocorrem as CNV que coincidem com regiões de interesse, tais como genes e “Quantitative trait loci” (QTL) previamente relacionados às características de sanidade, reprodução e produção de leite; (4) categorizar funcionalmente os genes sobrepostos às CNV e identificar as vias biológicas enriquecidas.

2 REVISÃO DE LITERATURA

2.1 A raça Gir

A raça zebuína Gir é originária da Índia, da península Kathiawar. O período de importação de animais da Índia compreendeu os anos de 1870 a 1962, em que foram trazidos cerca de 6.000 zebuínos, sendo menos de 700 bovinos Gir. Devido a sua adaptação ao clima do país, o gado zebuíno e o cruzamento desse com animais criolos se difundiu rapidamente (O'Brien et al., 2015).

Em 1938, a criação do livro genealógico das raças zebuínas pela Associação de Criadores de Zebuínos (ABCZ), em Uberaba, no estado de Minas Gerais, contribuiu para a regularização e disseminação da raça Gir no país. Na década de 1960, parte dos criadores praticava a seleção para dupla aptidão e outro pequeno grupo apenas para produção leiteira. Finalmente, após décadas de seleção fenotípica, em 1985, foi criado o Programa Nacional de Melhoramento do Gir Leiteiro (PNMGL). No mesmo ano, foi iniciado o teste de progênie, contribuindo, assim, para o crescimento populacional e aumento de produção da raça (Santana et al., 2014).

Os estudos moleculares na raça começaram em 2001, quando iniciou-se a colheita de sangue e sêmen para a formação do banco de DNA do Gir Leiteiro. A partir disso, estudos moleculares usando marcadores microssatélite foram desenvolvidos para os genes das proteínas do leite Kappa-caseína e Beta lactoglobulina. As análises para doenças hereditárias foram introduzidas no ano de 2013, tais como DUMPS (Síndrome da Deficiência de Síntese de Uridina Monofosfatase), CVM (Má-formação do Complexo Vertebral) e BLAD (Deficiência de Adesão Leucocitária Bovina). Pesquisas sobre o gene da beta-caseína também foram realizadas e a informação sobre os touros que transmitem os alelos A1 ou A2 da beta-caseína está presente nos sumários (PANETTO et al., 2020). A Prova de Pré-Seleção de Touros para o Teste de Progênie foi iniciada em 2009 e, a partir de 2016, informações genômicas foram utilizadas para verificação de parentesco e para a escolha de candidatos, o que resultou em aumento das médias dos valores genéticos dos touros em avaliação pelo Teste de Progênie dos Touros a partir de 2017 (FERNANDES et al., 2020).

Em 2018, a inclusão da informação genômica na avaliação genética foi iniciada no Gir Leiteiro. Esse programa tem avaliado características associadas ao

temperamento, produção de leite e seus componentes (gordura, proteína e sólidos), conformação corporal, sanidade e longevidade. O PNMGL utiliza ainda marcadores moleculares para predição de valores genéticos genômicos de machos e de fêmeas (PANETTO et al., 2020), o que justifica a importância do desenvolvimento de estudos genômicos na raça.

O Gir Leiteiro está adaptado as características e particularidades da produção leiteira nos trópicos. A raça Gir é considerada a principal raça zebuína leiteira no Brasil (PRATA et al., 2015). No país, o Gir Leiteiro também é utilizado em cruzamento com a raça Holandesa ou ainda com raças localmente adaptadas. O Gir Leiteiro.

2.2 Variações Estruturais do DNA

Variações estruturais (SV - “structural variants”) são alterações genômicas de, no mínimo, 50 pares de bases (bp – “base pairs”) (Figura 1). As SV podem ser divididas em balanceadas e não balanceadas. Dentre as balanceadas, estão as inversões e translocações. As SV não balanceadas envolvem perdas ou ganhos genômicos, como inserções, duplicações e deleções (Kosugi et al., 2019). Deleções e duplicações genômicas, quando em relação a um genoma referência, também podem ser chamadas de variações no número de cópias (CNV – “copy number variation”) (Feuk et al., 2006).

As CNV são consideradas variantes polimórficas e são herdadas entre gerações (MCCARROLL; ALTSHULER, 2007). Esse tipo de alteração genômica afeta grande parte do genoma, se comparada a outros tipos de variações, como os SNP e “indels”, uma vez que envolve vários pares de base.

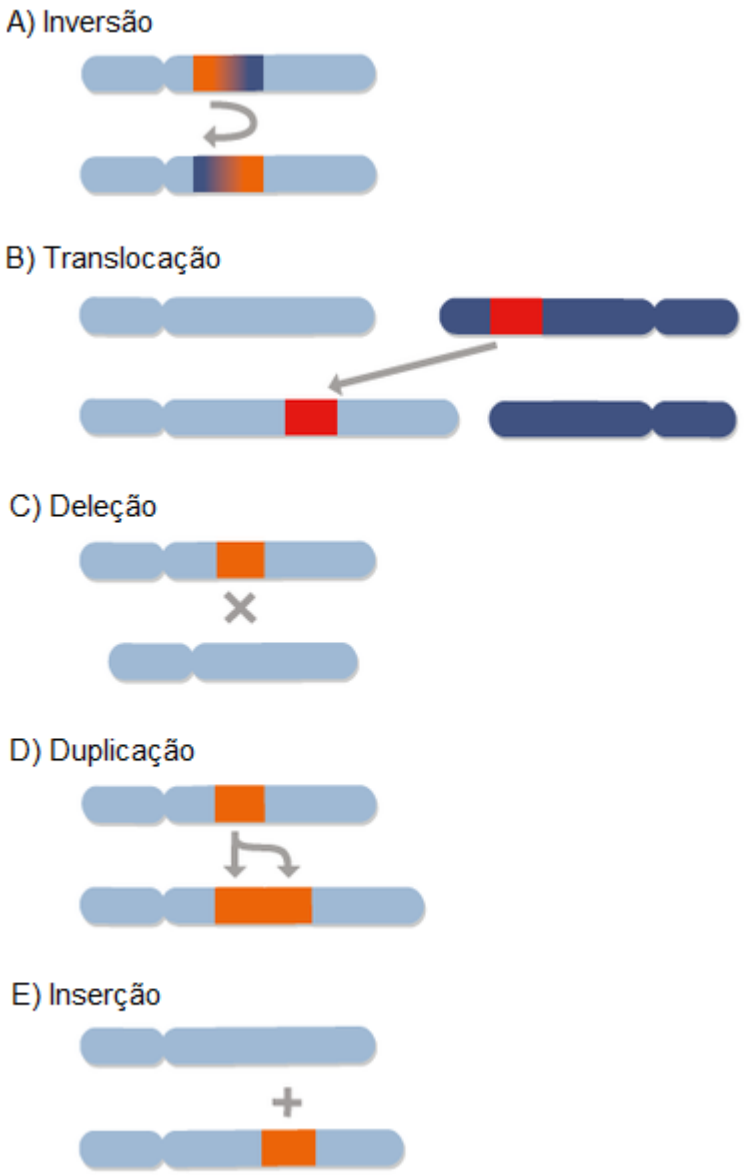


Figura 1. Variações Estruturais. A) A sequência de DNA é invertida de ponta a ponta; B) Um segmento de DNA é removido de um cromossomo e adicionado a outro; C) Um segmento de DNA é removido de um cromossomo; D) Um segmento de DNA é repetido no mesmo cromossomo; E) Um segmento de DNA é adicionado a um cromossomo. Adaptado de Kinghorn Centre for Clinical Genomics (2018). Disponível em: https://www.garvan.org.au/research/kinghorn-centre-for-clinical-genomics/learn-about-genomics/dna-base/collection1/structural-variation

Essas variações são abundantes e podem influenciar a expressão gênica e variação fenotípica por meio de mecanismos moleculares, como desregulação e alteração da dosagem gênica, fusão gênica, interrupção gênica, efeitos de posição, desmascaramento do alelo recessivo ou polimorfismos funcionais e efeitos de

transvecção. As CNV podem provocar fusão gênica devido à união de dois genes ou suas sequências regulatórias. A interrupção gênica ocorre quando o ponto de quebra de uma CNV está localizado dentro de um gene e essa CNV interrompe o gene, o que causa inativação e consequente, perda de função. O efeito de posição ocorre quando, por exemplo, ao remover ou alterar determinada sequência reguladora, a CNV provoca efeitos na expressão ou na regulação do gene que não está na região afetada pela CNV. O desmascaramento do alelo recessivo ocorre devido à deleção do alelo não recessivo e assim, o alelo recessivo passa a ser expresso. A transvecção ocorre quando uma sequência reguladora localizada em um dos cromossomos homólogos regula a unidade de transcrição no outro homólogo, nesses casos, CNV provocam alteração por meio deste mecanismo de trans regulação (Zhang et al., 2009).

Quando o termo CNV ainda não era estabelecido, a primeira evidência de que alterações no número de cópia de genes poderiam influenciar fenótipos nos humanos foi relatada em estudos que relacionavam rearranjos genômicos à doenças genômicas (Inoue e Lupski, 2002). As CNV são capazes de provocar efeitos sobre a expressão gênica. Por exemplo, se o ponto de quebra de uma deleção, inserção ou duplicação está localizado em um gene, isso pode causar a perda de sua função devido ao rompimento ou dissociação de promotores ou outros elementos regulatórios, ou ainda impactar a estrutura da cromatina (Zhang et al., 2009). Esse tipo de SV pode, indiretamente, afetar elementos reguladores da transcrição gênica mediante um efeito posicional, levando a níveis alterados de expressão gênica. Caso haja deleção de um elemento repressor, essa pode ocasionar uma alta taxa de expressão do gene associado, ao passo que as duplicações de sequências presentes na posição 3’ em relação ao promotor, podem levar a uma menor taxa de transcrição (Rodriguez-Revenga et al., 2007).

Se as CNV são vantajosas, essas podem evoluir sob seleção positiva ou negativa, entretanto, quando essas CNV envolvem genes essenciais podem sofrer seleção negativa a depender do seu impacto (Liu e Bickhart, 2012). Famílias de genes envolvidas na manutenção e controle do ciclo celular e da transcrição são normalmente desprovidas de eventos de CNV (Hou et al., 2011). Em adição, CNV

geralmente se localizam em regiões não gênicas e fora de elementos ultra conservados (Redon et al., 2006)

2.3 Mecanismo de formação de SV

Para a melhor compreensão da complexidade e da importância de SV, faz-se necessário o estudo sobre sua formação e detecção. Para isso, devem ser consideradas algumas teorias sobre os mecanismos envolvidos na formação de SV a saber: recombinação homóloga não-alélica (NAHR), ligação das extremidades não homólogas (NHEJ), “Fork Stalling and Template Switching” (FoSTeS), e retrotransposição mediada por LINE (“Long Interspersed Nuclear Elements”) (Zhang et al., 2009). NAHR e NHEJ são mecanismos de reparação do DNA. A NAHR ocorre entre duplicações segmentares (DS), que são consideradas catalisadoras e “hotspots” para a formação de CNV (Sharp et al., 2005). A NAHR ocorre pelo alinhamento entre DS não alélicas resultando em rearranjos genômicos, tais como duplicações, deleções, inversões ou mesmo translocações cromossômicas quando a recombinação ocorre em cromossomos não homólogos (Kim et al., 2008; Shaw e Lupski, 2004). As DS, também conhecidas como sequências “low copy-repeat, são blocos de DNA que variam de 1 a 400 Kb de comprimento. Diferentemente das CNV, as DS normalmente possuem um alto nível de identidade de sequência (> 90%) (Eichler, 2001) e geralmente são encontradas em regiões pericentroméricas e subteloméricas (Bailey et al., 2001).

A NHEJ é um mecanismo utilizado pelas células para reparar a quebra da dupla fita de DNA (DSB) que pode ocorrer em todas as etapas do ciclo celular. Esse mecanismo envolve a ligação das extremidades de DNA e permite a inserção ou deleção de nucleotídeos no local de ligação das fitas quebradas, gerando diversos resultados (Lieber, 2008). O mecanismo FoSTeS ocorre durante a replicação do DNA, quando a fita recém-sintetizada pode se desprender da fita molde de DNA e se anelar em outra forquilha de replicação com micro homologia entre as regiões, continuando a síntese da fita na forquilha invadida (Lee et al., 2007). A FoSTeS pode gerar duplicações genômicas grandes com muitas megabases (Mb) e também implicar na duplicação/triplicação gênica e rearranjo de éxons simples. A duplicação

gênica e o embaralhamento de éxons são os principais eventos que atuam na evolução gênica e genômica (Zhang et al., 2009).

As SV também podem ocorrer pela inserção de retrotransposons (LINE), são elementos transponíveis da classe I, que utilizam um mecanismo baseado na transcrição e na transcrição reversa e contam com o auxílio da enzima integrase. Com isso, a molécula de DNA móvel integra-se às regiões dos cromossomos e, desde que a cópia original do cromossomo seja mantida no mesmo local de origem, este mecanismo causa duplicação. Assim, tal processo gera um grande número de cópias de retro elementos no genoma de diversas espécies (Böhne et al., 2008).

2.4 Detecção de SV: influência do genoma referência, identificação e validação de SV

A versão ARS-UCD1.2 da montagem do genoma referência bovino (Rosen et al., 2020) utiliza diferentes tecnologias de sequenciamento e por isso possui “gaps” menores, sendo mais completa, o que fornece maior confiabilidade na posição de genes e marcadores genéticos (Rosen et al., 2020). Assim, pode haver diferenças entre as CNV identificadas entre as diferentes versões do genoma referência. Isso permite a identificação mais acurada de pequenas variações (SNP e “indels”) e SV, tais como as CNV. O genoma referência ARS-UCD1.2 em comparação com a montagem UMD3.1.1 (Zimin et al., 2009) possui melhoria de 10X na acurácia por base, 200X na continuidade da sequência, apenas 345 “contigs” (UMD3.1.1: 72.264 “contigs”), N50 de 25.8 Mb (UMD3.1.1: N50 de 0.092 Mb) e L50 de 32 “contigs”. “Contigs” são sequências maiores de DNA que são montadas por meio de sobreposição de leituras - “reads”. N50 é o comprimento mínimo de “contigs” necessário para cobrir 50% do genoma. L50 é o número de “contigs” necessários para atingir N50.

Lee et al. (2020) identificaram CNV a partir de dados do painel de SNP Illumina BovineHD BeadChip (Illumina, Inc., San Diego, CA, USA), em que foram utilizadas duas versões do genoma referência. Comparando-se os resultados obtidos a partir da montagem ARS-UCD1.2 com a montagem UMD3.1.1, o número de CNV por indivíduo foi 42% menor, mais SNP foram necessários para se detectar uma CNV e o tamanho de CNVR complexas ou mistas (CNVR que continham

ambos eventos de deleção e duplicação) foi menor utilizando a montagem ARS-UCD1.2. Os autores investigaram a região (BTA12:70-77 Mb) em que houve grande diferença entre as duas montagens de referência. Nessa região, observou-se que cerca de 43% dos SNP localizados na UMD3.1.1 foram movidos para “contigs” não mapeados ou não estavam definidos na ARS-UCD1.2. Dessa forma, as duas montagens referência diferem e isso pode levar a resultados discordantes na detecção de CNV (Lee et al., 2020), em relação ao número e tamanho de eventos.

A detecção das CNV pode ser realizada por diferentes técnicas moleculares, como hibridização genômica comparativa mediante microarranjos de DNA (CGH), painéis de genotipagem de SNP (“beadchips”) e HTS ou sequenciamento de nova geração (Bickhart e Liu, 2014). Arranjos de CGH foram muito utilizados para a detecção de CNV. Entretanto, com o desenvolvimento da tecnologia genômica, os painéis de SNP e o sequenciamento, que possuem maior poder de detecção de variantes estruturais, passaram a ser utilizados. A detecção de variantes a partir de micro arranjo de CGH depende do número, tamanho e qualidade das sondas presentes.

A detecção de CNV a partir de painéis de SNP depende da densidade e a distribuição de marcadores no genoma, por exemplo, CNV podem estar presentes em regiões não cobertas por SNP ou não abrangerem o número mínimo de marcadores necessários para serem detectadas. As CNV menores podem ser capturadas como uma única grande CNV de tamanho superestimado em painéis de SNP de média ou baixa densidade devido as distâncias entre marcadores. A escolha de SNP para a confecção de painéis é influenciada pela presença e frequência de marcadores nas raças estudadas e pela qualidade de genotipagem de SNP. Os SNP que não atendem aos padrões de herança mendeliana esperados e os com baixo agrupamento de genotipagem (por exemplo, escore “GenTrain”) tendem a ser considerados erros de genotipagem e podem não ser incluídos nos painéis de genotipagem. No entanto, a presença de CNV ou indels também podem provocar inconsistências Mendelianas e baixo agrupamento de genotipagem (Berry et al., 2019; Rafter et al., 2020).

A detecção de CNV a partir de dados de painéis de SNP é baseada na intensidade do sinal de fluorescência emitido para cada marcador, em que reduções

na intensidade do sinal podem representar deleções e o aumento da intensidade indica duplicações (Cassese et al., 2014; Wang et al., 2007). Essa detecção é baseada em duas principais medidas do sinal da fluorescência dos marcadores, pela Log da razão de R ("Log R Ratio" - LRR) e pela frequência do alelo B ("B allele frequency" - BAF). A LRR é uma medida normalizada da intensidade do sinal para cada SNP, em que as intensidades de sinal para os alelos A e B podem ser representadas por X e Y e, calculando-se o valor de R, tem-se que $R_{observado}$= X + Y. A LRR é calculada como $LRR = log2(R_{observado}/R_{esperado})$ (Diskin et al., 2008). A BAF refere-se a uma medida normalizada da taxa de intensidade relativa de sinal de cada alelo para cada marcador (Wang et al., 2007).

Existem diversos programas computacionais para a detecção de CNV a partir dos dados de genotipagem de painéis de SNP, tais como QuantiSNP (Colella et al., 2007), PennCNV (Wang et al., 2007), GADA (Pique-Regi et al., 2008), GenoCN (Sun et al., 2009) e MixHMM (Liu et al., 2010b). Por exemplo, o PennCNV (Wang et al., 2007) incorpora múltiplas fontes de informação para identificar os números de cópias, o LRR e a BAF para cada marcador SNP para cada indivíduo, a frequência populacional do alelo B (PFB – "population frequency of B allele"), o arquivo HMM (HMM – "hidden Markov model") e a informação de "pedigree", quando disponível, em modelo oculto de Markov. O arquivo HMM, que é fornecido pelo próprio programa, contem valores esperados para cada estado de número de cópias e a probabilidade de transição esperada para diferentes estados de número de cópias com base na distância entre marcadores vizinhos. Os estados ocultos de Markov identificados por esse programa e o número de cópias correspondentes estão descritos na Tabela 1.

A detecção a partir de painéis de SNP pode ser afetada pela baixa densidade dos arranjos e hibridização cruzada de sequências repetidas (Bickhart et al., 2012). Estudos de detecção de CNV em bovinos têm sido publicados, indicando o HTS como principal técnica molecular para identificação de CNV, devido ao maior poder de detecção e cobertura do genoma (Bickhart et al., 2012; Keel et al., 2016; Letaief et al., 2017).

Tabela 1. Estados ocultos de Markov, número de cópias, descrição para os autossomos e genótipos fornecidos pelo programa PennCNV

Estado	Número de cópias	Descrição para os autossomos	Genótipos
1	0	Deleção dupla	Nulo
2	1	Deleção de uma cópia	A, B
3	2	Normal	AA, AB, BB
4	2	Cópia neutras com perda de heterozigose (LOH)	AA, BB
5	3	Duplicação simples	AAA, AAB, ABB, BBB
6	4	Duplicação dupla	AAAA, AAAB, AABB, ABBB, BBBB

As análises a partir de dados de HTS possibilitaram o aprimoramento da identificação de CNV, uma vez que permitem a identificação mais acurada de pequenas CNV e os pontos de quebra (limites) das variações. O sequenciamento é capaz de prover resultados com alta resolução e sensibilidade (Bickhart et al., 2012). No entanto, a identificação de CNV a partir de dados de sequenciamento requer infraestrutura computacional para armazenar e analisar os dados. Além disso, os resultados podem ser influenciados pelo comprimento das leituras, eficiência do mapeamento, cobertura de alinhamento ao genoma e viés do conteúdo de guanina e citosina (GC) da tecnologia de sequenciamento (Alkan et al., 2011).

Diversas metodologias foram desenvolvidas para permitir a detecção de CNV a partir de dados de sequenciamento. As quatro principais metodologias disponíveis para a detecção a partir de dados de HTS são: “read-pair” (RP), “split-read” (SR), “read depth” (RD) e montagem, e também há a combinação dessas. Cada um desses métodos possui diferentes vantagens e desvantagens em sua aplicação e adequação para dados de HTS, e nenhum é capaz de identificar toda a variação presente no DNA. Na busca pelo melhor aproveitamento dos diferentes métodos, ferramentas computacionais foram desenvolvidas a partir de abordagem combinatória, associando as metodologias, com o intuito de aumentar a acurácia dessas análises (Pirooznia et al., 2015).

Pelo método RP compara-se a distribuição do comprimento das leituras mapeadas no genoma com a distribuição esperada no próprio genoma, em que uma diferença de distância entre as extremidades indica deleção ou duplicação. Se as extremidades dos fragmentos são mapeadas em uma orientação diferente da esperada, isso pode indicar inversão (Feuk, 2010). Tal método detecta SV de tamanho médio, mas pode ser insensível para as pequenas, devido a sua dificuldade em detectar precisamente pequenas alterações (Medvedev et al., 2009). Os programas computacionais que incluem métodos RP são: BreakDancer (Chen et al., 2009), PEMer (Korbel et al., 2009) e Ulysses (Gillet-Markowska et al., 2015).

O PRISM (Jiang et al., 2012), o SVseq2 (Zhang et al., 2012) e o Gustaf (Trappe et al., 2014) são alguns dos programas computacionais que utilizam o método baseado em SR, consistindo no mapeamento de apenas uma das leituras no genoma de referência. A metodologia SR utiliza leituras de sequenciamento “paired-end” em que apenas uma das leituras do par é mapeada de forma confiável e a outra é parcialmente ou completamente não mapeada no genoma (Gong et al., 2021). Tal método é capaz de identificar pontos de quebra das SV de forma acurada, mas possui capacidade limitada para identificar variações estruturais em larga escala (Pirooznia et al., 2015).

O método RD depende da densidade de alinhamento das leituras ao longo dos cromossomos, em que regiões de alta cobertura podem representar duplicações e as que apresentarem baixa cobertura, deleções (Pirooznia et al., 2015). Alguns dos programas computacionais que utilizam tal método são: CNV-seq (Xie e Tammi, 2009), CNVnator (Abyzov et al., 2011), ReadDepth (Miller et al., 2011) e cn.MOPs (Klambauer et al., 2012). O RD possui desempenho superior na detecção de CNV longos, que são mais difíceis de detectar por RP e SR. Além disso, esse método permite identificar os número de eventos de cada CNV (estados), diferentemente dos outros métodos (Yoon et al., 2009). Por exemplo, no programa CNVnator, o número de cópias pode ser obtido pela função *genotype*, em que o valor do sinal de RD < 0,5 representa deleção dupla, 0,5 ≤ RD <1,5 representa deleção de uma cópia e RD ≥ 1,5 refere-se ao número de cópias real (como valor arredondado).

O método de montagem é o menos utilizado, pois requer muitos recursos computacionais e é menos acurado (Pirooznia et al., 2015). Ainda há programas

computacionais que utilizam os métodos combinados, tais como DELLY (Rausch et al., 2012), LUMPY (Layer et al., 2014), SRBreak (Nguyen et al., 2016) e Parliament2 (Zarate et al., 2020).

Para a confirmação de cada CNV encontrada, se faz necessário validar as regiões encontradas por meio de técnicas moleculares, como qPCR (PCR em tempo real) e FISH (Hibridização *in situ* fluorescente) (Bickhart et al., 2012). Entretanto, esses testes moleculares demandam tempo, possuem custo elevado quando aplicados à um grande número de amostras e de CNV, e necessitam de quantidade suficiente de material biológico. Uma alternativa que tende a aumentar a confiabilidade da detecção das CNV é a verificação *in silico* dessas variantes estruturais por meio da utilização de diferentes de painéis de SNP e HTS.

Diferentes métodos de verificação e validação *in silico* tem sido aplicados na detecção de CNV nos bovinos, tais como a utilização de mais de uma técnica molecular (Butty et al., 2020; Chud, 2018; Letaief et al., 2017; Zhan et al., 2011), combinação de distintas metodologias de detecção (Keel et al., 2017; Pirooznia et al., 2015; Yang et al., 2021; Zhan et al., 2011), abordagem de herança mendeliana (Chen et al., 2017) e uso de mais de um sequenciamento do mesmo animal (“twice-sequenced animals”) (Chen et al., 2017). A comparação de resultados entre estudos de detecção de CNV pode ser realizada, comprovando que determinada CNV ou CNVR foi encontrada anteriormente. Entretanto, a sobreposição entre os conjuntos de CNV previamente descritas pode ser baixa. Isso se deve às diferenças entre os estudos de diferentes raças, tamanhos de amostra, metodologia, técnica molecular (Keel et al., 2017), e versão da montagem do genoma referência utilizada.

Todos os métodos de detecção de CNV possuem limitações e nenhum é capaz de identificar toda a variação presente no DNA (Pirooznia et al., 2015). A detecção de CNV a partir de painéis de SNP possui diversas limitações, assim, CNV detectadas por esse tipo de plataforma de genotipagem podem ser apenas uma fração daqueles que realmente existem e podem ser detectados por dados de sequenciamento (Rafter et al., 2020).

Independentemente da metodologia, as que são baseadas no sequenciamento de leituras curtas (menores do que 300 bp, como as da tecnologia Illumina) podem sofrer alta taxa de falsos positivos devido aos erros que ocorrem na

determinação das bases e no alinhamento (“mismappings”), especialmente em regiões repetitivas, que não são possíveis de serem abrangidas com esse tipo de tamanho de leitura. Para superar as dificuldades da detecção de CNV a partir de sequenciamento de leituras curtas, pode ser utilizado o sequenciamento de leituras longas (maiores do que 10.000 bp, como as das tecnologias Pacbio e Oxford Nanopore). No entanto, o alto custo e o baixo rendimento dessa tecnologia pode limitar seu uso (Kosugi et al., 2019). Assim, a combinação de diferentes métodos de detecção e plataformas pode ser considerada uma alternativa eficiente e menos onerosa para contornar as dificuldades na detecção de CNV e diminuir o número de falsos positivos e aumentar a probabilidade de detecção de CNV confiáveis.

2.5 Fenótipos de interesse relacionados à CNV e importância das CNV

A análise de associação de CNVR com fenótipos de interesse e a identificação de genes e QTL que coincidem com CNVR encontradas no genoma podem ser utilizadas para compreender os processos biológicos de características complexas e a importância das SV, que são capazes de implicar em variação fenotípica. A metodologia da análise de associação entre CNVR e características de interesse não é tão bem estabelecida quanto a do estudo de associação ampla com SNP (GWAS) (Kim et al., 2012). Vários métodos estatísticos podem ser empregados para esse tipo de estudo com CNVR, como por exemplo regressão linear (Butty et al., 2021; Xu et al., 2014; Zhou et al., 2018, 2016), regressão bayesiana (Ben Sassi et al., 2016) e modelos mistos (Aguiar et al., 2018; Sasaki et al., 2021). Ainda, a anotação do genoma e as informações depositadas nos bancos de dados genômicos podem ser limitantes para a identificação de CNV de importância fenotípica.

Xu et al. (2014) realizaram estudo de análise de associação ampla utilizando CNV detectadas em amostras de 26.363 animais Holandeses genotipados com Illumina BovineSNP50 array version 1 (Illumina Inc., San Diego, CA). Esses autores encontraram 34 CNV associadas a pelo menos uma dessas características: produção de leite, produção de gordura, produção de proteína, porcentagem de gordura e porcentagem de proteína.

Sasaki et al. (2021) encontraram uma CNV de 44 Kb (BTA5:103.317.687-103.361.802) associada a mortalidade de bezerros de um a 180 dias de idade em animais Wagyu. A região deletada compreendia o gene *C1RL* (C*omplement C1r subcomponent like*), um pseudogene e um IncRNA, que é um componente da via clássica de ativação do sistema complemento. A ativação dessa via é iniciada por complexos antígeno-anticorpo e resulta em complexos imunes contra patógenos infecciosos. Nos animais que possuíam a deleção em homozigose, a atividade média desse sistema no sétimo dia após o nascimento foi significativamente (p-valor = 0,002) menor do que a de animais de tipo selvagem e heterozigoto. Isso atribui importância na suscetibilidade a infecções associada essa CNV.

Butty et al. (2021) realizaram estudo de análise de associação ampla com sanidade do casco utilizando CNV detectadas em amostras de 5.845 animais da raça Holandesa genotipados com plataformas de diferentes densidades. O total de 14 CNVR foram associadas a pelo menos uma dessas características: podo dermatite circunscrita (úlcera de sola), dermatite digital, dermatite interdigital, erosão do calcanhar, lesão de linha branca, hemorragia de sola e hiperplasia interdigital.

Szyda et al. (2019) identificaram CNV em grupos divergentes para susceptibilidade à mastite. Amostras de sequenciamento de 13 pares de vacas Holandesas meio irmãs foram utilizadas. Cada par era constituído por uma fêmea susceptível a mastite e outra que não possuía histórico desse tipo de afecção. Em comparação, 191 CNV em estado de deleção foram encontradas no grupo susceptível, mas estavam presentes no grupo saudável. Essas CNV compreendiam genes envolvidos na resposta imunológica (ex: *CELSR2*, *ZBTB10*). Esses autores sugerem que pode haver relação entre deleção dessa categoria de gene e susceptibilidade a mastite clínica.

As CNV podem ser utilizadas para estudos de genética de populações (Xu et al., 2016). Algumas CNV podem surgir de forma independente nas raças bovinas de forma a contribuir para as diferenças raciais, estando, assim, associadas à domesticação, formação racial (Liu et al., 2010a; Liu e Bickhart, 2012) e adaptação. Bickhart et al. (2012), detectaram 1.265 CNVR no genoma bovino de seis indivíduos (cinco taurinos e um zebuíno). Nesse estudo, genes relacionados a resistência à patógenos e à parasitas (*CATHL4* e *ULBP17*) estavam presentes em maior número

de cópias no indivíduo da raça Nelore em comparação com os bovinos taurinos, que apresentaram maior número de cópias para os genes envolvidos no transporte e metabolismo de lipídeos. Isso pode evidenciar maior adaptabilidade e rusticidade de zebuínos em relação a taurinos.

Aguiar et al. (2018) detectaram uma duplicação (BTA5:47.900.000–48.200.000) associada ao comprimento do umbigo ao sobreano na raça Nelore, localizada no terceiro íntron do gene *HMGA2*. Esses autores relataram que essa CNV é antiga, específica de animais zebuínos e que pode ter tido valor adaptativo nesta subespécie, pois também foi encontrada em zebuínos africanos. Hu et al. (2020) realizaram identificação de CNV baseadas na estatística F (F_{st}) de diferenciação populacional a partir de dados de sequência de 73 animais de dez raças diferentes (quatro taurinos e seis zebuínos). As CNV que representaram 1% top diferenciais sobrepuseram genes relacionados à estresse térmico (*DNAJC18*), processamento metabólico de lipídeos e ATP (*PLCXD3* e *MUSK*), e regulação da diferenciação muscular (*CTNNA1*, *MUSK*, *PKN2* e *ENSBTAG00000004415*). Todas se encontravam em estado de deleção nos indivíduos zebuínos, sugerindo que estas CNV são específicas dessa subespécie.

Estudos de detecção e caracterização de variantes no número de cópias são importantes para os programas de melhoramento genético de bovinos leiteiros, pois auxiliam na compreensão dos processos biológicos envolvidos na expressão fenotípica de características produtivas, reprodutivas e de sanidade. Entretanto, as técnicas de detecção ainda precisam ser aprimoradas, pois estas apresentam um “trade-off” (balanço entre duas características desejáveis, mas incompatíveis) entre altas taxas de descoberta e baixas taxas de falsos positivos (Butty et al., 2020). Em adição, o aprimoramento da anotação do genoma, a realização de estudos funcionais para caracterização de genes, QTL e RNA não codificantes e o abastecimento e curadoria das informações encontradas nos bancos de dados genômicos são fundamentais para a identificação de CNV e CNVR relacionadas a expressão de características de interesse para a cadeia produtiva leiteira.

3 REFERÊNCIAS BIBLIOGRÁFICAS

Abyzov A, Urban AE, Snyder M, Gerstein M (2011) CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. **Genome Research** 21:974–984.

Agro em Dia (2020). Embrapa: Pandemia desajusta comportamento do mercado de leite no Brasil. Disponível em: <https://agroemdia.com.br/2020/06/30/embrapa-pandemia-desajusta-comportamento-do-mercado-de-leite-no-brasil/>. Acesso em: 5 dez. 2021.

Aguiar TS et al. (2018) Association of Copy Number Variation at Intron 3 of HMGA2 With Navel Length in Bos indicus. **Frontier Genetics** 9:627.

Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. **Nature Reviews Genetics** 12:363–376.

Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE (2001) Segmental Duplications: Organization and Impact Within the Current Human Genome Project Assembly. **Genome Research** 11:1005–1017.

Beckmann JS, Kashi Y, Hallerman EM, Nave A, Soller M (1986) Restriction fragment length polymorphism among Israeli Holstein ‐ Friesian dairy bulls. **Animal Genetics** 17:25‐38.

Sassi NB, González-Recio Ó, de Paz-del Río R, Rodríguez-Ramilo ST, Fernández AI (2016) Associated effects of copy number variants on economically important traits in Spanish Holstein dairy cattle. **Journal of Dairy Science** 99:6371–6380.

Berry DP, McHugh N, Wall E, McDermott K, O'brien AC (2019) Low-density genotype panel for both parentage verification and discovery in a multi-breed sheep population. **Irish Journal of Agricultural and Food Research** 58:1–12.

Bickhart DM et al. (2012) Copy number variation of individual cattle genomes using next-generation sequencing. **Genome Research** 22:778–790 .

Bickhart DM, Liu GE (2014) The challenges and importance of structural variation detection in livestock **Frontier Genetics** 5:37.

Böhne A, Brunet F, Galiana-Arnoux D, Schultheis C, Volff JN (2008) Transposable elements as drivers of genomic and biological diversity in vertebrates. **Chromosome Research** 16:203–215.

Butty AM et al. (2021) Genome-wide association study between copy number variants and hoof health traits in Holstein dairy cattle. **Journal of Dairy Science** 104:8050–8061.

Butty AM, Chud TCS (2020) High confidence copy number variants identified in Holstein dairy cattle from whole genome sequence and genotype array data.**Scientifc Reports** 10:1–13.

Cassese A, Guindani M, Tadesse MG, Falciani F, Vannucci M (2014) A hierarchical bayesian model for inference of copy number variants and their association to gene expression. **Annals of Applied Statistics** 8:148–175.

Cayuela H, Dorant Y, Mérot C, Laporte M, Normandeau E, Gagnon-Harvey S, Clément M, Sirois P, Bernatchez L (2021) Thermal adaptation rather than demographic history drives genetic structure inferred by copy number variants in a marine fish. **Molecular Ecology** 30:1624–1641.

Chardon P, Vaiman M, Kirszenbaum M, Geffrotin C, Renard C, Cohen D (1985) Restriction fragment length polymorphism of the major histocompatibility complex of the pig. **Immunogenetics** 21:161–171.

Chen K et al. (2009) BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. **Nature Methods** 6:677–681.

Chen L, Chamberlain AJ, Reich CM, Daetwyler HD, Hayes BJ (2017) Detection and validation of structural variations in bovine whole-genome sequence data. **Genetics Selection Evolution** 49:1-13.

Chud TCS (2018) **Identificação de regiões com variações no número de cópias dos segmentos de DNA em bovinos de leite**. 127f. Tese (Doutorado em Genética e Melhoramento Animal) – Unesp, Jaboticabal.

Cole JB, Silva MVGB (2016) Genomic selection in multi-breed dairy cattle populations. **Revista Brasileira de Zootecnia** 45:195–202.

Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, Bassett AS, Seller A, Holmes CC, Ragoussis J (2007) QuantiSNP: An objective bayes hidden-markov model to detect and accurately map copy number variation using SNP genotyping data. **Nucleic Acids Research** 35: 2013–2025.

Collins FS, Brooks LD, Chakravarti A (1998) A DNA polymorphism discovery resource for research on human genetic variation. **Genome Research** 8:1229–1231.

Diskin SJ, Li M, Hou C, Yang S, Glessner J, Hakonarson H, Bucan M, Maris JM, Wang K (2008) Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. **Nucleic Acids Research** 36:e126.

Eichler EE (2001) Recent duplication, domain accretion and the dynamic mutation of the human genome. **Trends in Genetics** 17:661–669.

Food and Agriculture Organization of the United Nations (FAO) (2020) **DAIRY MARKET REVIEW: Emerging trends and outlook**. Rome, 10p.

Feng Z, Li X, Cheng J, Jiang R, Huang R, Wang D, Huang Y, Pi L, Hu L, Chen H (2020) Copy number variation of the pigy gene in sheep and its association analysis with growth traits. **Animals** 10:688.

Fernandes AR et al. (2020) **Programa Nacional de Melhoramento do Gir Leiteiro - 11a prova de pré-seleção de touros - touros pré-selecionados por meio de avaliação genômica**. Juiz de Fora: Embrapa Gado de Leite, 22p. (Embrapa Gado de Leite. Documentos, 245).

Feuk L (2010) Inversion variants in the human genome: Role in disease and genome architecture. **Genome Medicine** 2:1–8.

Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. **Nature Reviews Genetics** 7:85-97.

Gamazon ER, Stranger BE (2015) The impact of human copy number variation on gene expression. **Briefings in Functional Genomics** 14:352–357.

García-Ruiz A, Cole JB, VanRaden PM, Wiggans GR, Ruiz-López FJ, Van Tassell CP (2016) Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. **Proceedings of the National Academy of Sciences of the United States of America** 113:E3995–E4004.

Georges M, Lequarré AS, Hanset R, Vassart G (1987) Genetic variation of the bovine thyroglobulin gene studied at the DNA level. **Animal Genetics** 18:41–50.

Gillet-Markowska A, Richard H, Fischer G, Lafontaine I (2015) Ulysses: Accurate detection of low-frequency structural variations in large insert-size sequencing libraries. **Bioinformatics** 31:801–808.

Gong T, Hayes VM, Chan EKF (2021) Detection of somatic structural variants from short-read next-generation sequencing data. **Briefings in bioinformatics** 22:bbaa056.

Hou Y et al. (2011) Genomic characteristics of cattle copy number variations. **BMC Genomics** 12:1-11.

Hu Y et al. (2020) Comparative analyses of copy number variations between Bos taurus and Bos indicus. **BMC Genomics** 21:1-11.

Inoue K, Lupski JR (2002) Molecular mechanisms for genomic disorders. **Annual Review of Genomics and Human Genetics** 3:199–242.

Jiang Y, Wang Y, Brudno M (2012) PRISM: Pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. **Bioinformatics** 28:2576–2583.

Keel BN, Keele JW, Snelling WM (2017) Genome-wide copy number variation in the bovine genome detected using low coverage sequence of popular beef breeds. **Animal Genetics** 48:141–150.

Keel BN, Lindholm-Perry AK, Snelling WM (2016) Evolutionary and functional features of copy number variation in the cattle genome. **Frontiers in Genetics** 7:207.

Khatri B, Kang S, Shouse S, Anthony N, Kuenzel W, Kong BC (2019) Copy number variation study in Japanese quail associated with stress related traits using whole genome re-sequencing data. **PLoS One** 14: e0214543.

Kim JH, Hu HJ, Yim SH, Bae JS, Kim SY, Chung YJ (2012) CNVRuler: A copy number variation-based case-control association analysis tool. **Bioinformatics** 28:1790–1792.

Kim PM, Lam HYK, Urban AE, Korbel JO, Affourtit J, Grubert F, Chen X, Weissman S, Snyder M, Gerstein MB (2008) Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history. **Genome Research** 18:1865–1874.

Klambauer G, Schwarzbauer K, Mayr A, Clevert DA, Mitterecker A, Bodenhofer U, Hochreiter S (2012) CnMOPS: Mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. **Nucleic Acids Research** 40:e69-e69.

Korbel JO, Abyzov A, Mu XJ, Carriero N, Cayting P, Zhang Z, Snyder M, Gerstein MB (2009) PEMer: A computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. **Genome Biology** 10:1-14.

Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y (2019) Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. **Genome Biology** 20:1–18.

Layer RM, Chiang C, Quinlan AR, Hall IM (2014) LUMPY: A probabilistic framework for structural variant discovery. **Genome biology** 15:1-19.

Lee JA, Carvalho CMB, Lupski JR (2007) A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders. **Cell** 131:1235–1247.

Lee YL, Bosse M, Mullaart E, Groenen MAM, Veerkamp RF, Bouwman AC (2020) Functional and population genetic features of copy number variations in two dairy cattle populations. **BMC Genomics** 21:1-15.

Letaief R et al. (2017) Identification of copy number variation in French dairy and beef breeds using next-generation sequencing. **Genetics Selection Evolution** 49:1–15.

Lieber MR (2008) The mechanism of human nonhomologous DNA End joining. **Journal of Biological Chemistry** 283:1–5.

Liu GE, Bickhart DM (2012) Copy number variation in the cattle genome. **unctional & integrative genomics** 12:609–624.

Liu GE, Hou Y (2010a) Analysis of copy number variations among diverse cattle breeds **Genome Research** 20:693–703.

Liu Z, Li A, Schulz V, Chen M, Tuck D (2010b) MixHMM: Inferring copy number variation and allelic imbalance using SNP arrays and tumor samples mixed with stromal cells. **PLoS One** 5:e10909.

Medvedev P, Stanciu M, Brudno M (2009) Computational methods for discovering structural variation with next-generation sequencing. **Nature Methods** 6:S13–S20.

Miller CA, Hampton O, Coarfa C, Milosavljevic A (2011) ReadDepth: A parallel R package for detecting copy number alterations from short sequencing reads. **PLoS One** 6:e16327.

Nguyen HT, Boocock J, Merriman TR, Black MA (2016) SRBreak: A read-depth and split-read framework to identify breakpoints of different events inside simple copy-number variable regions. **Frontiers in Genetics** 7:160.

O'Brien AMP, Höller D (2015) Low levels of taurine introgression in the current Brazilian Nelore and Gir indicine cattle populations. **Genetics Selection Evolution** 47:1-7.

OECD/FAO (2020) Dairy and dairy products. In.: **OECD-FAO Agricultural Outlook 2020-2029.** Paris: OECD Publishing, p 178.

Panetto JC do C et al. (2020) **Programa Nacional de Melhoramento do Gir Leiteiro Sumário Brasileiro de Touros 3ª Avaliação Genômica de Touros**

Resultado do Teste de Progênie - Maio 2020. Juiz de Fora: Embrapa Gado de Leite, 104p. (Embrapa Gado de Leite. Documentos, 244).

Pique-Regi R, Monso-Varona J, Ortega A, Seeger RC, Triche TJ, Asgharzadeh S (2008) Sparse representation and Bayesian detection of genome copy number alterations from microarray data. **Bioinformatics** 24:309–318.

Pirooznia M, Goes F, Zandi PP (2015) Whole-genome CNV analysis: Advances in computational approaches. **Frontiers in Genetics** 6:138.

Prata MA, Faro LE, Moreira HL, Verneque RS, Vercesi Filho AE, Peixoto MGCD, Cardoso VL (2015) Genetic parameters for milk production traits and breeding goals for Gir dairy cattle in Brazil. **Genetics and Molecular Research** 14:12585–12594.

Rafter P, Gormley IC, Parnell AC, Kearney JF, Berry DP (2020) Concordance rate between copy number variants detected using either high- or medium-density single nucleotide polymorphism genotype panels and the potential of imputing copy number variants from flanking high density single nucleotide polymorphism haplotypes in cattle. **BMC Genomics** 21:1-10.

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO (2012) DELLY: Structural variant discovery by integrated paired-end and split-read analysis. **Bioinformatics** 28:i333–i339.

Redon R, Ishikawa S (2006) Global variation in copy number in the human genome. **Nature** 444:444–454.

Rodriguez-Revenga L, Mila M, Rosenberg C, Lamb A, Lee C (2007) Structural variation in the human genome: the impact of copy number variants on clinical diagnosis. **Genetics in Medicine** 9:600–606.

Rosen BD, Bickhart DM (2020) De novo assembly of the cattle reference genome with single-molecule sequencing. **Gigascience** 9:1–9.

Rubin CJ, Megens HJ (2012) Strong signatures of selection in the domestic pig genome. **Proceedings of the National Academy of Sciences of the United States of America** 109:19529–19536.

Santana ML, Pereira RJ, Bignardi AB, El Faro L, Tonhati H, Albuquerque LG (2014) History, structure, and genetic diversity of Brazilian Gir cattle. **Livestock Science** 163:26–33.

Sasaki S, Miki Y, Ibi T, Wakaguri H, Yoshida Y, Sugimoto Y, Suzuki Y (2021) A 44-kb deleted-type copy number variation is associated with decreasing complement component activity and calf mortality in Japanese Black cattle. **BMC Genomics** 22:1-10.

Sharp AJ, Locke DP (2005) Segmental duplications and copy-number variation in the human genome. **American Journal of Human Genetics** 77:78–88.

Shaw CJ, Lupski JR (2004) Implications of human genome architecture for rearrangement-based disorders: The genomic basis of disease. **Human Molecular Genetics** 13:R57–R64.

Sun W, Wright FA, Tang Z, Nordgard SH, Van Loo P, Yu T, Kristensen VN, Perou CM (2009) Integrated study of copy number states and genotype calls using high-density SNP arrays. **Nucleic Acids Research** 37:5365–5377.

Szyda J, Mielczarek M, Fraęszczak M, Minozzi G, Williams JL, Wojdak-Maksymiec K (2019) The genetic background of clinical mastitis in Holstein-Friesian cattle. **Animal** 13:2156–2163.

Trappe K, Emde AK, Ehrlich HC, Reinert K (2014) Gustaf: Detecting and correctly classifying SVs in the NGS twilight zone. **Bioinformatics** 30:3484–3490.

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SFA, Hakonarson H, Bucan M (2007) PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. **Genome Research** 17:1665–1674.

Watson JD, Crick FHC (1953) Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. **Nature** 171:737–738.

Xie C, Tammi MT (2009) CNV-seq, a new method to detect copy number variation using high-throughput sequencing. **BMC Bioinformatics** 10:1-9

Xu L, Cole JB, Bickhart DM, Hou Y, Song J, VanRaden PM, Sonstegard TS, Van Tassell CP, Liu GE (2014) Genome wide CNV analysis reveals additional variants associated with milk production traits in Holsteins. **BMC Genomics** 15:1-10.

Xu L, Hou Y, Bickhart DM, Zhou Y, Hay EHA, Song J, Sonstegard TS, Van Tassell CP, Liu GE (2016) Population-genetic properties of differentiated copy number variations in cattle. **Scientific reports** 6:1-8.

Yang L et al. (2021) Genomic sequencing analysis reveals copy number variations and their associations with economically important traits in beef cattle. **Genomics** 113:812–820.

Yoon S, Xuan Z, Makarov, V, Ye, K, Sebat, J (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. **Genome Research** 19:1586–1592.

Zarate S, Carroll A, Mahmoud M, Krasheninina O, Jun G, Salerno WJ, Schatz MC, Boerwinkle E, Gibbs RA, Sedlazeck FJ (2020) Parliament2: Accurate structural variant calling at scale. **Gigascience** 9:1–9.

Zhan B, Fadista J, Thomsen B, Hedegaard J, Panitz F, Bendixen C (2011) Global assessment of genomic variation in cattle by genome resequencing and high-throughput genotyping. **BMC Genomics** 12:1–20.

Zhang F, Gu W, Hurles ME, Lupski JR (2009) Copy number variation in human health, disease, and evolution. **Annual Review of Genomics and Human Genetics** 10:451–481.

Zhang J, Wang J, Wu Y (2012) An improved approach for accurate and efficient calling of structural variations with low-coverage sequence data. **BMC Bioinformatics** 13:1-11.

Zheng X, Zhao P, Yang K, Ning C, Wang H, Zhou L, Liu J (2020) CNV analysis of Meishan pig by next-generation sequencing and effects of AHR gene CNV on pig reproductive traits. **Journal of Animal Science and Biotechnology** 11:1-11.

Zhou Y, Connor EE, Wiggans GR, Lu Y, Tempelman RJ, Schroeder SG, Chen H, Liu GE (2018) Genome-wide copy number variant analysis reveals variants associated with 10 diverse production traits in Holstein cattle. **BMC Genomics** 19:1-9.

Zhou Y et al. (2016) Genome-wide CNV analysis reveals variants associated with growth traits in Bos indicus. **BMC Genomics** 17:1-9.

Zimin AV et al. (2009) A whole-genome assembly of the domestic cow, *Bos taurus*. **Genome Biology** 10:1-10.

CAPÍTULO 02 – IDENTIFICAÇÃO DE VARIAÇÕES ESTRUTURAIS DO GENOMA DE BOVINOS GIR LEITEIRO

RESUMO - O Gir Leiteiro é um bovino zebuíno que possui tolerância ao calor, às doenças e aos parasitas tropicais. Essas características adaptativas tornam o Gir Leiteiro um recurso genético importante para a produção de leite nos trópicos. Por isso, o estudo e caracterização de variantes presentes no DNA desses indivíduos é de interesse para o melhoramento genético da raça Gir. Os objetivos deste trabalho foram: (1) detectar variações no número de cópias (CNV) em bovinos Gir Leiteiro; (2) definir CNVR de alta confiança por meio de dois métodos in silico; (3) determinar as regiões genômicas em que ocorrem as CNVR de alta confiança que coincidem com genes e "Quantitative trait loci" (QTL) previamente relacionados às características de interesse para a cadeia produtiva de leite. Para a detecção de CNV, foram utilizados, após o controle de qualidade, 38 animais sequenciados e amostras de 545 indivíduos genotipados com o painel Illumina BovineHD BeadChip, totalizando 547 animais Gir Leiteiro, dos quais 36 possuíam informação de painel de SNP e sequenciamento. Para aumentar a acurácia nas variantes estruturais detectadas, dois conjuntos de regiões de CNV (CNVR) de alta confiança foram definidos, contendo variantes encontradas nos dados de painéis de SNP e de sequenciamento. Esses conjuntos são relativos à população (n = 545) (CNVR_POP) e aos indivíduos representativos da população Gir Leiteiro (n = 36) (CNVR_ANI). No conjunto CNVR_POP, foram encontradas dez CNVR, representando 1,05 Mb do genoma bovino. No conjunto CNVR_ANI, foram encontradas 45 CNVR, cobrindo 4,4 Mb do genoma bovino. Os dois conjuntos de CNVR de alta confiança foram unidos para análise funcional resultando em 48 CNVR únicas e de alta confiança. Essas se sobrepuseram a 69 genes, incluindo os genes 69 genes, incluindo os genes *FILIP1*, *SENP6*, *CA5A*, *BANP*, *HERC2*, *RHOU*, *GBP2*, *GBP4*, *GBP6*, *BLA-DQB*, *ENSBTAG00000037605* e de receptores olfativos e receptores olfativos. O total de 44 "Quantitative trait loci" (QTL) foram significativamente associados a características de produção, reprodução, conformação, sanidade, leite e carne e carcaça foram encontrados nas CNVR únicas e de alta confiança. Nossos resultados abrangem a identificação e caracterização de 48 CNVR de alta confiança no genoma de bovinos Gir Leiteiro, o que pode contribuir para a elaboração de um mapa de SV na raça e para o melhor entendimento do genoma dos zebuínos.

Palavras-chave: Deleções, duplicações, variação no número de cópias, zebuínos.

1 INTRODUÇÃO

O Gir Leiteiro representa um recurso genético importante para a produção de leite nos trópicos (Santana et al., 2014). Os animais Gir Leiteiro possuem tolerância ao calor, às doenças e aos parasitas tropicais (Panetto et al., 2020). Santana et al. (2014) ainda indicaram que, em virtude das constantes mudanças climáticas, o Gir pode ganhar importância fora dos trópicos, mesmo em cruzamentos com animais taurinos. Variantes de DNA são utilizadas no Gir Leiteiro para auxiliar na identificação de reprodutores geneticamente superiores, análise de doenças hereditárias e na seleção genética de touros e vacas.

Dentre as variações que ocorrem no genoma, as variações no número de cópias (CNV) envolvem deleções e duplicações maiores que, em geral, 50 pares de base (bp) entre dois indivíduos de uma espécie (Mills et al., 2011). As CNV podem contribuir funcionalmente para o processo de domesticação e formação racial em bovinos (Liu et al., 2010; Liu and Bickhart, 2012), diferenciação entre subespécies, por exemplo, animais zebuínos e taurinos (Aguiar et al., 2018; Hu et al., 2020) e podem conferir vantagem adaptativa a indivíduos (Bickhart et al., 2012; Aguiar et al., 2018). Em estudos prévios com bovinos, CNV e regiões de CNV (CNVR) foram relacionadas a produção leiteira (Xu et al., 2014) e consumo alimentar residual (Hou et al., 2012a) na raça holandesa, estatura em animais de raças chinesas (Cao et al., 2018), comprimento de umbigo em zebuínos (Aguiar et al., 2018), mortalidade de bezerros em bovinos Wagyu (Sasaki et al., 2021), características de sanidade do casco (Butty et al., 2021). As CNVR são formadas pelo agrupamento de CNV sobrepostas (Redon et al., 2006).

A partir da informação genômica obtida por sequenciamento do genoma completo ou por painéis de genotipagem de polimorfismo de nucleotídeo único (SNP) diferentes CNV podem ser identificadas, que variam quanto ao número, ao comprimento e à distribuição no genoma (Zhan et al., 2011; Butty et al., 2020). A detecção de CNV a partir de painéis de SNP é baseada principalmente em duas medidas: o Log da razão de R (LRR - “Log R Ratio”) e a frequência do alelo B (BAF - “B allele frequency”) que são oriundos do processo de genotipagem (Wang et al., 2007). Em dados de sequenciamento do genoma completo, SV são preditas a partir de padrões anormais de alinhamento sugestivos de pontos de quebra de rearranjo

genômico. As principais metodologias são montagem, “read-pair”, “read depth” e “split-read”, (Zhao et al., 2013; Pirooznia et al., 2015). Apesar das metodologias que utilizam dados de sequenciamento serem mais precisas e acuradas, essas também apresentam limitações. Diante disso, a combinação de diferentes métodos de detecção e técnicas moleculares, como sequenciamento e painéis de SNP, pode ser uma alternativa para contornar as dificuldades na detecção de CNV e, assim, diminuir o número de falsos positivos e aumentar a probabilidade de detecção de CNV confiáveis.

Apesar das CNVR compreenderem entre 2 a 7% do genoma bovino (Keel et al., 2016), a seleção genômica nessa espécie tem sido direcionada para a utilização de SNP e pequenas inserções ou deleções (“indels”) e pouca atenção foi destinada às variações maiores, como as CNV e outras SV (Couldrey et al., 2017). Sendo assim, a predição genômica integrando SNP e CNV pode oferecer novos conhecimentos para elucidar características complexas e para compreender a proporção da variação genética que não é explicada pelos SNP (“missing heritability”), mas que está presente e pode ser predita pela estimativa de herdabilidade (Hay et al., 2018). Os mesmos autores relataram ainda que a inclusão de CNV na predição genômica provocou um pequeno aumento da acurácia para algumas características em bovinos Nelore. Todavia, os autores ressaltam que a inclusão de genótipos de CNV na predição genômica poderá aumentar a acurácia das predições dos valores genéticos genômicos e promover ganhos genéticos adicionais em animais de interesse zootécnico (Hay et al., 2018).

O primeiro passo para inclusão de SV, tais como as CNV, nas predições genômicas e nos estudos de associação com fenótipos é a detecção e mapeamento desse tipo de variante genômica. Ademais, o estudo de CNV pode ser útil para entender mecanismos de adaptação, domesticação, desenvolvimento de doenças e expressão de fenótipos de interesse para a cadeia produtiva leiteira. Assim, os objetivos deste trabalho foram: (1) detectar variações no número de cópias (CNV) em bovinos Gir Leiteiro; (2) definir CNVR de alta confiança por meio de dois métodos *in silico*; (3) determinar as regiões genômicas em que ocorrem as CNVR de alta confiança que coincidem com genes e “Quantitative trait loci” (QTL) previamente relacionados às características de interesse para a cadeia produtiva de leite.

2 MATERIAL E MÉTODOS

2.1 Amostras, alinhamento e preparação de dados de sequenciamento

Touros do Programa Nacional de Melhoramento do Gir Leiteiro (PNMGL), conduzido em parceria pela Associação Brasileira de Criadores de Gir Leiteiro (ABCGIL) e Embrapa Gado de Leite, foram ordenados após estudo de desempenho e número de progênie no PNMGL. Os touros mais bem classificados, ou seja, representativos da população foram selecionados para sequenciamento do genoma completo. Para esse estudo, foram utilizadas amostras de 43 touros Gir Leiteiro. As amostras de um a 13 foram cedidas por meio de parceria com a Embrapa Gado de Leite (processos: Embrapa SEG 02.13.05.011.00.00; CNPq 310199/2015–8; MCTI/CNPq/INCT-Ciência Animal e FAPEMIG CVZ PPM-00606/16). As amostras 14 a 43 foram sequenciadas mediante auxílio financeiro concedido pelo CNPq (processo 431629/2016-1).

Amostras de sangue e sêmen foram utilizadas para a extração de DNA genômico. Para as amostras um a 13 a extração foi realizada utilizando DNeasy Blood & Tissue Kit (Qiagen, Valencia, CA, USA), conforme recomendações do fabricante. O DNA extraído foi quantificado e avaliado pelo método de espectrofotometria (NanoDrop 1000, Thermo Scientific, Wilmington, DE, USA). O sequenciamento do genoma dessas amostras foi realizado pelas Illumina HiSeq2000 (Illumina Inc., San Diego, CA, USA). As bibliotecas utilizadas para o sequenciamento foram do tipo ‘paired-end’, em que foram produzidas leituras com tamanho de 2 x 100 bp e 2 x 200 bp, com cobertura média de sequenciamento de 13,9X.

Para as amostras 14 a 43, a extração de DNA foi realizada com protocolo utilizando tampão salino e purificação com fenol/clorofórmio brevemente descrito por Machado et al. (2010). A determinação da qualidade e a normalização do material extraído foi avaliada por meio de fluorescência no Qubit fluorometer 2.0 (Life technologies, Grand Island, NY). Na preparação da biblioteca foi utilizado o Illumina TruSeq Nano kit (Illumina Inc., San Diego, CA, USA). O sequenciamento das amostras foi realizado na Illumina NovaSeq 6000 (Illumina Inc., San Diego, CA, USA). Leituras com tamanho de 2 x 150 bp foram produzidas, com cobertura média de sequenciamento de 16,7X por amostra. Em todas as 43 amostras, a construção

das bibliotecas foi conduzida de acordo com os protocolos recomendados pelo fabricante.

A qualidade dos dados de sequenciamento foi verificada pela ferramenta FastQC (v. 0.11.8) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Com base nos resultados do FastQC, foram aplicados aos dados os critérios de controle de qualidade das leituras por meio do programa SeqyClean (ZHBANNIKOV et al., 2017), conforme parâmetros recomendados pelo protocolo do "1000 Bull Genomes Project". Assim, foram removidas (1) leituras com três ou mais bases não identificadas (N) nas sequências; (2) leituras com média de qualidade para "phred score" inferior ou igual a 20 ou seja, a probabilidade média de que as bases estejam incorretas foi de, no mínimo, 0,01; (3) leituras com comprimento menor do que 50 bases nas sequências.

Além da remoção das leituras com baixa qualidade, foram removidas as sequências de adaptadores e possíveis contaminantes. Com a remoção dos adaptadores, busca-se evitar que estas sequências sejam incorporadas no alinhamento do genoma, gerando alinhamento errôneos.

Mediante recomendação de parâmetros do 1000 Bull Genomes Project (http://www.1000bullgenomes.com/), as sequências foram alinhadas ao genoma referência bovino ARS-UCD 1.2 (https://sites.ualberta.ca/~stothard/1000_bull_genomes/) por meio do algoritmo BWA opção *mem* (v. 0.7.15-r1144-dirty) (Li and Durbin, 2009). A conversão para o formato binário, ordenação e indexação foi realizada pelo Samtools (v. 1.8) (Li et al., 2009; Li, 2011), por meio das opções *view*, *sort* e *index*, respectivamente. As duplicatas ópticas e de PCR foram removidas pela opção *MarkDuplicates* do Picard Tools (v. 2.18.2-SNAPSHOT) (Picard toolkit, 2019).

A junção dos arquivos de amostras que foram sequenciadas em múltiplas "lanes" (um a 13) foi realizada pela opção *merge* do Samtools (v. 1.8) (Li et al., 2009; Li, 2011). A recalibração do escore de qualidade das bases foi realizada pelo *BaseRecalibrator* e *PrintReads* do Genome Analysis Toolkit (GATK, v. 3.8-1-0-gf15c1c3ef), resultando em arquivos com maior confiabilidade por base. Todas as etapas seguiram as recomendações de parâmetros do 1000 Bull Genomes Project (http://www.1000bullgenomes.com/) (Figura 1). O conjunto de variantes conhecidas

fornecido pelo consórcio do projeto 1.000 Bull Genomes foi usado para recalibração de qualidade de base. A opção *flagstat* do Samtools (v. 1.8) (Li et al., 2009; Li, 2011) e um "script" em linguagem perl em conjunto a opção *mpileup* do Samtools (v. 1.8) (Li et al., 2009; Li, 2011) foram utilizados para o cálculo de estatísticas do alinhamento e da cobertura de alinhamento no genoma, respectivamente.

Figura 1. Fluxograma do alinhamento e preparação das amostras de sequenciamento

2.2 Amostras de genotipagem

Os dados de painéis de SNP deste estudo foram cedidos pela Embrapa Gado de Leite, situada na cidade de Juiz de Fora, em Minas Gerais, Brasil e pertencem ao projeto "Seleção genômica de raças leiteiras no Brasil". Amostras de 566 animais Gir

Leiteiro foram genotipadas com o painel Illumina BovineHD BeadChip (Illumina Inc., San Diego, CA, USA) que consiste em 777.962 marcadores distribuídos ao longo do genoma, com distância média entre SNP igual a 3,43 quilobases (Kb) e mediana igual a 2,68 Kb. Para a detecção de CNV, marcadores SNP com escore “GenCall” abaixo de 0,15 foram removidos (Illumina, 2014).

Dos 566 animais genotipados, 36 também foram sequenciados. A análise de componentes principais (PCA) foi realizada utilizando a matriz de genótipos apenas dos animais genotipados. O objetivo dessa análise foi verificar se existe alguma estrutura de população entre os animais e avaliar a representatividade dos indivíduos que foram sequenciados, ou seja, se eles formaram uma amostra aleatória da população genotipada. O mapa de SNP utilizado foi baseado no genoma referência ARS-UCD1.2 (https://www.animalgenome.org/repository/cattle/UMC_bovine_coordinates/, último acesso em 03/03/2020), em que foram considerados somente SNP localizados nos cromossomos autossômicos e de posição conhecida na montagem ARS-UCD1.2 (720.731 marcadores SNP). A PCA foi realizada pelo programa PLINK (v.1.9) (SHAUN PURCELL; CHRISTOPHER CHANG; CHANG et al., 2015). O controle de qualidade dos genótipos para a PCA removeu SNP com “minor allele frequency” (--*maf*) menor que 5%, genótipos perdidos por *locus* acima de 10% (--*geno*) e genótipos faltantes por animal acima de 10% (--*mind*).

2.3 CNV identificadas a partir de dados de sequenciamento

2.3.1 CNVnator

O CNVnator (v. 0.4.1) (Abyzov et al., 2011) foi utilizado para a identificação de CNV. Este programa possui alta sensibilidade na detecção de CNV, principalmente para deleções, e baixa FDR (“False Discovery Rate”). O CNVnator utiliza o método Read Depth (RD) e, por isso, é menos sensível a detecção de duplicações, se comparado a detecção de deleções. Para o cálculo do sinal de RD, o algoritmo utilizado pelo CNVnator segmenta todo o genoma em janelas não sobrepostas de tamanho igual (“bins”) e utiliza a contagem de leituras mapeadas dentro de cada segmento como o sinal RD. Após a partição em segmentos consecutivos, há a correção para o conteúdo GC e então as CNV são identificadas (Abyzov et al.,

2011). RD é a metodologia mais comum na detecção de CNV, no entanto, é menos robusta quanto a precisão da definição dos pontos de quebra (“breakpoints”) de CNV (Zhao et al., 2013).

A detecção de CNV ocorreu apenas nos cromossomos autossômicos, em que foi utilizado “bin size” (tamanho de janela) de 250 bp e o sinal RD médio foi de 4,12, o que se adequa as recomendações de ABYZOV et al. (2011). Apenas foram considerados CNV maiores do que 1 Kb e menores do que 5 Mb (Hay et al., 2018), significativas ($p<0,05$) para o teste estatístico t, em que a hipótese nula é se a média de sinal de profundidade das leituras na região de CNV é a mesma da média de sinal de profundidade na amostra, e CNV com fração de leituras mapeadas com baixa qualidade menor do que 0,5 ($q_0<0,5$).

Com base na distribuição dos dados, as amostras que apresentaram valores muito discrepantes de eventos (acima 7.000 eventos) e sinal de RD abaixo do recomendado ou valores muito discrepantes de eventos (acima 7.000 eventos) e número desproporcional de duplicações e deleções foram avaliadas. Essas amostras tiveram sua qualidade de alinhamento avaliada pela opção *qc* do programa ALFRED (Rausch et al., 2019). Também foi considerado o efeito da inclusão ou exclusão dessas amostras no cálculo de correlação linear simples de Pearson entre cobertura de alinhamento no genoma e número de eventos (soma do número de deleções e duplicações) detectados e, a partir desse resultado, as amostras foram excluídas

2.3.2 DELLY

Com o intuito de aumentar a confiabilidade, a detecção de CNV também foi realizada pelo programa DELLY (v. 0.7.6) (Rausch et al., 2012), que utiliza a combinação das metodologias “read-pair” (RP) e “split-read” (SR). As CNV são detectadas pelo método RP que permite alta sensibilidade e secundariamente, pelo método SR que aumenta a especificidade da detecção. O algoritmo RP analisa bibliotecas de leituras em busca de pares de leitura mapeados de forma discordante. Em seguida, o método SR é utilizado para refinar a definição dos pontos de quebras das SV preditas pelo método RP (Rausch et al., 2012).

O programa foi utilizado porque permite a detecção de eventos de duplicação e deleção em todos os indivíduos de forma simultânea, dado que CNV identificadas em apenas um indivíduo (“singletons” CNV) são mais prováveis de serem falsos positivas em comparação com CNV identificadas em vários indivíduos (Redon et al., 2006). A detecção de deleções e duplicações ocorreu apenas nos cromossomos autossômicos. A opção de qualidade mínima de mapeamento (que é a probabilidade de que uma leitura esteja alinhada no lugar errado) (*-q*) foi utilizada com valor de 20, seguindo os critérios de Khan et al. (Khan et al., 2018). Após a detecção, recomendações de estudos anteriores foram adotadas e apenas foram consideradas CNV maiores do que 1 Kb e menores que 5 Mb (Hay et al., 2018) e CNV com suporte de mais de 4 pares de leitura (“paired-end support”) (Khan et al., 2018).

2.4 CNV identificadas a partir de painéis de SNP

A detecção de CNV a partir de painéis de SNP foi realizada por meio do PennCNV (v. 1.0.5) (Wang et al., 2007). Esse programa utiliza a aplicação das metodologias bayesianas do modelo oculto de Markov, em que são aplicadas duas medidas da intensidade do sinal de fluorescência emitido para cada SNP, o LRR e a BAF. Também são utilizadas a distância entre SNP vizinhos e a frequência do alelo B na população (PFB). O arquivo PFB foi criado a partir do valor de BAF de cada marcador em todas as amostras.

Para reduzir a taxa de resultados falsos positivos, os valores de LRR de cada SNP foram ajustados para a dispersão da intensidade do sinal (“genomic waves”) ao longo das regiões genômicas de acordo com o conteúdo de GC esperado no genoma bovino, considerando uma região de 500 Kb ao redor de cada SNP (Diskin et al., 2008). Foram mantidas CNV com: mais de dez SNP, valores de desvio-padrão do LRR menor do que 0,30, BAF “drift” menor do que 0,01 e fator de dispersão da intensidade do LRR (“waviness factor”) menor do que 0,05 e CNV maiores do que 1 Kb e menores do que 5 Mb (Hay et al., 2018).

2.5 CNVR de alta confiança

As CNVR identificadas a partir dos resultados de métodos moleculares diferentes podem ser consideradas como de alta confiança (Zhan et al., 2011). A fim

de prover resultados mais confiáveis, foram estabelecidos dois conjuntos de CNVR de alta confiança baseado no trabalho de Butty et al. (2020), em que foram definidos dois conjuntos de CNVR de alta confiança, um baseado nas CNV identificadas para os indivíduos mais representativos da população Gir Leiteiro (CNVR_ANI) e o outro na população estudada (CNVR_POP). As CNVR foram determinadas pelo agrupamento de CNV sobrepostas por, no mínimo, 1 bp. Isso foi realizado pela opção *merge* do programa Bedtools (Quinlan and Hall, 2010).

Para a construção do conjunto CNVR_POP, foram utilizadas as CNVR detectadas a partir de dados de sequenciamento (CNVR_SEQ) e de painéis de SNP (CNVR_GEN). O conjunto de CNVR detectadas por meio de dados de painéis de SNP (CNVR_GEN) foi formado pelo agrupamento, com critério de sobreposição de, no mínimo, 1 bp, entre as CNV detectadas por meio da genotipagem. Como na detecção de CNV a partir de dados de HTS foram utilizadas mais de uma metodologia, as CNV resultantes das diferentes metodologias foram agrupadas dentro do resultado de cada programa, seguindo os mesmos critérios de sobreposição de, no mínimo, 1 bp. Em seguida, foram selecionadas apenas CNVR reciprocamente sobrepostas entre os dois programas com critério de, no mínimo, 50% (CNVR_SEQ). Definidos os conjuntos de CNVR_GEN e CNVR_SEQ, foram selecionadas CNVR sobrepostas entre esses dois conjuntos, com critério de sobreposição recíproca mínima de 50%. A partir disso, foram escolhidas apenas as CNVR que estavam presentes em mais de 5% da população utilizada no estudo, finalmente formando o conjunto CNVR_POP (Figura 2). Os agrupamentos de CNV com critério de sobreposição de, no mínimo, 1 bp foi realizado pela opção “merge” do programa Bedtools (Quinlan and Hall, 2010). A verificação de sobreposição recíproca mínima de 50% entre CNVR foi realizada por meio da opção Intersect do Bedtools (Quinlan and Hall, 2010).

Para o estabelecimento do conjunto CNVR_ANI, foram utilizados apenas os 36 indivíduos que foram sequenciados e genotipados. Para cada um desses indivíduos, foram encontradas CNV detectadas em comum a partir de dados de painéis de SNP e de sequenciamento, sendo utilizado o mesmo critério de sobreposição recíproca mínima de 50%. Apenas os resultados dos programas PennCNV e CNVnator foram utilizados nessa etapa, pois a detecção de CNV foi

realizada por amostra. Ao final, estas CNVR foram sobrepostas com critério mínimo de 1 bp, formando o conjunto CNVR_ANI (Figura 2).

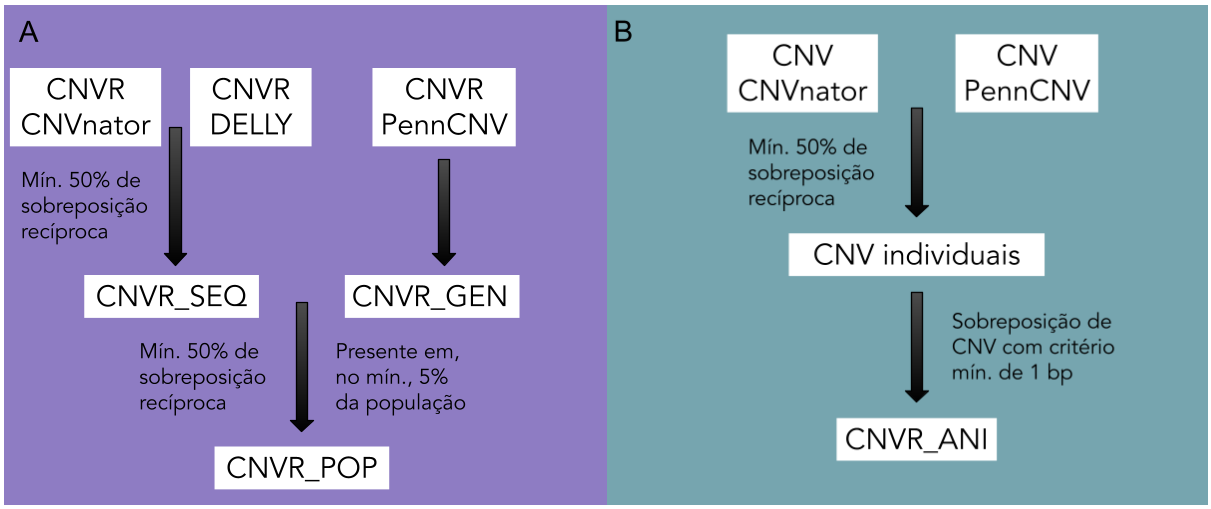


Figura 2. Fluxograma da construção dos conjuntos de regiões de variações no número de cópias (CNVR) de alta confiança. A) Conjunto CNVR_POP. B) Conjunto CNVR_ANI

2.6 Análise Funcional

Genes e QTL foram recuperados do banco de dados Ensembl Genes (Ensembl Release 104, acesso em 11/05/2021) (https://www.ensembl.org/) e do Animal Genome database (acesso em 11/05/2021) (https://www.animalgenome.org/cgi-bin/QTLdb/BT/index), respectivamente. O pacote GALLO (Fonseca et al., 2020) do programa R (R Core Team, 2021) foi utilizado para identificar genes e QTL localizados nas mesmas regiões genômicas que as CNVR únicas e de alta confiança. A partir disso, termos do banco de dados Gene Ontology (GO) e vias biológicas preditas pelo banco de dados Kyoto Encyclopedia of Genes and Genomes (KEGG) (https://www.genome.jp/kegg/) foram enriquecidos (FDR<0,05) por meio do pacote WebGestaltR (Wang and Liao, 2020) do programa R (R Core Team, 2021). Para ambas as análises, o enriquecimento foi realizado pelo teste hipergeométrico ORA ("Over-Representation Analysis"). Os termos biológicos do Gene Ontology são divididos em três grupos: Componentes Celulares, Processos Biológicos e Funções Moleculares.

Termos da plataforma Medical Subject Headings (MeSH) (https://www.ncbi.nlm.nih.gov/mesh) foram utilizados para análises de enriquecimento de conjunto de genes (p-ajustado<0,05) por meio dos pacotes

meshes (Yu, 2018) do programa R (R Core Team, 2021), em que foi utilizada a opção de banco de dados “gene2pubmed”. Os termos MeSH utilizados foram Anatomia (A), Doenças (C), Drogas e Químicos (D) e Fenômenos e Processos (G). Informações sobre os genes anotados foram obtidas no RefSeq Genes (https://www.ncbi.nlm.nih.gov/refseq/rsg/) e GeneCards (https://www.genecards.org/).

3 RESULTADOS

3.1 Alinhamento e pré-processamento de dados de sequenciamento

Após a remoção de duplicatas ópticas e de PCR, o sequenciamento “paired-end” produziu 15.183.484.455 leituras, em que a média do número total de leituras por amostra foi de 353.104.290 (variando de 245.377.907 a 486.209.902, com mediana igual a 356.956.710 e desvio padrão - SD - equivalente a 48.976.689), das quais, em média 93,64% (variando de 72,94% a 97,98%, com mediana igual a 95,76% e SD equivalente a 5,69%) foram corretamente mapeadas no genoma referência ARS-UCD1.2. A média da cobertura de alinhamento no genoma foi de 16,06X (variando de 10,2X a 25X com mediana igual a 15,8X e SD equivalente a 2.99X) (Tabela 1).

Tabela 2. Número total de leituras (bp), porcentagem de leituras mapeadas (%), porcentagem de leituras corretamente mapeadas (%) e cobertura de alinhamento no genoma (X) por amostra após remoção de duplicatas

Amostra*	Número total de leituras (bp)	Leituras mapeadas (%)	Leituras corretamente mapeadas (%)	Cobertura de alinhamento no genoma (X)
1 (741)	356.625.327	99,72	87,27	13,6
2 (824)	362.701.796	99,79	97,82	15,0
3 (837)	385.353.621	99,81	97,48	15,9
4 (879)	374.801.680	99,80	97,48	15,6
5 (861)	386.325.249	99,85	97,54	16,0
6 (887)	368.701.320	99,77	97,97	15,2
7 (894)	306.516.816	99,79	97,56	12,6
8 (920)	345.575.226	99,79	97,04	14,0

9 (958)	245.377.907	99,75	97,49	10,2
10 (1024)	394.641.626	99,80	97,70	16,1
11 (1037)	395.470.276	99,83	97,96	16,2
12 (1437)	260.192.208	99,64	96,74	10,3
13 (851)	343.774.440	99,87	97,79	14,8
14 (1200)	364.206.965	99,79	93,89	17,7
15 (1295)	452.832.085	99,81	95,98	21,5
16 (1438)	367.083.050	99,72	84,38	15,6
17 (1439)	325.143.647	99,79	86,33	14,1
18 (1440)	366.598.393	99,87	95,76	17,7
19 (1442)	307.936.476	97,79	95,76	15,2
20 (1447)	359.056.254	99,70	90,39	16,7
21 (702)	486.209.902	99,78	97,17	25,0
22 (707)	305.000.681	99,66	94,08	14,8
23 (713)	380.210.484	99,78	96,15	18,8
24 (714)	335.406.862	99,71	92,29	16,2
25 (752)	340.249.508	99,79	95,49	17,0
26 (753)	372.228.475	99,64	94,32	17,1
27 (756)	405.614.167	99,87	96,80	20,1
28 (789)	404.099.122	99,72	97,80	20,9
29 (791)	312.239.045	99,63	88,37	13,9
30 (805)	328.843.430	97,56	95,54	16,7
31 (823)	365.787.842	99,79	97,30	18,8
32 (846)	323.707.726	96,59	89,11	15,0
33 (863)	401.851.858	99,81	97,98	20,3
34 (870)	276.449.451	99,80	96,56	13,9
35 (890)	321.650.759	99,85	94,35	15,8
36 (893)	273.119.499	99,79	95,04	13,6
37 (906)	395.116.189	99,78	95,81	19,3
38 (907)	434.008.561	99,88	91,13	20,3
39 (1258)	310.223.610	99,79	92,94	15,1
40 (1501)	313.768.952	99,74	75,59	11,6
41 (721)	356.956.710	99,81	72,94	12,3

42 (754)	346.287.110	99,87	85,93	13,9
43 (831)	325.540.150	99,70	93,32	16,0

*Número entre parênteses corresponde a identificação original das amostras.

3.2 CNV identificadas a partir de dados de sequenciamento

3.2.1 CNVnator

Após a avaliação dos resultados, optou-se por retirar das análises as amostras de cinco indivíduos (39, 40, 41, 42, 43) (Apêndice B) que apresentaram número muito alto de eventos e baixo sinal de RD ou número discrepante de duplicações com maior proporção de duplicações do que deleções, dado que é esperado maior número de deleções do que duplicações. Devido ao método, o programa pode confundir “missmapping” em regiões de repetição com duplicações (Abyzov et al., 2011). Sugere-se que o resultado da detecção de variantes nessas amostras continha grande número de falsos positivos e isso pode ocorrer em virtude da qualidade e/ou da origem da amostra utilizada para extração do DNA, de fatores laboratoriais que podem interferir na qualidade do sequenciamento, como por exemplo a qualidade do DNA extraído e a amplificação excessiva na etapa de montagem das bibliotecas.

Os resultados da avaliação de qualidade de alinhamento pela opção *qc* programa Alfred (Rausch et al., 2019) para as amostras que apresentaram número muito alto de eventos estão descritos no Apêndice C. A correlação linear simples de Pearson entre o número de CNV e a cobertura de alinhamento no genoma, quando foram incluídas todas as amostras não foi significativa (-0,18, p=0,23). Entretanto, ao retirar as amostras heterogêneas, obteve-se correlação positiva e significativa (0,34, p=0,04), sendo esse resultado esperado devido a metodologia de detecção (“read depth”). Diante disso, optamos por retirar essas amostras (39, 40, 41, 42, 43) de todas as análises de detecção de CNV.

Após a filtragem e controle de qualidade, para as 38 amostras restantes, com cobertura média de alinhamento ao genoma de 16,35X, foram detectadas, em média, 2.143 CNV por animal (variando de 1.554 a 3.844, com mediana igual a 1.940 e SD equivalente a 564,93). O total de CNV foi de 81.447, sendo 53.876 deleções e 27.571 duplicações. O tamanho médio das CNV foi de 17.239 bp

(variando de 1.249 bp a 1.791.499 bp, com mediana igual a 7.999 bp e SD equivalente a 42.662,99 bp).

3.2.2 DELLY

Como a detecção de deleções e duplicações ocorreu de forma simultânea para todas as amostras, foram detectadas CNV em mais de um indivíduo (CNV populacionais) e CNV detectadas em apenas um indivíduo (“singletons” CNV). A detecção múltipla de 38 indivíduos gerou o total de 20.888 variantes (20.351 CNV populacionais e 537 “singletons” CNV). Foram detectadas 14.571 deleções (14.186 CNV populacionais e 385 “singletons”) e 6.317 duplicações (6.165 populacionais e 152 “singletons”). O tamanho médio das CNV foi de 179.007 bp (variando de 1.000 bp a 4.983.990 bp com mediana igual a 11.518 bp e SD equivalente a 551.161,4 bp).

3.3 Amostras de genotipagem

Após o controle de qualidade das destinado à análise de detecção de CNV, em que foram removidos SNP com valor de escore “GenCall” abaixo de 0,15, o valor médio de SNP por animal foi de 770.125 (variando de 666.135 a 774.163, com mediana igual a 772.024 e SD equivalente a 10.090,94).

Após o controle de qualidade da PCA, restaram 433.015 SNP e cinco animais foram removidos. Não foi observada estratificação na população. Os animais que também foram sequenciados estão distribuídos de forma aleatória no gráfico bidimensional, representando a diversidade de distâncias genéticas dentro da população que foi genotipada (Figura 3).

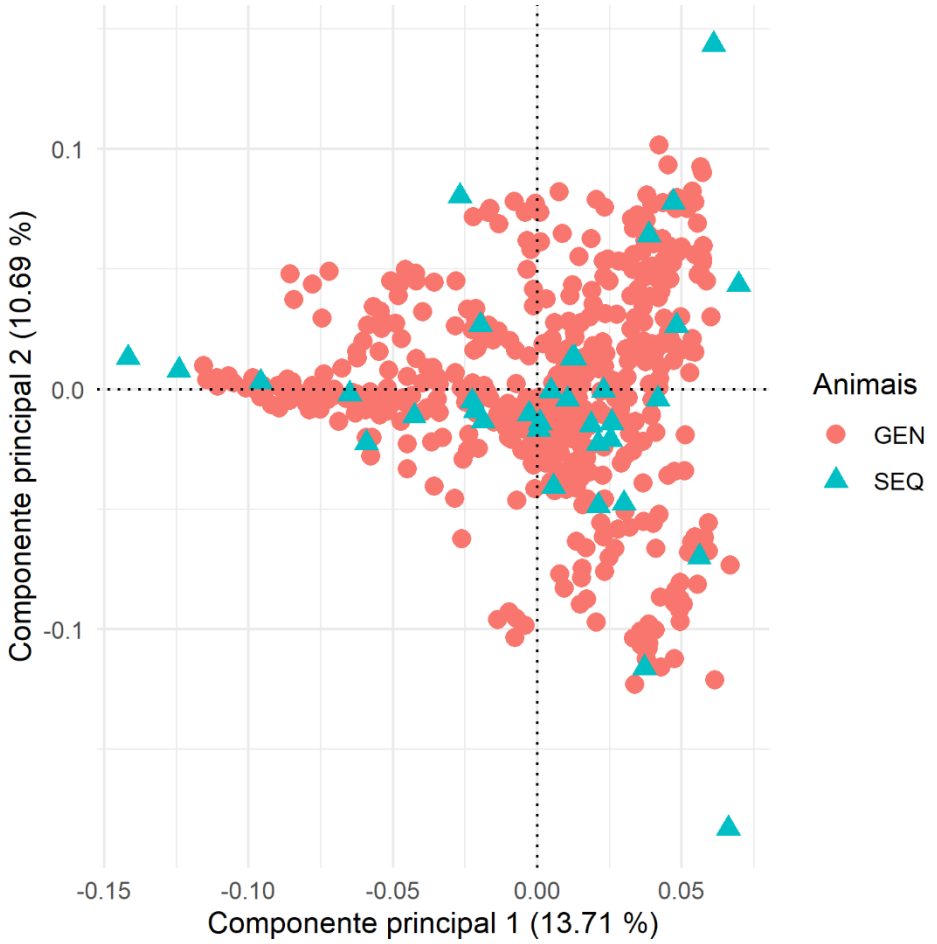


Figura 3. Análise de componentes principais dos animais genotipados com painel Illumina BovineHD BeadChip. Nesta figura estão representados os animais que foram apenas genotipados (GEN) e os que também foram sequenciados (SEQ)

3.4 CNV identificadas a partir de painéis de SNP

O mapa de SNP utilizado foi baseado no genoma referência ARS-UCD1.2 (https://www.animalgenome.org/repository/cattle/UMC_bovine_coordinates/, último acesso em 03/03/2020), em que foram considerados somente SNP localizados nos cromossomos autossômicos e de posição conhecida na montagem ARS-UCD1.2 (720.731 marcadores SNP).

No mapa de SNP, 7,35% dos SNP foram removidos por não pertencerem a cromossomos autossômicos ou não possuírem posição conhecida no genoma referência ARS-UCD1.2 e, durante a análise, outros 9,46% SNP não foram utilizados devido ao seu baixo escore de “GenCall” na população. Uma amostra foi removida da análise de detecção de CNV, pois apresentou número muito discrepante de CNV,

com base em interpretação gráfica (Apêndice A), o que poderia refletir alto número de eventos falso positivos.

Após controle de qualidade, CNV foram identificadas em 545 animais, utilizando 652.560 marcadores SNP. O total de 4.162 CNV foram detectadas, com média de 7,6 CNV por animal (variando de um a 90, com mediana igual a sete e SD equivalente a sete), sendo 2.510 deleções e 1.650 duplicações. O número médio de marcadores SNP em cada CNV foi de 25,16 (variando de dez a 293, com mediana igual a 17 e SD equivalente a 19,26). O tamanho médio das CNV foi de 122.807 bp (variando de 10.180 bp a 1.371.933 bp, com mediana igual a 58.988 bp e SD equivalente a 120.392,8 bp).

3.5 CNVR de alta confiança

Nas CNVR oriundas de dados de genotipagem de SNP (CNVR_GEN), foram detectadas 489 CNVR, com média de tamanho 95.170 bp (variando de 10.714 bp a 1.410.517 bp, com mediana igual a 50.517 bp e SD equivalente a 127.725,6 bp), somando 46.538.246 bp, sendo 428 CNVR compostas apenas por eventos de deleção, 55 por eventos de duplicação e seis consideradas complexas, em que ocorreram ambos os eventos.

Dentre as CNV oriundas de dados de sequenciamento, no programa CNVnator foram detectadas 13.725 CNVR, sendo 7.204 compostas apenas por deleções, 4.961 apenas por duplicações e 1.560 complexas. No programa DELLY foram detectadas 5.714 CNVR, sendo 4.003 compostas apenas por deleções, 443 apenas por duplicações e 1.268 complexas. No conjunto CNVR_SEQ, foram identificadas 960 CNVR, de tamanho médio 22.786 bp (variando de 1.111 bp a 2.006.399 bp, com mediana igual a 3.346 bp e SD equivalente a 104.755,6 bp), somando 21.874.126 bp, sendo 728 CNVR compostas apenas por eventos de deleção, 63 CNVR apenas por duplicações e 169 CNVR complexas.

Dentre as CNVR de alta confiança, no conjunto CNVR_POP, com todas as 547 amostras estudadas, foram encontradas dez CNVR em oito cromossomos, de tamanho médio igual a 104.943 bp (variando de 14.879 bp a 521.437 bp, com mediana igual a 52.933 bp e SD equivalente a 151.104,4 bp), somando 1.049.430 bp, sendo quatro CNVR compostas apenas por eventos de deleção, duas CNVR

apenas por duplicações e quatro CNVR complexas (Tabela 2). Quatro CNVR foram identificadas em mais de 10% da população e uma em mais de 30%.

Tabela 2. Cromossomo, posição inicial e final, tamanho em pares de base (bp) e tipo do conjunto de CNVR de alta confiança (CNVR_POP)

Cromossomo	Posição inicial	Posição final	Tamanho (bp)	Tipo
2	123735242	123851299	116057	DELEÇÃO
3	54329751	54851188	521437	COMPLEXA
6	3202792	3240026	37234	DUPLICAÇÃO
9	5051796	5177690	125894	DELEÇÃO
9	29399118	29413997	14879	DELEÇÃO
9	30698315	30726606	28291	DELEÇÃO
15	44870278	44942116	71838	COMPLEXA
18	13328574	13397206	68632	DUPLICAÇÃO
19	23956716	23987626	30910	COMPLEXA
26	23374431	23408689	34258	COMPLEXA

No conjunto CNVR_ANI foram detectadas 240 CNV oriundas de dados de painéis de SNP e 77.582 oriundas de dados de sequenciamento. Após a sobreposição das CNV oriundas dos dados de painéis de SNP e de sequenciamento, foram identificadas 45 CNVR em 21 cromossomos, de tamanho médio de 97.931 bp (variando de 12.003 bp a 355.151 bp, com mediana igual a 53.140 bp e SD equivalente a 96.949,66 bp), somando 4.406.887 bp, sendo 23 CNVR compostas apenas por eventos de deleção e 22 CNVR apenas por duplicação (Tabela 3).

Tabela 3. Cromossomo, posição inicial e final, tamanho em pares de base (bp) e tipo do conjunto de CNVR de alta confiança (CNVR_ANIMAL)

Cromossomo	Posição inicial	Posição final	Tamanho (bp)	Tipo
1	18360408	18384034	23626	DUPLICAÇÃO
1	130963070	130991359	28289	DUPLICAÇÃO
2	123765001	123851299	86298	DELEÇÃO
2	719378	745361	25983	DUPLICAÇÃO

2	117790751	117904500	113749	DUPLICAÇÃO
2	134624266	134933500	309234	DUPLICAÇÃO
3	54367531	54651000	283469	DELEÇÃO
3	20917796	20944085	26289	DUPLICAÇÃO
4	82698947	82728750	29803	DUPLICAÇÃO
4	105218001	105292500	74499	DUPLICAÇÃO
5	7733251	7765707	32456	DELEÇÃO
6	11393501	11436703	43202	DELEÇÃO
7	9455783	9693750	237967	DELEÇÃO
7	9739213	9793250	54037	DELEÇÃO
7	10055082	10135500	80418	DELEÇÃO
7	41582849	41938000	355151	DELEÇÃO
9	5054168	5177690	123522	DELEÇÃO
9	29399118	29411121	12003	DELEÇÃO
9	15095199	15271750	176551	DUPLICAÇÃO
11	83535731	83559396	23665	DELEÇÃO
11	26400251	26444703	44452	DUPLICAÇÃO
12	59242099	59433070	190971	DELEÇÃO
12	70538501	70738500	199999	DELEÇÃO
12	71894273	71953261	58988	DELEÇÃO
12	167702	262500	94798	DUPLICAÇÃO
12	71187501	71259000	71499	DUPLICAÇÃO
12	71334251	71418750	84499	DUPLICAÇÃO
13	2199336	2238554	39218	DELEÇÃO
13	53461848	53511604	49756	DELEÇÃO
13	12487232	12761250	274018	DUPLICAÇÃO
14	79478001	79499712	21711	DUPLICAÇÃO
15	44881987	44933173	51186	DELEÇÃO
16	32607925	32655581	47656	DELEÇÃO
17	26898751	26929822	31071	DELEÇÃO
18	13344360	13397500	53140	DUPLICAÇÃO
18	58916664	59054123	137459	DUPLICAÇÃO
18	64384251	64406577	22326	DUPLICAÇÃO

20	3549957	3609244	59287	DELEÇÃO
20	57454844	57467750	12906	DELEÇÃO
21	58680616	58696778	16162	DUPLICAÇÃO
23	25679501	25705975	26474	DELEÇÃO
26	23378751	23408689	29938	DUPLICAÇÃO
28	123251	413750	290499	DELEÇÃO
28	627488	934000	306512	DUPLICAÇÃO
28	6398983	6451134	52151	DUPLICAÇÃO

Após a sobreposição das CNV contíguas e contínuas dos conjuntos CNVR_POP e CNVR_ANI, o conjunto resultante consistiu em 48 CNVR únicas e de alta confiança (Apêndice D) (Figura 4) que foram utilizadas para análise funcional. Sete CNVR (70% do conjunto CNVR_POP) foram encontradas em comum entre CNVR_POP e CNVR_ANI.

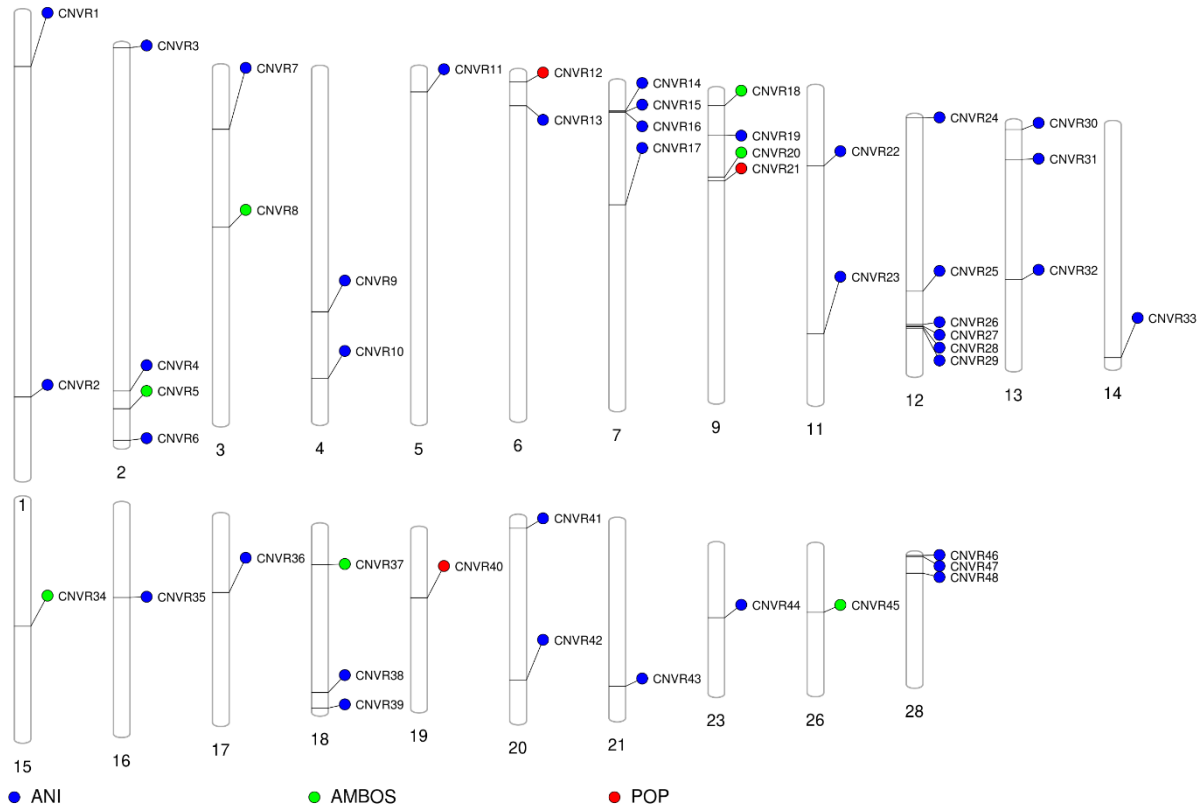


Figura 4. Distribuição de regiões de variação do número de cópia (CNVR) únicas e de alta confiança no genoma bovino. Estão representados os conjuntos CNVR_ANI (ANI), CNVR_POP (POP) e CNVR presentes nos dois conjuntos (AMBOS). Apenas os cromossomos com CNVR estão representados

3.6 Análise Funcional

De acordo com o RefSeq Genes e o Gene Cards, 69 genes e dois pseudogenes foram anotados em 31 CNVR únicas e de alta confiança (64,58%) (Apêndice E). Dentre esses, foram encontrados 21 genes e dois pseudogenes pertencentes a família de receptores olfativos (ex: *OR2L13*, *OR2L2*, *OR1P1*) nas CNVR14 (BTA7: 9455783-9693750), CNVR16 (BTA7:10055082-10135500), CNVR17 (BTA7:41582849-41938000), CNVR34 (BTA15:44870278-44942116), CNVR40 (BTA19:23956716-23987626), CNVR46 (BTA28:123251-413750). Três genes das proteínas de ligação de guanilato (GBP – "guanylate binding proteins") (*GBP2*, *GBP4*, *GBP6)* foram encontrados na CNVR8 (BTA3:54329751-54851188), as GBP participam da imunidade inata contra diversos patógenos intracelulares (Praefcke, 2018). Outros seis genes relacionados a imunidade (*HERC2*, *CLEC5A,*

SIRPB1, *BANP*, *BLA-DQB*, *ENSBTAG00000037605* ou *DQA1*) foram encontrados nas CNVR3 (BTA2:719378-745361), CNVR10 (BTA4:105218001-105292500), CNVR32 (BTA13:53461848-53511604), CNVR37 (BTA18:13328574-13397500) e CNVR44 (BTA23:25679501-25705975). As CNVR sobrepuseram regiões exônicas em todos os genes e pseudogenes encontrados. As regiões em que houveram apenas eventos de deleção se sobrepuseram a um pseudogene e 27 genes (39,43%), regiões que houveram apenas duplicação se sobrepuseram a 28 (39,43%) genes, e regiões complexas se sobrepuseram a um pseudogene e 14 genes (21,13%).

Em 14 CNVR (29,17%) únicas e de alta confiança foram encontrados 156 QTL, em que 44 QTL foram significativamente associados ($p<0,05$) a características de produção (29,54%), reprodução (22,73%), conformação (18,18%), sanidade (13,64%), leite (13,63%), carne e carcaça (2,27%) (Apêndice E). A maioria dos QTL (52,27%) se sobrepôs a regiões em que houve apenas eventos de duplicação, 43,18% dos QTL se sobrepuseram a regiões em que ocorreram eventos de deleção e 4,54% a regiões complexas.

Na análise de enriquecimento gênico de termos GO significativos (FDR<0,05), na categoria Processos Biológicos foi encontrado o termo detecção de estímulos (GO:0051606) e na categoria Funções Moleculares foi encontrado o termo atividade do receptor olfativo (GO:0004984). Os dois termos encontrados foram relacionados a cinco genes (*OR1P1*, *OR5D18K*, *OR2L13*, *OR2T22*, *OR2M16*). Nenhum termo significativamente enriquecido (FDR>0,05) foi encontrado para a categoria Componentes Celulares. Na análise de enriquecimento de vias biológicas preditas pelo banco de dados KEGG não foi encontrado nenhuma via significativamente enriquecida (FDR>0,05).

Na análise de enriquecimento de termos MeSH significativos (p-ajustado<0,05), foi encontrado um termo na categoria Anatomia, três termos na categoria Fenômenos e Processos, e 13 na categoria Químicos e Drogas. Esses termos foram relacionados à pelos menos um dos genes *BLA-DQB*, *ENSBTAG00000037605* (*DQA1*) e *GBP4* (Apêndice F). Nenhum termo significativamente enriquecido (p-ajustado>0,05) foi encontrado na categoria Doenças.

4 DISCUSSÃO

Após o controle de qualidade, dados de sequenciamento do genoma completo de 38 touros Gir Leiteiro representativos da população foram alinhados e as posições de painéis de SNP de alta densidade de amostras de 545 animais Gir Leiteiro foram baseadas no genoma referência ARS-UCD1.2. O total de 547 animais foram utilizados neste estudo, dentre esses, 36 possuíam informação de sequenciamento e de painel de SNP disponível. As CNV foram detectadas por essas técnicas moleculares e por diferentes metodologias de detecção. A partir da combinação de resultados, foram utilizados dois métodos *in silico* para a identificação de CNVR de alta confiabilidade, relativos aos indivíduos e a população estudada, em que se obteve 45 e dez CNVR de alta confiabilidade, que cobrem 4,4 Mb e 1,05 Mb, respectivamente. A análise funcional das regiões cobertas por CNVR revelou genes relacionados a características de interesse para a cadeia produtiva leiteira.

Nas CNV identificadas nos mesmos animais, mas a partir de diferentes fontes. de dados, foram encontradas 325 vezes mais CNV oriundas dos dados de sequenciamento do que aquelas oriundas de painéis de SNP. Em bovinos, Butty et al. (2020) e Zhan et al. (2011) também encontraram diferenças no número de CNV detectadas entre os resultados de painéis de SNP e de sequenciamento. Isso ocorre, pois, essas técnicas moleculares diferem em suas capacidades de detecção e de resolução de pontos de quebras de CNV. Certas CNV que são detectadas apenas a partir de dados de sequenciamento, podem ser verdadeiras, entretanto são pouco prováveis ou impossíveis de serem detectadas por painéis de SNP de alta densidade (Rafter et al., 2020), devido a quantidade e distribuição (Wang et al., 2007), e a posição pré-estabelecida de marcadores (Klambauer et al., 2012).

Os parâmetros utilizados nos programas de detecção de CNV e a estratégia para identificar conjuntos de CNVR de alta confiabilidade (CNVR_POP e CNVR_ANI) podem ter diminuído o número de CNVR identificadas. Entretanto, o foco deste trabalho foi a qualidade em termos de confiabilidade na detecção. As CNV podem ser parcialmente validadas quando a mesma região de variação é detectada com o uso do sequenciamento e dados de painéis de SNP (Butty et al.,

2020). Devido as chamadas de falso-positivas inerentes às metodologias de detecção de CNV e as limitações da validação experimental dessas em um grande número de animais, a combinação de diferentes técnicas moleculares e metodologias pode oferecer identificação de SV com alta confiança (Zhan et al., 2011). A acurácia da detecção de CNV e da definição de pontos de quebra pode ser aumentada pela utilização do sequenciamento de leituras longas (Couldrey et al., 2017). No entanto, o alto custo e o baixo rendimento dessa tecnologia de sequenciamento pode limitar seu uso em larga escala (Kosugi et al., 2019).

No conjunto CNVR_POP, além das CNV oriundas de dados de painéis de SNP, foram utilizadas de forma combinada aquelas detectadas pelas metodologias RD, SR e PE. A integração desses métodos pode auxiliar na diminuição da taxa de falsos positivos durante a detecção de CNV, em comparação com a utilização de uma única metodologia (Hu et al., 2020). Zhan et al (2011) relataram aumento da acurácia na detecção de CNV a partir da utilização de dados de dados de painéis de SNP e de sequenciamento, com mais de uma metodologia. A principal limitação da metodologia RD é a determinação dos pontos de quebras das SV, que pode ser superada pela utilização das metodologias RP e SR (Zhao et al., 2013; Pirooznia et al., 2015).

O conjunto CNVR_POP pode ser utilizado como critério de escolha de CNV a serem validadas na população. Ademais de serem confiáveis, essas CNVR podem ser consideradas como polimorfismos no número de cópia, pois estão presentes em mais de 1% da população estudada. Para validar as CNV encontradas, as técnicas moleculares de qPCR (PCR em tempo real) e FISH (Hibridização *in situ* Fluorescente) podem ser utilizadas (Bickhart et al., 2012). Entretanto, essas análises demandam tempo, possuem custo elevado e necessitam de quantidade suficiente de material biológico.

O conjunto CNVR_ANI foi definido para detectar CNVR de alta confiança presentes nos touros representativos e que, por isso, podem estar presentes na população de animais Gir Leiteiro. Esse conjunto foi obtido pela verificação entre as CNV encontradas a partir de dados de painéis de SNP e por meio da metodologia RD nos dados de sequenciamento, em que a detecção de CNV foi realizada por amostra. Essas metodologias utilizam formas semelhantes de detecção, a

quantidade de DNA presente em determinada região é utilizada para, de forma indireta, identificar CNV em cada amostra (Butty et al., 2020). Na metodologia RD, nos dados de sequência, isso é medido indiretamente pela cobertura de cada segmento (Abyzov et al., 2011). Nos painéis de SNP, a intensidade de sinal de fluorescência para cada sonda no momento da genotipagem também reflete a quantidade de DNA (Wang et al., 2007). Em bovinos, Zhan et al (2011) e Butty et al. (2020) também utilizaram dados de painéis de SNP e de sequenciamento para aumentar a acurácia na detecção de CNV.

Dentre as CNVR únicas e de alta confiança, CNVR foram encontradas adjacentes as extremidade dos cromossomos, outros estudos também encontraram CNVR próximas ou localizadas em regiões teloméricas em bovinos (Hou et al., 2012b; Butty et al., 2020; Sasaki et al., 2021). Duplicações segmentares são consideradas regiões “hotspots” para CNV, essas são encontradas em regiões teloméricas e subtelomérica (Liu et al., 2009). As CNV e CNVR encontradas neste estudo estabelecem a base para futuras pesquisas com SV em zebuíno, essas podem auxiliar no desenvolvimento de painéis de genotipagem para o Gir leiteiro, pela escolha de SNP que estejam sobrepostos às CNVR. Além do mais, a utilização de CNV na escolha de touros para o teste de progênie e a inclusão de CNVR na avaliação genômica pode provocar aumento no ganho genético e deve ser verificada.

Análises adicionais são necessárias para investigar a relação entre CNVR e características economicamente importantes. Alguns dos genes encontrados nas CNVR únicas e de alta confiança foram previamente relacionados a características reprodutivas e de sanidade. As CNVR presentes nos genes *FILIP1* (*Filamin A Interacting Protein 1*) e *SENP6* (*SUMO Specific Peptidase 6*) foram associados ao número de progênies por gestação em ovelhas (Salehian-Dehkordi et al., 2021). Os genes *FILIP1*, *SENP6*, *CA5A* (*Carbonic anhydrase 5A*) e *BANP* (*BTG3 Associated Nuclear Protein*) foram relacionados a característica de longevidade em bovinos da raça Holandesa (Zhang et al., 2021). Em ovelhas, o gene *SENP6* foi relacionado a mortalidade embrionária (Pokharel et al., 2020) e *HERC2* (*HECT And RLD Domain Containing E3 Ubiquitin Protein Ligase 2*) foi associado a resistência à parasitas gastrointestinais (Al Kalaldeh et al., 2019).

Na CNVR37 foi mapeado o gene *CA5A*. As enzimas anidrases carbônicas catalisam a conversão reversível de dióxido de carbono em bicarbonato, liberando prótons e também auxiliam no transporte dessas moléculas por meio de membranas biológicas (Hassan et al., 2013). Em bovinos, o gene *CA5A* foi encontrado presente em regiões de assinatura de seleção e pode estar relacionado a adaptação ao ambiente em raças iraquianas (Alshawi et al., 2019). Esse gene também foi relacionado a mecanismos reprodutivos em ovelhas (Hernández-Montiel et al., 2019; Pokharel et al., 2020). Na CNVR46 foi mapeado o gene *RHOU* (*ras homolog family member U*), sendo que esse gene codifica uma proteína da família RHO de GTPases (guanina trifosfatases), que regula processos fundamentais para o desenvolvimento da glândula mamária (Bray et al., 2011).

Os genes das proteínas de ligação de guanilato (GBP) *GBP2*, *GBP4*, *GBP6* foram encontrados nas CNVR8. GBP são importantes na eliminação de parasitas intracelulares, que é mediada pelo IFN-γ (interferon-γ) durante a resposta imune inata (Sasai et al., 2018). Park et al. (2016) relataram que o gene GBP6 possui papel importante na resposta imunológica contra *Mycobacterium avium* subespécie *paratuberculosis* em bovinos, sugerindo que esse gene pode atuar na morte intracelular desse patógeno. Cao et al. (2018) encontraram uma região de polimorfismo no número de cópias de natureza complexa presente no gene *GBP4*, que foi negativamente associada a estatura em bovinos de raças chinesas. Hou et al. (2012b) encontraram CNV nos genes da família GBP (*GBP2*, *GBP4*, *GBP5 e GBP7*) associadas a consumo alimentar residual em vacas Holandesas. Em adição, Ghoreishifar et al. (2020) em um estudo com gado suíço, encontraram regiões de assinaturas de seleção sobrepostas aos genes *GBP2*, *GBP4*, *GBP6*.

Na CNVR44 foram encontrados os genes *BLA-DQB* (*MHC class II antigen*) (*DQB1*, *ENSBTAG00000019588*) e *ENSBTAG00000037605* (*DQA1*) pertencentes a região de complexo principal de histocompatibilidade bovina (MHC) classe II. Termos MeSH relacionados a sistema imunológico e a duplicação gênica foram enriquecidos nestes dois genes. Moléculas da classe II são expressas em células que apresentam epítopos de antígenos, como as células dendríticas, aos linfócitos T CD4+ que ativam macrófagos e linfócitos B provocando resposta inflamatória e produção de anticorpos, respectivamente (Behl et al., 2012). O *ENSBTAG00000037605* (*DQA1*)

foi associado a carga pró viral na infecção pelo vírus da leucemia bovina. Essa carga pode ser considerada como índice de diagnóstico para a determinação da progressão e risco de transmissão dessa doença (Takeshima et al., 2019).

Dentre os genes encontrados, 30,43% pertencem a família olfativa e esses foram encontrados nas CNVR14, 17, 18 e 34 em que ocorreram eventos de deleção ou eventos complexos. Ainda, dois termos GO foram enriquecidos. A expressão e regulação dos genes dos receptores olfativos (OR) está relacionada a recepção de informações sobre o meio ambiente e a comunicação e comportamento entre bovinos pelo sentido do olfato, por meio de feromônios (Samuel and Dinka, 2020). Genes da família de OR possuem papel evolutivo e estão sob pressão seletiva em animais (Bickhart and Liu, 2014). Variações genômicas no genes olfativos, tais como SNP e CNV, estão associados à estresse em humanos (Melroy-Greif et al., 2017) e afecções no casco em bovinos de raça Holandesa (Butty et al., 2021), respectivamente. CNVR que englobam genes de OR também foram encontradas em outros estudos com bovinos (Butty et al., 2020).

5 CONCLUSÃO

Nossos resultados abrangem a identificação e caracterização de 48 CNVR de alta confiança no genoma de bovinos Gir Leiteiro, relativas aos indivíduos e a população estudada. Isso contribui para a elaboração de um mapa de SV na raça Gir e para o melhor entendimento do genoma dos zebuínos. As CNVR identificadas neste estudo podem afetar potencialmente genes que estão envolvidos no processo evolutivo e no controle fenotípico de característica de interesse para a cadeia produtiva leiteira, como imunidade, lactação, reprodução, reconhecimento de estímulos e sanidade.

6 REFERÊNCIAS

Abyzov, A., A.E. Urban, M. Snyder, and M. Gerstein. 2011. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 21:974–984. doi:10.1101/gr.114876.110.

Aguiar, T.S., R.B.P. Torrecilha, M. Milanesi, A.T.H. Utsunomiya, B.B. Trigo, A. Tijjani, H.H. Musa, F.L. Lopes, P. Ajmone-Marsan, R. Carvalheiro, H.H. de R. Neves, A.S. Do Carmo, O. Hanotte, T.S. Sonstegard, J.F. Garcia, and Y.T. Utsunomiya. 2018. Association of copy number variation at intron 3 of hmga2 with navel length in bos indicus. Front. Genet. 9. doi:10.3389/fgene.2018.00627.

Alshawi, A., A. Essa, S. Al-Bayatti, and O. Hanotte. 2019. Genome Analysis Reveals Genetic Admixture and Signature of Selection for Productivity and Environmental Traits in Iraqi Cattle. Front. Genet. 10. doi:10.3389/fgene.2019.00609.

Behl, J.D., N.K. Verma, N. Tyagi, P. Mishra, R. Behl, and B.K. Joshi. 2012. The Major Histocompatibility Complex in Bovines: A Review. ISRN Vet. Sci. 2012:1–12. doi:10.5402/2012/872710.

Bickhart, D.M., Y. Hou, S.G. Schroeder, C. Alkan, M.F. Cardone, L.K. Matukumalli, J. Song, R.D. Schnabel, M. Ventura, J.F. Taylor, J.F. Garcia, C.P. Van Tassell, T.S. Sonstegard, E.E. Eichler, and G.E. Liu. 2012. Copy number variation of individual cattle genomes using next-generation sequencing. Genome Res. 22:778–790. doi:10.1101/gr.133967.111.

Bickhart, D.M., and G.E. Liu. 2014. The challenges and importance of structural variation detection in livestock. Front. Genet. 5:37. doi:10.3389/FGENE.2014.00037.

Bray, K., C. Brakebusch, and T. Vargo-Gogola. 2011. The Rho GTPase Cdc42 is required for primary mammary epithelial cell morphogenesis in vitro. Small GTPases 2:247–258. doi:10.4161/sgtp.2.5.18163.

Butty, A.M., T.C.S. Chud, D.F. Cardoso, L.S.F. Lopes, F. Miglior, F.S. Schenkel, A. Cánovas, I.M. Häfliger, C. Drögemüller, P. Stothard, F. Malchiodi, and C.F. Baes. 2021. Genome-wide association study between copy number variants and hoof health traits in Holstein dairy cattle. J. Dairy Sci. 104:8050–8061. doi:10.3168/jds.2020-19879.

Butty, A.M., T.C.S. Chud, F. Miglior, F.S. Schenkel, A. Kommadath, K. Krivushin, J.R. Grant, I.M. Häfliger, C. Drögemüller, A. Cánovas, P. Stothard, and C.F. Baes. 2020. High confidence copy number variants identified in Holstein dairy cattle from whole genome sequence and genotype array data. Sci. Rep. 10:1–13. doi:10.1038/s41598-020-64680-3.

Cao, X.-K., Y.-Z. Huang, Y.-L. Ma, J. Cheng, Z.-X. Qu, Y. Ma, Y.-Y. Bai, F. Tian, F.-P. Lin, Y.-L. Ma, and H. Chen. 2018. Integrating CNVs into meta-QTL identified GBP4 as positional candidate for adult cattle stature. Funct. Integr. Genomics 18:559–567. doi:10.1007/s10142-018-0613-0.

Chang, C.C., C.C. Chow, L.C. Tellier, S. Vattikuti, S.M. Purcell, and J.J. Lee. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4:7. doi:10.1186/S13742-015-0047-8.

Couldrey, C., M. Keehan, T. Johnson, K. Tiplady, A. Winkelman, M.D. Littlejohn, A. Scott, K.E. Kemper, B. Hayes, S.R. Davis, and R.J. Spelman. 2017. Detection and assessment of copy number variation using PacBio long-read and Illumina sequencing in New Zealand dairy cattle. J. Dairy Sci.. doi:10.3168/jds.2016-12199.

Diskin, S.J., M. Li, C. Hou, S. Yang, J. Glessner, H. Hakonarson, M. Bucan, J.M. Maris, and K. Wang. 2008. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Res. 36:126. doi:10.1093/nar/gkn556.

Fonseca, P.A.S., A. Suárez-Vega, G. Marras, and Á. Cánovas. 2020. GALLO: An R package for genomic annotation and integration of multiple data sources in livestock for positional candidate loci. Gigascience 9:1–9. doi:10.1093/gigascience/giaa149.

Ghoreishifar, S.M., S. Eriksson, A.M. Johansson, M. Khansefid, S. Moghaddaszadeh-Ahrabi, N. Parna, P. Davoudi, and A. Javanmard. 2020. Signatures of selection reveal candidate genes involved in economic traits and cold acclimation in five Swedish cattle breeds. Genet. Sel. Evol. 52:1–15. doi:10.1186/s12711-020-00571-5.

Hassan, M.I., B. Shajee, A. Waheed, F. Ahmad, and W.S. Sly. 2013. Structure, function and applications of carbonic anhydrase isozymes. Bioorganic Med. Chem. 21:1570–1582. doi:10.1016/j.bmc.2012.04.044.

Hay, E.H.A., Y.T. Utsunomiya, L. Xu, Y. Zhou, H.H.R. Neves, R. Carvalheiro, D.M. Bickhart, L. Ma, J.F. Garcia, and G.E. Liu. 2018. Genomic predictions combining SNP markers and copy number variations in Nellore cattle. BMC Genomics 19. doi:10.1186/s12864-018-4787-6.

Hernández-Montiel, W., R.C. Collí-Dula, J.P. Ramón-Ugalde, M.A. Martínez-Núñez, and R. Zamora-Bustillos. 2019. RNA-seq transcriptome analysis in ovarian tissue of pelibuey breed to explore the regulation of prolificacy. Genes (Basel). 10:358. doi:10.3390/genes10050358.

Hou, Y., D.M. Bickhart, H. Chung, J.L. Hutchison, H.D. Norman, E.E. Connor, and G.E. Liu. 2012a. Analysis of copy number variations in Holstein cows identify potential mechanisms contributing to differences in residual feed intake. Funct. Integr. Genomics 12:717–723. doi:10.1007/s10142-012-0295-y.

Hou, Y., D.M. Bickhart, M.L. Hvinden, C. Li, J. Song, D.A. Boichard, S. Fritz, A. Eggen, S. DeNise, G.R. Wiggans, T.S. Sonstegard, C.P. Van Tassell, and G.E. Liu. 2012b. Fine mapping of copy number variations on two cattle genome assemblies using high density SNP array. BMC Genomics 13. doi:10.1186/1471-2164-13-376.

Hu, Y., H. Xia, M. Li, C. Xu, X. Ye, R. Su, M. Zhang, O. Nash, T.S. Sonstegard, L. Yang, G.E. Liu, and Y. Zhou. 2020. Comparative analyses of copy number variations between Bos taurus and Bos indicus. BMC Genomics 21:1–11. doi:10.1186/s12864-020-07097-6.

Illumina. 2014. Infinium Genotyping Data Analysis. 2014. Pub. No. 970-2007-005. Available online at: < https://www.illumina.com/Documents/products/technotes/technote_infinium_genotyping_data_analysis.pdf>.

Al Kalaldeh, M., J. Gibson, S.H. Lee, C. Gondro, and J.H.J. Van Der Werf. 2019. Detection of genomic regions underlying resistance to gastrointestinal parasites in Australian sheep. Genet. Sel. Evol. 51:37. doi:10.1186/s12711-019-0479-1.

Keel, B.N., A.K. Lindholm-Perry, and W.M. Snelling. 2016. Evolutionary and functional features of copy number variation in the cattle genome. Front. Genet. 7:207. doi:10.3389/fgene.2016.00207.

Khan, F.F., P.E. Melton, N.S. McCarthy, B. Morar, J. Blangero, E.K. Moses, and A. Jablensky. 2018. Whole genome sequencing of 91 multiplex schizophrenia families reveals increased burden of rare, exonic copy number variation in schizophrenia probands and genetic heterogeneity. Schizophr. Res. 197:337–345. doi:10.1016/j.schres.2018.02.034.

Klambauer, G., K. Schwarzbauer, A. Mayr, D.A. Clevert, A. Mitterecker, U. Bodenhofer, and S. Hochreiter. 2012. Cn.MOPS: Mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Res. 40:e69–e69. doi:10.1093/nar/gks003.

Kosugi, S., Y. Momozawa, X. Liu, C. Terao, M. Kubo, and Y. Kamatani. 2019. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biol. 20:1–18. doi:10.1186/s13059-019-1720-5.

Li, H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987–2993. doi:10.1093/bioinformatics/btr509.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760. doi:10.1093/bioinformatics/btp324.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. doi:10.1093/bioinformatics/btp352.

Liu, G.E., and D.M. Bickhart. 2012. Copy number variation in the cattle genome.. Funct. Integr. Genomics 12:609–624. doi:10.1007/s10142-012-0289-9.

Liu, G.E., Y. Hou, B. Zhu, M.F. Cardone, L. Jiang, A. Cellamare, A. Mitra, L.J. Alexander, L.L. Coutinho, M.E. Dell'Aquila, L.C. Gasbarre, G. Lacalandra, R.W. Li, L.K. Matukumalli, D. Nonneman, L.C.D.A. Regitano, T.P.L. Smith, J. Song, T.S. Sonstegard, C.P. Van Tassell, M. Ventura, E.E. Eichler, T.G. McDaneld, and J.W. Keele. 2010. Analysis of copy number variations among diverse cattle breeds. Genome Res. 20:693–703. doi:10.1101/gr.105403.110.

Liu, G.E., M. Ventura, A. Cellamare, L. Chen, Z. Cheng, B. Zhu, C. Li, J. Song, and E.E. Eichler. 2009. Analysis of recent segmental duplications in the bovine genome. BMC Genomics 10. doi:10.1186/1471-2164-10-571.

Machado, M.A., A.L.S. Azevedo, R.L. Teodoro, M.A. Pires, M.G.C. Peixoto, C. de Freitas, M.C.A. Prata, J. Furlong, M.V.G. da Silva, S.E. Guimarães, L.C. Regitano, L.L. Coutinho, G. Gasparin, and R.S. Verneque. 2010. Genome wide scan for quantitative trait loci affecting tick resistance in cattle (Bos taurus × Bos indicus). BMC Genomics 2010 111 11:1–11. doi:10.1186/1471-2164-11-280.

Melroy-Greif, W.E., K.C. Wilhelmsen, R. Yehuda, and C.L. Ehlers. 2017. Genome-wide association study of post-traumatic stress disorder in two high-risk populations. Twin Res. Hum. Genet. 20:197–207. doi:10.1017/thg.2017.12.

Mills, R.E., et al. 2011. Mapping copy number variation by population-scale genome sequencing. Nature 470:59–65. doi:10.1038/nature09708.

Panetto, J.C. do C., M.V.G.B. Silva, R. da S. Verneque, M.A. Machado, A.R. Fernandes, M.F. Martins, D.R. de L. Reis, W.A. Arbex, J.C. de Oliveira, H.T. Ventura, and M.A. Pereira. 2020. Programa Nacional de Melhoramento Do Gir Leiteiro Sumário Brasileiro de Touros 3ª Avaliação Genômica de Touros Resultado Do Teste de Progênie - Maio 2020. Embrapa Gado de Leite, Juiz de Fora.

Park, H.E., M.K. Shin, H.T. Park, M. Jung, Y. Il Cho, and H.S. Yoo. 2016. Gene expression profiles of putative biomarker candidates in Mycobacterium avium subsp. paratuberculosis-infected cattle. Pathog. Dis. 74. doi:10.1093/femspd/ftw022.

Picard toolkit. 2019. . Broad Institute, GitHub Repos.

Pirooznia, M., F. Goes, and P.P. Zandi. 2015. Whole-genome CNV analysis: Advances in computational approaches. Front. Genet. 6:138. doi:10.3389/fgene.2015.00138.

Pokharel, K., J. Peippo, M. Weldenegodguad, M. Honkatukia, M.H. Li, and J. Kantanen. 2020. Gene expression profiling of corpus luteum reveals important insights about early pregnancy in domestic sheep. Genes (Basel). 11:415. doi:10.3390/genes11040415.

Praefcke, G.J.K. 2018. Regulation of innate immune functions by guanylate-binding proteins. Int. J. Med. Microbiol. 308:237–245. doi:10.1016/j.ijmm.2017.10.013.

Quinlan, A.R., and I.M. Hall. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. doi:10.1093/bioinformatics/btq033.

R Core Team. 2021. R: A Language and Environment for Statistical Computing.

Rafter, P., I.C. Gormley, A.C. Parnell, J.F. Kearney, and D.P. Berry. 2020. Concordance rate between copy number variants detected using either high-or medium-density single nucleotide polymorphism genotype panels and the potential of imputing copy number variants from flanking high density single nucleotide polymorphism haplotyp. BMC Genomics 21:1–10. doi:10.1186/s12864-020-6627-8.

Rausch, T., M. Hsi-Yang Fritz, J.O. Korbel, and V. Benes. 2019. Alfred: Interactive multi-sample BAM alignment statistics, feature counting and feature annotation for long- and short-read sequencing. Bioinformatics 35:2489–2491. doi:10.1093/bioinformatics/bty1007.

Rausch, T., T. Zichner, A. Schlattl, A.M. Stütz, V. Benes, and J.O. Korbel. 2012. DELLY: Structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28:i333–i339. doi:10.1093/bioinformatics/bts378.

Redon, R., S. Ishikawa, K.R. Fitch, L. Feuk, G.H. Perry, T.D. Andrews, H. Fiegler, M.H. Shapero, A.R. Carson, W. Chen, E.K. Cho, S. Dallaire, J.L. Freeman, J.R. González, M. Gratacòs, J. Huang, D. Kalaitzopoulos, D. Komura, J.R. MacDonald, C.R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M.J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J.

Zhang, L. Armengol, D.F. Conrad, X. Estivill, C. Tyler-Smith, N.P. Carter, H. Aburatani, C. Lee, K.W. Jones, S.W. Scherer, and M.E. Hurles. 2006. Global variation in copy number in the human genome. Nature 444:444–454. doi:10.1038/nature05329.

Salehian-Dehkordi, H., Y.X. Xu, S.S. Xu, X. Li, L.Y. Luo, Y.J. Liu, D.F. Wang, Y.H. Cao, M. Shen, L. Gao, Z.H. Chen, J.T. Glessner, J.A. Lenstra, A. Esmailizadeh, M.H. Li, and F.H. Lv. 2021. Genome-Wide Detection of Copy Number Variations and Their Association With Distinct Phenotypes in the World's Sheep. Front. Genet. 12:670582. doi:10.3389/fgene.2021.670582.

Samuel, B., and H. Dinka. 2020. In silico analysis of the promoter region of olfactory receptors in cattle (Bos indicus) to understand its gene regulation. Nucleosides, Nucleotides and Nucleic Acids 39:853–865. doi:10.1080/15257770.2020.1711524.

Santana, M.L., R.J. Pereira, A.B. Bignardi, L. El Faro, H. Tonhati, and L.G. Albuquerque. 2014. History, structure, and genetic diversity of Brazilian Gir cattle. Livest. Sci. 163:26–33. doi:10.1016/j.livsci.2014.02.007.

Sasai, M., A. Pradipta, and M. Yamamoto. 2018. Host immune responses to Toxoplasma gondii. Int. Immunol. 30:113–119. doi:10.1093/intimm/dxy004.

Sasaki, S., Y. Miki, T. Ibi, H. Wakaguri, Y. Yoshida, Y. Sugimoto, and Y. Suzuki. 2021. A 44-kb deleted-type copy number variation is associated with decreasing complement component activity and calf mortality in Japanese Black cattle. BMC Genomics 22:1–10. doi:10.1186/s12864-021-07415-6.

Shaun Purcell, and Christopher Chang. PLINK 1.9. Accessed August 26, 2021. https://www.cog-genomics.org/plink/1.9/.

Takeshima, S.N., A. Ohno, and Y. Aida. 2019. Bovine leukemia virus proviral load is more strongly associated with bovine major histocompatibility complex class II DRB3 polymorphism than with DQA1 polymorphism in Holstein cow in Japan. Retrovirology 16:14. doi:10.1186/s12977-019-0476-z.

Wang, J., and Y. Liao. 2020. WebGestaltR: Gene Set Analysis Toolkit WebGestaltR.

Wang, K., M. Li, D. Hadley, R. Liu, J. Glessner, S.F.A. Grant, H. Hakonarson, and M. Bucan. 2007. PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. 17:1665–1674. doi:10.1101/gr.6861907.

Xu, L., J.B. Cole, D.M. Bickhart, Y. Hou, J. Song, P.M. VanRaden, T.S. Sonstegard, C.P. Van Tassell, and G.E. Liu. 2014. Genome wide CNV analysis reveals additional variants associated with milk production traits in Holsteins. BMC Genomics 15:1–10. doi:10.1186/1471-2164-15-683.

Yu, G. 2018. Using meshes for MeSH term enrichment and semantic analyses. Bioinformatics 34:3766–3767. doi:10.1093/bioinformatics/bty410.

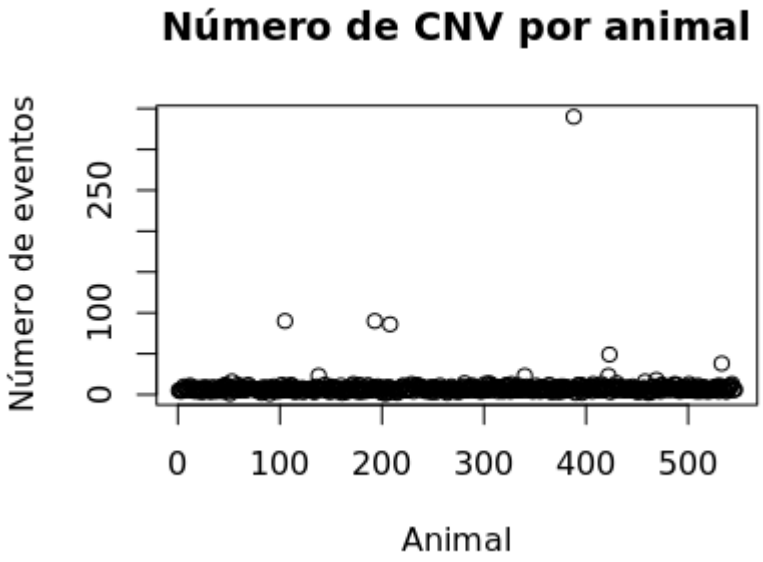
Zhan, B., J. Fadista, B. Thomsen, J. Hedegaard, F. Panitz, and C. Bendixen. 2011. Global assessment of genomic variation in cattle by genome resequencing and high-throughput genotyping. BMC Genomics 12:1–20. doi:10.1186/1471-2164-12-557.

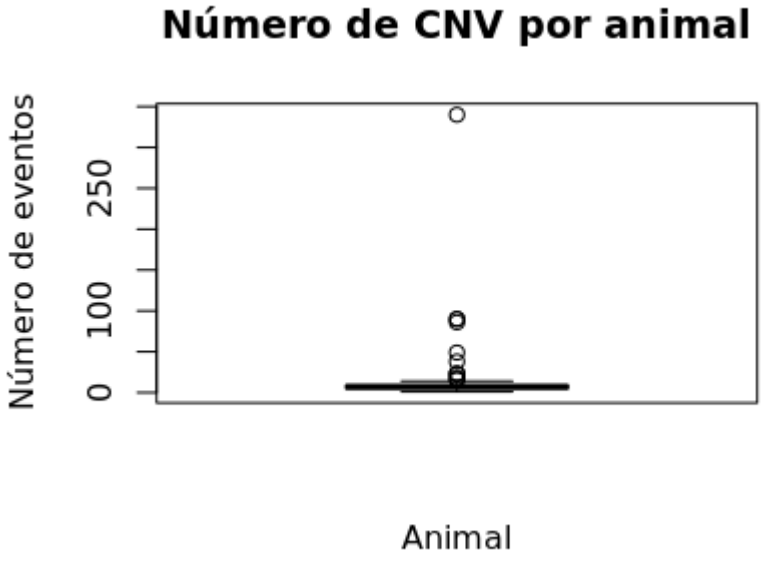
Zhang, H., A. Liu, Y. Wang, H. Luo, X. Yan, X. Guo, X. Li, L. Liu, and G. Su. 2021. Genetic Parameters and Genome-Wide Association Studies of Eight Longevity Traits Representing Either Full or Partial Lifespan in Chinese Holsteins. Front. Genet. 12:231. doi:10.3389/fgene.2021.634986.

Zhao, M., Q. Wang, Q. Wang, P. Jia, and Z. Zhao. 2013. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: Features and perspectives. BMC Bioinformatics 14. doi:10.1186/1471-2105-14-S11-S1.

APÊNDICES

Apêndice A. Gráficos de número de eventos de CNV detectados por animal pelo programa PennCNV





Apêndice B. Sinal de “Read Depth” (RD), número de eventos de CNV, de deleções e de duplicações de amostras que foram excluídas após a etapa de detecção pelo CNVnator

Amostra*	Sinal RD	Número de eventos	Número de deleções	Número de duplicações
39 (1258)	3,2	14.649	3.973	10.676
40 (1501)	2,5	8.186	1.000	7.186
41 (721)	2,2	13.114	2.124	10.990
42 (754)	1,9	8.927	4.279	4.648
43 (831)	3,6	7.389	2.737	4.652

*Número entre parênteses corresponde a identificação original das amostras.

Apêndice C. Estatísticas de alinhamento (pares de leituras mapeados ao genoma, pares de leitura mapeados ao mesmo cromossomo e fração de pares mapeados ao mesmo cromossomo (%) calculadas pelo programa Alfred opção qc

Amostra*	Pares de leituras mapeados	Pares mapeados ao mesmo cromossomo	Fração de pares mapeados ao mesmo cromossomo (%)
39 (1258)	154663600	143874037	92,7
40 (1501)	156291209	117619247	75
41 (721)	177981383	128139404	71,8
42 (754)	172792124	144119506	83.2
43 (831)	162124231	151696186	93,2

*Número entre parênteses corresponde a identificação original das amostras.

Apêndice D. Identificação (CNVR), Cromossomo (BTA), posição inicial e final, tamanho em pares de base (bp), classificação do tipo de CNVR e número de indivíduos identificados nas CNVR dos conjuntos CNVR_POP e CNVR_ANI das CNVR únicas e de alta confiança

CNVR	BTA	Posição inicial (bp)	Posição final (bp)	Tamanho (bp)	Tipo CNVR_POP	Tipo CNVR_ANI	Amostras CNVR_POP*	Amostras CNVR_ANI
CNVR1	1	18360408	18384034	23626	-	DUPLICAÇÃO	-	5
CNVR2	1	130963070	130991359	28289	-	DUPLICAÇÃO	-	1
CNVR3	2	719378	745361	25983	-	DUPLICAÇÃO	-	2
CNVR4	2	117790751	117904500	113749	-	DUPLICAÇÃO	-	2
CNVR5	2	123735242	123851299	116057	DELEÇÃO	DELEÇÃO	51 (9,32%)	3
CNVR6	2	134624266	134933500	309234	-	DUPLICAÇÃO		2
CNVR7	3	20917796	20944085	26289	-	DUPLICAÇÃO		1
CNVR8	3	54329751	54851188	521437	COMPLEXA	DELEÇÃO	185 (33,82%)	3
CNVR9	4	82698947	82728750	29803	-	DUPLICAÇÃO	-	10
CNVR10	4	105218001	105292500	74499	-	DUPLICAÇÃO	-	1
CNVR11	5	7733251	7765707	32456	-	DELEÇÃO	-	1
CNVR12	6	3202792	3240026	37234	DUPLICAÇÃO	-	28 (5,12%)	-
CNVR13	6	11393501	11436703	43202	-	DELEÇÃO		3
CNVR14	7	9455783	9693750	237967	-	DELEÇÃO	-	2
CNVR15	7	9739213	9793250	54037	-	DELEÇÃO	-	4

Apêndice D. Continuação

CNVR16	7	10055082	10135500	80418	-	DELEÇÃO	-	4
CNVR17	7	41582849	41938000	355151	-	DELEÇÃO	-	6
CNVR18	9	5051796	5177690	125894	DELEÇÃO	DELEÇÃO	32 (5,85%)	12
CNVR19	9	15095199	15271750	176551	-	DUPLICAÇÃO	-	13
CNVR20	9	29399118	29413997	14879	DELEÇÃO	DELEÇÃO	72 (13,16%)	2
CNVR21	9	30698315	30726606	28291	DELEÇÃO	-	40 (73,13%)	-
CNVR22	11	26400251	26444703	44452	-	DUPLICAÇÃO	-	7
CNVR23	11	83535731	83559396	23665	-	DELEÇÃO	-	1
CNVR24	12	167702	262500	94798	-	DUPLICAÇÃO	-	1
CNVR25	12	59242099	59433070	190971	-	DELEÇÃO	-	2
CNVR26	12	70538501	70738500	199999	-	DELEÇÃO	-	1
CNVR27	12	71187501	71259000	71499	-	DUPLICAÇÃO	-	2
CNVR28	12	71334251	71418750	84499	-	DUPLICAÇÃO	-	3
CNVR29	12	71894273	71953261	58988	-	DELEÇÃO	-	4
CNVR30	13	2199336	2238554	39218	-	DELEÇÃO	-	1
CNVR31	13	12487232	12761250	274018	-	DUPLICAÇÃO	-	13
CNVR32	13	53461848	53511604	49756	-	DELEÇÃO	-	14
CNVR33	14	79478001	79499712	21711	-	DUPLICAÇÃO	-	2
CNVR34	15	44870278	44942116	71838	COMPLEXA	DELEÇÃO	68 (12,43%)	1

Apêndice D. Continuação

CNVR35	16	32607925	32655581	47656	-	DELEÇÃO	-	1
CNVR36	17	26898751	26929822	31071	-	DELEÇÃO	-	11
CNVR37	18	13328574	13397500	68926	DUPLICAÇÃO	DUPLICAÇÃO	98 (17,91%)	1
CNVR38	18	58916664	59054123	137459	-	DUPLICAÇÃO	-	1
CNVR39	18	64384251	64406577	22326	-	DUPLICAÇÃO	-	2
CNVR40	19	23956716	23987626	30910	COMPLEXA	-	39 (7,13%)	-
CNVR41	20	3549957	3609244	59287	-	DELEÇÃO	-	1
CNVR42	20	57454844	57467750	12906	-	DELEÇÃO	-	3
CNVR43	21	58680616	58696778	16162	-	DUPLICAÇÃO	-	1
CNVR44	23	25679501	25705975	26474	-	DELEÇÃO	-	1
CNVR45	26	23374431	23408689	34258	COMPLEXA	DUPLICAÇÃO	36 (6,58%)	3
CNVR46	28	123251	413750	290499	-	DELEÇÃO	-	6
CNVR47	28	627488	934000	306512	-	DUPLICAÇÃO	-	1
CNVR48	28	6398983	6451134	52151	-	DUPLICAÇÃO	-	1

* Frequência relativa representada entre parênteses

Apêndice E. Identificação (CNVR), genes e pseudogenes encontrados (Genes e pseudogenes) e tipo QTL e características significativamente associadas (p<0,05) a esse QTL das CNVR únicas e de alta confiança

CNVR	Genes e pseudogenes	Tipo de QTL e característica associada*
1	-	-
2	-	-
3	*HERC2*	-
4	*ENSBTAG00000049805, TRIP12*	-
5	-	-
6	*ACTL8*	Produção (66904) - Ganho de peso; Leite (125440) - Persistência da Lactação
7	*MGC134040*	-
8	*ENSBTAG00000002416, GBP4, ENSBTAG00000024272, GBP6, ENSBTAG00000038233, GBP2, U2*	Reprodução (181182, 181446) - Taxa de concepção, Inseminações por concepção
9	*ENSBTAG00000002859, ENSBTAG00000055254*	-
10	*CLEC5A, TAS2R38, MGAM*	-
11	-	-
12	-	-
13	-	-
14	*OR7A95, ENSBTAG00000050759, OR7A78, OR7A99, OR7A97, ENSBTAG00000054398*	-
15	*ENSBTAG00000047589*	-
16	*OR7A112*	-

Apêndice E. Continuação

17	*OR2M16, OR2L13, OR2AJ9, OR2T22, OR2AJ10P* (pseudogene)*, OR2L2C, OR2L2B, OR2L2, OR2L3C*	Leite (64049, 64050) - Conteúdo de riboflavina no leite
18	-	-
19	*FILIP1, SENP6, ENSBTAG00000032382*	Saúde (167889) - Susceptibilidade a tuberculose bovina
20	-	-
21	-	-
22	*ENSBTAG00000016794, U6, ENSBTAG00000054517*	-
23	-	Produção (45767, 45772, 45775, 45780) – Profundidade corporal, PTA Tipo, Mérito Líquido, Largura da garupa; Reprodução (45768, 45781) -Facilidade de parto como característica da mãe, Facilidade de parto; Conformação (45769, 45770, 45773, 45778, 45782, 45783) - Ângulo do casco, Conformação de pés e pernas, Inserção do úbere, Posição das pernas - vista por trás, Estatura, Força
24	-	-
25	-	-
26	*ENSBTAG00000046041*	-
27	*ENSBTAG00000049836*	-
28	*ENSBTAG00000026070*	Saúde (211939) - susceptibilidade à *M. paratuberculosis*
29	*ENSBTAG00000052990*	-

Apêndice E. Continuação

30	*ENSBTAG00000054174*	-
31	*ECHDC3, USP6NL*	Carne e Carcaça (36961) – Rendimento de carne magra
32	*SIRPB1, ENSBTAG00000054594*	-
33	-	-
34	*OR10AB6, OR10AB2, OR5P90P* (pseudogene)*, OR5P1C, OR5P76B*	-
35	ENSBTAG00000044066	-
36	-	-
37	*CA5A, BANP*	-
38	*ENSBTAG00000015899, ENSBTAG00000052265, ENSBTAG00000054310*	Produção (123784, 123053, 123250, 123656, 122886, 122887, 123291) – Duração da vida produtiva; Reprodução (147014, 147022) – Facilidade de parto como característica da mãe, Facilidade de parto
39	*ENSBTAG00000050946*	-
40	*OR1P1, U6*	-
41	-	Reprodução (139139) – circunferência escrotal; Leite (158256) – Tempo de ordenha
42	-	Saúde (179855, 179913) - Cetose
43	*ENSBTAG00000031834*	Conformação (125875, 125954) – Conformação de pés e pernas, Qualidade óssea
44	*BLA-DQB, ENSBTAG00000037605 (DQA1)*	Saúde (153931) – Susceptibilidade ao vírus da leucemia bovina

Apêndice E. Continuação

45	*WBP1L*	-
46	*OR5AS1, OR5D18K, OR5L20*	-
47	*RHOU*	Saúde (179050) - Susceptibilidade a tuberculose bovina; Reprodução (53666, 53667, 53673) – Facilidade de parto como característica da mãe, Natimortos como característica da mãe, Facilidade de parto; Produção (53670) – Mérito Líquido
48	-	Leite (173141, 173244) – Conteúdo de lactose do leite, Espectro infravermelho médio do leite

*Número entre parênteses corresponde à identificação do QTL.

Apêndice F. Identificação, descrição, número de genes (número) e genes relacionados aos termos MeSH significativamente enriquecidos (p-ajustado<0,05)

Identificação	Descrição	Número	Genes
	Anatomia		
D015496	Linfócitos T CD4+	2	*BLA-DQB, ENSBTAG00000037605*
	Fenômenos e Processos		
D017951	Apresentação do antígeno	2	*BLA-DQB, ENSBTAG00000037605*
D020131	Genes duplicados	1	*BLA-DQB*
D056915	Variações do número de cópias do DNA	1	*GBP4*
	Químicos e Drogas		
D000953	Antígenos de protozoários	2	*BLA-DQB, ENSBTAG00000037605*
D006683	Antígenos HLA-DQ	2	*BLA-DQB, ENSBTAG00000037605*
D018122	Antígeno B7-1	2	*BLA-DQB, ENSBTAG00000037605*
D001425	Proteínas bacterianas da membrana externa	2	*BLA-DQB, ENSBTAG00000037605*
D051940	Antígeno B7-2	2	*BLA-DQB, ENSBTAG00000037605*
D000949	Antígenos de histocompatibilidade classe II	2	*BLA-DQB, ENSBTAG00000037605*
D059866	Cadeias beta HLA-DQ	1	*BLA-DQB*
D059848	Cadeias alfa HLA-DQ	1	*ENSBTAG00000037605*
D006684	Antígenos HLA-DR	1	*BLA-DQB*
D021382	Sinais de classificação de proteínas	1	*BLA-DQB*
D000939	Epítopos	1	*BLA-DQB*
D000911	Anticorpos Monoclonais	1	*BLA-DQB*
D019204	Proteínas de ligação GTP	1	*GBP4*