**Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”**

**Identificação e uso de CNVs na seleção genômica de vacas leiteiras criadas em
regiões tropicais**

Sofia Andrade e Silva

Dissertação apresentada para obtenção do título de
Mestra em Ciências. Área de concentração: Ciência
Animal e Pastagens

**Piracicaba
2021**

Sofia Andrade e Silva
Zootecnista

Identificação e uso de CNVs na seleção genômica de vacas leiteiras criadas em regiões tropicais

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:
Prof. Dr. **GERSON BARRETO MOURÃO**

Dissertação apresentada para obtenção do título de
Mestra em Ciências. Área de concentração: Ciência
Animal e Pastagens

Piracicaba
2021

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Silva, Sofia Andrade e

Identificação e uso de CNVs na seleção genômica de vacas leiteiras criadas em regiões tropicais / Sofia Andrade e Silva. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2021.

47 p.

Dissertação (Mestrado) - USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Gado leiteiro 2. Variação do número de cópias 3. GEBV 4. GWAS I. Título

DEDICATÓRIA

À minha mãe, Andréia Andrade, por me incentivar, me apoiar incondicionalmente nos momentos mais difíceis e por sempre acreditar em mim. Você é essencial, obrigada por me ensinar tanto sobre a vida.

Ao meu pai Waldimar Silva, por me ensinar a lutar pelos meus sonhos e me fazer acreditar que é possível alcançá-los com muito esforço e dedicação. O senhor é minha fonte de inspiração.

Aos meus irmãos Ana Clara, Isabelly e Samuel, por toda demonstração de carinho e por compreenderem a minha ausência. Mesmo longe eu prometo que “estarei sempre por perto”, pois vocês são o meu bem mais precioso.

AGRADECIMENTOS

Termina aqui mais um ciclo da minha vida. Ninguém disse que seria fácil, e não foi, mas foi graças a muita dedicação e ao apoio de diversas pessoas esse sonho se tornou realidade.

Primeiramente gostaria de agradecer ao meu orientador, Prof. Dr. Gerson Barreto Mourão, pelo apoio, conselhos, críticas e por compartilhar seus conhecimentos. Eles foram essenciais para o meu crescimento profissional e pessoal.

À Escola Superior de Agricultura “Luiz de Queiroz” pela oportunidade, e ao programa de Pós Graduação em Ciência Animal e Pastagens, que permitiu a minha dedicação exclusiva a esse projeto.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), à Comissão de Aperfeiçoamento do Nível Superior (CAPES) e à Fundação de Estudos Agrários Luiz de Queiroz (FEALQ), pelo apoio à pesquisa em distintos momentos da minha formação.

A Clínica do Leite e aos demais colaboradores pela parceria.

Às professoras Doutoras Juliana Petrini e Sandra Maria Simonelli por aceitarem fazer parte da minha banca.

Aos meus amigos e colegas, do Grupo de Estudos em Estatística e Melhoramento Animal (GEMA) por toda ajuda e troca de conhecimentos, em especial ao Fabrício Pilonetto, Giovanni Ladeira, Izally Gervásio, Paola Boscollo e Luíz Firmino.

Aos meus amigos que estiveram comigo ao longo destes anos e me ajudaram de uma forma ou de outra, em especial à Isabela, Felipe, Samira, Taynara, Débora, Patrícia e a Anna Carolina, Letícia, Leonardo, Laura.

À Deus, pelas realizações e oportunidades que me concedeu.

“As palavras têm a leveza do vento e a força da tempestade”

Victor Hugo

SUMÁRIO

RESUMO	7
ABSTRACT	8
1. INTRODUÇÃO	9
1.1. Objetivo geral.....	10
1.2. Objetivos específicos.....	11
2. REVISÃO BIBLIOGRÁFICA.....	13
2.1. Estresse térmico em bovinos	13
2.2. Variação do número de cópias (CNV)	16
2.3. Regiões de CNVs (CNVRs).....	17
2.4. Seleção genômica (GS) e Associação genômica ampla (GWAS) utilizando CNVs.....	13
3. MATERIAL E MÉTODOS	19
3.1. Origem dos dados	19
3.2. Índice de temperatura e umidade (THI).....	20
3.3. Detecção das variações do número de cópias.....	24
3.4. Detecção de CNVRs.	20
3.5. Associação genômica	21
4. RESULTADOS E DISCUSSÃO	27
4.1. Descritivas	27
4.2. Detecção de CNVs	34
4.3. Detecção de CNVRs.	27
4.4. Associação genômica	35
5. CONCLUSÃO	41
REFERÊNCIAS.....	44

RESUMO

Identificação e uso de CNVs na seleção genômica de vacas leiteiras criadas em regiões tropicais

O uso e descobertas de novas metodologias no melhoramento têm favorecido o progresso genético. No entanto, não há relatos científicos sobre a associação de variação do número de cópias (CNVs) com características influenciadas pelo estresse térmico (sob clima tropical), principalmente com enfoque genômico, o que torna indispensável o desenvolvimento de trabalhos neste segmento. Diante disso, foram utilizadas informações genotípicas de 389 vacas genotipadas com chip de média densidade (50k SNP). Os fenótipos foram ajustados aos efeitos fixos e aleatórios por meio de um modelo linear, e posteriormente foram calculados os valores genéticos genômicos. A detecção foi realizada utilizando o algoritmo PennCNV, desenvolvido com base nos modelos ocultos de Markov (HMM). Um total de 1.231 CNVs foram identificados no genoma das vacas leiteiras, posteriormente foram identificadas 219 regiões de variação do número de cópias (CNVRs), utilizadas na associação com as características de produção de leite, ácidos graxos e gordura, na população. A detecção de CNVs foi realizada por meio do algoritmo PennCNV, e as CNVRs pelo pacote do R, denominado de CNVRanger. Foram encontrados 2 CNVs significativamente associados à produção de leite. Adicionalmente, o desdobramento deste projeto continuará contribuindo para o progresso científico no estudo genômico em bovinos leiteiros.

Palavras-chave: Gado leiteiro, Variação do número de cópias, GEBV, GWAS

ABSTRACT

Identification and use of CNVs in the genomic selection of dairy cows reared in tropical regions

The use and discoveries of new methodologies in breeding have favored genetic progress. However, there are no scientific reports on the association of copy number variation (CNVs) with characteristics influenced by heat stress (under tropical climate), especially with a genomic focus, which makes the development of work in this segment indispensable. Therefore, genotypic information from 389 cows genotyped with medium density chip (50k SNP) were used. Detection was performed using the PennCNV algorithm, developed based on the hidden Markov models (HMM). The phenotypes were adjusted to fixed and random effects using a linear model, and then the genomic breeding values were calculated. A total of 1,231 CNVs were identified in the genome of dairy cows, later 219 copy number variation regions (CNVRs) were identified, used in association with the characteristics of milk production, fatty acids, fat, in the population. The detection of CNVs was performed using the PennCNV algorithm, and the CNVRs using the R package, called CNVRanger. The phenotypes were adjusted to fixed and random effects using a general linear model, and then the genomic breeding values were calculated. 2 CNVs were found significantly associated with milk production. Additionally, the development of this project will continue to contribute to scientific progress in the genomic study of dairy cattle.

Keywords: Dairy cattle, Copy number variation, GEBV, GWAS

1. INTRODUÇÃO

No decorrer dos anos houve uma intensificação no crescimento populacional, o que interfere diretamente no consumo de alimentos. Diante desse cenário, pesquisas apontam que o consumo de leite, por exemplo, aumentará em 57% até o ano de 2030 (Steinfeld and Gerber, 2010). Como as vacas da raça Holandesa são as maiores produtoras de leite no mundo, melhorar seu desempenho produtivo, reprodutivo, de crescimento e resistência a doenças é de suma importância.

Assim sendo, técnicas de melhoramento genético animal tradicional vêm sendo aplicadas no gado leiteiro, o qual é submetido à intensa seleção a fim de aumentar a produção de leite (Hayes et al., 2009). No entanto, com a disponibilidade do sequenciamento bovino em 2009 (Elsik et al., 2009), observou-se um grande avanço tecnológico que permitiu o uso da genotipagem para marcadores genéticos do tipo SNP (*Single Nucleotide Polymorphism*) e a observação de variações no número de cópias de alelos no genoma (*Copy Number Variation* – CNV), na identificação da variabilidade genética entre os animais e, conseqüentemente, na identificação de animais geneticamente superiores.

CNVs podem ser definidos como variações estruturais que produzem alterações no número de cópias de uma determinada região genômica. Essas alterações consistem em inserções, deleções, duplicações e podem variar de um quilo pares de bases (Kb) até várias megabases (Mb) (Breheny et al., 2012). Em comparação com os SNPs, os CNVs geralmente envolvem regiões genômicas maiores e têm efeitos potencialmente mais evidentes sobre a função do genoma, incluindo mudanças na estrutura e dosagem dos genes, alterando a regulação gênica e a exposição de alelos recessivos (Zhang et al., 2009). Neste caso, CNVs constituem uma fonte de diversidade e variabilidade genética populacional que tem mostrado importante papel na evolução das espécies (Carvalho and Lupski, 2016), uma vez que podem influenciar na expressão gênica (Zhang et al., 2009) e conseqüentemente modificar o fenótipo de características de interesse econômico, por exemplo, susceptibilidade a doenças e produtividade (Breheny et al., 2012; Wang et al., 2012).

O uso de novas metodologias no melhoramento em conjunto com as informações genômicas, tem auxiliado no processo de identificação de animais geneticamente superiores e mais adaptados a condições de clima e manejo. A raça Holandesa apresenta origem

européia, e por esta razão ainda é muito afetada pelo estresse térmico. Diante desse fato, como uma tentativa de manter a temperatura corporal constante em elevadas temperaturas ambientais, distúrbios fisiológicos são desencadeados nestes animais, o que contribui com a redução no desempenho produtivo, causando, por exemplo, alterações no perfil proteico e de ácidos graxos no leite (West, 2003). Para avaliar os efeitos do estresse térmico no animal, pode-se utilizar o índice de temperatura e umidade (*Temperature Humidity Index – THI*), comumente adotado em estudos desta magnitude (Hayes et al., 2003; Bernabucci et al., 2014). A partir desta metodologia, é possível identificar as variações do fenótipo em relação ao THI, por exemplo (Hayes et al., 2009; Nguyen et al., 2017). No entanto, técnicas mais específicas como a identificação de CNVs podem explicar regiões no genoma que estejam associadas a características influenciadas pela presença ou ausência de estresse térmico.

Assim sendo, esse estudo contribui na identificação de alterações no número de cópias nos genomas das vacas leiteiras que estejam diretamente associadas às características genéticas de interesse econômico na bovinocultura leiteira. Adicionalmente, identificar variantes estruturais, associadas às características de interesse econômico e prever os valores genéticos genômicos dos animais é uma ferramenta que pode possibilitar compreender melhor a estrutura genética dessas características, assim como a sua inclusão nos programas de melhoramento pode promover o progresso genético, a favor da necessidade do produtor rural e da indústria.

1.1. Objetivo geral

Identificar alterações no número de cópias (CNV) com base em informações genômicas de bovinos leiteiros da raça Holandesa e estudar a associação genética entre os CNVs para as características de produção leite, ácidos graxos saturados, ácidos graxos saturados, proteína e gordura, em ambientes com e sem estresse térmico, fundamentado pelo THI. Também prever o valor genético genômico dos animais, baseado nas informações anteriores.

1.2. Objetivos específicos

- (a) Predizer os valores genéticos genômicos dos indivíduos que apresentam CNVs associados às características responsivas a presença e ausência de estresse térmico.
- (b) Identificar as CNVs no genoma bovino das vacas leiteiras.
- (c) Identificar as sobreposições de CNVs, ou seja, identificar CNVRs (Regiões de CNVs).
- (d) Associar os CNVs às características de interesse: produção de leite, ácidos graxos saturados, ácidos graxos insaturados, proteína e gordura, na população na presença e ausência de estresse térmico.

2. REVISÃO BIBLIOGRÁFICA

2.1. Estresse térmico em bovinos

O clima é um dos principais fatores que afetam o desempenho fenotípico em bovinos leiteiros, principalmente temperaturas acima de 24°C podem causar decréscimos, economicamente importantes, nos índices produtivos e reprodutivos dos animais (Renaudeau et al., 2012). Isto ocorre devido à necessidade dos animais em manter o equilíbrio térmico corporal, que desencadeia respostas fisiológicas que são refletidas negativamente no desempenho do indivíduo (West, 2003; Bernabucci et al., 2014). Diante disso, é utilizado de modo geral o índice de temperatura e umidade (THI) para estudar o estresse térmico em animais de produção (Ravagnolo et al., 2000; Hayes et al., 2003; Bernabucci et al., 2014; Hammami et al., 2015; Santana et al., 2016).

Hayes et al. (2003) identificaram em bovinos leiteiros da raça Holandesa, uma interação com correlações de 0,90 entre genótipo e ambiente com o THI para as características de rendimento de proteína e gordura. Adicionalmente, Ravagnolo et al. (2000), mostraram que a produção de leite também é afetada na medida em que aumenta a temperatura do ambiente, os autores identificaram que, em THI maiores que 72, houve decréscimo de 0,2 kg de produção de leite, no total de 263 kg de rendimento no dia do teste, e 0,012 de gordura e 0,008 kg de proteína (0,92 e 0,85 kg de rendimento no dia do teste, respectivamente). Estes resultados indicam que a inclusão do THI na seleção dos animais, é um importante fator a ser considerado em regiões que apresentam elevada temperatura (Ravagnolo et al., 2000). Em estudo com bovinos da raça Holandesa sob clima tropical, Santana et al. (2016) apontam que os melhores animais, aqueles superiores geneticamente para produção de leite, sob conforto térmico (THI < 66), não são, necessariamente, aqueles com o melhor desempenho fenotípico em THI acima de 66 (condição de estresse térmico).

Os efeitos da variação climática, observada no fenótipo dos animais, podem servir como bioindicador no estresse térmico. Hammami et al. (2015), identificaram aumento na produção de alguns ácidos graxos, principalmente c18: 1 cis-9, e o escore de células somáticas, em condições de elevado THI (acima de 62 – condição de estresse apontada pelos autores), e devido a variância entre parâmetros em relação ao THI, indicam a possibilidade

de utilizar tais fenótipos como indicador de estresse em bovinos leiteiros. Li et al. (2017), em estudos in vitro com células epiteliais da glândula mamária em bovinos, apontaram efeito do estresse térmico na síntese dos constituintes do leite, que reforça a possibilidade de indicar a condição de estresse térmico por meio das informações fenotípicas dos animais.

2.2. Variação do número de cópias (CNV)

CNVs são fontes de variações genéticas estruturais conhecidas em um segmento de DNA no genoma de um indivíduo, como uma ou mais deleções ou duplicações, que resulta em um número diferente de cópias de um único polimorfismo (SNP; Breheny et al., 2012), os quais podem alterar o nível de expressão gênica. Uma representação destas variações está na figura 1.

(Adaptado de Dierssen et al., 2009)

Figura 1. Ilustração das classes de variantes estruturais, as quais, geralmente, se referem às alterações genômicas maiores que 1 Kb de comprimento. O desenho esquemático ilustra deleções, inserções e duplicações quando comparada com genoma de referência.

Estudos indicam que os CNVs aparecem preferencialmente em regiões de sequências repetitivas e podem ser transmitidos entre gerações (Conrad, et al. 2010). Porém, enquanto alguns aparecem de maneira recorrente, os mais raros podem ser patogênicos, ou seja, estarem associados diretamente ou indiretamente a doenças (Beckmann et al., 2007;

Stankiewicz and Lupski, 2010); Stankiewicz, 2015). Diante disso, estudos da análise e caracterização sistemática das CNVs é uma importante ferramenta para o entendimento da herdabilidade de características complexas e alterações fenotípicas (Liu et al., 2013), uma vez que associações do tipo SNP não explicam toda a herdabilidade estimada por métodos tradicionais.

Diferentes metodologias podem ser utilizadas na identificação de CNVs, sendo as principais: o *array*, genotipagem e sequenciamento de DNA (Xu et al., 2013). Porém, apesar do crescimento do uso de sequenciamento, a metodologia por genotipagem comparado a outras tecnologias detecta uma vasta informação, devido à presença de milhares de SNPs no chip, proporcionar boa cobertura do genoma e apresentarem baixo custo quando comparada a outras metodologias.

Os arranjos com milhares de SNPs é um método que detecta as variações no número de cópias (5-10 kb), e proporciona para os dois alelos o valor normalizado da intensidade total do sinal de hibridização (relação Log R - LRR) e a relação das intensidades alélicas (frequência do alelo B – BAF). Essas informações permitem que os CNVs sejam investigados tanto na produção animal quanto na agricultura.

Com a descoberta da genotipagem, vários algoritmos foram capazes de detectar CNVs no genoma bovino diante da intensidade do sinal fluorescente dos arranjos de SNPs, sendo o mais utilizado o software PennCNV (Wang et al., 2007) e em seguida o QuantiSNP (Colella et al., 2007). No PennCNV, a detecção de uma CNV é determinada pela comparação dos sinais das amostras em relação à sequência de DNA de referência. Como quando há diminuição dos valores de LRR e ausência de heterose de BAF ou quando ocorre um aumento nos valores de LRR e uma separação do genótipo heterozigoto em dois grupos, conferindo a deleção de CNVs.

Embora estudos de CNVs ainda sejam pouco explorados, em bovinos, foram identificados vários CNVs usando abordagens como genotipagem, sequenciamento e hibridização genômica (Hou et al., 2012; Upadhyay et al., 2017). Devido à variedade de abordagens, Jiang et al. (2013) apontam grande variação no número e tamanho de CNVs em bovinos da raça Holandês, como por exemplo, para características de produção, ângulo de garupa e rendimento de gordura.

Segundo Jiang et al. (2013), as possíveis inconsistências entre os resultados de CNVs podem acontecer devido a diferenças no tamanho e na estrutura das populações estudadas,

plataforma e algoritmo para detecção das CNVs. Este fato indica que os CNVs oriundos de painel de 50K (aproximadamente ~50.000 SNP) são muito maiores do que aqueles identificados por *aCGH* (*Array Comparative Genomic Hybridization*) painel de SNP 777K (aproximadamente ~777.000 SNP), ou sequenciamento, devido à maior distância entre as sondas no painel de genotipagem. Embora o *array* CGC apresenta maior sensibilidade para detectar pequenas diferenças no número de cópias, em vacas leiteiras, vários estudos têm sido realizados para detectar CNVs associados a fenótipos usando matrizes de SNP.

Em suma, a descoberta dos CNVs no genoma de bovinos é uma fonte de informação que poderia explicar variações em características importantes (Breheny et al., 2012). Esse fato permite explorar mais informações do genoma, ou seja, fontes de variabilidade, a fim de explicar o componente genético de vários fenômenos de interesse econômico, uma vez que a frequência que ocorre no genoma, tamanhos, localizações e processos de hereditariedade ainda é pouco compreendida.

2.3. Regiões de CNVs (CNVRs)

CNVRs são denominadas como agrupamento e sobreposições dos CNVs de pelo menos um par de bases. Diante disso, podem apresentar uma maior confiabilidade comparada às CNVs identificadas individualmente, pois podem detectar resultados falso-positivos uma vez que as CNVs apresentam uma maior taxa de ruído na intensidade do sinal (Cassese et al., 2014). Além disso, a associação de CNVRs com fenótipos, tem sido utilizada em estudos com o objetivo de entender como a expressão das características se comporta.

As ferramentas mais utilizadas para identificação das CNVRs são o pacote CNVRanger do software R/Bioconductor (REF) e o software CNVRuler (REF), aos quais após a detecção, as CNVRs são classificadas em ganho, perda ou ambos os acontecimentos, de acordo com a variação do número de cópias.

Embora possam apresentar maior confiabilidade, pesquisadores também utilizam as CNVs na associação de características complexas de valor econômico, por exemplo, para características de produtividade, na tentativa de entender melhor os processos biológicos envolvidos na expressão das características fenotípicas, uma vez que estudos dessa magnitude ainda não estão bem definidos quanto à associação ampla em SNPs (Xu et al., 2013). Adicionalmente, Chud (Chud, 2018) em um estudo com CNVs em bovinos leiteiros

identificou 2.611 CNVs nos cromossomos autossômicos em 961 fêmeas da raça Girolando, em seguida encontrou 668 CNVRs, sendo 278 deleções, 329 duplicações e 61 regiões mistas, com variação de 36,06 a 4.629,02 kb com média e mediana de 184,50 kb e 132, 48 kb, respectivamente, sendo esses valores correspondentes a 4% do genoma bovino.

Assim sendo, tanto as CNVs quanto as CNVRs podem ser utilizadas na pré-seleção de indivíduos por meio de regiões específicas relacionadas à característica de interesse. Entretanto, para uma melhor confiabilidade, além das análises de associação ampla ao fenótipo deve se fazer uma PCR em tempo real (Da Silva et al., 2016).

2.4. Seleção genômica (GS) e Associação genômica ampla (GWAS) utilizando CNVs

A seleção genômica (GS, *Genomic Selection*) é uma ferramenta inovadora proposta inicialmente por Meuwissen et al. (Meuwissen et al., 2001), utilizada usualmente para estimar os valores genéticos genômicos dos indivíduos e para aumentar a acurácia de seleção com a utilização de painéis de marcadores moleculares do tipo SNP. Desse modo, os valores genéticos esperados são calculados com base nos efeitos dos QTLs (*Quantitative Trait Loci*) utilizando equações de predições baseadas na associação de dados fenotípicos e genotípicos de uma população referência (Goddard et al., 2010).

Embora métodos tradicionais tenham sido eficientes, o uso da seleção genômica em bovinos leiteiros permite a realização de estudos de associação genômica ampla com características mais complexas (GWAS, *Genome-Wide Association Study*) a fim de identificar correlações genótipo-fenótipo. A princípio a seleção genômica era usualmente mais utilizada na seleção de touros, no entanto, estudos implementados para seleção de fêmeas é uma realidade (Weigel et al., 2012). Pois as informações de identificação de fêmeas com valores genéticos genômicos superiores pode influenciar as decisões de seleção permitindo economizar com a criação de novilhas não promissoras, devido as informações de rastreamento do pedigree, determinação da composição racial, além da participação do progresso genético, mesmo que em menor escala. Dessa forma, o uso de genômica em vacas leiteiras possibilita identificar regiões do genoma e genes candidatos que estejam associados a alguma característica economicamente importante, como, produtividade, longevidade e fertilidade (Pryce and Daetwyler, 2012).

O estudo de associação com características fenotípicas dos animais, representa um papel importante na determinação da expressão das características (Rendon, et al., 2006), devido ao entendimento da associação dos marcadores moleculares com valores genéticos genômicos (Silva et al., 2019). Segundo Breheny et al. (2012) e Silva et al. (2019) após estimar o CNVs presentes em todos os segmentos do genoma do indivíduo é possível realizar o teste de associação genética com cada segmento para o qual existe a variabilidade no número de cópias e identificar marcadores genéticos relacionados a características fenotípicas. Desse modo, pesquisas apontam que a utilização de CNVs tem alto poder de detectar associações em comparação com a utilização de SNPs, especialmente quando envolvem segmentos longos e em baixa frequência.

Cada CNV pode ser analisado por meio da associação com cada característica, separadamente (Alkan et al., 2011). Xu, et al. (2014) utilizando dados de chip de SNPs de 50k, detectou CNVs significativamente associados à produção de leite, produção de gordura e proteína. Diante disso, conclui-se que essas metodologias genômicas são ferramentas importantes para identificar regiões no genoma importantes na determinação de expressão das características, bem com o entendimento dos processos biológicos que confere melhorias da produção das vacas leiteiras (Nicolini et al., 2018; Wang et al., 2019).

3. MATERIAL E MÉTODOS

3.1. Origem dos dados

Foram utilizadas informações fenotípicas de produção e qualidade do leite, obtidas com a colaboração da Clínica do Leite, do Departamento de Zootecnia da Escola Superior de Agricultura “Luiz de Queiroz”, da Universidade de São Paulo.

Os dados de produção mensais das vacas, foram advindos de quatro fazendas brasileiras, entre os meses de maio de 2012 e dezembro de 2016. As fazendas apresentam sistema *free-stall*, onde os animais são ordenhados em sistema mecanizado três vezes ao dia. Nestes rebanhos, a alimentação das vacas é principalmente composta por silagem de milho, feno de capim, caroço de algodão, farelo de milho e soja, casca de soja, polpa cítrica, vitaminas e minerais.

Três das fazendas estão localizadas no estado de São Paulo, região caracterizada por temperaturas médias de mínima e máxima de 12 a 28 °C e precipitação média mensal de 40 a 237 mm. A quarta fazenda está localizada no estado do Paraná, que tem temperaturas médias de mínima e máxima entre 8 e 27 °C, e precipitação média mensal de 150 mm.

O banco de dados era composto por 127.430 informações fenotípicas de desempenho em número total de 7.660 vacas mensuradas, sendo que 768 foram genotipadas em painel de baixa densidade (Illumina Bovine LD BeadChip, 6K), com 6.909 SNPs, e 388 vacas genotipadas em painel de média densidade, composto por 47.843 SNPs (geneSeek Genomic Profile Bovine 50K). Sendo estas referentes a informações individuais de data de nascimento, data de parto, data de coleta da amostra de leite, dias em lactação no controle leiteiro e ordem da lactação, juntamente com as mensurações fenotípicas de produção de leite (PROD, em kg/dia), contagem de células somáticas (CCS), teor de gordura (GORD, %), de proteína (PROT, %), de lactose (LACT, %), de caseína (CAS, %) e conteúdo de ácidos graxos no leite (saturados, AGSAT; insaturados, AGINSAT; monoinsaturados, MONO; poli-insaturados, POLI; ácido graxo palmítico, C16:0; esteárico, C18:0; e oléico, C18:1).

Os componentes do leite foram mensurados por meio de espectrometria em infravermelho médio (Delta Instruments CombiScope™ Filter, Advanced Instruments, Inc., Norwood, EUA), utilizando as amostras de leite refrigeradas e conservadas com bronopol, conforme descrito por Rodriguez et al. (2014).

Os grupos de contemporâneos foram formados concatenando-se fazenda, data de parto e da coleta das amostras de leite, destes, foram eliminados animais pertencentes a grupos de contemporâneos com menos de cinco animais. Além disso, foi realizada uma análise de variância, ANOVA, foi por meio do *software R* versão 4.1.0 para testar o nível de significância das variáveis incluídas nos grupos formados.

A fim de eliminar informações inconsistentes, foram excluídos do banco de dados os registros de animais: com nenhum dado mensurado, sem data de nascimento, sem data da coleta da amostra de leite, sem data de parto, com idade superior a 10 anos, sem ordem de lactação ou com esta superior a 7, com dias em lactação no controle leiteiro (DEL) menor que cinco e maior que 365 dias.

As informações de *pedigree* são referentes a 16.632 animais, que incluíam 903 touros e 8.242 reprodutoras de 7,28 gerações, considerando a primeira geração como um (Brinks et al., 1961).

Abaixo, a tabela 1 mostra os valores mínimos e máximos para a manutenção da medida no banco de dados para as características estudadas.

Tabela 1. Mínimo e máximo aceitável para a manutenção da medida no banco de dados para produção de leite (PROD), teor de proteína (PROT), de gordura (GOR), de proteína (PROT), ácido graxo saturado (AGsat) e ácido graxo insaturado (AGinsat).

Característica	Min	Máx
PROD (kg/dia)	5	68,4
PROT (%)	1,77	5,83
GORD (%)	1,01	11,92
AGsat (%)	0,71	7,16
AGinsat (%)	0,02	4,22

3.1. Índice de temperatura e umidade (THI)

O THI é uma importante ferramenta utilizada para avaliar o estresse térmico dos animais. Assim sendo, foi calculado o valor de THI para as 127.430 informações fenotípicas contida no banco de dados, que posteriormente foram separadas em THI maiores que 72

(indicativo de estresse térmico) e THI menores que 72 (indicativo de conforto térmico) (Ravagnolo et al., 2000; Hayes et al., 2003).

Uma vez que as fazendas não mensuram informações meteorológicas locais, foram utilizados os dados de temperatura e umidade preditas e disponíveis na base de dados da NASA (*Prediction of Worldwide Energy Resource - Power*, 2017) de acordo com as informações de latitude e longitude de cada fazenda incluída no estudo. Com base em evidências que THI de dois dias que antecedem o controle leiteiro afetam a produção (Santana et al., 2016), foi realizado o cálculo da média da umidade relativa e temperatura deste período para cada data de coleta. Adicionalmente, estas informações foram utilizadas na equação descrita por *National Research Council* (NRC, 2017).

$$\text{THI} = (1,8 \times \text{temp} + 32) - [(0,55 - 0,0055 \times \text{rh}) \times (1,8 \times \text{temp} - 26)]$$

onde **temp** é a temperatura em °C, e **rh** é a umidade relativa em %.

3.2. Predição dos valores genéticos genômicos

Foram preditos os valores genéticos genômicos (*Genomic Estimated Breeding Value - GEBV*) para cada característica estudada: de produção leite (PROD, kg/dia), ácidos graxos saturados (AGsat, %), ácidos graxos saturados (AGinsat, %), proteína (PROT, %) e gordura (GORD, %), foi realizada a associação com os CNVs identificados.

Para calcular os GEBVs, foram 127.430 informações fenotípicas de desempenho de, em número total de 7.660 vacas, genotipadas, apresentando 16.632 informações de *pedegree*. Após calculado dos GEBVs, foi realizada a identificação dos animais que apresentaram CNVs identificadas.

Os GEBVs de cada animal e para cada característica, foram baseados nas estimativas do *single-step GBLUP* (*ssBLUP*), sendo estes separados em dois grupos: GEBVs dos animais em que o THI da característica se apresenta maior que 72 e animais que apresentavam THI da característica menor que 72. Diante disso, objetivou-se identificar os indivíduos que apresentam maior potencial genético para cada característica em diferentes ambientes.

Para estimar os componentes de (co)variância, foi utilizado o método de máxima verossimilhança restrita, no software AIRMLF90 (Misztal et al., 2018). A matriz de relacionamento genômico **G** foi utilizada seguindo um modelo genético univariado ajustado:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \boldsymbol{\varepsilon}$$

em que $\mathbf{y}$ é o vetor das características analisadas; $\mathbf{b}$ é o vetor de soluções do efeito fixo; $\mathbf{a}$ é o vetor dos efeitos fixos; $\mathbf{X}$, $\mathbf{Z}$, são as matrizes de incidência dos efeitos fixos e aleatórios, respectivamente; $\boldsymbol{\varepsilon}$ é o vetor de resíduos aleatórios assumindo $\boldsymbol{\varepsilon} \sim \mathbf{N}(0, I\sigma^2)$.

No programa BLUPF90, a matriz $\mathbf{A}$ de relacionamento aditivo de pedigree é combinada com a matriz $\mathbf{G}$ para criar uma matriz única de relacionamento genômico $\mathbf{H}$ (Legarra, et al., 2009; Aguilar et al., 2010).

Assim, a inversa de $\mathbf{H}^{-1}$ pode ser expressa como:

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}_w^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

em que $\mathbf{G}_w = w\mathbf{G} + (1 - w)\mathbf{A}_{22}$; $\mathbf{G}_w$ é a matriz final dos relacionamentos genômicos; w o peso dos marcadores moleculares, que representa a proporção da variância genética aditiva dos SNPs; $\mathbf{A}_{22}$ uma submatriz de $\mathbf{A}$ para os animais genotipados; $\mathbf{G}$ a matriz de relacionamento genômico.

Foram excluídos os marcadores com *call-rate* menor que 90%, frequência alélica menor que 0,05 (*MAF* - *minor allele frequency* *MAF*) e os que apresentavam frequência de heterozigotos maior que 0,15 do que o esperado pelo equilíbrio de Hardy-Weinberg. Esse controle de qualidade foi realizado no *default* do software AIREMLF90. Diante disso, restaram aproximadamente 47843 SNPs. Esse controle de qualidade foi realizado no *default* do software AIREMLF90, por meio do programa BLUPF90 (Misztal et al., 2018).

3.3. Detecção das variações do número de cópias

Para a detecção de CNVs foram utilizadas informações genotípicas de 388 vacas genotipadas em painel de média densidade (*GeneSeek Genomic Profiler Bovine 50k*), composto por 47.843 SNPs (Iung et al., 2019), no qual todos os animais apresentavam informações de genealogia, data de nascimento, data de parto, data de coleta de amostra de leite, e dias em lactação.

As análises de detecção de CNVs foram realizadas utilizando o algoritmo PennCNV, conforme descrito por Wang et al. (2007). Basicamente, este algoritmo foi desenvolvido com base nos modelos ocultos de Markov (HMM), que é amplamente usado para detectar CNVs com base na matriz de SNP devido à sua taxa de falso positivo ser relativamente baixa (Dellinger et al., 2010). Desse modo, o PennCNV tem como função avaliar a combinação de LRR (valor normalizado da intensidade total do sinal de hibridização) e o BAF (relação das intensidades alélicas) em cada marcador SNP para inferir as alterações no número de cópias de fragmentos do genoma, sendo estas informações geradas a partir das plataformas de genotipagem e derivadas da montagem da sequência do genoma UMD3.1 bovino.

Os primeiros passos para obtenção dos arquivos de entrada foi a extração dos valores de LRR e BAF dos arquivos de genotipagem, os quais foram desmembrados em arquivo separados por amostra, contendo as informações do nome do SNP (*Name*), identificação da amostra (*Sample ID*), LRR e BAF, a fim de gerar o arquivo de PFB (frequência alélica de B da população). Diante disso, o PFB foi estimado a partir dos valores de BAF individuais do marcador, por meio do executável *compile_pfb.pl*, disponível no *software* PennCNV (Wang et al., 2007).

De acordo com os valores de LRR e BAF é possível identificar as mudanças no número de cópias no genoma. Abaixo, a tabela 2 mostra um exemplo dos valores de LRR, variando de -0,5 a 0,3 o que confere a identificação da deleção ou inserção do número de cópias no genoma, e as possibilidades genotípicas variam em Deleção de dois alelos (DD), Deleção de uma cópia (A, B), Normal (AA, AB, BB), Duplicação Simples (AAA, AAB, ABB, BBB) de acordo com os valores de BAF.

Tabela 2. Classificação dos genótipos.

Genótipos	Números de cópias	LRR
DD	0	-0,5
A,B	1	-0,45
AA, AB, BB	2	0
AAA, AAB, ABB, BBB	3	0,3

Fonte: (Illumina, 2017)

Para normalizar o LRR de cada SNP, foi usado o arquivo de correção GC padrão (*GC Reference bos_taurus_UMD3.1.gc_degest.dsff*) por meio do executável *cal_gc_snp.pl*, uma vez que a proporção de GC em torno de cada marcador pode ter influência sobre a força do sinal, criando as chamadas ondas genômicas (Disk et al., 2008). Adicionalmente, a correção do conteúdo de GC do segmento genômico foi realizada em torno de 500kb de cada lado do marcador.

Para as análises de detecção foram considerados os arquivos gerados anteriormente, por meio do executável *detect_cnv.pl*, contido no PennCNV. A fim de reduzir os resultados falso-positivos, alguns critérios de controle de qualidade padrão foram aplicados nas CNVs detectadas. Sendo assim, foram consideradas apenas CNVs com desvio padrão de LRR (>0.3), BAF *drift* (>0.1), fator de dispersão (“*waviness factor*”) (>0.05) e CNVs com pelo menos três SNPs (Chud, 2018). Foram excluídas das análises CNVs com tamanhos maiores que 1 kilobase (kb), assim como os cromossomos sexuais (Quinlan, H., 2010; Silva, V. H., 2015). Posteriormente, as CNVs foram utilizadas na associação com cada característica, ou seja, de forma individual (Alkan et al., 2011).

3.4. Detecção de CNVRs

A identificação das regiões de CNVs (CNVRs) foi realizada pelo método de agrupamento da sobreposição recíproca de no mínimo um par de bases utilizado a função *PopulaçãoRanges* (*grl*, densidade = 0,1) do pacote *CNVRanger* do *software* R versão 4.1.0 (Silva et al., 2019), no qual foram filtradas apenas as CNVRs significativas ($p < 0.05$). Diante disso, as CNVRs foram classificadas em deleções, em que a amostra apresentou uma região com perda do segmento cromossômico, duplicações em que ocorreu uma duplicação da região, e mistas, ou seja, onde alguns indivíduos apresentaram deleções e outras duplicações na mesma região cromossômica do DNA.

3.1. Associação genômica

As análises de associação dos CNVRs com os GEBVs foram realizadas utilizando o pacote *CNVRanger R/Bioconductor software* R versão 4.1.0 (Silva, V., 2019), o qual realiza

uma regressão linear padrão de fenótipo. Diante disso, CNVs significativas foram contadas no nível de significância de (p-valor $<0,05$) após correção de FDR.

4. RESULTADOS E DISCUSSÃO

4.1. Descritivas

4.1.2 Fenótipos

Após a edição do banco de dados, restaram 93.449 informações fenotípicas correspondentes a 6.485 vacas. A estatística descritiva das características avaliadas são apresentadas na tabela 3.

Tabela 3. Número de observações (N), média, desvio-padrão (DP), valores mínimos (MIN) e máximos (MAX) obtidos para produção de leite (PROD), teor de proteína (PROT), de gordura (GOR), de proteína (PROT), ácido graxo saturado (AGsat) e ácido graxo insaturado (AGinsat).

Característica	Média	DP	MIN	MAX
PROD (kg/dia)	35,9723	10,7108	5,00	68,4
PROT (%)	3,1294	0,3165	1,77	5,83
GOR (%)	3,5134	0,7941	1,01	11,92
AGsat (%)	1,83	0,527	0,71	7,16
AGinsat (%)	0,9866	0,3165	0,02	4,22

Os valores de média e desvio-padrão para todas as características são semelhantes aos reportados por Petrini et al. (2016) e Carrara et al. (2018), em que se avaliou animais da mesma população.

4.2.3 THI

Os valores para THI variaram de 54,911 a 78,326, representando valores médios de 66,6189. A descrição estatística dos THI separados em THI maior que 72, THI menor que 72, estão apresentados na Tabela 4.

Tabela 4. Número de observações (N), média, desvio-padrão (DP), coeficiente de variação (CV, %), valores mínimos (MIN) e máximos (MAX) obtidos para THI considerados para produção de leite (THIprod), teor de gordura (THIlgord), de proteína (THIprot), ácidos graxos saturados (THIagsat), insaturados (THIaginsat) em THI superior a 72.

Característica	Média	DP	CV	MIN	MAX
THIprod	74,245	1,6862	2,8433	72,350	78,326
THIprot	73,980	1,5772	2,4878	72,102	78,326
THIlgord	74,230	1,7337	2,4878	72,102	77,271
THIagsat	74,012	1,7337	3,0057	72,102	77,271
THIaginsat	74,012	1,7337	3,0057	72,102	77,271

Com a queda na produção do leite relatada na época do calor e de acordo com estudos, Thon (1959) propôs que THI maiores que 72 conferem estresse térmico aos animais. Baseados na combinação de que os fatores ambientais, tais como temperatura e umidade do ar podem causar o estresse calórico, investir no controle ambiental por meio do uso de ventiladores e aspersores pode ser uma alternativa.

Pereira (2018) trabalhando com animais em *free stall* sem ventiladores e aspersores, observou que o THI apresentava valores médios de acima de 72 em 37% do tempo, sugerindo que esse ambiente sem o uso dessas tecnologias pode ocasionar uma condição de indução de estresse por calor. Diante disso, do ponto que o impacto do calor na produtividade em vacas leiteiras de regiões tropicais é realidade, estudos para entender e justificar o investimento econômico dessas tecnologias de controle ambiental se fazem necessários.

Tabela 5. Número de observações (N), média, desvio-padrão (DP), coeficiente de variação (CV, %), valores mínimos (MIN) e máximos (MAX) obtidos para THI considerados para produção de leite (THIprod), teor de gordura (THIlgord), de proteína (THIprot), ácidos graxos saturados (THIagsat), insaturados (THIaginsat) em THI inferior a 72.

Característica	Média	DP	CV	Min	Máx
THIprod	64,532	5,2086	27,1303	54,911	71,732
THIprot	64,982	4,8162	23,1963	54,911	71,732
THIlgord	64,982	4,8162	23,1963	54,911	71,732
THIagsat	66,938	3,1167	9,71400	59,405	71,732
THIaginsat	66,938	3,1167	9,71400	59,405	71,732

Os valores reportados acima, estão de acordo com Carrara et al. (2018), que utilizou informações da mesma população de estudo. Para THT para as mesmas características neste estudo, foi relatado THI variando em torno de 78,15 a 51,558. Embora apresente valores próximos, as diferenças encontradas podem ter ocorrido devido aos critérios adotados de edição dos dados.

Pereira (2021) encontrou valores de THI com magnitude variando entre 77,67 e 66,41; 80,5 e 65,2; 75,7 e 64,4; 77,4 e 65,5 em quatro fazendas com instalações do tipo *free stall*. Observa-se similaridade dos valores máximos de THI deste estudo, o que pode indicar que a utilização de informações de temperatura e umidade preditas e disponíveis na base de dados da NASA pode ser uma alternativa de uso, quando a local não possui uma estação meteorológica ou não faz a mensuração diária dentro das instalações.

4.2.4 GEBV

A descrição estatística dos GEBVs separados de acordo com a variação no THI estão apresentados na Tabela 6. Desse modo, o intuito foi avaliar os GEBVs dos animais em ambientes diferentes, ou seja, com THI maior 72 e THI menor que 72.

Tabela 6. Componentes de variâncias e coeficientes de herdabilidade estimados sob abordagem genômica por meio de análises baseadas no pedigree, para produção leite (PROD), teor de proteína (PROT), de gordura (GORD), ácidos graxos saturados (AGsat), ácidos graxos saturados (AGinsat), referentes a THI maiores que 72 (THI>72) e THI menores que 72 (THI<72).

Característica	σ_p (THI>72)	h^2 (THI>72)	σ_p (THI<72)	h^2 (THI<72)
PROD (kg/dia)	74,191	0,11671	78,446	0,15949
PROT (%)	0,06045	0,16541	0,06918	0,21210
GORD (%)	0,62791	0,22654	0,50620	0,20257
AGsat (%)	0,26810	0,23229	0,20184	0,38454
AGinsat (%)	0,77851	0,06013	0,56097	0,13839

σ_p = variância fenotípica; h^2 = coeficiente de herdabilidade.

Para o ambiente de estresse térmico, todas as características estudadas apresentaram coeficientes de herdabilidade moderado, exceto para ácido graxo insaturado ($h^2 = 0,06$). Nesse caso, quanto maior o efeito do ambiente, menos exata será a estimativa do fenótipo. Apesar de ser raramente incluídas nos programas de melhoramento genético, o perfil dos ácidos graxos no leite deve ser monitorado, uma vez que possuem efeito na saúde do consumidor.

De acordo com estudos na mesma população, foram encontrados valores de herdabilidade de 0,13; 0,19; 0,25; 0,08, para produção de leite, gordura, ácido graxo saturado e ácido graxo insaturado, respectivamente (Petrini et al., 2016). O discreto aumento das herdabilidades encontradas pelos autores, e neste estudo, pode ter ocorrido devido à diferença entre o conjunto de dados utilizado.

Autores relatam que o declínio da produção de leite devido ao aumento do estresse térmico varia entre os animais, e que a herdabilidade desta característica pode variar de baixa a moderada (Bernabucci et al., 2014). Diante disso, a variação dos resultados entre os estudos pode ser explicada pela densidade do painel de SNP utilizado, número de amostra, grupo de contemporâneos e fatores ambientais, como temperatura do ambiente e a nutrição (Boddhireddy et al., 2013; Fernandes, 2015).

O conteúdo de ácido graxo saturado apresentou coeficiente de herdabilidade alto de 0,38. No entanto, para produção de leite, proteína e gordura, foram moderadas (0,15; 0,21; 0,20; respectivamente). Estimativas de herdabilidade com valores de moderado a alto é um bom indicativo do valor genético do animal para determinada característica. Desse modo, podem ser mais facilmente selecionadas e transmitidas para a próxima geração. Além disso, estudos apontam que a disponibilidade de um grande número de informações fenotípicas e genotípicas e herdabilidade de moderada a alta, apresentam maior confiabilidade nos resultados de GEBVs.

Comparando as tabelas 6 e 7, observa-se que em condições de conforto térmico (Tabela 7), os valores referentes a variância fenotípica e herdabilidade são maiores. Porém, o aumento do valor desse parâmetro possivelmente ocorreu devido a maior proporção da variância aditiva, ou seja, em condições de conforto térmico os animais conseguiram expressar melhor o seu potencial genético, e consequentemente podem ter apresentado uma maior produtividade. Entretanto, em clima tropical, deseja-se selecionar indivíduos mais robustos e que apresentam GEBVs maiores para condições de estresse térmico. Uma possível alternativa para aumentar a adaptação dos animais, seria introduzir alelos adaptados ao estresse térmico através do cruzamento.

O conteúdo de ácido graxo saturado e insaturado apresentou variância genética aditiva extremamente baixa, variando de 0,04 a 0,07 para ambos ambientes, assim como encontrado na literatura (Lung et al., 2019; Penasa et al., 2015). Esses valores indicam que quanto menores, menos exata será a estimativa do genótipo do animal, devido ao maior o efeito do ambiente.

Como os componentes genéticos são influenciados por frequências gênicas e podem diferir de uma população para outra, a variabilidade dos resultados entre os estudos pode ser explicada pelo número de amostra, grupo de contemporâneos, fatores ambientais, como temperatura do ambiente e a nutrição, pela densidade do painel de SNP utilizado, (Boddhireddy et al., 2013; Fernandes, 2015).

4.2. Detecção das CNVs

Um total de 2474 CNVs foram detectadas inicialmente, no entanto, após o controle de qualidade restaram 1231 CNVs autossômicos, em 244 amostras. Observa-se que a exclusão dos cromossomos sexuais acarretou na perda de quase metade das CNVs, uma vez que os fragmentos eram excessivamente grandes (>1mb), foram considerados não confiáveis.

Os 1241 CNVs foram encontrados em 308 regiões diferentes ao longo do genoma, onde a média do número de CNVs identificados por amostras foi igual a 6. Adicionalmente, apresentaram comprimento mínimo de 99,701 bp (*base pairs - bp*) e máximo de 1,042,448 bp, o que corresponde a uma cobertura de 4,09% do genoma bovino (UMD_3.1). Os resultados estão de acordo com Hou et al., (2011), que utilizando plataforma de genotipagem de 50k encontrou 3666 CNVs em 521 amostras da raça Holandesa, sendo um total médio de 7,09 eventos de ganho e perda por amostra, com uma cobertura de 5,49% do genoma bovino de referência (UMD_3.1). Portanto, as proporções dos CNVs identificados estão de acordo com a literatura, o que confere que o protocolo de identificação de marcadores e o controle de qualidade realizado foram adequados.

Em outros estudos utilizando informações oriundas do chip *Bovine HD (700k)*, Upadhyay et al., (2017) relatam ter encontrado 9994 CNVs, em 149 amostras. Observa-se que os resultados desse estudo apresentam maior número de CNVs por indivíduo, o que pode ser explicado pela diferença da plataforma genômica utilizada entre os estudos. Isso acontece, pois, o *BovineSNP50* possui um limite de resolução sob os critérios do PennCNV devido à baixa densidade e distribuição não uniforme dos SNPs, explicando assim, a diferença da variação do comprimento e números de CNVs em diferentes estudos.

A figura 2, representa um exemplo escolhido de forma aleatória, da deleção de um número de cópia de CNV ocorrendo no cromossomo 14, da amostra sample.2814.txt. Essa representação é realizada e identificada pelo *software PennCNV* através da variação dos valores de LRR e BAF, em apenas uma amostra por vez. Diante disso, observa-se que os valores de LRR encontram-se predominantemente abaixo de zero e de BAF ao redor de 0 e 1, classificado este evento como uma deleção do número de cópias.

Figura 2. Representação da deleção de 1 cópia de CNV da amostra sample.2814.txt.

Gao et al. (2017) após a filtragem e aplicação do controle de qualidade dos 2.236 CNVs, encontrou 1.834 CNV em 8 touros da raça Holandesa, com base na metodologia de sequenciamento de nova geração (*NGS – Next Generations Sequencing*). Assim sendo, estes autores utilizaram o software *CNVnator* para detectar as variações ao longo do genoma, enquanto que neste estudo foi utilizado o *PennCNV*. Os dois fatores, tanto a plataforma quanto os diferentes softwares podem interferir nas respostas, principalmente o uso de distintos softwares, que podem ocasionar resultados discrepantes quando comparados. Sendo assim, segundo Breheny, et al. (2012) apesar de parecer uma boa estratégia utilizar múltiplos algoritmos para a identificação de CNVs, ela pode aumentar a taxa de falso-positivos dos resultados, conferindo resultados não confiáveis.

De forma complementar, Jiang et al. (2012) e Jiang et al. (2013) detectaram 99 e 367 CNVs, em vacas da raça Holandesa, utilizando dados de chip *Illumina Bovine 50k* e chip *BovineHD*, respectivamente. Diante deste fato, outra consideração a se fazer é sobre a diferença dos resultados quando comparados com a literatura, os quais demonstram a

existência da alta variabilidade da detecção de CNVs entre as plataformas, métodos, e tamanho da população bovina.

4.3. Detecção de CNVRs

Os CNVRs representam a sobreposição de CNVRs, em pelo menos um par de base entre as amostras deste estudo. Deste modo, dos 1231 CNVs detectados pelo PennCNV foram mapeados 219 CNVRs ($p < 0.05$) distribuídos em 22 cromossomos da população sob estudo, representando um total de 1,32% do comprimento total do genoma bovino de referência. Considerando o total de CNVRs obtidas, foram caracterizadas como regiões de inserção de números de cópias 147 (67,12%), deleção 47 (21,46%), e com ambos os eventos na mesma região (inserção e deleção) apresentando um valor de 18 (8,22%).

Na figura 3, observa-se os eventos de ganho, perda e ambos que ocorreram no cromossomo 3, mais precisamente, detectados em 25 amostras. Essas informações foram geradas pelo pacote *CNVRuler R/ Bioconductor*, de acordo com o nível de significância de $p < 0.05$.

Figura 3. Eventos de deleção, inserção e ambos CNVRs localizados no cromossomo 3 do genoma bovino e representado por linhas verticais (azuis, rosa e verde). A Barra cinza na vertical indica posição cromossômica em que está ocorrendo as CNVRs.

Os valores de CNVRs são semelhantes aos reportados por Bae et al. (2010) utilizando o método de genotipagem com painel de 50K para identificação de CNVs, no qual relataram ter encontrado 368 regiões de CNV em 265 bovinos da raça Holandesa. Do mesmo modo reportado em estudos com CNVs, a variação dos valores encontrados na literatura para

regiões de CNVs pode ser explicado por um viés que ocorre devido a diferença do número e posições existentes entre os SNPs e painéis de genotipagem, além de fatores como: seleção natural da população, seleção recente e a hereditariedade dessas informações.

Adicionalmente, Prado (2019), relatou ter encontrado 318 CNVRs em 418 animais, foram agrupadas em 203 CNVRs, sendo 100 deleções e 15 duplicações concordantes. Também utilizando marcadores moleculares para identificação de CNVs. Diante disso, em geral, o número de CNVs identificados em cada amostra é consistente com as estimativas de SNPs encontrados na raça Holandesa.

Realizando o teste de permutação de sobreposição pelo pacote *regioner* *R/Bioconductor*, foram detectadas das 219 regiões de CNVs, 14 CNVRs que se sobrepõem a regiões codificadoras de proteína ($p < 0.05$). Essas regiões são altamente conservadas, desse modo, a detecção desses CNVs explica a importância e influência do CNV sob a expressão fenotípica.

Figura 4. Sobreposição de regiões de CNVs para proteína do leite.

4.4. Associação genômica

A identificação de CNVs ao longo do genoma bovino e a associação de CNVRs com características de produtividade em bovinos da raça Holandesa podem auxiliar no entendimento da arquitetura biológica dessas características complexas. Além de

possibilitarem o desenvolvimento de novas estratégias de seleção. Desse modo, neste estudo foi realizada uma análise de associação do genoma entre as características de desempenho e 308 segmentos de CNVs, usando o pacote CNVRanger R/ Bioconductor. Diante disso, gráficos plot Manhattan para segmentos de CNV nos 29 cromossomos, associados às características de interesse do estudo, em ambiente com THI superior e inferior a 72, foram representados nas Figuras 5 e 6, respectivamente.

(a)

(b)

(c)

(d)

(e)

Figura 5. Gráfico plot Manhattan para segmentos de CNV nos 29 cromossomos, associados à (a) gordura (GORD, %), (b) ácidos graxos insaturados (AG,sat, %), (c) produção de leite (PROD, Kg/dia), (d) proteína (PROT, %), e (e) ácidos graxos saturados (AGinsat, %), em ambiente com THI superior a 72. O eixo x representa os cromossomos somáticos e o eixo Y mostra o valor $-\log_{10}(p)$ correspondente. As linhas vermelhas e azuis indicam valores p corrigidos por FDR de 0,05 e 0,1, respectivamente.

(a)

(b)

(c)

Figura 6. Gráfico plot Manhattan para segmentos de CNV nos 29 cromossomos, associados à (a) gordura (GORD, %), (b) ácidos graxos insaturados (AG,sat, %), (c) produção de leite (PROD, Kg/dia), (d) proteína (PROT, %), e (e) ácidos graxos saturados (AGinsat, %), em ambiente com THI inferior a 72. O eixo x representa os cromossomos somáticos e o eixo Y mostra o valor $-\log_{10}(p)$ correspondente. As linhas vermelhas e azuis indicam valores p corrigidos por FDR de 0,05 e 0,1, respectivamente.

Cada ponto representa um segmento de CNVs variando entre deleção e inserção. Foram observados dois segmentos de CNVs significativamente associados à PROD, que se encontram sobrepostos a QTLs nos cromossomos 2 e 13. O mesmo comportamento para o CNV foi observado pelos autores Cheere et al.2019.

Xu et al. 2014, também utilizando genótipos de painel de 50k encontrou 29 CNVs associadas a produção de leite e porcentagem de gordura no leite em bovinos da raça Holandesa. Destes, 11 foram significativamente associados à produção de leite, sendo o de maior significância localizado no Chr17: 73.055.503-75.058.715. Além disso, no mesmo estudo foram detectados 29 CNVs. Da mesma forma, observa-se que a maioria dos CNVs se

sobrepõe a QTL conhecidos, o que demonstra que as CNVs podem alterar a função dos genes e, portanto, alterar os fenótipos.

Segundo Bickhart et. (2012), a herdabilidade dos CNVs pode estar relacionada a características economicamente importantes. Desse modo, observa-se que o uso da informação genômica na seleção é uma importante ferramenta para entender o comportamento expressivo dos fenótipos, principalmente em ambiente de clima temperado, onde os estudos dessa magnitude ainda são escassos.

5. CONCLUSÃO

Foram identificadas 1234 variações estruturais ao longo do genoma bovino, apresentando 308 segmentos de CNVs diferentes. Dentre elas, conclui-se que são importantes e podem estar associadas a características complexas, principalmente porque têm sido localizadas em regiões conservadas do genoma, como no caso da produção de leite. Além disso, o uso da associação genômica de CNVs e GEBVs pode contribuir para a identificação de marcadores associados ao fenótipo em ambientes de conforto e estresse térmico em bovinos leiteiros, assim como os SNPs.

REFERÊNCIAS

- Aguilar, I, Misztal, I.; Jhonson, D. L.; Legarra, A.; Tsuruta, S.; Lawlor, T. J. 2010. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *Journal of Dairy Science*. 93:743-75.
- Alkan, C., B.P. Coe, and E.E. Eichler. 2011. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* 12:363–376. doi:10.1038/nrg2958.
- Bae, J.S., H.S. Cheong, L.H. Kim, S. NamGung, T.J. Park, J.Y. Chun, J.Y. Kim, C.F.A. Pasaje, J.S. Lee, and H.D. Shin. 2010. Identification of copy number variations and common deletion polymorphisms in cattle. *BMC Genomics* 11. doi:10.1186/1471-2164-11-232.
- Beckmann, J.S., X. Estivill, and S.E. Antonarakis. 2007. Copy number variants and genetic traits: Closer to the resolution of phenotypic to genotypic variability. *Nat. Rev. Genet.* 8:639–646. doi:10.1038/nrg2149.
- Bernabucci, U., S. Biffani, L. Buggiotti, A. Vitali, N. Lacetera, and A. Nardone. 2014. The effects of heat stress in Italian Holstein dairy cattle. *J. Dairy Sci.* 97:471–486. doi:10.3168/jds.2013-6611.
- 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.
- Breheny, P., P. Chalise, A. Batzler, L. Wang, and B.L. Fridley. 2012. Genetic association studies of copy-number variation: Should assignment of copy number states precede testing?. *PLoS One* 7:1–9. doi:10.1371/journal.pone.0034262.
- Carvalho, C.M.B., and J.R. Lupski. 2016. Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.* 17:224–238. doi:10.1038/nrg.2015.25.
- Cassese, A., M. Guindani, M.G. Tadesse, F. Falciani, and M. Vannucci. 2014. A hierarchical bayesian model for inference of copy number variants and their association to gene expression. *Ann. Appl. Stat.* 8:148–175. doi:10.1214/13-AOAS705.
- Chud, T.C.S. 2018. Identificação de regiões com variações no número de cópias dos segmentos de DNA em bovinos de leite. Universidade Estadual Paulista,.

- Colella, S., C. Yau, J.M. Taylor, G. Mirza, H. Butler, P. Clouston, A.S. Bassett, A. Seller, C.C. Holmes, and J. Ragoussis. 2007. QuantiSNP: An objective bayes hidden-markov model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res.* 35:2013–2025. doi:10.1093/nar/gkm076.
- Dellinger, A.E., S.M. Saw, L.K. Goh, M. Seielstad, T.L. Young, and Y.J. Li. 2010. Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. *Nucleic Acids Res.* 38:1–14. doi:10.1093/nar/gkq040.
- Dierssen, M., Herault, Y., Estivil, X. 2009. Aneuploidy: from a physiological mechanism of variance to Down syndrome. *Physiol Rev.* 89(3):887-920. doi: 10.1152/physrev.00032.2007.
- Gao, Y., J. Jiang, S. Yang, Y. Hou, G.E. Liu, S. Zhang, Q. Zhang, and D. Sun. 2017. CNV discovery for milk composition traits in dairy cattle using whole genome resequencing. *BMC Genomics* 18:1–12. doi:10.1186/s12864-017-3636-3.
- Goddard, M.E., B.J. Hayes, and T.H.E. Meuwissen. 2010. Genomic selection in livestock populations. *Genet. Res. (Camb).* 92:413–421. doi:10.1017/S0016672310000613.
- Hammami, H., J. Vandenplas, M.L. Vanrobays, B. Rekik, C. Bastin, and N. Gengler. 2015. Genetic analysis of heat stress effects on yield traits, udder health, and fatty acids of Walloon Holstein cows. *J. Dairy Sci.* 98:4956–4968. doi:10.3168/jds.2014-9148.
- Hayes, B.J., P.J. Bowman, A.J. Chamberlain, and M.E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. *J. Dairy Sci.* 92:433–443. doi:10.3168/jds.2008-1646.
- Hayes, B.J., M. Carrick, P. Bowman, and M.E. Goddard. 2003. Genotype x environment interaction for milk production of daughters of Australian dairy sires from test-day records. *J. Dairy Sci.* 86:3736–3744. doi:10.3168/jds.S0022-0302(03)73980-0.
- 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. 2012. Fine mapping of copy number variations on two cattle genome assemblies using high density SNP array. *BMC Genomics* 13:1–10. doi:10.1186/1471-2164-13-376.
- Hou, Y., G.E. Liu, D.M. Bickhart, M.F. Cardone, K. Wang, E. soo Kim, L.K. Matukumalli, M. Ventura, J. Song, P.M. VanRaden, T.S. Sonstegard, and C.P. Van Tassell. 2011. Genomic characteristics of cattle copy number variations. *BMC Genomics* 12. doi:10.1186/1471-2164-12-127.

- lung, L.H.S., J. Petrini, J. Ramírez-Díaz, M. Salvian, G.A. Rovadoscki, F. Pilonetto, B.D. Dauria, P.F. Machado, L.L. Coutinho, G.R. Wiggans, and G.B. Mourão. 2019. Genome-wide association study for milk production traits in a Brazilian Holstein population. *J. Dairy Sci.* 102:5305–5314. doi:10.3168/jds.2018-14811.
- Jiang, L., J. Jiang, J. Wang, X. Ding, J. Liu, and Q. Zhang. 2012. Genome-Wide Identification of Copy Number Variations in Chinese Holstein. *PLoS One* 7. doi:10.1371/journal.pone.0048732.
- Jiang, L., J. Jiang, J. Yang, X. Liu, J. Wang, H. Wang, X. Ding, J. Liu, and Q. Zhang. 2013. Genome-wide detection of copy number variations using high-density SNP genotyping platforms in Holsteins. *BMC Genomics* 14:23–26. doi:10.1186/1471-2164-14-131.
- Li, L., Y. Wang, C. Li, and G. Wang. 2017. Proteomic analysis to unravel the effect of heat stress on gene expression and milk synthesis in bovine mammary epithelial cells. *Anim. Sci. J.* 88:2090–2099. doi:10.1111/asj.12880.
- Liu, R., Y. Sun, G. Zhao, F. Wang, D. Wu, M. Zheng, J. Chen, L. Zhang, Y. Hu, and J. Wen. 2013. Genome-Wide Association Study Identifies Loci and Candidate Genes for Body Composition and Meat Quality Traits in Beijing-You Chickens. *PLoS One* 8. doi:10.1371/journal.pone.0061172.
- Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829. doi:10.1093/genetics/157.4.1819.
- Nguyen, T.T.T., P.J. Bowman, M. Haile-Mariam, G.J. Nieuwhof, B.J. Hayes, and J.E. Pryce. 2017. Short communication: Implementation of a breeding value for heat tolerance in Australian dairy cattle. *J. Dairy Sci.* 100:7362–7367. doi:10.3168/jds.2017-12898.
- Petrini, J.; lung, L. H. S.; Rodriguez, M. A. P.; Salvian, M.; Pértille, F.; Ravadoscki, G. A.; Cassoli, L. D.; Coutinho, L. L.; Machado, P. F.; Wiggans, G. R.; Mourão, G. B. 2016. *Animal Breeding and Genetics* 133:384-395.
- Pryce, J.E., and H.D. Daetwyler. 2012. Designing dairy cattle breeding schemes under genomic selection: A review of international research. *Anim. Prod. Sci.* 52:107–114. doi:10.1071/AN11098.
- R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrived from <https://www.R-project.org/>

- Ravagnolo, O., I. Misztal, and G. Hoogenboom. 2000. Genetic component of heat stress in dairy cattle, development of heat index function. *J. Dairy Sci.* 83:2120–2125. doi:10.3168/jds.S0022-0302(00)75094-6.
- Renaudeau, D., A. Collin, S. Yahav, V. De Babilio, J.L. Gourdin, and R.J. Collier. 2012. Adaptation to hot climate and strategies to alleviate heat stress in livestock production. *Animal* 6:707–728. doi:10.1017/S1751731111002448.
- Rodriguez, M.A.P., J. Petrini, E.M. Ferreira, L.R.M.B. Mourão, M. Salvian, L.D. Cassoli, A.V. Pires, P.F. Machado, and G.B. Mourão. 2014. Concordance analysis between estimation methods of milk fatty acid content. *Food Chem.* 156:170–175. doi:10.1016/j.foodchem.2014.01.092.
- Salomon-Torres, R., R. Villa-Angulo, and C. Villa-Angulo. 2016. Analysis of copy number variations in Mexican Holstein cattle using axiom genome-wide Bos 1 array. *Genomics Data* 7:97–100. doi:10.1016/j.gdata.2015.12.007.
- Santana, M.L., A.B. Bignardi, R.J. Pereira, A. Menéndez-Buxadera, and L. El Faro. 2016. Random regression models to account for the effect of genotype by environment interaction due to heat stress on the milk yield of Holstein cows under tropical conditions. *J. Appl. Genet.* 57:119–127. doi:10.1007/s13353-015-0301-x.
- Da Silva, V.H., L.C. De Almeida Regitano, L. Geistlinger, F. Pértile, P.F. Giachetto, R.A. Brassaloti, N.S. Morosini, R. Zimmer, and L.L. Coutinho. 2016. Genome-wide detection of CNVs and their association with meat tenderness in Nelore cattle. *PLoS One* 11:1–22. doi:10.1371/journal.pone.0157711.
- Stankiewicz, P., and J.R. Lupski. 2010. Structural variation in the human genome and its role in disease. *Annu. Rev. Med.* 61:437–455. doi:10.1146/annurev-med-100708-204735.
- Steinfeld, H., and P. Gerber. 2010. Livestock production and the global environment: Consume less or produce better?. *Proc. Natl. Acad. Sci. U. S. A.* 107:18237–18238. doi:10.1073/pnas.1012541107.
- Upadhyay, M., V.H. da Silva, H.J. Megens, M.H.P.W. Visker, P. Ajmone-Marsan, V.A. Bălțeanu, S. Dunner, J.F. Garcia, C. Ginja, J. Kantanen, M.A.M. Groenen, and R.P.M.A. Crooijmans. 2017. Distribution and functionality of copy number variation across European cattle populations. *Front. Genet.* 8:1–12. doi:10.3389/fgene.2017.00108.

- Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W.M. Muir. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. *Genet. Res. (Camb)*. 94:73–83. doi:10.1017/S0016672312000274.
- 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.
- Weigel, K.A., P.C. Hoffman, W. Herring, and J.J. Lawlor. 2012. Potential gains in lifetime net merit from genomic testing of cows, heifers, and calves on commercial dairy farms. *J. Dairy Sci.* 95:2215–2225. doi:10.3168/jds.2011-4877.
- West, J.W. 2003. Effects of heat-stress on production in dairy cattle. *J. Dairy Sci.* 86:2131–2144. doi:10.3168/jds.S0022-0302(03)73803-X.
- 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.
- Xu, L., Y. Hou, D. Bickhart, J. Song, and G. Liu. 2013. Comparative Analysis of CNV Calling Algorithms: Literature Survey and a Case Study Using Bovine High-Density SNP Data. *Microarrays* 2:171–185. doi:10.3390/microarrays2030171.
- Zhang, F., W. Gu, M.E. Hurles, and J.R. Lupski. 2009. Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.* 10:451–481. doi:10.1146/annurev.genom.9.081307.164217.