**UNIVERSIDADE ESTADUAL PAULISTA - UNESP**

**CÂMPUS DE JABOTICABAL**

# ESTRUTURA GENÔMICA POPULACIONAL DE BOVINOS LEITEIROS GIR

**Tássia Souza Bertipaglia**

Zootecnista

**2017**

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

# ESTRUTURA GENÔMICA POPULACIONAL DE BOVINOS LEITEIROS GIR

**Tássia Souza Bertipaglia**
**Orientadora: Profa. Dra. Vera Fernanda Martins Hossepian de Lima**
**Coorientadores: Prof. Dr. Danísio Prado Munari**
**Dr. Rodrigo Pelicioni Savegnago**
**Dr. Marcos Vinícius Gualberto Barbosa da Silva**

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

**2017**

Bertipaglia, Tássia Souza
B544e Estrutura genômica populacional de bovinos leiteiros Gir / Tássia Souza Bertipaglia. – – Jaboticabal, 2017
v, 69 p. : il. ; 29 cm

Tese (doutorado) - Universidade Estadual Paulista, Faculdade de Ciências Agrárias e Veterinárias, 2017
Orientadora: Vera Fernanda Martins Hossepian de Lima
Coorientadores: Danísio Prado Munari, Rodrigo Pelicioni Savegnago, Marcos Vinícius Gualberto Barbosa da Silva
Banca examinadora: Lenira El Faro, Joslaine Noely dos Santos Gonçalves Cyrillo, Ricardo da Fonseca, João Ademir de Oliveira
Bibliografia

1. Análise multivariada. 2. Bos taurus indicus. 3. Bovino de leite. 4. Linhagens. 5. Painéis de SNP. 6. Variabilidade genética I. Título. II. Jaboticabal-Faculdade de Ciências Agrárias e Veterinárias.

CDU 636.082:636.2

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

UNIVERSIDADE ESTADUAL PAULISTA

**Câmpus de Jaboticabal**

## CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: ESTRUTURA GENÔMICA POPULACIONAL DE BOVINOS LEITEIROS GIR

**AUTORA: TÁSSIA SOUZA BERTIPAGLIA**
**ORIENTADORA: VERA FERNANDA MARTINS HOSSEPIAN DE LIMA**
**COORIENTADOR: MARCOS VINÍCIUS GUALBERTO BARBOSA DA SILVA**
**COORIENTADOR: DANISIO PRADO MUNARI**
**COORIENTADOR: RODRIGO PELICIONI SAVEGNAGO**

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

Pós-doutorando RODRIGO PELICIONI SAVEGNAGO
Departamento de Ciências Exatas / FCAV / UNESP - Jaboticabal

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

Pesquisadora Dra. JOSLAINE NOELY DOS SANTOS GONÇALVES CYRILLO
Instituto de Zootecnia / Sertãozinho/SP

Prof. Dr. RICARDO DA FONSECA
Departamento de Zootecnia / FCAT / UNESP - Dracena

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

Jaboticabal, 08 de maio de 2017

## DADOS CURRICULARES DO AUTOR

**TÁSSIA SOUZA BERTIPAGLIA**, nascida em Tupi Paulista – SP em 11 de agosto de 1986, filha de José Alipio Bertipaglia e Rosani Maria de Souza Bertipaglia. Iniciou o curso de Zootecnia em agosto de 2006 na Universidade Estadual Paulista "Júlio de Mesquita Filho" (FCAT/Unesp), Dracena – SP, e obteve o título de bacharel em Zootecnia em julho de 2011. Em agosto de 2011, ingressou no curso de mestrado do programa de Pós-Graduação em Genética e Melhoramento Animal na Faculdade de Ciências Agrárias e Veterinárias (FCAV/UNESP), Jaboticabal - SP, sob orientação do professor Dr. Ricardo da Fonseca, tornando-se mestre em julho de 2013. Em agosto de 2013, ingressou no curso de doutorado do programa de Pós-Graduação em Genética e Melhoramento Animal na Faculdade de Ciências Agrárias e Veterinárias (FCAV/ UNESP), Jaboticabal - SP, sob orientação do professora Dra. Vera Fernanda Martins Hossepian de Lima e coorientação do Prof. Dr. Danísio Prado Munari, Dr. Rodrigo Pelicioni Savegnago e Dr. Marcos Vinícius Gualberto Barbosa da Silva.

“Não é sobre ter todas as pessoas do mundo para si.
É sobre saber que em algum lugar alguém zela por ti.
É sobre cantar e poder escutar mais do que a própria voz.
É sobre dançar na chuva de vida que cai sobre nós.
É saber se sentir infinito num universo tão vasto e bonito, é saber sonhar.
Então fazer valer a pena cada verso daquele poema sobre acreditar.
Não é sobre chegar no topo do mundo e saber que venceu.
É sobre escalar e sentir que o caminho te fortaleceu.
É sobre ser abrigo e também ter morada em outros corações.
E assim ter amigos contigo em todas as situações.
A gente não pode ter tudo, qual seria a graça do mundo se fosse assim?
Por isso eu prefiro sorrisos e os presentes que a vida trouxe para perto de mim.
Não é sobre tudo que o seu dinheiro é capaz de comprar.
E sim sobre cada momento, sorriso a se compartilhar.
Também não é sobre correr contra o tempo para ter sempre mais.
Porque quando menos se espera, a vida já ficou para trás!
Segura teu filho no colo, sorria e abraça os teus pais enquanto estão aqui.
Que a vida é trem bala, parceiro, e a gente é só passageiro prestes a partir...”

Ana Vilela

Aos meus pais José Alipio e Rosani,

A minha irmã Camila, cunhado Gustavo e sobrinhos Cecilia e Antonio,

Ao meu amor Fabio,

Pelo apoio, força, amor e por me guiarem nos momentos de dificuldade.

Dedico.

## AGRADECIMENTOS

À Deus, pela vida, por me proteger e me guiar em toda a caminhada.

Ao Fabio, meu amor, marido, amigo, que foi essencial para este trabalho, pela ajuda, ensinamentos, paciência, companheirismo e por colaborar sempre que precisei. Muito obrigada!

Aos meus familiares, avós, tios, primos, cunhadas, sogros, e amigos, da graduação, da pós-graduação e demais amigos, que sempre torceram por mim.

Ao professor Dr. Danísio Prado Munari por me receber no seu grupo, pela orientação, ensinamentos e confiança no meu trabalho. À professora Dra. Vera Fernanda Martins Hossepian de Lima pela orientação e confiança. Aos coorientadores Dr. Rodrigo Penilioni Savegnago e Dr. Marcos Vinícius Gualberto Barbosa da Silva, pela coorientação e colaboração com o trabalho. Muito obrigada.

Aos professores da pós-graduação Dr. Adhemar Sanches, Dr. Anibal Eugênio Vercesi Filho, Dr. Danísio Prado Munari, Dr. Euclides Malheiros, Dr. Fernando Sebastian Baldi Rey, Dr. Guilherme Rosa, Dr. Henrique Nunes de Oliveira, Dr. João Ademir de Oliveira, Dra. Lúcia Galvão de Albuquerque, Dr. Ricardo da Fonseca, Dr. Roberto Carvalheiro e Dra. Sandra Aidar de Queiroz. Obrigada pelos ensinamentos.

Aos membros da banca de qualificação e defesa, Dra. Lenira El Faro, Dra. Joslaine Noely dos Santos Gonçalves Cyrillo, Dr. João Ademir de Oliveira, Dr. Ricardo da Fonseca, Dr. Rodrigo Pelicioni Savegnago, Dra. Nedenia Stafuzza. Obrigada por colaborarem com o meu trabalho. Em especial ao Dr. João Ademir, que colaborou diretamente com este trabalho. Obrigada pela disponibilidade.

Ao pessoal do grupo EAGMA, Alejandro, Ana Paula, Guilherme Bio, Jaque, Jorge, Letícia, Luara, Marcos, Natalia, Nedenia, Priscila, Rafael, Rebeka, Rodrigo, Salvador e Tati. Obrigada pela ajuda e pela companhia nos momentos do café.

À UNESP, ao Programa de Pós-graduação em Genética e Melhoramento Animal, aos funcionários da Unesp, aos queridos funcionários do Departamento de Ciências Exatas, à CAPES pela bolsa de estudos concedida.

À Embrapa CNPGL, principalmente ao Dr. Marcos Vinícius e Dr. João Claudio Panetto, e ABCGIL, por cederem os dados para as minhas análises e pela disponibilidade em colaborar com o trabalho.

A todos vocês minha gratidão!

# SUMÁRIO

**Página**

LISTA DE TABELAS ..... ii
LISTA DE FIGURAS ..... iii
RESUMO ..... iv
ABSTRACT ..... v
1 INTRODUÇÃO ..... 1
2 REVISÃO DE LITERATURA ..... 2
2.1 História da raça Gir no Brasil ..... 2
2.2 Variabilidade genética ..... 4
2.3 Controle de qualidade dos genótipos ..... 5
2.4 Estrutura da população ..... 11
2.5 Técnicas para análise de agrupamento ..... 13
2.6 Análises de componentes principais (PCA) ..... 18
2.7 Parâmetros populacionais baseados em informações genômicas ..... 21
3 MATERIAL E MÉTODOS ..... 28
3.1 Dados genômicos ..... 28
3.2 Controle de qualidade ..... 29
3.3 Análise dos dados genômicos ..... 30
4 RESULTADOS E DISCUSSÃO ..... 36
5 CONCLUSÕES ..... 57
6 REFERÊNCIAS ..... 58
7 APÊNDICES ..... 66

## LISTA DE TABELAS

**Página**

Tabela 1. Análise descritiva dos marcadores com base nos cromossomos .....36
Tabela 2. Variâncias das distâncias genéticas dos indivíduos dos cluster 1, 2, 3 e 4 nos três primeiros componentes principais (PC1, PC2 e PC3) .....39
Tabela 3. Média do desvio de parentesco genômico dentro (diagonal) e entre cluster (acima da diagonal) e desvio-padrão (entre parênteses e abaixo da diagonal) .....51
Tabela 4. Média, mediana, desvio-padrão (DP), coeficiente de variação (CV), valores mínimo e máximo de heterozigosidade esperada (He), heterozigosidade observada (Ho), coeficiente de endogamia e desequilíbrio de ligação (LD) para a população estudada .....52
Tabela 5. Probabilidade de indivíduos endogâmicos (x) e coeficiente de endogamia (e) de acordo com a Distribuição de Qui-Quadrado (Tabela IV).....53
Tabela 6. Desequilíbrio de ligação ($r^2$) entre pares de SNP (N) localizados em diferentes distâncias (em Kb) .....55

## LISTA DE FIGURAS

**Página**

Figura 1. Eixos do componente principal (CP1) e secundário ($CP_2$), perpendicular, sendo a elipse a densidade. ..... 20
Figura 2. Etapas utilizadas para a análise de componentes principais. ..... 31
Figura 3. Determinação do número ótimo de cluster obtidos pelos Métodos de Ward (a), Elbow (b), pela minimização da soma de quadrados, e Silhouette (c), pelo maior grau de ajuste. ..... 38
Figura 4. Dendograma da matriz de distância genética de alelos compartilhados colorido para os diferentes cluster. ..... 40
Figura 5. Projeção dos indivíduos em 3D, considerando os três primeiros componentes principais (PC1, PC2, PC3) em duas faces, anterior (a) e esquerda (b), coloridas para os diferentes cluster, e os touros mais influentes na variância dos componentes principais, enumerados. ..... 41
Figura 6. Proporção das variâncias dos dez primeiros componentes principais (barras) e proporção da variância acumulada de todos os componentes principais (linha contínua). ..... 43
Figura 7. Projeção das distâncias genéticas dos indivíduos nos dois primeiros componentes principais. ..... 43
Figura 8. Projeção das variâncias das distâncias genéticas dos indivíduos nos dois primeiros componentes principais coloridos considerando a contribuição individual. ..... 44
Figura 9. Projeção dos indivíduos nos três primeiros componentes principais em 3D considerando as diferentes fazendas, identificadas por diferentes colorações. ..... 45
Figura 10. Projeção dos vinte touros mais (a) e menos (b) influentes na variância dos dois primeiros componentes principais. ..... 46
Figura 11. Matriz de parentesco genômico entre os animais distribuídos em seus respectivos cluster. ..... 49
Figura 12. Distribuição das estimativas de parentesco genômico dos animais nos diferentes cluster ..... 50
Figura 13. Desequilíbrio de ligação (LD) ao longo das distâncias entre SNP em Megabase (Mb). ..... 56

**ESTRUTURA GENÔMICA POPULACIONAL DE BOVINOS LEITEIROS GIR**

**RESUMO** – Considerando painéis de SNP (do inglês *Single Nucleotide Polymorphism*, polimorfismo de nucleotídeo único) objetivou-se descrever o perfil genômico populacional de bovinos leiteiros da raça Gir por meio da caracterização de linhagens desta população e inferir sobre a variabilidade genética utilizando informações de animais provenientes de diversas fazendas do Brasil. Após o controle de qualidade de genótipos e amostras, foram obtidos 1987 SNP bialélicos em cromossomos autossomos presentes nos painéis HD e 50K SNP, em um painel resultante de SNP em comum dos dois painéis, 21K SNP. Distâncias genéticas entre os marcadores foram obtidas e, por meio da decomposição espectral, foram gerados os componentes principais. A análise multivariada de componentes principais, em um contexto de estrutura populacional, permite inferir sobre a variabilidade genética utilizando informações de parentesco entre os animais. A determinação do número de agrupamentos foi realizada com o procedimento k-*means*, utilizando a distância genética entre os animais, que permite identificar subgrupos genéticos na população, com moderada-alta qualidade de agrupamento, com coeficiente correlação cofenética equivalente a 0,66. O parentesco, heterozigosidade esperada (He) e observada (Ho), coeficientes de endogamia, desequilíbrio de ligação (LD) e tamanho efetivo populacional, baseados exclusivamente em informações genômicas também foram obtidos. Os resultados evidenciaram quatro diferentes agrupamentos genéticos, definidos como linhagens da raça Gir leiteira do Brasil, indicando que o uso de poucos reprodutores ou de seus descendentes com maior intensidade e também devido a diferentes origens dos ancestrais, formou os diferentes agrupamentos genéticos, os quais são compostos por animais de várias fazendas. Há pouco parentesco entre os animais e isso se refletiu na média do coeficiente de endogamia, sendo que a média da endogamia obtida para todos os indivíduos foi de 0,017, evidenciando o fluxo gênico nas diferentes fazendas, o que evitou o acasalamento entre animais aparentados, ou ainda devido à escolha dos animais para genotipagem. O valor médio de He foi 0,25, próximo a média de Ho, revelando que não houve perda e nem fixação de alelos com efeitos indesejáveis. O LD foi obtido para todos os pares de SNP adjacentes com janelas de 50 Mb, revelando que o LD é maior em menores distâncias entre os marcadores, com valor médio de LD de 0,17 ± 0,26 para até 200 Kb, sendo assim, o LD estimado para espaçamento até 40 Kb seria suficiente para se obter associação com *Quantitative Trait Loci*. O tamanho efetivo populacional (Ne) foi 92,84 animais, sendo assim, cerca de 93 indivíduos nesta população contribuirão em termos de variância genética para a próxima geração desta população. Portanto, concluiu-se que o fluxo gênico nas propriedades mantém a variabilidade genética dos rebanhos. Esses parâmetros populacionais poderão auxiliar nas decisões de acasalamentos dos programas de melhoramento genético da raça a fim de aumentar ou manter a variabilidade genética dos rebanhos.

**Palavras-chave**: análise multivariada, *Bos taurus indicus*, bovino de leite, linhagens, painéis de SNP, variabilidade genética

## GENOMIC STRUCTURE OF DAIRY CATTLE GIR POPULATION

**ABSTRACT –** The objective was to describe the genomic structure of Gir dairy cattle population through the characterization of lineages of this population and infer about the genetic variability considering information from animals from several farms in Brazil using panels of SNP (Single Nucleotide Polymorphism). After quality control of genotypes and samples, we obtained 1987 biallelic SNPs on autosomal chromosomes present in the HD panels and 50K SNP, in a resultant panel of SNPs in common from to two panels, 21K SNP. Genetic distances between the markers were obtained and, through spectral decomposition, the principal components were generated. Multivariate analysis of principal components, in a context of population structure, allows inferring about genetic variability using relatedness information between animals. The determination of the number of clusters was performed using the k-means procedure, using the genetic distance between the animals, which allows the identification of genetic subgroups in the population, with moderate-high quality of grouping, with a coefficient correlation coefficient equivalent to 0.66. Expected heterozygosity (He) and observed (Ho), coefficients of inbreeding, linkage disequilibrium (LD) and effective population size (Ne) based exclusively on genomic information were also obtained. The results evidenced four different genetic groups, defined as breeding lineages of the Gir dairy breed of Brazil, indicating that the use of a few breeders or their descendants with greater intensity and also due to the different origins of the ancestors, formed the different genetic groups, which are composed of animals from various farms. There was little relationship between the animals and this was reflected in the average inbreeding coefficient, and the average inbreeding obtained for all the individuals was 0.017, evidencing the gene flow in the different farms, which avoided the mating between related animals, or due to the choice of animals for genotyping. The mean value of He was 0.25, close to the mean of Ho, revealing that there is no loss or fixation of alleles with undesirable effects. The LD was obtained for all adjacent SNP pairs with 50 Mb windows, revealing that the LD is larger at smaller distances between the markers, with a mean LD value of 0.17 ± 0.26 for up to 200 Kb, therefore, the estimated LD for spacing up to 40 Kb would be sufficient to obtain association with quantitative trait loci (QTL). The effective population size was 92.84 animals, thus, about 93 individuals in this population will contribute in terms of genetic variance for the next generation of this population. Therefore, it was concluded that the gene flow in the properties maintains the genetic variability of the herds. These population parameters may aid in the breeding decisions of breed breeding programs in order to increase or maintain the genetic variability of the herds.

**Keywords**: *Bos taurus indicus*, dairy cattle, genetic variability, lineages, multivariate analysis, SNP panels

## 1 INTRODUÇÃO

O agronegócio do leite e seus derivados ocupam uma posição de destaque na economia nacional. Além do leite ser um dos produtos mais importantes da pecuária brasileira, é uma fonte de alimento de alto valor nutritivo e representa importante papel social na geração de empregos no país. De acordo com a Organização das Nações Unidas para Alimentação e Agricultura, o Brasil é um fundamental agroexportador, principalmente em relação aos produtos do setor pecuário, sendo responsável por 40% da produção de leite da América Latina. A produção de leite no Brasil dobrou entre os anos 2000 e 2015, o que colocou o Brasil entre os cinco maiores produtores mundiais de leite de vaca no *ranking* mundial, superado pela Índia, Estados Unidos, China e Paquistão (FAO, 2015).

A raça Gir é a principal zebuína leiteira e também a mais utilizada em cruzamentos com a raça Holandesa no país (SILVA; MACHADO, 2011). Apesar da pequena população da raça quando comparada às demais implementadas no país, a contribuição genética do Gir leiteiro para os rebanhos comerciais tem aumentado. O aumento no uso de animais da raça Gir foi atribuído à implementação do Programa Nacional de Melhoramento de Gir Leiteiro (PNMGL) em 1985, com rápido crescimento desde aquele ano.

Para o desenvolvimento dos programas de melhoramento genético da raça, um fator importante a ser considerado é o conhecimento da estrutura genômica populacional. No contexto de estrutura populacional é possível inferir sobre a variabilidade genética da população, fator fundamental, pois além de fatores evolutivos, está aliada à expressão do potencial genético da população e à facilidade de seleção de animais geneticamente superiores. Além disso, o estudo da estrutura genômica populacional é uma importante ferramenta por permitir maior compreensão sobre o histórico da população como endogamia, animais que contribuem na variabilidade genética da população, introdução de novo material genético, identificar subgrupos genéticos, agrupá-los por semelhança genética, detectar desequilíbrio de ligação entre marcadores e QTL e conhecer as relações de parentesco.

O estudo da estratificação populacional para identificar possíveis linhagens, que são subestruturas genéticas, permite considerar esses efeitos para se obter maior acurácia nas análises utilizando informações genômicas, por alocar os indivíduos de diferentes grupos genéticos nos grupos de treinamento e validação nas análises genômicas. Ao considerar os diferentes subgrupos genéticos há a correção para a estratificação (PRICE et al., 2006) por ajustar o efeito dos agrupamentos e, com isso, elimina a ocorrência de falsos SNP significativos para as características de interesse econômico de forma mais eficiente, por abranger todos os grupos genéticos.

A compreensão da estrutura genômica populacional é importante para os estudos de predição genômica, seleção e associação, e poderá auxiliar nas decisões dos acasalamentos dentro dos programas de melhoramento genético da raça Gir. Portanto, o objetivo deste estudo foi descrever o perfil genômico populacional por meio da caracterização de linhagens desta população de bovinos leiteiros da raça Gir e inferir sobre a variabilidade genética, por meio das análises de componentes principais, análise de agrupamento, endogamia, heterozigosidade, desequilíbrio de ligação e tamanho efetivo populacional, baseados exclusivamente em informações genômicas contidas em painéis de SNP.

## 2 REVISÃO DE LITERATURA

### 2.1 História da raça Gir no Brasil

Na pecuária leiteira de países de clima tropical, em que há predominância de produção de leite à pasto e com forragens de baixa qualidade, raças taurinas (*Bos taurus taurus*) não conseguem expressar todo o seu potencial produtivo que apresentariam se estivessem produzindo em clima temperado. Desta forma, é recomendável a utilização de raças zebuínas (*Bos taurus indicus*) com propósito leiteiro, por serem adaptadas ao clima, com o intuito de se obter aumento de produtividade (MADALENA et al., 1990).

No Brasil, o rebanho bovino foi formado originalmente por gado de origem taurina trazidos da Índia, Península Ibérica, na época da colonização. Com as

importações que ocorreram naquela época, o rebanho nacional sofreu grandes modificações pelo processo de cruzamento absorvente (VERCESI FILHO et al., 2010). Em 1970, cerca de 32% do rebanho nacional era leiteiro. Em 1980, caiu para 20%, com aumento da contribuição na produção de leite dos bovinos de corte, com maior destaque para os bovinos mistos, responsáveis por 60% da produção de leite. Os bovinos mistos são produtos de cruzamentos com zebu e pelas raças zebuínas com propósito leiteiro (AMARAL et al., 1988).

Os bovinos da raça Gir são originários das regiões de Gir na Índia, da península de Kathiaware onde eram utilizados para a produção de leite. Possuem coloração vermelha, branca ou manchada, sendo os primeiros animais introduzidos no Brasil em importações que ocorreram entre os anos de 1918 e 1962 (MASON, 2002). As importações mais importantes ocorreram em 1930, 1960 e 1962 (VERCESI FILHO et al., 2010), com grande contribuição para a formação gênica da raça nos dias atuais. Inicialmente eram utilizados para duplo propósito, ou seja, carne e leite. No final da década de 1930, foram iniciados os primeiros trabalhos de seleção do Gir para leite no Brasil (ACGZ, 2016), sendo que a subdivisão genética definitiva da raça ocorreu entre 1993 e 2002 (SANTANA JUNIOR et al., 2014).

A raça Gir é a primeira zebuína leiteira do mundo a ter touros provados pelo teste de progênie e a segunda que mais possui controle leiteiro oficial no Brasil (LEÃO et al., 2013). Além disso, se apresenta como a melhor opção em sistemas de produção à pasto, com forragens de baixa qualidade na maior parte do ano, por ser altamente adaptada às condições ambientais brasileiras.

O programa de melhoramento genético da raça Gir no Brasil é conduzido pela Associação Brasileira dos Criadores de Gir Leiteiro (ABCGIL) e pela Empresa Brasileira de Pesquisa Agropecuária (Embrapa) Gado de Leite localizada em Juiz de Fora, Minas Gerais. Estas instituições desenvolvem há 37 anos o Programa Nacional de Melhoramento do Gir Leiteiro (PNMGL), com o objetivo de identificar, por meio dos teste de progênie, os touros geneticamente superiores para características de produção e qualidade do leite. Os touros em teste de progênie são avaliados e selecionados para características de produção (leite, gordura, proteína e sólidos totais), conformação corporal e do sistema mamário, e facilidade

de ordenha e temperamento, visando o melhoramento genético da raça para aptidão leiteira (PANETTO et al., 2016).

Os rebanhos anteriormente fechados geneticamente começaram a participar da avaliação genética de touros do programa de melhoramento da raça, o que aumentou as trocas gênicas nos rebanhos. Os acasalamentos entre os rebanhos aumentaram a partir de 2002, o que contribuiu para a redução da endogamia (SANTANA JUNIOR et al., 2014). Para maximizar o ganho genético em uma população de Gir no Brasil seria necessário aumentar o tamanho da população nos rebanhos, controlar endogamia por meio de acasalamentos dirigidos, evitando o uso intensivo de poucos touros (REIS FILHO et al., 2010).

Por ser uma raça adaptada, o Gir leiteiro apresenta menores infestações de ectoparasitas e endoparasitas e menores incidências de doenças do que raças de clima temperado, quando mantidos sob nossas condições ambientais (ABCGIL, 2017). Apesar do pequeno tamanho da população, cerca de 150.000 animais no Brasil (ABCGIL, 2017), a sua contribuição genética no país tem aumentando por mostrar-se como a raça preferencialmente utilizada em cruzamentos com taurinos leiteiros, como a raça Holandesa (NEIVA, 2017), contribuindo com produção de leite e rusticidade, características fundamentais para a produção econômica de leite em países de clima tropical.

## 2.2 Variabilidade genética

A variabilidade genética é definida como a variedade de genótipos, haplótipos e alelos presentes em determinada população (TORO; VILLANUEVA; FERNÁNDEZ, 2014). Essa diversidade determina a capacidade da população em evoluir, a fim de atender as estratégias dos programas de melhoramento genético. Uma preocupação importante dos programas de melhoramento é a manutenção da variabilidade genética, pois os ganhos genéticos dependem diretamente da variabilidade genética da população. A variabilidade genética permite que os programas de melhoramento apresentem alternativas nas mudanças dos objetivos de seleção, pois é preciso adaptação às mudanças ambientais e no interesse dos consumidores, que variam ao longo do tempo.

Entretanto, a seleção dos animais realizada por programas de melhoramento genético diminui a variabilidade genética pelo favorecimento de alguns alelos. Além disso, os indivíduos selecionados podem ser altamente aparentados, elevando a taxa de endogamia e, consequentemente, afetando a variabilidade genética. Um dos principais fatores associados à perda de variabilidade genética é a endogamia, que como consequência pode levar à perda do desempenho produtivo, em termos de reprodução, sobrevivência e resistência a doenças, causadas pela depressão endogâmica (FALCONER; MACKAY, 1996). Para isso, é necessário monitoramento do fluxo gênico e inclusão de material genético novo na população, o que possibilita novas trocas gênicas entre os indivíduos.

Uma gestão adequada de escolha de animais para acasalamento nos programas de melhoramento genético podem auxiliar na manutenção ou acréscimo da diversidade genética das espécies a longo prazo. Porém, na realidade, o que é observado é que alguns poucos touros são utilizados repetidamente, contribuindo menos para a variabilidade genética. Como consequência, tem causado um “gargalo” no pedigree, o que pode levar a perda da variabilidade genética nos rebanhos ao longo das gerações. Portanto, a maior compreensão das relações genéticas entre as populações de gado são prioridades para a gestão da diversidade genética dos animais (NOTTER, 1999).

## 2.3 Controle de qualidade dos genótipos

A utilização de informações genômicas é cada vez mais utilizada em programas de melhoramento genético de animais domésticos. Sendo assim, painéis de marcadores foram desenvolvidos por meio da genotipagem dos indivíduos e o uso das informações contidas nos painéis de SNP pode contribuir para os estudos de seleção e de associação genômica (GWAS) para o aprimoramento da raça no Brasil.

A genotipagem dos animais é o processo que determina os genótipos de uma região do DNA (AA, AB ou BB) de acordo com a leitura de sinais definida no processo de leitura dos genótipos (ZIEGLER; KONIG; THOMPSON, 2008; ANNEY et al., 2008). Como a atribuição dos genótipos é realizada por meio de

procedimentos de leitura automatizados, consequentemente, está sujeita a erros (COLEMAN et al., 2016). Apesar da baixa taxa de erros nas análises laboratoriais, por qualidade ou fonte do DNA, erros podem acontecer, o que podem afetar a intensidade dos sinais de fluorescência utilizadas para identificar os genótipos (ANNEY et al., 2008).

O processo de atribuição dos genótipos produz painéis de marcadores SNP de baixa ou alta densidade de genótipos, compostos por três grupos de sinais. Por padrão da Illumina (ILLUMINA, 2009), a determinação dos genótipos das amostras é resultante da intensidade de luz, gerando regiões de coloração vermelha, roxo e azul, que tem os genótipos denominados AA, AB e BB, respectivamente. Se a genotipagem dos SNP falhar em muitos indivíduos, a sua qualidade é questionável. Para remover ou minimizar esses erros pode-se realizar um controle de qualidade dos SNP e das amostras antes de realizar as análises estatísticas. Esse filtro tem por objetivo eliminar SNP que não se enquadram no controle de qualidade, o que melhora a qualidade do conjunto de dados a ser analisado e, consequentemente, torna os dados mais consistentes.

Existem diferentes critérios de controle de qualidade que podem ser aplicados em dados genômicos. A ausência do controle de qualidade pode gerar falsos SNP significativos (falso negativos ou falso positivos) em estudos de associação genômica (TURNER et al., 2011). Porém, a utilização de critérios rigosos pode eliminar muitos SNP, o que reduz as chances de encontrar SNP significativos que poderiam estar associados com regiões do genoma responsáveis pela determinação de fenótipos de interesse econômico, por excluir parte da variância genética explicada pelos SNP.

Portanto, as informações dos painéis de SNP devem ser verificados minunciosamente com o objetivo de eliminar possíveis erros. Para o controle de qualidade dos dados genômicos existem diferentes critérios que podem ser aplicados aos SNP e amostras individuais, os quais serão descritos a seguir.

- ***Missing genotype* ou genótipos ausentes:**

Amostras mal genotipadas podem ser evidenciadas no filtro de *missing genotype*, causada também pela baixa qualidade de DNA e devem ser removidas

das análises. Esses genótipos podem surgir quando os SNP da plataforma de genotipagem não são analisados corretamente, quando os SNP de interesse não estão na plataforma ou quando a variação que ocorre no DNA é determinada apenas em uma pequena fração de indivíduos (LIN; HU; HUANG, 2008). A remoção dos SNP com uma maior proporção de genótipos ausentes pode resultar em um conjunto de SNP com maior confiabilidade.

- **SNP mapeados em cromossomos sexuais e mitocondrial:**

Devido a falhas na montagem do genoma bovino, alguns SNP podem ainda não ter posição definida no genoma. Para análises genômicas são considerados apenas os cromossomos autossomos, pois os alossomos, cromossomos sexuais X e Y, estão em haploidia nos machos. Sendo assim, estes cromossomos são diferentes entre os sexos e devem ser excluídos da amostra de dados. Além disso, os SNP localizados nos cromossomos mitocondriais são excluídos para retirar o efeito das matrizes, pois o DNA mitocondrial é transmitido apenas pelas fêmeas (GILES et al., 1980).

- **GC *score* (*genotype call score*):**

Este escore é uma medida de qualidade e confiabilidade na leitura dos genótipos. O GC *score* é calculado por meio informações do agrupamento das amostras, sendo que cada SNP é avaliado com base nas seguintes características: ângulo, dispersão, sobreposição e intensidade e é atribuído a cada leitura do genótipo.

Em geral, quando o GC *score* é baixo, os genótipos estão localizados mais distantes do agrupamento de genótipos e, portanto, há menor confiabilidade. Não há interpretação global do GC *score*, pois a pontuação depende do agrupamento das amostras para cada SNP, que é afetado por muitas variáveis, incluindo a qualidade das amostras e dos *loci*. A plataforma Illumina (ILLUMINA, 2009) recomenda que seja utilizado o GC *score* de corte acima de 0,15 para painéis *Infinium*, como é o caso do HDBovineSNP BeadChip.

- ***Call rate* ou taxa de genotipagem:**

Essa taxa é determinada pelo número de SNP que recebem um genótipo homozigótico ou heterozigótico em relação ao número total dos *loci* observados. O *call rate* é um indicador da qualidade ou eficiência de genotipagem, pois determina o percentual de SNP e de amostras que foram lidos de maneira eficiente. Sendo assim, por este critério, são eliminados os SNP que não foram genotipados em uma dada proporção das amostras, incluindo SNP monomórficos, ou amostras que não tiveram altas taxas de leitura.

As amostras com baixa porcentagem de leitura indicam que o processo de genotipagem não teve êxito, podendo assumir valores que podem variar de acordo com a taxa a qualidade de genotipagem, sendo recomendáveis taxas de leitura de SNP acima de 97%. Se a taxa de leitura for considerada de qualidade pelo laboratório, o valor para amostras dos indivíduos pode chegar a 90% (ZIEGLER; KONIG; THOMPSON, 2008).

- **Frequência de alelos menos comuns ou raros (MAF):**

A MAF refere-se à frequência do alelo menos comuns na população e é usada como um filtro de dados no controle de qualidade dos SNP por eliminar possíveis erros de leitura dos genótipos. Os SNP com baixa MAF (*minor allele frequency*) são mais propensos a erros, uma vez que menos amostras estariam dentro de um agrupamento de genótipo. Os SNP com baixa MAF são geralmente menos informativos e podem prejudicar as análises de GWAS por reduzir o poder de detecção de associação genótipo-fenótipo (TEO, 2008).

A MAF é um evento raro e a significância entre os SNP raros e as características em estudo tem maiores chances de ocorrer devido a erros (LAM et al., 2007). Porém, estes alelos raros podem explicar parte das variações genéticas que não são explicadas pelos alelos frequentes (CIRULLI; GOLDSTEIN, 2010). Os SNP são frequentemente excluídos das análises se a MAF é baixa, sendo que o critério de eliminação tipicamente usado varia entre 1% e 5% (ZIEGLER; KONIG; THOMPSON, 2008).

- **Equilíbrio de Hardy-Weinberg (HWE):**

O HWE (*Hardy Weinberg Equilibrium*) ocorre quando as frequências alélicas e genotípicas em uma população grande e com acasalamento aleatório, permanecem estáveis por muitas gerações, havendo constância entre as frequências alélicas e genotípicas. O desvio a partir do HWE é calculado para cada marcador utilizando-se a comparação entre o valor esperado e observado, pelo teste de qui-quadrado de Pearson (COX; KRAFT, 2006), sendo a exclusão realizada com base em p-valores significativos ($P < 0,05$), sendo excluídos SNP altamente significativos. Desvios observados no HWE podem indicar erros de genotipagem (TEO et al., 2007) pela maior proporção de um alelo observado em relação ao esperado.

O desvio do equilíbrio de Hardy-Weinberg pode também ser indício de estratificação da população provocada por seleção ou endogamia devido a agrupamentos de indivíduos que são, em média, mais relacionados entre si do que outros membros da população. Portanto, desvios do HWE podem ser considerados como indicadores da existência de uma subestrutura genética no conjunto de dados da amostra. Além disso, os desvios podem indicar uma tendência de atribuir genótipos homozigotos como heterozigotos. Por evidenciar erros no processo de genotipagem, deve-se excluir os SNP que desviam do HWE (ZIEGLER; KONIG; THOMPSON, 2008).

Alguns autores questionam o mérito de testes de HWE nos estudos de associação genômica (COX; KRAFT, 2006; TEO et al., 2007) e afirmam que este critério não auxilia na identificação de SNP com erros de genotipagem. Os mesmos autores afirmaram que possíveis erros de leitura não são susceptíveis de serem detectados por meio do teste de HWE e podem muitas vezes não ser a causa de desvio extremo de equilíbrio. Além disso, considerar o HWE no filtro de qualidade de genótipos pode eliminar poucos SNP.

- ***Pruning* ou SNP correlacionados:**

A verificação do *pruning* ocorre por meio dos alelos idênticos por descendência (IBD, *Identical by Descent*) e ocorrem quando os indivíduos têm

sequências de nucleotídeos idênticos em um determinado segmento de DNA. Também pode ocorrer por meio da comparação aos pares de determinada janela, sendo que cada possível par de variantes é comparado e é removido quando o LD (*linkage disequilibrium*) entre eles está acima de um dado limite.

A presença de estrutura é inferida a partir da análise dos genótipos do genoma e a presença de LD pode aumentar ou reduzir as semelhanças, mesmo que as duas regiões sejam de mesmo tamanho. Consequentemente, é necessário considerar o *pruning* entre os alelos para avaliação de IBD e estratificação populacional (COLEMAN et al., 2016). O uso indevido de amostras relacionadas pode ocorrer em estudos de grande escala e o *pruning* de SNP deve ser considerado para evitar superestimação das correlações entre os marcadores baseados na proporção de IBD.

A presença de semelhança entre indivíduos, independente do fenótipo, pode ser uma potencial fonte de viés nos resultado de associação genômica (COLEMAN et al., 2016). Além disso, SNP altamente correlacionados podem ter grande influência sobre alguns componentes principais durante a avaliação da estratificação pela análise de componentes principais (PCA). Há autores que recomendam que esses SNP devem ser retirados do arquivo de dados para que a estratificação da população seja avaliada sem viés (LAURIE et al., 2010).

- **Heterozigosidade**

A heterozigosidade é definida como a proporção entre o número de *loci* heterozigotos em relação ao total de *loci* do indivíduo, ou genótipos diplóides constituídos por dois alelos distintos. Sendo assim, estima-se a média e o desvio-padrão da heterozigosidade em todos os animais em estudo para excluir os que não estão dentro de três desvios-padrão a partir da média (GONDRO, 2015), ou seja, 0,3% dos dados da amostra. Uma alternativa comum é eliminar valores atípicos, *outliers*, com base no histograma de heterozigosidade (ZIEGLER; KONIG; THOMPSON, 2008).

A heterozigosidade é um indicador de boa qualidade da genotipagem dos SNP. Se a heterozigosidade for alta, próximo a 0,5, pode indicar contaminação da amostra de DNA em rebanhos puros (ZIEGLER; KONIG; THOMPSON, 2008),

resultando em uma quantidade desproporcional de genótipos heterozigotos, ou ainda indicar amostras erroneamente genotipadas. Se a heterozigosidade for baixa, pode indicar alta semelhança genética na população. A abordagem típica relatada pelos autores anteriormente citados é de estimar a média e desvio-padrão da heterozigosidade para todos os indivíduos da população e excluir aqueles indivíduos com heterozigosidade abaixo ou acima de três desvios-padrão em relação à média.

## 2.4 Estrutura da população

O efeito da estratificação da população ou de agrupamentos genéticos (subpopulação) nos estudos de associação genômica (GWAS) tem sido amplamente discutido na literatura e é citado como um dos principais fatores de viés na interpretação dos resultados quando este efeito não é considerado nas análises (PRICE et al., 2006; WANG et al., 2009; LAURIE et al., 2010). Em GWAS, para corrigir os efeitos de estratificação da população, diferenças devido à subpopulação dentro da mesma raça podem ser consideradas nas análises (ALEXANDER; NOVEMBRE; LANGE, 2009).

Um dos principais fatores de viés na interpretação dos resultados de GWAS é a presença de estrutura na população causada pela presença de subgrupos genéticos na população, tornando importante o estudo da estratificação populacional. Estes efeitos devem ser considerado nos estudos de GWAS (PRICE et al., 2006; WANG et al., 2009; LAURIE et al., 2010), pois o maior efeito da subestrutura populacional é levar a associações falso positivas em GWAS (MARCHINI et al., 2004), então tais efeitos devem ser considerados nas análises genômicas.

Mesmo um pequeno grau de estratificação da população pode afetar as análises de GWAS devido aos grandes tamanhos amostrais necessários para detectar SNP envolvidos na expressão de características complexas. Além disso, a diferença alélica entre as populações ancestrais pode acarretar em uma associação significativa entre alelos e fenótipos, entretanto, sem efeito causal direto ou indireto com o fenótipo (CARDON; PALMER, 2003).

Em espécies domesticadas, a estrutura da população pode ser influenciada ao longo do tempo pela seleção natural ou pelos métodos de melhoramento utilizados pelo homem, sendo assim, alguns SNP diferem na sua frequência alélica mais em uma população do que em outras. As diferenças genéticas entre os animais podem evidenciar a presença de diferentes agrupamentos genéticos em uma população, que ocorrem por haver diferenças na ancestralidade dos indivíduos que compõem a população em estudo (PRICE et al., 2006).

Se houver estratificação populacional, é importante distinguir as diferenças entre as subpopulações que são devidas a fatores de deriva genética e aquelas que surgiram a partir de divergência populacional (PRICE et al., 2010). O desafio central nas análises de dados é explorar se há evidências de que as amostras de uma população são estruturadas (JOMBART; DEVILLARD; BALLOUX, 2010; PATTERSON; PRICE; REICH, 2006) e, a partir disso, explorar os parâmetros genéticos e genômicos para se conhecer a história da população.

A raça Gir foi reportada como geneticamente subdividida, a partir da avaliação do pedigree de animais nascidos entre os anos de 1938 a 1998 (FARIA et al., 2009). Inicialmente, no Brasil, os animais da raça Gir eram utilizados para duplo propósito, carne e leite. No final da década de 1930 foram iniciados os primeiros trabalhos de seleção do Gir para propósito de leite (ACGZ, 2016), sendo que entre 1993 e 2002 houve a definitiva subdivisão da raça (SANTANA JUNIOR et al., 2014). Na década de 1960 ocorreram importações importantes (VERCESI FILHO et al., 2010), com grande contribuição para a formação gênica da raça nos dias atuais. Em estudos sobre a estrutura populacional genômica realizados por meio de informações contidas em painéis 32K SNP revelaram diferentes agrupamentos genéticos resultantes das raças envolvidas na composição da raça Canchim (BUZANSKAS et al., 2017). Em estudos sobre a estrutura populacional genômica de bovinos Hereford e Braford foi possível obter parâmetros populacionais, como desequilíbrio de ligação e tamanho efetivo populacional por meio de painel 41K SNP (BIEGELMEYER et al., 2016). Porém, nenhum estudo foi realizado utilizando informações genômicas de rebanhos leiteiros do Brasil, sobretudo para a raça Gir.

## 2.5 Técnicas para análise de agrupamento

A proposta das análises de agrupamento é colocar, em um mesmo grupo, informações que sejam similares de acordo com algum critério pré-determinado. O critério da análise de agrupamento baseia-se em uma função que contêm as distâncias genéticas a cada par de indivíduos. A ideia deste método é a divisão do conjunto de dados para a definição de grupos distintos e a minimização da variação dentro dos grupos.

A análise de agrupamento ou clusterização (*clustering*) permite classificar um grupo de informações em subgrupos (*cluster*) de maneira a maximizar a homogeneidade dentro dos grupos e a maximizar a heterogeneidade entre os grupos. Sendo assim, torna-se possível estudar a relação entre os subgrupos. A técnica consiste em separar os dados em grupos distintos, baseando-se nas características destes dados para simplificação.

- **Método de *Elbow***

Este método é talvez o mais antigo e conhecido método de determinação do número de grupos. Por meio do método de *Elbow* avalia-se o número ideal de grupos por meio da porcentagem da variância explicada em função do número de agrupamentos dentro dos *cluster* (1), de modo que a adição de um novo agrupamento não acrescentará informação relevante em termos de variância.

Por este método, há o acréscimo da quantidade de *cluster* e a análise é realizada a cada incremento. A porcentagem de variância explicada é a relação entre a variação entre grupos em relação à variância total, visando definir o número aceitável de *cluster* a serem criados com base nos dados da amostra. Quando o acréscimo deixa de ser relevante, a trajetória passa a apresentar movimento retilíneo, pois a diferença da distância entre os grupos é quase insignificante. Quando é determinada a quantidade de k *cluster*, os dados são segmentados em subgrupos.

$$\sum_{k=1}^{k} W(C_k) \qquad (1)$$

Em que, $C_k$ é o *k-ésimo cluster* e $\mathrm{W(C_k)}$ é a variação dentro do *cluster*.

- **Método de *Silhouette***

Este método mede a qualidade de um agrupamento em um grau de confiança dentro de um intervalo, entre -1 e 1. O maior valor *Silhouette* indica o número ótimo de *cluster*, ou seja, quando está mais próximo de 1 (2). Por meio da observação do gráfico de *Silhouette* é possível definir a formação de agrupamentos baseados em dissimilaridades (ROUSSEEUW, 1987).

$$\frac{1}{N} * \sum_{i=1}^{N} \left(\frac{b_i - a_i}{\max(a_i, b_i)}\right) \quad (2)$$

Em que, $N$ é o número total de informações, $a_i$ é a dissimilaridade média das observações i em relação às demais dentro do *cluster*; $b_i$ é dissimilaridade média da observação i em relação às demais observações do *cluster* vizinho mais próximo.

A validação da qualidade do agrupamento dos dados pode ser realizada pela estimação do coeficiente de correlação cofenética (CCC) (SOKAL; ROHLF, 1962) (3), definida como uma correlação linear de Pearson entre o coeficiente de parentesco e a proporção de alelos compartilhados que originam o gráfico dendograma. Sendo assim, mede-se o grau de ajuste entre a matriz de dissimilaridade ou distância genética (matriz fenética F) e a matriz resultante da simplificação devido ao método de agrupamento genético (matriz cofenética C). Por este método, os valores próximos à unidade indicam melhor agrupamento (CRUZ; CARNEIRO; REGAZZI, 2014). Quando CCC é maior que 0,7, o método de agrupamento é considerado adequado.

$$CCC = \frac{\widehat{Cov}(F,C)}{\sqrt{\hat{V}(F).\hat{V}(C)}} \quad (3)$$

Em que, $CCC$ é o coeficiente de correlação cofenética; $\mathrm{\widehat{Cov}(F,C)}$ é a covariância estimada entre as matrizes de dissimilaridade ou fenética $\mathrm{(F)}$ e a matriz

simplificada devido ao agrupamento genético ou cofenética ($C$); $\widehat{V}(F)$ variância estimada da matriz F e $\widehat{V}(C)$ a variância estimada da matriz $C$.

Há outros algoritmos desenvolvidos para análise de agrupamentos, subdividindo-se em dois métodos: hierárquicos e não hierárquicos.

- **Método hierárquico**

A aglomeração hierárquica se caracteriza pelo estabelecimento de uma hierarquia ou estrutura em forma de árvore (SNEATH; SOKAL, 1973). A aglomeração hierárquica interliga os grupos produzindo o dendrograma, uma representação gráfica em que os grupos semelhantes, segundo as variáveis estudadas, são agrupados entre si. No método hierárquico uma matriz de semelhança entre as informações estudadas e grupos se forma a partir das observações mais próximas, definidas nos processos de aglomeração ou divisão (SNEATH; SOKAL, 1973).

Após a escolha das variáveis para o processamento da análise de agrupamentos, três procedimentos devem ser realizados: (i) Padronização dos dados. A padronização da matriz de dados é realizada para que as unidades associadas às observações não afetem no grau de semelhança entre os grupos. Além disso, a padronização faz com que as informações em estudo contribuam com o mesmo peso no cálculo do coeficiente de semelhança entre grupos. As variáveis padronizadas passam a ter média zero e variância unitária; (ii) Escolha do coeficiente de semelhança. O coeficiente é importante para que se quantifique o quanto duas informações são semelhantes, sendo medidos de duas maneiras: Coeficiente de similaridade, que estabelece os padrões que servirão para a análise e o agrupamento dos grupos, sendo considerado coeficiente de similaridade aquele cujo maior valor observado representa a maior proximidade entre as informações estudadas; Coeficiente de dissimilaridade que indica que quanto maior o valor observado, menor é a proximidade e menos parecidos são os objetos (HAIR et al., 2005). (iii) Escolha da estratégia de agrupamento. Dentre os métodos de agrupamento mais utilizados, destacam-se aqueles que se caracterizam por serem sequenciais, aglomerativos, hierárquicos e sem sobreposição.

Outros algoritmos ou coeficientes de mensuração podem ser utilizados para quantificar a distância entre os objetos e analisar a similaridade entre eles. A escolha deve ser mediante os objetivos do estudo. Alguns coeficientes se adaptam melhor a determinadas análises, sendo os coeficientes de dissimilaridade mais adequados para variáveis quantitativas e os coeficientes de similaridade para as variáveis qualitativas.

**- Método de *Ward***

Para a determinação do número ótimo de *cluster* há alguns métodos que são mais utilizados. O Método de *Ward* ou método da soma de quadrados (SQD) dentro de *cluster* (4) tem como objetivo a minimização da soma de quadrados entre dois agrupamentos, realizada para as distâncias de todas as variáveis em estudo. Sendo assim, a distância é definida como a soma de quadrados de dois grupos (HAIR et al., 2005). Em cada estágio do procedimento de agrupamento, a soma interna de quadrados é minimizada sobre todas as partições que podem ser obtidas pela combinação de dois grupos do estágio anterior. Este método tende a combinar grupos com um pequeno número de observações e também tende a produzir grupos com aproximadamente o mesmo número de observações.

$$SQD(v) = \sqrt{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ijv} - \overline{x}_{jv})^2} \quad (4)$$

Em que, $SQD(v)$ é a soma de quadrados dentro dos *cluster* para a variável v, v = 1,..., p; $\overline{x}_{jv}$ a média amostral da variável v, no grupo *j-ésimo* grupo, $x_{ijv}$ a *i-ésima* observação da variável v, no *j-ésimo* grupo.

**- Distância Euclidiana**

A distância Euclidiana é o coeficiente de dissimilaridade mais conhecido e utilizado para quantificar a distância entre dois pontos quando todas as variáveis são quantitativas (SEIDEL et al., 2008). Quando o número de observações é

grande, a visualização gráfica fica comprometida. Sendo assim, torna-se importante a construção de um coeficiente que permita quantificar o quão próximo um objeto está de outro. Uma maneira de se obter a distância geométrica e o cálculo do coeficiente que distância física entre cada um deles é por meio da distância Euclidiana. Por exemplo, a distância entre objetos A e B em uma análise, de coordenadas respectivas ($x_1$, $y_1$) e ($x_2$, $y_2$), é definida (5):

$$d_{(A)(B)} = \sqrt{(x_{1(A)} - x_{2(B)})^2 + (y_{1(A)} - y_{2(B)})^2} \quad (5)$$

Em que, $d_{(A)(B)}$ é a distância Euclidiana entre as observações A e B; $x_{1(A)}$ e $y_{1(A)}$ são as coordenadas do ponto A; $x_{2(B)}$ e $y_{2(B)}$ são as coordenadas do ponto B.

Generalizando:

$$d_{(A)(B)} = \sqrt{\sum_{i=1}^{p}(x_{i(A)} - x_{i(B)})^2}$$

Em que, $d_{(A)(B)}$ é a distância Euclidiana; $x_{i(A)}$ e $x_{i(B)}$ são as coordenadas dos pontos considerados.

Aplicando-se para todos os pares da matriz de dados obtém-se a matriz de distâncias. A partir desta informação, é calculada a distância genética entre os indivíduos, sendo recalculadas as distâncias a cada par de indivíduos. Em seguida, a distância entre dois grupos deve ser calculada por meio da média entre os valores das distâncias de cada um dos grupos com os valores do outro, reagrupando as informações por similaridade em um *loop.* A distância entre quaisquer dois objetos não é afetada pela inserção de outros objetos ao conjunto de dados de análise e é baseado na média das distâncias (6):

$$d_{(C)(AB)} = \frac{d_{(C)(A)} + d_{(C)(B)}}{2} \quad (6)$$

- **Método não hierárquico**

No método de agrupamento não hierárquico primeiramente é determinado um centro de agrupamento e, em seguida, agrupam-se todos os objetos que estão a menos de um valor pré-estabelecido do centro.

**- K-*means* (*k*-médias)**

Este algoritimo visa alocar os objetos em grupos previamente definidos, buscando a melhor solução na formação de grupos, sendo o k o número de *cluster* e os centroides são pontos centrais dos grupos. Uma das formas de iniciar o processo é inserir o *k*, um número em um ponto observado no gráfico de determinação do número de *cluster*.

O ponto escolhido pode ser aleatório em qualquer lugar do plano e em seguida começam as iterações para se obter as soluções. A partir desse *k* inicial, as diferenças vão se reduzindo. A primeira iteração do algoritmo é para calcular a distância média de todos os pontos que estão associados ao centroide. Alguns pontos mudam de centroide e o *loop* ocorre até que nenhum ponto mude de centroide, ou seja, até que os centroides estejam na posição central da distância entre os pontos.

A utilização dos métodos hierárquicos conjuntamente com o não hierárquico é uma solução, pois ao utilizar o número de agrupamentos encontrado pelo método *Ward*, pode-se definir o número de agrupamentos que devem ser formados pelo método k-*means*, de modo a encontrar grupos internamente homogêneos (SEIDEL et al., 2008).

## 2.6 Análises de componentes principais (PCA)

A análise de componentes principais (PCA, do inglês *Principal Component Analyses*), foi introduzida por Karl Pearson em 1901 e fundamentada por Hotelling em 1933. Em 2003, foi introduzida pela primeira vez para o estudo de dados genéticos por Cavalli-Sforza e Feldman, com o objetivo de caracterizar a variação do DNA em humanos. A PCA permite estimar o grau de diferenciação populacional

(PATTERSON; PRICE; REICH, 2006) por meio da estrutura da variância e da covariância de um vetor aleatório, formado por p-variáveis aleatórias, por combinações lineares das variáveis originais denominadas de componentes principais e não correlacionadas entre si.

As análises multivariadas têm sido utilizadas há décadas para extrair vários tipos de informações a partir de dados genéticos (PATTERSON; PRICE; REICH, 2006; NOVEMBRE; STEPHENS, 2008). Em particular, a análise de componentes principais (PEARSON, 1901; HOTELLING, 1933) é um dos métodos estatísticos de múltiplas variáveis que visa reduzir a dimensionalidade dos dados capturando o máximo de variância possível. Além deste método, há a análise de agrupamento, que permite agrupar os animais por meio da similaridade genética e, assim, conhecer a estrutura da população e seus possíveis agrupamentos genéticos. Ambos os métodos utilizam informações de matriz de distâncias genéticas para detectar a estrutura dentro da população de forma menos paramétrica por meio da estrutura de covariâncias entre as características estudadas.

Utilizar todos os SNP do genoma em uma matriz é computacionalmente oneroso, sendo necessário otimizar as análises computacionais de grande conjunto de dados (JOMBART; DEVILLARD; BALLOUX, 2010). No entanto, todo o genoma contém aglomerados de SNP altamente correlacionados e a presença de agrupamento genético de indivíduos pode ter influência sobre os componentes principais.

A análise de componentes principais é uma metodologia mais complexa do que a análise de agrupamento que visa a redução de dimensão dos dados (JAMES et al., 2013), menor que a original, pela escolha das formas mais representativas por eliminar as sobreposições (JOMBART; DEVILLARD; BALLOUX, 2010). Os dados reduzidos por este método são aqueles que explicam a maior parte da variação, preservando a máxima informação contida nas variáveis originais. Cada componente principal é uma combinação linear das variáveis originais.

Por meio da PCA há a transformação das variáveis discretas e a variação reflete na variação da população, por representar a estrutura da população com base nas relações genéticas entre os indivíduos. Quando o estudo é com base em valores genéticos ou em informações genômicas, a variação dos componentes

principais representa as relações genéticas entre os indivíduos. Além disso, é medido o poder de cada variável com seu respectivo componente. As unidades amostrais são distribuídas em gráficos em que os eixos ortogonais são os componentes principais.

A PCA é uma maneira de identificar a relação entre características extraídas dos dados e é útil quando os vetores das características têm muitas dimensões e quando uma representação gráfica não é possível. O componente principal é o arranjo que melhor representa a distribuição dos dados e o componente secundário é perpendicular ao componente principal (Figura 1). A primeira direção dos componentes principais é aquela em que há maior variação das observações estudadas (JAMES et al., 2013). A medida da quantidade de informação explicada por cada componente principal é a sua variância. Por esta razão, os componentes principais são ordenados de acordo com a variância por ela explicada, sendo o componente principal que contém mais informação é o primeiro. Sendo assim, o último é aquele componente principal com menos informação.

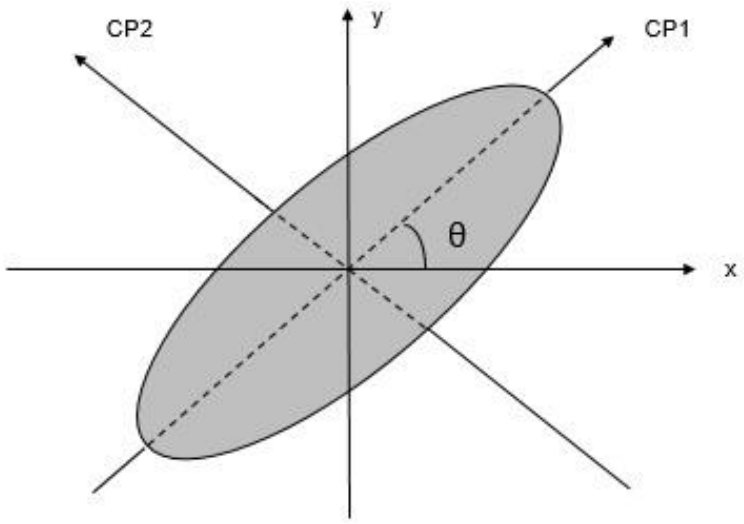


**Figura 1**. Eixos do componente principal (CP1) e secundário ($CP_2$), perpendicular, sendo a elipse a densidade.

## 2.7 Parâmetros populacionais baseados em informações genômicas

### 2.7.1 Parentesco genômico

Com o advento da tecnologia de genotipagem tornou-se possível a utilização de relações genômicas que podem estimar melhor a proporção de segmentos compartilhados pelos indivíduos. A genotipagem identifica alelos idênticos por descendência (IBD, do inglês *Identical by Descent*) e que podem ser compartilhados por meio dos ancestrais comuns, informações ausentes nos registros de pedigree.

A matriz de parentesco genômica contém as relações com estimativas da similaridade entre um grupo de indivíduos com base em dados de SNP, sendo uma medida de relação aditiva entre os indivíduos, ou ainda, uma matriz de correlação entre indivíduos corrigidos para a variação nas frequências alélicas (GONDRO, 2015). O mapa de calor (*heatmap*), construído por essas relações e parentesco, auxilia no entendimento da conexão entre indivíduos da mesma população e entre populações (GONDRO; VAN DER WERF; HAYES, 2013).

A utilização de matriz de parentesco genômica, denominada matriz G ou GRM (do inglês *Genomic Relationship Matrix*) (FORNI; AGUILAR; MISZTAL, 2011), é uma matriz de parentesco real ou realizada com base nas informações genômicas, sendo mais acurada que a matriz de parentesco com base nas informações de pedigree (matriz A), que é uma matriz de parentesco médio ou esperado, que pode conter erros de registros.

### 2.7.2 Endogamia

A endogamia é consequência do acasalamento entre animais aparentados capaz de alterar a constituição genética da população, aumentando a homozigose e a chance de dois alelos no mesmo *loci* serem idênticos por descendência, diminuindo a variabilidade populacional. Consequentemente, há diminuição da heterozigose, alterando, a frequência genotípica da população, mas não as frequências gênicas (QUEIROZ; ALBUQUERQUE; LANZONI, 2000).

Uma consequência importante no acasalamento entre indivíduos com ancestral em comum é que ambos podem carregar cópias de um gene do mesmo locos presente no ancestral. Se esses indivíduos se acasalarem podem passar essas cópias para sua progênie, aumentando o número de indivíduos autozigotos, ou seja, indivíduos com genótipos homozigotos idênticos por descendência (FALCONER; MACKAY, 1996).

A endogamia em uma população tem como consequência a redução do desempenho fenotípico médio do indivíduo, valores associados com a queda da capacidade reprodutiva ou eficiência fisiológica, fenômeno conhecido como depressão endogâmica (FALCONER; MACKAY, 1996). As perdas devido à depressão endogâmica podem ser recuperadas quando duas linhagens de raças puras são cruzadas em busca da diversidade genética, por meio do vigor híbrido (heterose), que é o fenômeno oposto da depressão por endogamia.

Nos programas de melhoramento de bovinos de leite tem-se constatado melhoria da produção, porém a endogamia pode ser um problema, tendo em vista os efeitos da depressão endogâmica, que além de reduzir a variabilidade genética, está associada com o aumento do risco de genes deletérios recessivos que causam doenças letais (KRISTENSEN; SORENSEN, 2005). Portanto, é necessário monitorar a endogamia e conhecer a estrutura genética e diversidade de raças, pois as populações sob seleção só têm aumentado (GUTIÉRREZ et al., 2003).

### 2.7.3 Heterozigosidade

A heterozigosidade é um importante parâmetro sobre a diversidade genética em uma população, pois valores baixos de heterozigosidade podem ser indício de “gargalo” ou de baixo fluxo gênico, o que pode ocorrer em rebanhos fechados. A alta heterozigosidade é indício de alta variabilidade genética na população (TORO; VILLANUEVA; FERNÁNDEZ, 2014). Além disso, há uma relação direta entre o aumento da endogamia e a diminuição da heterozigose para um dado locos em uma população fechada, não selecionada e de tamanho finito (WRIGHT, 1931).

A heterozigosidade é definida como a proporção de genótipos diploides constituídos por dois alelos distintos. A heterozigosidade se apresenta inversamente proporcional à herdabilidade, sendo assim, quanto maior a heterozigosidade, maior o efeito de dominância e menor o efeito aditivo. Pode ser utilizada como uma ferramenta rápida de melhoramento em características de baixa herdabilidade, como as características reprodutivas.

A heterozigosidade é máxima em cruzamento de raças puras. O cruzamento de *Bos taurus taurus* x *Bos taurus indicus* tem se destacado em rentabilidade em sistemas de produção no Brasil devido à presença de heterose significativa para as características economicamente importantes, como leite, rendimento, eficiência reprodutiva, longevidade produtiva e capacidade de sobrevivência (MADALENA et al., 1990). A redução da heterozigosidade está relacionada com os efeitos da endogamia, fatores como deriva genética e diferenciação genética entre as populações.

Nas análises de heterozigosidade são utilizadas informações de heterozigosidade esperada em vez de heterozigosidade observada, pois a esperada é calculada com base nas frequências alélicas dos SNP, permitindo assim, fazer inferência sobre a diversidade genética da população, diferentemente da heterozigosidade observada, que é a proporção de indivíduos heterozigotos em determinado locos em relação ao número total de indivíduos.

De acordo com o Teorema de Hardy-Weinberg, a heterozogosidade observada menor que a esperada (Ho < He) é indicativo de fixação de alelos ou alta endogamia, sendo que o oposto (Ho > He) é indício de alta variabilidade genética pelo maior fluxo gênico nos rebanhos. Se a população estiver em Equilíbrio de Hardy-Weinberg, os valores de He e Ho serão próximos. A maioria das mutações novas são perdidas, mesmo as que são favoráveis ao aumento da plasticidade das populações e as poucas mutações que se mantém na primeira geração têm probabilidade de se fixar (HARTL; CLARK, 2010). A probabilidade de fixação de genes na primeira geração é conhecida, sendo $\frac{1}{2N}$, sendo N o tamanho da população (GRIFFITHS et al., 2015). Então, quanto maior o tamanho da população, menor a taxa de fixação de alelos na primeira geraçao por mutação.

### 2.7.4 Desequilíbrio de ligação (LD)

Compreendido como uma associação não aleatória de alelos em dois ou mais *loci*, o LD é uma informação importante na biologia evolutiva, pois muitos fatores o afetam e são afetados por ele, além de fornecer informações sobre eventos passados. O estudo de LD na população permite a identificação e mapeamento dos QTL, pois pode ocorrer associações entre o SNP e o QTL (GODDARD, 1991; CAVALLI-SFORZA; FELDMAN, 2003). Ao longo do genoma, o LD reflete a história da população, o sistema de criação e o padrão de subdivisão geográfica (SLATKIN, 2008). O LD em cada região genômica reflete a história de seleção de genes, mutação e outras forças que afetam a evolução da frequência dos genes.

O estudo do LD pode proporcionar melhor entendimento da arquitetura genômica e da estrutura histórica da população (HAYES et al., 2009), além de ser relacionada com o tamanho da população efetiva e fornecer informações sobre a diversidade genética das populações (BIEGELMEYER et al., 2016). Sendo assim, o sucesso do mapeamento das regiões em LD depende do equilíbrio apropriado entre a extensão do LD e a densidade dos SNP (SARGOLZAEI et al., 2008). Além disso, o entendimento do padrão de LD nas populações é fundamental para determinar a densidade dos SNP necessária para se obter acurácia no GWAS e seleção genômica (BIEGELMEYER et al., 2016).

A seleção cria um estado de desequilíbrio de ligação e a recombinação genética tende a restaurar o equilíbrio de ligação (associação aleatória de genes nos gametas), sendo que a taxa de recombinação é correlacionada com a taxa de mutação. Desse modo, as regiões de reduzida recombinação têm menos mutações (HARTL; CLARK, 2010). Os fatores que afetam o LD entre um par particular de *loci* ou em uma região genômica são taxas de recombinação e número de gerações (7), redução do tamanho efetivo da população, processos naturais de deriva genética e cruzamento de populações distantes, alterando, assim, a extensão e variação do LD (HARTL; CLARK, 2010; O'BRIEN et al., 2014).

$$LD = (1 - r)^t * LD_0 \quad (7)$$

Em que, $LD$ é medido em uma geração, $LD_0$ é o LD inicial; $t$ é o número de gerações de cruzamento ao acaso; $r$ é a taxa de recombinação, variando de 0 a 0,5. O LD decairá rapidamente tendendo a zero se nenhum outro processo além da recombinação estiver agindo na população. Se a taxa de $r$ for pequena, o LD decairá lentamente.

O nível de LD é estimado a partir das frequências de alelos e de haplótipos em uma amostra. A definição do coeficiente de LD entre dois SNP pode ser realizada pelo $r^2$ (8), variando entre 0 e 1, como definido a seguir (HILL; ROBERTSON, 1968; HARTL; CLARK, 2010), ou ainda por D' (9). O $r^2$ tem a vantagem de ser uma medida mais precisa e exigir menor tamanho da amostra para detectar a associação entre um QTL e um SNP (O'BRIEN et al., 2014; PRITCHARD; PRZEWORSKI, 2001).

$$r^2 = \frac{(freq.AB * freq.ab - freq.Ab * freq.aB)^2}{freq.A * freq.a * freq.B * freq.b} \quad (8)$$

Em que, $freq.A$, $freq.a$, $freq.B$, $freq.b$ são as frequências dos alelos A, a, B e b, respectivamente, e $freq.AB$, $freq.ab$, $freq.Ab$ e $freq.aB$ são as frequência dos haplótipos AB, ab, Ab e aB na população, respectivamente.

Se dois *loci* são independentes, a frequência esperada do haplótipo AB, por exemplo, é calculada pelo produto entre $freq.A$ e $freq.B$. Se a $freq.AB$ observada for maior ou menor que o valor esperado, isso indica que esses dois *loci* tendem a segregar juntos e estão em LD. Quando a taxa de recombinação é máxima, 0,5 em *loci* de cromossomos diferentes, o decaimento de LD (*decay*) será rápido. Em novas populações, há maior extensão de LD, ou seja, maior distância observada entre dois *loci* em desequilíbrio, quando comparado com populações formadas há mais tempo. Em *loci* separados por longas distâncias, o número de gerações necessárias para atingir o equilíbrio é muito menor e o LD será quebrado.

$$D' = P_{AB} - (P_A * P_B) \quad (9)$$

Em que, D’ é desequilíbrio de ligação; $P_{AB}$ é a frequência do haplótipos; $P_A * P_B$ é o produto das frequências dos dois alelos A e B.

As medidas mais comumente empregadas para calcular o LD são D' (LEWONTIN, 1964), r² (HILL; ROBERTSON, 1968) e o $X^2$ padronizado (YAMAZAKI, 1977). No caso do marcadores bialélico, os dois últimos são equivalentes. Embora D' tenha sido comumente utilizado em estudos de LD em bovinos, esta medida tende a ser tendenciosa nos casos de pequenos tamanhos de amostras e marcadores com baixas frequências alélicas (SARGOLZAEI et al., 2008). Em simulações realizadas (ZHAO; NETTLETON; DEKKERS, 2007), o D' superestimou a quantidade de LD para marcadores bialélicos, portanto r² seria o preditor mais adequado para estimar LD.

No momento há pouco conhecimento sobre o LD do bovino Gir brasileiro, informação de grande importância para determinação do melhoramento da raça no país.

### 2.7.5 Tamanho efetivo populacional ($Ne$)

A preocupação com as perdas econômicas relacionadas com o fenômeno da depressão endogâmica, definido como declínio do valor fenotípico de uma característica (FALCONER; MACKAY, 1996) e a diminuição da variabilidade genética, é de grande interesse e tem conduzido a estudos sobre a estrutura da população envolvendo os parâmetros populacionais. A distribuição e quantidade da variância genética dentro de populações dependem não só do tamanho real ou número total de indivíduos da população, mas do tamanho da população geneticamente efetiva ($Ne$), sempre de menor tamanho que a população real.

Para o $Ne$ são considerados o número de indivíduos que irão contribuir com a manutenção da diversidade genetica para a geração seguinte. Este número de indivíduos são aqueles que de fato dão origem à variação da amostragem calculada ou ainda à taxa de endogamia (FALCONER; MACKAY, 1996). O Ne é um parâmetro populacional que auxilia o entendimento sobre a evolução, expansão ou redução da população e no entendimento da estrutrura populacional.

Por meio do conhecimento de $Ne$ é possível predizer a taxa de redução da variabilidade genética em uma população isolada e o tempo necessário para que

essa variação reduzida se torne uma ameaça à variabilidade genética. O tamanho efetivo da população é diretamente afetado por fatores genéticos, como seleção, mutação, deriva, migração e acasalamentos não aleatórios, além de fatores não genéticos.

Há diversas maneiras de calcular Ne (RELETHFORD, 2012):

(i) Considerando-se o número de gerações, uma maneira de calcular é por meio da média harmônica:

$$Ne = \frac{t}{\sum\left(\frac{1}{N}\right)}$$

Em que, para $^{1}/_{N}$ são somados todos os indivíduos da geração; $t$ é número de gerações. Dessa forma são considerados mudanças no tamanho da população (N) ao longo das gerações.

(ii) Considerando-se a fertilidade. Uma suposição simplificada realizada, considerando a deriva genética, é que a variação do número de descendentes é aleatória.

$$Ne = \frac{4\,N - 4}{2 + V}$$

Em que, N é o tamanho da população em acasalamento; V é a variância no número de filhos.

iii) Em razão do sexo. Se a população reprodutiva consiste de $Nm$ machos e $Nf$ fêmeas, o tamanho efetivo populacional será:

$$Ne = \frac{4\,Nm\,Nf}{Nm + Nf}$$

A perda de heterozigosidade por geração será ½ $N$, em que $N$ é o número de indivíduos de uma população. O $Ne$ também determina a taxa de deriva genética. Por exemplo, um $Ne$ < 20 significa que menos de 20% dos machos

acasalaram com todas as fêmeas. O menor Ne sugere baixa diversidade genética, o que representa uma preocupação para a viabilidade das espécies.

iv) Em razão do LD. O LD pode ser usado para estimar $Ne$ se a taxa de recombinação é conhecida (HAYES et al., 2003). O LD em grandes distâncias de recombinação estima com maior precisão o $Ne$ em um passado recente, quando o LD é comparado em distâncias curtas de recombinação.

Usar LD em vez da homozigose de marcador individual tem a vantagem de a taxa de recombinação ser mais controlável do que a taxa de mutação, então o $Ne$ mais recente pode ser estimado, pois as taxas de recombinação podem ser muito maiores do que as taxas de mutação (HAYES et al., 2003).

$$Ne = \frac{\left(\left(\frac{1}{r^{2*}}\right)-1\right)}{4c}$$

Em que, c é a distância média de recombinação e $r^{2*}$ é igual a $r^2$ em todos os pares de SNP em dado alcance de $c$, distância do marcador em Morgans, em que 1 Morgan é igual a 100.000.000 pares de bases). Dado que 1 centimorgan (cM) é igual a 1 megabase (Mb), a média de $r^2$ é calculada em pares de base (pb) para todos os SNP e convertida em Mb, multiplicando $r^2$ por $1x10^{-8}$.

Em 2009, no projeto consorciado para estudos do genoma completo de bovinos foi constatado que o Ne recente tem diminuído para todas as raças, como consequência do fenômeno denominado “gargalo”, em que há a redução do número de fundadores devido à domesticação, formação de raças e seleção para produção. Apesar do declínio do Ne, entre os subgrupos taurinos e zebuínos a diversidade não é considerada baixa (ELSIK et al., 2009).

## 3 MATERIAL E MÉTODOS

### 3.1 Dados genômicos

Os dados genômicos deste estudo foram cedidos pela Embrapa Gado de Leite, situada em Juiz de Fora, Minas Gerais e pertencem ao projeto “Seleção

genômica de raças leiteiras no Brasil". As análises foram realizadas utilizando informações genômicas de duas gerações de 2.279 bovinos da raça Gir genotipados com painel de alta densidade, sendo 601 touros genotipados com o painel 777K SNP, considerando o mapa BovineHD BeadChip, e 1.678 vacas genotipadas com o painel 50K SNP, considerando o mapa BovineSNP50 BeadChip versão 2, ambos genotipados pela plataforma Illumina (ILLUMINA, 2009).

Para a genotipagem foram escolhidos todos os touros que participam ou participaram do teste de progênie ou pré-teste e que possuíam DNA disponível para as análises. Para a genotipagem das vacas foram escolhidas aquelas que possuíam DNA disponível, registros de lactação e, preferencialmente, oriundas de fazendas com maior número de animais.

### 3.2 Controle de qualidade

A edição dos dados foi realizada por meio do software estatístico R versão 3.3.1 (R CORE TEAM, 2016), sendo considerados para as análises somente SNP em cromossomos autossômicos e SNP com posição conhecida no mapa UMD_3.1 bovine assembly (ZIMIN et al., 2009).

Para o controle de qualidade dos genótipos foram excluídos os SNP com *call rate* inferior a 0,95 e *genotype calling score* (GC score) inferior a 0,85. Além disso, foram excluídos os SNP com frequência do alelo raro (MAF) menor que 1%, SNP monomórficos e sem posição no genoma. Foram considerados os *pruning* de SNP (SNP redundantes) por meio de uma correlação de Pearson para evitar superestimação das correlações entre os marcadores, baseados na proporção de alelos idênticos por estado (IBS, do inglês *Identical by State*), sendo eliminados aqueles com correlação acima de 98%.

Para o controle de qualidade das amostras foram excluídos animais com *call rate* inferior a 0,90 e desvios de heterozigosidade com mais ou menos três desvios-padrões em relação à heterozigose média esperada. Após o controle de qualidade permaneceram no arquivo informações de 1.987 animais, sendo 525 touros nascidos entre 1960 e 2012 e 1.462 fêmeas nascidas entre 1982 e 2006, provenientes em 154 fazendas (Apêndice 1) de diferentes regiões do Brasil. Foram

utilizados apenas SNP em comum entre os painéis HD e 50K SNP, em um total de 21.727 SNP.

## 3.3 Análise dos dados genômicos

### 3.3.1 Matriz de distância genética

Para a obtenção agrupamentos genéticos entre os indivíduos e dos componentes principais, por meio de informações de SNP, foram realizados os seguintes passos:

(i) Identificação dos alelos IBS a cada par de SNP,

(ii) Cálculo da distância genética entre indivíduos,

(iii) Cálculo do coeficiente de similaridade e dissimilaridade,

(iv) Técnicas de agrupamento e análise de componentes principais,

(v) Cálculo do coeficiente de correlação cofenética.

O cálculo do número total de alelos idênticos por estado (IBS) para todos os pares de todos os SNP bialélicos foi realizado por meio do software estatístico R (R CORE TEAM, 2016), pacote “snpStats”, função “IBScount”. A função “IBSdist”, do mesmo pacote e software, transforma a matriz de distância Euclidiana em uma matriz de distâncias genéticas e os menores valores da matriz agrupam os indivíduos, reduzindo assim, o número de dimensões dos dados pela reunião de pares semelhantes.

### 3.3.2 Análise de agrupamento

A determinação do número de agrupamentos ou *cluster* foi realizada utilizando a distância genética entre os indivíduos obtida por meio dos painéis de SNP reduzidos a 21K SNP. Três métodos foram utilizados para a clusterização: Método *Ward* (soma de quadrados dentro dos *cluster*), Método de *Elbow* e Método de *Silhouette*, por meio do algoritmo *k-means* implementados pelo software estatístico R (R CORE TEAM, 2016). O pacote “factoextra” foi utilizado nas análises de agrupamento para verificar subestruturas na população.

### 3.3.3 Análise de componentes principais (PCA)

A PCA foi utilizada para investigar a estrutura da população estudada a partir da matriz de distâncias genéticas entre os indivíduos. A análise foi realizada por meio da função “prcomp” do software estatístico R (R CORE TEAM, 2016), com o objetivo de estimar o grau de diferenciação populacional (PATTERSON; PRICE; REICH, 2006). Não houve a necessidade de padronização dos dados, que pode ser feita com média zero e variância um, ou com variância um e qualquer média, pois as unidades de medidas das variáveis observadas são as mesmas.

Os componentes principais foram gerados a partir da matriz de correlações dos dados, variâncias e covariâncias (Figura 2). Por meio da decomposição espectral, autovalores e autovetores foram gerados. Os autovalores estão relacionados com o comprimento e a variação dos componentes principais. Associado a cada autovalor existe um vetor denominado autovetor, que está associado à intensidade e direção. Os vetores são fatores de ponderação (ou escores) e sentido da contribuição das variáveis originais em cada componente principal.

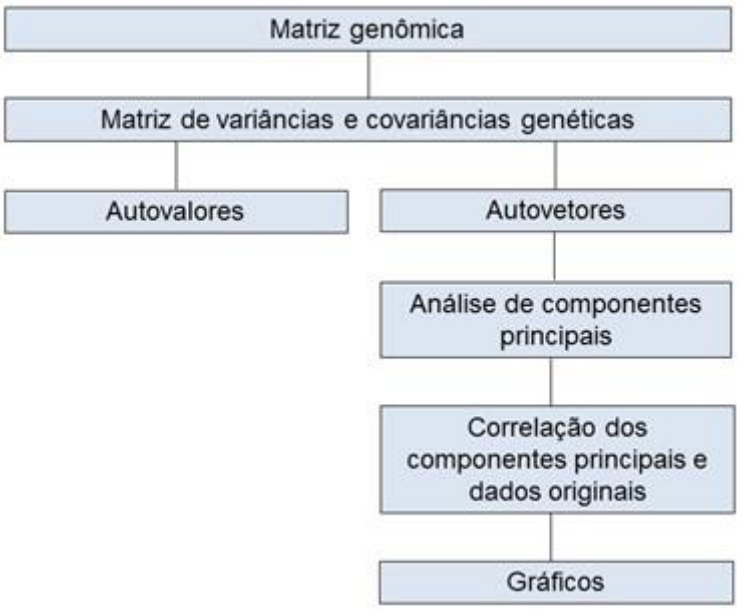


**Figura 2.** Etapas utilizadas para a análise de componentes principais.

Para a caracterização dos vinte touros que mais e menos contribuíram para a variância dos componentes principais, buscou-se informações sobre o número de filhos genotipados, idades, *cluster*, PTA (do inglês *Predicted Transmiting Hability*, habilidade prevista de transmissão) para produção de leite ajustado para 305 dias e STA (do inglês *Standandized Transmiting Hability*, capacidade prevista padronizada) das características de conformação do sistema mamário, informações utilizadas para complementar a discussão dos resultados deste estudo (Apêndices 2 e 3).

A PTA é uma estimativa do potencial genético de um touro para a característica de produção ou de tipo, que poderá ser transmitida à sua progênie em comparação com a base genética (SANTOS; CORRÊA, 2000). Desta forma, é uma medida do desempenho esperado das filhas do touro em relação à média genética da população (PANETTO et al., 2016), gerada pelas avaliações dos teste de progênie dos touros leiteiros e deve serutilizada para auxiliar na seleção e descarte de animais. Para tornar as diferentes características de produção e de tipo comparáveis, esses índices estimados são padronizados (STA), dividindo-se a PTA do touro pelo desvio-padrão da PTA da característica obtida para todos os touros avaliados.

A partir da PCA, os touros foram selecionados por meio de um *ranking* considerando um índice ponderado para os três primeiros componentes principais, de acordo com a proporção na contribuição da variância de cada componente principal na variância total. Os valores dos componentes principais são independentes do sinal (positivo ou negativo), que indica a direção dos componentes principais em um espaço p-dimensional. A indicação do sinal não tem efeito na avaliação da variação dos componentes principais (JAMES et al., 2013), que são avaliados apenas pela sua magnitude.

Em adição, dois touros importados da Índia, “Naidu Imp”, nascido em 1960, e “Gaiolão DC”, nascido em 1977, foram prospectados pela sua importância na contribuição da variância dos componentes principais. Esses touros são parte da população base e suas progênies são parte da população base de outros plantéis.

### 3.3.4 Parâmetros populacionais

Neste estudo os parâmetros populacionais foram obtidos a partir de informações genômicas de: matriz de parentesco genômico, coeficiente de endogamia, heterozigosidade esperada e observada, desequilíbrio de ligação e tamanho efetivo populacional. Para isso, foram consideradas informações do painel de SNP 21K SNP e não foram incluídas as informações de registros de pedigree.

- **Parentesco genômico**

A matriz de parentesco genômico (H ou GRM, do inglês *Genomic Relationship Matrix*), foi calculada por meio do software estatístico R (R CORE TEAM, 2016), pacote “BGLR”, em que se aplica modelo de regressão bayesiana paramétrica. A representação gráfica da matriz de parentesco foi obtida por meio do pacote “pheatmap” (KOLDE, 2016) do mesmo software. Na diagonal da matriz foram representadas estimativas dos coeficientes de parentesco genômico dos indivíduos com eles mesmos, ou seja, endogamia, e fora da diagonal, estimativas do parentesco genômico para os todos os pares de indivíduos. Os dendogramas nos limites da matriz agruparam os animais de acordo com o parentesco genômico.

Por meio do teste de Smith – Satterthwaite (10) foi possível verificar diferença significativa entre a distribuição da probabilidade dos indivíduos dos diferentes *cluster*.

$$t = \frac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{s_x^2/n_x+s_y^2/n_y}} \qquad (10)$$

Em que, sob $H_0$ verdadeira ($\mu_x = \mu_y = 0$), t aproxima-se do teste t de *Student*, tal que $\bar{x}$ e $\bar{y}$ são as médias dos subgrupos, $s_x^2$ e $s_y^2$ são as variâncias dos subgrupos, $n_x$ e $n_y$ são os tamanhos das amostras dos subgrupos populacionais.

- **Heterozigosidade**

A heterozigosidade esperada (He) e observada (Ho) foram calculadas por meio da função “hardy” para cada SNP bialélico do programa PLINK v.1.9 (PURCELL et al., 2007). A Ho é calculada pela contagem da frequência de genótipos (11) e a He é calculada pela frequência de alélica, 2pq.

$$Ho = (N(NM) - O(Hom))/N(NM) \quad (11)$$

Em que, $N(NM)$ é o número total de genótipos não perdidos, $O(Hom)$ é o genótipo homozigoto observado.

- **Coeficiente de endogamia**

O coeficiente de endogamia genômico (F) foi estimado por meio da função “het” do programa PLINK v.1.9 (PURCELL et al., 2007), em que se utiliza a contagem de genótipos homozigotos autossomos esperados e observados. Os valores de F estimados fortemente negativos, com taxas de homozigotos mais baixas que o esperado, podem indicar contaminação de amostra (PURCELL et al., 2007), sendo considerados apenas valores dentro de dois e meio desvios-padrão (|F| < 0,10), variando de 0 a 1 (12):

$$F = \frac{(homozigoto\ observado - homozigoto\ esperado)}{(total\ de\ SNP - homozigoto\ esperado)} \quad (12)$$

Para o cálculo da frequência de animais endogâmicos e de coeficiente de endogamia foram considerados seis possíveis valores de F (0,05, 0,01, 0,02, 0,03, 0,04, 0,05) e de N% (0,50, 1,00, 5,00, 10,00, 25,00, 50,00), respectivamente. Duas formas de calcular por meio da Distribuição de Qui-Quadrado (Tabela IV) (13) foram utilizadas: (i) fixando valores de endogamia ($F$) para obter a porcentagem de animais endogâmicos ($x$) e (ii) fixando a probabilidade da porcentagem da animais (N%) para obter os valores de endogamia (e).

$$1 - P(X_n^2 \leq x) \quad (13)$$

Em que, a probabilidade ($P$) é calculada considerando que o valor $X_n^2$ (valor tabelado) é menor ou igual aos valores especificados de endogamia ($x$).

- **Desequilíbrio de ligação (LD)**

A extensão do LD foi mensurada para todos os pares de SNP com tamanho de janela de 50 Mb, por meio do coeficiente $r^2$ com frequência alélica padronizada (HILL; ROBERTSON, 1968). Por meio da função "$r^2$" do programa PLINK v.1.9 (PURCELL et al., 2007), correlações da contagem de alelos variantes foram calculados para cada par de SNP adjacente bialélico.

Para analisar o decaimento de LD (*decay*), os pares de SNP foram classificados em classes com intervalos baseados na distância física entre os marcadores. A representação gráfica do decaimento de LD em função do aumento das distâncias entre os marcadores foi obtida por meio do software estatístico R (R CORE TEAM, 2016). A distância de 100 Kb foi adotada para cálculo de média de $r^2$.

- **Tamanho efetivo populacional (Ne)**

O Ne foi estimado em razão do LD (14) por meio do software estatístico R (R CORE TEAM, 2016), em que c é a distância média de recombinação e $r^{2*}$ é o $r^2$ médio em todos os pares de SNP em dado alcance de c. Dado que 1 centimorgan (cM) é igual a 1 megabase (Mb), a média de $r^2$ calculada em pares de bases (pb) para todos os SNP foi convertida em Mb (dividida por $1x10^6$) e depois convertida em cM (dividida por $1x10^2$).

$$Ne = \frac{\left(\left(\frac{1}{r^{2*}}\right) - 1\right)}{4c} \quad (14)$$

## 4 RESULTADOS E DISCUSSÃO

Na análise descritiva dos SNP (Tabela 1), os tamanhos dos cromossomos foram obtidos por meio da plataforma do *National Center for Biotechnology Information* (NCBI, 2017). Observou-se maior número de SNP no cromossomo 1, o de maior tamanho, sendo que a densidade de SNP variou entre os cromossomos, de 375 a 1.498 SNP. A média de MAF por cromossomo variou de 0,1682 a 0,1938, sendo que o cromossomo 5 apresentou a maior média (Tabela 1).

**Tabela 1**. Análise descritiva dos marcadores com base nos cromossomos

| CHR[1] | Tamanho CHR[2] | nº SNP/CHR[3] | % SNP/CHR[4] | MAF [5] |
|---|---|---|---|---|
| 1 | 161,11 | 1498 | 6,89 | 0,1706 |
| 2 | 140,68 | 1144 | 5,26 | 0,1743 |
| 3 | 127,87 | 1007 | 4,63 | 0,1826 |
| 4 | 124,43 | 1027 | 4,73 | 0,1752 |
| 5 | 125,64 | 852 | 3,92 | 0,1938 |
| 6 | 122,65 | 1246 | 5,74 | 0,1892 |
| 7 | 111,95 | 919 | 4,23 | 0,1739 |
| 8 | 116,94 | 1024 | 4,71 | 0,1690 |
| 9 | 108,10 | 924 | 4,25 | 0,1853 |
| 10 | 106,31 | 919 | 4,23 | 0,1711 |
| 11 | 110,26 | 927 | 4,27 | 0,1834 |
| 12 | 85,44 | 708 | 3,26 | 0,1682 |
| 13 | 84,43 | 719 | 3,31 | 0,1883 |
| 14 | 81,41 | 758 | 3,49 | 0,1850 |
| 15 | 84,80 | 682 | 3,14 | 0,1855 |
| 16 | 77,91 | 744 | 3,43 | 0,1731 |
| 17 | 76,52 | 660 | 3,04 | 0,1809 |
| 18 | 65,95 | 554 | 2,55 | 0,1816 |
| 19 | 65,32 | 539 | 2,48 | 0,1755 |
| 20 | 75,86 | 676 | 3,11 | 0,1753 |
| 21 | 69,31 | 571 | 2,63 | 0,1798 |
| 22 | 61,89 | 557 | 2,56 | 0,1878 |
| 23 | 53,33 | 455 | 2,09 | 0,1892 |
| 24 | 65,02 | 564 | 2,59 | 0,1732 |
| 25 | 44,04 | 375 | 1,73 | 0,1900 |
| 26 | 51,86 | 450 | 2,07 | 0,1803 |
| 27 | 48,75 | 420 | 1,93 | 0,1879 |
| 28 | 46,11 | 400 | 1,84 | 0,1880 |
| 29 | 52,13 | 410 | 1,89 | 0,1732 |

[1]Cromossomo, [2]Tamanho do cromossomo, em megabases (Mb), [3]Número de SNP por cromossomo, [4]Porcentagem de SNP por cromossomo, [5]Média da frequência do alelo raro (MAF) por cromossomo.

- **Análise de agrupamento**

Nas análises de agrupamento utilizando o painel de 21K SNP foi possível identificar estrutura populacional (Figura 3), pois as análises utilizando poucos SNP poderá prejudicar a detecção de *cluster* na população (VENTURA et al., 2016).

Pelo método de *Ward*, ou método da soma de quadrados dentro de cluster (SQD), houve a minimização da soma de quadrados dentro dos agrupamentos a cada *loop*, pela combinação dos dados (Figura 3a). O método de *Elbow* avaliou a porcentagem da variância explicada em função do número de agrupamentos e, quando o acréscimo de um novo grupo deixou de ser relevante, a trajetória entrou em movimento retilíneo, pois a diferença entre os dados foi quase insignificante (Figura 3b). O método de *Silhouette* mede a qualidade de cada objeto por meio de um grau de confiança, e o ápice no gráfico indicou o número ótimo de cluster (Figura 3c), ou seja, quanto mais próximo de 1, mais próximo ao número provável de agrupamento.

Nos métodos de *Ward* e *Elbow*, quando a população foi dividida em até 4 grupos, ou seja, em 2, 3 e 4, as SQD apresentaram diferenças maiores (Figura 2). Quando a população foi dividida de quatro a 10 grupos, as mudanças nas SQD se estabilizaram, sugerindo que a melhor escolha para a divisão da população foi de quatro grupos. Sendo assim, nas análises genômicas como seleção e associação, deve-se levar em conta esses agrupamentos.

O método de *Ward* considera em suas análises os *cluster* que têm tamanhos diferentes, mas ainda assim, as trajetórias dos dois métodos foram parecidas. Apesar de serem análises subjetivas, por estarem sujeitas a interpretações dos gráficos, todas apontaram para o mesmo resultado (Figura 3). Portanto, quatro é o número provável de *cluster*.

Os agrupamentos genéticos são compostos por 237, 287, 474 e 989 animais nos *cluster* 1, 2, 3 e 4, respectivamente, em um total de 1.987 animais. As variâncias dos três primeiros componentes dentro de cada *cluster* foram baixas, próximas de zero (Tabela 2), ou seja, as distâncias genéticas dos indivíduos dentro de cada *cluster* foram pequenas, sendo maior para o *cluster* 3 no primeiro componente principal, aquele de maior variância. Então, o *cluster* 3 foi o

agrupamento genético que mais se destacou dos demais, com maior dissimilaridade genética em relação aos demais agrupamentos genéticos.

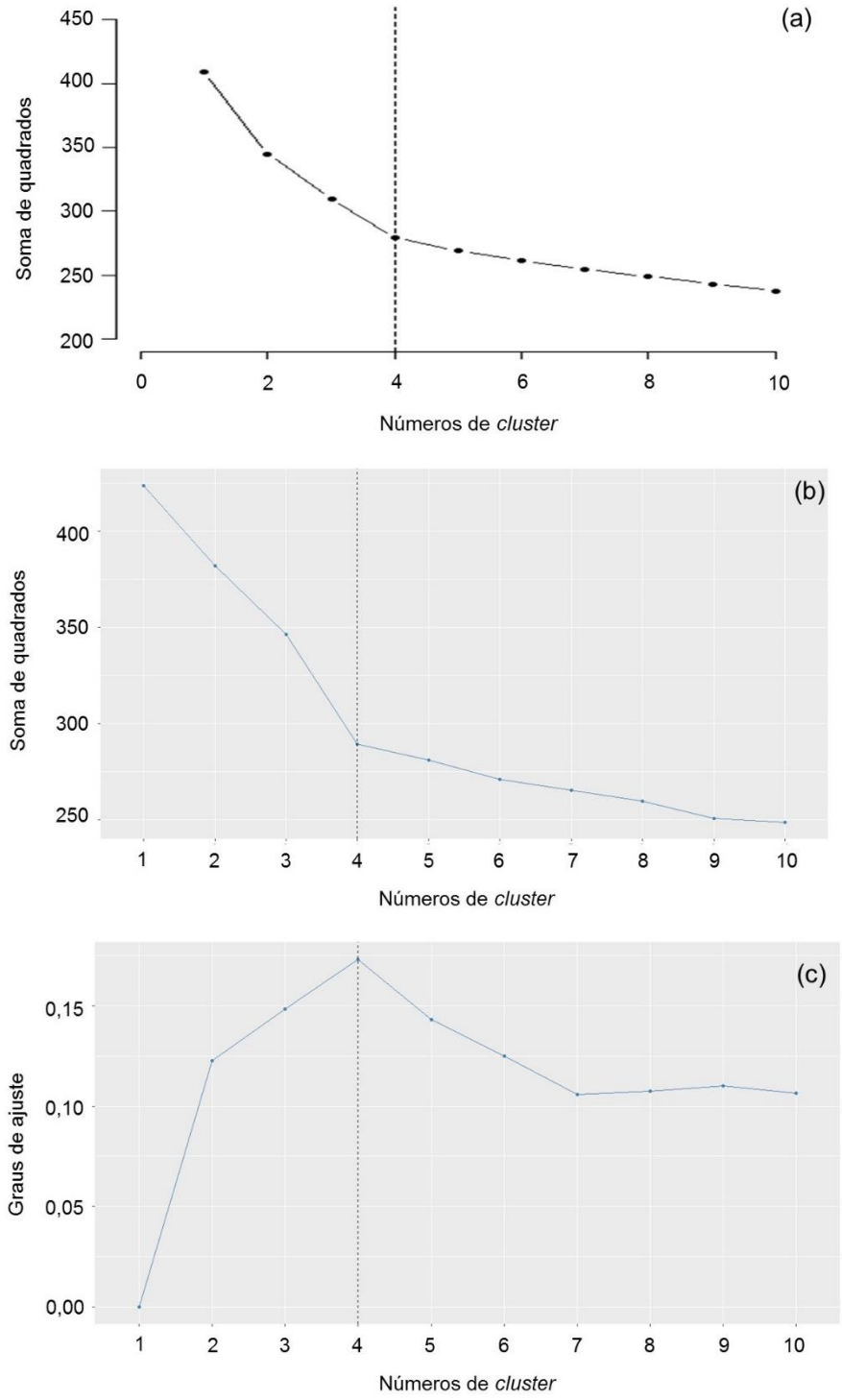


**Figura 3**. Determinação do número ótimo de *cluster* obtidos pelos Métodos de *Ward* (a), *Elbow* (b), pela minimização da soma de quadrados, e *Silhouette* (c), pelo maior grau de ajuste.

O coeficiente de correlação cofenética, calculado com base nas distâncias genéticas e agrupamentos dos animais, foi equivalente a 0,66, indicando moderada-alta qualidade dos agrupamentos, de acordo com a metodologia utilizada (SOKAL; ROHLF, 1962).

**Tabela 2**. Variâncias das distâncias genéticas dos indivíduos dos *cluster* 1, 2, 3 e 4 nos três primeiros componentes principais (PC1, PC2 e PC3)

| | PC1 | PC2 | PC3 |
|---|---|---|---|
| *Cluster* 1 | $3 \times 10^{-5}$ | $1 \times 10^{-4}$ | $2 \times 10^{-4}$ |
| *Cluster* 2 | $5 \times 10^{-5}$ | $8 \times 10^{-5}$ | $2{,}4 \times 10^{-5}$ |
| *Cluster* 3 | $1{,}5 \times 10^{-4}$ | $3{,}8 \times 10^{-5}$ | $5{,}1 \times 10^{-5}$ |
| *Cluster* 4 | $2{,}3 \times 10^{-5}$ | $3 \times 10^{-5}$ | $3{,}6 \times 10^{-5}$ |

Nas projeções do dendograma (Figura 4) observou-se as relações genéticas entre os quatro agrupamentos. Na análise de componentes principais foi possível observar a relação da distância genética com os três primeiros eixos na nova dimensionalidade dos dados. O *cluster* 2, de coloração verde, contém os animais com maior influência sobre o primeiro componente e o *cluster* 3, de coloração azul, sobre o segundo componente. Os animais do *cluster* 3 foram aqueles com menor distância genética do *cluster* 4 e os animais dos *cluster* 1 e 2 foram os mais distantes geneticamente do *cluster* 4.

A média da distância genética obtida por IBS foi 0,2563, valor próximo aos encontrados na literatura para bovinos africanos Bonsmara (0,24), Angus e Holandês (0,25) (MAKINA et al., 2014). Sendo assim, em média, a dissimilaridade genética dos indivíduos é de aproximadamente 26%, então esses indivíduos compartilham entre si cerca de 74% dos SNP. A diferenciação genética de 26% é o que diverge um indivíduo do outro e essa diferença pode ser atribuída à diferentes ancestrais, animais que fizeram parte da população base, à seleção, que pode ter favorecido alguns grupos de alelos para um grupo de indivíduos, ou ainda devido às mutações causais.

As distâncias genéticas são reflexo da variabilidade genética na população, o que permite que os programas de melhoramento apresentem alternativas para as mudanças nos objetivos de seleção, dado que a população possui diversidade genética.

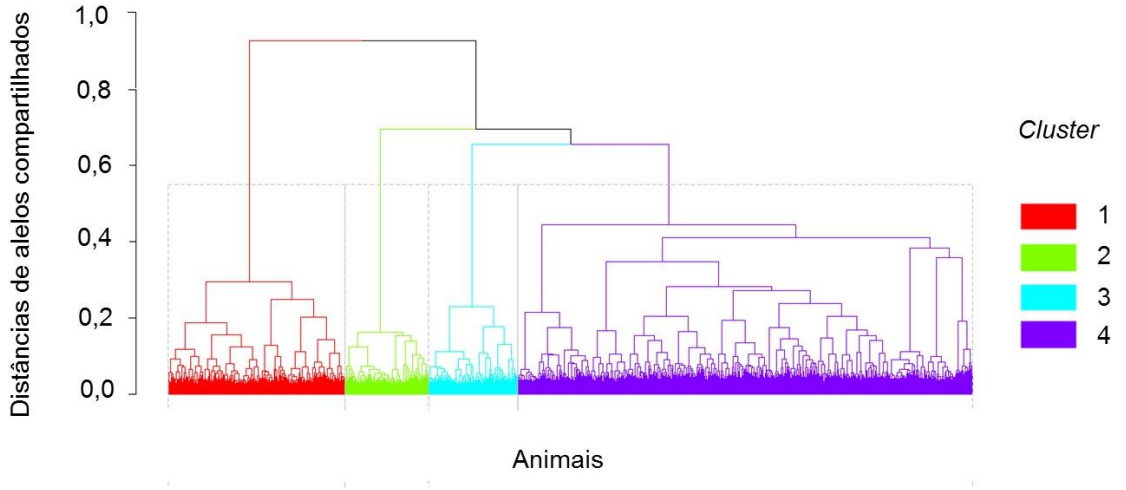


**Figura 4**. Dendograma da matriz de distância genética de alelos compartilhados colorido para os diferentes *cluster*.

Nas projeções das distâncias genéticas dos indivíduos nos componentes principais apresentadas pelas coordenadas rotacionadas em 3D (Figura 5) há quatro agrupamentos genéticos destacados, representados nesta população, sendo possível observar a formação de uma pirâmide a partir da distribuição das distâncias genéticas, sendo as arestas da pirâmide os diferentes agrupamentos genéticos, definidos como linhagens. A formação dos agrupamentos pode ser atribuída à origem dos progenitores com constituições genéticas distintas, embora sejam da mesma raça.

O uso intensivo dos mesmos reprodutores pode também ter contribuído para a formação dos diferentes *cluster*. A escolha dos progenitores pode ter causado os diferentes direcionamentos na estrutura da população, pois os touros com adequado desempenho para atender os objetivos de seleção em testes de progênie são provavelmente os mais procurados nas centrais de inseminação artificial.

Portanto, para a manutenção da variabilidade genética em futuras gerações de bovinos leiteiros da raça Gir, deve-se investir em estratégias de acasalamento que reduzam a endogamia e que não haja o uso maciço de apenas alguns reprodutores de alto valor genético (REIS FILHO et al., 2010). A utilização em larga escala de poucos touros incrementa o intervalo de geração e a variância do

tamanho da família, que é uma das principais causas da diminuição do tamanho efetivo da população (FALCONER;  MACKAY, 1996).

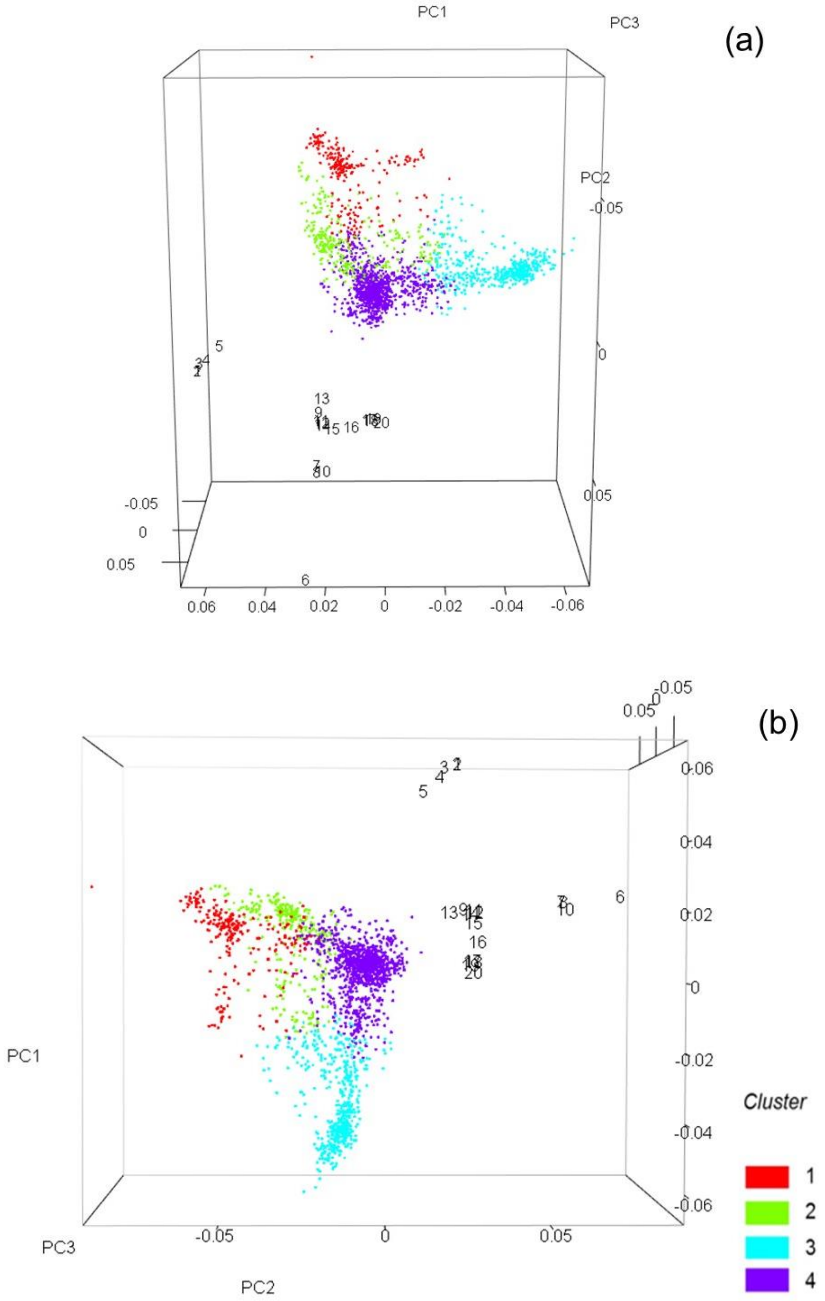


**Figura 5**. Projeção dos indivíduos em 3D, considerando os três primeiros componentes principais (PC1, PC2, PC3) em duas faces, anterior (a) e esquerda (b), coloridas para os diferentes cluster, e os touros mais influentes na variância dos componentes principais, enumerados.

A utilização em conjunto dos resultados das análises de agrupamento nas predições genômicas poderá construir os grupos de animais de forma mais eficiente para as análises (VENTURA et al., 2016), sendo que a distribuição dos animais nos grupos de treinamento e validação abrangeria todos os grupos genéticos de indivíduos da população estudada.

- **Análise dos componentes principais (PCA)**

Por meio das relações dos indivíduos com os componentes principais foi possível observar a variabilidade genética da população (Figura 5). Os indivíduos do *cluster* 4 estão mais próximos dos eixos zero dos três componentes principais, portanto são os que menos influenciaram na variância dos componentes principais. O *cluster* 4 contém alguns animais da população base do Gir, como os importados, que são os touros mais antigos desta população. Conforme os animais foram utilizados na reprodução, houve aumento da variabilidade genética entre grupos e essa dispersão do material genético determinou as diferentes linhagens leiteiras.

O primeiro componente principal (PC1) explicou 19,8% da variância das distâncias genéticas e o segundo componente principal (PC2) explicou 13,0% (Figura 6). Os três primeiros componentes principais foram escolhidos para realização das demais análises por somarem grande parte da variação dos dados (42,4%).

Na projeção dos dois primeiros componentes principais (Figura 6) é possível observar a distribuição dos animais em uma nova dimensionalidade das distâncias genéticas e que estas ocupam os quatro quadrantes subdivididos no espaço amostral 2D. A distribuição dos dados no espaço bidimensional e a proporção da variância explicada pelos três componentes principas são indicativo de que a amostra da população escolhida para ser genotipada, ou seja, o delineamento para a genotipagem dos animais foi suficiente para se conhecer os agrupamentos genéticos nessa população, considerando que os touros escolhidos para genotipagem participam ou participaram do teste de progênie. Sendo assim, são touros amplamente disseminados na população e com menor parentesco entre si.

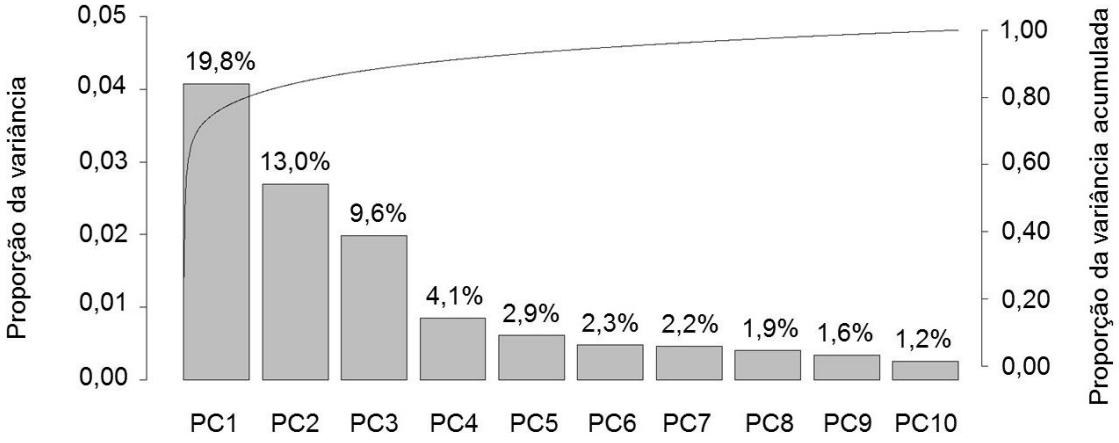


**Figura 6**. Proporção das variâncias dos dez primeiros componentes principais (barras) e proporção da variância acumulada de todos os componentes principais (linha contínua).

Em um plano espacial 2D observou-se a contribuição genética dos animais na variância dos dois primeiros componentes principais (Figura 7). Os animais concentrados em torno dos eixos zero dos dois primeiros componentes principais são aqueles que possuem menor contribuição na variância dos componentes principais.

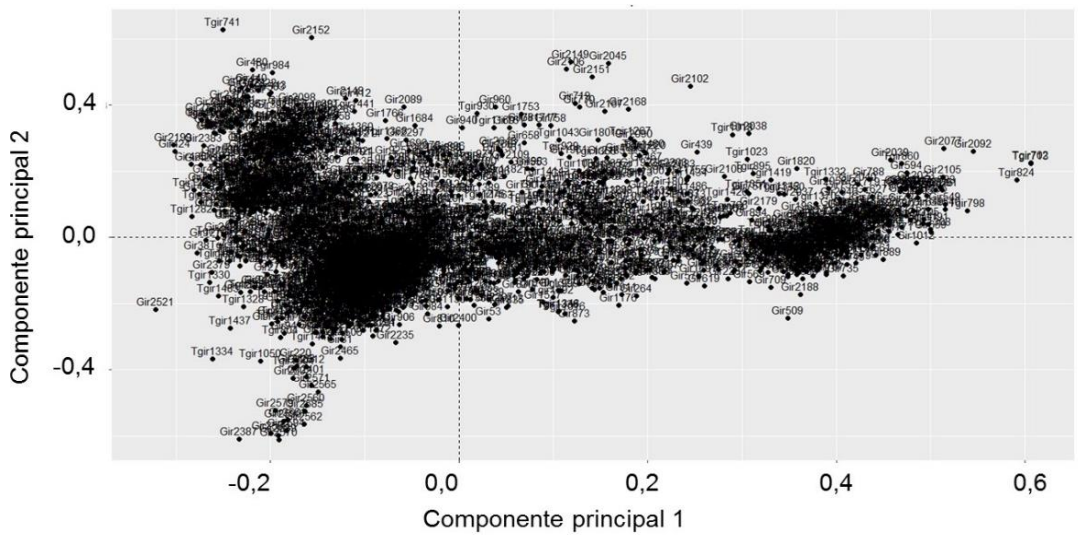


**Figura 7**. Projeção das distâncias genéticas dos indivíduos nos dois primeiros componentes principais.

Os indivíduos que estão nos extremos da distribuição são os animais de maior contribuição na variância dos componentes principais (Figura 8). Os animais mais influentes na variância dos componentes principais são aqueles mais distantes dos eixos zero dos dois primeiros componentes principais, coloridos de tonalidade azul claro para melhor visualização.

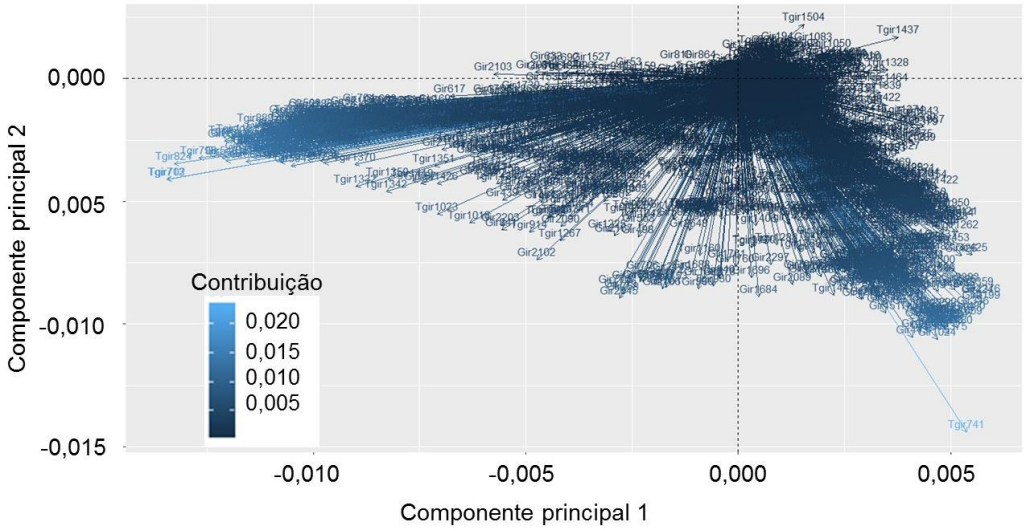


**Figura 8**. Projeção das variâncias das distâncias genéticas dos indivíduos nos dois primeiros componentes principais coloridos considerando a contribuição individual.

Considerando as 154 fazendas (Apêndice 1) do arquivo de dados é possível verificar que as distâncias genéticas entre os animais estão bem distribuídas nas diferentes propriedades, evidenciadas pela distribuição da variância por fazenda no espaço tridimensional (Figura 9). Não houve formação de um agrupamento com uma única coloração, o que evidencia o fluxo gênico nas diferentes fazendas. Ou seja, a formação dos *cluster* não está associada com o agrupamento genético formado em uma propriedade, mas pelo uso dos descendentes dos mesmos reprodutores. Sendo assim, os *cluster* são formados por indivíduos de diversas fazendas e nas fazendas há indivíduos de vários *cluster*.

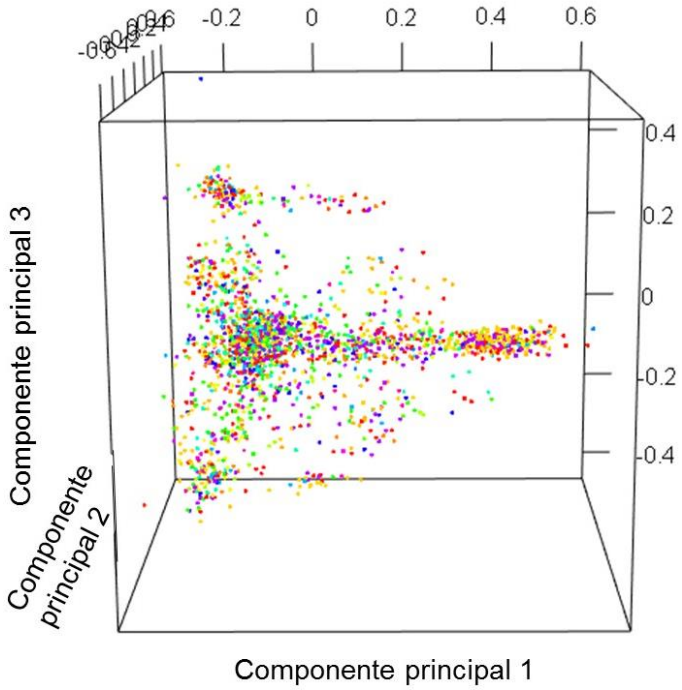


**Figura 9**. Projeção dos indivíduos nos três primeiros componentes principais em 3D considerando as diferentes fazendas, identificadas por diferentes colorações.

Os vinte touros mais influentes em relação à variância dos componentes principais, que representam 1% da população total, e as suas variâncias genéticas estão bem distribuídos nas coordenadas dos componentes principais (Figura 10a). Estes touros mais influentes na variância dos componentes principais estão presentes nos extremos das projeções (Figuras 5 e 10a), com as maiores magnitudes para o primeiro e segundo componentes principais.

Os vinte touros mais influentes nasceram entre os anos 1981 e 2006, sendo que o touro mais influente (“Benfeitor Raposo da CAL”) (Apêndice 4) é o mais distante geneticamente dos demais representados na projeção, contribuindo com mais de 2%, 8% e 8% na variância do primeiro, segundo e terceiro componentes principais, respectivamente.

O segundo touro mais influente na variância dos componentes principais (“Caju de Brasília”) é o segundo mais velho entre os vinte, contribuindo com mais de 6% na variância do primeiro componente principal e 2% na variância do segundo componente principal. O seu progenitor (“Vale Ouro de Brasília”) é o touro mais velho entre os vinte e o terceiro mais influente na variância dos

componentes principais, com contribuição semelhante a de seu filho anteriormente citado (Apêndice 4).

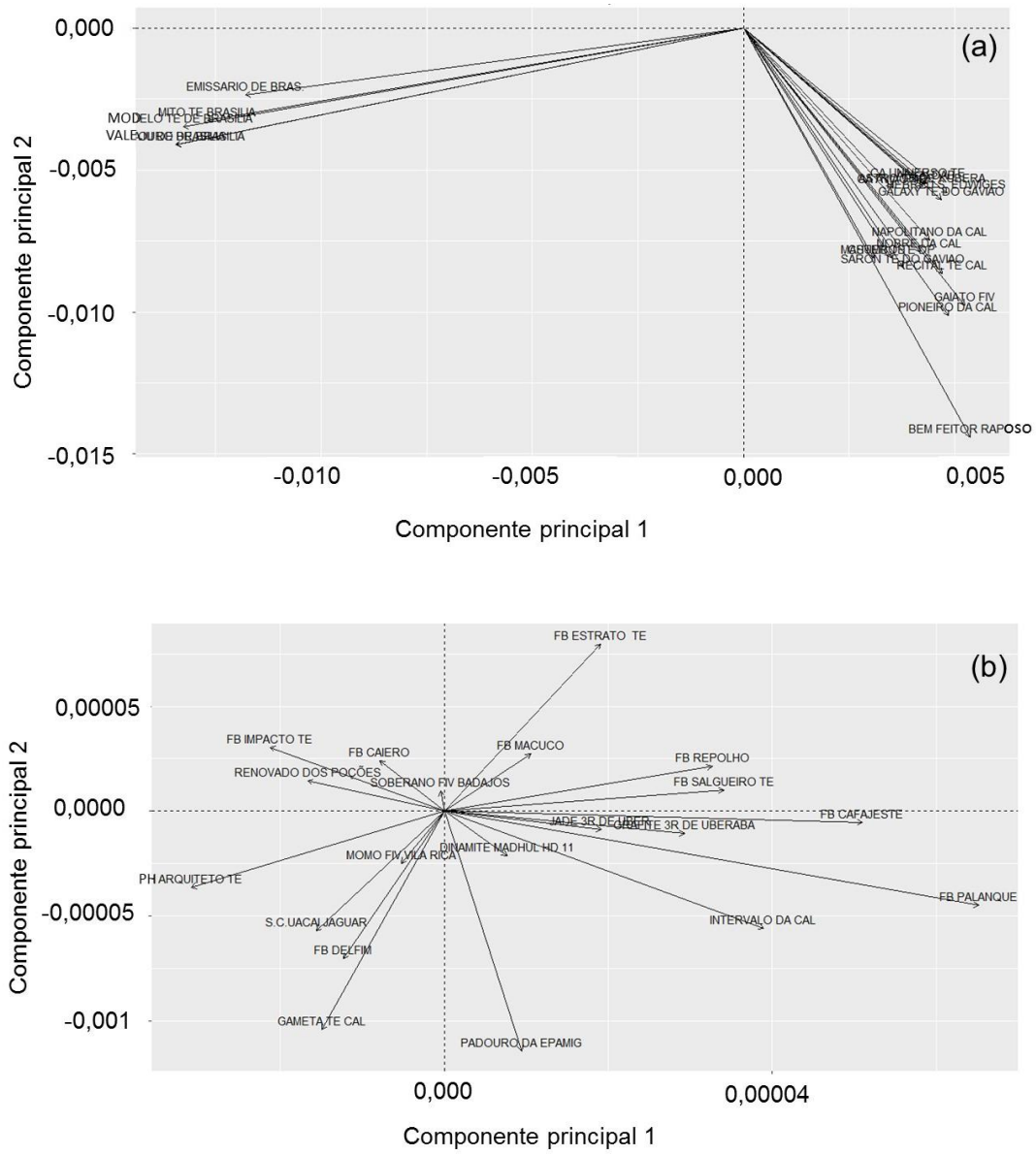


**Figura 10**. Projeção dos vinte touros mais (a) e menos (b) influentes na variância dos dois primeiros componentes principais.

Os três touros mais influentes são os que têm maior número de progênies genotipadas na população, entre 45 e 151 filhos, e pertencem ao *cluster* 3 (Apêndice 2), sendo que o mesmo *cluster* é o que contém animais de maior variabilidade genética. Estes touros nasceram entre os anos 1981 e 2005, ou seja, a variabilidade nos *cluster* não depende da idade dos animais. Esses animais são

os mais influentes na variância dos componentes principais, pois são aqueles com maior variabilidade genética em relação à população (Figura 10a).

Os touros denominados “mais influentes” são aqueles que mais influenciaram na variância dos componentes principais. Todavia, isso não significa que esses animais são os mais importantes na população em termos de potencial genético (PTA). A interpretação a ser feita é que esses touros são os mais distantes geneticamente entre os animais do conjunto de dados e essa distância refletiu na maior variabilidade dos valores dos componentes principais (Figuras 5, 8, 10a).

Os touros menos influentes na variância dos componentes principais nasceram entre os anos 1983 e 2009. A baixa contribuição na variância dos componentes principais ocorreu pela menor variabilidade genética em relação aos demais animais na população (Figura 10b). Estes touros têm menor amplitude no número de filhos genotipados presentes no arquivo de dados, entre 0 e 13, e pertencem principalmente ao *cluster* 4 (Apêndice 3). Portanto, o número de filhos genotipados determinou a contribuição dos touros na variância dos componentes principais.

Em relação aos “famosos” touros Gir, o “C.A. Sansão” não está presente nas análises por ter sido genotipado por outra plataforma, Affymetrix, sendo que para as análises foram considerados apenas os animais genotipados pela Illumina (ILLUMINA, 2009). O touro “C.A. Everest” também foi genotipado pela plataforma Affymetrix, porém alguns de seus filhos estão presentes nas análises deste estudo. Oito dos vinte touros mais influentes são filhos de “C.A. Everest” (“CA Universo TE”, “Napolitano da CAL”, “Hebreu S. Edwiges”, “Nobre da CAL”, “Galaxy TE do Gavião”, “CA Avião TE”, “Astro TE de Kubera”, “Askai DAB”), pertencentes ao *cluster* 1 ou 2. O touro “Radar dos Poções” está presente nas análises, porém ele e seus filhos não estão entre os vinte mais ou menos influentes na variância dos componentes principais.

Os touros que mais contribuíram para a variância dos componentes principais são aqueles que possuem PTA positiva para leite (Apêndice 4), com PTA média de 155,14 kg, considerando aqueles com informação disponível. Os touros que menos contribuíram para a variância dos componentes principais são principalmente os touros com as menores PTA para leite, com média de -23,93 kg

(Apêndice 5), ou seja, são os touros que possuem, em média, habilidade prevista de transmissão da característica produção de leite abaixo da média da população.

Os valores de PTA se refletem na classificação geral pela PTA de produção de leite (PTAL), sendo que os animais mais influentes são, em média e em amplitude, melhores colocados no *ranking* da PTAL do sumário de touros (95º), quando comparados com os animais menos influentes (169,6º) (Apêndices 4 e 5, respectivamente). O mesmo ocorre para a STA da característica largura do úbere posterior (LUP), com média superior para os mais influentes (1,01 e -0,32, respectivamente). Os valores de PTA e STA mais altos dos animais mais influentes é consequência da seleção, que ocorre para maiores produções de leite e escores de LUP. Portanto, animais com maiores variabilidades genéticas são aqueles com PTA favoráveis para produção de leite e largura do úbere posterior.

Em relação às demais características, ligamento do úbere anterior (LUA) (-0,40 e -0,32), profundidade de úbere (PU) (0,77 e 0,87), comprimento de tetos (CPT) (-0,83 e 2,09) e diâmetro de tetos (DT) (-0,26 e 1,49) (Apêndices 4 e 5, respectivamente), as STA foram inferiores para os animais mais influentes na variância dos componentes principais. Porém, para todas essas características, é desejável que os valores de escore sejam intermediários. Tomando como base as informações obtidas junto aos painéis de SNP não é possível inferir sobre aqueles animais que possam ter maior valor genético aditivo. Embora, os vinte touros mais influentes na variância dos componentes principais possuem maiores PTA para leite, quando comparados aos vinte menos influentes na variância dos componentes principais.

Os touros importados “Naidu Imp” e “Gaiolão DC”, nascidos em 1960 e 1977, respectivamente, pertencem ao mesmo agrupamento, *cluster* 4, aquele de menor influência na variância dos componentes principais. Estes touros são dois dos mais antigos da população e foram importados da Índia para formação da população base do rebanho leiteiro, sendo que seus descendentes formaram a população base de outros plantéis. As contribuições genéticas dos touros importados foram importantes para a formação desta população, dada idade desses touros em relação aos demais touros da população.

A baixa magnitude do primeiro componente principal dos touros importados “Naidu Imp” e “Gaiolão DC”, (0,0094, 0,0130), e secundário (0,0011, 0,0005),

respectivamente, pode ter ocorrido devido à diluição do material genético desses animais. Ao longo de várias gerações sob seleção, diferentes alelos foram favorecidos, formando agrupamentos a partir dessa população base, que contém os touros importados. No *cluster* 4 estão presentes tanto animais jovens como animais mais velhos e que fizeram parte da população base, com o material genético disseminado em muitos indivíduos na população ou são aqueles que possuem número limitado de filhos genotipados nesta população.

- **Parâmetros populacionais**

**Parentesco genômico**

Por meio da matriz de parentesco genômico, representada pelo mapa de calor (*heatmap*) (Figura 11) foi possível observar que os animais possuem parentesco genômico semelhante, tanto dentro de *cluster* quanto entre *cluster.* A intensidade da coloração azul na matriz de parentesco genômico indicou menor coeficiente de parentesco genômico entre os animais e os dendogramas nos limites da matriz agruparam os indivíduos de parentesco genômicos mais próximos.

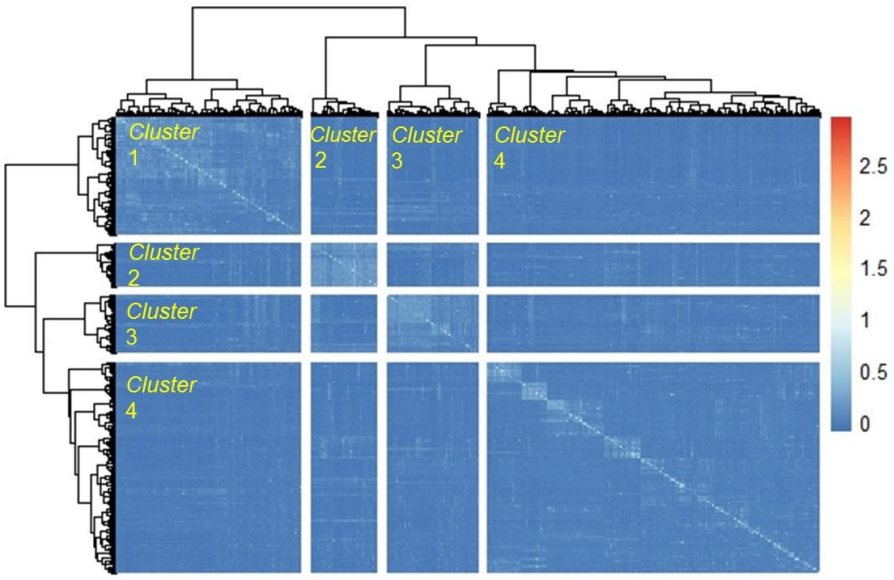


**Figura 11**. Matriz de parentesco genômico entre os animais distribuídos em seus respectivos cluster.

Nas estimativas de parentesco genômico (Figura 12) observou-se maior dispersão no *cluster* 1, apesar das semelhanças entre as dispersões dos *cluster* 1, 2 e 3 e menor no *cluster* 4. A distribuição do parentesco dos indivíduos do *cluster* 4 está mais à direita do zero quando comparada com os demais, indicativo de que os animais contidos no *cluster* 4 são mais aparentados entre si do que nos demais agrupamentos.

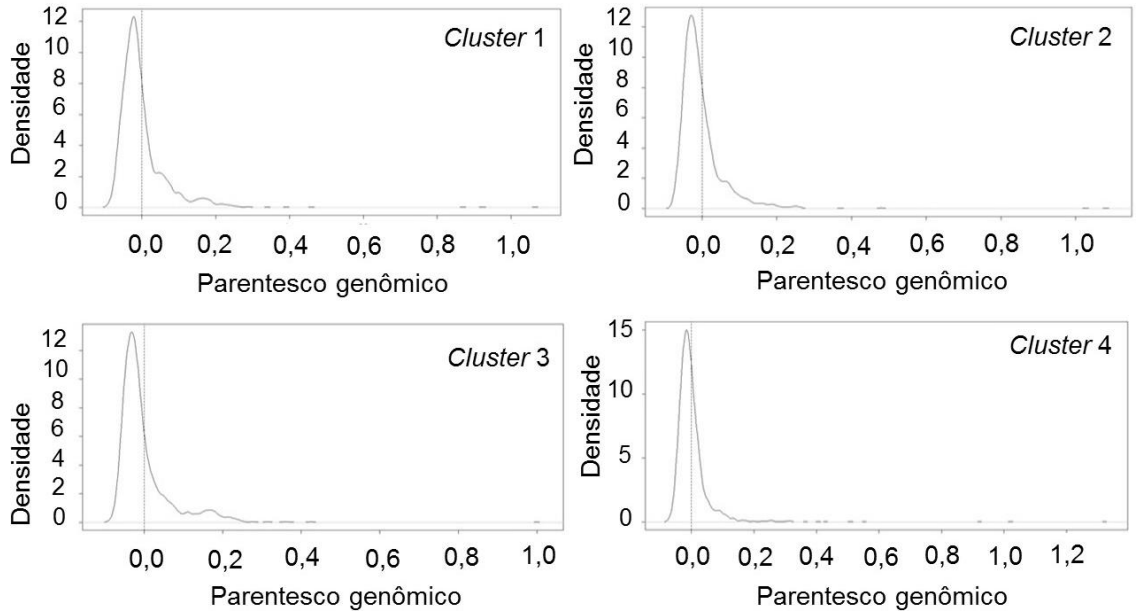


**Figura 12**. Distribuição das estimativas de parentesco genômico dos animais nos diferentes *cluster*.

Os valores associados à matriz de parentesco genômico são desvios da média, portanto com média zero (Tabela 3). Então, valores negativos são interpretados como animais menos aparentados do que a média da população (GONDRO; VAN DER WERF; HAYES, 2013). Porém, também existem estimativas de parentesco extremos, o que causou alta variação, indicada pelos altos valores dos desvios-padrão e CV (Tabela 4).

Por meio das médias de desvio de parentesco genômico dentro e entre os *cluster* (Tabela 3) foi possível observar que o parentesco é equivalente entre os animais. De acordo com o teste *t* de *Student*, as médias de parentesco entre os indivíduos do *cluster* 1 e 4 apresentaram diferença significativa ($P < 0,05$). Portanto, em média, esses são os agrupamentos com os animais com maiores

diferenças de parentesco. Esses resultados corroboram aquele verificado na matriz de parentesco genômico (Figura 11), no qual é possível observar maior distância entre os dois *cluster* pela estrutura do dendograma.

**Tabela 3**. Média do desvio de parentesco genômico dentro (diagonal) e entre *cluster* (acima da diagonal) e desvio-padrão (entre parênteses e abaixo da diagonal)

| | *Cluster* 1 | *Cluster* 2 | *Cluster* 3 | *Cluster* 4 |
|---|---|---|---|---|
| *Cluster* 1 | -0,0005 (0,0685) | 0,0013 | 0,0019 | 0,0009 |
| *Cluster* 2 | 0,0779 | -0,0022 (0,0629) | 0,0000 | -0,0002 |
| *Cluster* 3 | 0,0755 | 0,0749 | 0,0000 (0,0688) | -0,0004 |
| *Cluster* 4 | 0,0748 | 0,0739 | 0,0715 | 0,0017 (0,0655) |

**Heterozigosidade**

As médias de heterozigosidade esperada (He) e heterozigosidade observada (Ho) apresetaram altos coeficientes de variação (Tabela 4). Então, estes parâmetros poderiam ser melhor representados pela mediana, como medida de tendência central. Porém, as médias de He e Ho foram muito próximas às medianas.

As médias de He e Ho apresentaram-se iguais e o valor médio de He foi intermediário, sendo que o valor máximo foi 0,50 (FALCONER; MACKAY, 1996). A He é indicativo de variabilidade genética e quanto maior He, mais equilibradas são as frequências dos alelos. A proximidade de Ho e He indica que não houve fixação e nem perda de alelos com efeitos indesejáveis e que combinações heterozigóticas poderiam produzir resultados favoráveis a cada geração.

Em análise de variabilidade genética em bovinos leiteiros Girolando e Holandês com base em informações genômicas (VILAÇA et al., 2016), foi observado Ho menor que He, indicando desequilíbrio de Hardy-Weinberg na população e maior diversidade genética entre os indivíduos cruzados. Em populações de seis raças bovinas africanas (MAKINA et al., 2014) foram observadas He de 0,24 a 0,31, sendo que o valor mais alto indicou maior diversidade genética das raças Angus e Holandesa, resultado semelhante ao encontrado no presente estudo, e He de 0,24 indicou baixo nível de variabilidade

genética para a raça Afrikaner, sendo o valor mais baixo em relação às demais raças estudadas atribuído à forte seleção na raça, ao uso de touros elite e ao pequeno tamanho efetivo da população.

**Tabela 4**. Média, mediana, desvio-padrão (DP), coeficiente de variação (CV), valores mínimo e máximo de heterozigosidade esperada (He), heterozigosidade observada (Ho), coeficiente de endogamia e desequilíbrio de ligação (LD) para a população estudada

| | Média | Mediana | DP | CV (%) | Mínimo | Máximo |
|---|---|---|---|---|---|---|
| He | 0,25 | 0,24 | 0,15 | 60,00 | 0,01 | 0,50 |
| Ho | 0,25 | 0,23 | 0,15 | 60,00 | 0,01 | 0,76 |
| Endogamia | 0,017 | 0,00 | 0,029 | 163,30 | 0,00 | 0,10 |
| LD | 0,17 | 0,04 | 0,26 | 152,94 | 0,00 | 1,00 |

Com relação ao valor médio de He e Ho de todos os SNP (Tabela 4), observou-se que a população poderia estar em equilíbrio de Hardy-Weinberg (FALCONER; MACKAY, 1996). Entretanto, se a comparação for realizada com base nos valores de He e Ho de cada SNP separadamente, alguns poderão não estar em equilíbrio, como ocorre com os valores máximos observados em He e Ho (0,50 e 0,76, respectivamente).

A causa do desequilíbrio de Hardy-Weinberg para alguns SNP pode ser devido aos processos de seleção para as características de interesse econômico e que afeta determinadas regiões do genoma, o que causa a variação dos valores de He e Ho para alguns SNP (FALCONER; MACKAY, 1996).

**Endogamia**

A média da endogamia foi baixa, 0,017, com alto coeficiente de variação (Tabela 4). A estimativa do parentesco genômico reflete no coeficiente de endogamia, sendo que a magnitude da estimativa do parentesco apresentou-se baixa, devido ao delineamento para genotipagem dos animais desta população, sendo escolhidos indivíduos com menor parentesco entre si. Além disso, o valor da média do coeficiente de endogamia é influenciado pelas frequências dos alelos raros, sendo que, quando a MAF é moderada a alta (Tabela 1), os coeficientes de

endogamia tendem a ficar próximos de zero (ZHANG et al., 2015). A baixa magnitude do coeficiente de endogamia também pode ter ocorrido devido ao pequeno número de gerações genotipadas desta população.

As frequências de animais endogâmicos, com coeficiente de endogamia (F) iguais ou acima de 0,05, 0,01, 0,02, 0,03, 0,04, 0,05 variaram entre 27,28% e 42,01% da população (Tabela 5), diminuindo o número de animais conforme o aumento da endogamia. Na segunda forma de análise da endogamia, considerando 0,50 a 50,00% da população, foi possível observar maiores coeficientes de endogamia (e), 0,1002, para pequena amostra da população. Sendo assim, da mesma forma que o cálculo anteriormente citado, quanto menor foi o número de indivíduos, maiores foram os coeficientes de endogamia, por abranger o extremo direito da distribuição dos indivíduos endogâmicos da população.

**Tabela 5**. Probabilidade de indivíduos endogâmicos (x) e coeficiente de endogamia (e) de acordo com a Distribuição de Qui-Quadrado (Tabela IV)

| F | x | N% | e |
|---|---|---|---|
| 0,05 | 0,4201 | 0,50 | 0,1002 |
| 0,01 | 0,3790 | 1,00 | 0,0959 |
| 0,02 | 0,3351 | 5,00 | 0,0702 |
| 0,03 | 0,3081 | 10,00 | 0,0525 |
| 0,04 | 0,2884 | 25,00 | 0,0235 |
| 0,05 | 0,2728 | 50,00 | 0,0000 |

Média=0,017, desvio-padrão=0,029. À esquerda da tabela, fixou-se F para cálculo das probabilidades de indivíduos (x) com endogamias iguais ou superiores a esses valores (F). À direita da tabela fixou-se a porcentagem de indivíduos com maiores valores de endogamia (N%) para cálculo do coeficiente de endogamia (e).

Com base nos registros de pedigree de outros autores, em bovinos da raça Gir leiteiro foram obtidas médias de coeficientes de endogamia de 2,82% (REIS FILHO et al., 2010), e 2,14% (SANTANA JUNIOR et al., 2014). Por meio de análise de pedigree para estimativas dos efeitos da endogamia sobre a produção de leite no dia de controle e persistência da lactação envolvendo a raça Gir leiteira (PEREIRA et al., 2016), foi observado que o aumento da endogamia teve impacto

negativo sobre a produção de leite e níveis mais elevados de endogamia afetariam a persistência da produção de leite na lactação, devendo ser levado em consideração nas avaliações genéticas.

**Desequilíbrio de ligação**

Todos os pares de SNP com distância igual ou inferior a 100 Kb produziram 8.358 combinações de pares de SNP bialélicos para estimar o LD nos 29 cromossomos autossomos. A média geral de LD entre pares de marcadores foi 0,17 para $r^2$ (Tabela 4) e o decaimento (*decay*) da média geral do LD pela metade (0,08) ocorreu entre as distâncias 150 e 200 Kb (Tabela 6).

O LD diminuiu à medida que a distância física entre os marcadores aumentou, seguindo o padrão de decaimento de LD (Figura 13). Níveis moderados de $r^2$, entre 0,20 e 0,38, foram observados em distâncias entre marcadores inferiores a 40 Kb (Tabela 6). Quando a distância aumentou de 40 até 100 Kb, o $r^2$ médio para as classes diminuiu, entre 0,20 e 0,11. A alta variabilidade nas estimativas de $r^2$ gerou altos CV.

O decaimento de LD ao longo das maiores distâncias ocorreu porque quanto maior a distância entre os marcadores, maiores as chances de eventos de recombinação, *crossing over*, o que leva a redução ou quebra do LD entre os marcadores, com redução da média de LD tendendo a zero (HARTL; CLARK, 2010). O conhecimento sobre o padrão de LD nas populações é essencial para determinar a densidade dos SNP necessária para se obter acurácia no GWAS e seleção genômica, por haver maior conhecimento sobre a MAF, histórico de seleção, entrutura e tamanho efetivo da população (BIEGELMEYER et al., 2016).

Para a taxa de decaimento de LD, se reduzida à metade, será necessário o dobro de gerações para o mesmo decaimento de LD (MACKAY; POWEL, 2007). Este decaimento pode ser atribuído à intensa seleção ou ainda devido ao elevado número de gerações sem a inclusão de novo material genético, sendo que a última importação de indivíduos nesta população ocorreu na década de 1970.

O maior número de meioses resulta em maior quebra de LD, ou seja, quanto maior o número de gerações, maior será o decaimento de LD. A maior extensão do LD ocorre porque alelos em *loci* vizinhos tendem a ser herdados

juntos e tendem a ser associados em uma população segregante. A seleção reduz a variação genética na próxima geração e, como ocorre em *loci*, então *loci* vizinhos que estão em LD terão uma extensão maior em desequilíbrio de ligação, um efeito denominado *hitchhiking* (BULMER, 1971).

**Tabela 6**. Desequilíbrio de ligação ($r^2$) entre pares de SNP (N) localizados em diferentes distâncias (em Kb)

| Distância | N | Média | Mediana | DP | CV (%) |
|---|---|---|---|---|---|
| 0-10 | 74 | 0,38 | 0,26 | 0,40 | 105,26 |
| 10-20 | 189 | 0,21 | 0,05 | 0,29 | 138,09 |
| 20-30 | 2.987 | 0,21 | 0,07 | 0,30 | 142,85 |
| 30-40 | 2.469 | 0,20 | 0,06 | 0,28 | 140,00 |
| 40-50 | 2.141 | 0,18 | 0,05 | 0,28 | 155,55 |
| 50-60 | 2.206 | 0,16 | 0,04 | 0,25 | 156,25 |
| 60-70 | 2.190 | 0,15 | 0,04 | 0,24 | 160,00 |
| 70-80 | 2.182 | 0,14 | 0,03 | 0,21 | 150,00 |
| 80-100 | 4.279 | 0,13 | 0,03 | 0,21 | 161,53 |
| 100-150 | 10.591 | 0,11 | 0,02 | 0,20 | 181,81 |
| 150-200 | 10.491 | 0,08 | 0,02 | 0,16 | 200,00 |

A predição do valor genético genômico e seleção genômica poderia ocorrer principalmente considerando menores distâncias entre pares de SNP separados por distâncias de até 40 Kb, em que $r^2$ foi 0,20 (Tabela 6). Estimativas inferiores foram obtidas para bovinos leiteiros Gir (0,09, ±0,14) para distâncias entre pares de SNP de até 40 Kb (NEVES et al., 2015). Em outro estudo envolvendo três raças, o $r^2$ variou de 0,13 a 0,16 em distâncias de até 40 Kb (BUZANSKAS et al., 2017). Para bovinos Hanwoo Korean Cattle foi relatado média de $r^2$ de 0,23 em distâncias mais curtas, menores que 25 Kb, reduzindo a 0,1 em distâncias entre 40 e 60 Kb (LEE et al., 2011). As estimativas superiores obtidas neste estudo sugerem que a maior extensão de LD pode ter ocorrido devido à maior pressão de seleção nesta população, ou ainda devido ao menor tamanho efetivo populacional quando comparado às demais populações.

Os resultados deste estudo sugerem que maior LD em distâncias menores poderia ser indício de um tamanho efetivo populacional maior do que outras

populações estudadas. De acordo com a literatura (MCKAY et al., 2007; NEVES et al., 2015), os valores obtidos para $r^2$ para distâncias mais próximas entre os marcadores seriam suficientes para obter desequilíbrio de ligação entre SNP e QTL.

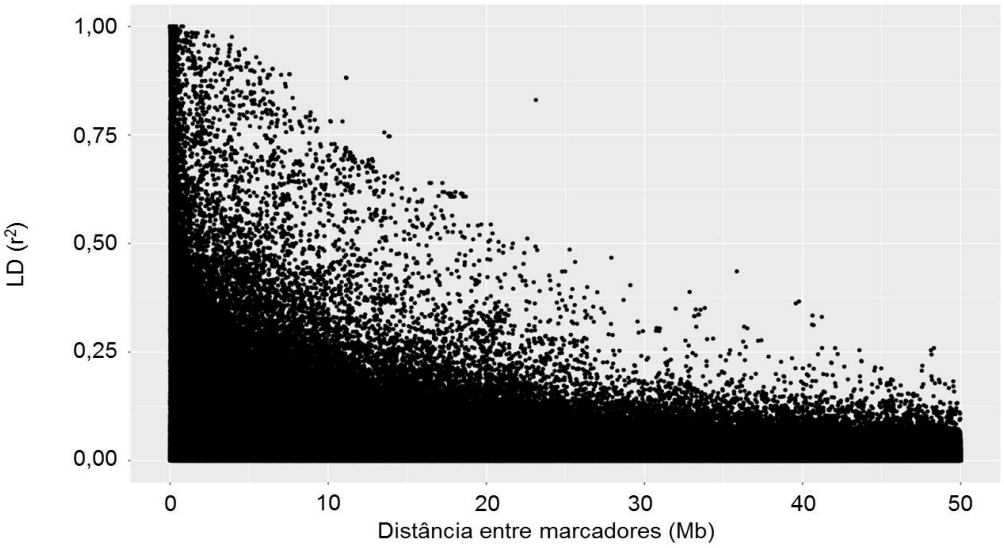


**Figura 13**. Desequilíbrio de ligação (LD) ao longo das distâncias entre SNP em Megabase (Mb).

## Tamanho efetivo populacional (Ne)

O tamanho efetivo populacional calculado considerando o desequilíbrio de ligação foi 92,84, pelo método de Hayes et al. (2003). Ou seja, aproximadamente 93 indivíduos nesta população irão contribuir com a manutenção da diversidade genética para a próxima geração. Por meio da análise de pedigree foi obtido tamanho efetivo populacional para bovinos Gir próximos ao deste estudo, de 94 animais (SANTANA JUNIOR et al., 2014), e superior, de 146 animais (REIS FILHO et al., 2010). Em análises do pedigree da raça Gir, outros autores relataram que houve redução de Ne de 70 para 45 entre os anos de 1979 e 1998 (FARIA et al., 2009).

Em análises de Ne considerando informações genômicas foi relatado Ne de 98,1 para bovinos Hanwoo Korean Cattle (LEE et al., 2011) e 99 indivíduos

para bovinos Charolês na primeira geração (BUZANSKAS et al., 2017), valores semelhantes ao deste estudo. Para bovinos Gir foi obtido valor inferior de Ne, 56 indivíduos (NEVES et al., 2015), e nas duas e quatro últimas gerações de bovinos Braford e Hereford, os valores de Ne foram de 220 e 153 animais, respectivamente (BIEGELMEYER et al., 2016). O tamanho efetivo populacional obtido neste estudo, que representou 4,67% da população estudada, sugere que para manter a variabilidade genética da raça será necessário considerar o tamanho efetivo populacional nas decisões de escolha dos touros para acasalamento.

A redução do tamanho efetivo populacional em gerações recentes dos rebanhos pode ser consequência da redução do número de ancestrais ao longo das gerações devido à intensa seleção (LEE et al., 2011; NEVES et al., 2015), atribuída à utilização de poucos touros “famosos”, cuja terminologia foi utilizada por Falconer e Mackay et al. (1996) para discriminar touros que são intensamente utilizados nos acasalamentos. Como a seleção é baseada nas PTA (diferença esperada na progênie), há maiores chances de indivíduos aparentados serem selecionados, o que pode causar aumento da endogamia e, consequentemente, redução do Ne, por aumentar as chances de acasalamentos entre animais aparentados.

Portanto, em um futuro próximo será necessário incluir novos materiais genéticos e reduzir a pressão de seleção nos rebanhos, para que assim o Ne possa aumentar nas próximas gerações. Este parâmetro populacional poderá auxiliar nas decisões dos acasalamentos para que se evite a redução da variabilidade genética nos rebanhos.

## 5 CONCLUSÕES

A população está estruturada pela presença de quatro agrupamentos genéticos, podendo ser definidos como linhagens desta população, as quais devem ser consideradas nas análises de GWAS e seleção genômica. Portanto, além do desequilíbrio de ligação, as análises genômicas para esta população devem levar em conta os efeitos de subgrupos genéticos.

Com base no parentesco genômico, heterozigosidade e coeficiente de endogamia, concluiu-se que há fluxo gênico dos indivíduos desta população nas

fazendas, proporcionado pelos acasalamentos entre animais pouco aparentados, mantendo a variabilidade genética nesta população, apesar do tamanho efetivo populacional pequeno. Esses resultados devem ser considerados nas decisões de escolha dos indivíduos para acasalamento nos programas de melhoramento genético da raça a fim de aumentar ou manter a variabilidade genética dos rebanhos.

## 6 REFERÊNCIAS

ABCGIL. Associação Brasileira de Criadores de Gir Leiteiro. 2017. Acesso em: 15 mar. 2017. Disponível em: <http://girleiteiro.org.br/>.

ACGZ. **Associação dos Criadores de Gaúchos de Zebu**. 2016. Disponível em: <http://www.acgz.com.br/secao_racas.php?pagina=5>. Acesso em: 3 nov. 2016.

ALEXANDER, D. H.; NOVEMBRE, J.; LANGE, K. Fast model-based estimation of ancestry in unrelated individuals. **Genome Research, Woodbury**, v. 19, n. 9, p. 1655-1664, 2009.

AMARAL, R.; VILLARES, J. B.; FARIA, R. S.; GESTAL, R. L. Melhoramento tecnológico: estímulo alimentar, no período de lactação de Gir leiteiro. In: CONGRESSO BRASILEIRO DE PESQUISA DE ZEBU, 1, 1988, Uberaba. **Anais...** Uberaba: Empresa de Pesquisa Agropecuária de Minas Gerais, 1988.

ANNEY, R. J. L; KENNY, E.; O'DUSHLAINE, C. T.; LASKY-SU, J.; FRANKE, B.; MORRIS, D. W.; NEALE, B. M.; ASHERSON, P.; FARAONE, S. V.; GILL, M. Non-Random Error in Genotype Calling Procedures: implications for family-based and case-control genome-wide association studies. **American Journal of Medic Genetics B**: Neuropsychiatric Genetics, Hoboken, v. 147, n. 8, p. 1379–1386, 2008.

BIEGELMEYER, P.; GULIAS-GOMES, C. C.; CAETANO, A. R.; STEIBEL, J. P.; CARDOSO, F. F. Linkage disequilibrium, persistence of phase and effective population size estimates in Hereford and Braford cattle. **BMC Genetics**, London, v. 17, n. 1, p. 32, 2016.

BULMER, M. G. The effect of selection on genetic variability. **The American Naturalist**, v. 105, n. 943, p. 201-211, 1971.

BUZANSKAS, M. E.; VENTURA, R. V.; CHUD, T. C. S.; BERNARDES, P. A.; SANTOS, D. J. A.; REGITANO, L. C. A.; ALENCAR, M. M.; MUDADU, M. A.; ZANELLA, R.; SILVA, M. V. B. G.; LI, C; SCHENKEL, F. S; MUNARI, D. P. Study on the introgression of beef breeds in Canchim cattle using single nucleotide polymorphism markers. **PloS One**, San Francisco, v. 12, n. 2, p. e0171660, 2017.

CARDON, L. R.; PALMER, L. J. Population stratification and spurious allelic association. **Lancet**, London, v. 361, n. 9357, p. 598-604, 2003.

CAVALLI-SFORZA, L. L.; FELDMAN, M. W. The application of molecular genetic approaches to the study of human evolution. Nature Genetics, v. 33, p. 266–275, 2003.

CIRULLI, E. T.; GOLDSTEIN, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. **Nature Reviews Genetics**, London, v. 11, n. 6, p. 415-425, 2010.

COLEMAN, J. R.; EUESDEN, J.; PATEL, H.; FOLARIN, A. A.; NEWHOUSE, S.; BREEN, G. Quality control, imputation and analysis of genome-wide genotyping data from the Illumina HumanCoreExome microarray. **Briefings in Functional Genomics**, Oxford, v. 15, n. 4, p. 298-304, 2016.

COX, D. G.; KRAFT, P. Quantification of the power of Hardy-Weinberg equilibrium testing to detect genotyping error. **Human Heredity**, New York, v. 61, n. 1, p. 10-14, 2006.

CRUZ, C. D.; CARNEIRO, P. C. S.; REGAZZI, A. J. **Modelos biométricos aplicados ao melhoramento genético**. Viçosa: Editora UFV, 2014.

ELSIK, C. G.; TELLAM, R. L.; WORLEY, K. C. The genome sequence of taurine cattle: a window to ruminant biology and evolution. **Science**, Washington, v. 324, n. 5926, p. 522-528, 2009.

FALCONER, D. S.; MACKAY, T. F. C. **Introdution to Quantitative Genetics**. London: Pearson United Kingdom, 1996. p.1-22, 57-72, 247-262.

FARIA, F. J. C.; VERCESI FILHO, A. E.; MADALENA, F. E.; JOSAHKIAN, L. A. Pedigree analysis in the Brazilian Zebu breeds. **Journal of Animal Breeding and Genetics**, Oxford, v. 126, n. 2, p. 148-153, 2009.

FAO. Food and Agriculture Organization of the United Nations. **Statistical Pocketbook. World food and agriculture**. 2015. Disponível em: <http://www.fao.org/3/a-i4691e.pdf>. Acesso em: 15 mar. 2017.

FORNI, S.; AGUILAR, I.; MISZTAL, I. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. **Genetics Selection Evolution**, Les Ulis, v. 43, n. 1, p. 1, 2011.

GILES, R.E.; BLANC, H.; CANN, H. M.; WALLACE, D. C. Maternal inheritance of human mitochondrial DNA. **Proceeding of the National Academy of Sciences of the United States of America**, v. 77, n. 11, p. 6715-6719, 1980.

GODDARD, M. Mapping genes for quantitative traits using linkage disequilibrium. **Genetics Selection Evolution**, Les Ulis, v. 23, n. Suppl 1, p. 131s-134s, 1991.

GONDRO, C. **Primer to analysis of genomic data using R**. Cham: Springer. 2015. cap. 3, p. 91.

GONDRO, C.; VAN DER WERF, J.; HAYES, B. **Genome-Wide Association Studies and Genomic Prediction**. Armidale: Humana Press, 2013. cap. 9, p. 218, cap. 19, p. 430-434.

GRIFFITHS, F; WESSLER, S. R.; CARROLL, S. B.; DOEBLEY, J. **Introdução em Genética**. Rio de Janeiro: Guanabara koogan, 2015. p. 170.

GUTIÉRREZ, J. P.; ALTARRIBA, J.; DÍAZ, C.; QUINTANILLA, R.; CAÑÓN, J.; PIEDRAFITA, J. Pedigree analysis of eight Spanish beef cattle breeds. **Genetics Selection Evolution**, Les Ulis, v. 35, n. 1, p. 43-64, 2003.

HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. **Análise multivariada de dados**. Porto Alegre: Bookman, 2005.

HARTL, D. L.; CLARK, A. G. **Princípios de Genética de Populações**. 4. ed. Porto Alegre: Artmed, 2010. cap. 9, p. 491-500.

HAYES, B. J.; BOWMAN, P. J.; CHAMBERLAIN, A. J.; GODDARD, M. E. Invited review: Genomic selection in dairy cattle: progress and challenges. **Journal of Dairy Science**, Champaign, v. 92, n. 2, p. 433-443, 2009.

HAYES, B. J.; VISSCHER, P. M.; MCPARTLAN, H. C.; GODDARD, M. E. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. **Genome Research**, Woodbury, v. 13, n. 4, p. 635-643, 2003.

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

HOTELLING, H. Analysis of a complex of statistical variables into principal components. **Journal of Educational Psychology**, Washington, v. 24, n. 6, p. 498-520, 1933.

ILLUMINA. **GenomeStudio: An integrated platform for data visualization and analysis**. 2009. Disponível em: <http://support.illumina.com/>. Acesso em: 13 set. 2016.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRAN, R. **An Introduction to Statistical Learning with Applications in R**. 2. ed. New York: Springer-Verlag New York, 2013. cap. 6, p. 230-237.

JOMBART, T.; DEVILLARD, S.; BALLOUX, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. **BMC Genetics**, London, v. 11, n. 1, p. 94, 2010.

KOLDE, R. **Package "pheatmap"**. R-project. Disponível em: <http://cran.r-project.org/web/packages/pheatmap/pheatmap.pdf>. Acesso em: 7 nov. 2016.

KRISTENSEN, T. N.; SORENSEN, A. C. Inbreeding – lessons from animal breeding, evolutionary biology and conservation genetics. **Animal Science**, Cambridge, v. 80, n. 2, p. 121-133, 2005.

LAM, A. C.; SCHOUTEN, M.; AULCHENKO, Y. S.; HALEY, C. S.; KONING, D. J. Rapid and robust association mapping of expression quantitative trait loci. **BMC Proceedings**, London, v. 1, n. 1, p. S144, 2007.

LAURIE, C. C.; DOHENY, K. F.; MIREL, D. B.; PUGH, E. W.; BIERUT, L. J.; BHANGALE, T.; BOEHM, F.; CAPORASO, N. E.; CORNELIS, M. C.; EDENBERG, H. J.; GABRIEL, S. B.; HARRIS, E. L.; HU, F. B.; JACOBS, K. B.; KRAFT, P.; LANDI, M. T.; LUMLEY, T.; MANOLIO, T. A.; MCHUGH, C.; PAINTER, I.; PASCHALL, J.; RICE, J. P.; RICE, K. M.; ZHENG, X.; WEIR, B. S.; FOR THE GENEVA INVESTIGATORS. Quality control and quality assurance in genotypic data for genome-wide association studies. **Genetic Epidemiology**, Malden, v. 34, n. 6, p. 591-602, 2010.

LEÃO, G. F. M.; PIVATO, D. R. D.; CARNIEL, H.; RODRIGUES, M. G. K.; BRAGA, R. A.; SILVA, M. R. H.; TEIXEIRA, P. P. M. Melhoramento genético em zebuínos leiteiros – uma revião. **Agropecuária Científica no Semiárido**, Campina Grande, v. 9, n. 4, p. 9-14, 2013.

LEE, S. H.; CHO, Y. M.; LIM, D.; KIM, H. C.; CHOI, B. H.; PARK, H. S.; KIM, O. H.; KIM, S.; KIM, T. H.; YOON, D.; HONG, S. K. Linkage disequilibrium and effective population size in Hanwoo Korean Cattle. **Asian-Australasian Journal of Animal Sciences**, Seoul, v. 24, n. 12, p. 1660-1665, 2011.

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

LIN, D. Y.; HU, Y.; HUANG, B. E. Simple and efficient analysis of disease association with missing genotype data. **The American Journal of Human Genetics**, Houston, v. 82, n. 2, p. 444-452, 2008.

MACKAY, I.; POWEL, W. Methods for linkage disequilibrium mapping in crops. **Trends in Plant Science**, v. 12, n. 2, p. 57-63, 2007.

MADALENA, F. E.; TEODORO, R. L.; LEMOS, A. M.; MONTEIRO, J. B. N.; BARBOSA, R. T. Evaluation of Strategies for Crossbreeding of Dairy Cattle in Brazil. **Journal of Dairy Science**, Champaign, v. 73, n. 7, p. 1887-1901, 1990.

MAKINA, S. O.; MUCHADEYI, F. C.; VAN MARLE-KÖSTER, E.; MACNEIL, M. D.; MAIWASHE, A. Genetic diversity and population structure among six cattle breeds in South Africa using a whole genome SNP panel. In: **Advances in Farm Animal Genomic Resources**, Lausanne:Frontiers in Genetics, 2014. p. 92-98.

MARCHINI, J.; CARDON, L. R., PHILLIPS, M. S.; DONNELLY, P. The effects of human population structure on large genetic association studies. **Nature Genetics**, New York , v. 36, n. 5, p. 512-517, 2004.

MASON, I. L. **Mason's World Dictionary of Livestock Breeds, Types and Varieties**. Wallingford: CAB Internacional, 2002. p. 33, 43.

MCKAY, S D; SCHNABEL, R. D.; MURDOCH, B. M.; MATUKUMALLI, L. K; AERTS, J.; COPPIETERS, W.; CREWS, D.; DIAS NETO, E.; GILL, C. A.; GAO, C.; MANNEN, H.; STOTHARD, P.; WANG, Z.; VAN TASSELL, C. P.; WILLIAMS, J. L.; TAYLOR, J.F.; MOORE, S. S. Whole genome linkage disequilibrium maps in cattle. **BMC Genetics**, London, v. 8, n. 1, 2007.

NCBI. National Center for Biotechnology Information Gene. **Bos indicus (zebu cattle)**. Disponível em: <https://www.ncbi.nlm.nih.gov/genome/?term=bos+indicus%5Borgn%5D>. Acesso em: 13 mar. 2017.

NEIVA, R. Genômica promove seleção mais veloz. **Revista XXI Ciência para a vida**, 2017, p. 21.

NEVES, H. H. R.; DESIDÉRIO, J. A.; PIMENTEL, E. C. G.; SCALEZ, D. C. B.; QUEIROZ, S. A. Preliminary study to determine extent of linkage disequilibrium and estimates of autozygosity in Brazilian Gyr dairy cattle. **Archivos de Zootecnia**, Cordoba, v. 64, n. 246, p. 99-107, 2015.

NOTTER, D. R. The importance of genetic diversity in livestock populations of the future. **Journal of Animal Science**, Champaign, v. 77, n. 1, p. 61-69, 1999.

NOVEMBRE, J.; STEPHENS, M. Interpreting principal component analyses of spatial population genetic variation. **Nature Genetics**, New York, v. 40, p. 646-649, 2008.

O'BRIEN, A. M. P.; MÉSZÁROS, G.; UTSUNOMIYA, Y. T.; SONSTEGARD, T. S.; GARCIA, F. J.; TASSELL, C. P. V.; CAVALHEIRO, R.; SILVA, M. V. B.; SOLKNER, J. Linkage disequilibrium levels in Bos indicus and Bos Taurus cattle using medium and high density SNP chip data and different minor allele frequency distributions. **Livestock Science**, Amsterdam, v. 166, n. 5, p. 646-649, 2014.

PANETTO, J. C. C.; VERNEQUE, R. S.; SILVA, M. V. G. B.; MACHADO, M. A.; MARTINS, M. F.; BRUNELI, F. A. T.; PEIXOTO, M. G. C. D.; SANTOS, G. G.; ARBEX, W. A.; REIS, D. R. L.; GERALDO, C. C.; MACHADO, C. H. C.; VENTURA, H. T.; PEREIRA, M. A.; HORTOLANI, B.; VERCESI FILHO, A. E.; MACIEL, R. S.; FERNANDES, A. R. **Programa Nacional de Melhoramento do Gir Leiteiro - Sumário brasileiro de touros - Resultado do teste de progênie 7ª prova de pré-seleção de touros**. Juiz de Fora: Embrapa: CNPGL, 2016. 86 p. (Embrapa-CNPGL. Documentos, 187).

PATTERSON, N.; PRICE, A. L.; REICH, D. Population Structure and Eigenanalysis. **PLoS Genetics**, Cambridge, v. 2, n. 12, p. e190, 2006.

PEARSON, K. On Lines and Planes of Closest Fit to Systems of Points in Space, **Philiosophical Magazin**, London, v. 2, p. 559-572, 1901.

PEREIRA, R. J.; SANTANA JÚNIOR, M. L.; AYRES, D. R.; BIGNARDI, A. B.; VERCESI FILHO, A. E. Depressão endogâmica na produção de leite no dia do controle de bovinos Gir leiteiro. **Pesquisa Agropecuária Brasileira**, Brasília, v. 51, n. 6, p. 751-758, 2016.

PRICE, A. L.; PATTERSON, N. J.; PLENGE, R. M.; WEINBLATT, M. E.; SHADICK, N. A.; REICH, D. Principal components analysis corrects for stratification in genome-wide association studies. **Nature Genetics**, New York, v. 38, n. 8, p. 904-909, 2006.

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

PRITCHARD, J. K.; PRZEWORSKI, M. Linkage disequilibrium in humans: models and data. **The American Journal of Human Genetics**, Houston, v. 69, n. 1, p. 1-14, 2001.

PURCELL, S; NEALE, B; TODD-BROWN, K; THOMAS, L; FERREIRA, M. A.; BENDER, D; MALLER, J.; SKLAR, P; BAKKER, P. I.; DALY, M. J.; SHAM, P. C. PLINK: a tool set for whole-genome association and population-based linkage analyses. **The American Journal of Human Genetics**, Houston, v. 81, n. 3, p. 559-575, 2007.

QUEIROZ, S. A.; ALBUQUERQUE, L. G.; LANZONI, N. A. Efeito da endogamia sobre características de crescimento de bovinos da raça Gir no Brasil. **Revista Brasileira de Zootecnia**, Viçosa, v. 29, n. 4, p. 1014-1019, 2000.

R CORE TEAM. **A language and environment for statistical computing, R Foundation for Statistical Computing**. Vienna, Austria. 2016.

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

RELETHFORD, J. H. **Human population genetics**. Hoboken: John Wiley & Sons, 2012. p. 112-119.

ROUSSEEUW, P. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, v. 20, n. 1, 1987.

SANTANA JUNIOR, M. L.; PEREIRA, R. J.; BIGNARDI, A. B.; EL FARO, L.; TONHATI, H.; ALBUQUERQUE, L. G. History, structure, and genetic diversity of Brazilian Gir cattle. **Livestock Science**, Amsterdam, v. 163, p. 26-33, 2014.

SANTOS, R.; CORRÊA, A. B. Como são feitos os testes de progênie (provas de touros). In: SIMPÓSIO NACIONAL DE MELHORAMENTO ANIMAL, 3, 2000, Belo Horizonte. **Anais eletrônicos...** Belo Horizonte: Sociedade Brasileira de Melhoramento Animal, 2000. Disponível em: <http://sbmaonline.org.br/anais/iii/palestras/pdfs/iiip26.pdf>. Acesso em: 9 jan. 2017.

SARGOLZAEI, M.; SCHENKEL, F. S.; JANSEN, G. B.; SCHAEFFER, L. R. Extent of linkage disequilibrium in Holstein cattle in North America. **Journal of Dairy Science**, Champaign, v. 91, n. 5, p. 2106-2117, 2008.

SEIDEL, E. J.; MOREIRA JÚNIOR, F. J.; ANSUJ, A. P.; NOAL, M. R. C. Comparação entre o método Ward e o método K-médias no agrupamento de produtores de leite. Ciência e Natura, v. 30, n. 1, p. 7-15, 2008.

SILVA, P. B. R.; MACHADO, C. H. C. Atribuições da subdivisão do julgamento da raça Gir em: Gir leiteiro e Gir dupla-aptidão. **Cadernos de Pós-Graduação da FAZU**, 2011.

SLATKIN, M. Linkage disequilibrium-understanding the evolutionary past and mapping the medical future. **Nature Reviews Genetics**, London, v. 9, n. 6, p. 477-485, 2008.

SNEATH, P. H. A.; SOKAL, R. R. **Numerical taxonomy**. San Francisco: W. H. Freeman & Co, 1973.

SOKAL, R. R.; ROHLF, F. J. The Comparison of Dendrograms by Objective Methods. **Taxon**, v. 11, n. 2, p. 33-40, 1962.

TEO, Y. Y. Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure. **Current Opinion in Lipidology**, v. 19, n. 2, p. 133-143, 2008.

TEO, Y. Y.; FRY, A. E.; CLARK, T. G.; TAI, E. S; SEIELSTAD, M. On the usage of HWE for identifying genotyping errors. **Annals of Human Genetics**, v. 71, n. 5, p. 701-703, 2007.

TORO, M. A.; VILLANUEVA, B.; FERNÁNDEZ, J. Genomics applied to management strategies in conservation programmes. **Livestock Science**, Amsterdam, v. 166, 48-53, 2014.

TURNER, S.; ARMSTRONG, L. L.; BRADFORD, Y.; CARLSON, C. S.; CRAWFORD, D. C.; CRENSHAW, A. T.; ANDRADE, M.; DOHENY, K. F.; HAINES, J. L.; HAYES, G.; JARVIK, G.; JIANG, L.; KULLO, I. J.; LI, R.; LING, H.; MANOLIO, T. A.; MATSUMOTO, M.; MCCARTY, C. A.; MCDAVID, A. N., MIREL, D. B.; PASCHALL J. E., PUGH, E. W.; RASMUSSEN, L. V.; WILKE, R. A.; ZUVICH, R. L.; RITCHIE M. D. Quality control procedures for genome wide association studies. **Current Protocols in Human Genetics**, Hoboken, p. 1-24, 2011.

VENTURA, R.; LARMER, S.; SCHENKEL, F. S.; MILLER, S. P; SULLIVAN, P. Genomic clustering helps to improve prediction in a multibreed population. **Journal of Animal Science**, Champaign, v. 94, n. 5, p. 1844-1856, 2016.

VERCESI FILHO, A. E.; DIAS, A. L.; CARDOSO, V. L.; EL FARO, L.; MERINGUE, G. K. F.; MEIRELLES, F. V. Caracterização de um rebanho Gir Leiteiro quanto à origem do DNA mitocondrial (mtDNA). **Boletim de Indústria Animal**, v. 67, n. 1, p. 91-95, 2010.

VILAÇA, L. F.; DINIZ, W. J. S.; MELO, T. F.; OLIVEIRA, J. C. V.; GUIDO, S. I.; BRITO, C. E. V. L.; COSTA, N. A.; SANTORO, K. R. Polimorfismos do gene BoLA-DRB3 em rebanhos bovinos leiteiros 5/8 Girolando e Holandês no estado de Pernambuco. **Archivos de Zootecnia**, Cordoba, v. 65, n. 249, p. 7-11, 2016.

WANG, D.; SUN, Y.; STANG, P.; BERLIN, J. A.; WILCOX, M. A.; QINGQIN, L. Comparison of methods for correcting population stratification in a genome-wide association study of rheumatoid arthritis: principal-component analysis versus multidimensional scaling. **BMC Proceedings**, London, v. 3, p. S109, 2009.

WRIGHT, S. Evolution in Mendelian Populations, **Genetics**, Bethesda, v. 16, n. 2, p. 97–159, 1931.

YAMAZAKI, T. The effects of overdominance on linkage in a multilocus system. **Genetics**, v. 86, n. 1, p. 227-236, 1977.

ZHANG, Q; CALLUS, M. P. L.; GULDBRANDTSEN, B.; LUND, M. S.; SAHANA, G. Estimation of inbreeding using pedigree, 50k SNP chip genotypes and full sequence data in three cattle breeds. **BMC Genetics**, London, v. 16, p. 1-11, 2015.

ZHAO, H.; NETTLETON, D.; DEKKERS, J. C. Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between single nucleotide polymorphisms. **Genetical Research**, Cambridge, v. 89, n. 1, p. 1-6, 2007.

ZIEGLER, A.; KONIG, I. R.; THOMPSON, J. R. Biostatistical aspects of genome-wide association studies. **Biometrical Journal**, Malden, v. 50, n. 1, p. 8-28, 2008.

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

## 7 APÊNDICES

Apêndice 1. Recodificação (Recod) e identificação das fazendas*.

| Recod | Fazenda | Recod | Fazenda | Recod | Fazenda | Recod | Fazenda |
|---|---|---|---|---|---|---|---|
| 1 | IMP** | 43 | ENA | 87 | K | 134 | SQP |
| 2 | A | 44 | ESA | 88 | KA | 134 | SQPA |
| 3 | AA | 45 | EUS | 89 | KAL | 135 | T |
| 4 | AB | 46 | EVPF | 90 | KB | 136 | TCA |
| 5 | ABPAA | 47 | FAN | 91 | KC | 137 | TCAT |
| 5 | ABP | 48 | FASA | 92 | KCA | 138 | TOE |
| 6 | ACFG | 49 | FBGO | 93 | KCAK | 139 | TOLA |
| 7 | ACOP | 49 | FBGA | 94 | KOK | 140 | TSF |
| 8 | ADAO | 50 | FCGA | 95 | L | 141 | TSFE |
| 9 | AEV | 51 | FCGO | 95 | LA | 142 | TZN |
| 10 | ANF | 52 | FGVL | 96 | LAC | 143 | U |
| 11 | ANV | 52 | FGVP | 97 | LANF | 144 | UDI |
| 11 | ANVA | 53 | FGVM | 98 | LBRY | 145 | UNIG |
| 12 | APAG | 54 | FJAG | 99 | LCRM | 146 | V |
| 13 | APPG | 55 | FJLS | 100 | LEAO | 147 | VRPG |
| 14 | AVB | 56 | FNE | 101 | LEIT | 148 | WALV |
| 15 | AVLA | 57 | FRFL | 102 | LFRB | 149 | WCBL |
| 16 | B | 58 | FSDS | 103 | LFTN | 150 | X |
| 17 | BASP | 59 | GAOM | 104 | LGR | 151 | YOYG |
| 18 | BCO | 60 | GAV | 105 | LGX | 152 | Z |
| 19 | BEY | 61 | GIL | 106 | LLB | 153 | ZAB |
| 20 | BJAS | 62 | GIVR | 107 | LMT | 154 | MCCV |
| 21 | BQP | 62 | GIVL | 108 | LUF | 130 | RMM |
| 21 | BQPF | 63 | GVCS | 109 | LUGO | 131 | RRP |
| 22 | BRTG | 64 | HCFG | 110 | MABG | 132 | RSSO |
| 23 | C | 65 | HCP | 111 | MAMJ | 133 | SDNA |
| 24 | CALL | 66 | HDD | 112 | MDB | 134 | SQP |
| 24 | CAL | 67 | HGS | 113 | MELM | 134 | SQPA |
| 25 | CEAP | 68 | HMQ | 114 | MILE | 135 | T |
| 26 | CGG | 69 | ISPG | 115 | MJJR | 136 | TCA |
| 27 | CKGL | 70 | IVAR | 116 | MPVV | 137 | TCAT |
| 28 | CLMD | 71 | JAS | 117 | MUT | 138 | TOE |
| 29 | CSLM | 72 | JCRF | 118 | NLT | 139 | TOLA |
| 30 | D | 73 | JCVL | 119 | OGM | 140 | TSF |
| 31 | DAB | 74 | JDRB | 120 | PARG | 141 | TSFE |
| 32 | DGLM | 74 | JDRC | 121 | PECG | 142 | TZN |
| 33 | DIAS | 75 | JFR | 122 | PHPO | 143 | U |
| 34 | DOBI | 76 | JFSA | 123 | PRAC | 144 | UDI |
| 35 | DPJ | 77 | JFSH | 124 | RBTT | 145 | UNIG |
| 36 | DQP | 78 | JGVA | 125 | RCBR | 146 | V |
| 36 | DQPL | 79 | JJJJ | 126 | RCPO | 147 | VRPG |
| 37 | E | 80 | JMCH | 127 | RIG | 148 | WALV |
| 38 | EFC | 81 | JMMA | 128 | RMB | 149 | WCBL |
| 38 | EFCA | 82 | JRDG | 129 | RMI | 150 | X |
| 39 | EGB | 83 | JRF | 130 | RMM | 151 | YOYG |
| 40 | ELPF | 84 | JRR | 131 | RRP | 152 | Z |
| 41 | ELZ | 85 | JRRG | 132 | RSSO | 153 | ZAB |
| 42 | EMGU | 86 | JWLJ | 133 | SDNA | 154 | MCCV |

*Fazendas definidas pelos prefixos dos RGDs (identificação) dos animais.
**Animais importados

Apêndice 2. Descrição dos vinte touros mais influentes na variância dos componentes principais.

| Nº | RGDs | Nomes | Filhos genotipados | *Cluster* |
|---|---|---|---|---|
| 1 | A7481 | Benfeitor Raposo | 151 | 3 |
| 2 | B58 | Caju de Brasília | 58 | 3 |
| 3 | A6796 | Vale Ouro de Brasília | 45 | 3 |
| 4 | B5213 | Modelo TE de Brasília | 16 | 3 |
| 5 | CAL5277 | Recital TE CAL | 0 | 3 |
| 6 | B5212 | Mito TE Brasília | 8 | 1 |
| 7 | KCA633 | CA Universo TE | 1 | 1 |
| 8 | GAV244 | Saron TE do Gavião | 0 | 1 |
| 9 | CAL4406 | Napolitano da CAL | 0 | 2 |
| 10 | RIG126 | Hebreu S. Edwiges | 0 | 1 |
| 11 | CAL4397 | Nobre da CAL | 6 | 2 |
| 12 | GAV171 | Galaxy TE do Gavião | 2 | 2 |
| 13 | CAL4762 | Pioneiro da CAL | 0 | 2 |
| 14 | KCA888 | CA Avião TE | 0 | 2 |
| 15 | ACFG50 | Astro TE de Kubera | 0 | 2 |
| 16 | JFR1734 | Master TE | 0 | 2 |
| 17 | RMM46 | Gaiato FIV | 0 | 2 |
| 18 | RRP5764 | Emissário de Brasília | 0 | 2 |
| 19 | DPJ373 | Chumbo TE DP | 0 | 2 |
| 20 | DAB6 | Askai DAB | 3 | 2 |

Apêndice 3. Descrição dos vinte touros menos influentes na variância dos componentes principais.

| Nº | RGD | Nome | Filhos genotipados | *Cluster* |
|---|---|---|---|---|
| 1 | LA34 | FB Caiero | 1 | 4 |
| 2 | B5594 | Dinamite Madhul HD 11 | 0 | 2 |
| 3 | APPG1294 | Renovado dos Poções | 0 | 4 |
| 4 | B4706 | Grafite 3R de Uberaba | 2 | 4 |
| 5 | LLB161 | Soberano FIV Badajoz | 0 | 4 |
| 6 | B6304 | FB Macuco | 7 | 4 |
| 7 | B4623 | Jade 3R de UBER. | 0 | 2 |
| 8 | B3563 | FB Impacto TE | 6 | 4 |
| 9 | FBGO343 | FB Salgueiro TE | 0 | 4 |
| 10 | LA405 | FB Repolho | 1 | 4 |
| 11 | GIVR307 | Momo FIV Vila Rica | 0 | 4 |
| 12 | LA429 | FB Delfim | 1 | 4 |
| 13 | LA35 | FB Cafajeste | 0 | 4 |
| 14 | PHPO357 | PH Arquiteto TE | 0 | 4 |
| 15 | K1557 | Intervalo da CAL | 2 | 4 |
| 16 | FBGO621 | FB Estrato TE | 0 | 4 |
| 17 | B4010 | SC Uaçaí Jaguar | 13 | 4 |
| 18 | B5032 | Gameta TE CAL | 4 | 4 |
| 19 | A9726 | Padouro da Epamig | 3 | 4 |
| 20 | B6317 | FB Palanque | 1 | 4 |

Apêndice 4. Habilidade prevista de transmissão (PTA, kg) para produção de leite (PL), classificação geral pela PTA de produção de leite (PTAL, kg) e STA (PTA padronizada) para as características de conformação, enumeradas, para os touros mais influentes na variância dos componentes principais.

| Animais | Características | | | | | | |
|---|---|---|---|---|---|---|---|
| | PL | PTAL | LUA[1] | LUP[2] | PU[3] | CPT[4] | DT[5] |
| A7481 | 24,2 | 187º | 3,2769 | -1,2843 | -2,2998 | -0,1223 | 1,2533 |
| B58 | 194,4 | 81º | 2,3474 | 2,6511 | 3,0537 | -3,9564 | -2,1801 |
| A6796 | 115,4 | 131º | 2,4892 | 0,1442 | 2,2347 | -3,7352 | -2,3389 |
| B5213 | 316,9 | 27º | 1,0319 | 1,2637 | 0,7051 | -1,4394 | 0,9003 |
| CAL5277* | - | - | - | - | - | - | - |
| B5212 | 17,1 | 198º | 0,6932 | 2,1566 | 2,2889 | -2,1188 | 0,2824 |
| KCA633 | 73,3 | 154º | -2,7413 | 2,4382 | 2,3106 | 0,1718 | -0,9179 |
| GAV244 | -274,8 | - | 0,7798 | -0,7898 | -1,1987 | 0,2603 | 0,6267 |
| CAL4406 | -32,7 | - | -1,6463 | 0,9409 | 1,177 | 0,2030 | -0,5207 |
| RIG126 | 335,8 | - | -2,7177 | 0,9478 | 1,0197 | -0,3670 | -1,0238 |
| CAL4397 | 281,0 | 42º | -3,6866 | 3,4684 | 2,6307 | 2,0459 | 2,2507 |
| GAV171 | 234,1 | 57º | -3,5133 | 1,2981 | 0,3417 | -0,1301 | -1,9771 |
| CAL4762 | 239,8 | 61º | -0,4017 | 2,0124 | -0,7214 | 0,0078 | 2,0212 |
| KCA888 | 209,2 | 70º | -2,4971 | 2,1429 | 2,1805 | -1,0099 | -1,6593 |
| ACFG50 | 152,2 | 107º | -1,1107 | 1,4011 | 0,2766 | -1,0099 | -0,8473 |
| JFR1734 | 188,2 | 83º | 0,9374 | -1,6277 | -1,1607 | -0,2239 | 0,3089 |
| RMM46 | 281,4 | - | 0,8429 | 0 | -0,9004 | -1,1921 | -0,0088 |
| RRP5764 | 204,0 | 76º | 0,7720 | 1,0714 | 1,8984 | -0,4946 | 0,2030 |
| DPJ373** | - | - | - | - | - | - | - |
| DAB6 | 233,0 | 56º | -2,0875 | -0,0412 | 0,1573 | -1,8871 | -1,0768 |
| Média | 155,1 | 95º | -0,4017 | 1,0107 | 0,7774 | -0,8332 | -0,2613 |

[1] ligamento do úbere anterior, [2] largura do úbere posterior, [3] profundidade de úbere, [4] comprimento de tetos, [5] diâmetro de tetos.

*Touros que não fazem parte do teste de progênie da ABCGIL, portanto a PTA para PL não pode ser divulgada.

**Resultados não apresentados no sumário.

Apêndice 5. Habilidade prevista de transmissão (PTA, kg) para produção de leite (PL), classificação geral pela PTA de produção de leite (PTAL, kg) e STA (PTA padronizada) para as características de conformação, enumeradas, para os touros menos influentes na variância dos componentes principais.

| Animais | Características | | | | | | |
|---|---|---|---|---|---|---|---|
| | PL | PTAL | LUA[1] | LUP[2] | PU[3] | CPT[4] | DT[5] |
| LA34 | -11,4 | - | 0,2048 | -0,5563 | 1,2801 | 5,4609 | 4,6779 |
| B5594 | -110,1 | - | -0,4569 | -0,1580 | 0,4719 | -0,3670 | -0,0265 |
| APPG1294 | 49,2 | 170º | -0,4096 | 2,0810 | 1,6001 | 2,1344 | 3,1421 |
| B4706 | -266,5 | - | 0,9453 | -1,250 | -0,1681 | 6,0960 | 4,6514 |
| LLB161*** | - | - | - | - | - | - | - |
| B6304 | 138,5 | 115º | -0,1969 | 1,9643 | 2,4788 | 4,3338 | 3,7070 |
| B4623* | - | - | - | - | - | - | - |
| B3563 | -209,0 | | -0,4963 | -1,5934 | -0,8516 | 2,0875 | 0,8473 |
| FBGO343 | -41,8 | - | -0,0394 | -0,3777 | 0,1953 | 1,8116 | 1,7034 |
| LA405* | - | - | - | - | - | - | - |
| GIVR307*** | - | - | - | - | - | - | - |
| LA429 | 51,4 | 171º | -0,5987 | 1,2912 | 1,8062 | 2,8502 | 0,6973 |
| LA35 | 12,7 | 203º | 0,3308 | -0,5495 | 0,6455 | 2,2697 | 0,8120 |
| PHPO357** | - | - | - | - | - | - | - |
| K1557 | -17,6 | | -1,221 | 0,7212 | 0,5966 | -1,3821 | -0,6708 |
| FBGO621 | 0 | - | -1,6936 | 0,0069 | 0,7811 | 1,2806 | 0,7149 |
| B4010 | 46,1 | 189º | -0,8507 | 3,5302 | 3,5961 | 1,3795 | 1,5711 |
| B5032 | 142,2 | 113º | -0,1891 | 1,6003 | -1,0902 | 0,7080 | -0,5296 |
| A9726 | -159,0 | - | 0,5908 | 0,4739 | 0,1519 | 1,1609 | 0,1677 |
| B6317 | 16,4 | - | -0,772 | 0,783 | 1,6435 | 1,5617 | 0,9709 |
| Média | -23,93 | 169,6º | -0,3235 | 0,5311 | 0,8758 | 2,0923 | 1,4957 |

[1] ligamento do úbere anterior, [2] largura do úbere posterior, [3] profundidade de úbere, [4] comprimento de tetos, [5] diâmetro de tetos.

*Touros que não fazem parte do teste de progênie da ABCGIL, portanto a PTA para PL não pode ser divulgada.

**Resultados não apresentados no sumário.

***Touros que fazem parte do teste de progênie da ABCGIL, porém ainda sem resultado divulgado.