UNIVERSIDADE ESTADUAL PAULISTA - UNESP

CÂMPUS DE JABOTICABAL

APLICAÇÕES DA INFORMAÇÃO GENÔMICA NO ESTUDO DE POPULAÇÕES SELECIONADAS

Amanda Marchi Maiorano

Zootecnista

2019

UNIVERSIDADE ESTADUAL PAULISTA - UNESP

CÂMPUS DE JABOTICABAL

APLICAÇÕES DA INFORMAÇÃO GENÔMICA NO ESTUDO DE POPULAÇÕES SELECIONADAS

Amanda Marchi Maiorano

Orientador: Prof. Dr. Josineudson Augusto II de Vasconcelos Silva

Coorientadoras: Profa. Dra. Daniela Andressa Lino Lourenço

Dra. Joslaine Noely dos Santos Gonçalves Cyrillo

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.

2019

M227a

Maiorano, Amanda Marchi

Aplicações da informação genômica no estudo de populações selecionadas / Amanda Marchi Maiorano. -- Jaboticabal, 2019

77 p. : tabs.

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

Orientador: Josineudson Augusto II de Vasconcelos Silva

Coorientadora: Daniela Andressa Lino Lourenço

1. Acurácia de predição. 2. Bovinos. 3. Estrutura da população. 4. Genes. 5. Suínos. I. Título.

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

Essa ficha não pode ser modificada.

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Jaboticabal

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: APLICAÇÕES DA INFORMAÇÃO GENÔMICA NO ESTUDO DE POPULAÇÕES SELECIONADAS

AUTORA: AMANDA MARCHI MAIORANO
ORIENTADOR: JOSINEUDSON AUGUSTO II DE VASCONCELLOS SILVA
COORIENTADORA: JOSLAINE NOELY DOS SANTOS GONÇALVES CYRILLO
COORIENTADORA: DANIELA ANDRESSA LINO LOURENÇO

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

Prof. Dr. JOSINEUDSON AUGUSTO II DE VASCONCELLOS SILVA
Departamento de Melhoramento e Nutrição Animal-FMVZ/UNESP / Botucatu/SP

Prof. Dr. HENRIQUE NUNES DE OLIVEIRA
Departamento de Zootecnia / FCAV / Unesp - Jaboticabal

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

Pesquisadora Dra. NEDENIA BONVINO STAFUZZA
Instituto de Zootecnia-APTA / Sertãozinho/SP

Prof. Dr. RUSBEL RAUL ASPILCUETA BORQUIS
Departamento de Zootecnia / Universidade Federal da Grande Dourados - Dourados/MS

Jaboticabal, 14 de janeiro de 2019

DADOS CURRICULARES DO AUTOR

Amanda Marchi Maiorano – nascida no município de São Carlos, São Paulo, Brasil, em 06 de setembro de 1990, filha de Luis Carlos Maiorano e Wilneide do Carmo Marchi Maiorano. Zootecnista, graduada pela Faculdade de Medicina Veterinária e Zootecnia, Universidade Estadual Paulista (UNESP), Câmpus de Botucatu, em julho de 2013, com a monografia intitulada "Comparação de análise genética da característica musculosidade realizada com modelo linear e não linear", sob orientação do Prof. Dr. Josineudson Augusto II de Vasconcelos Silva. Em agosto de 2013, iniciou seu mestrado no Programa de Pós-graduação em Genética e Melhoramento Animal, da Faculdade de Ciências Agrárias e Veterinárias (FCAV), Universidade Estadual Paulista (Unesp), Câmpus de Jaboticabal. Realizou parte experimental de seu projeto de mestrado, intitulado "Análise genética da espessura do couro, do crescimento pós desmama e das resistências ao carrapato e à babesiose em bovinos de corte", no Instituto de Zootecnia (Centro APTA Bovinos de Corte) de Sertãozinho, SP, e na Embrapa Pecuária Sudeste (CPPSE) de São Carlos, SP. Recebeu apoio financeiro do CNPq durante o mestrado e, em fevereiro de 2015, obteve o título de mestre em Genética e Melhoramento Animal. No mesmo ano iniciou o doutorado no Programa de Pós-graduação em Genética e Melhoramento Animal da FCAV/Unesp, sob orientação do Prof. Dr. Josineudson Augusto II de Vasconcelos Silva, sendo bolsista CNPq durante todo o período. Realizou parte experimental de seu projeto de doutorado no Centro APTA Bovinos de Corte, IZ, sob supervisão da Dra. Joslaine Noely dos Santos Gonçalves Cyrillo. Durante o período de Abril/2017 à Março/2018 foi bolsista pelo programa de doutorado sanduíche da CAPES (processo número 88881.131671/2016-01), desenvolvendo parte do projeto na Universidade de Georgia em Athens, Georgia, Estados Unidos, sob orientação da Prof. Dra. Daniela Andressa Lino Lourenço.

“If you can dream it, you can do it”

Walt Disney

Dedico

Aos meus pais, Wilneide e Luis, pelo apoio e amor que me proporcionaram em todas as etapas de minha vida.

AGRADECIMENTOS

Ao professor **Josineudson Augusto II de Vasconcelos Silva,** por ter me orientado durante o doutorado, pelas oportunidades que me ofereceu, confiança, parceira, ensinamentos e conselhos.

Aos pesquisadores**, Joslaine Noely dos Santos Gonçalves Cyrillo** e **Anibal Eugênio Vercesi Filho**, e demais funcionários do Instituto de Zootecnia, pelo esforço e contribuição na coleta dos dados.

À professora **Daniela Lourenco,** por ter me orientado durante meu doutorado sanduíche em Athens, pelos conhecimentos adquiridos e pela atenção que me foi dada durante todo o período.

Aos membros das bancas de qualificação e defesa do meu doutorado, pelas sugestões e contribuições.

Aos pesquisadores da Universidade da Georgia, **Shogo Tsuruta, Yutaka Masuda** e **Breno Fragomeni,** pelos ensinamentos e conversas, e ao **Dr. Misztal** e **Dr. Rekaya** pelas discussões em aulas e seminários que contribuíram para o meu aprendizado.

À **Nedenia Bonvino Stafuzza** e ao **Rafael Medeiros de Oliveira Silva** por terem me ajudado nos primeiros dias em Athens.

Aos alunos (**Yvette Steyn, Ashley Ling, Andre Garcia, Pattrapoljk Sumreddee (Pete), Heather Braford, Ivan Pocrnic, Hinayah Rojas, Alula Aemayehu, Aline Guarini, Gustavo Garcia Santiago** e **Axelle Mineur**) do "Department of Animal and Dairy Science" que tive a oportunidade de conhecer e conviver durante minha estadia nos Estados Unidos.

Aos alunos, **Ricardo Antônio da Silva Faria** e **Alejandra Maria Toro Ospina**, pela amizade e, principalmente, pelo apoio que me foi dado durante esta fase, pelas conversas e parcerias.

Aos colegas e amigos de departamento da UNESP, câmpus de Botucatu, **Luiz Eduardo Correia, Camilla Nogueira, Beatriz Pressi, Ana Beatriz Menezes**, **Felipe Krieck**, **Matheus Henrique Vargas**, **Jessica Moraes Malheiros** e demais alunos, por todos os momentos compartilhados.

As bibliotecárias da UNESP, câmpus de Jaboticabal, pela atenção e correções das referências.

Ao querido secretário **Carlão**, *in memoriam*, pelo tempo em que esteve disposto a nos ajudar e não media esforços para isso.

À minha família e aos meus amigos, do Brasil e dos Estados Unidos, por estarem sempre presentes e pelo significado que cada um tem em minha vida.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001 (processo nº 88881.131671/2016-01), da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP; processo nº 2015/ 06686-7) e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq; processo nº 142373/2015-0).

SUMÁRIO

Página

APLICAÇÕES DA INFORMAÇÃO GENÔMICA NO ESTUDO DE POPULAÇÕES SELECIONADAS

RESUMO – A disponibilidade de painéis de polimorfismos de nucleotídeos únicos (SNPs) a baixo custo tem viabilizado estudos genômicos em espécies de interesse econômico. Os objetivos do presente estudo foram conduzir análises de estruturação da população e assinaturas de seleção em duas populações de bovinos Gir selecionadas para diferentes critérios de seleção (corte ou leite); e investigar se a estimativa do efeito materno para característica medida na progênie apresenta maior acurácia com uso de "pool" de sêmen e informação genômica em diferentes cenários simulados de suínos. Análise de componentes principais (PCA) e Análise Discriminante de Componentes Principais (DAPC) foram utilizadas para verificar a estrutura da população nas populações Gir analisadas, e três métodos diferentes foram utilizados na detecção de assinaturas de seleção: índice de fixação (Fst), Escore de Integração dos Haplótipos (iHS) e Homozigose do Haplótipo Estendido Entre Populações (XP-EHH). Diferenciação genética entre as duas populações de bovinos foi observada nos resultados de PCA e DAPC. Um total de 282 genes foram identificados sob seleção, com base nos testes Fst, iHS e XP-EHH, dos quais 35 genes estão associados com QTL previamente descritos em bovinos. Os padrões de variação genética nos bovinos Gir foram consistentes com a presença de pressões seletivas em algum momento na história das populações de corte e leite. Estes resultados fornecem informações complementares sobre regiões genômicas de interesse para estudos funcionais, estudos de associação ampla do genoma e implementação de programas de melhoramento objetivando o melhoramento genético e conservação de espécies domésticas. No estudo de simulação com suínos, o programa QMSim foi utilizado para simular diferentes cenários de irmãos completos e meios-irmãos maternos. Os componentes de variâncias, herdabilidades direta e materna foram estimados com uso do programa AIREMLF90. Os preditores foram calculados por meio das metodologias Melhor Predição Linear Não-viesada (BLUP) e "single step genomic" BLUP (ssGBLUP). Correlações entre os valores genéticos verdadeiros e estimados da população de validação foram usadas para avaliar as acurácias de predição dos efeitos direto (acc_{direta}) e materno (acc_{mat}). As estimativas de herdabilidade direta e materna variaram de 0,04 a 0,06 e 0,16 a 0,20, respectivamente. Em geral, as estimativas de acc_{direta} e acc_{mat} foram menores para o cenário de pai único comparado com os cenários de "pool" de sêmen. Maiores acurácias foram observadas nos modelos ssGBLUP, independentemente do cenário, evidenciando a importância da inclusão da informação genômica nas avaliações genéticas.

Palavras-chave: acurácia de predição, bovinos, estrutura da população, genes, suínos

APPLICATIONS OF GENOMIC INFORMATION IN THE STUDY OF SELECTED POPULATIONS

ABSTRACT – The availability of commercial SNP panels at a low cost has made it possible the application of genomic studies in domestic species. The aim of this study was to conduct population structure and signatures of selection analyzes in two populations of Gir cattle selected for different criteria (beef or dairy); and to investigate whether maternal direct effect for litter birth weight can be improved when pooled semen and genomic information are used in different simulated scenarios in pigs. Principal Component Analysis (PCA) and Discriminant Analysis of Principal Components (DAPC) were used to verify population structure in the studied Gir cattle populations, and three different methods were used to detect signatures of selection: fixation index (Fst), Integrated Haplotype Score (iHS), and Cross-Population Extend Haplotype Homozygosity (XP-EHH). Genetic differentiation between the two cattle populations was observed in the PCA and DAPC results. A total of 282 genes were identified under selection based on the Fst, iHS and XP-EHH tests, of which 35 genes are associated with QTL previously described in cattle. The patterns of genotypic variation in Gir cattle were consistent with the presence of selective pressures at some point in the history of the beef and dairy populations. These findings can provide complementary information on genomic regions of interest for functional genomic studies, genome-wide associations, and the implementation of breeding schemes aiming genetic improvement and conservation of livestock populations. For the simulation study in pigs, QMSim program was used to simulate different scenarios of full sibling and maternal half sibling families. Variance components, direct and maternal heritabilities were estimated using AIREMLF90. Predictors were computed via Best Linear Unbiased Prediction (BLUP) and single-step genomic BLUP (ssGBLUP). Correlations of estimated and true breeding values from the validation population was used to evaluate accuracies of prediction for direct (acc_{direct}) and maternal (acc_{mat}) effects. The estimates of h^2_{direct} and h^2_{mat} ranged from 0.04 to 0.06 and 0.16 to 0.20, respectively. The estimated direct heritability and maternal heritability ranged from 0.04 to 0.06 and from 0.16 to 0.20, respectively. In general, acc_{direct} and acc_{mat} estimates were lower in single-boar scenario compared to pooled semen scenarios. Higher accuracies were found for ssGBLUP when compared to BLUP independently of the scenario, which is a clear evidence it is important to account for genomic information in genetic evaluations.

Keywords: accuracy of prediction, cattle, population structure, genes, pig

CAPÍTULO I – CONSIDERAÇÕES GERAIS

1. Introdução

Com a disponibilidade de painéis de polimorfismos de nucleotídeos únicos (SNPs) a baixo custo no mercado, estudos de estrutura da população, assinaturas de seleção e seleção genômica podem contribuir para o melhoramento genético animal. Os painéis de SNPs fornecem informações sobre a distância genética entre os animais (VanRaden et al., 2011), a qual é importante para verificar a estrutura genômica de populações bem como o parentesco entre animais.

Os estudos de estrutura da população possibilitam identificar similaridades entre os indivíduos. É esperado que animais expostos a diferentes critérios de seleção permaneçam em diferentes grupos, denominados “clusters” (Chen et al., 2016). As regiões do genoma que foram selecionadas nos animais também podem ser investigadas. Os estudos de assinaturas de seleção fornecem maior entendimento acerca dos genes que estão associados as regiões selecionadas e suas funções. As regiões sob seleção são regiões do genoma em que os marcadores apresentam-se em alto desequilíbrio de ligação, sendo denominados de “selective sweeps” (Griffiths et al., 2012).

A inclusão da informação molecular nos programas de melhoramento também permite calcular corretamente o parentesco (Legarra, 2017) e a endogamia (VanRaden, 2008; Forutan et al., 2018). A estimação da matriz genômica de parentesco e seu uso nas avaliações genéticas garantem aumento na acurácia das estimativas dos valores genéticos (EBV) e aumento da estimativa de herdabilidade em características de baixa herdabilidade ou de difícil mensuração (Meuwissen et al., 2001). Segundo os autores, maior ganho genético pode ser esperado quando a informação genômica é considerada nas avaliações dado a redução do intervalo de geração.

A inclusão da informação genômica em pesquisas tem sido possível até mesmo em pequenas populações e países mais pobres devido à redução no custo da

genotipagem. Por esta razão, atenção deve ser dada às novas técnicas com uso de painéis de SNPs, aplicadas ao melhoramento e à conservação dos animais domésticos.

2. Revisão de literatura

2.1. Estrutura genômica da população

O estudo da estrutura da população é importante para conhecimento histórico de populações e orientação em programas de conservação ou melhoramento genético. A Análise de Componentes Principais (PCA) e a Análise Discriminante de Componentes Principais (DAPC) possibilitam verificar a estrutura da população e compreender o parentesco entre e dentro de populações em nível molecular (Jombart et al., 2010). A PCA possibilita a classificação dos indivíduos por meio de reduzido número de importantes componentes principais (PC) ortogonais (Dunteman, 1989). Cada PC está relacionado ao seu autovalor que descreve o total da inércia explicada pelo componente. Quando utilizados no contexto de estudo genômico populacional, o autovalor indica parte da variabilidade genética representada por cada PC associado. O primeiro PC está relacionado a uma alta quantidade de inércia reproduzindo a estrutura dos dados genéticos (Jombart e Pontier, 2009).

É possível que a PCA induza a estratificação de populações de acordo com as diferenças demográficas históricas de cada população, ou seja, animais representantes de um mesmo rebanho ou mesma região do país sejam colocados em um mesmo “cluster”. A DAPC pode ser utilizada de forma complementar aos resultados da PCA. A estratificação dos dados gerada pela DAPC não é influenciada por “clusters” pré-definidos geograficamente. A DAPC pode ser utilizada para testar o número exato de “clusters” (K) entre todos os “clusters” obtidos pela PCA. Nesta metodologia, a escolha do valor de K é realizada por meio do menor valor associado de BIC (“Bayesian Information Criterion”) (Jombart et al., 2010). A estimação de K pela metodologia DAPC vem mostrando êxito na inferência sobre o número exato de “clustes” e diversidade genética em estudos de estrutura da população em bovinos (Wang et al., 2015;

Kukučkova et al., 2017). Além disso, esta metodologia possibilita a visualização do resultado da estrutura genética da população por meio de gráficos das funções discriminantes da DAPC.

A pressão de seleção durante a domesticação dos bovinos resultou no aumento da diversidade genética entre as populações recentes. Populações de mesma raça com históricos de seleção divergentes apresentam diferenças alélicas, e, portanto, diversidade genética entre os indivíduos e estratificação das populações. A seleção artificial seria um dos fatores responsáveis por gerar variabilidade genética entre indivíduos (Chen et al., 2016).

2.2. Índice de fixação

A diversidade genética entre as populações recentes de bovino faz com que existam diferenças morfológicas, fisiológicas, produtivas e reprodutivas entre elas (Chen et al., 2016). Os padrões de variações genéticas entre as populações são de particular interesse para os estudos de domesticação das espécies, formação das raças e para compreender as consequências da seleção (Kijas et al., 2012).

Uma das medidas mais utilizadas para identificar diferenciação genética entre populações é o índice de fixação (Fst), proposto por Wright em 1951, o qual possibilita identificar padrões de variação genética entre duas populações à nível de *loci*. Além disso, o Fst é uma das medidas mais comuns para detectar assinaturas de seleção em animais quando calculado a partir de informações de SNPs. Este método é mais apropriado para detectar eventos que ocorreram em um passado mais distante (Cadzow et al., 2014).

É esperado que o nível de diferenciação genética entre populações seja baixo em regiões neutras do genoma ou em regiões de seleção balanceada. O Fst é utilizado para quantificar as diferenças alélicas entre as populações, e, teoricamente, varia de zero a um, sendo que zero indica nenhuma diferenciação e um indica completa diferenciação em que as subpopulações estão fixadas para diferentes alelos. Os valores de Fst por *loci* podem ser utilizados para identificar assinaturas de seleção entre diferentes grupos, ou

seja, *loci* nos quais alelos estão fixados de forma diferente entre os grupos (Qanbari et al., 2011) e também podem ser utilizados para entender como a seleção para diferentes critérios pode afetar os padrões genômicos nestes grupos (Zhao et al., 2015). A detecção de assinaturas de seleção é realizada por meio de diferenças nas frequências alélicas, sendo que um limiar de corte deve ser estabelecido e os SNPs com valores de Fst que ultrapassam esse limiar são considerados como regiões com evidências de assinaturas de seleção.

2.3. Haplótipos e assinaturas de seleção

Métodos baseados em haplótipos também têm sido aplicados em estudos que buscam identificar assinaturas de seleção. A recombinação quebra o desequilíbrio de ligação ao redor de uma mutação, diminuindo o tamanho do haplótipo no qual a mutação está localizada (Gautier et al., 2017). Em 2002, Sabeti et al. introduziram o método da Homozigose do Haplótipo Estendido (EHH) para explorar o decaimento da homozigose do haplótipo como uma função da distância genética de determinado SNP. Posteriormente, métodos complementares ao EHH surgiram. O Escore de Integração dos Haplótipos (iHS), proposto por Voight et al. (2006), e o teste da Homozigose do Haplótipo Estendido Entre Populações (XP-EHH), proposto por Sabeti et al. (2007), apresentam acurada detecção de regiões do genoma que são evidências de seleção. O iHS é baseado na padronização do log da razão das integrais do decaimento observado de EHH calculado para os alelos ancestral e derivado em um SNP alvo. O XP-EHH é utilizado para comparar os perfis de EHH de duas populações. Recentemente, o teste iHS tem sido aplicado para detectar assinaturas de seleção em humanos (Gautier et al., 2017), plantas (He et al., 2017) e animais domésticos (Zhu et al., 2015; Bahbahani et al., 2017; Msalya et al., 2017).

Diferente do Fst, o teste iHS possibilita detectar seleção positiva recente, enquanto que o XP-EHH é sensível em detectar *loci* completamente ou próximo de serem fixados (Sabeti et al., 2007). No teste iHS, as informações genômicas são diretamente comparadas com a informação de um ancestral, de forma que este teste tem aplicação

na detecção de assinaturas de seleção dentro da população. Já o XP-EHH possibilita a detecção de regiões candidatas à seleção entre populações pela comparação dos perfis de EHH de uma população referência e uma população observada (Sabeti et al., 2007). A ideia principal do XP-EHH é testar se determinado *locus* do genoma é homozigoto em uma população e polimórfico em outra. A informação genômica de ancestral não é necessária para a realização do teste XP-EHH. Devido as propriedades de cada teste, os resultados dos métodos de identificação de assinaturas de seleção devem ser utilizados de forma complementar. A incorporação destes métodos no estudo da diferenciação genética e regiões sob seleção poderia contribuir para a compreensão do fluxo gênico e histórico de populações geneticamente distantes ou próximas.

Os *loci* identificados como assinaturas de seleção pelas técnicas baseadas no EHH podem ser estudados com intuito de verificar se estes sobrepõem com regiões de genes, e estas informações podem atuar como informações complementares aos estudos recentes de busca por genes associados com características de interesse econômico, que seriam os Estudos de Associação Ampla do Genoma (GWAS). Por meio da comparação das regiões identificadas como evidências de seleção nos testes baseados em EHH e resultados de GWAS, é possível testar a contribuição de genes na variabilidade genética de fenótipos de interesse econômico, que são utilizados na seleção genômica (Chen et al., 2016).

2.4. Aplicações de estudos de assinaturas de seleção

O testes baseados no Fst e em haplótipos têm sido utilizados para detectar regiões caracterizadas como fortes evidências de assinaturas de seleção em humanos (Gautier et al., 2017), plantas (He et al., 2017) e diversas espécies de interesse econômico como bovinos (Zhao et al., 2015; Urbinati et al., 2016; Cardoso et al., 2018) e ovinos (Zhu et al., 2015).

Estas metodologias vêm sendo aplicadas em populações bovinas selecionadas no Brasil. Urbinati et al. (2016) usaram o teste iHS e encontraram regiões com evidências de assinaturas de seleção nos cromossomos 5 e 14 da raça Canchim. O estudo de

Urbinati et al. (2016) é amplo em abordar as regiões do genoma que teriam colaborado com a formação da raça Canchim e fornece informações sobre haplótipos que teriam sido, possivelmente, selecionados nos animais fundadores desta raça. Na raça Nelore, Cardoso et al. (2018) identificaram *loci* localizados em regiões gênicas previamente descritas para características de crescimento. As assinaturas de seleção encontradas na raça Nelore se sobrepuseram com os genes *TBC1D1, CRH* e *XKR4,* que estão relacionados com crescimento, e novas regiões candidatas (*COL14A1, PSAPL1* e *CPT1C1*) foram descritas. Estes estudos foram importantes em fornecer informações sobre genes que controlam características quantitativas sob seleção nestas duas populações, evidenciando a importância e aplicação destas análises no melhoramento genético.

2.5. Seleção genômica e matriz genômica de parentesco

A seleção tradicional utiliza modelos mistos para a análise de dados de fenótipos, juntamente com registros de pedigree, para se obter as soluções das estimativas dos valores genéticos dos candidatos a seleção. Em 2001, Meuwissen et al. fizeram algumas previsões sobre o benefício da inclusão da informação genômica na avaliação genética dos animais e previam que essa inclusão possibilitaria o cálculo do valor genético estimado (EBV) com base em milhares de SNPs sem a necessidade de usar dados de fenótipos ou pedigree para os animais genotipados, e aumentaria a acurácia de predição dos EBVs e a acurácia para características de baixa herdabilidade e difícil mensuração e selecionaria animais precocemente, reduzindo o intervalo de geração. Pesquisas com seleção genômica em dados reais mostraram que algumas das suposições de Meuwissen et al. (2001) foram atingidas. Contudo, os desafios dos estudos de seleção genômica ainda incluem aumentar a acurácia de predição do valor genético genômico (GEBV) e tornar a avaliação mais acessível em termos de custo de genotipagem.

O Melhor Preditor Linear Não Viesado (BLUP) é uma metodologia baseada na semelhança entre parentes devido à efeitos genéticos. O parentesco entre todos os animais incluídos na avaliação genética geralmente advém das informações de pedigree.

Com as informações de parentesco, é possível calcular a proporção da variação fenotípica devido a efeitos genéticos e ambientais. Uma única matriz de parentesco (H) que combina a matriz de parentesco baseada no pedigree (A) e matriz genômica de parentesco (G), incluindo, portanto, informação de SNPs, pode ser aplicada em um procedimento BLUP, sendo então chamado de “Single Step Genomic” BLUP (ssGBLUP) (Legarra et al., 2014). Este procedimento tornou-se a melhor opção para a avaliação genômica.

O desenvolvimento da metodologia ssGBLUP foi realizado, de forma paralela, por Legarra et al. (2009) e Christensen and Lund (2010). As ideias iniciaram de pontos de vistas diferentes e levaram ao desenvolvimento da expressão que representa a matriz H, expressa como:

$$H = \begin{pmatrix} A_{11} + A_{12}A_{22}^{-1}A_{21} + A_{12}A_{22}^{-1}GA_{22}^{-1}A_{21} & A_{12}A_{22}^{-1}G \\ GA_{22}^{-1}A_{21} & G \end{pmatrix}$$

e sua inversa desenvolvida, posteriormente, por Aguilar et al. (2010):

$$H^{-1} = A^{-1} + \begin{pmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{pmatrix}$$

As informações dos genótipos dos indivíduos, ao serem consideradas no ssGBLUP, influenciam tanto o coeficiente de parentesco dos indivíduos genotipados como o coeficiente de parentesco dos indivíduos não genotipados (Legarra et a., 2014).

2.6. Parentesco entre irmãos completos e meios-irmãos

Os marcadores moleculares podem ser utilizados para correção de erros no pedigree. As correções no pedigree devem melhorar a acurácia de predição dos valores genéticos e, consequentemente, aumentar o ganho genético (VanRaden, 2007; Munoz et al., 2013). Além disso, os marcadores moleculares permitem o cálculo do parentesco realizado, o qual difere do parentesco esperado no pedigree (Legarra, 2017). Quando somente informação de pedigree é considerada, assume-se que dois irmãos completos compartilham metade do genoma, uma vez que uma população base não relacionada é considerada no cálculo do parentesco. Neste caso, assume-se que a população base ou fundadores não compartilham alelos em comum. Com o uso da informação presente no

DNA, acessada como, por exemplo, por painéis de alta densidade de SNPs, as proporções herdadas dos genomas parentais podem variar entre os dois irmãos completos, determinando se serão mais ou menos semelhantes com base na proporção compartilhada dos genomas ou de alelos compartilhados para os *loci* que afetam uma característica (Legarra, 2017).

VanRaden (2007) simulou o parentesco entre irmãos completos e meios-irmãos para verificar quão relacionados são os indivíduos quando o número de SNPs varia. Os resultados encontrados pelo autor mostraram que os valores de covariância entre parentes são diferentes dos valores esperados no pedigree, que seria de 0,50 entre irmão completos e 0,25 entre meios-irmãos. Segundo o autor, os resultados de simulações quando o número de *loci* é maior que 10,000 seriam semelhantes ao resultado encontrado para 100 *loci*, quando os *loci* estão aleatoriamente distribuídos ao longo de 30 cromossomos. A implicação disso para o melhoramento genético animal é que maior acurácia de predição dos valores genéticos pode ser obtida com a inclusão dos parentescos genômicos na avaliação genética, especialmente se o número de indivíduos genotipados for grande.

2.7. Aplicação da seleção genômica e importância da simulação de dados

Como descrito anteriormente, uma das vantagens da aplicação da seleção genômica no melhoramento genético animal é a obtenção de maior acurácia de predição dos valores genéticos. Resultados satisfatórios foram obtidos por pesquisadores em populações reais selecionadas. Lourenco et al. (2015) estudaram características de crescimento em bovinos Angus e obtiveram ganho de 10 a 21% quando estimaram o valor genético genômico (GEBV). A acurácia do GEBV aumentou de 0,32 para 0,35. O primeiro valor foi obtido quando os genótipos de 2.000 animais foram incluídos nas análises, e o segundo foi obtido com o uso de genótipos de 33.000 animais. Isto evidencia que o aumento do número de animais genotipados nas análises proporciona aumento na acurácia do GEBV.

Outra questão acerca da aplicação da informação genômica no melhoramento animal seria se o aumento do número de marcadores beneficiaria a acurácia de predição dos GEBV. VanRaden et al. (2011) responderam esta questão estudando a acurácia de predição do GEBV em bovinos da raça Holandesa em diferentes cenários, com de painéis de 3K, 50K e 500k SNPs. Os autores identificaram pequeno aumento na acurácia do GEBV com o aumento da densidade dos painéis de SNPs. Os resultados encontrados tanto por Lourenco et al. (2015) como por VanRaden et al. (2011) sugerem que seria mais importante aumentar o número de animais genotipados nas análises do que aumentar o número de SNPs no painel. Desta forma, as informações contidas nos painéis de baixa-moderada densidades seriam suficientes para determinar o parentesco entre os indivíduos.

Dados obtidos por VanRaden et al. (2011) para o cenário de 500K SNPs foram simulados de acordo com a estrutura da população e o desequilíbrio de ligação obtidos em dados reais de 33.414 animais Holandeses genotipados com o painel de 50K SNPs. O primeiro trabalho com seleção genômica que surgiu na literatura, desenvolvido por Meuwissen et al. (2001), também fez uso da simulação de dados para comparar acurácias de predição dos valores genéticos obtidos por meio das metodologias de quadrados mínimos, BLUP e Bayesiana. Desta forma, a simulação de dados pode ser utilizada no melhoramento genético para testar hipóteses e programas. Além disso, a simulação possibilita obter grande conjunto de dados para serem investigados nas avaliações genéticas.

A hipótese testada por Tonussi et al. (2017) foi em relação ao uso do ssGBLUP para obtenção de estimativas de GEBV mais acurados e menos viesados em uma população simulada de bovinos. O uso de touros múltiplos em acasalamentos com monta natural é uma prática comum nas criações de bovinos de corte no Brasil. Apesar de trazer vantagens relacionadas ao manejo do rebanho, o uso de touros múltiplos não possibilita a identificação de paternidade das progênies e favorece o aumento de ocorrência de registros faltantes no pedigree, comprometendo a confiabilidade da avaliação genética. De acordo com os autores, o uso de dados simulados foi importante para avaliar a viabilidade técnica do ssGBLUP em diferentes cenários com incerteza de paternidade.

3. Objetivos

I – Verificar a estruturação dos dados genômicos de duas populações da raça Gir selecionadas para corte ou leite, usando SNPs arrays; realizar análises de assinaturas de seleção por meio de três métodos diferentes (Fst, iHS e XP-EHH); e realizar análises funcionais de genes presentes dentro das regiões identificadas como assinaturas de seleção nesta raça para verificar suas funções biológicas.

II – Investigar se a estimativa do efeito materno em suínos para característica medida na progênie pode ser melhorada com uso de “pool” de sêmen em diferentes cenários simulados, considerando que o parentesco entre possível pai e leitegada possa ser acessado pela informação genômica.

4. Referências bibliográficas

Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ (2010) Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. **Journal of Dairy Science** 93:743–752.

Bahbahani H, Tijjani A, Mukasa C, Wragg D, Almathen F, Nash O, Akpa GN, Mbole-Kariuki M, Malla S, Woolhouse M, Sonstegard T, Van Tassell C, Blythe M, Huson H, Hanotte O (2017) Signatures of Selection for Environmental Adaptation and Zebu×Taurine Hybrid Fitness in East African Shorthorn Zebu. **Frontiers in Genetics** 8:1–20.

Cardoso DF, Albuquerque LG, Reimer C, Qanbari S, Erbe M, Nascimento AV, Venturini GC, Scalez DCB, Baldi F, Camargo GMF, Mercadante MEZ, Cyrillo JNSG, Simianer H, Tonhati H (2018) Genome-wide scan reveals population stratification and footprints of recent selection in Nelore cattle. **Genetics Selection Evolution** 50:22.

Cadzow M, Boocock J, Nguyen HT, Wilcox P, Merriman TR, Black MA (2014) A bioinformatics workflow for detecting signatures of selection in genomic data. **Frontiers in Genetics** 5:293.

Chen M, Pan D, Ren H, Fu J, Li J, Su G, Wang A, Jiang L, Zhang Q, Liu JF (2016) Identification of selective sweeps reveals divergent selection between Chinese Holstein and Simmental cattle populations. **Genetics Selection Evolution** 48:76.

Dunteman GH (Eds.) (1989) **Principal components analysis**: quantitative applications in the Social Sciences. Newbury: Sage Publications, 96p.

Forutan M, Mahyari SA, Baes C, Melzer N, Schenkel FS, Sargolzaei M (2018) Inbreeding and runs of homozygosity before and after genomic selection in North American Holstein cattle. **BMC Genomics** 19:98.

Gautier M, Klassmann A, Vitalis R (2017) Rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. **Molecular Ecology Resources** 17:78–90.

Griffiths AJF, Wessler S, Carroll SB, Doebley J (Eds.) (2012) **Introduction to Genetic Analysis.** 10th ed. New York: W. H. Freeman. 864 p.

He Q, Kim KW, Park YJ (2017) Population genomics identifies the origin and signaturesof selection of Korean weedy rice. **Plant Biotechnology Journal** 15:357–366.

Jombart T, Pontier D, Dufour AB (2009) Genetic markers in the playground of multivariate analysis. **Heredity** 102:330–341.

Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. **BMC Genetics** 11:94.

Kijas JW, Lenstra JA, Hayes B, Boitard S, Neto LRP, Cristobal MS, Servin B, McCulloch R, Whan V, Gietzen K, Paiva S, Barendse W, Ciani E, Raadsma H, McEwan J, Dalrymple B (2012) Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. **PLoS Biology** 10:1–14.

Kukučkova V, Moravčıkova N, FerenčakovićM, SimčičM, Meszaros G, Solkner J, Trakovická A, Kadlečík O, Curik I, Kasarda R (2017) Genomic characterization of Pinzgau cattle: genetic conservation and breeding perspectives. **Conservation Genetics** 18:893–910.

Legarra A, Christensen OF, Aguilar I, Misztal I (2014) Single Step, a general approach for genomic selection. **Livestock Science** 166:54-65.

Legarra A (2017) **Bases for genomic prediction**. Version 0.9.6. Castanet Tolosan, France: INRA: Animal Genetics Department, 81p. (INRA-Animal Genetics Department. Documentos).

Lourenco DA, Tsuruta S, Fragomeni BO, Masuda Y, Aguilar I, Legarra A, Bertrand JK, Amen TS, Wang L, Moser DW, Misztal I (2015) Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. **Journal of Animal Science** 93:2653-2662.

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. **Genetics** 157(4):1819-29.

Msalya G, Kim E, Laisser ELK, Kipanyula MJ, Karimuribo ED, Kusiluka LJM, Chenyambuga SW, Rothschild MF (2017) Determination of genetic structure and signatures of selection in three strains of Tanzania Shorthorn Zebu, Boran and Friesian Cattle by Genome-Wide SNP Analyses. **PLoS One** 1–18.

Munoz PR, Resende Jr. MFR, Huber DA, Quesada T, Resende MDV, Neale DB, Wegrzyn JL, Kirst M, Peter GF (2013) Genomic relationship matrix for correcting pedigree errors in breeding populations: Impact on genetic parameters and genomic selection accuracy. **Crop Science** 53:1115–1123.

Qanbari S, Gianola D, Hayes B, Schenkel F, Miller S, Moore S, Thaller G, Simianer H (2011) Application of site and haplotype frequency based approaches for detecting selection signatures in cattle. **BMC Genomics** 12:318.

Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES (2002) Detecting recent positive selection in the human genome from haplotype structure. **Nature** 419:832–837.

Tonussi RL, Silva RMO, Magalhães AFB, Espigolan R, Peripolli E, Olivieri BF, Feitosa FLB, Lemos MVA, Berton MP, Chiaia HLJ, Pereira ASC, Lobo RB, Bezerra LAF, Magnabosco CU, Lourenco DAL, Aguilar I, Baldi F (2017) Application of single step genomic BLUP under different uncertain paternity scenarios using simulated data. **PLOS ONE** 1-14.

Urbinati I, Stafuzza NB, Oliveira MT, Chud TCS, Higa RH, Regitano LC, Alencar MM, Buzanskas ME, Munari DP (2016) Selection signatures in Canchim beef cattle. **Journal of Animal Science and Biotechnology** 7:29.

Van Raden PM (2007) Genomic measures of relationship and inbreeding. **Interbull Bulletin** 37:33-36.

VanRaden PM (2008) Efficient methods to compute genomic predictions. **Journal of Dairy Science** 91:4414-4423.

VanRaden PM, O'Connell JR, Wiggans GR, Weigel KA (2011) Genomic evaluations with many more genotypes. **Genetics Selection Evolution** 43:10.

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. **PLoS Biology** 4:446–659.

Wang MD, Dzama K, Hefer CA, Muchadeyi FC (2015) Genomic population structure and prevalence of copy number variations in South African Nguni cattle. **BMC Genomics** 16:894.

Wright S (1951) The genetical structure of populations. **Annals of Eugenics** 15: 323-354.

Zhao F, McParland S, Kearney F, Du L, Berry DP (2015) Detection of selection signatures in dairy and beef cattle using high-density genomic information**. Genetics Selection Evolution** 47:49.

Zhu C, Fan H, Yuan Z, Hu S, Zhang L, Wei C, Zhang Q, Zhao F, Du L (2015) Detection of selection signatures on the X Chromosome in three sheep breeds. International Journal of Molecular Sciences 16:20360–20374.Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ (2010) Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. **Journal of Dairy Science** 93:743–752.

CHAPTER II - ASSESSING GENETIC ARCHITECTURE AND SIGNATURES OF SELECTION OF DUAL PURPOSE GIR CATTLE POPULATIONS USING GENOMIC INFORMATION [a]

Abstract - Gir is one of the main cattle breeds raised in tropical South American countries. Strong artificial selection through its domestication resulted in increased genetic differentiation among the countries in recent years. Over the years, genomic studies in Gir have become more common. However, studies of population structure and signatures of selection in divergent Gir populations are scarce and need more attention to better understand genetic differentiation, gene flow, and genetic distance. Genotypes of 173 animals selected for growth traits and 273 animals selected for milk production were used in this study. Clear genetic differentiation between beef and dairy populations was observed. Different criteria led to genetic divergence and genetic differences in allele frequencies between the two populations. Gene segregation in each population was forced by artificial selection, promoting isolation, and increasing genetic variation between them. Results showed evidence of selective forces in different regions of the genome. A total of 282 genes were detected under selection in the test population based on the fixation index (Fst), integrated haplotype score (iHS), and cross-population extend haplotype homozygosity (XP-EHH) approaches. The QTL mapping identified 35 genes associated with reproduction, milk composition, growth, meat and carcass, health, or body conformation traits. The investigation of genes and pathways showed that quantitative traits associated to fertility, milk production, beef quality, and growth were involved in the process of differentiation of these populations. These results would support further investigations of population structure and differentiation in the Gir breed.
Keywords: allele frequency, fixation index, genes, haplotypes,

[a] Article published in the “PLOS ONE” Journal: doi: 10.1371/journal.pone.0200694

1. Introduction

Gir is one of the main *Bos indicus* cattle breeds native from India that was first introduced in Brazil in late 1800s. Although Gir is an important breed for milk production in tropical South American countries, small populations are still found in North America and Australia. Few animals (<700) were imported in Brazil between 1870 and 1962. In 1938, the Brazilian Association of Zebu Cattle Breeders (ABCZ) started the registration of Gir cattle [1] and contributed to the dissemination of this breed in Brazil. In the 1960s, a number of Brazilian breeders started selecting for dual-purpose (milk and meat), and others selected only for milk production. Since 1993, the majority of the Gir breeders have been selecting mainly for milk yield [1].

Because of the importance for livestock and conservational systems, along with the reduced cost of whole-genome genotyping, genomic studies in Gir have become more common in the latest years. Most of the recent studies in Gir cattle are related to genetic structure, genetic diversity and inbreeding levels based on pedigree data in dairy herds [1], whole genome sequencing [2,3] for studying regions under selection for environmental adaptation [4], and genetic differentiation compared with other breeds [5]. Another study in a Gir population identified low levels of taurine introgression during the formation of this breed [6]. These studies were relevant for understanding the Gir history and exploiting the genetic background of the population. In 2013, three selective sweeps were identified and overlapped with the *ST6GALNAC5* gene in Gir [7]. Recently, Peripolli et al. (2018) [8] reported 14 runs of homozygosity (ROH) islands in a Brazilian Gir dairy population, including the animals used by Utsunomiya et al. (2013) [7]. ROH islands are regions of the genome with high levels of homozygosity and have been used to assess regions under selective pressure.

Strong artificial selection through domestication resulted in increased diversity among recent cattle populations. Diversity includes variation in morphology, physiology, production, and fertility [9]. Assessing patterns of genetic variation is of particular interest for studying domestication, breed formation, population structure, and consequences of selection [10]. Several measures of genetic differentiation have been proposed, and one of them is the fixation index (Fst; Wright (1951) [11]). Several studies have used Fst as a

tool for identifying patterns of genetic variation at a locus among populations relative to that within populations. Thus, Fst has also been used as a test for identifying signatures of selection in populations [11, 12] by using high-throughput SNP information. In fact, Fst is one of the most commonly used metrics for detecting signatures of selection in animals.

Besides Fst, haplotype-based methods have also been used in signatures of selection studies. The integrated haplotype score (iHS) proposed by Voight et al. (2006) [13] and the cross-population extend haplotype homozygosity (XP-EHH) test proposed by Sabeti et al. (2007) [14] can help to spot evidence of selection with high detection power. Recombination has the power to break down linkage disequilibrium around a mutation, decreasing the length of haplotypes on which a mutation is located [15]. Sabeti et al. (2002) [16] introduced the extended haplotype homozygosity (EHH) to exploit the decay of haplotype homozygosity as a function of genetic distance from a focal SNP. Later, iHS was proposed as a complementary approach to the EHH method and is based on the standardized log ratio of the integrals of the observed decay of EHH computed for the ancestral and the derived alleles at the focal SNP [13]. The iHS test has been currently applied for detecting selection signatures in humans [15], plants [17], and domestic animals [4, 18, 19]. The XP-EHH is also based on the EHH method and is used to detect selection footprints between populations by comparing their EHH profiles.

The Fst and the haplotype-based approaches are somehow different. The time scale over which selection has been occurred has a major impact on the ability of each method to detect evidence of selection. The Fst method is best suited for detection in events occurring in the more distant past [20] whereas the iHS test is best suited for detection in recent positive selection [15]. The XP-EHH is useful for detection of entirely or approximately fixed loci [14]. As the Fst, iHS, and XP-EHH approaches are complementary, incorporating such measures in the study of genetic differentiation and selection would strongly contribute to understanding gene flow and genetic makeup of dual-purpose (meat and milk production) Gir cattle populations.

So far, a limited number of studies have been conducted to identify population structure and signatures of selection in Gir. O'Brien et al. (2015) [6] sampled bulls that were widely used for artificial insemination in dairy breeding systems, reflecting only the

top animals in the breed. A study based on the iHS test did not detect significant numbers of signatures of selection in the breed probably because very few individuals were used (N=53). Only the *ST6GALNAC5* gene was found harboring evidence of selection in Gir [7]. A sample of animals selected for different purposes from production herds across the country would be useful to understand genetic differentiation and selection footprints in the Brazilian Gir populations.

In this study, we present a comprehensive analysis of population structure and signatures of selection in two populations of Gir cattle selected either for beef or milk production, using high-throughput genomic information. Three methods (Fst, iHS, and XP-EHH) were implemented to scan the whole genome of those two populations. Afterwards, we performed a functional study of the genes identified within the regions harboring signatures of selection in each population to explain the biological importance of selection footprints. The information from our research can be useful for future GWAS studies, conservation, and genetic improvement of the Gir breed.

2. Material and Methods

2.1. Data resource

The experiment was conducted in accordance with animal welfare guidelines according to State Law No. 11.977 of the São Paulo state, Brazil. All animal procedures were approved by the Ethics and Animal Handling Committee of the Instituto de Zootecnia, Nova Odessa, SP, Brazil.

Genotyped animals from two distinct populations were used in this study. A set of 273 female Gir were obtained from the Brazilian Program of Dairy Gir Genetic Improvement (PNMGL), which is a breeding program. These animals belonged to five farms located in Minas Gerais and Sao Paulo states. More details of the history and genetic background of this population can be found in Santana Jr et al. (2014) [1]. A set of 173 males and females were obtained from the Animal Science Institute (IZ; Sertaozinho, SP), which started a research program for growth traits in indicine breeds in 1976, including Gir.

2.2. Differentiation between dairy and beef populations

Among all the Gir breeding or research programs in Brazil, two are of special interest: a) PNMGL aims improving milk production traits; b) IZ, which is considered unique, aims improving beef production traits in a closed herd scheme. The PNMGL was created in 1985 to conduct progeny testing programs [1], where the main objective was to identify genetically superior bulls for milk production, based on progeny performance. Traits such as milk production, milk composition (fat and protein), somatic cell counts, handling (e.g., ease of milking and temperament), and body conformation (e.g., ligament last udder, rear udder height, rear udder width, and length and diameter of the teats) are included in the breeding program. The mean 305-day milk yield of the current dairy Gir population is 3000 ± 1500 kg [1]. Because of the well-designed breeding program, semen of genetically proven bulls from PNMGL are commercialized worldwide. Recently, genomic information started being used in genetic evaluation of the Brazilian Gir cattle.

The IZ started the first breeding season in 1976 with different lines that were introduced in the Gir herd to increase genetic variability and avoid problems with inbreeding. Since the first breeding season, only animals from this herd were used for mating, making it a closed herd. The choice for founders was based on their yearling body weight (YBW) at 550 days. The first genetic evaluation was performed in 1981 [21] and the selection criterion for sires was the final body weight at 378 days of age, after 168-day feedlot performance tests. Females remained on pasture, and the selection criterion for females was their YBW at 550 days of age. For this herd, the generation interval was 5.65 years, and genetic gains were 2.88 kg/year for the weight of sires after the weight gain test and 2.80 kg/year for female considering the adjusted weight at 550 days [22]. Records of milk production were not available in this herd, animals were selected for growth traits.

2.3. SNP array data and quality control

Animals used in this study were genotyped with the GGP Bovine LDv4 (Illumina, San Diego, CA) with approximately 33K SNP. Only SNPs located in the 29 autosomes remained for the analysis. Quality control was performed to ensure high quality of genomic data. Therefore, SNPs were removed when they were monomorphic, had a call rate lower

than 0.90, and had a minor allele frequency (MAF) lower than 0.01, and when the difference between expected (Hardy-Weinberg equilibrium) and observed allele frequencies at a given locus was greater than 0.15. Samples with a call rate lower than 0.90 were also removed. After the quality control, a total of 442 animals and 23,275 SNPs were retained for further analyses. Filtering the SNPs based on MAF may affect the probability of identifying alleles related to selection [23]. For this reason, a low threshold for MAF (<0.01) was imposed as a criterion for the SNP quality control.

Imputation of missing SNPs and phasing were performed via Beagle 3.3.2 [24]. The phased data were annotated with ancestral reference alleles, and then the haplotype file and the physical map were used for the iHS analysis.

2.4. Genetic structure of the population

To access the Gir population structure and to understand the relationship within and between populations in a genomic level, Principal Component Analysis (PCA) and Discriminant Analysis of Principal Components (DAPC) were performed using the adegenet package for the R software [25]. The PCA approach allows classifying individuals based on the reduced number of important orthogonal principal components (PC) [26]. Each PC relates its eigenvalue that describes the amount of total inertia explained on the component. When used in the context of genomic analysis of population structure, the eigenvalues indicate part of the total genetic variability represented by the associated PC. The first PC related to a high amount of inertia reproduces the structuring of the genetic data [27]. To perform PCA with different subsets of SNP markers, the default options for the *glPCA* function were used, allowing compensating for differences in variance among allele frequencies [27].

The DAPC was applied to select the optimal number of clusters (K), among all output clusters from PCA. The choice of K was made on the basis of the lowest associated Bayesian Information Criterion (BIC; Jombart et al. (2010) [25]). When the optimal number of clusters is ambiguous, K increases as long as it resulted in a noticeable improvement in BIC. DAPC uses K-means clustering of PC in order to infer the actual number of populations [25]. The DAPC was run using 100 PC and 10 discriminant functions.

2.5. Divergent selection between populations

To verify the genetic divergence between two populations, genetic statistics such as observed heterozygosity (Ho), genetic diversity within population (Hs), total genetic diversity (Ht), coefficient of inbreeding (Fis), and fixation index (Fst) were computed using the HierFstat R package [28]. The fixation index (Fst), as defined in Nei (1987) [29], was used as a measure of genetic differentiation between populations being derived from the equation:

$$Fst = \frac{Ht - Hs}{Ht}$$

Fst quantifies differences in allele frequencies between populations, and theoretically its value ranges from zero to one, meaning that there is no differentiation or complete differentiation in which subpopulations are fixed for different alleles, respectively. Fst values are effective for identifying selection signatures between different groups, i.e., loci in which alleles are fixed differently in different groups [30] and allow determining how the divergent selection can affect the genomic pattern of these groups [31].

As a way to compare Fst values for each SNP, measures of centrality and dispersion such as mean and standard deviation were considered. The negative Fst values observed for 2,794 SNPs were set to zero, since negative values have no biological interpretation [32]. Loci were plotted relative to their physical position within each autosome. The threshold to call an SNP outlier was defined as three standard deviations above the mean. This methodology allows identifying SNPs with Fst values that stand out from the others, and that could be related to genes affecting adaptive and/or economical important traits for beef or milk production, meaning evidence of selection signatures. Similar approaches have been used in other studies for identifying selection signatures [12, 31]. A control chart approach allows the partition of Fst variation into a component due to selection and also allows the discovery of significant SNPs based on control limits set at three standard deviations from the mean [12]. Kijas et al. (2012) [10] and Zhao et al. (2015) [31] used the top 0.1% Fst values to represent selection signature.

The approach adopted in the present study is a simplistic technique that allows to access the variation patterns on many loci and considers values above the cut-off as

evidence of divergent selection. Assuming that the data follow a normal distribution, a Fst value above three standard deviations from the mean indicate that the locus has a value higher than 99.8% of total SNPs in the chromosome.

2.6. Within population test

The iHS test was used to detect strong footprints of selection within the studied population. This test is based on the standardized log ratio of integrals of the observed decay of extended haplotype homozygosity (EHH), computed for both ancestral and derived alleles at the focal SNP.

Phased haplotypes produced by Beagle 3.3.2 and ancestral allele information were submitted to iHS-based tests. For each SNP, ancestral and derived alleles were determined according to the 50K SNP annotation file provided in Rocha et al. (2014) [33], which included information of putative ancestral alleles of bovine SNPs. The number of SNPs in common between the GGP LDv4 SNP chip and the 50K SNP was 26,058. The phased haplotypes were annotated with ancestral allele information (via R package rehh 2.0 [15]). For each allele in dairy and beef populations, information was defined as missing if the code was incompatible with ancestral and derived alleles. The iHS score was computed for each autosomal SNP using the R package rehh 2.0 [17]. Default options were generally used, except for the minimal threshold on the minor allele frequency that was set to 0.01 and the percentage of retained haplotypes that were changed to suit the Gir haplotype data. In addition, the standardization of iHS was performed with allele frequency bins of 0.01, which is controlled by the ihh2ihs function in the package [15].

According to Voight et al. (2006) [13], the iHS of a given focal SNP (iHS) was defined as

$$iHS = \frac{UniHS - \mu_{UniHS}^{p_s}}{\sigma_{UniHS}^{p_s}}$$

where UniHS = log($iHH_{ancestral}$/$iHH_{derived}$) (where iHH is the integrated allele-specific extended haplotype homozygosity for core SNP alleles (ancestral and derived)), and $\mu_{UniHS}^{p_s}$ and $\sigma_{UniHS}^{p_s}$ are the average and the standard deviation of the UniHS, respectively, computed over all the SNPs with a derived allele frequency p_s similar to that of the core

SNPs. The iHS was constructed to approximately follow a standard Gaussian distribution and to enable comparisons among SNPs regardless of their underlying allele frequencies.

The iHS is transformed into piHS, as shown by [34]:

$$p_iHS = -log_{10}(1 - 2|\Phi(iHS) - 0.5|)$$

where $\Phi(x)$ represents the Gaussian cumulative distribution function. Assuming that most of the iHS values follow the Gaussian distribution under neutrally, piHS may be interpreted as a two-sided *p*-value in a -log10 scale [15].

To control for false positives, the fndr.cutoff function available in the *fdrtool* R package [35] was used with its default options for calculating a *p*-value, which defines the cut-off point chosen according to the false non-discovery rate. After the false discovery rate (FDR) adjustment within population, the genome-wide significance level was equal to approximately 0.006 and 0.008 for dairy and beef cattle populations, respectively. This methodology has been used to identify regions displaying strong footprints of selection in humans [15], cattle [31, 36] and other animal populations [18].

2.7. Cross-population test

To compare EHH profiles between the two populations, we used the XP-EHH statistics. The main idea of XP-EHH is to test if the genome site is homozygous in one population but polymorphic in another population through the comparison of EHH score of two populations on one core SNP [37]; ancestral information is not needed when performing this test. In our analysis, the genome of the beef population, as an observed population, was compared with the dairy population, as a reference population. The XP-EHH score was computed for each autosomal SNP using the same R package used to compute iHS, rehh 2.0 [15].

The XP-EHH of a given focal SNP (XP-EHH) was defined and standardized, according to Sabeti et al. (2007) [14] and Gautier et al. (2017) [15] as

$$XP - EHH = \frac{\mathrm{LRiES} - med_{\mathrm{LRiES}}}{\sigma_{\mathrm{LRiES}}}$$

where LRiES = log($iEH_{population1}/iEH_{population2}$); iEH is the integrated allele-specific extended haplotype homozygosity; med_{LRiES} and σ_{LRiES} are the median and the standard deviation

of the LRiES, respectively. Populations 1 and 2 are considered as a reference population and an observed population, respectively. The XP-EHH is transformed into p_{XP-EHH} as shown by Gautier et al. (2011) [34]:

$$p_{XP-EHH} = -log_{10}(1 - 2|\Phi(XP - EHH) - 0.5|)$$

where $\Phi(x)$ represents the Gaussian cumulative distribution function. As p_{XP-EHH} may be interpreted as a two-sided *p*-value in a -log10 scale, regions with a *p*-value lower than 0.01 (1%) were considered signatures of selection in the test populations. Negative XP-EHH scores suggest selection happened in the reference population, otherwise happened in the observed population.

2.8. Candidate genes and functional analysis

A genomic region was considered as being under selection if it contained significant SNPs based on Fst or iHS values. Windows of 500 kb (250 kb upstream and 250 kb downstream of the significant SNP) were investigated to verify overlapping gene segments. Additionally, genes were compared with QTL regions previously identified and present in the Cattle QTL database (https://www.animalgenome.org/cgi-bin/QTLdb/BT/search). Gene annotation was performed using the UMD3.1 bovine genome assembly from the BioMart (www.ensembl.org/biomart) and NCBI (https://www.ncbi.nlm.nih.gov) databases. The orthologous genes from primates and sheep were used when the annotation information for bovine genes was not available.

A database for Annotation, Visualization, and Integrated Discovery (DAVID) v6.8 tool [38, 39] was used to identify significant ($p < 0.05$) Gene Ontology (GO) terms and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways using a list of genes with significant SNPs based on Fst and iHS values and the *Bos taurus* annotation file as a reference genome.

3. Results and Discussion

Eigenvalues of the PCA are shown in Fig 1a. The genetic structure of the dataset, i.e. the variance of the data, was captured mainly by the first PC. The eigenvalue for the first PC was 134.11, and the eigenvalue for the second PC was 39.16 (Fig 1a). Although

eigenvalues are absolute variances of the corresponding PC, it is common to express them as a percentage of the total variation in the data. In this case, percentages of variance explained by first and second PCs were 6.79 and 1.98, respectively. The percentage of the variance explained by PC was calculated as the eigenvalue times 100 divided by the sum of all the eigenvalues.

In Fig 1b, it is clear that both groups were completely separated according to the first principal component (PC1). The obtained results are clear and consistent with the history of both populations, which were under different breeding programs; therefore, were subjected to the intentional segregation of genes within each population, promoting the complete isolation and genetic variation between them. The dairy population presented a more scattered cluster based on the second component (PC2), indicating higher genetic distance and lower relatedness between individuals of the population, compared with the beef population cluster. Because animals from the dairy population were from commercial herds located in different regions of Brazil, it is possible that the PCA analysis clustered individuals of the dairy population according to their different demographic histories, providing a more scattered pattern. The proximity between animals from the beef population could be due to the fact that it was characterized as a closed herd in which only breeding animals of this herd were used in the mating. Mating of related animals favored the increase of inbreeding and reduced the genetic difference among individuals of this group.

Results of PCA were based on the clustering of individuals into geographically pre-defined groups. Discriminant Analysis of Principal Components based on the K-means algorithm, implemented in the R package adegenet [25] was further used to test the greatest number of clusters and to better evaluate the hypothesis on the two populations. An inspection of BIC values when the number of clusters (K) ranged from 1 to 100 showed that two clusters should be considered (Fig 1c); given this was the number with the lowest BIC. This test reflects the minimum number of clusters after which the BIC increases or decreases by a negligible amount [25]. The estimation of K by the DAPC has been markedly successful for the inference of the exact number of clusters in recent population genetic studies [40, 41]. The plots of the first discriminant function of DAPC

with different colors for different groups provided a visual assessment of between-population genetic structure (Fig 1d). In the present study, DAPC can also be considered well suited to describe the genetic diversity of the genotyped animals in the Brazilian Gir population.

According to PCA and DAPC analyses, there was a clear genetic difference between beef and dairy populations, and this division reflects the variation existent between the populations. Because of that, the next step was to compute genetic statistics taking into account the stratification of the dataset into beef and dairy populations. Average heterozygosity by chromosome was different between the two populations (Fig 2). The greatest discrepancy between means was observed on chromosome 23 (0.06). The allele frequency-dependent diversity estimate, such as observed heterozygosity (H_o), is a measure of genetic variation and can be very useful in comparison between populations [10]. The dairy population displayed higher heterozygosity levels compared with the beef population, which most likely reflects relative levels of genetic diversity. The highest heterozygosity level was observed on chromosome 17 (0.38), and the exception was for chromosome 20, where the beef population displayed the higher mean. Fis was slightly negative for dairy (-0.008) and beef (-0.0322) populations. Negative values of Fis could indicate an excess of heterozygosity beyond that expected under Hardy-Weinberg equilibrium within those populations. Negative values of Fis in Gir and other cattle breeds have been reported in the literature [6, 42].

Genetic differentiation varied throughout the genome (Fig 3). The average value for Fst was 0.033. Although the two populations in the present study were selected for different purposes, some levels of proximity between them were expected and were indicated by the low Fst values. In addition, the average Fst value was close to that observed in different cattle breeds [9, 30].

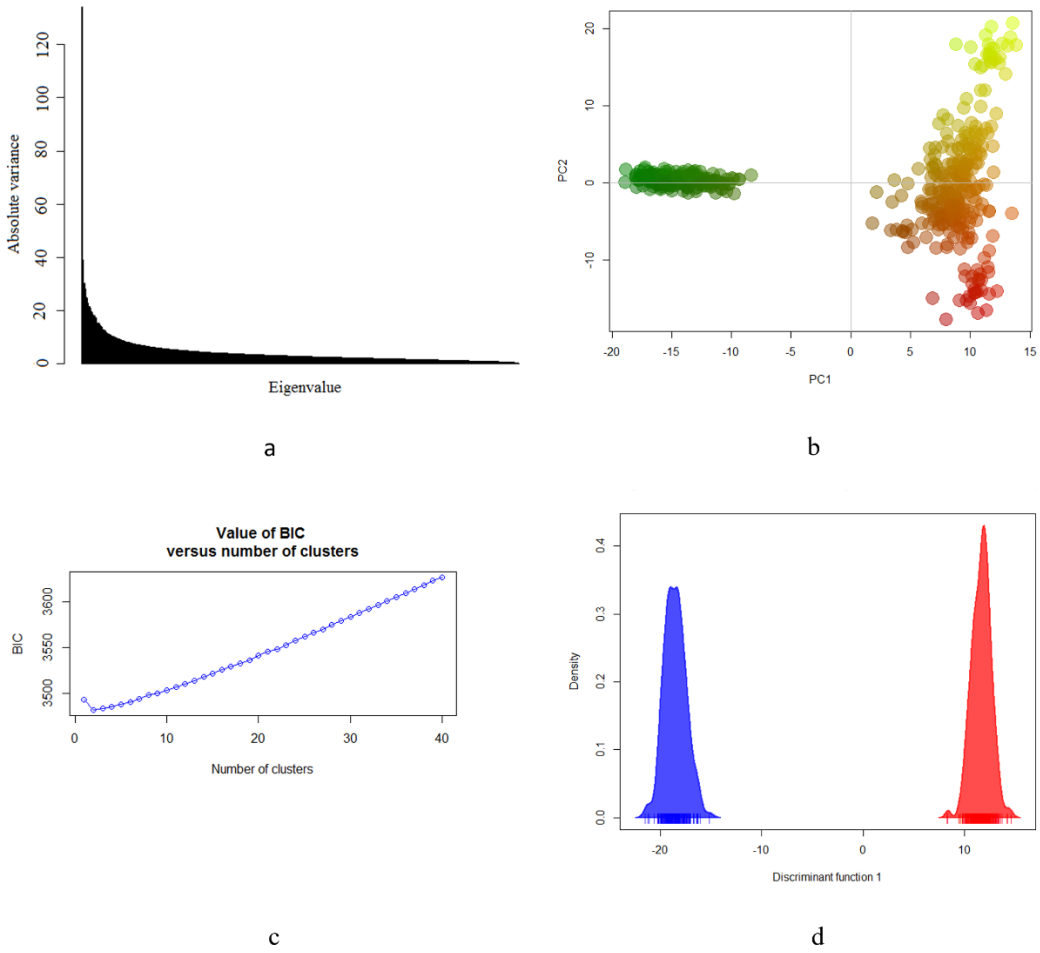


Fig 1. Genetic structure of the Brazilian Gir population.

(a) Eigenvalues of principal component analysis (PCA). (b) PCA scatter-plots of the first two principal components (PC) showing clearly separation between beef population (green color) and dairy population (red-yellow colors). (c) Inference of the number of clusters in Gir cattle based on K-means algorithm. (d) Plots of the first two discriminant functions of discriminant analysis of PC algorithm.

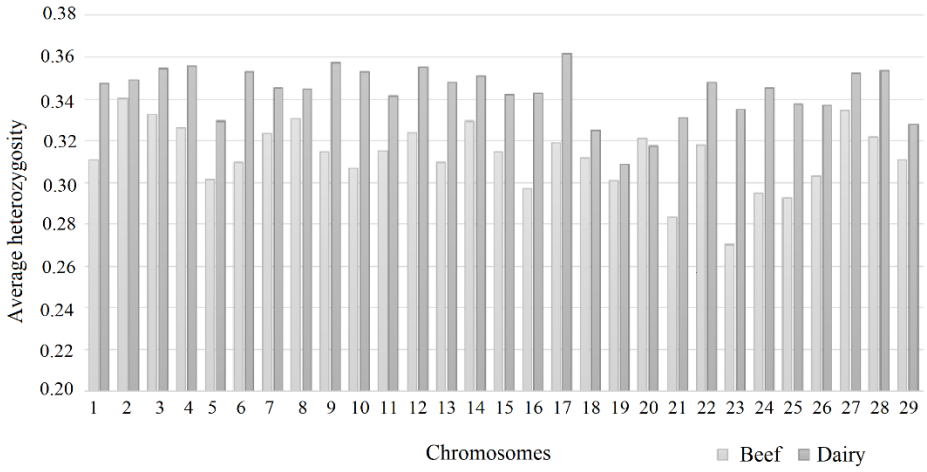


Fig 2. Average heterozygosity per chromosome in two populations. Light grey = beef cattle; dark grey = dairy cattle.

The level of genetic differentiation between populations is expected to be low in neutral regions of the genome or in regions of balanced selection, and divergent in regions subject to directional selection. When a Fst value is zero, there is no genetic differentiation between the populations under comparison. In the present study, low Fst values were possibly identified because the two populations belonged to the same breed and were originated from the same recent ancestral population; therefore, many alleles were expected to be commonly fixed in both populations. Gir is originally a dual-purpose breed. During the 1960s, part of the Brazilian breeders were artificially selecting animals for meat and milk simultaneously [1], which could explain the expected levels of proximity and low Fst values for the two populations. These populations only started being selected for different purposes in the last decades. Under artificial selection, allele frequencies can shift over time in the direction of the desired phenotypes producing signatures of selection.

The Fst approach enables the detection of selection signatures based on differences in allele frequencies across populations. This method revealed 488 SNPs as relevant loci for divergent selection among the total number of loci evaluated. Although being from the same breed, some changes in allele frequencies were observed between the two populations (Fig. 4). This change in allele frequencies and the genetic diversity observed were possibly caused by recent selection for different criteria and other non-

reported genetic events that occurred in the past. These findings are supported by the population structure results. The higher allele frequencies of these SNPs are representative of differences in selection, neutrality or other processes used in breeding programs [40].

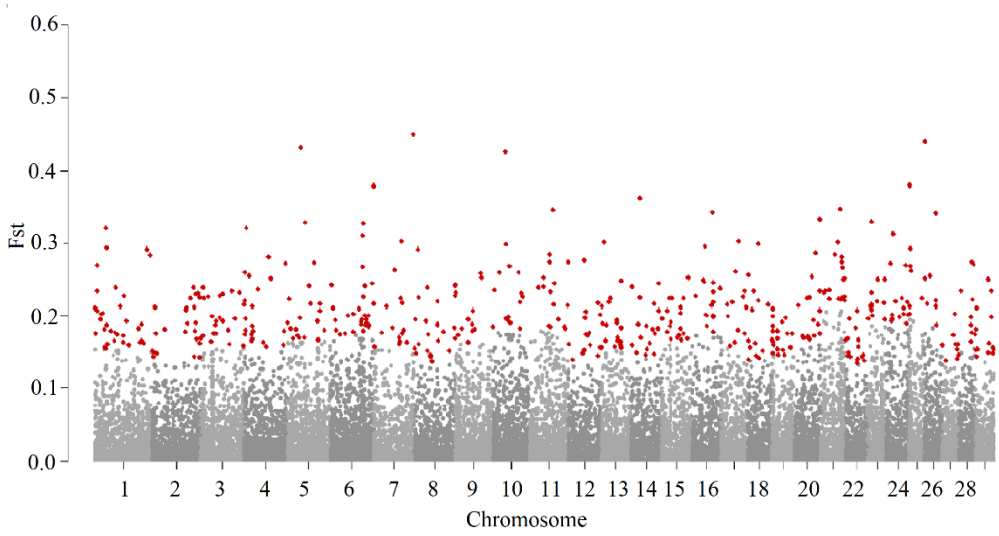


Fig 3. Genomic distribution of Fst values. SNPs were plotted relative to their physical positions within each autosome. The cutoff to call SNP outliers was defined as three standard deviations above the mean for each autosome. Red dots are SNPs with Fst beyond the cutoff value.

The SNPs identified as outliers based on the Fst method can be strong evidence of signatures of selection. Four SNPs had Fst values higher than 0.4; they were located on chromosomes 5, 7, 10, and 26. In total, 19 SNPs had Fst values ranging from 0.3 to 0.4. The greatest number of outliers was obtained on chromosome 1 (n = 33, with Fst values ranging from 0.15 to 0.32) and the lowest was obtained on chromosome 25 (n=7, with Fst values ranging from 0.20 to 0.38) (S1 Fig).

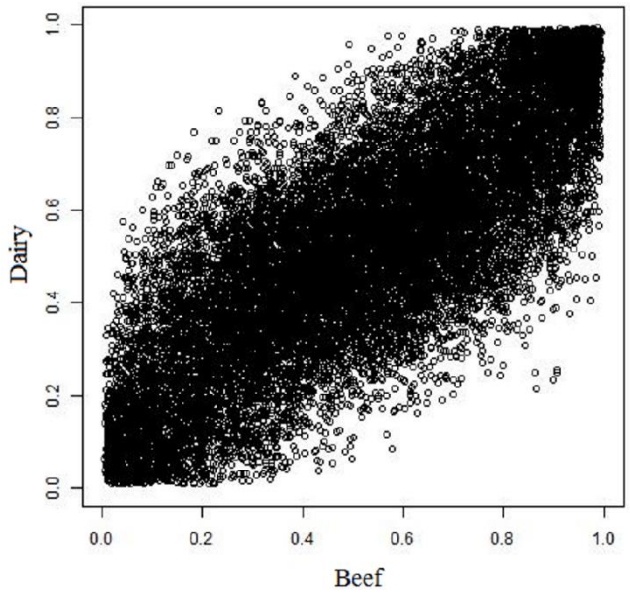


Fig 4. Scatter plots for population-specific allele frequency (dairy x beef). The data were displayed as a collection of points; each point represents an SNP having the allele frequency for the dairy population determining the position on the vertical axis and the allele frequency for the beef population determining the position on the horizontal axis. It reveals a positive linear relationship between allele frequencies of two populations.

Considering the threshold of three standard deviations to report an outlier Fst, the SNPs elected as outliers had Fst values higher than 99.8% of the total SNPs in the chromosome, which identified a variation pattern within the chromosome. Cadzow et al. (2014) [20] acknowledged that distinct methods use different patterns of genetic variation to identify evidence of selection. In addition, they highlighted the importance of utilizing multiple methodologies for investigating selection among populations. Thus, the iHS test was used as a complementary approach to identify regions that exhibit evidence of selection. This is a linkage disequilibrium-based method that provides increased power for assessment of signatures of selection within the population using high-throughput molecular information (e.g., SNP arrays).

After adjustment for FDR within the population, SNPs displaying piHS values greater than or equal to 2.10 (which approximately corresponded to p-value <0.008) were considered significant for the beef population. For the dairy population, the cutoff value

was 2.20 (which approximately corresponded to p-value <0.006). The "fdrtool" package allows the computation of local FDR values from p-values while taking into account an empirical null model [35]. The application of the FDR adjustment within each population avoids the detection of false positive selection signatures [31]. The chromosome-wide scans of iHS for beef and dairy populations are shown in Fig 5. The plots show clear evidence of selective forces in different regions of the genome. A total of 82 and 129 SNPs harbored signatures of selection in beef and dairy populations, respectively. The most significant SNP mapped to chromosome 6 (95,171,308 bp) for the beef population (iHS = -4.22) and to chromosome 16 (25,933,132 bp) for the dairy population (iHS = -4.43).

To further identify regions displaying strong signatures of selection, 500 kb windows containing the significant SNPs were investigated. A total of 23 and 43 candidate genes overlapped with significant windows for beef and dairy populations, respectively. The description of the top ten most significant iHS genomic regions per population is shown in Table 1. The *SLC24A4* (solute carrier family 24 member 4) gene located on chromosome 21 was common in the two populations, which can be clearly confirmed by the iHS plots (Fig. 5). It has been shown that the *SLC24A4* gene belongs to a group of the potassium-dependent sodium/calcium exchanger proteins which has been associated with hair, skin and eye pigmentation in humans [43, 44]. Nayeri et al. (2016) [45] conducted a genome-wide association study (GWAS) and identified the *SLC24A4* gene in a significant region associated with success in fertility traits such as calving interval and days open in female Canadian dairy Holstein cattle.

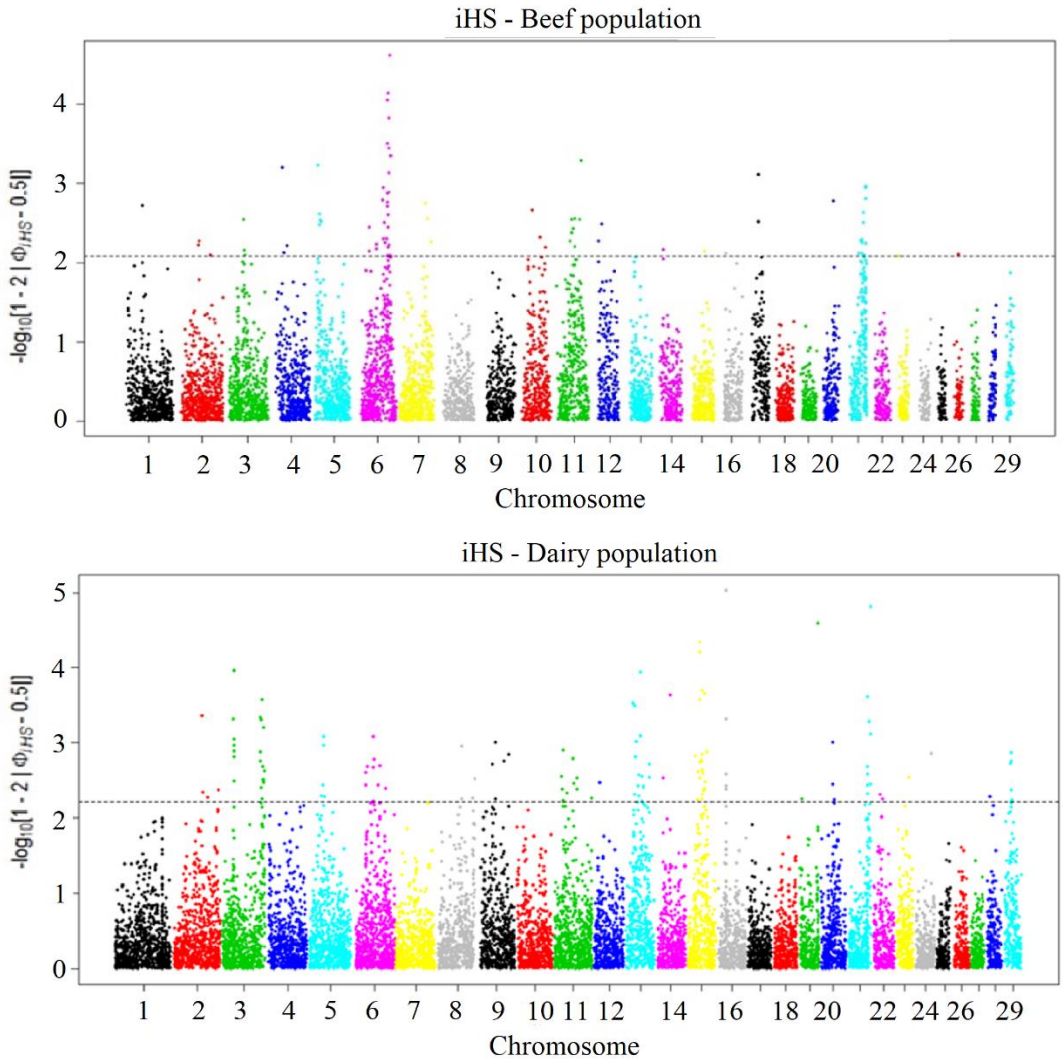


Fig 5. Genome-wide distribution of |iHS| values for Gir cattle. Upper = beef population; bottom = dairy population.

The description of the top ten most significant XP-EHH genomic regions per population is shown in Table 2. The most significant SNP was mapped to chromosome 22 (58,886,462 bp) for the beef population (XP-EHH = 2.44) and to chromosome 21 (57,721,222 bp) for the dairy population (XP-EHH = -7.02). We detected 28 SNPs as candidates of selection footprints in the beef population when the dairy population was used as a reference in the XP-EHH test, but only 10 out of 28 SNPs were mapped in gene regions. For the beef population, seven genes (*DNAH7, FBLN2, HECW2, SLC8A2,*

TRRAP, UGT1A1, and WNT7A) were identified based on the XP-EHH statistics. Two SNPs were located in the region of the *UGT1A1* gene*,* and two SNPs were located in the region of the *WNT7A* gene*.* The *UGT1A1* gene was located on chromosome 3 and encodes a critical enzyme that transforms small lipophilic molecules, such as steroids, hormones, and drugs, into excretable metabolites maintaining homeostasis [46]. The *WNT7A* gene regulates several cells and developmental pathways that affect the development of female reproductive tract and maintain uterine function in adults [47]. The *WNT7A* gene enhances muscle regeneration, increases the satellite stem cells and stimulates myogenic stem cell motility [48, 49]. Xue et al. (2013) [50] identified polymorphisms in the WNT7A gene associated with growth traits in Chinese Qinchuan cattle.

For the dairy population, 655 negative values presented p-values higher than 0.01, but only 197 out of 655 outliers were mapped to gene regions. When searching for genes, only 94 were identified for the dairy population. The *SLC24A4, DUSP10*, *CTNNA2, ADAMTS3,* and *TTC12* genes were located in regions with the highest number of SNPs identified per window, being 29, 12, 9, 7, and 7 SNP per region, respectively. The *SLC24A4* were also detected with iHS for the beef and dairy population, and the *DUSP10* gene was also detected with iHS approach. The *CTNNA2* gene was detected with the Fst, iHS and XP-EHH approaches.

The *DUSP10* gene located on chromosome 16 negatively regulates the activation of mitogen-activated protein (MAP) kinases and has a principal function in both innate and adaptive immune responses as an inhibitor of inflammation [51]. This gene has been described playing an important role in regulating the balance of energy thought the control of brown adipocyte differentiation [52]. Huang et al. (2017) [53] comparatively analyzed the transcriptome of subcutaneous adipose tissue between Wagyu and Holstein breeds with difference in fat deposition and identified the *DUSP10* gene up-regulated in Wagyu, which could be a key gene associated with fat metabolism and adipogenesis. The *CTNNA2* (catenin alpha 2) gene codifies a protein that plays an important role in the catabolism of collagen, cell adhesion and myogenesis [54]. This gene was identified in a QTL region associated with Lipomatous Myopathy in Piedmontese beef cattle, a disease

characterized by degeneration and infiltration of the muscular tissue characterized by replacement of myofibers with adipose tissue [55].

The *ADAMTS3* gene was mapped on chromosome 6 and encodes a member of the procollagen aminopropeptidase subfamily of proteins that play a role in the processing of type II fibrillar collagen in articular cartilage which is crucial for embryonic development and regulates placental angiogenesis [56]. Mészáros et al. (2014) [57] performed a genome-wide association study in Fleckvieh bulls and identified the *ADAMTS3* gene associated with longevity.

In total, 157 out of 488 significant Fst values were located in 151 regions. Three SNPs with significant Fst values overlapped with the *FCHSD2* (FCH and double SH3 domains 2) gene, which is located on chromosome 15 and regulates cell growth, migration and adhesion [58]. Two Fst peaks were within the genomic regions of *NSG1* (neuronal vesicle trafficking associated 1), *RALGPS1* (Ral GEF with PH domain and SH3 binding motif 1), and *HS3ST5* (heparan sulfate-glucosamine 3-sulfotransferase 5) genes, which are located on chromosomes 6, 11 and 9, respectively.

The *NSG1* gene encodes a small transmembrane protein highly expressed specifically in neurons which plays a critical role in the trafficking and polarization of several proteins [59-61]. Lee et al. (2016) [62] reported the *NSG1* gene as a region containing highly selective SNPs, i.e. SNPs that have higher probability of being selected in the next generation for milk production traits (milk yield and fat and protein contents) in Holsteins.

The *HS3ST5* gene encodes a member of the heparan sulfate 3-O-sulfotransferases protein family that catalyzes the biosynthesis of heparan sulfate, which regulates blood coagulation [63]. This gene has been associated with reproductive seasonality in sheep (total days of anoestrus and progesterone cycling months) [64].

Three genes (*CTNNA2, SLC24A4*, and T*MEM117*) were detected based on more than one approach. In the dairy population, three genes (*CNTN3, STK33*, and *TMEM117*) were detected with Fst and iHS, four genes (*DUSP10, NCAM1, SLC24A4,* and *TMEM117*) were detected with iHS and XP-EHH, and thirteen genes (*AFP, ANKRD17, CSNK1G1, C4orf22, CTNNA2, GALNT14, MAP1A, TBC1D22B, TMEM117, TMPRSS11E, UNC79,*

VPS13C, and *ZFAND3*) were detected with Fst and XP-EHH methods. Eight genes (*ADAMTS3, BCL11A, CAPN13, CTNNA2, EPHA5, GC, SLC24A4*, and *SLC25A21*) were detected based on the iHS for the beef population and XP-EHH for the dairy population.

Table 1. Top ten significant iHS genomic regions harboring signatures of selection in beef and dairy Gir cattle.

Population	Position[a]	Peak position	iHS[b]	piHS[c]	Candidate gene (position)
Beef	Chr4:35784894-35785394	35785144	-3.42	3.20	*SEMA3D* (35563538-35793762)
	Chr5:4589482-14589982	14589732	3.44	3.23	*SLC6A15* (14572186-14629483)
	Chr6:88541997-88542497	88542247	-3.92	4.05	*n.a*
	Chr6:88936051-88936551	88936301	-3.61	3.51	n.a.
	Chr6:90771212-90771712	90771462	3.97	4.14	*LOC101906736*[d] (90743814-90771489)
	Chr6:93023933-93024433	93024183	3.57	3.45	*CCDC158* (92975081-93058373)
	Chr6:93906353-93906853	93906603	3.79	3.82	n.a.
	Chr6:95171058-95171558	95171308	-4.22	4.62	n.a.
	Chr6:98735539-98736039	98735789	-3.51	3.34	n.a.
	Chr11:75228982-75229482	75229232	-3.47	3.29	*ATAD2B* (75141063-75263986)
Dairy	Chr3:33004500-33005000	33004750	3.86	3.95	n.a.
	Chr13:41168962-41169462	41169212	-3.85	3.93	n.a.
	Chr14:41835997-41836497	41836247	3.68	3.63	n.a.
	Chr15:37760348-37760848	37760598	4.00	4.20	*LOC104974236*[d] (37754671-37769288)
	Chr15:37871231-37871731	37871481	4.08	4.34	n.a.
	Chr15:44711168-44711668	44711418	3.72	3.70	*STK33* (44548054-44742390)
	Chr15:52440542-52441042	52440792	-3.69	3.65	n.a.
	Chr16:25932882-25933382	25933132	-4.43	5.03	*DUSP10* (25895748-25936856)
	Chr19:56552331-56552831	56552581	-4.21	4.59	*RECQL* (556549762-56580017)
	Chr21:66999575-67000075	66999825	4.32	4.81	*WDR25* (66917744-67066001)

n.a., not available.

[a] All positions are given in base pairs (bp) according to the Bovine UMD3.1 assembly.

[b] iHS is a value of integrated haplotype score.

[c] piHS is a value of transformed integrated haplotype score.

[d] non-coding RNA.

Table 2. Top ten significant XP-EHH genomic regions harboring signatures of selection in beef and dairy Gir cattle.

Population	Position[a]	Peak position	XP-EHH[b]	pXP-EHH[c]	Candidate gene (position)
Beef	Chr5:62492901-62493401	62493151	2.20	1.56	n.a.
	Chr15:80280407-80280907	80280657	2.24	1.60	n.a.
	Chr18: 54862897- 54863397	54863147	2.33	1.71	*SLC8A2* (54853503-54881796)
	Chr18: 54955789- 54956289	54956039	2.19	1.55	n.a.
	Chr19:11998622-11999122	11998872	2.22	1.58	n.a.
	Chr19:12009008-12009508	12009258	2.29	1.66	n.a.
	Chr22:58836201-58836701	58836451	2.28	1.65	*WNT7A* (58809372-58873170)
	Chr22:58886212-58886712	58886462	2.44	1.83	n.a.
	Chr22:58896176- 58896676	58896426	2.22	1.58	n.a.
	Chr25:38067748- 38068248	38067998	2.17	1.53	n.a.
Dairy	Chr21:57531747-57532247	57531997	-6.66	10.56	n.a.
	Chr21:57532846-57533346	57533096	-6.72	10.75	*SLC24A4* (57596461-57783306)
	Chr21:57619270-57619770	57619520	-6.64	10.51	*SLC24A4* (57596461-57783306)
	Chr21:57619838-57620338	57620088	-6.67	10.61	*SLC24A4* (57596461-57783306)
	Chr21:57720972-57721472	57721222	-7.02	11.65	*SLC24A4* (57596461-57783306)
	Chr21:57723114-57723614	57723364	-6.90	11.29	*SLC24A4* (57596461-57783306)
	Chr21:57723895-57724395	57724145	-6.84	11.11	*SLC24A4* (57596461-57783306)
	Chr21:57726974-57727474	57727224	-6.79	10.96	*SLC24A4* (57596461-57783306)
	Chr21:57729925-57730425	57730175	-6.81	11.02	*SLC24A4* (57596461-57783306)
	Chr21:57730971-57731471	57731221	-6.65	10.54	*SLC24A4* (57596461-57783306)

n.a., not available.

[a] All positions are given in base pairs (bp) according to the Bovine UMD3.1 assembly.

[b] XP-EHH is a value of cross-population extend haplotype homozygosity score.

[c] pXP-EHH is a value of transformed cross-population extend haplotype homozygosity score.

The *CNTN3* (contactin 3) gene located on chromosome 22 has been described as a candidate gene for tenderness in Nelore beef cattle [65]. The *STK33* (serine/threonine kinase 33) gene located on chromosome 15 encodes a member of the calcium/calmodulin-dependent kinase family, which exhibits a non-ubiquitous and a low level of expression in most tissues [66]. The *TMEM117* (transmembrane protein 11) located on chromosome 5 codifies an uncharacterized multi-pass transmembrane protein. Veerkamp et al. (2012) [67] performed a GWAS for feed utilization complex in Holstein–Friesian dairy cows and identified an SNP in *TMEM117* gene with a large effect in body condition scores. In another study also with GWAS, Zhu et al. (2017) [68] identified this gene associated with saturated fatty acids composition in Simmental cattle. The *NCAM1* was mapped on chromosome 15 and plays important functional roles in cell migration and plasticity changes in the developing and adult nervous system [69], insulin signaling and adipocyte differentiation in mouse [70].

A total of 282 genes were detected under selection in the dairy and beef Gir populations based on Fst, iHS, and XP-EHH approaches (see S1 File for details). Chen et al. (2016) [9] did not find common candidate regions in Chinese Holstein and Simmental cattle populations by using Fst and XP-EHH methods. The authors attributed it to different features of these methods and recommended the integration of various methods to increase the detection sensitivity of signatures of selection. As Fst, iHS, and XP-EHH tests assume different methodologies, more confidence is provided when a common genomic region is obtained by various methods. Fst, iHS, and XP-EHH must be used as complementary approaches in the detection of signatures of selection.

We compared the gene list from each test with previous regions of selection signatures in Gir. Three significant SNPs harboring signatures of selection based on the iHS were identified for Gir dairy cattle by Utsunomiya et al. (2013) [7]; only one of the SNPs identified by the authors was located in the intergenic region of the *ST6GALNAC5* gene.

The *ST6GALNAC5* gene was not identified in our analysis. However, six genes identified in our study (*ARHGEF7, DNAH7, EPB41, RELL1*, *TRAPPC9* and *ZFHX4*) were located within the ROH islands reported by Peripolli et al. (2018) [8] in Brazilian Gir dairy

cattle. The ROH islands are genomic regions with high homozygosity around a selected locus, i.e., with reduced genetic diversity. These islands of the genome might harbor targets of positive selection.

Next, we compared our gene list with QTL regions previously reported in cattle and available for online search. In total, 35 genes with signatures of selection overlapped with QTL regions. Mapping *LCORL* and *NCAPG* genes to the cattle QTL database showed that both genes were associated with reproduction, growth, and meat and carcass traits. They are all located on chromosome 6. A number of QTL terms have been reported for cattle in the region of *NCAPG* and *LCORL* genes [71, 72]. The *LCORL* and *NCAPG* were identified based on iHS test in the dairy population.

Other genes with selection signatures that had overlapped with QTL terms were found being associated with production (*KCNIP4, ANXA4*, *FTO, EGFR,* and *PAK1*), reproduction (*VAV3, KCNIP4, FRAS1, C4orf22, TRAPPC9*, and *APBB1*), milk composition (*ROBO1, NRCAM, KCNIP4, DAAM1, TRAPPC9, NF2*, and *FTO*), meat and carcass (*FAM184B, NCAM1, NF2, CTSD, TOX*, and *VSTM2L*), health (*TMTC2, IL1RN, BFSP1, TOX*, *CNTN3*, and *TRAPPC9*), and body conformation (*GC, ANXA4*, and *NF2*). More details can be found in the supplementary S1 File. The above findings suggest that these genes might play an important role for traits that have been selected in both dairy and beef populations.

The dataset used for DAVID v.6.8 comprised 142, 23, 43, and 94 genes identified based on Fst, iHS beef, iHS dairy, and XP-EHH dairy tests, respectively. Regarding to the XP-EHH beef test analysis, only seven genes (*DNAH7, FBLN2, HECW2, SLC8A2, TRRAP, UGT1A1, and WNT7A*) were identified in those genomic regions, which is not suitable for a functional analysis.

The significant ($P<0.05$) GO terms (biological process, cellular component, and molecular function) and KEGG pathways identified with Fst, iHS dairy, and XP-EHH dairy tests are described in Table 3. No significant terms were detected to the iHS beef cattle analysis, probably due to the small number of genes on this dataset. Results from the functional enrichment analysis with the Fst test revealed a total of six KEGG pathways,

four GO cellular components, six GO biological processes, and one GO molecular function (Table 3).

The regulation of the actin cytoskeleton (bta04810) and tight junction (bta04530) are pathways from the cellular process class, related to cellular community and cell motility, respectively. The actin cytoskeleton is a dynamic skeletal cell structure that contains actin and associated proteins, which has been associated with important functions in several biological processes including early development of oocyte organization and maturation [73], tenderness of *Longissimus dorsi* muscle from Qinchuan cattle [74], *Longissimus dorsi* intramuscular adipose tissue in Hanwoo cattle [75], response to diet [76] and calf birth weight in Holstein cattle [77], and response to several pathogens such as *Trypanosoma congolense* [78], *Salmonella enterica* [79], and *Mycobacterium bovis* [80].

Tight junctions are essential for establishing a selectively permeable barrier to diffusion through the paracellular space between cells, which are composed of transmembrane proteins that are involved in junction assembly, barrier regulation, cell polarity, gene transcription, and among other pathways. Tight junctions are known to be related to milk mammary gland development and milk secretion, controlling the transport of lactose and K+ to the extracellular fluid, whereas Cl− and Na+ are transported into milk [81, 82]. Since decreasing in the tight junction permeability results in increasing milk secretion, the tight junction pathway (bta04530) has been reported involved in milk production and quality traits in dairy cattle [83-85].

The functional enrichment analysis with the XP-EHH dairy test results revealed one GO molecular function and three GO biological processes (Table 3). The GTPase activator activity (GO:0005096) is a molecular function related to GTPase-activating proteins (GAP), a family of regulatory proteins that can bind to activate G-proteins and stimulate its intrinsic GTPase activity. Regulation of G-proteins is important because these proteins are involved in important cellular processes and physiological functions [86]. Regarding to GO biological processes identified by the XP-EHH dairy test, all of them were involved in the embryonic development and specification of patterns of cell differentiation. The embryonic limb morphogenesis (GO:0030326) is a biological process

that occur in the embryo by which the anatomical structures of the limp are generated and organized; the anterior/posterior pattern specification (GO:0009952) is defined as a regionalization process in which specific areas of cell differentiation are determined along the anterior-posterior axis (line that runs from the head to the tail of an organism), while the collagen biosynthetic process (GO:0032964) is related to chemical reactions and pathways resulting in the formation of collagen that is a group of proteins that form the main component of connective tissue in animals.Considering the intrinsic features of each test, genes identified by the iHS approach are potentially involved in recent selection pressures. Conversely, genes identified by the Fst method are likely related with events occurring further in the past. However, there is no guarantee that captured patterns of genotypic variation are a result of selection alone. It can be a result of other unrelated ancestral events [20] or even it may represent false positive results. To control the false positive rate, we applied the FDR adjustment within each population as recommended in the literature [31]. For the Fst, we used a technique similar to a control chart, but in a simplistic way that was based on measures of central tendency and dispersion.

Even though we used a tool available in the fdrtool R package [35] to control for false positives in the iHS test, it is difficult to completely rule out potential false positives. Teshima et al. (2006) [87] advised that error rates can be further decreased by combining several statistics. Even when a stringent cutoff of 1% for significance is defined, a fraction of loci identified as targets of selection may, in fact, be false discoveries [87]. In this way, it is possible that the target of selection that stood out in the Fst, iHS, and XP-EHH methods may contain a fraction of false discoveries. Genetic historical events different from artificial selection such as bottleneck, genetic drift, and migration could shape the population genomic variation [88, 89]. In addition to the effects of positive selection, genetic drift might be the other cause of genetic sweep altering the genetic structure of the two cattle populations. Further investigations need to be done to distinguish demographic history from selection in the two populations used in this study. Methods to compare models of positive selection relative to nonequilibrium models have been proposed by Jensen et al., 2007 [88] to accurately define the demographic events in a population.

Table 3. Gene Ontology terms and KEGG pathways enrichment analysis.

Test	Term	P	Genes
Fst	***KEGG Pathway***		
	bta04530:Tight junction	0.0150	*MAGI3, EPB41, ACTN1, AMOTL1, TJP2*
	bta04810:Regulation of actin cytoskeleton	0.0160	*EGFR, ARHGEF4, VAV3, ARHGEF7, DIAPH3, ACTN1*
	bta04970:Salivary secretion	0.0200	*KCNMA1, PLCB4, ADCY9, PRKG2*
	bta04540:Gap junction	0.0240	*EGFR, PLCB4, ADCY9, PRKG2*
	bta00240:Pyrimidine metabolism	0.0332	*NME5, PRIM2, POLR3A, POLR2D*
	bta00230:Purine metabolism	0.0340	*NME5, ADCY9, PRIM2, POLR3A, POLR2D*
	Gene Ontology Cellular Component		
	0045211~postsynaptic membrane	0.0072	*KCNMA1, GRIK2, CLSTN1, NSG1, HTR3B*
	0030027~lamellipodium	0.0103	*CORO1C, NF2, ARHGEF7, ACTN1, AMOTL1*
	0005925~focal adhesion	0.0105	*CORO1C, EGFR, SYNE2, ARHGEF7, LPP, GIT2, ACTN1, SYNPO2*
	0016328~lateral plasma membrane	0.0287	*CORO1C, NSG1, ACTN1*
	Gene Ontology Biological Process		
	0019228~neuronal action potential	0.0151	*KCNMA1, GRIK2, SCN5A*
	0035023~regulation of Rho protein signal transduction	0.0224	*ARHGEF4, VAV3, ARHGEF7, ARHGEF10L*
	0030032~lamellipodium assembly	0.0240	*ARHGEF4, VAV3, ARHGEF7*
	0045184~establishment of protein localization	0.0240	*CORO1C, SMYD3, MCC*
	0030036~actin cytoskeleton organization	0.0306	*CORO1C, NF2, DIAPH3, DAAM1*
	0042384~cilium assembly	0.0486	*CEP295, NME5, TTBK2, AHI1*
	Gene Ontology Molecular Function		
	0050518~2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase activity	0.0147	*ISPD*
iHS dairy	***Gene Ontology Cellular Component***		
	0097431~mitotic spindle pole	0.0174	*NUMA1, EML1*

	Gene Ontology Biological Process		
	0007420~brain development	0.0201	*EML1, RAB18, SLC6A17*
	Gene Ontology Molecular Function		
	0043140~ATP-dependent 3'-5' DNA helicase activity	0.0114	*RECQL5, ASCC3*
	0016887~ATPase activity	0.0166	*ABCG5, MACF1, CLPB*
	0005524~ATP binding	0.0247	*STK33, ABCG5, RECQL5, ASCC3, CLPB, PAK1, CPS1*
XP-EHH dairy	***KEGG Pathway***		
	Gene Ontology Biological Process		
	GO:0030326~embryonic limb morphogenesis	0.0107	*FRAS1, PBX1, ALX1*
	GO:0032964~collagen biosynthetic process	0.0250	*TRAM2, ADAMTS3*
	GO:0009952~anterior/posterior pattern specification	0.04771	*GRSF1, PBX1, ALX1*
	Gene Ontology Molecular Function		
	GO:0005096~GTPase activator activity	0.103	*RASAL2, ELMOD3, RALGAPA1, TBC1D22B, ARHGAP25*

The number of animals used in the present study is not a major concern in the structure and signatures of selection analysis. Several studies have focused on reduced number of animals per population, for example, <150 animals [18, 90] and <550 animals [7, 36]. This issue can be solved by the choice of appropriate methods among the large diversity of existing methods in the literature. As we have seen, the approaches used in the present study are largely recommended and powerful for comparison and studying genotypic variation in different populations.

The use of non-related animals in population structure studies is difficult in livestock species, although the choice for non-related animals allows for high genetic differentiation. For a small livestock population, the choice for animals to be genotyped are usually made by their importance in the herd. It is possible that the genotyped animals are related because breeding programs use tools as non-random mating systems and

reproductive technologies, which can increase the inbreeding level in the population. In the case of Gir cattle, the level of inbreeding is high and the effective population size is small for the dairy [1] and the beef populations. We consider that each breeding program was well represented by the chosen animals. Further improvements for the present study can be probably achieved by using a greater variety of SNP in the genotyping panel.

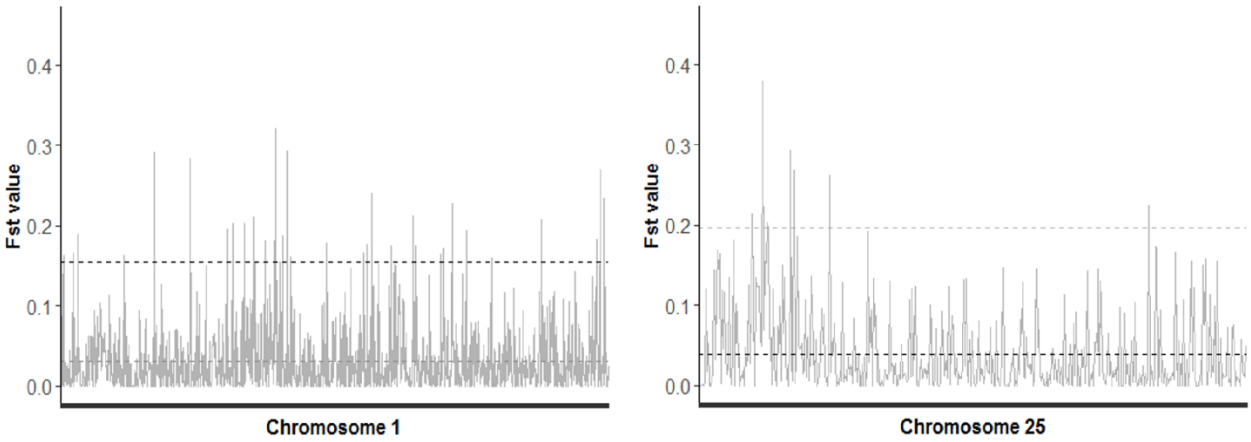
We found evidence of signatures of selection for two Gir cattle populations artificially selected for different purposes (e.g., meat or milk production). The detection of selection signatures used here can act as complementary information to current gene mapping approaches (GWAS). By comparing candidate gene regions found through the identification of selection signatures and GWAS, it is possible to test the contribution of genes under selection to phenotypes, which can subsequently be used in genomic selection [9]. The findings of the present study provided preliminary details about the recent adaptation of Brazilian Gir cattle. The loci that we identified as selection signatures provided information about genes and pathways in which the two Gir populations have adapted to the selective pressures. Strong selection on each population led to specialization of these populations. As we have seen in the investigation of genes and pathways, traces associated with fertility, milk production, beef quality, and growth were involved in this process. Furthermore, these selective signals indicated the presence of genetic variants that must affect complex phenotypes of particular interest for conservation and genetic improvement of the Brazilian Gir cattle. As a whole, the results found here give basic support for further investigations in the Gir breed.

4. Conclusion

High-throughput genomic information such as SNP markers can be successfully used to study the population structure and to identify genomic regions undergone divergent selection in Gir cattle. The difference in breeding history was able to imprint a degree of genomic differentiation for both beef and dairy populations. The patterns of genotypic variation in Gir cattle were consistent with the presence of selective pressures at some point in the history of the beef and dairy populations. These findings can provide complementary information on genomic regions of interest for functional genomic studies,

genome-wide associations, and the implementation of breeding schemes aiming genetic improvement and conservation of livestock populations.

5. Supporting information



S1 Fig. Distribution of Fst values for chromosomes 1 and 25.

S1 File. Intergenic regions harboring signatures of selection in the Brazilian Gir cattle.

Fst:

index	ensembl_gene_id	entrezgene	hgnc_symbol	chromosome_name	start_position	end_position	
1	ENSBTAG00000019404	767864	IL10RB	1	1563137	1591758	
2	ENSBTAG00000000597	282009	TMPRSS15	1	18058709	18207251	
3	ENSBTAG00000013662	538564	COL8A1	1	43541936	43717619	
4	ENSBTAG00000044001	524132	LPP	1	79041610	79725431	
5	ENSBTAG00000013219	531027	PEX5L	1	87792267	87930510	
6	ENSBTAG00000023426	616795	FAIM	1	131548459	131564836	
7	ENSBTAG00000015834	505852	ARHGEF4	2	1609041	1792901	
8	ENSBTAG00000010225	617422	POLR2D	2	4554334	4565801	
9	ENSBTAG00000005180	526135	RBMS1	2	35960130	36178719	
10	ENSBTAG00000012647	522419	ERBB4	2	99660620	99660620	
11	ENSBTAG00000018199	541221	ZBTB8A	2	121847012	121901991	
12	ENSBTAG00000006667	281753	EPB41	2	125014818	125113212	
13	ENSBTAG00000015240	514561	UBR4	2	134039040	134176134	
14	ENSBTAG00000000684	529043	ARHGEF10L	2	135557345	135672091	
15	ENSBTAG00000010158	539686	NOS1AP	3	7105925	7463616	

16	ENSBTAG00000037565	788142	LOC788142	3	20871006	20901647
17	ENSBTAG00000019611	532952	MAGI3	3	29849955	30117829
18	ENSBTAG00000031575	521961	VAV3	3	35428177	35861571
19	ENSBTAG00000020958	539204	LRRC8D	3	53475696	53606716
20	ENSBTAG00000014482	531770	FAF1	3	95937279	96434760
21	ENSBTAG00000005499	654401	ZPBP	4	5727483	5820313
22	ENSBTAG00000044068	784924	VWC2	4	5908623	6036680
23	ENSBTAG00000015073	104968457	ISPD	4	24492021	24886037
24	ENSBTAG00000011336	531410	OTOGL	5	9787259	9960051
25	ENSBTAG00000001810	538632	SCAF11	5	34430403	34480031
26	ENSBTAG00000032148	614719	TMEM117	5	36197272	36807308
27	ENSBTAG00000011352	541051	TBC1D30	5	49181675	49265076
28	ENSBTAG00000012794	510583	PAH	5	66950501	67040705
29	ENSBTAG00000011395	540892	DNM1L	5	77514854	77575339
30	ENSBTAG00000006434	540340	SYNPO2	6	7388728	7590933
31	ENSBTAG00000010564	533333	ELOVL6	6	16376642	16510240
32	ENSBTAG00000038520	100335739	TMPRSS11E	6	85614106	85716159
33	ENSBTAG00000048013	100138004	LOC100138004	6	85926576	85948595
34	ENSBTAG00000004912	508405	ANKRD17	6	89935677	90102798
35	ENSBTAG00000017131	506011	AFP	6	90258976	90280522
36	ENSBTAG00000019478	515375	NAAA	6	92508029	92531911
37	ENSBTAG00000035776	615662	C4orf22	6	96794894	97503220
38	ENSBTAG00000002978	533330	PRKG2	6	97652569	97735626
39	ENSBTAG00000005711	523110	NSG1	6	106483716	107356158
40	ENSBTAG00000017296	523362	POLN	6	108563826	108711716
41	ENSBTAG00000043368	#N/A	RF00026	7	1748890	1748996
42	ENSBTAG00000031231	789216	IRF1	7	23235653	23243697
43	ENSBTAG00000008752	785439	NME5	7	51153752	51182289
44	ENSBTAG00000020491	538232	SOX30	7	71295642	71337207
45	ENSBTAG00000035007	615819	GALNTL6	8	3816704	5330369
46	ENSBTAG00000043987	535671	KDM4C	8	37980606	38175047
47	ENSBTAG00000015607	541094	ERMP1	8	39132930	39188274
48	ENSBTAG00000011770	407101	TJP2	8	45598159	45696886
49	ENSBTAG00000009806	529698	INVS	8	65597445	65734342
50	ENSBTAG00000000188	540355	HS3ST5	9	36551274	36844165
51	ENSBTAG00000033153	615226	GRIK2	9	48476204	49218551
52	ENSBTAG00000017958	528398	AHI1	9	74330947	74544060
53	ENSBTAG00000017525	767903	ATG12	10	4619383	4799774
54	ENSBTAG00000006792	505206	EHD4	10	37412316	37496211
55	ENSBTAG00000006470	541215	TTBK2	10	38159317	38248606
56	ENSBTAG00000016823	527889	CSNK1G1	10	45717384	45867730
57	ENSBTAG00000038920	783566	VPS13C	10	48088919	48268869
58	ENSBTAG00000005984	513742	DAAM1	10	71730213	71916529
59	ENSBTAG00000025450	540504	SYNE2	10	76362023	76699213

60	ENSBTAG00000018255	524770	ACTN1	10	81023526	81121590
61	ENSBTAG00000031669	527492	CTNNA2	11	54723190	55906462
62	ENSBTAG00000021735	540500	GALNT14	11	68745032	68963585
63	ENSBTAG00000023843	616166	RALGPS1	11	97674642	97975506
64	ENSBTAG00000004663	511988	TRUB2	11	98919236	98930920
65	ENSBTAG00000012443	525628	DIAPH3	12	2130595	2695362
66	ENSBTAG00000014752	533105	AKAP11	12	12459630	12495973
67	ENSBTAG00000020726	524352	ARHGEF7	12	89451354	89552212
68	ENSBTAG00000013116	281985	PLCB4	13	2211089	2410692
69	ENSBTAG00000005915	528635	SFMBT2	13	16366946	16640879
70	ENSBTAG00000014194	524061	ADARB2	13	46383731	46630127
71	ENSBTAG00000006531	535136	DIP2C	13	47041229	47177126
72	ENSBTAG00000013955	533451	TRAPPC9	14	4229512	4616552
73	ENSBTAG00000004954	525888	TOX	14	26631190	26941726
74	ENSBTAG00000033268	539762	ZFHX4	14	41988047	42189775
75	ENSBTAG00000017165	781953	MATN2	14	68478180	68651684
76	ENSBTAG00000018945	541209	AMOTL1	15	15794576	15884804
77	ENSBTAG00000034193	514135	BCO2	15	22838240	22905944
78	ENSBTAG00000031330	541055	HTR3B	15	24774060	24815850
79	ENSBTAG00000000977	529873	CADM1	15	26101411	26451509
80	ENSBTAG00000011910	528309	STK33	15	44552624	44726438
81	ENSBTAG00000019062	539601	FCHSD2	15	53161555	53404313
82	ENSBTAG00000033180	616050	SMYD3	16	31589396	32333109
83	ENSBTAG00000033056	617120	PLD5	16	35407318	35635138
84	ENSBTAG00000011823	514447	CLSTN1	16	44633323	44671659
85	ENSBTAG00000013715	100335482	BRINP2	16	59954087	59977621
86	ENSBTAG00000000854	404548	SEC16B	16	60747108	60792164
87	ENSBTAG00000010306	532277	RXFP1	17	41318286	41446129
88	ENSBTAG00000046256	512602	TMEM132C	17	49445324	49739129
89	ENSBTAG00000006506	516048	GIT2	17	65556431	65596948
90	ENSBTAG00000007993	515798	CORO1C	17	66467521	66542216
91	ENSBTAG00000013153	540479	NF2	17	70895108	70966968
92	ENSBTAG00000015894	618792	WWOX	18	5283909	6196774
93	ENSBTAG00000012501	618622	FTO	18	22118201	22541532
94	ENSBTAG00000025200	617930	ASIC2	19	16353233	17562209
95	ENSBTAG00000020992	535623	RHBDL3	19	18160728	18192076
96	ENSBTAG00000019347	536989	PLXDC1	19	40226036	40293157
97	ENSBTAG00000018661	505575	TMC6	19	54647278	54666987
98	ENSBTAG00000001001	533461	ANKRD55	20	23007391	23117580
99	ENSBTAG00000023529	100337039	SPEF2	20	38404694	38606719
100	ENSBTAG00000002845	541259	CDH10	20	48874612	49060950
101	ENSBTAG00000012338	508724	FAH	21	26609988	26643266
102	ENSBTAG00000010963	506314	IL16	21	27527684	27606711
103	ENSBTAG00000017565	100140328	SCFD1	21	41692842	41804341

104	ENSBTAG00000014001	515593	MAP1A	21	55810061	55829462
105	ENSBTAG00000008017	537528	UNC79	21	58679832	58924124
106	ENSBTAG00000003526	520207	SYNE3	21	61855015	61963799
107	ENSBTAG00000020192	505757	PPP2R5C	21	68377171	68473921
108	ENSBTAG00000012143	530121	TECPR2	21	68837087	68901733
109	ENSBTAG00000002336	506525	RCOR1	21	68981058	69090482
110	ENSBTAG00000011628	407217	EGFR	22	892005	1069280
111	ENSBTAG00000003952	616212	FBXL2	22	7645493	7720326
112	ENSBTAG00000009155	282061	SCN5A	22	12015691	12118718
113	ENSBTAG00000007615	526697	CNTN3	22	27286727	27534171
114	ENSBTAG00000015556	512647	UBA3	22	32583221	32607984
115	ENSBTAG00000009541	537713	SUCLG2	22	33977312	34251728
116	ENSBTAG00000007937	514682	PRIM2	23	2475807	2809862
117	ENSBTAG00000017617	618875	C6orf222	23	10228126	10240589
118	ENSBTAG00000020595	514967	KCTD20	23	10355458	10391460
119	ENSBTAG00000014253	615878	TBC1D22B	23	11133036	11206519
120	ENSBTAG00000000527	532641	ZFAND3	23	11712802	12040961
121	ENSBTAG00000014063	525414	DNAH8	23	12550367	12883939
122	ENSBTAG00000019234	617566	BMP6	23	47355128	47510231
123	ENSBTAG00000008158	504721	CCDC68	24	54728527	54786445
124	ENSBTAG00000016663	525095	PIGN	24	60978529	61087325
125	ENSBTAG00000016629	525482	ADCY9	25	3242404	3344129
126	ENSBTAG00000001078	513912	SRL	25	3392668	3404630
127	ENSBTAG00000004038	280969	ABAT	25	7647412	7689263
128	ENSBTAG00000015272	524543	TSC22D4	25	36614550	36623767
129	ENSBTAG00000021195	100336715	A1CF	26	8523899	8580001
130	ENSBTAG00000012719	506002	UBTD1	26	18575002	18625622
131	ENSBTAG00000007709	511417	SHOC2	26	31762792	31849699
132	ENSBTAG00000009408	281599	ADAM2	27	34249449	34353369
133	ENSBTAG00000013300	282573	KCNMA1	28	32807351	33587986
134	ENSBTAG00000013259	540308	POLR3A	28	33875820	33924983
135	ENSBTAG00000001902	532848	CEP295	29	1067600	1121206
136	ENSBTAG00000004081	536281	FAT3	29	1965869	2605125
137	ENSBTAG00000019001	507760	PRDM10	29	36684172	36749050
138	ENSBTAG00000019173	337906	PAG12	29	38238677	38248855
139	ENSBTAG00000000292	513684	TCIRG1	29	46214074	46223038
140	ENSBTAG00000007622	282883	CTSD	29	50352064	50361497
141	ENSBTAG00000047713	100851924	LMNTD2	29	51036147	51077863
142	ENSBTAG00000017861	535868	BRSK2	29	51057469	51115452

iHS_dairy:

index	ensembl_gene_id	entrezgene	hgnc_symbol	chromosome_name	start_position	end_position
1	ENSBTAG00000016662	510506	CPS1	2	98758036	98898492
2	ENSBTAG00000005105	282299	PAFAH2	2	127684837	127720396
3	ENSBTAG00000001981	407174	SLC6A17	3	33337960	33373481
4	ENSBTAG00000021024	506730	MACF1	3	107180352	107389223
5	ENSBTAG00000047685	#N/A	KIAA0754	3	107242359	107246829
6	ENSBTAG00000005784	100139189	CSMD2	3	112558420	112898742
7	ENSBTAG00000014079	506913	GIGYF2	3	113064293	113194137
8	ENSBTAG00000016504	107131270	AGAP1	3	115843770	116226449
9	ENSBTAG00000011534	617017	RAMP1	3	117913716	117955245
10	ENSBTAG00000032148	614719	TMEM117	5	36197272	36807308
11	ENSBTAG00000021582	531234	NCAPG	6	38765969	38812051
12	ENSBTAG00000046561	540095	LCORL	6	38840894	38992112
13	ENSBTAG00000047743	614299	KCNIP4	6	41709362	43021626
14	ENSBTAG00000000841	539634	LZTS1	8	67861051	67867745
15	ENSBTAG00000004997	515749	CTNNAL1	8	100221980	100277722
16	ENSBTAG00000025782	574056	TNFSF8	8	105841516	105870631
17	ENSBTAG00000020482	538416	ASCC3	9	49714504	50103135
18	ENSBTAG00000016365	515536	ABCG5	11	26124239	26156411
19	ENSBTAG00000032519	786620	CAMKMT	11	26602922	27032485
20	ENSBTAG00000023216	783682	MALRD1	13	21049546	21386152
21	ENSBTAG00000023920	538123	GPR158	13	26229727	26526726
22	ENSBTAG00000010387	538509	MINDY3	13	30604489	30684245
23	ENSBTAG00000009871	511160	RAB18	13	37303894	37331999
24	ENSBTAG00000021867	281644	BFSP1	13	38182322	38220659
25	ENSBTAG00000004750	527903	VSTM2L	13	67460324	67498847
26	ENSBTAG00000005710	281941	NCAM1	15	24093189	24165578
27	ENSBTAG00000043981	510498	SBF2	15	42970987	43480821
28	ENSBTAG00000011910	528309	STK33	15	44552624	44726438
29	ENSBTAG00000015655	505096	APBB1	15	47109758	47267567
30	ENSBTAG00000018449	513091	NUMA1	15	52419502	52450409
31	ENSBTAG00000001280	520639	CLPB	15	52661473	52804186
32	ENSBTAG00000014847	507159	BBOX1	15	58532712	58609229
33	ENSBTAG00000001729	541175	DUSP10	16	25895751	25936864
34	ENSBTAG00000009968	508750	TBX4	19	11865527	11891936
35	ENSBTAG00000011715	512590	RECQL5	19	56550486	56580017
36	ENSBTAG00000006620	615739	SLC24A4	21	57596461	57783306
37	ENSBTAG00000013491	512904	EML1	21	66358950	66556801
38	ENSBTAG00000007233	527651	WDR25	21	66917780	67066000
39	ENSBTAG00000039855	536435	SUMF1	22	21994945	22088653
40	ENSBTAG00000007615	526697	CNTN3	22	27286727	27534171
41	ENSBTAG00000018901	615119	CTIF	24	48769621	49013492

42	ENSBTAG00000044197	518089	EDARADD	28	9128813	9195608
43	ENSBTAG00000010191	533729	PAK1	29	18684328	18780241

iHS_beef:

index	ensembl_gene_id	entrezgene	hgnc_symbol	chromosome_name	start_position	end_position
1	ENSBTAG00000008039	515030	ACMSD	2	62692769	62753692
2	ENSBTAG00000018739	529184	RPAP2	3	51195435	51281129
3	ENSBTAG00000024394	536417	SEMA3D	4	35563538	35789896
4	ENSBTAG00000006732	534500	NRCAM	4	49526574	49615802
5	ENSBTAG00000018432	281952	SLC6A15	5	14572186	14604741
6	ENSBTAG00000005932	523874	FAM184B	6	38614370	38672306
7	ENSBTAG00000004329	768210	RELL1	6	58599882	58660258
8	ENSBTAG00000009438	100337226	EPHA5	6	82560093	82962887
9	ENSBTAG00000013718	530076	GC	6	88695940	88739180
10	ENSBTAG00000006507	534082	ADAMTS3	6	89162542	89460195
11	ENSBTAG00000011935	520101	RASSF6	6	90373220	90400848
12	ENSBTAG00000019716	280828	CXCL8	6	90559882	90563647
13	ENSBTAG00000010954	407147	ART3	6	92607226	92755605
14	ENSBTAG00000012157	534614	CCDC158	6	92975363	93058338
15	ENSBTAG00000018585	540657	GABRB2	7	75103433	75393342
16	ENSBTAG00000021577	536671	ZFYVE16	7	82875305	82908670
17	ENSBTAG00000016534	538680	BCL11A	11	43071977	43174031
18	ENSBTAG00000031669	527492	CTNNA2	11	54723190	55906462
19	ENSBTAG00000008946	526038	CAPN13	11	69045268	69128762
20	ENSBTAG00000017255	518785	ATAD2B	11	75141459	75261008
21	ENSBTAG00000000125	100297992	SPATA5	17	34878951	35145717
22	ENSBTAG00000019350	513423	SLC25A21	21	47310562	47831598
23	ENSBTAG00000006620	615739	SLC24A4	21	57596461	57783306

XPEHH_dairy:

index	ensembl_gene_id	entrezgene	hgnc_symbol	chromosome_name	start_position	end_position
1	ENSBTAG00000009851	536815	ROBO1	1	25620170	26210494
2	ENSBTAG00000001686	536513	LRRC52	3	3425059	3446915
3	ENSBTAG00000013801	540688	PBX1	3	4228463	4527283
4	ENSBTAG00000010206	535325	UAP1	3	6900267	6932034
5	ENSBTAG00000002888	100337258	TMTC2	5	12460330	12563124
6	ENSBTAG00000014977	539689	ALX1	5	14991697	15013533
7	ENSBTAG00000026966	615588	RASSF9	5	15546053	15579641
8	ENSBTAG00000032148	614719	TMEM117	5	36197272	36807308
9	ENSBTAG00000002775	767878	C12orf71	5	82653406	82656449
10	ENSBTAG00000009438	100337226	EPHA5	6	82560093	82962887

11	ENSBTAG00000000962	507296	STAP1	6	85013014	85043206
12	ENSBTAG00000020298	517049	TMPRSS11A	6	85337482	85400112
13	ENSBTAG00000004412	785556	TMPRSS11F	6	85473854	85512994
14	ENSBTAG00000038520	100335739	TMPRSS11E	6	85614106	85716159
15	ENSBTAG00000001249	521920	SULT1B1	6	86917981	86942952
16	ENSBTAG00000018531	280821	JCHAIN	6	87759438	87768834
17	ENSBTAG00000008577	527169	GRSF1	6	87922395	87941062
18	ENSBTAG00000012397	530642	DCK	6	88049498	88077488
19	ENSBTAG00000002348	282360	SLC4A4	6	88182303	88541046
20	ENSBTAG00000013718	530076	GC	6	88695940	88739180
21	ENSBTAG00000006507	534082	ADAMTS3	6	89162542	89460195
22	ENSBTAG00000004912	508405	ANKRD17	6	89935677	90102798
23	ENSBTAG00000017131	506011	AFP	6	90258976	90280522
24	ENSBTAG00000010716	107131184	FRAS1	6	94711652	95055569
25	ENSBTAG00000010153	518050	ANXA3	6	95065256	95136945
26	ENSBTAG00000035776	615662	C4orf22	6	96794894	97503220
27	ENSBTAG00000003043	281203	GNG2	10	44711724	44842190
28	ENSBTAG00000016823	527889	CSNK1G1	10	45717384	45867730
29	ENSBTAG00000019390	614323	CA12	10	46749740	46768704
30	ENSBTAG00000003667	528252	TLN2	10	47303876	47796327
31	ENSBTAG00000038920	783566	VPS13C	10	48088919	48268869
32	ENSBTAG00000005141	100337503	PELI2	10	68778347	68974093
33	ENSBTAG00000016534	538680	BCL11A	11	43071977	43174031
34	ENSBTAG00000019665	281860	IL1RN	11	46699166	46706152
35	ENSBTAG00000006482	509502	PTCD3	11	48598061	48629215
36	ENSBTAG00000021935	535421	ELMOD3	11	49442120	49473288
37	ENSBTAG00000006075	509983	SUCLG1	11	50400445	50453412
38	ENSBTAG00000031669	527492	CTNNA2	11	54723190	55906462
39	ENSBTAG00000006027	534067	USP34	11	59791217	59987854
40	ENSBTAG00000004531	613336	CCT4	11	60397422	60410035
41	ENSBTAG00000018046	515585	B3GNT2	11	60654634	60687730
42	ENSBTAG00000018401	507989	APLF	11	66840544	66933742
43	ENSBTAG00000023976	534994	ARHGAP25	11	67066107	67158230
44	ENSBTAG00000012949	407211	GKN1	11	67283597	67289445
45	ENSBTAG00000007808	616010	ANTXR1	11	67334898	67590481
46	ENSBTAG00000009343	532546	AAK1	11	67812869	67964370
47	ENSBTAG00000001105	281625	ANXA4	11	68070487	68144492
48	ENSBTAG00000001042	514626	MXD1	11	68232862	68255606
49	ENSBTAG00000015865	515741	CAPN14	11	68678749	68715047
50	ENSBTAG00000021735	540500	GALNT14	11	68745032	68963585
51	ENSBTAG00000008946	526038	CAPN13	11	69045268	69128762
52	ENSBTAG00000005014	520912	LCLAT1	11	69204181	69321396
53	ENSBTAG00000007378	515213	CLIP4	11	70656465	70733268
54	ENSBTAG00000018193	519188	KCNK3	11	72831969	72868413

55	ENSBTAG00000005710	281941	NCAM1	15	24093189	24165578
56	ENSBTAG00000010008	515758	TTC12	15	24209210	24254092
57	ENSBTAG00000001729	541175	DUSP10	16	25895751	25936864
58	ENSBTAG00000015537	540692	RASAL2	16	61121078	61314123
59	ENSBTAG00000044109	100299729	ASIC5	17	44427939	44483562
60	ENSBTAG00000003840	282433	GUCY1B1	17	44498805	44569575
61	ENSBTAG00000014576	281216	GUCY1A1	17	44600509	44665571
62	ENSBTAG00000007988	519148	STX2	17	47324800	47352599
63	ENSBTAG00000009797	527119	RIMBP2	17	47630080	47712174
64	ENSBTAG00000044121	614641	NUBPL	21	42475716	42719201
65	ENSBTAG00000017719	537810	AKAP6	21	43132861	43630707
66	ENSBTAG00000001282	618573	RALGAPA1	21	46202963	46376749
67	ENSBTAG00000019350	513423	SLC25A21	21	47310562	47831598
68	ENSBTAG00000000655	528380	MIPOL1	21	47852815	48169316
69	ENSBTAG00000015095	517396	TTC6	21	48390858	48445207
70	ENSBTAG00000003684	539366	LRFN5	21	51772760	52095083
71	ENSBTAG00000039117	540222	TOGARAM1	21	55208034	55285668
72	ENSBTAG00000014476	512667	ADAL	21	55610774	55626668
73	ENSBTAG00000014478	512668	ZSCAN29	21	55630358	55639007
74	ENSBTAG00000014479	510865	TUBGCP4	21	55640209	55681852
75	ENSBTAG00000014001	515593	MAP1A	21	55810061	55829462
76	ENSBTAG00000000838	515039	CCDC88C	21	56629747	56773304
77	ENSBTAG00000018709	617124	CATSPERB	21	56887882	57008398
78	ENSBTAG00000003375	534419	TC2N	21	57075109	57117586
79	ENSBTAG00000026953	786759	TRIP11	21	57300674	57331304
80	ENSBTAG00000006620	615739	SLC24A4	21	57596461	57783306
81	ENSBTAG00000010416	539951	RIN3	21	57859148	57953844
82	ENSBTAG00000009845	518488	ITPK1	21	58182391	58347037
83	ENSBTAG00000008017	537528	UNC79	21	58679832	58924124
84	ENSBTAG00000039808	513746	SERPINA6	21	59517404	59530028
85	ENSBTAG00000014253	615878	TBC1D22B	23	11133036	11206519
86	ENSBTAG00000000527	532641	ZFAND3	23	11712802	12040961
87	ENSBTAG00000014063	525414	DNAH8	23	12550367	12883939
88	ENSBTAG00000009112	617287	TRAM2	23	24637969	24702830
89	ENSBTAG00000019300	511644	TPMT	23	39209390	39232415
90	ENSBTAG00000026290	534815	PIK3C3	24	14182757	14347226
91	ENSBTAG00000014284	510218	ALPK2	24	58119582	58248068
92	ENSBTAG00000027916	504994	MALT1	24	58282742	58346780
93	ENSBTAG00000013364	517290	LIPM	26	10504440	10540469
94	ENSBTAG00000013368	504991	ANKRD22	26	10542379	10570091

XPEHH_beef:

index	ensembl_gene_id	entrezgene	hgnc_symbol	chromosome_name	start_position	end_position
1	ENSBTAG00000016784	517781	DNAH7	2	84773343	85028058
2	ENSBTAG00000020696	531691	HECW2	2	85136811	85285250
3	ENSBTAG00000026181	751790	UGT1A1	3	113907720	114031371
4	ENSBTAG00000010856	535881	SLC8A2	18	54853503	54881796
5	ENSBTAG00000001668	533782	WNT7A	22	58809372	58873170
6	ENSBTAG00000004014	511854	FBLN2	22	58990282	59038424
7	ENSBTAG00000007113	507169	TRRAP	25	37807383	37895047

QTLdb:

ensembl_gene_id	entrezgene	hgnc_symbol	chromosome_name	start_position	end_position	identified by
ENSBTAG00000009851	536815	ROBO1	1	25620170	26210494	XP-EHH (dairy)
ENSBTAG00000031575	521961	VAV3	3	35428177	35861571	Fst
ENSBTAG00000006732	534500	NRCAM	4	49526574	49615802	iHS (beef)
ENSBTAG00000002888	100337258	TMTC2	5	12460330	12563124	XP-EHH (dairy)
ENSBTAG00000005932	523874	FAM184B	6	38614370	38672306	iHS (beef)
ENSBTAG00000021582	531234	NCAPG	6	38765969	38812051	iHS (dairy)
ENSBTAG00000046561	540095	LCORL	6	38840894	38992112	iHS (dairy)
ENSBTAG00000047743	614299	KCNIP4	6	41709362	43021626	iHS (dairy)
ENSBTAG00000013718	530076	GC	6	88695940	88739180	iHS (beef) and XP-EHH (dairy)
ENSBTAG00000011935	520101	RASSF6	6	90373220	90400848	iHS (beef)
ENSBTAG00000010716	107131184	FRAS1	6	94711652	95055569	XP-EHH (dairy)
ENSBTAG00000035776	615662	C4orf22	6	96794894	97503220	Fst and XP-EHH (dairy)
ENSBTAG00000005984	513742	DAAM1	10	71730213	71916529	Fst
ENSBTAG00000019665	281860	IL1RN	11	46699166	46706152	XP-EHH (dairy)
ENSBTAG00000001105	281625	ANXA4	11	68070487	68144492	XP-EHH (dairy)
ENSBTAG00000021867	281644	BFSP1	13	38182322	38220659	iHS (dairy)
ENSBTAG00000004750	527903	VSTM2L	13	67460324	67498847	iHS (dairy)
ENSBTAG00000013955	533451	TRAPPC9	14	4229512	4616552	Fst

ENSBTAG00000004954	525888	TOX	14	26631190	26941726	Fst
ENSBTAG00000033268	539762	ZFHX4	14	41988047	42189775	Fst
ENSBTAG00000017165	781953	MATN2	14	68478180	68651684	Fst
ENSBTAG00000005710	281941	NCAM1	15	24093189	24165578	iHS (dairy) and XP-EHH (dairy)
ENSBTAG00000015655	505096	APBB1	15	47109758	47267567	iHS (dairy)
ENSBTAG00000019062	539601	FCHSD2	15	53161555	53404313	Fst
ENSBTAG00000033180	616050	SMYD3	16	31589396	32333109	Fst
ENSBTAG00000000125	100297992	SPATA5	17	34878951	35145717	iHS (beef)
ENSBTAG00000013153	540479	NF2	17	70895108	70966968	Fst
ENSBTAG00000012501	618622	FTO	18	22118201	22541532	Fst
ENSBTAG00000025200	617930	ASIC2	19	16353233	17562209	Fst
ENSBTAG00000011628	407217	EGFR	22	892005	1069280	Fst
ENSBTAG00000007615	526697	CNTN3	22	27286727	27534171	Fst and iHS (dairy)
ENSBTAG00000016663	525095	PIGN	24	60978529	61087325	Fst
ENSBTAG00000044197	518089	EDARADD	28	9128813	9195608	iHS (dairy)
ENSBTAG00000010191	533729	PAK1	29	18684328	18780241	iHS (dairy)
ENSBTAG00000007622	282883	CTSD	29	50352064	50361497	Fst

6. References

1. SantanaJr ML, Pereira RJ, Bignardia AB, Faro L El, Tonhati H, Albuquerque LG. History, structure, and genetic diversity of Brazilian Gir cattle. Livest Sci. 2014;163: 26-33. doi: 10.1016/j.livsci.2014.02.007.

2. Liao X, Peng F, Forni S, McLaren D, Plastow G, Stothard P. Whole genome sequencing of Gir cattle for identifying polymorphisms and loci under selection. Genome. 2013;56: 592–598. doi: 10.1139/gen-2013-0082.

3. Stafuzza NB, Zerlotini A, Lobo FP, Yamagishi MEB, Chud TCS, Caetano AR, et al. Single nucleotide variants and InDels identified from whole-genome re-sequencing of Guzerat, Gyr, Girolando and Holstein cattle breeds. 2017 Mar 21. doi: 10.1371/journal.pone.0173954.

4. Bahbahani H, Tijjani A, Mukasa C, Wragg D, Almathen F, Nash O, et al. Signatures of Selection for Environmental Adaptation and Zebu × Taurine Hybrid Fitness in East African Shorthorn Zebu. Front Genet, 2017;8: 1-20. doi: 10.3389/fgene.2017.00068.

5. Sharma A, Lee S-H, Lim D, Chai HH, Choi BH, Cho Y. A genome-wide assessment of genetic diversity and population structure of Korean native cattle breeds. Genet. 2016;17: 139. doi: 10.1186/s12863-016-0444-8.

6. O'Brien AMP, Höller D, Boison SA, Milanesi M, Bomba L, Utsunomiya YT, et al. Low levels of taurine introgression in the current Brazilian Nelore and Gir indicine cattle populations. Genet Sel Evol. 2015;47: 31. doi: 10.1186/s12711-015-0109-5.

7. Utsunomiya YT, O'Brien AMP, Sonstegard TS, Van Tassell CP, Carmo AS, Mészáros G, et al. Detecting loci under recent positive selection in dairy and beef cattle by combining different genome-wide scan methods. PLoS One. 2013,8:e64280. doi: 10.1371/journal.pone.0064280

8. Peripolli E, Stafuzza NB, Munari DP, Lima ALF, Irgang R, Machado MA, et al. Assessment of runs of homozygosity islands and estimates of genomic inbreeding in Gyr (*Bos indicus*) dairy cattle. BMC Genomics. 2018;19: 34. doi: 10.1186/s12864-017-4365-3.

9. Chen M, Pan D, Ren H, Fu J, Li J, Su G, et al. Identification of selective sweeps reveals divergent selection between Chinese Holstein and Simmental cattle populations. Genet Sel Evol. 2016,48: 76. doi: 10.1186/s12711-016-0254-5.

10. Kijas JW, Lenstra JA, Hayes B, Boitard S, Neto LRP, Cristobal MS, et al. Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. PLoS Biol. 2012;10: 1-14.

11. Wright S. The genetical structure of populations. Ann Eugen. 1951;15: 323–354.

12. Pintus E, Sorbolini S, Albera A, Gaspa G, Dimauro C, Steri R, et al. Use of locally weighted scatterplot smoothing (LOWESS) regression to study selection signatures in Piedmontese and Italian Brown cattle breeds. Anim Genet. 2014;45(1): 1-11. doi: 10.1111/age.12076.

13. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol, 2006;4: 446-659. doi: 10.1371/journal.pbio.0040072.

14. Sabeti 2007 Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C., et al. Genome-wide detection and characterization of positive selection in human populations. Nature, 2007:449: 913–918.

15. Gautier M, Klassmann A, Vitalis R. rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. Mol Ecol Resour. 2017;17: 78–90. doi: 10.1111/1755-0998.12634.

16. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002;419: 832–837. doi:10.1038/nature01140.

17. He Q, Kim KW, Park YJ. Population genomics identifies the origin and signaturesof selection of Korean weedy rice. Plant Biotechnology Journal. 2017;15: 357–366. doi:10.1111/pbi.12630.

18. Zhu C, Fan H, Yuan Z, Hu S, Zhang L, Wei C, et al. Detection of selection signatures on the X Chromosome in three sheep breeds. Int J Mol Sci. 2015;16: 20360-20374. doi:10.3390/ijms160920360.

19. Msalya G, Kim E, Laisser ELK, Kipanyula MJ, Karimuribo ED, Kusiluka LJM, Chenyambuga SW, Rothschild MF. Determination of genetic structure and signatures of selection in three strains of Tanzania Shorthorn Zebu, Boran and Friesian Cattle by Genome-Wide SNP Analyses. PLoS One. 2017;1-18. doi:10.1371/journal.pone.0171088.

20. Cadzow M, Boocock J, Nguyen HT, Wilcox P, Merriman TR, Black MA. A bioinformatics workflow for detecting signatures of selection in genomic data. Front Genet. 2014;5: 293. doi: 10.3389/fgene.2014.00293.

21. Razook AG, Figueiredo LA, Bonilha Neto LM. Selection for yearling weight in Nelore and Guzerá zebu breeds: selection applied and response in 15 years of progeny. WORLD CONGRESS ON GENETICS APPLIED TO LIVESTOCK PRODUCTION. 1998.

22. Knackfuss FB, Razook AG, Mercadante MEZ, Cyrillo JNSG, Figueiredo LA, Tonhati H. Selection for growth traits in Gyr cattle. 2. Estimates of variances and genetic parameters due to direct and maternal effects. R Bras Zootec. 2006;35(3): 726-732. doi: 10.1590/S1516-35982006000300013.

23. Iso-Touru T, Tapio M, Vilkki J, Kiseleva T, Ammosov I, Ivanova Z, et al .Genetic diversity and genomic signatures of selection among cattle breeds from Siberia, eastern and northern Europe. Anim Genet. 2016;47: 647. doi: 10.1111/age.12473.

24. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. Am J Hum Genet. 2007;81: 1084-1097. doi:10.1086/521987.

25. Jombart T, Devillard S, Balloux F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genetics. 2010;11: 94. doi: 10.1186/1471-2156-11-94

26. Dunteman, GH. Principal Components Analysis. Quantitative Applications in the Social Sciences. Newbury: Sage Publications; 1989.

27. Jombart T, Pontier D, Dufour AB. Genetic markers in the playground of multivariate analysis. Heredity. 2009;102(4): 330-341.

28. Goudet J. HIERFSTAT, a package for R to compute and test hierarchical F – statistics. Mol Ecol Notes. 2005;5: 184–186. doi: 10.1111/j.1471-8278.2004.00828.x.

29. Nei M. Molecular Evolutionary Genetics. Reprint edition. New York: Columbia University Press; 1987.

30. Qanbari S, Gianola D, Hayes B, Schenkel F, Miller S, Moore S, et al. Application of site and haplotype-frequency based approaches for detecting selection signatures in cattle. BMC Genomics. 2011;12: 318.

31. Zhao F, McParland S, Kearney F, Du L, Berry DP. Detection of selection signatures in dairy and beef cattle using high-density genomic information. Genet Sel Evol. 2015;47: 49. doi: 10.1186/s12711-015-0127-3.

32. Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. Interrogating a high-density SNP map for signatures of natural selection. Genome Res. 2002;12: 1805–1814. doi: 10.1101/gr.631202.

33. Rocha D, Billerey C, Samson F, Boichard D, Boussaha M. Identification of the putative ancestral allele of bovine single-nucleotide polymorphisms. J Anim Breed Genet. 2014;131: 483–48.

34. Gautier M, Naves M. Footprints of selection in the ancestral admix-ture of a new world creole cattle breed. Mol Ecol. 2011;20: 3128–3143.

35. Strimmer, K. fdrtool: A unified approach to false discovery rate estimation. BMC Bioinformatics. 2008;9: 303. doi:10.1186/1471-2105-9-303.

36. Urbinati I, Stafuzza NB, Oliveira MT, Chud TCS, Higa RH, Regitano LCA, et al. Selection signatures in Canchim beef cattle. J Anim Sci Biotechnol. 2016;7: 29. doi: 10.1186/s40104-016-0089-5.

37. Ma Y, Zhang H, Zhang Q, Ding X. Identification of selection footprints on the X chromosome in pig. PLoS One. 2014;9(4): e94911. doi: 10.1371/journal.pone.0094911.

38. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009a;4: 44–57.

39. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009b;37: 1–13.

40. Wang M D, Dzama K, Hefer CA, Muchadeyi FC. Genomic population structure and prevalence of copy number variations in South African Nguni cattle. BMC Genomics. 2015;16:894.

41. Kukučková V, Moravčíková N, Ferenčaković M, Simčič M, Mészáros G, Sölkner J, et al. Genomic characterization of Pinzgau cattle: genetic conservation and breeding perspectives. Conserv. Genet. 2017;18: 893–910. doi: 10.1007/s10592-017-0935-9.

42. Ben Jemaa S, Boussaha M, Ben Mehdi M, Lee JH, Lee SH. Genome-wide insights into population structure and genetic history of tunisian local cattle using the illumina bovinesnp50 beadchip. BMC Genomics. 2015;16: 677. doi: 10.1186/s12864-015-1638-6.

43. Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Magnusson KP, et al. Genetic determinants of hair, eye and skin pigmentation in Europeans. Nat Genet. 2007;39(12): 1443-1452.

44. Han J, Kraft P, Nan H, Guo Q, Chen C, Qureshi A, et al. A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. PLoS Genet. 2008;4(5): e1000074. doi: 10.1371/journal.pgen.1000074.

45. Nayeri S, Sargolzaei M, Abo-Ismail MK, May N, Miller SP, Schenkel F, et al. Genome-wide association for milk production and female fertility traits in Canadian dairy Holstein cattle. BMC Genet. 2016;17(1): 75. doi: 10.1186/s12863-016-0386-1.

46. Radominska-Pandya A, Czernik PJ, Little JM, Battaglia E, Mackenzie PI. Structural and functional studies of UDP-glucuronosyltransferases. Drug Metab Rev. 1999;31(4): 817-99. doi: 10.1081/DMR-100101944.

47. Miller C, Sassoon DA. Wnt-7a maintains appropriate uterine patterning during the development of the mouse female reproductive tract. Development. 1998;125: 3201-3211.

48. Le Grand F, Jones AE, Seale V, Scime A, Rudnicki MA. Wnt7a activates the planar cell polarity pathway to drive the symmetric expansion of satellite stem cells. Cell Stem Cell. 2009;4: 535-547.

49. Bentzinger CF1, Von Maltzahn J, Dumont NA, Stark DA, Wang YX, Nhan K, Frenette J, Cornelison DD, Rudnicki MA. Wnt7a stimulates myogenic stem cell motility and engraftment resulting in improved muscle strength. J Cell Biol. 2014;205(1):97-111. doi: 10.1083/jcb.201310035.

50. Xue J, Sun Y, Guo W, Yang Z, Tian H, Zhang C, Lei C, Lan X, Chen H. Haplotypes and effects on growth traits of bovine Wnt7a gene in Chinese Qinchuan cattle. Gene. 2013;524(2):241-5. doi: 10.1016/j.gene.2013.04.013.

51. Zhang Y, Blattman JN, Kennedy NJ, Duong J, Nguyen T, Wang Y, Davis RJ, Greenberg PD, Flavell RA, Dong C. Regulation of innate and adaptive immune responses by MAP kinase phosphatase 5. Nature. 2004;430(7001):793-7.

52. Choi HR, Kim WK, Kim EY, Han BS, Min JK, Chi SW, Park SG, Bae KH, Lee SC. Dual-specificity phosphatase 10 controls brown adipocyte differentiation by modulating the phosphorylation of p38 mitogen-activated protein kinase. PLoS One. 2013;8(8):e72340. doi: 10.1371/journal.pone.0072340.

53. Huang W1, Guo Y1, Du W1, Zhang X1, Li A1, Miao X. Global transcriptome analysis identifies differentially expressed genes related to lipidmetabolism in Wagyu and Holstein cattle. Sci Rep. 2017;7(1):5278. doi: 10.1038/s41598-017-05702-5.

54. Malik A, Lee EJ, Jan AT, Ahmad S, Cho KH, Kim J, et al. Network Analysis for the Identification of Differentially Expressed Hub Genes Using Myogenin Knock-down Muscle Satellite Cells. PLoS One. 2015;10(7): e0133597. doi: 10.1371/journal.pone.0133597.

55. Peletto S, Strillacci MG, Capucchio MT, Biasibetti E, Modesto P, Acutis PL, et al. Genetic basis of Lipomatous Myopathy in Piedmontese beef cattle. Livest Sci. 2017;206: 9-16. doi: 10.1016/j.livsci.2017.09.027.

56. Janssen L, Dupont L, Bekhouche M, Noel A, Leduc C, Voz M, Peers B, Cataldo D, Apte SS, Dubail J, Colige A. ADAMTS3 activity is mandatory for embryonic lymphangiogenesis and regulates placental angiogenesis. Angiogenesis. 2016 ;19(1):53-65. doi: 10.1007/s10456-015-9488-z.

57. Mészáros G, Eaglen S, Waldmann P, Sölkner J. A Genome Wide Association Study for Longevity in Cattle. Open J Genet. 2014; 4: 46-55.

58. Liu S, Xiong X, Zhao X, Yang X, Wang H. F-BAR family proteins, emerging regulators for cell membrane dynamic changes-from structure to human diseases. J Hematol Oncol. 2015;8: 47. doi: 10.1186/s13045-015-0144-2.

59. Steiner P, Sarria JC, Glauser L, Magnin S, Catsicas S, Hirling H. Modulation of receptor cycling by neuron-enriched endosomal protein of 21 kD. J Cell Biol. 2002;157(7): 1197-209.

60. Yap CC, Wisco D, Kujala P, Lasiecka ZM, Cannon JT, Chang MC, et al. The somatodendritic endosomal regulator NEEP21 facilitates axonal targeting of L1/NgCAM. J Cell Biol. 2008;180(4): 827-842. doi: 10.1083/jcb.200707143.

61. Norstrom EM, Zhang C, Tanzi R, Sisodia SS. Identification of NEEP21 as a ß-amyloid precursor protein-interacting protein in vivo that modulates amyloidogenic processing in vitro. J Neurosci. 2010;17;30(46): 15677-15685. doi: 10.1523/JNEUROSCI.4464-10.2010.

62. Lee YS, Shin D, Lee W, Taye M, Cho K, Park KD, et al. The prediction of the expected current selection coefficient of single nucleotide polymorphism associated with Holstein milk yield, fat and protein contents. Asian-Australas J Anim Sci. 2016;29(1): 36-42. doi: 10.5713/ajas.15.0476.

63. Duncan MB, Chen J, Krise JP, Liu J. The biosynthesis of anticoagulant heparan sulfate by the heparan sulfate 3-O-sulfotransferase isoform 5. Biochim Biophys Acta. 2004;1671(1-3): 34-43.

64. Martinez-Royo A, Alabart JL, Sarto P, Serrano M, Lahoz B, Folch J, et al. Genome-wide association studies for reproductive seasonality traits in Rasa Aragonesa sheep breed. Theriogenology. 2017;99: 21-29. doi: 10.1016/j.theriogenology.2017.05.011.

65. Carvalho ME, Baldi FS, Santana MHA, Ventura RV, Oliveira GA, Bueno R. S, et al. Identification of genomic regions related to tenderness in Nellore beef cattle. Adv Anim Biosci. 2017;8: 42–44. doi: 10.1017/S2040470017001674.

66. Mujica AO, Brauksiepe B, Saaler-Reinhardt S, Reuss S, Schmidt ER. Differential expression pattern of the novel serine/threonine kinase, STK33, in mice and men. FEBS J. 2005;272(19): 4884-4898.

67. Veerkamp RF, Coffey M, Berry D, de Haas Y, Strandberg E, Bovenhuis H, et al. Genome-wide associations for feed utilisation complex in primiparous Holstein-Friesian dairy cows from experimental research herds in four European countries. Animal. 2012;6(11): 1738-1749.

68. Zhu B, Niu H, Zhang W, Wang Z, Liang Y, Guan L, et al. Genome wide association study and genomic prediction for fatty acid composition in Chinese Simmental beef cattle using high density SNP array. BMC Genomics. 2017;18(1): 464. doi: 10.1186/s12864-017-3847-7.

69. Kiss JZ, Troncoso E, Djebbara Z, Vutskits L, Muller D. The role of neural cell adhesion molecules in plasticity and repair. Brain Res Brain Res Rev. 2001;36(2-3): 175-84. doi: 10.1016/S0165-0173(01)00093-5.

70. Yang HJ, Xia YY, Wang L, Liu R, Goh KJ, Ju PJ, Feng ZW. A novel role for neural cell adhesion molecule in modulating insulin signaling and adipocyte differentiation of mouse mesenchymal stem cells. J Cell Sci. 2011;124: 2552-2560. doi: 10.1242/jcs.085340.

71. Lindholm-Perry AK, Sexten AK, Kuehn LA, Smith TP, King DA, Shackelford SD, et al. Association, effects and validation of polymorphisms within the NCAPG - LCORL locus located on BTA6 with feed intake, gain, meat and carcass traits in beef cattle. BMC Genet. 2011 Dec 14;12: 103. doi: 10.1186/1471-2156-12-103.

72. Setoguchi K, Watanabe T, Weikard R, Albrecht E, Kühn C, Kinoshita A, et al. The SNP c.1326T>G in the non-SMC condensin I complex, subunit G (NCAPG) gene encoding a p.Ile442Met variant is associated with an increase in body frame size at puberty in cattle. Anim Genet. 2011 Dec;42(6):650-5. doi: 10.1111/j.1365-2052.2011.02196.x.

73. Sun QY, Schatten H. Regulation of dynamic events by microfilaments during oocyte maturation and fertilization. Reproduction. 2006;131(2): 193-205.

74. Zhang Y, Zan L, Wang H. Screening candidate genes related to tenderness trait in Qinchuan cattle by genome array. Mol Biol Rep. 2011;38(3):2007-14. doi: 10.1007/s11033-010-0323-8.

75. Lee HJ, Park HS, Kim W, Yoon D, Seo S. Comparison of metabolic network between muscle and intramuscular adipose tissues in Hanwoo beef cattle using a systems biology approach. Int J Genomics. 2014;2014:1-6.

76. Kong RS, Liang G, Chen Y, Stothard P, Guan LL. Transcriptome profiling of the rumen epithelium of beef cattle differing in residual feed intake. BMC Genomics. 2016;17: 592. doi: 10.1186/s12864-016-2935-4.

77. Cole JB, Waurich B, Wensch-Dorendorf M, Bickhart DM, Swalve HH. A genome-wide association study of calf birth weight in Holstein cattle using single nucleotide polymorphisms and phenotypes predicted from auxiliary traits. J Dairy Sci. 2014;97(5): 3156-72. doi: 10.3168/jds.2013-7409.

78. Noyes H, Brass A, Obara I, Anderson S, Archibald AL, Bradley DG, et al. Genetic and expression analysis of cattle identifies candidate genes in pathways responding to Trypanosoma congolense infection. Proc Natl Acad Sci U S A. 2011;108(22): 9304-9. doi: 10.1073/pnas.1013486108.

79. Lawhon SD, Khare S, Rossetti CA, Everts RE, Galindo CL, Luciano SA, et al. Role of SPI-1 secreted effectors in acute bovine response to Salmonella enterica Serovar Typhimurium: a systems biology analysis approach. PLoS One. 2011;6(11): e26869. doi: 10.1371/journal.pone.0026869.

80. Blanco FC, Soria M, Bianco MV, Bigi F. Transcriptional response of peripheral blood mononuclear cells from cattle infected with Mycobacterium bovis. PLoS One. 2012;7(7): e41066.

81. Nguyen DA, Neville MC. Tight junction regulation in the mammary gland. J Mammary Gland Biol Neoplasia. 1998;3(3): 233-246.

82. Stelwagen K, Singh K. The role of tight junctions in mammary gland function. J Mammary Gland Biol Neoplasia. 2014r;19(1): 131-138. doi: 10.1007/s10911-013-9309-1.

83. Suchocki T, Wojdak-Maksymiec K, Szyda J. Using gene networks to identify genes and pathways involved in milk production traits in Polish Holstein dairy cattle. Czech J Anim Sci. 2016;61(11): 526–538. doi: 10.17221/43/2015-CJAS.

84. Dadousis C, Pegolo S, Rosa GJM, Bittante G, Cecchinato A. Genome-wide association and pathway-based analysis using latent variables related to milk protein composition and cheesemaking traits in dairy cattle. J Dairy Sci. 2017a;100(11): 9085-9102. doi: 10.3168/jds.2017-13219.

85. Dadousis C, Pegolo S, Rosa GJ, Gianola D, Bittante G, Cecchinato A. Pathway-based genome-wide association analysis of milk coagulation properties, curd firmness, cheese yield, and curd nutrient recovery in dairy cattle. J Dairy Sci. 2017b;100(2): 1223-1231. doi: 10.3168/jds.2016-11587.

86. Syrovatkina V, Alegre KO, Dey R, Huang XY. Regulation, Signaling, and Physiological Functions of G-Proteins. J Mol Biol. 2016;428(19):3850-3868. doi: 10.1016/j.jmb.2016.08.002.

87. Teshima KM, Coop G, Przeworski M. How reliable are empirical genomic scans for selective sweeps? Genome Res. 2006;16(6): 702–712. doi: 10.1101/gr.5105206.

88. Jensen JD, Thornton KR, Bustamante CD, Aquadro CF. On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in nonequilibrium populations. Genetics. 2007;176: 2371–2379. doi: 10.1534/genetics.106.069450.

89. Wang MS, Otecko NO, Wang S, Wu DD, Yang MM, Xu YL, et al. An evolutionary genomic perspective on the breeding of dwarf chickens. Mol Biol Evol. 2017;34(12): 3081-3088. doi: 0.1093/molbev/msx227.

90. Upadhyay MR, Chen W, Lenstra JA, Goderie CRJ, MacHugh DE, Park SDE, et al. European cattle genetic diversity consortium, Crooijmans RPMA. Genetic origin, admixture and population history of aurochs (*Bos primigenius*) and primitive European cattle. Heredity. 2017;118: 169–176. doi:10.1038/hdy.2016.79.

CHAPTER III - IMPROVEMENT IN ESTIMATION OF MATERNAL EFFECTS IN PIGS BY USING POOLED SEMEN: A SIMULATION STUDY

ABSTRACT - The use of pooled semen is a common technique in swine production systems. Assuming five simulated scenarios where paternity can be assessed from genotyping piglets and potential boars, the aim of the present study was to investigate whether maternal direct effect for litter birth weight can be improved when pooled semen is used. Different scenarios of pooling semen were emulated by allowing the same female to mate different sires such that the female offspring size summing across all sires was 12 (i.e., emulating a unique litter with size 12 but potentially sired by multiple boars). The genome was constructed based on the size of the 18 autosomes of the swine, considering 60.000 SNPs evenly distributed and 1.080 QTLs on predefined positions across the genome. Direct and maternal true breeding values (TBV) were computed as the sum of the effects of QTLs. Phenotypes were constructed as the sum of direct TBV, maternal TBV, a mean of 1.25kg, and a residual effect. The variance components, direct heritability, and maternal heritability were estimated using AIREMLF90. Predictors were computed via Best Linear Unbiased Prediction (BLUP) and single-step genomic BLUP (ssGBLUP). Pedigree information for all animals from the last generation were considered known in the first scenario with single-boar semen. Unknown boars identification for the last generation was also used for pooled semen scenarios. The animal model included an overall mean of litter birth weight as fixed effect and random effects of direct additive genetic effect, maternal additive genetic, and residual. Correlations of estimated and true breeding values from the validation population was used to evaluate accuracies of prediction for direct (acc_{direct}) and maternal (acc_{mat}) effects under BLUP and ssGBLUP methods. The estimated direct heritability and maternal heritability ranged from 0.04 to 0.06 and from 0.16 to 0.20, respectively. When pedigree records where known, acc_{direct} ranged from 0.21 to 0.26 in BLUP method and from 0.38 to 0.43 in ssGBLUP. For unknown sire identification scenarios, acc_{direct} ranged from 0.06 to 0.08 in BLUP and from 0.41 to 0.43 in ssGBLUP. In the same order, acc_{mat} ranged from 0.52 to 0.60, 0.58 to 0.66, 0.53 to 0.60, and 0.62 to 0.66. In general, acc_{direct} and acc_{mat} estimates were lower in single-boar scenario compared to pooled semen scenarios, indicating disadvantages of litters composed of full-sib piglets alone. Higher accuracies were found for ssGBLUP when compared to BLUP independently of the scenario, which is a clear evidence that accounting for genomic information improved EBV inferences.

Keywords: genomic prediction, maternal ability, multiple sire mating, prediction accuracy

1. Introduction

The use of pooled semen is a common technique in swine production systems because it enhances the fertility success ratio of inseminations. Semen doses containing sperm from several boars increase fertility by 0.3 piglets compared with semen from one boar (Pedersen, 2013).

The boar has an influence on his offspring phenotype by the transmitted genes, while the sow may influence her offspring through the genes and maternal environmental condition. The environmental proportioned by the female to the offspring occurs early in life by the cytoplasm of the egg, intra-uterine environment or after the birth by the milk production and mothering ability (Robison, 1972). The offspring phenotype is influenced by individual's own genotype (direct genetic effect) and environment and by the maternal ability of the sow, which is also influenced by the joint action of genotype and environment. Maternal effects were defined as the causal influence of the maternal genotype or phenotype on the offspring phenotype, although other definitions are focused on the detection or partitioning of maternal effects in other causal components of phenotypic variation (Wolf and Wade, 2009). An understanding of the genetic variation in maternal effects (maternal genetic effect) will facilitate the formulation of optimum breeding programs. Studies reported the importance of the inclusion of maternal effects in a model to improve accuracy of evaluation for economical traits (Roehe and Kennedy, 1993a, Jasouri et al., 2017). Nonetheless, complications arise in the context of accurate estimates of maternal effects.

The complex interrelationships between indirect environmental factors and the underlying genetic milieu make understanding the key contributors to phenotypic variation difficult, particularly when it is not possible to control genetic background or conduct controlled mating experiments (Noble et al., 2014). Mating designs considering full-sibs in pigs does not adequately allow the estimation of maternal effect because this effect can be confounded with dominance genetic variance. Dominance is a non-additive variance component in the genetic variation and arises from the interaction between alleles within a locus.

Dominance genetic variance is not expected to contribute to phenotypic resemblance among half-sibs (Charmantier et al., 2014). Different levels of multiple paternity provide an opportunity to assess the role of heritable and nongenetic maternal effects on offspring traits because families of maternal half siblings better allow the statistical separation of additive genetic and maternal effects than full sibling families (Noble et al., 2014). Information from multiple paternity can be incorporated in a swine production system through the use of pooled semen. This is a real situation in some swine production systems, where piglets in a litter can possibly be from different boars (i.e., maternal half-sibs).

The maternal effect is recognized as an important source of variation in the phenotypic variation of piglets for litter birth weight (Ahlschwede and Robison, 1971). The use of genomic selection became common for livestock because it enhances the genetic gain for economical traits. It was shown genomic selection has the potential to increase genetic gain for maternal related traits such as litter weight and litter size (Lillehammer et al., 2011). The beneficial use of genomics is increasing the accuracy of estimation of the genetic effects by adding molecular genetic information to pedigree and phenotypic information (Hayes et al., 2009). In addition, genomic allows identifying maternal half-sibs as different from full-sibs in a molecular basis.

Quantifying and accounting for maternal effects in a model to evaluate the animals could increase the accuracy of animal breeding programs (Lourenco et al., 2013; Josh et al., 2018) and improve the overall genetic merit in maternal traits (Roehe and Kennedy, 1993b). Few studies have investigated accuracy of the estimation of maternal effect by different mating designs in pigs (Eisen, 1967; Eisen 1966). They took into account only pedigree records. In view of the importance that genomic has in application to animal breeding, it is important to study the impact of the inclusion of genomic information on the estimation of maternal effects, in addition to direct genetic effects. Assuming a scenario where paternity can be assessed from genotyping piglets and potential sires, the aim of the present study was to investigate whether maternal direct effect for litter-related traits can be improved when pooled semen is used.

2. Materials and Methods

2.1. Scenarios

Phenotypes and genotypes were simulated by QMSim (Sargolzaei and Schenkel, 2009). This software does not explicitly allow for a pooled semen scheme. However, we simulated the populational effects of pooling semen by allowing the same female to mate different boars such that the female offspring size summing across all boars was 12 (i.e., emulating a unique litter with size 12 but potentially sired by multiple boars). The following simulation scenarios were applied:

1) Each sow was mated with a unique boar under simulated litter size = 12.
2) Each sow was mated with 2 boars under simulated litter size = 6.
3) Each sow was mated with 3 boars under simulated litter size = 4.
4) Each sow was mated with 4 boars under simulated litter size = 3.
5) Each sow was mated with 6 boars under simulated litter size = 2.

2.2. Simulation of data

As the main objective of this study was to investigate the impact of pooled semen schemes on the maternal ability, both direct and maternal effects had to be simulated. However, QMSim (Sargolzaei and Schenkel, 2009) can only generate one additive genetic effect a time; therefore, the software was run twice using the same seed file and same parameters, but different phenotypic variances as described in Lourenco et al. (2013). In a nutshell, the first run generated the additive genetic direct effect and the second generated the additive genetic maternal effect. Phenotypic variance for the first run was 0.005 and for the second 0.017, whereas heritabilities for direct and maternal effects were assumed to be 0.99; therefore, additive genetic variances approached phenotypic variances. To mimic observed heritabilities for birth weight in pigs (i.e., 0.06 for direct and 0.20 for maternal), a residual effect (e ~ N(0, σ_e^2=0.07)) was simulated using a standalone software. For simplicity, permanent environmental effect was not simulated, therefore, cannot be estimated in this study. Using a method described in VanVleck (1994), the genetic correlation between QTL effects for direct and maternal was set to -0.25. Five replicates for each scenario were simulated and analyzed.

The population structure started with a historical population to generate initial linkage disequilibrium, mutation and drift. This historical population was created considering 100 generations with an expansion in size from 100 to 3000 individuals. The number of males in the last historical generation was 100. A recent population started from 50 males and 1000 females randomly chosen from the historical population. This recent population underwent natural selection for 4 generations, with a replacement rate of 0.6 for boars and 0.45 for sows. After that, 50 males randomly chosen and 100 females chosen based on high phenotypes were mated mimicking single-boar or multiple-boar semen. For the single-boar scheme, each sow was mated to 1 boar and had a litter size of 12. For the pooled semen schemes, the population effect of pooling semen from different boars was emulated by allowing a multi-mating structure, given that QMSim (Sargolzaei and Schenkel, 2009) cannot handle the explicit pooled semen structure.

The total number of animals (61,050) was the same in all scenarios. Only animals from generations 3, 4, and 5 were included in the genotype file to be used in the genetic evaluations. In each generation, all 12,000 piglets had genotypic information. Generation 5 was considered as a validation population.

Parents from generations 3 and 4, together with all 12,000 piglets in generation 5 were genotyped. The Sus scrofa domesticus assembly was used as a reference genome for simulating the genotypes. According to Groenen et al. (2012), the Sus scrofa domesticus genome size is approximately 2596 cM with eighteen autosomes of different lengths. We defined chromosome 1 as the largest among the autosomes according to Groenen et al. (2012) with a length of 298 cM and 7600 evenly distributed SNP. Chromosome 18 was defined as the smallest among the autosomes with a length of 61 cM and 1,790 SNP. Overall, the total number of evenly distributed, biallelic SNP was 60,000 SNPs and the number of biallelic QTL was 1080. The QTL effects were simulated from a Gamma distribution with shape parameter equal to 0.40, which resulted in several QTL with small effects and few QTL with large effect. All SNP and QTL had initial allele frequency of 0.5; however, this value changed along generations. The recurrent mutation was 2.5 × 10-5 mutations per meiosis per loci. Additive genetic variance was completely explained by the 1080 QTL. Direct and maternal true breeding values (TBV) were

computed as the sum of the effects of pre-defined QTL. Phenotype for an animal was constructed as the sum of its direct TBV, maternal TBV from its dam, residual effect, and a mean of 1.25kg.

Before the availability of genomic information, using pooled semen would increase the missingness of pedigree because individual boars could not be identified. Regarding to this issue, we investigated two scenarios. In the first one we used genomic information to construct the genomic relationship matrix, however, we considered missing boar information in the pedigree. In the second scenario we assumed genomic information was used to construct the genomic relationship matrix and to identify individual boars, so pedigrees were considered complete for both ssGBLUP and BLUP.

2.3. Statistical analysis

To check the quality of the simulated data, variance components, direct heritability, and maternal heritability were estimated using the AIREMLF90. Predictors were computed via traditional BLUP and single-step GBLUP (ssGBLUP) methods.

When single-boar semen scheme was used, pedigree information was complete for both BLUP and ssGBLUP. Under the pooled semen scheme, traditional BLUP considered missing boar information in the pedigree and ssGBLUP considered either missingness of pedigree or pedigree assignment using genomic information.

The following animal model was applied for BLUP and ssGBLUP methods:

$$\mathrm{y} = \mu + \mathrm{Da} + \mathrm{Mm} + \mathrm{e}$$

where **y** is the vector of phenotype, **μ** is the vector of an overall mean of the phenotype, **a** is the vector of direct additive genetic effect, **m** is the vector of maternal additive genetic effect, **e** is the vector of random residual effect; **D** and **M** are incidence matrices for a and m, respectively. The assumptions for the random effects in BLUP were a~N(0,Aσ_a^2), m~N(0,Aσ_m^2), in which A is the pedigree relationship matrix; in ssGBLUP the assumptions were a~N(0,Hσ_a^2) and m~N(0,Hσ_m^2), in which H is the realized relationship matrix that combines pedigree and genomic information. The inverse of H, as used in the mixed model equations as the following form (Aguilar et al., 2010):

$$\mathrm{H}^{-1} = \mathrm{A}^{-1} + \begin{pmatrix} 0 & 0 \\ 0 & \mathrm{G}^{-1} - \mathrm{A}_{22}^{-1} \end{pmatrix}$$

where A_{22}^{-1} is the inverse of the relationship matrix among genotyped animals and G^{-1} is the inverse of the genomic relationship matrix. The G matrix was created as in VanRaden (2008):

$$\mathrm{G} = 0.95\mathrm{G}^{*} + 0.05\mathrm{A}_{22}$$

with

$$G^{*} = KK'; \text{ with } K = \frac{(M-P)}{\sqrt{2\sum_{j=1}^{n} p_j(1-p_j)}}$$

where *M* is a matrix of n SNP genotypes for each animal, *P* is a matrix of 2 times the allele frequency of the second allele *p* at locus *j* (p_j).

Accuracy of predicting EBV and genomic EBV (GEBV) for the direct effect was calculated as correlation between TBV and (G)EBV for genotyped animals in generation 5 that had phenotypes removed from the analysis. For the maternal effect, correlation between maternal TBV of genotyped dams of animals in generation 5 and their (G)EBV for maternal effect. Accuracies were compared over all boar schemes.

3. Results and Discussion

The average (SE) estimates of direct and maternal heritability were 0.05 (0.004) and 0.19 (0.006), respectively, over all boar schemes. These results are in agreement with the parametric values used in the simulation, showing the simulation process was effective in generating the desired data structure for this study. As a common practice, variance components obtained based only on phenotypes and pedigree were used for both BLUP and ssGBLUP.

Accuracies for traditional and genomic evaluations for single-boar semen and pooled semen considering unknown and known boar information are in Table 1. For the traditional BLUP method, accuracy of direct effect was greater when boars were known, compared to the unknown boar scenario. Missingness of pedigree can negatively affect predictions increasing bias and decreasing accuracy (Quaas, 1988; VanRaden, 1992). For ssGBLUP, the inclusion of pedigree records for boars that sired the validation piglets

did not change the accuracy for direct and maternal effects. The main reason is that genomic relationships in ssGBLUP provide accurate information, and although coefficients in $\boldsymbol{A^{-1}}$ are zero between an animal and its real but missing boar, $\boldsymbol{G^{-1}}$ has non-zero coefficients that reflect the proportion of alleles shared between them. Genomic EBV of validation animals are composed of parent average (PA) and direct genomic value (DGV). When boars are unknown, PA is based only on sows which have usually less information, therefore, less accurate EBV than boars. Additionally, when validation animals have genotyped parents, most likely the information in $\boldsymbol{A^{-1}}$ cancels out with $\boldsymbol{A_{22}^{-1}}$, remaining only the information in $\boldsymbol{G^{-1}}$. Therefore, ssGBLUP can assess the realized relationship between a boar and its offspring in the absence of recorded pedigree.

From the second scenario, simulation was used to mimic a population of maternal half-sibs. As mentioned by Noble et al. (2014) families of maternal half siblings better allow the statistical separation of additive genetic and maternal effects than full sibling families.

In general, accuracies of estimating both direct and maternal effects were less in the single-boar scenario compared to pooled semen scenarios, indicating disadvantages of litters composed of full-sib piglets. However, a decrease in accuracy for the maternal effect was observed when each female offspring had only pairs of full-sib piglets (i.e., when using 6 boars). In a study with genetic associative effect, Cheng et al. (2009) reported that full sib assignment resulted in bias in the estimation of variance components compared to random assignment or mixed sib assignment. The authors mentioned the genetic direct effect and genetic associative effect may face a problem of identifiability, which means parameters are weakly identified because the data provide little information. In our study, the direct and maternal effects are associated and this correlation was given by the relationship matrix and the additive genetic covariances.

Table 1. Average accuracy of prediction for direct and maternal effects for BLUP and ssGBLUP

	known boar				unknown boar			
	BLUP		ssGBLUP		BLUP[1]		ssGBLUP[2]	
	Direct	Maternal	Direct	Maternal	Direct	Maternal	Direct	Maternal
1 boar	0.21 ± 0.02	0.52 ± 0.02	0.38 ± 0.02	0.58 ± 0.01	-	-	-	-
2 boars	0.24 ± 0.01	0.55 ± 0.02	0.42 ± 0.02	0.63 ± 0.01	0.07 ± 0.01	0.55 ± 0.02	0.42 ± 0.02	0.63 ± 0.01
3 boars	0.27 ± 0.02	0.60 ± 0.02	0.43 ± 0.01	0.65 ± 0.01	0.08 ± 0.02	0.60 ± 0.02	0.43 ± 0.01	0.65 ± 0.01
4 boars	0.22 ± 0.01	0.57 ± 0.04	0.41 ± 0.01	0.66 ± 0.03	0.06 ± 0.01	0.59 ± 0.03	0.41 ± 0.01	0.66 ± 0.03
6 boars	0.26 ± 0.02	0.53 ± 0.02	0.43 ± 0.01	0.63 ± 0.02	0.06 ± 0.01	0.53 ± 0.02	0.43 ± 0.01	0.62 ± 0.02

[1,2]In the evaluations, boar was considered unknown among the list of 2-6 potential boars, i.e., only for the validation population.

Greater accuracies for direct effect were found in ssGBLUP when compared to BLUP, independently of the boar scheme, which is a clear evidence that accounting for genomic information improved accuracy of breeding values. The gain in accuracy between BLUP and ssGBLUP models was greater for direct effect using unknown boars. In the scenarios that consider 4 boars and 6 unknown boars, accuracy of GEBV increased from 0.06 to 0.41 and from 0.06 to 0.43, respectively, being approximately 7 times higher compared to the accuracies obtained in BLUP model. Part of this increase is likely due to the inclusion of animal's own genotype in the evaluations, which has a greater impact on the accuracy of its own GEBV. Similar results were found by Tonussi et al. (2017) studying scenarios with multiple sires and different levels of missing pedigree records in a simulated beef cattle population. They reported prediction accuracy for direct effect of 5 to 6 times greater in ssGBLUP compared to traditional BLUP in a scenario with all unknown sires.

Our findings showed that not only the direct effect is benefiting by the inclusion of genomic information in the evaluations, but also the maternal effect. Lourenco et al. (2013) recommend the inclusion of maternal effect in the model to reduce bias in genomic evaluation of maternally affected traits.

The greatest accuracy for maternal effect was obtained when 3 to 4 boars were used, and results were similar for both known and unknown boar scenarios. This similarity is because the maternal effect reflects the genetic ability of the sow to provide a nurturing environment for the offspring, independently if the boars are identified or not. Under traditional BLUP and using only 1 boar, accuracy was 8 points less than when using 3 boars, and 1 point less than using 6 boars. When using genomic information, the greatest advantage over traditional BLUP was seen in the 6 boars scheme (10 points). For ssGBLUP the greatest increase in accuracy compared to using single-boar was 8 points in the 4 boars scheme. This shows that having a large number of boars is not beneficial because the size of the maternal half-sib family is reduced, not allowing a good estimation of maternal breeding values. Studies that clearly show results from the comparison between traditional EBV and genomic EBV were still not reported in the literature for pigs. Some studies have reported the importance of the use of genomic evaluation to increase

accuracies of selection (Cleveland et al., 2010; Guo et al., 2016) and the importance of genotype imputation as a strategy for genomic evaluation in pigs (Badke et al., 2014).

As genotyping cost has been decreasing, genotyping is already being considered by industry in many domestic animals, including swine. In the future, genotyping will be more affordable for small populations and countries with limited economic power. Many are the reason for the beneficial use of genomic information in animal genetic evaluations: improvement in accuracy of EBV, improvement in accuracy of traits with low heritability or traits hard of measuring, and higher genetic gain and consequently reduction of the generation interval (Meuwissen et al., 2001). Lillehammer et al. (2011) observed that genotyping dams in addition to the male selection candidates should increase accuracy of the genomic breeding value, genetic gain, and decrease rate of inbreeding in pigs. Our findings suggested that breeding program with pooled semen should continue the evaluations with genomic information so that predicted GEBV can be verified.

Overall, we can conclude that using pooled semen from few boars can help to improve accuracy of predicting maternal and direct effects when the size of maternal half-sib family is greater than 2. Even when paternity tests are not carried out and boar information is missing in the pedigree, using genomic information in a single-step genomic BLUP helps to keep maximum levels of prediction accuracy.

4. References

Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ (2010) Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. **Journal of Dairy Science** 93:743–752.

Ahlschwede WT, Robison OW (1971) Maternal effects on weights and backfat of swine. **Journal of Animal Science** 33:1206-1211.

Badke YM, Bates RO, Ernst CW, Fix J, Steibel JP (2014) Accuracy of estimation of genomic breeding values in pigs using low-density genotypes and imputation. **Genomic Selection** 4:623-631.

Charmantier A, Garant D, Kruuk LEB (2014) Quantitative genetics in the wild. Oxford: Oxfor university Press. 288p.

Cheng J, Janssens S, Buys N (2009)Full sib pens of pigs are not suitable to identify variance component of associative effect: a simulation study using Gibbs Sampling. **BMC Genetics** 10:9.

Cleveland MA, Forni S, Garrick DJ, Deeb N (2010) Prediction of genomic breeding values in a commercial pig population. In: WORLD CONGRESS ON GENETICS APPLIED TO LIVESTOCK PRODUCTION. **Proceedings…** Leipzig, Germany: International Committee of WCGALP. p. 0266.

Eisen EJ, Bohren BB, McKean HE (1966). Sex-linked and maternal effects in the diallel cross. **Australian Journal of Biological Science** 19:1061-1071.

Eisen EJ (1967) Mating designs for estimating direct and maternal genetic variances and direct-maternal genetic covariances. **Canadian Journal of Genetics and Cytology** 9:13-22.

Groenen MAM, Archibald AL, Uenishi H, Tuggle CK, Takeuki Y, Rthshild MF, Rogel-Gailard C, Park C, Milan D, Megens H, Li S, Larkin DM, Kim H, Frantz LAF, Caccamo M, Ahn H, Aken BL, Anselmo A, Anthon C, Auvil L, Badaoui B, Beattie CW, Bendixen C, Berman D, Blecha F, Blomberg J, Bolund L, Bosse M, Botti S, Bujie Z, Bystrom M, Capitanu B, Carvalho-Silva D, Chardon P, Chen C, Cheng R, Choi SH, Chow W, Clark RC, Clee C, Crooijmans RPMA, Dawson HD, Dehais P, De Sapio F, Dibbits B, Drou N, Du ZQ, Eversole K, Fadista J, Fairley S, Faraut T, Faulkner GJ, Fowler KE, Fredholm M, Fritz E, Gilbert JGR, Giuffra E, Gorodkin J, Griffin DK, Harrow JL, Hayward A, Howe K, Hu ZL, Humphray SJ, Hunt T, Hornshøj H, Jeon JT, Jern P, Jones M, Jurka J, Kanamori H, Kapetanovic R, Kim J, Kim JH, Kim KW, Kim TH, Larson G, Lee K, Lee KT, Leggett R, Lewin HA, Li Y, Liu W, Loveland JE, Lu Y, Lunney JK, Ma J, Madsen O, Mann K, Matthews L, McLaren S, Morozumi T, Murtaugh MP, Narayan J, Nguyen DT, Ni P, Oh SJ, Onteru S, Panitz F, Park EW, Park HS, Pascal G, Paudel Y, Perez Enciso M, Ramirez-Gonzalez R, Reecy JM, Rodriguez-Zas S, Rohrer GA, Rund L, Sang Y, Schachtschneider K, Schraiber JG, Schwartz J, Scobie L, Scott C, Searle S, Servin B, Southey BR, Sperber G, Stadler P, Sweedler JV, Tafer H, Thomsen B, Wali R, Wang J, Wang J, White S, Xu X, Yerle M, Zhang G, Zhang J, Zhang J, Zhao S, Rogers J, Churcher C, Lawrence B (2012) Analyses of pig genomes provide insight into porcine demography and evolution. **Nature** 491:393–398.

Guo X, Christensen OF, Ostersen T, Wang Y, Lund MS, Su G (2016) Genomic prediction using models with dominance and imprinting effects for backfat thickness and average daily gain in Danish Duroc pigs. **Genetic Selection Evolution** 48:67.

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Invited review: genomic selection in dairy cattle: Progress and challenges. **Journal of Dairy Science** 92:433–443.

Jasouri M, Zamani P, Alijani S (2017) Dominance genetic and maternal effects for genetic evaluation of egg production traits in dual-purpose chickens. **British Poultry Science** 58:498–505.

Joshi R, Woolliams JA, Meuwissen THE, Gjøen HM (2018) Maternal, dominance and additive genetic effects in Nile tilapia; influence on growth, fillet yield and body size traits. **Heredity** 120:452–462.

Lillehammer M, Meuwissen THE, Sonesson AK (2011) Genomic selection for maternal traits in pigs. **Journal of Animal Science** 89:3908-3916.

Lourenco DAL, Misztal I, Wang H, Aguilar I, Tsuruta S, Bertrand JK (2013) Prediction accuracy for a simulated maternally affected trait of beef cattle using different genomic evaluation models. **Journal of Animal Science** 91:4090–4098.

Meuwissen THE, Hayes BJ, Goddard ME (2001). Prediction of total genetic value using genome-wide dense marker maps. genetics. **Genetics** 157:1819–1829.

Noble DWA, McFarlane SE, Keogh JS, Whiting MJ (2014) Maternal and additive genetic effects contribute to variation in offspring traits in a lizard. **Behavioral Ecology** 25:633–640.

Pedersen, ML (2013) **Fertility higher with pooled duroc semen than with semen from one boar**. Denmark: Pig Research Centre. 7p. (Trial Report, 969). Disponível em: https://pdfs.semanticscholar.org/20db/5111f25aacc0c966c8b9f9667bbcc7725179.pdf?_ga=2.82947861.2003103432.1548764279-1563262216.1548764279. Acesso em: 29 jan. 2019.

Quaas RL (1988) Additive genetic model with groups and relationships. **Journal of Dairy Science** 71:1338–1345.

Robison OW (1972) The role of maternal effects in animal breeding: V. Maternal effects in swine. **Journal of Animal Science** 35:1303-1315.

Roehe R, Kennedy BW (1993a) The influence of maternal effects on accuracy of evaluation of litter size in swine. **Journal of Animal Science** 71:2353-2364.

Roehe R, Kennedy BW (1993b) Effect of selection for maternal and direct genetic effects on genetic improvement of litter size in swine. **Journal of Animal Science** 71:2891-2904.

Sargolzaei M, Schenkel FS (2009) QMSIM: a large-scale genome simulator for livestock. **Bioinformatics** 25:680–681.

Tonussi RL, Silva RMO, Magalhães AFB, Espigolan R, Peripolli E, Olivieri BF, Feitosa FLB, Lemos MVA, Berton MP, Chiaia HLJ, Pereira ASC, Lobo RB, Bezerra LAF, Magnabosco CU, Lourenco DAL, Aguilar I, Baldi F (2017) Application of single step genomic BLUP under different uncertain paternity scenarios using simulated data. **PLOS ONE** 1-14.

VanRaden PM (1992) Accounting for inbreeding and crossbreeding in genetic evaluation of large populations. **Journal of Dairy Science** 75:3136–3144.

VanRaden PM (2008) Efficient methods to compute genomic predictions. **Journal of Dairy Science** 91:4414–4423.

Van Vleck LD (1994) Algorithms for simulation of animal models with multiple traits and with maternal and non-additive genetic effects. **Revista Brasileira de Genetica** 17:53–57.

Wolf JB, Wade MJ (2009) What are maternal effects (and what are they not)? **Philosophical Transactions of the Royal Society B: Biological Sciences** 364:1107–1115.