UNIVERSIDADE ESTADUAL PAULISTA - UNESP

CÂMPUS DE JABOTICABAL

ESTUDO DE ASSOCIAÇÃO GENÔMICA AMPLA E ANÁLISE FUNCIONAL COM DADOS DE SEQUENCIAMENTO COMPLETO PARA HIPOPLASIA TESTICULAR E CARACTERÍSTICAS DE CONFORMAÇÃO DE PERNAS E PÉS EM BOVINOS DA RAÇA NELORE

Thales de Lima Silva

Médico Veterinário

2023

UNIVERSIDADE ESTADUAL PAULISTA - UNESP

CÂMPUS DE JABOTICABAL

ESTUDO DE ASSOCIAÇÃO GENÔMICA AMPLA E ANÁLISE FUNCIONAL COM DADOS DE SEQUENCIAMENTO COMPLETO PARA HIPOPLASIA TESTICULAR E CARACTERÍSTICAS DE CONFORMAÇÃO DE PERNAS E PÉS EM BOVINOS DA RAÇA NELORE

Thales de Lima Silva

Orientador: **Dr. Roberto Carvalheiro**

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

2023

S586e

Silva, Thales de Lima

Estudo de associação genômica ampla e análise funcional com dados de sequenciamento completo para hipoplasia testicular e características de conformação de pernas e pés em bovinos da raça Nelore / Thales de Lima Silva. -- Jaboticabal, 2023

108 p. : il., tabs.

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

Orientador: Roberto Carvalheiro

1. Bovinos de corte Melhoramento Genético. 2. Nelore (Bovino). 3. Genômica. 4. SNPs. 5. Malformações congênitas. 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.

unesp

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Jaboticabal

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: ESTUDO DE ASSOCIAÇÃO GENÔMICA AMPLA E ANÁLISE FUNCIONAL COM DADOS DE SEQUENCIAMENTO COMPLETO PARA HIPOPLASIA TESTICULAR E CARACTERÍSTICAS DE CONFORMAÇÃO DE PERNAS E PÉS EM BOVINOS DA RAÇA NELORE

AUTOR: THALES DE LIMA SILVA
ORIENTADOR: ROBERTO CARVALHEIRO

Aprovado como parte das exigências para obtenção do Título de Doutor em Genética e Melhoramento Animal, pela Comissão Examinadora:

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

Prof. Dr. CEDRIC GONDRO (Participaçao Virtual)
Michigan State University / Michigan/EUA

Prof. Dr. DANISIO PRADO MUNARI (Participaçao Presencial)
Departamento de Engenharia e Ciencias Exatas DECEx / FCAV UNESP Jaboticabal

Prof. Dr. DELVAN ALVES DA SILVA (Participaçao Virtual)
Departamento de Zootecnia / Universidade Federal de Viçosa

gov.br Documento assinado digitalmente
DELVAN ALVES DA SILVA
Data: 19/04/2023 06:22:57-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. HENRIQUE NUNES DE OLIVEIRA (Participaçao Presencial)
Departamento de Zootecnia / FCAV UNESP Jaboticabal

gov.br Documento assinado digitalmente
HENRIQUE NUNES DE OLIVEIRA
Data: 19/04/2023 10:58:51-0300
Verifique em https://validar.iti.gov.br

Jaboticabal, 31 de março de 2023

DADOS CURRICULARES DO AUTOR

Thales de Lima Silva – Nasceu em Brumado, Bahia em 05 de março de 1981, filho de Aparecido Souza Silva e Deuza de Lima Silva. Em agosto de 2007 iniciou o curso de Medicina Veterinária na Pontifícia Universidade Católica de Minas Gerais em Betim como bolsista do Programa Universidade para Todos – PROUNI, do governo federal, graduando-se em julho de 2012. Entre fevereiro de 2013 e julho de 2015 cursou MBA em Agronegócios pela Escola Superior de Agricultura Luiz de Queiroz – ESALQ/USP, tendo seu trabalho de conclusão de curso "Estatística como ferramenta para mitigar o risco de preço sobre o hedge de boi gordo" publicado na Revista iPecege em abril de 2016. Em março de 2017 iniciou o mestrado pelo Programa de Pós-Graduação em Zootecnia da Universidade Estadual do Sudoeste da Bahia – UESB como bolsista CAPES, dedicando-se ao estudo da interação genótipo-ambiente na análise da variabilidade genética de características reprodutivas na raça Nelore, sob a orientação do professor Dr. Carlos Henrique Mendes Malhado, recebendo o título de mestre em Fevereiro de 2019. No mesmo ano iniciou o doutorado no Programa de Pós-Graduação em Genética e Melhoramento Animal da Faculdade de Ciências Agrárias e Veterinárias da UNESP – Campus de Jaboticabal, sob a orientação do pesquisador Dr. Roberto Carvalheiro, sendo bolsista CAPES. Entre dezembro de 2021 e novembro de 2022 realizou doutorado sanduíche na Michigan State Univesity, EUA, sob orientação do prof. Dr. Cedric Gondro e com apoio financeiro do Programa Institucional de Internacionalização - PRINT, da CAPES.

“O conhecimento não é predeterminado pela hereditariedade; não é predeterminado nas coisas que nos rodeiam - em conhecer as coisas ao seu redor, o sujeito sempre acrescenta algo a elas.”

Jean Piaget

DEDICO

A todos aqueles que têm a coragem de sonhar e correr o risco de viver seus sonhos, mesmo que tenham que pagar um alto preço por isso.

AGRADECIMENTOS

Aos meus pais Deuza e Aparecido pelo seu amor incondicional e pelo sacrifício e dedicação para que eu e meus irmãos tivéssemos uma educação de excelência.

Aos meus irmãos – Moab, Naiara e Bruno, aos meus sobrinhos – Heitor e Júlia, ao Rafael, Anayara e Aline, por todo apoio, confiança e respeito. Amo todos vocês!

Ao meu orientador Dr. Roberto Carvalheiro por acreditar no meu potencial e pela serenidade e profissionalismo com que me orientou apesar das adversidades que enfrentamos durante os anos da realização deste doutorado.

Ao Prof. Dr. Cedric Gondro e sua esposa Simone, pelo acolhimento, gentileza e ensinamentos durante minha estadia em Michigan e pela humanidade com que me trataram durante o momento mais difícil deste doutorado. Serei eternamente grato!

Aos Profs. Drs. Lúcia Galvão de Albuquerque, Danísio Prado Munari e Henrique Nunes de Oliveria e ao Diego Mafra (STPG), pelo exemplo de paciência, atenção, suporte, e colaboração.

Aos Drs. Delvan Silva, Gabriel Campos, Giovana Vargas, Pablo Fonseca e Haroldo Neves pelos valiosos ensinamentos e colaboração tanto na fase de análise dos dados quanto na redação desta tese.

Aos amigos da UNESP – Ivan, Caio, Willian, Larissa, Baltasar, Gabi, Patrícia, Bruna, Rafael e Diego, e da MSU – Lee, Rodrigo S., Hanna, Lautaro, Fei, Ken, Bayode, Yasir, Gabriel, Bia e Junjie, pelo companheirismo, apoio e cumplicidade.

Aos meus colegas da república MK2.0 – Rabera, David, Dodô, Hugo, Thallison, Bernardo e Andrus. E aos meus *roommates* no Spartan Village – Luiz, Piush e Luyando. Conviver com vocês foi a melhor parte desta jornada.

Aos amigos que fiz ao longo do caminho e que quero levar para a vida toda – Marcos, Viviane, Dan Winter, Thomas McCurdy, Hélio & Hyllana, Maiyara, Rodrigo C., Tiago, Eduardo & Sabrina, Bia & Bernard, Simone & Cris, Thainá, Leo Ziccardi, Leo Volp & Bella, Amanda S. & Juliano S., João, Alisson, Rod Friesen, Terry, Mylena, Amanda & Juliano G., Ricardo & Vanessa, Paulo C., Pedro.

À FCAV/UNESP e à Michigan State University pela oportunidade de desenvolver este trabalho.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. e à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP – processos nº 2015/25449-6 e 2017/03331-9), pelo apoio financeiro.

SUMÁRIO

ESTUDO DE ASSOCIAÇÃO GENÔMICA AMPLA E ANÁLISE FUNCIONAL COM DADOS DE SEQUENCIAMENTO COMPLETO PARA HIPOPLASIA TESTICULAR E CARACTERÍSTICAS DE CONFORMAÇÃO DE PERNAS E PÉS EM BOVINOS DA RAÇA NELORE

RESUMO – A Hipoplasia Testicular (TH) e a Malformação de Pernas e Pés (FLM) estão entre os principais problemas funcionais observados em gado de corte devido à importância da correlação da medidas testiculares com características relacionadas a fertilidade, bem como o bem-estar animal, desempenho e longevidade associados à má formação das pernas e pés. Através de estudos de associação genômica ampla (GWAS) é possível encontrar segmentos do genoma que expliquem uma parte substancial da variação em doenças ou características de interesse econômico. Quando realizados com dados de sequenciamento completo (WGS) espera-se alcançar um refinamento na caracterização de variantes causais possibilitando uma predição de valores genômicos mais precisos. Os objetivos deste estudo foram: i) Estimar parâmetros e tendências genéticas para TH e FLM na raça Nelore; ii) investigar possíveis genes funcionais e vias metabólicas associadas à TH e FLM; iii) compreender a arquitetura genética destas características. Foram utilizados os fenótipos de 1.036.960 animais da raça Nelore e genótipos imputados para HD (~777.000 SNPs) e sequência tendo como referência 151 animais provenientes da base de dados Aliança Nelore. A herdabilidade estimada para TH (0,16) e FLM (0,19) na raça Nelore concorda com valores já relatados na literatura e sugere uma redução da incidência destas malformações por meio da seleção. A tendência fenotípica evidencia a importância de se intensificar esta seleção, apesar do ganho genético anual favorável (TH = -0.12; FLM = -0.25) no período avaliado. Dados de HD e sequencia passibilitaram a identificação de diferentes regiões cromossômicas associadas principalmente à espermatogênese, síntese de esteróides, metabolismo lipídico e esqueletogênese. Novos estudos podem vir a investigar o conjunto de variantes genéticas significativas aqui relatadas, buscando melhorar as decisões de seleção e acasalamento para reduzir a incidência destas em bovinos da raça Nelore.

Palavras chave: acurácia de predição, *Bos indicus*, característica funcional, genes candidatos, tendência genética.

GENOME-WIDE ASSOCIATION STUDY AND FUNCTIONAL ANALYSIS WITH WHOLE-GENOME SEQUENCING DATA FOR TESTICULAR HYPOPLASIA AND FEET AND LEG MALFORMATION TRAITS IN NELORE CATTLE

ABSTRACT – Testicular Hypoplasia (TH) and Feet and Legs Malformation (FLM) are important functional problems observed in beef cattle because of the correlated testicular measurements with fertility-related traits, as well as welfare, performance and longevity related to healthy feet and legs. Through Genomic-Wide Association Studies (GWAS), it is possible to find genomic fragments that explain a significant portion of variation in diseases or traits of economic interest. Improvements in the characterization of causal variants are expected when we use Whole-Genome sequencing (WGS) data to predict genomic values more accurately. The objectives of this study were: i) to estimate the genetic parameters and trends of TH and FLM in the Nellore cattle; ii) to investigate possible functional genes and pathways associated with TH and FLM; iii) to understand the genetic architecture of these traits. We used phenotypes of 1,036,960 Nellore animals and genotypes imputed for HD (~777,000 SNPs) and WGS using as reference 151 sequenced animals from the Aliança Nelore database. Estimated heritability for TH (0.16) and FLM (0.19) in the Nellore breed was consistent with values previously reported in the literature, suggesting a reduction in the incidence of these malformations by selection. Phenotypic trend indicated the importance of reinforcing this selection, despite a favorable annual genetic gain over the study period (TH = -0.12; FLM = -0.25). Sequence and HD data allowed identification of distinct chromosomal regions primarily associated with spermatogenesis, synthesis of steroids, lipid metabolism, and skeletogenesis. New research may examine the important genetic variants reported here with the aim of improving selection and mating decisions to reduce their incidence in Nellore cattle.

Key words: *Bos indicus*, functional trait, candidate genes, genetic trend, prediction accuracy.

CAPÍTULO 1 - Considerações gerais

1.1. Introdução

De acordo com dados da FAO (Food and Agriculture Organization of the United Nations) o Brasil possuía o maior rebanho bovino do mundo em 2020 com 217 milhões de cabeças, representando 14,3% do rebanho mundial. Neste mesmo ano, o país também passou a ocupar o posto de maior exportador de carne bovina 14,4% do mercado internacional, ou 2,2 milhões de toneladas. Entretanto, o país ocupa apenas a terceira colocação em valor (US$ 7 bilhões), sendo superado pela Austrália (US$ 8.5 bilhões) e EUA (US$ 8.1 bilhões) (Aragão & Contini, 2021). Desta forma, o Brasil ainda tem muito a explorar neste mercado em termos de valor. Isto será possíel dando um maior enfoque na qualidade da carne bovina produzida no país.

Para melhorar a qualidade da carne bovina exportada, de modo a alcançar mercados mais exigentes, o rebanho brasileiro precisa ser melhorado geneticamente. Entretanto, o método de monta natural ainda é extensamente utilizado nos rebanhos comerciais, e o país ainda não consegue produzir touros geneticamente superiores em número suficiente para atender sua demanda de reposição anual. Na raça Nelore, que corresponde a 80% do rebanho nacional, mesmo que fossem utilizados todos os machos à desmama dos programas de seleção, com controle de genealogia e todos os touros portadores de CEIP (Certificado Especial de Identificação e Produção), ainda haveria um déficit de cerca de 48% de touros para reposição (Rosa et al., 2021).

O CEIP permite que animais geneticamente superiores sejam certificados como melhoradores (MAPA, 2006). O que é necessário para que animais não registrados possam ser inscritos como doadores de sêmen e garantir benefícios fiscais na comercialização. Touros com CEIP são fontes de receita e importantes ferramentas para melhorar a eficiência de produção dos rebanhos que os utilizam. Entre 20% e 30% dos melhores animais de cada safra são candidatos a receber a certificação, mas alguns acabam não recebendo devido a problemas morfológicos identificados na inspeção final. Esse déficit de touros melhoradores no país é agravado pelas perdas devido a características desclassificantes.

Dois dos principais problemas funcionais observados em gado de corte nos programas CEIP são a hipoplasia testicular (Testicular Hypoplasia - TH) e a malformação de pernas e pés (Feet and Legs Malformation - FLM). Os animais

acometidos geralmente são descartados por contribuírem para a diminuição da eficiência em sistemas de produção de carne pela sua incapacidade de desempenhar satisfatoriamente funções produtivas e reprodutivas. Estudos recentes relacionados a estes atributos morfológicos/funcionais mencionam a ocorrência de perdas econômicas consideráveis na raça Nelore (Neves et al., 2019; Vargas et al., 2017), e indicam a possibilidade de melhorias por meio da seleção de reprodutores.

Os Estudos de Associação Genômica Amplos (Genome-Wide Association Study – GWAS) e as análises funcionais podem ajudar a identificar quais genes chave reguladores, ou subjacentes, e processos biológicos estão envolvidos com a ocorrência destas malformações contribuido para o processo de seleção. Dados o aumento considerável do número de animais genotipados nas bases de dados da raça Nelore a recente atualização no mapa de referência genômica da espécie bovina, se faz necessário trazer novas considerações para o conhecimento destas características objetivando animais com um maior potencial genético produtivo e uma menor incidência de problema funcionais.

1.2. Objetivo Geral

Conduzir análise de associação genômica ampla para identificar genes candidatos e/ou potenciais regiões no genoma relacionadas à variação do valor genético para TH e FLM, utilizando dados imputados para sequenciamento completo, buscando melhorar a precisão da seleção em bovinos da raça Nelore.

1.3. Objetivos Específicos

- Estimar os parâmetros genéticos para TH e FLM, incluindo a correlação genética entre TH e perímetro escrotal (SC) e entre FLM e peso ao sobreano (YW). Estimar as tendências genéticas para estas características na raça Nelore.
- Investigar possíveis genes funcionais e vias metabólicas associadas às regiões do genoma em que as variantes foram significativas na análise de associação ampla do genoma.
- Compreender a arquitetura genética destas características e fornecer subsídios aos programas de melhoramento genético da raça que possam contribuir para a seleção a fim de reduzir a ocorrência destes problemas nos animais.

1.4. Revisão de literatura

1.4.1. Hipoplasia testicular

A hipoplasia testicular (TH) é uma alteração genética caracterizada pelo desenvolvimento incompleto do epitélio germinativo dos túbulos seminíferos devido ao número inadequado de células germinativas no testículo, levando à diminuição da concentração espermática e aumento das alterações morfológicas dos espermatozoides, bem como da azoospermia (McEntee, 1990). O parênquima testicular é o principal responsável pelo volume gonadal, sendo composto, basicamente, pelos túbulos seminíferos e o tecido intersticial. Os túbulos seminíferos contêm o tecido germinativo (espermatogônias) e um conjunto de células especializadas para o suporte da produção de espermatozoides, as células de Sertoli. O tecido intersticial é formado por uma rede de tecido conjuntivo frouxo, que preenche o espaço entre os túbulos seminíferos e os vasos sanguíneos. O principal tipo celular que reside no tecido intersticial são as células de Leydig. Estas células organizam-se em clusters no tecido intersticial, sendo responsáveis pela produção de andrógenos nos testículos (Fonseca, 2018).

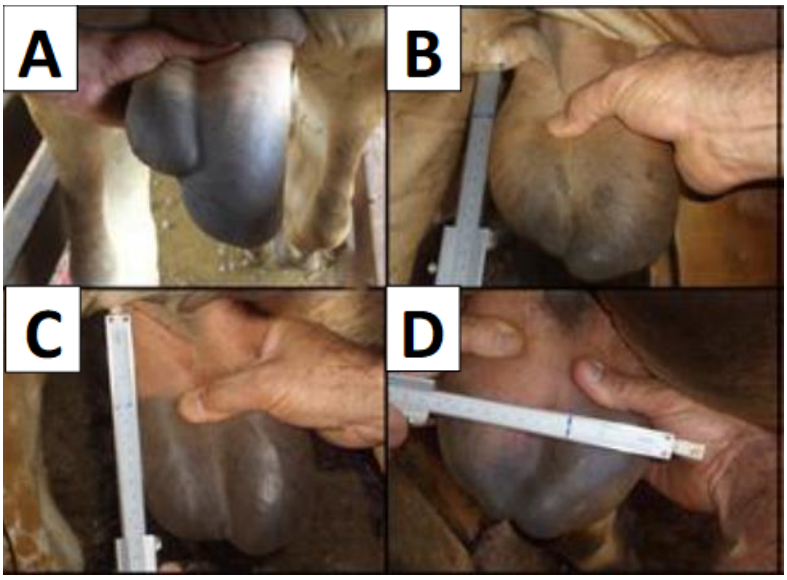

Figura 1. Características morfológicas testiculares de animais hipoplásicos e saudáveis. Os animais A, B e C apresentam hipoplasia testicular com graus de assimetria diferentes, enquanto o animal D é clinicamente normal. Fonte: Adaptado de (Moura, 2013).

A TH pode ser uni ou bilateral, parcial ou total (Steffen, 1997). Touros com hipoplasia unilateral parcial ou total, ou bilateral parcial apresentam libido, capacidade de fecundar e habilidade de serviço normal, apesar de não serem recomendados para uso na reprodução (Figura 1) (Moura, 2013). Devido à maior frequência de TH ser do tipo unilateral parcial, a maioria dos touros apresenta-se subfértil, transmitindo a condição indesejável aos seus descendentes (Bicudo et al., 2007). A subfertilidade é caracterizada por um sucesso reprodutivo inferior à média geral da população, porém, sem a presença de infertilidade no rebanho, chegando a compor de 20-40% de alguns rebanhos bovinos (Gnoth et al., 2005; Kastelic and Thundathil, 2008).

O componente ambiental influi fortemente no desenvolvimento das características reprodutivas em bovinos. Altas temperaturas podem causar estresse térmico devido a falhas nos mecanismos de termorregulação e dissipação de calor (Armstrong, 1994; Maraia et al., 2008; West, 2003). Os testículos também podem ter seu tamanho reduzido devido à má-nutrição, trauma e inflamação crônica. Por isso, a hipoplasia deve ser diferenciada da atrofia e degeneração adquiridas (Steffen, 1997). Apesar de, em alguns casos, distúrbios reprodutivos serem visualmente de fácil detecção, existem situações em que isto não é possível, sendo realizada a identificação somente no momento da primeira reprodução. O diagnóstico diferencial pode ser feito por exames de biometria testicular e espermiogramas periódicos com a confecção da curva de produção gamética nos touros a partir dos 7 meses de idade, o que é importante para a seleção precoce de reprodutores (Bicudo et al., 2007; Neves et al., 2019).

Medidas testiculares apresentam maior herdabilidade quando comparadas a medidas espermáticas, bem como correlação positiva, como no caso de concentração (0,54) e motilidade (0,36) espermáticas e porcentagem de esperma normal (0,33) (Fonseca, 2018). Além disso, medidas testiculares apresentam correlação positiva com diversas medidas de produção, como por exemplo, peso à desmama (0,56; 0,36), área do músculo longissimus dorsi (0,28) e espessura da gordura na região dorsal (0,17) (Boligon et al., 2010; Lunstra et al., 1988; Smith et al., 1989; Yokoo et al., 2010). Desta forma, tanto a seleção direta para medidas testiculares, quanto a seleção para alguma destas características de produção podem resultar no aumento das medidas testiculares, com um potencial aumento da eficiência reprodutiva.

Devido à importância das medidas testiculares, seja pela grande variedade de processos biológicos envolvidos, ou pela correlação com outras características de

elevada importância, somado à grande frequência de alterações testiculares nos rebanhos, o estudo dos fatores causadores destas alterações é fundamental (Fonseca, 2018). Deste modo, mensurações do perímetro escrotal ao sobreano, estimação da concentração espermática e do volume ejaculado são cruciais na seleção dos reprodutores e um padrão mínimo aceitável destes parâmetros deve ser estabelecido e adotado para a redução dos casos de hipoplasia e, consequentemente, dos prejuízos econômicos nos rebanhos (Steffen, 1997).

1.4.2. Malformação de Pernas e Pés

Outra característica funcional de elevada importância é a malformação de pernas e pés (AP), relacionada à qualidade dos aprumos, que têm a função de sustentar solidamente o corpo do animal e permitir seu fácil deslocamento. Os aprumos podem ser definidos como a direção normal dos membros em toda a sua extensão e são avaliados através das proporções, direções, angulações e articulações dos membros anteriores e posteriores (Mota et al., 2013). Sua má formação pode comprometer o bem-estar, desempenho e longevidade dos animais. Na questão reprodutiva a conformação das pernas e pés é de fundamental importância, tanto para que o macho possa efetuar a monta quanto para que a fêmea possa suportar o reprodutor. Dadas as condições extensivas de produção da maioria dos rebanhos brasileiros, animais que apresentam melhor conformação de pernas e pés são favorecidos, enquanto que, animais com má formação de pernas e pés tendem a perder peso e acabarão sofrendo pressão de seleção por subfertilidade ou por aprumos inadequados (Mota et al., 2013).

A malformação de pernas e pés pode ser avaliada como característica binária, ou seja, apresentar ou não defeito (Vargas et al., 2017). Animais de programas de melhoramento que recebem pontuação 1 (positivo) não podem ser candidatos a receber o CEIP. O diagnóstico se dá pela simples observação de falhas de locomoção e/ou má formação e as avaliações devem ser realizadas por técnicos treinados que atribuem pontuações visuais à estrutura geral dos pés e das pernas em dois momentos diferentes, por exemplo, à desmama e ao sobreano. Recomenda-se que o mesmo técnico seja responsável pelas avaliações e registro de todos os animais de determinado grupo de manejo. Na raça Nelore, as pontuações são geralmente atribuídas com base em padrões de medição e diretrizes especificadas pela

Associação Brasileira dos Criadores de Zebu (ABCZ, http://www.abcz.org.br/) (Vargas et al., 2018).

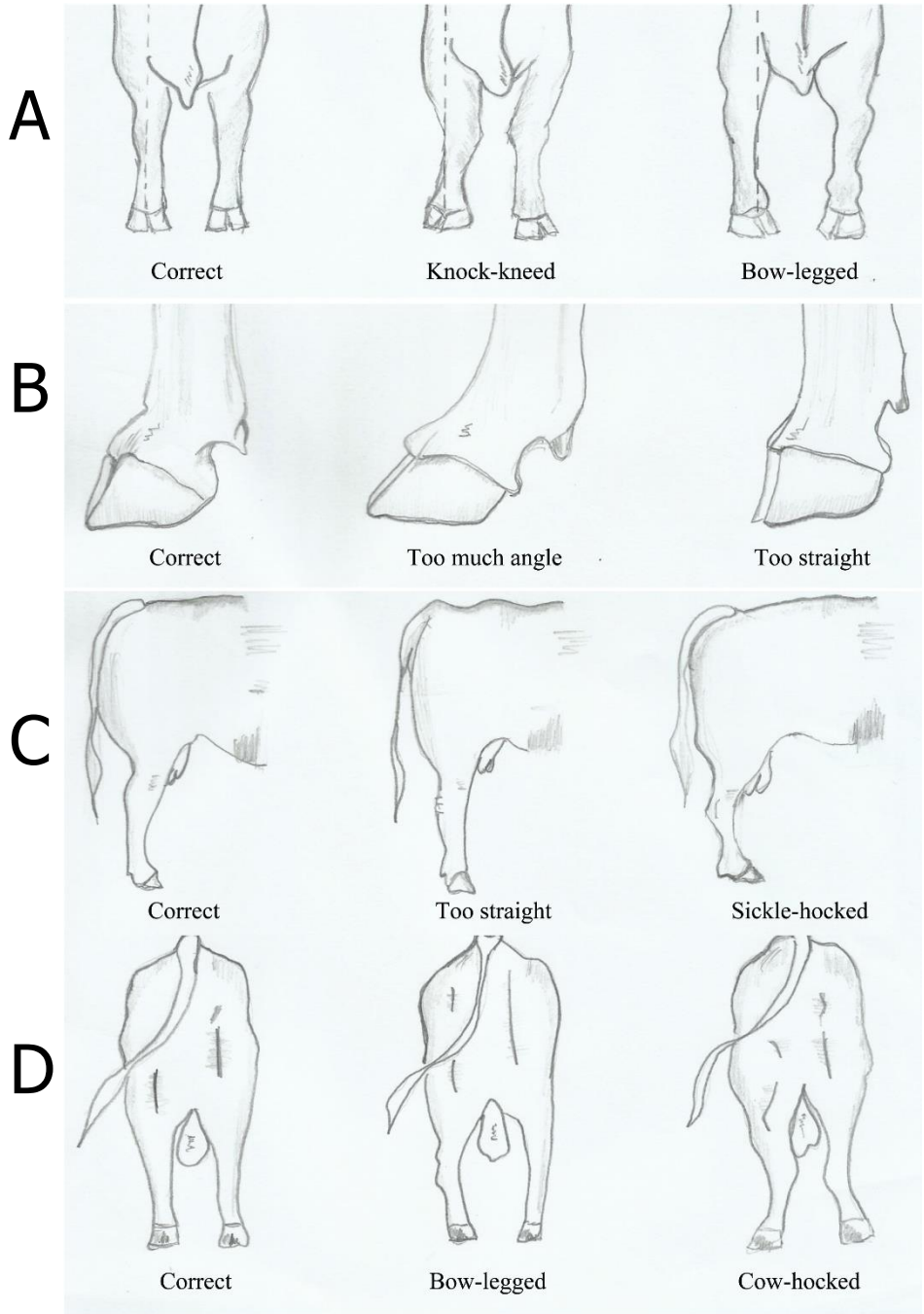

Figura 2. Ilustração da estrutura dos pés e pernas quando vistos de frente, de trás e de lado. Fonte: adaptado de Vargas et al. (2018).

As irregularidades posturais mais comumente observadas nos aprumos dos membros torácicos ou anteriores vistos de frente são: elevada ou reduzida distância entre joelhos, denominadas de aprumos cambotas (*knock-kneed*) e cambaios (*bow-legged*) (Figura 2.A), respectivamente, assim como elevada ou reduzida distância

entre as pinças dos cascos, denominadas de pinças para fora e pinças para dentro. Também podem ocorrer desvios nos boletos, para fora ou para dentro.

Sobre a angulação dos cascos, talões baixos e quartelas arqueadas causam cascos “achinelados”, enquanto que talões altos e quartelas fortes causam cascos “encastelados” (Figura 2.B). Vistos de lado, aprumos dos membros posteriores irregulares não tem um efeito ideal de flexão e absorção de choque. A elevada ou reduzida angulação do jarrete (articulação tíbio-társica) são chamadas de perna perna reta (*straight-legged*) e em foice (*sickle hocked*), respectivamente (Figura 2.C). Vistos por trás, os desvios mais comumente observados nos aprumos dos membros pélvicos ou posteriores são a elevada ou reduzida distância entre jarretes, denominadas “jarrete de vaca” (cow hocked) quando os jarretes são rodados para dentro e os cascos rodados para fora, e “pernas arqueadas” (*bow-legged*) quando as pernas são largas nos jarretes, mas as patas estão viradas para dentro (Figura 2.D). Patas longas ou excessivamente curtas podem indicar ângulos de metacarpos excessivos ou insuficientes, levando a um crescimento ou desgaste excessivos.

1.4.3. Estudo de Associação Genômica Amplo

Métodos de genotipagem de baixo custo combinados com a descoberta dos marcadores SNP (Single Nucleotide Polymorphism) tornaram possível uma abordagem experimental para o mapeamento de variantes causais, também conhecidas como Quantitative Trait Loci (QTL), com o emprego de marcadores densos distribuídos ao longo do genoma (Meuwissen et al., 2001). Desta forma, a seleção de animais para fins reprodutivos baseada em métodos tradicionais, que faziam uso de informações fenotípicas e de pedigree, deu lugar à seleção genômica que utiliza informações genotípicas de marcadores moleculares distribuídos ao longo dos cromossomos, tornando possível estimar com precisão o valor genético até mesmo de animais que não têm registro fenotípico próprio e nenhuma progênie.

O Estudo de Associação Genômica Amplo (Genome-Wide Association Study – GWAS) explora o conceito de desequilíbrio de ligação (LD), uma associação não aleatória a nível de população entre alelos em diferentes loci ou entre SNPs e QTLs (Hayes, 2013), que surge porque grandes segmentos cromossômicos idênticos (haplótipos) são herdados por múltiplos descendentes de ancestrais comuns. O

tamanho efetivo da população (N_e) é a principal causa de LD em rebanhos de animais de produção (Goddard and Hayes, 2009), embora o LD também possa ser causado por diferentes eventos (ex.: migração, mutação, seleção, etc.). Já a quebra de LD é primariamente dirigida pela recombinação. Em bovinos, aproximadamente um evento de recombinação ocorre por cromossomo por geração (Arias et al., 2009). Desta forma, LD entre loci próximos tenderão a se manter e LD entre loci distantes não persistem ao longo das gerações devido à baixa e alta probabilidades de recombinação, respectivamente.

A importância do LD se dá pelo auxílio na interpretação dos efeitos de marcadores em estudos de associação. A estimativa do efeito de marcador se refere a uma associação indireta entre o SNP e o fenótipo, que não tem necessariamente uma relação causal. Na verdade, o que é estimado é uma associação direta entre QTL e fenótipo captada pelo SNP via medida de LD. Diferentes medidas de LD são propostas pela literatura (Qanbari, 2020). Na quantificação de LD desenvolvida por Lewontin and Kojima (1960), "D" é a diferença entre o produto das frequências gaméticas em fase de ligação pelo produto das frequências gaméticas em fase de repulsão. Esta medida é muito sensível à frequência alélica, não sendo adequada para comparar LD de diferentes sítios. A medida r^2, preferida em estudos de associação, é indicativa da correlação que um marcador pode ter com o gene de interesse, em outras palavras ela mede a proporção da variação causada por alelos do QTL que é explicada pelos marcadores (Bohmanova et al., 2010; Hill and Robertson, 1968).

O Manhattan plot (Figura 3) é um gráfico usado para apresentar os resultados de estudos de GWAS. Ele mostra a posição dos SNPs nos cromossomos e sua importância para a característica estudada. Os picos no gráfico representam regiões com efeito pronunciado para a característica de interesse (Utsunomiya et al., 2013).

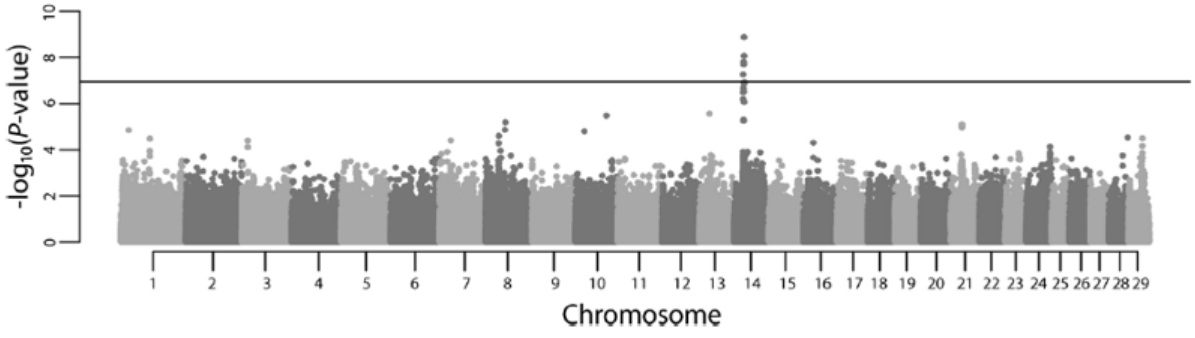

Figura 3. Exemplo de Manhatan plot. Fonte: adaptado de Utsunomiya et al. (2013).

Dependendo do critério utilizado para avaliar a importância dos marcadores, nem sempre os picos são os mesmos. No geral, há uma coincidência nos picos ao selecionar com base no p-value ou na proporção da variância genética explicada pelos SNPs (Lee et al., 2013). Mas, em alguns casos, SNPs de maior efeito não explicam a maior proporção da variância genética dos dados, porque a variância genética é uma função não só do efeito do SNP mas também da frequência alélica dos SNPs. Assim, se um SNP está próximo de ser fixado, ou seja, tem uma *minor allele frequency* (MAF) muito baixa, possivelmente ele vai tender a explicar uma variância aditiva menor mesmo que ele tenha um efeito pronunciado.

Diferentes métricas são usadas para avaliar a importância dos SNPs em um GWAS, como a proporção da variância genética, o p-valor, o Bayes Factor e a Posterior Inclusion Probability (Vidal et al., 2005; Johnson,2013; Saatchi et al., 2013). Estudos de meta-análise em GWAS são úteis para validar regiões significativas do genoma e produzir estudos de associação mais precisos (Melo et al., 2018), especialmente para fenótipos complexos, além de identificar estratégias de seleção genômica que possam funcionar entre raças (Fortes et al., 2013).

1.4.4. Modelagem em GWAS

No GWAS, uma amostra de animais com registros para uma característica de interesse é testada com um painel de marcadores distribuídos pelo genoma para detectar associações estatísticas entre a característica e qualquer um dos marcadores (Goddard and Hayes, 2009). Dentre os métodos estatísticos que exploram associações entre marcadores e QTLs, o mais simples é o de associação do genoma usando regressão simples de marcador único ou método de quadrados mínimos ordinários (Ordinary Least Squares - OLS). Este modelo assume que o marcador afetará a característica apenas se ele estiver em LD com o suposto QTL (Resende et al., 2012), e pode ser representado matricialmente pela equação:

$$y = Wb + Xg + e,$$

onde y é um vetor de fenótipos, W é a matriz de incidência para os efeitos fixos dos fenótipos, b é um vetor de efeitos fixos (ex.: média, efeitos de estrutura populacional e idade), X é a matriz de incidência dos efeitos de marcador, g é o efeito do marcador, e e é um vetor de resíduos $e_{ij} \sim N(0, \sigma_e^2)$, onde σ_e^2 é a variância do erro. Neste modelo

o efeito de marcador é tratado como fixo e o modelo é aditivo, de modo que as soluções representam o efeito de substituição alélica do alelo “A” pelo alelo “B”, ou vice-versa dependendo de como os genótipos são codificados (Hayes, 2013). A solução para média e para os efeitos de substituição alélica são obtidas pela resolução das seguintes equações:

$$\begin{bmatrix}\hat{\mu}\\ \hat{g}\end{bmatrix} = \begin{bmatrix}W'W & W'X\\ X'W & X'X\end{bmatrix}^{-1}\begin{bmatrix}W'y\\ X'y\end{bmatrix}$$

Com um número suficiente de observações é possível estimar os efeitos dos marcadores aplicando o OLS. Uma desvantagem desta abordagem é que um grande número de SNPs pode estar em LD com o mesmo QTL. De modo que pode ser difícil de definir a região contendo uma mutação verdadeira pela observação de vários SNPs significativos abrangendo uma ampla região do genoma (Pryce et al., 2010). Por isso, a utilização do modelo de regressão simples para estimar os efeitos de SNPs únicos, haplótipos ou regiões únicas do genoma nas análises é questionável (Hayes, 2013).

Ao concluírem que no método OLS os marcadores significativos têm efeitos superestimados, Goddard et al. (2009) propuseram uma abordagem integrada para estimar todos os SNPs simultaneamente. No entanto, o problema ao modelar marcadores múltiplos é que o número de marcadores é quase sempre maior que o número de observações, o que gera um problema estatístico pela falta de graus de liberdade suficientes para estimar os parâmetros. Neste caso, é necessário usar algum parâmetero de regularização, ou alguma informação a priori.

1.4.4.1. Métodos Frequentistas

Na regressão de cumeeira (*Ridge Regression*) a estimativa dos efeitos dos marcadores pode ser obtida pela solução do seguinte sistema de equações:

$$\begin{bmatrix}\hat{\mu}\\ \hat{g}\end{bmatrix} = \begin{bmatrix}1_n{}'1_n & 1_n{}'X\\ X'1_n & X'X + I\lambda\end{bmatrix}^{-1}\begin{bmatrix}1_n{}'y\\ X'y\end{bmatrix}$$

Neste sistema é adicionado o parâmetro de regularização $(I\lambda)$ na diagonal da matriz $X'X$ para conseguir obter uma inversa única. Sobre o valor a ser utilizado para λ, deve-se lembrar que quanto maior este valor, maior o “*shrinkage*” das estimativas dos efeitos dos marcadores. Isso reduz o erro de predição, porém, aumenta o “viés”

destas estimativas em direção à zero. Neste caso, menor importância seria atribuída à informação contida nos dados e mais à uma fonte de informação externa (λ).

O método Ridge Regression BLUP (RRBLUP) aplica a regressão de cumeeira usando a derivação BLUP em que considera o efeito do marcador como aleatório. Neste método é feita uma regressão múltipla para todos os SNPs de forma simultânea considerando-os como efeitos aleatórios do modelo:

$$y = \mu + WXg_i + e$$

Agora g é um vetor dos efeitos dos SNPs estimado de forma simultânea. Num RRBLUP a suposição mais comum é que os efeitos dos SNPs são derivados de uma distribuição normal (Meuwissen et al., 2001). Então, $g \sim N\left(0, I\sigma_g^2\right)$ onde I é uma matriz identidade e existe uma σ_g^2 para cada SNP. O λ a ser adicionado neste caso é uma função da variância residual pela variância dos marcadores ($\lambda = \sigma_e^2/\sigma_g^2$). Neste método, quando o número de marcadores é muito grande a regressão dos efeitos dos SNPs em direção a zero é bem forte, introduzido algum viés, o que torna difícil a sua utilização para a obtenção de estimativas de marcadores de grande efeito (de los Campos et al., 2013).

1.4.4.2. Métodos Bayesianos

Outra alternativa para estimar efeitos de marcadores múltiplos é utilizar métodos Bayesianos. Segundo o Teorema de Bayes (Bayes, 1764), a probabilidade a *posteriori* dos parâmetros (θ) condicional aos dados (y) é proporcional à probabilidade dos dados (y) condicional aos parâmetros (θ) (função de verossimilhança dos dados) multiplicada pela probabilidade *a priori* dos parâmetros (θ):

$$P(\theta|y) \propto P(y|\theta)P(\theta)$$

Em comparação aos métodos frequentistas temos que: (1) todos os parâmetros do modelo são considerados como efeitos aleatórios, (2) é utilizada uma fonte de informação externa aos dados (*a priori*), (3) é obtida uma distribuição de probabilidade dos parâmetros ao invés de um único valor. Nos métodos Bayesianos, o conhecimento da *priori* pode ainda ser usado como um mecanismo de atualização de informação ou atualização do conhecimento, ou seja, as distribuições *a posteriori* tornam-se

distribuições *a priori* em novos estudos. Sendo assim, pode-se sair de uma distribuição *a priori* não informativa e chegar a uma *priori* informativa (Silva et al., 2013).

O Amostrador de Gibbs é o método mais utilizado para se estimar a distribuição *a posteriori*. Neste, um vetor de parâmetros de interesse recebe valores iniciais aleatórios (chute inicial) e então são realizadas amostragens de cada parâmetro, via processo iterativo n vezes, em função de todos os valores dos outros parâmetros em função dos dados (Casella and George, 1992). Definindo n como o número total de iterações, se $n \to \infty$, a propriedade da cadeia de Markov garante que após descartar algumas iterações iniciais, os valores gerados para um determinado parâmetro, θ_k, são caracterizados como amostras de sua distribuição marginal *a posteriori*, $P(\theta_k|y)$ (Silva et al., 2013). É importante frisar que o erro de Monte Carlo é inversamente proporcional ao tamanho da cadeia, porém, problemas de delineamento, de coleta de dados, de especificação do modelo, entre outros, não são resolvidos aumentando o tamanho da cadeia amostral.

A convergência da cadeia para a distribuição de interesse depende de uma série de fatores, dentre os quais os "chutes iniciais". Como estes chutes são aleatórios, utiliza-se um período de aquecimento ou "*burn-in*" que representa o descarte das informações entre o chute inicial e o momento em que os valores das iterações começam a se estabilizar. Também se utiliza o parâmetro intervalo de amostragem, ou "*thin*", para a obtenção de amostragens independentes, visto que amostras sucessivas apresentam correlação serial. A Tabela 1 traz exemplos das distribuições *a priori* assumidas em alguns dos métodos Bayesianos.

Table 1. Métodos Bayesianos quanto à distribuição *a priori* assumida para os marcadores.

Método	Priori condicional	Prioris "variâncias"	Priori marginal
Bayesian RR	$p(\beta_j\|\sigma^2_\beta)$ Normal	σ^2_β: Chi-quadrado escalonada invertida	$p(\beta_j)$ Normal
Bayesian LASSO	$p(\beta_j\|\sigma^2_\varepsilon,\tau^2_j)$ Normal	σ^2_ε: Chi-quadrado escalonada invertida τ^2_j: exponencial	$p(\beta_j\|\lambda)$ Exponencial dupla
Bayes A	$p(\beta_j\|\sigma^2_{\beta_j})$ Normal	$\sigma^2_{\beta_j}$: Chi-quadrado escalonada invertida	$p(\beta_j\|v,S^2)$ *t*-escalar

Na Figura 4 estão representadas as curvas das distribuições a priori para os efeitos de marcadores assumidos em cada um dos métodos listados na Tabela 1. Pode-se observar que o Bayesian LASSO tem maior densidade em zero e caudas mais longas, assim, é mais provável encontrar algum marcador com efeito maior neste método em comparação ao Bayes A.

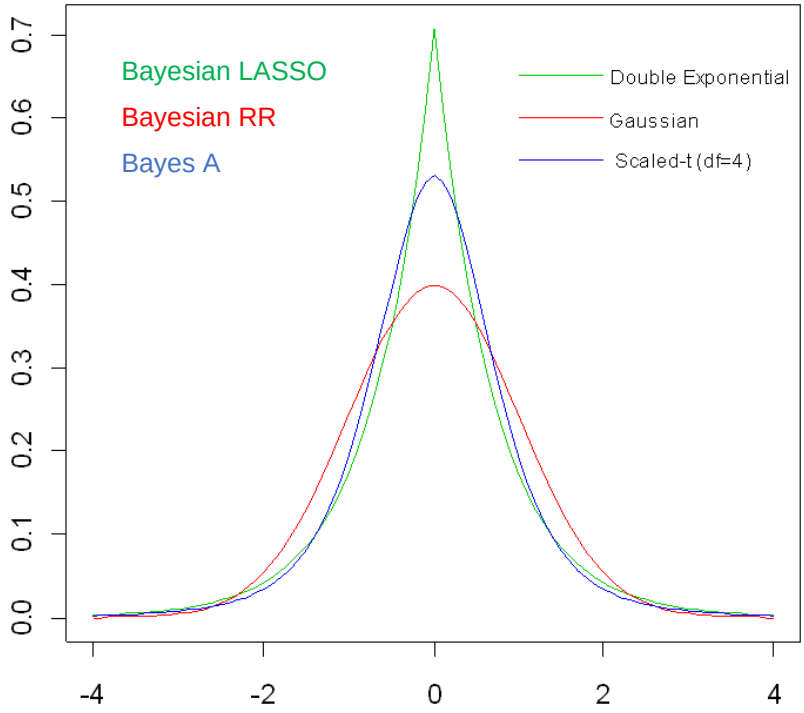

Figura 4. Distribuições *a priori* nos métodos Bayesian RR, Bayesian LASSO e Bayes A.

Em outro método, Bayes B, uma fração p de marcadores tem efeito nulo e uma distribuição *t*-escalar é assumida para os demais $(1-p)$, o que faz com que a curva fique mais achatada permitindo capturar marcadores com efeito mais pronunciado. No método Bayes C uma fração p de marcadores também tem efeito nulo e uma distribuição normal é assumida para os demais $(1-p)$ (Habier et al., 2011; Legarra et al., 2015; van den Berg et al., 2013).

Existem inúmeros métodos Bayesianos disponíveis para estimar parâmetros na seleção genética assistida por marcadores. Tantos que, Gianola et al. (2009), cunharam o termo “alfabeto Bayesiano” para se referirem a estes métodos. Vale resaltar que, o que os diferencia é a *priori* assumida para os efeitos dos marcadores (de los Campos et al., 2013). Uma crítica em relação à utilização dos métodos Bayesianos para estimar os efeitos de um grande número de marcadores tendo poucos animais genotipados é que a *priori* acaba tendo importância muito grande.

Nestes casos, com uma *priori* mais informativa, ou de maior peso, o resultado final para o parâmetro de interesse se aproxima da *priori*. Por outro lado, com uma *priori* menos informativa, próxima de uma distribuição uniforme, o resultado tende a ser dominado pelos dados (Colombani et al., 2013; Gianola et al., 2009; van den Berg et al., 2013).

1.4.4.3. Multiple-Step Genomic Selection

Os métodos bayesianos podem ser usados como uma ferramenta estatística para combinar diferentes fontes de informações genômicas no método *"multiple-step"*. Inicialmente as metodologias aplicadas a predição genômica permitiam utilizar apenas informações de animais com fenótipo (ou pseudo-fenótipo) e genótipo. Entretanto, na prática, existem animais com informação de genótipo e sem fenótipo ou com fenótipo e sem genótipo. Dessa forma, a utilização do método "*multiple-step*" é uma alternativa que visa usar toda a informação disponível, tanto de genótipo quanto de fenótipo, para estimar o mérito genético do animal em várias etapas.

Num primeiro passo busca-se predizer valor genético estimado (EBV) dos animais pelo método BLUP clássico, utilizando o parentesco baseado no pedigree.

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \alpha A^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

Em seguida são calculados os pseudo-fenótipos utilizando tanto o EBV desregredido (sem a média dos pais) quanto o *Daughter Yield Deviation* (DYD), que é a informação das produções das filhas ajustadas para os efeitos fixos e EBVs das matrizes combinadas num único valor. A remoção do efeito da média dos pais (PA) como parte do processo de desregressão é justificada por dois motivos. Primeiro, porque alguns animais podem ter EBV sem informação própria ou de sua progênie e, portanto, não podem contribuir de forma útil para a predição genômica. Segundo, porque se algum dos pais estiver segregando um efeito principal, cerca de metade da prole herdará o alelo favorável e os outros herdarão o alelo desfavorável. No entanto, o EBV de toda a progênie será regredido em direção à média dos pais (Garrick et al., 2009). O uso de EBVs desregredidos evita a dupla contagem (*double-counting*) de informações e a dupla redução (*shrinkage*) dos valores genômicos diretos (DGV) além de contribuir para a redução do viés e aumento da acurácia dos DGVs em comparação com os EBVs (de Oliveira et al., 2018).

O terceiro passo do método *multiple-step* é utilizar os pseudo-fenótipos calculados no segundo passo para estimar os efeitos dos marcadores ou, a depender da formulação do modelo, calcular mérito genético utilizando os genótipos.

Finalmente, no último passo os diferentes componentes associados ao mérito genético predito dos animais são combinados numa predição única (*Genomic Estimated Breeding Value* – GEBV). Se necessário, pode-se subtrair valores referentes à prova genômica dos pais (PI) para evitar *double-counting* (Misztal et al., 2020).

$$GEBV = W_1 PA + W_2 DGV - W_3 PI$$

Uma crítica ao método de *multiple-step* é a possibilidade de viés nas estimativas dos méritos genéticos pois o EBV de um animal contém informações de outros indivíduos, embora nem todos os SNPs sejam compartilhados entre eles. Um dos principais problemas está em subestimar as tendências genéticas pois este método tende a penalizar animais selecionados genomicamente subestimando o mérito genético dos candidatos a seleção jovens (Misztal et al., 2020).

1.4.4.4. Single-step Genomic BLUP (ssGBLUP)

O método "*single-step* GBLUP" (ssGBLUP) é mais robusto a este tipo de viés pois utiliza todas as informações disponíveis, de pedigree, de fenótipos e de SNPs, num único passo para se obter as predições dos méritos genéticos (Legarra et al., 2014; VanRaden et al., 2012). A principal vantagem deste método é a capacidade de incorporar fenótipos de animais não genotipados, aparentados e não aparentados, diretamente em uma abordagem do tipo BLUP sem computar pseudo-fenótipos. Isto é possível pela solução de um sistema de equações no qual a matriz A^{-1} é substituída pela matriz H^{-1}.

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \alpha H^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

Onde:

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

A matriz H^{-1} é uma combinação da inversa da matriz de parentesco (A^{-1}) com uma outra matriz que decompõe uma parte dos animais que têm fenótipo e não tem genótipo ($0s$) e uma parte central ($G^{-1} - A_{22}^{-1}$) que são os animais que apresentam genótipo e que estão na matriz de parentesco clássica (Aguilar et al., 2010; Christensen and Lund, 2010). A matriz de parentesco genômico é subtraída da matriz de parentesco clássica para evitar *double-counting* da informação de parentesco destes animais.

Um ponto importante na implementação deste método, por estar utilizando matrizes genômica e de pedigree, é buscar a compatibilidade entre estas duas matrizes pois nem sempre as bases genéticas são as mesmas. Enquanto os parentescos genômicos respondem indiretamente por todos os ancestrais, mas têm uma escala arbitrária dependendo das frequências alélicas, o parentesco clássico tem uma escala bem definida, mas são limitados pela profundidade e integridade do pedigree (Misztal et al., 2020). Deve-se buscar a compatibilidade em casos de erro de pedigree, quando as populações base são geneticamente muito distantes ou no caso de muitas informações faltantes na matriz de parentesco clássica. O software BLUPF90 (Misztal et al., 2018) oferece opções para tratar este problema e garantir que as matrizes G e A sejam compatíveis.

O método ssGBLUP, desenvolvido para implementação da seleção genômica, modela todos os marcadores simultaneamente e faz um forte *shrinkage* dos seus efeitos em direção a zero. Isto equivale a pressupor a *priori* o modelo poligênico para as características, o que acaba sendo uma desvantagem ao tentar mapear QTLs com maior efeito. As soluções dos efeitos dos SNPs ($\hat{u}$) são estimadas por backsolving, após se obter o mérito genético predito dos animais ($\hat{a}_g$) pela resolução do sistema de equações.

$$\hat{u} = \frac{\sigma_u^2}{\sigma_a^2} DZ'G^{-1}\hat{a}_g$$

O ssGBLUP equivale a usar D como uma matriz identidade, que considera o mesmo peso a priori para os efeitos de todos os marcadores. Entretanto é possível utilizar uma ponderação específica para cada marcador ($D = diag(W_i)$), dando maior peso para os marcadores que explicam a maior parte da proporção da variância da característica e regredindo mais em direção a zero os marcadores que explicam pouco

da variabilidade da característica. Diferentes formas de se calcular os elementos da diagonal da matriz D são relatados na literatura (VanRaden, 2008; Wang et al., 2012). A consequência de se ponderar diferentemente os SNPs é uma melhor resolução no mapeamento. Assim, os picos relacionados aos QTLs podem ser identificados de forma mais clara.

O método ssGBLUP com marcadores poderados ou "Weighted single-step GBLUP" (WssGBLUP) (Wang et al., 2012) é uma alternativa à falta de dados suficientes para GWAS, ou seja, quando muitos animais são fenotipados mas apenas uma proporção reduzida deles é genotipada, situação comum em sistemas de produção de gado comercial (Vargas et al., 2018). Ele tenta reduzir problemas comuns como a falta de poder na detecção e a baixa resolução do mapeamento de QTLs sem aumentar os custos de genotipagem. Entretanto vale ressaltar que, com muitos dados a diferença entre os resultados de ssGBLUP e WssGBLUP não é tão grande (Lourenco et al., 2017).

1.4.5. Imputação de Genótipos

A imputação de genótipos é utilizada para predizer genótipos faltantes ou quando amostras de uma mesma população são genotipadas com painéis de diferentes densidades (ex.: painéis de 50k e 700k). Isto é possível porque longos segmentos cromossômicos são herdados de ancestrais comuns. Assim, uma vez que ancestrais comuns são genotipados com painéis de maior densidade, pode-se utilizar painéis menos densos para genotipar as gerações mais recentes e completar a informação faltante (*"missing"*) via métodos de imputação.

O comprimento médio de segmentos cromossômicos herdados de um ancestral distante t gerações é de $1/(t+1)M$, onde 1 Morgan (M) equivale a aproximadamente 100 milhões de pares de bases (bp). O genoma bovino possui em torno de 3 bilhões de bp com $1M$ por cromossomo e ocorre em média apenas uma recombinação por cromossomo por geração. É possível, inclusive, predizer onde houveram os eventos de recombinação (Sargolzaei et al., 2008).

O primeiro passo da imputação de genótipos é o faseamento da informação genotípica ("*phase genotypes*") para determinar qual haplótipo foi herdado do pai e qual foi herdado da mãe. No caso de genótipos não ambíguos, ou seja, em que todos os SNPs do haplótipo são homozigotos ou apenas um dentre eles é heterozigoto, fica

fácil inferir que um dos alelos veio do pai e o outro da mãe. Entretanto, para os genótipos ambíguos, em que mais de um SNP é heterozigoto, o animal pode apresentar mais que 4 possíveis configurações de haplótipos e isto dificulta o faseamento. Em seguida é construída uma biblioteca de haplótipos de referência com os haplótipos identificados nos ancestrais. E finalmente é feita a predição dos genótipos faltantes nos animais genotipados com painel de menor densidade.

O principal motivo para se fazer imputação é o custo de genotipagem. Dado que o preço dos painéis aumenta em função do aumento da densidade. Outro motivo é a confiabilidade ou acurácia das predições genômicas. Painéis com maior densidade permitem uma maior acurácia das predições genômicas, independente no número de animais na população de referência (VanRaden et al., 2011). Khatkar et al. (2012) calcularam a predição genômica para cinco características num conjunto de dados reais de bovinos leiteiros e mostraram que a acurácia de predição do valor genético num cenário em que parte dos animais foram genotipados com painel de 3 mil SNPs e imputados para um painel de 50 mil SNPs e parte foi genotipada com painel de 50 mil SNPs foi bem semelhante à acurácia no cenário em que todos os animais foram genotipados para 50 mil SNPs.

Em relação aos estudos de associação, o motivo para se imputar genótipos é que painéis mais densos permitem uma maior resolução no mapeamento de QTLs (Marchini et al., 2007). Já nos estudos de LD para diferentes raças, o padrão é mais consistente quando se utiliza painéis mais densos (Pérez O'Brien et al., 2014). Ferenčaković et al. (2013) demonstraram que painéis com apenas 50 mil marcadores superestimavam o nível de endogamia via número de corridas de homozigosidade em segmentos pequenos de bovinos.

A validação da imputação é uma etapa importante tanto para se avaliar o nível de precisão quanto para evitar a presença de ruído nas análises posteriores. Nesta etapa, primeiro se determina quais animais serão genotipados com painel de maior densidade (HD) e quais painéis de baixa densidade serão testados. Então, parte dos animais genotipados com painel HD são testados com diferentes painéis de menor densidade para se identificar a densidade mínima que garantiria uma acurácia de imputação razoável (Carvalheiro et al., 2014). Os animais genotipados para HD são separados em grupos de referência e de validação (aleatoriamente, em sequência, por distanciamento genético, categoria, etc.). As informações do grupo de validação são mascaradas, fingindo que estes foram genotipados com painel de menor

densidade. Para avaliar a eficiência do método, os genótipos imputados no grupo de validação são então comparados com os genótipos observados.

Diferentes estatísticas podem ser calculadas para avaliar a eficiência da imputação, tanto por animal quanto por SNP. Dentre as quais: a porcentagem dos genótipos imputados; a porcentagem dos genótipos corretamente imputados; a correlação entre imputados e observados, conhecida como acurácia de imputação; a taxa média de erro alélico. Vale lembrar que estas estatísticas devem ser calculadas apenas para os genótipos mascarados, ou seja, que foram omitidos no painel de menor densidade e imputados. Dados faltantes no genótipo original também serão imputados, porém como a informação original é desconhecida, estes devem ser desconsiderados no cálculo da eficiência de imputação.

Afim de definir a melhor estratégia de genotipagem dos animais deve-se conhecer os fatores que afetam a acurácia da imputação. São eles: a proporção de genótipos a ser imputada, que é a diferença entre o painel de maior densidade e o de menor densidade (Hickey et al., 2012a); o número de animais genotipados com painel de mais alta densidade, sendo o quanto maior melhor (Druet et al., 2010); o grau de parentesco entre os animais presentes nos grupos de referência e validação, sendo este fator mais expressivo que a densidade do painel e quanto menor o grau de parentesco menor a acurácia de imputação (Carvalheiro et al., 2014); a localização do SNP no cromossomo, pois SNPs localizados nas extremidades tem menor acurácia que aqueles localizados em posições centrais porque os métodos de faseamento dos haplótipos usam janelas e existe pouca informação nas extremidades (Steibel et al., 2013); a MAF, como parte do processo de imputação é baseado em probabilidade dificilmente o software vai indicar um alelo raro como sendo o verdadeiro (Steibel et al., 2013); a estratégia de imputação, pois a acurácia aumenta quando amostras são genotipadas com um painel de densidade intermediária e a imputação é feita em três passos (ex.: 3K → 50K → 800K) – a estratégia vai ser tão mais importante quanto maior o número de animais genotipados nas camadas intermediárias (Khatkar et al., 2012); finalmente, os algoritmos de imputação mostram diferenças na acurácia da imputação por cromossomo, por isso, recomenda-se que o estudo de validação seja realizado em diferentes softwares (Sun et al., 2012).

Dentre os softwares mais utilizados para imputação genômica no momento em que este trabalho é escrito estão o AlphaImpute (Hickey et al., 2012b), Fimpute3.0 (Sargolzaei et al., 2014), findhap.f90 (Vanraden and Sun, 2014), Impute5 (Rubinacci

et al., 2020) e Beagle5.2 (Browning et al., 2018). Para o faseamento de genótipos, o Beagle5.2 se mostrou altamente preciso em amostras típicas de gado de leite (Oget-Ebrad et al., 2022). O FImpute3.0 reconstrói segmentos extremamente longos sem erros, se mostrando um software muito eficiente em termos de acurácia de imputação e de eficiência computacional (Oget-Ebrad et al., 2022; Sargolzaei et al., 2014).

1.4.6. Sequenciamento Completo do Genoma

Variantes genéticas podem ser genotipadas por diferentes tecnologias, incluindo painéis de marcadores de maior ou de menor densidade ou sequenciamento completo do genoma (Whole-Genome Sequencing – WGS) (Tam et al., 2019). A tecnologia de escolha para identificar e entender genes candidatos e vias metabólicas associados com a expressão da característica deve considerar a melhor acurácia de seleção, o número de animais e de marcadores. Tecnologias de sequenciamento completo permitem gerar diariamente grandes volumes de dados com alta qualidade, possibilitando a realização de avaliações abrangentes das variantes nas regiões de interesse usando um método eficiente em termos de tempo e com alto rendimento (Sun et al., 2016).

A sequência genômica está disponível para várias espécies domésticas, incluindo os bovinos, e um grande número de SNPs foram descobertos como subproduto do sequenciamento (Goddard e Hayes, 2009). Para Daetwyler et al. (2014), a caracterização de variantes da sequência causal tornará a seleção genômica muito mais robusta do que o uso de marcadores.Para Druet et al. (2014), o uso de dados de sequenciamento completo em estudos como os de predição genômica pode ser vantajoso porque a acurácia da predição não é mais limitada pela extensão do LD entre os marcadores e as mutações causais que afetam a característica, uma vez que que as mutações causais estão no conjunto de dados.

Os custos da execução de WGS em amostras muito grandes são proibitivos dado o número de indivíduos necessários com o fenótipo e a sequência genômica (Daetwyler et al., 2014). Assim, a genotipagem com painéis SNP de baixa densidade combinados com imputação estatística de genótipos observados nos painéis de referência da população é o método mais utilizado em GWAS (Tam et al., 2019). Uma abordagem econômica para atingir um grande número de indivíduos com dados de sequência é usar um tamanho efetivo da população relativamente pequeno –

resultado da intensa seleção e inseminação artificial. Uma vez que os ancestrais forem sequenciados, os descendentes precisam apenas ser genotipados com painéis SNP densos e os dados imputados para sequência a fim de inferir com precisão sua sequência de genoma, rastreando os grandes segmentos de cromossomos herdados (Daetwyler et al., 2014; Druet et al., 2014).

À medida em que os custos do sequenciamento diminuírem e a quantidade de dados de sequência gerados aumentar, novos paradigmas como armazenamento e análise de dados serão cada vez mais importantes. Segundo Muir et al. (2016), o comportamento relativo de escala dessas tecnologias em evolução impactará a pesquisa genômica no futuro. Para Tam et al. (2019), embora provavelmente a mudança para o WGS seja inevitável com a redução dos custos de sequenciamento, a escolha por painéis SNP ou WGS no GWAS deverá levar em consideração outros fatores, como a confiabilidade da tecnologia de genotipagem, cobertura desejada de variantes genéticas, recursos disponíveis e os objetivos da pesquisa.

1.5. Referências

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

Aragão, A., Contini, E., 2021. O agro no Brasil e no Mundo: uma síntese do período de 2000 a 2020. Embrapa SIRE.

Arias, J.A., Keehan, M., Fisher, P., Coppieters, W., Spelman, R., 2009. A high-density linkage map of the bovine genome. https://doi.org/10.1186/1471-2156-10-18

Armstrong, D. V., 1994. Heat Stress Interaction with Shade and Cooling. Journal of Dairy Science 77, 2044–2050. https://doi.org/10.3168/jds.S0022-0302(94)77149-6

Bayes, T., 1764. An Essay Toward Solving a Problem in the Doctrine of Chances. Philosophical Transactions of the Royal Society of London 53, 370–418.

Bicudo, S., Siqueira, J., Meira, C., 2007. Patologias do sistema reprodutor de touros. Biológico 69, 43–48.

Bohmanova, J., Sargolzaei, M., Schenkel, F.S., 2010. Characteristics of linkage disequilibrium in North American Holsteins, BMC Genomics.

Boligon, A.A., Silva, J.A.V., Sesana, R.C., Sesana, J.C., Junqueira, J.B., Albuquerque, L.G., 2010. Estimation of genetic parameters for body weights, scrotal circumference, and testicular volume measured at different ages in Nellore cattle. Journal of Animal Science 88, 1215–1219. https://doi.org/10.2527/jas.2008-1719

Browning, B.L., Zhou, Y., Browning, S.R., 2018. A One-Penny Imputed Genome from Next-Generation Reference Panels. Am. J. Hum. Genet. 103, 338–348.

Carvalheiro, R., Boison, S.A., Neves, H.H.R., Sargolzaei, M., Schenkel, F.S., Utsunomiya, Y.T., O'Brien, A.M.P., Sölkner, J., McEwan, J.C., Van Tassell, C.P., Sonstegard, T.S., Garcia, J.F., 2014. Accuracy of genotype imputation in Nelore cattle. Genet. Sel. Evol. 46, 1–11. https://doi.org/10.1186/s12711-014-0069-1

Casella, G., George, E.I., 1992. Explaining the Gibbs Sampler. The American Statistician 46, 167. https://doi.org/10.2307/2685208

Christensen, O.F., Lund, M.S., 2010. Genomic prediction when some animals are not genotyped. Genetics Selection Evolution 42, 2. https://doi.org/10.1186/1297-9686-42-2

Colombani, C., Legarra, A., Fritz, S., Guillaume, F., Croiseau, P., Ducrocq, V., Robert-Granié, C., 2013. Application of Bayesian least absolute shrinkage and selection operator (LASSO) and BayesCπ methods for genomic selection in French Holstein and Montbéliarde breeds. Journal of Dairy Science 96, 575–591.

Daetwyler, H.D., Capitan, A., et al., 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat. Genet. 46, 858–865.

de los Campos, G., Hickey, J.M., Pong-Wong, R., Daetwyler, H.D., Calus, M.P.L., 2013. Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. Genetics 193, 327–345. https://doi.org/10.1534/genetics.112.143313

Druet, T., Macleod, I.M., Hayes, B.J., 2014. Toward genomic prediction from whole-genome sequence data : impact of sequencing design on genotype imputation and accuracy of predictions. Heredity (Edinb). 112, 39–47.

Druet, T., Schrooten, C., de Roos, A.P.W., 2010. Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. J. Dairy Sci. 93, 5443–5454. https://doi.org/10.3168/jds.2010-3255

Ferenčaković, M., Sölkner, J., Curik, I., 2013. Estimating autozygosity from high-throughput information: Effects of SNP density and genotyping errors. Genet. Sel. Evol. 45, 42. https://doi.org/10.1186/1297-9686-45-42

Fonseca, P.A. de S., 2018. Investigação de genes candidatos para características reprodutivas e efeito pleiotrópico para características de produção em bovinos.

Fortes, M.R.S., DeAtley, K.L., Lehnert, S.A., Burns, B.M., Reverter, A., Hawken, R.J., Boe-Hansen, G., Moore, S.S., Thomas, M.G., 2013. Genomic regions associated with fertility traits in male and female cattle: Advances from microsatellites to high-density chips and beyond. Animal Reproduction Science 141, 1–19.

Garrick, D.J., Taylor, J.F., Fernando, R.L., 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genetics Selection Evolution 41, 1–8. https://doi.org/10.1186/1297-9686-41-55

Gianola, D., De Los Campos, G., Hill, W.G., Manfredi, E., Fernando, R., 2009. Additive genetic variability and the Bayesian alphabet. Genetics 183, 347–363. https://doi.org/10.1534/genetics.109.103952

Gnoth, C., Godehardt, E., Frank-Herrmann, P., Friol, K., Tigges, J., Freundl, G., 2005. Definition and prevalence of subfertility and infertility. Human Reproduction 20, 1144–1147. https://doi.org/10.1093/humrep/deh870

Goddard, M.E., Hayes, B.J., 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nature Reviews Genetics. https://doi.org/10.1038/nrg2575

Goddard, M.E., Wray, N.R., Verbyla, K., Visscher, P.M., 2009. Estimating Effects and Making Predictions from Genome-Wide Marker Data. Statistical Science 24, 517–529. https://doi.org/10.1214/09-sts306

Habier, D., Fernando, R.L., Kizilkaya, K., Garrick, D.J., 2011. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12:186. https://doi.org/https://doi.org/10.1186/1471-2105-12-186

Hayes, B., 2013. Genome-Wide Association Studies and Genomic Prediction, Genome-Wide Association Studies and Genomic Prediction, Methods in Molecular Biology. Humana Press, Totowa, NJ. https://doi.org/10.1007/978-1-62703-447-0

Hickey, J.M., Crossa, J., Babu, R., de los Campos, G., 2012a. Factors Affecting the Accuracy of Genotype Imputation in Populations from Several Maize Breeding Programs. Crop Sci. 52, 654–663. https://doi.org/10.2135/cropsci2011.07.0358

Hickey, J.M., Kinghorn, B.P., Tier, B., Van Der Werf, J.H., Cleveland, M.A., 2012b. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genet. Sel. Evol. 44, 1–11. https://doi.org/10.1186/1297-9686-44-9

Hill, W.G., Robertson, A., 1968. Linkage disequilibrium in finite populations. Theoretical and applied genetics 38, 226–231. https://doi.org/10.1080/03071848408522227

Johnson, V.E., 2013. Revised standards for statistical evidence. Proceedings of the National Academy of Sciences of the United States of America 110, 19313–19317.

Kastelic, J.P., Thundathil, J.C., 2008. Breeding Soundness Evaluation and Semen Analysis for Predicting Bull Fertility. Reproduction in Domestic Animals 43, 368–373.

Khatkar, M.S., Moser, G., Hayes, B.J., Raadsma, H.W., 2012. Strategies and utility of imputed SNP genotypes for genomic analysis in dairy cattle. BMC Genomics 13, 1–12.

Lee, S.H., Choi, B.H., Lim, D., Gondro, C., Cho, Y.M., Dang, C.G., Sharma, A., Jang, G.W., Lee, K.T., Yoon, D., Lee, H.K., Yeon, S.H., Yang, B.S., Kang, H.S., Hong, S.K., 2013. Genome-Wide Association Study Identifies Major Loci for Carcass Weight on BTA14 in Hanwoo (Korean Cattle). PLoS ONE 8.

Legarra, A., Christensen, O.F., Aguilar, I., Misztal, I., 2014. Single Step, a general approach for genomic selection. Livestock Science 166, 54–65.

Legarra, A., Croiseau, P., Pierre Sanchez, M., Teyssèdre, S., Sallé, G., Allais, S., Fritz, S., Moreno, C.R., Ricard, A., Elsen, J.-M., 2015. A comparison of methods for whole-genome QTL mapping using dense markers in four livestock species. Genetics Selection Evolution 47. https://doi.org/10.1186/s12711-015-0087-7

Lewontin, R.C., Kojima, K., 1960. THE EVOLUTIONARY DYNAMICS OF COMPLEX POLYMORPHISMS. Evolution 14, 458–472. https://doi.org/10.2307/2405995

Lourenco, D.A.L., Fragomeni, B.O., Bradford, H.L., Menezes, I.R., Ferraz, J.B.S., Aguilar, I., Tsuruta, S. and Misztal, I., 2017. Implications of SNP weighting on single-step genomic predictions for different reference population sizes. Journal of animal breeding and genetics, 134(6), 463-471. https://doi.org/10.1111/jbg.12288

Lunstra, D.D., Gregory, K.E., Cundiff, L. V., 1988. Heritability estimates and adjustment factors for the effects of bull age and age of dam on yearling testicular size in breeds of bulls. Theriogenology 30, 127–136. https://doi.org/10.1016/0093-691X(88)90270-1

MAPA - MINISTÉRIO DA AGRICULTURA PECUÁRIA E ABASTECIMENTO, B., 2006. Instrução Normativa no 53, de 27 de setembro de 2006 [WWW Document]. URL http://sistemasweb.agricultura.gov.br/sislegis/action/detalhaAto.do?method=visualizarAtoPortalMapa&chave=735332411 (accessed 10.28.19).

MAPA - MINISTÉRIO DA AGRICULTURA PECUÁRIA E ABASTECIMENTO, B., 1995. Instituir o Certificado Especial de Identificação e Produção – CEIP, para bovinos resultantes de cruzamento planificado e para animais de raças puras com aptidão para corte. Portaria Ministerial no 267, de 04 de Maio de 1995.

Maraia, I.F.M., El-Darawanya, A.A., Fadielc, A., Abdel-Hafezb, M.A.M., 2008. Reproductive performance traits as affected by heat stress and its alleviation in sheep. Tropical and Subtropical Agroecosystems 8, 209–234.

Marchini, J., Howie, B., Myers, S., McVean, G., Donnelly, P., 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. 39, 906–913. https://doi.org/10.1038/ng2088

McEntee, K., 1990. Scrotum, spermatic cord, and testis: proliferative lesions, in: Reproductive Pathology of Domestic Mammals. Academic Press, San Diego, CA, pp. 279–300.

Melo, T.P., Fortes, M.R.S., Bresolin, T., Mota, L.F.M., Albuquerque, L.G., Carvalheiro, R., 2018. Multitrait meta-analysis identified genomic regions associated with sexual precocity in tropical beef cattle. Journal of Animal Science 96, 4087–4099. https://doi.org/10.1093/jas/sky289

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819–1829.

Misztal, I., Lourenco, D., Legarra, A., 2020. Current status of genomic evaluation. Journal of Animal Science 98, 1–14. https://doi.org/10.1093/jas/skaa101

Misztal, I., Tsuruta, S., Lourenco, D., Aguilar, I., Legarra, A., Vitezica, Z., 2018. Manual for BLUPF90 family of programs. university of Georgia, Athens, USA 125.

Mota, L.F.M., Pires, A.V., Bonafé, C.M., Abreu, L.R.A., 2013. Escores visuais corporais como auxílio à seleção em bovinos de corte, Boletim Técnico PPGZOO UFVM. Diamantina.

Moura, G.S., 2013. Estudo clínico-andrológico e genético-molecular da subfertilidade em touros da raça Gir Leiteiro (Bos taurus indicus) no Estado de Minas Gerais. UNIVERSIDADE FEDERAL DE MINAS GERAIS.

Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D.J., Salichos, L., Zhang, J., Weinstock, G.M., Isaacs, F., Rozowsky, J., Gerstein, M., 2016. The real cost of sequencing: scaling computation to keep pace with data generation. Genome Biol. 17, 1–9. https://doi.org/10.1186/s13059-016-0917-0

Neves, H.H.R., Vargas, G., Brito, L.F., Schenkel, F.S., Albuquerque, L.G., Carvalheiro, R., 2019. Genetic and genomic analyses of testicular hypoplasia in Nellore cattle. Plos One 14, e0211159. https://doi.org/10.1371/journal.pone.0211159

Oget-Ebrad, C., Kadri, N.K., Moreira, G.C.M., Karim, L., Coppieters, W., Georges, M., Druet, T., 2022. Benchmarking phasing software with a whole-genome sequenced cattle pedigree. BMC Genomics 23, 1–16. https://doi.org/10.1186/s12864-022-08354-6

Oliveira, H.R., Silva, F.F., Brito, L.F., Guarini, A.R., Jamrozik, J., Schenkel, F.S., 2018. Comparing deregression methods for genomic prediction of test-day traits in dairy cattle. Journal of Animal Breeding and Genetics 135, 97–106. https://doi.org/10.1111/jbg.12317

Pérez O'Brien, A.M., Mészáros, G., Utsunomiya, Y.T., Sonstegard, T.S., Garcia, J.F., Van Tassell, C.P., Carvalheiro, R., da Silva, M.V.B., Sölkner, J., 2014. Linkage disequilibrium levels in Bos indicus and Bos taurus cattle using medium and high density SNP chip data and different minor allele frequency distributions. Livest. Sci. 166, 121–132. https://doi.org/10.1016/j.livsci.2014.05.007

Pryce, J.E., Bolormaa, S., Chamberlain, A.J., Bowman, P.J., Savin, K., Goddard, M.E., Hayes, B.J., 2010. A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length haplotypes. Journal of Dairy Science 93, 3331–3345. https://doi.org/10.3168/jds.2009-2893

Qanbari, S., 2020. On the Extent of Linkage Disequilibrium in the Genome of Farm Animals. Frontiers in Genetics. https://doi.org/10.3389/fgene.2019.01304

Rosa, A.N., Menezes, G.R.O., Egito, A.A., Nogueira, E., 2021. Genética aplicada para pequenos e médios produtores de gado de corte. Embrapa Gado de Corte-Documentos (INFOTECA-E).

Rubinacci S., Delaneau O., Marchini J. 2020. Genotype imputation using the Positional Burrows Wheeler Transform. PLoS Genet 16, e1009049. https://doi.org/10.1371/journal. pgen.1009049

Saatchi, M., Garrick, D.J., Tait, R.G., Mayes, M.S., Drewnoski, M., Schoonmaker, J., Diaz, C., Beitz, D.C., Reecy, J.M., 2013. Genome-wide association and prediction of direct genomic breeding values for composition of fatty acids in Angus beef cattle. BMC Genomics 14, 730. https://doi.org/10.1186/1471-2164-14-730

Sargolzaei, M., Chesnais, J.P., Schenkel, F.S., 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15, 478. https://doi.org/10.1186/1471-2164-15-478

Sargolzaei, M., Scnenkel, F.S., Jansen, G.B., Schaeffer, L.R., 2008. Extent of linkage disequilibrium in Holstein cattle in North America. J. Dairy Sci. 91, 2106–2117. https://doi.org/10.3168/jds.2007-0553

Silva, F.F., Viana, J.M.S., Faria, V.R., Resende, M.D.V., 2013. Bayesian inference of mixed models in quantitative genetics of crop species. Theoretical and Applied Genetics 126, 1749–1761. https://doi.org/10.1007/s00122-013-2089-6

Smith, B.A., Brinks, J.S., Richardson, G. V., 1989. Estimation of genetic parameters among breeding soundness examination components and growth traits in yearling bulls. Journal of animal science 67, 2892–2896. https://doi.org/10.2527/jas1989.67112892x

Steffen, D., 1997. Genetic causes of bull infertility. Veterinary Clinics of North America: Food Animal Practice 13, 243–253. https://doi.org/10.1016/S0749-0720(15)30338-8

Steibel, J.P., Badke, Y.M., Bates, R.O., Ernst, C.W., Schwab, C., Fix, J., Van Tassell, C.P., 2013. Methods of tagSNP selection and other variables affecting imputation accuracy in swine, BMC Genetics.

Sun, C., Wu, X.L., Weigel, K.A., Rosa, G.J.M., Bauck, S., Woodward, B.W., Schnabel, R.D., Taylor, J.F., Gianola, D., 2012. An ensemble-based approach to imputation of moderate-density genotypes for genomic selection with application to Angus cattle. Genet. Res. (Camb). 94, 133–150. https://doi.org/10.1017/S001667231200033X

Sun, X., Sui, W., Wang, X., Hou, X., Ou, M., Dai, Y., Xiang, Y., 2016. Whole-genome re-sequencing for the identification of high contribution susceptibility gene variants in patients with type 2 diabetes. Mol. Med. Rep. 13, 3735–3746. https://doi.org/10.3892/mmr.2016.5014

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., Meyre, D., 2019. Benefits and limitations of genome-wide association studies. Nat. Rev. Genet. 20, 467–484. https://doi.org/10.1038/s41576-019-0127-1

Utsunomiya, Y.T., do Carmo, A.S., et al., 2013. Genome-wide association study for birth weight in Nellore cattle points to previously described orthologous genes affecting human and bovine height. BMC Genetics 14, 52. https://doi.org/10.1186/1471-2156-14-52

van den Berg, I., Fritz, S., Boichard, D., 2013. QTL fine mapping with Bayes C(π): a simulation study. Genetics Selection Evolution 45. https://doi.org/10.1186/1297-9686-45-19

VanRaden, P.M., 2008. Efficient methods to compute genomic predictions. Journal of dairy science 91, 4414–4423. https://doi.org/10.3168/jds.2007-0980

VanRaden, P.M., O'Connell, J.R., Wiggans, G.R., Weigel, K.A., 2011. Genomic evaluations with many more genotypes. Genet. Sel. Evol. 43, 1–11. https://doi.org/10.1186/1297-9686-43-10

VanRaden, P.M., Wright, J.R., Cooper, T.A., 2012. Adjustment of selection index coefficients and polygenic variance to improve regressions and reliability of genomic evaluations. J. Dairy Sci. 95, 446–447.

Vanraden, P.M., Sun, C., 2014. Proceedings, 10 th World Congress of Genetics Applied to Livestock Production Fast Imputation Using Medium-or Low-Coverage Sequence Data, Proceedings of the World Congress on Genetics Applied to Livestock Production. World Congress on Genetics Applied to Livestock Production.

Vargas, G., Neves, H.H.R., Camargo, G.M.F., Cardoso, V., Munari, D.P., Carvalheiro, R., 2018. Genome-wide association study and functional analysis of feet and leg conformation traits in Nellore cattle. Journal of Animal Science 96, 1617–1627. https://doi.org/10.1093/jas/sky079

Vargas, G., Neves, H.H.R., Cardoso, V., Munari, D.P., Carvalheiro, R., 2017. Genetic analysis of feet and leg conformation traits in Nelore cattle. Journal of Animal Science 95, 2379–2384. https://doi.org/10.2527/jas2016.1327

Vidal, O., Noguera, J.L., Amills, M., Varona, L., Gil, M., Jiménez, N., Dávalos, G., Folch, J.M., Sánchez, A., 2005. Identification of carcass and meat quality quantitative trait loci in a Landrace pig population selected for growth and leanness. Journal of Animal Science 83, 293–300. https://doi.org/10.2527/2005.832293x

Wang, H., Misztal, I., Aguilar, I., Legarra, A., Muir, W.M., 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genetics Research 94, 73–83. https://doi.org/10.1017/S0016672312000274

West, J.W., 2003. Effects of heat-stress on production in dairy cattle. Journal of Dairy Science 86, 2131–2144. https://doi.org/10.3168/jds.S0022-0302(03)73803-X

Yokoo, M.J., Lobo, R.B., Araujo, F.R.C., Bezerra, L.A.F., Sainz, R.D., Albuquerque, L.G., 2010. Genetic associations between carcass traits measured by real-time ultrasound and scrotal circumference and growth traits in Nelore cattle. Journal of Animal Science 88, 52–58. https://doi.org/10.2527/jas.2008-1028

CAPÍTULO 2 - Testicular hypoplasia in Nellore cattle: genetic analysis and functional analysis of genome-wide association study results[1]

Abstract – Characterized by the incomplete development of the germinal epithelium of the seminiferous tubules, Testicular Hypoplasia (TH) leads to decreased sperm concentration, increased morphological changes in sperm and azoospermia. Economic losses resulting from the disposal of affected bulls reduce the efficiency of meat production systems. A genome-wide association study and functional analysis were performed to identify genomic windows and the underlying positional candidate genes associated with TH in Nellore cattle. Phenotypic and pedigree data from 207,195 animals and genotypes (461,057 Single Nucleotide Polymorphism, SNP) from 17,326 sires were used in this study. TH was evaluated as a binary trait measured at 18 months of age. A possible correlated response on TH resulting from the selection for scrotal circumference was evaluated by using a two-trait analysis. Thus, estimated breeding values were calculated by fitting a linear-threshold animal model in a Bayesian approach. The SNP effects were estimated using the weighted single-step genomic BLUP method. Twelve non-overlapping windows of 20 adjacent SNP that explained more than 1% of the additive genetic variance were selected for candidate gene annotation. Functional and gene prioritization analysis of the candidate genes identified genes (KHDRBS3, GPX5, STAR, ERLIN2) which might play an important role in the expression of TH due to their known roles in the spermatogenesis process, synthesis of steroids and lipid metabolism.

Key words: *Bos taurus indicus*, candidate genes, fertility, prioritization, sperm, testis.

[1] Este capítulo corresponde ao artigo científico publicado na revista Journal of Animal Breeding and Genetics, 00, 1–13, 2022. https://doi.org/10.1111/jbg.12747

2.1. Introduction

In most beef producing countries reduced fertility is the mayor restriction to efficient production and sound economic results. Brazil is not foreign to such problem and, although it occupies a prominent position in the world beef market (Aragão and Contini, 2021), the national herd still has great potential for genetic improvement. The Nellore breed comprises 80% of the national herd but the system cannot produce enough genetically superior bulls to meet its annual replacement demand. Even if all bulls with CEIP (Special Certificate of Identification and Production) (MAPA, 2006) were used, as well as all young bulls in the selection programs, with paternity check, there would still be a 48% deficit of replacement bulls (Rosa et al., 2021). In breeding programs authorized to issue CEIP, every year between 20% and 30% of the best animals are evaluated at weaning and yearling to become candidates for certification. Besides being genetically superior, the animals potentially eligible to receive the CEIP must not present serious morphological and functional defects or disorders. However, many males do not receive the certificate due to morphological problems identified in the final inspection that leads to their disqualification. These disqualifying morphological traits are a major contributor to the deficit of breeding bulls in the country.

Testicular hypoplasia (TH) is one of the main morphological problems observed in Nellore bulls in CEIP programs, with an incidence of approximately 5% in yearlings (Neves et al., 2019). This amounts to a large number of certificate candidates being excluded from the program even though they have high genetic merit for other traits of economic importance (e.g., reproductive, growth and carcass quality traits). This genetic alteration is characterized by the incomplete development of the germinal epithelium of the seminiferous tubules due to an inadequate number of germ cells in the testis that leads to a decrease in sperm concentration and an increase in sperm morphological alterations, as well as azoospermia (McEntee, 1990). TH can be unilateral or bilateral, partial or total (Steffen, 1997). Bulls with unilateral hypoplasia, partial or total, or partial bilateral hypoplasia, may have libido and reproductive ability, making it difficult to identify them in a herd (Moura, 2013). Due to the higher frequency of TH being of the partial unilateral type, most bulls are subfertile, passing on the undesirable condition to their descendants (Bicudo et al., 2007). The subfertility is characterized by a reproductive success lower than the

general average of the population and can affect 20-40% of bulls in some cattle herds (Gnoth et al., 2005; Kastelic and Thundathil, 2008).

Genome-wide association analyses (GWAS) are a powerful approach to identify candidate regions associated with diseases from genetic deficiencies. Here, a GWAS was performed using the weighted single-step GBLUP method (WssGBLUP) (Wang et al., 2012), which allows combining all available information, even the one originated in non-genotyped animals.

A better understanding of the genetic architecture of TH can help reduce the incidence of this problem in the Nellore herd and increase their productive and reproductive efficiency. This study aimed to: (1) estimate the genetic parameters of TH, including the genetic correlation between TH and scrotal circumference (SC); (2) estimate the genetic trends of these traits in Nellore cattle; (3) find windows of adjacent SNPs that are significantly associated with TH using the WssGBLUP methodology and (4) identify candidate genes supported by functional evidence for a potential role in the incidence of this trait.

2.2. Material and methods

Approval from the Welfare and Animal Use Committee was not required for this study as the data was obtained from the Nellore Alliance database of phenotypic and genotypic records of Nellore cattle.

2.2.1. Phenotypic and pedigree data

Phenotypic and genealogical data of animals with TH records born between 2001 and 2017 were obtained from the Nellore Alliance database (GenSys, 2019). TH measurements were performed at yearling and assigned on a binary scale (1 for affected animals and 0 for non-affected) by trained technicians. Contemporary groups (CG) were defined by concatenating the effects of year and season of birth, farm at weaning and yearling, management group at weaning and yearling, and evaluation date at yearling. Only CG with at least 10 records and with variability for the trait were used. Connectedness among CG was verified using the AMC software (Roso and Schenkel, 2006) and only CG with at least 10 genetic links to other CG were kept for analysis. After editing the data, 207,195 phenotypic records were kept for further

analyses. Of these, 60,013 (29%) had no sire information due to the use of multiple sires during the breeding season.

2.2.2. Genetic parameter estimates

Estimates of variance components and estimated breeding values (EBVs) for each animal were obtained by Bayesian inference using the THRGIBBS1F90 software (Misztal, 2020). A linear-threshold animal model was fitted in a two-trait analysis with SC and TH data. The objective of the two-trait analysis was to evaluate a possible correlated response on TH resulting from the selection for SC. The statistical model fitted to the data for the two-trait analysis can be written in matrix notation as:

$$l = X\beta + Za + e, \quad (1)$$

where l is the vector of observations (or liabilities for the trait), β is the vector of fixed effects (including CG and age of the animal at the time of evaluation as linear and quadratic covariates), a is the vector of direct additive genetic effects and e a vector of random error effects, where X and Z are the incidence matrices relating the elements in l to the vectors β and a, respectively. We assumed that $a \sim N(0, A\sigma_a^2)$ and $e \sim N(0, I\sigma_e^2)$, where A is the numerator relationship matrix, I is an identity matrix, σ_a^2 and σ_e^2 are the additive genetic and error variance components, respectively. As σ_e^2 is not estimable in threshold models for binary traits (Gianola and Foulley, 1983), the parameterization $\sigma_e^2 = 1$ was used (Sorensen and Gianola, 2002).

In the threshold model, it is assumed that categorical traits are determined by continuous unobservable variables in an underlying scale (liabilities), in such a way that a set of $(j-1)$ thresholds corresponds to the discontinuities verified in the observed scale, for a trait with j ordered and mutually exclusive categories. Thus, initial values for the thresholds are defined as: $t_1 < t_2 < \cdots < t_{j-1}$, under the assumption that $t_0 = -\infty$ and $t_j = +\infty$ (Gianola and Foulley, 1983). In the case of binary traits, the relationship between the score for the i-th animal (y_i) and its liability (l_i) can be represented by:

$$\begin{aligned} &\{y_i = 0, if\ l_i \leq t_1 \\ &\ y_i = 1, if\ l_i > t_1\}; \end{aligned} \quad (2)$$

A Markov chain of 500,000 cycles with an initial discard period of 60,000 cycles and a sampling interval of 68 cycles was generated for the SC-TH two-trait analysis. The convergence of the chains was assessed using the Geweke (1992) and the Heidelberger and Welch (1983) tests, using the packages CODA (Plummer et al., 2006) and BOA (Smith, 2007) in R version 4.1.2 (R Development Core Team, 2017), in addition to visual inspection.

Genetic trends were estimated by linear regression of the posterior means of the EBVs for each trait over the year of birth of animals with known paternity. The posterior means of the EBV for each trait were standardized using the estimated posterior means of the respective additive genetic standard deviation.

2.2.3. Genotypic data and quality control

Data from 6,624 bulls genotyped with the Illumina BovineHD Genotyping BeadChip high-density panel (HD; Illumina, Inc., San Diego, CA, USA), with approximately 780,000 SNP markers, were used as a reference panel to impute 28,155 bulls genotyped with medium density panels (20k to 90k SNPs). These animals are part of the genomic reference population of the Nellore Alliance (GenSys, 2019) and have their own performance record and/or progeny evaluated for TH. The reference panel used allows for a high accuracy of imputation (Carvalheiro et al., 2014; Ferraz et al., 2020). Imputation was performed using the FImpute V3 software (Sargolzaei et al., 2014) and mapped onto the ARS-UCD1.2 assembly coordinates (Hu et al., 2022).

The QCF90 software (Masuda et al., 2019) was used for the quality control (QC) of the genotypic and pedigree datasets. SNP markers with p-value<10^{-5} for the Hardy-Weinberg equilibrium test, call rate less than 0.90 and minor allele frequency (MAF) less than 0.02 were removed from the dataset. After these checks, 461,057 SNPs remained. All genotyped samples had a call rate per individual greater than 0.9. Genotypes from 17,326 animals passed QC and were used in the GWAS analyses.

2.2.4. Genome-wide association study

The procedure WssGBLUP (Wang et al., 2012) allows calculating the genetic values and the effects of SNPs through iterative weighting and was used to estimate the percentage of variance explained by non-overlapping windows of 20 adjacent

SNPs. A single-trait threshold animal model (1) was fitted using the inverse of the H matrix (Aguilar et al., 2010):

$$H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{bmatrix} \quad (3)$$

where G is the genomic relationship matrix using current allele frequencies (VanRaden, 2008), and A_{22} is a numerator relationship matrix for genotyped animals. It was assumed that $a \sim N(0, H\sigma_a^2)$ and $e \sim N(0, I\sigma_e^2)$. The matrices G^{-1} and A_{22}^{-1} were obtained with the PREGSF90 software (Misztal, 2020). To obtain the genomic estimated breeding values (GEBVs), a single chain with 10,000 cycles was generated, applying a discard period of 2,000 initial cycles and a sampling interval of 1 cycle, using the THRGIBBS1F90 software (Misztal, 2020). The variance components were estimated with a single-trait analysis that included only pedigree and phenotypes.

The SNP effect estimates ($\hat{u}$) were obtained with the POSTGSF90 software (Misztal, 2020) as a function of the GEBVs, using a weighted genomic relationship matrix, through the equation:

$$\hat{u} = DZ'[ZDZ']^{-1}\hat{a}_g, \quad (4)$$

where D is a diagonal matrix with weights for the SNPs; Z is an incidence matrix of genotypes for each locus; and $\hat{a}_g$ is the vector of GEBVs of genotyped animals. Estimates of SNP effect, which corresponds to the diagonal elements of D, were used to estimate the variance of each SNP effect (Zhang et al., 2010):

$$\hat{\sigma}_{u,i}^2 = \hat{u}_i^2 2p_i(1 - p_i) \quad (5)$$

where $\hat{u}_i^2$ is the SNPi effect estimates, and p_i is the allele frequency observed for the second allele of SNPi. Using the formulas above, the vector $\hat{u}$ and the matrix D were iteratively recalculated through the algorithm created by Wang et al. (2012).

The S2 method (looping for step 2) was used, in which the animal and SNP effects are recalculated at each iteration. In iteration 0 (single-step GBLUP), weight 1 was assigned to each SNP, which means that all SNPs have the same weight. In the 1st and 2nd iterations of wssGBLUP, the weights are SNP-specific variances that are calculated using the estimate of the SNP allelic substitution effect from the previous iteration and the allele frequencies of the corresponding SNPs (Vallejo et al., 2019). In this study, results from the second iteration of WssGBLUP were used, because they

generally provide greater accuracy for genomic predictions (Lourenco et al., 2020; Vallejo et al., 2016) and marker effects (Irano et al., 2016; Melo et al., 2016; Wang et al., 2012). Marker weights were assigned based on SNP solutions:

$$w_i = 2p_i q_i \hat{u}_i^2, \quad (6)$$

These weights were reused in $D = diag(w_i)$ after standardization. The procedure was restarted with the new D and repeated until the results of the 2nd iteration were obtained (Masuda, 2019).

2.2.5. Gene annotation

The criterion to identify potentially important genomic regions was the sum of the proportion of variance explained by the SNP effects. It was estimated by the POSTGSF90 software (Misztal, 2020), within non-overlapping windows of 20 adjacent SNPs, approximately 130 Kb (average density of one SNP every 6.5 Kb). The windows that explained more than 1% of the additive genetic variance were selected for annotation of candidate genes using the file Bos_taurus.ARS-UCD1.2.104.gtf, corresponding to the bovine gene annotation of the reference map ARS-UCD1.2 (Hu et al., 2022), and the GALLO package (Fonseca et al., 2020). These windows were also visually inspected with the Genome Data Viewer (NCBI, 2022), in case the candidate gene annotation package failed to detect any associated gene(s) close to the coordinates.

We also evaluated the presence of genes in adjacent 100 Kb windows upstream and downstream of the main windows (average annotation interval of 330 Kb), because their effect can be captured by neighboring SNPs that are in linkage disequilibrium (LD). In cattle, the average recombination block is estimated to be 1 Mb across the entire genome (Arias et al., 2009; Weng et al., 2014). Furthermore, LD values (r^2) > 0.08 are observed up to 400 Kb between marker pairs in the Nellore breed, with higher LD estimates (r^2 > 0.15) observed in the first 100 Kb (Pérez O'Brien et al., 2014).

2.2.6. Gene prioritization and gene ontology enrichment analysis

We created a "trained list" of the top 100 genes associated with relevant trait keywords ("hypogonadism", "scrotal circumference", "scrotal", "testicular hypoplasia", "testicular", "testicle", "testicles", "sperm", "semen", "spermatozoon",

“spermatogenesis” and “fertility”) using GUILDify (Guney et al., 2014) and a species-specific (Homo sapiens) interaction network. This list of “trained” genes was used in the ToppGene Suite (Chen et al., 2009) simultaneously with a list of “test” genes, located within windows that explained more than 1% of the genetic variance. Gene Ontology (GO) terms (molecular function, biological process, and cellular component), human and mouse phenotypes, metabolic pathways, PubMed publications, coexpression patterns and diseases were used to retrieve the functional information for genes in training and testing lists (Fonseca et al., 2018).

The prioritized significant genes were selected based on a multiple correction false discovery rate of 5% (p-value ≤ 10^{-3}). This means that these genes have the same functional profile as the genes on the “trained” list (Chen et al., 2009). The R packages ClusterProfiler (Yu et al., 2012) and enrichplot (Wu et al., 2021) were used for enrichment analysis, graphic representation and functional clustering of GO terms for the list of “test” genes. Genes and GO terms were considered enriched when the False-Discovery Rate (FDR) was < 0.05. The function pairwise_termsim() from enrichplot was used to calculate the Jaccard correlation coefficient (Jaccard, 1901) among GO terms, resulting in a similarity matrix. Subsequently, we group the terms functionally using the similarity matrix to identify classes of closely related terms and reduce redundancy across terms. Furthermore, we used the cnetplot() function to identify cluster of genes connected by similar GO terms in order to identify potential candidate biological processes and genes for fertility related-traits.

2.3. Results

The pedigree file contained 368,948 animals. The 4.6% value for total incidence considers animals with phenotypic records after editing the dataset (Table 1). Mean TH frequency increases over years in the population studied, reaching values close to 5% (Figure 1) in the most recent years.

Table 1. Summary of Nellore data used in the SC-TH bi-trait analysis to estimate the variance components and calculate heritability.

Number of phenotypic records	207,195
Number of sires with progeny with phenotypic records	4,466
Mean age of animals at the time of evaluation (in days)	518
Number of contemporary groups	4,771
TH incidence (%)	4.6

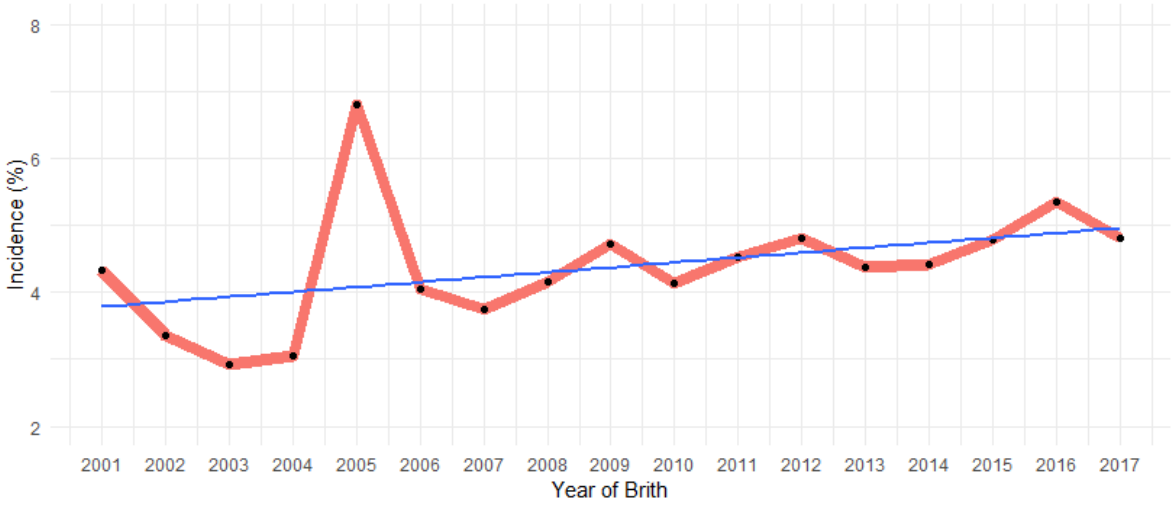

Figure 1. Incidence of TH measured at yearling by birth year.

The posterior mean of the heritability estimate (L95%; U95%) for TH was between low and moderate at 0.16 (0.13; 0.22). In a single-trait analysis, performed for comparison purposes, the average heritability estimate for TH was 0.15 (0.12; 0.18). The posterior mean estimate of the heritability of SC was much higher at 0.46 (0.44; 0.48) (Table 2), while the posterior mean estimate of the genetic correlation between SC and TH was of -0.0445 (±0.0227).

Table 2. Estimates of additive (σ_a^2) and residual (σ_e^2) variance components, covariance (σ), heritability (h^2) and correlation (r) obtained in SC-TH bi-trait analysis in Nellore cattle.

Parameter*	σ_a^2(SC)	σ(SC,TH)	σ_a^2(TH)	σ_e^2(SC)	σ_e^2(TH)	h^2(SC)	h^2(TH)	r(SC,TH)
Mean	3.1742	-0.0996	0.2044	3.7002	1.0075	0.4617	0.1680	-0.0445
L95%	3.0190	-0.1998	0.1588	3.5740	0.9990	0.4411	0.1301	-0.0888
U95%	3.3300	-0.0011	0.2755	3.8250	1.0160	0.4820	0.2148	-0.0005
Median	3.1740	-0.0991	0.1994	3.7010	1.0070	0.4617	0.1652	-0.0444
Mode	3.1750	-0.0900	0.1850	3.7250	1.0070	0.4625	0.1550	-0.0350
SD	0.0784	0.0511	0.0338	0.0640	0.0044	0.0103	0.0230	0.0227
p-value	0.2055	0.1781	0.0525	0.5081	0.9696	0.3039	0.0580	0.1785
ESS	1074.9	93.8	71.3	1276.8	6470.0	1093.8	37.1	82.5

*Posterior mean and respective lower (L95%) and upper (U95%) limits for the 95% credibility interval for scrotal circumference (SC) and testicular hypoplasia (TH) at yearling, posterior Median, posterior Mode, Standard Deviation (SD), Z-Geeke (p-value), Effective Sample Size (ESS).

Annual genetic trends of 0.25 and -0.12 units of genetic standard deviation were estimated for SC and TH, respectively (Figure 2).

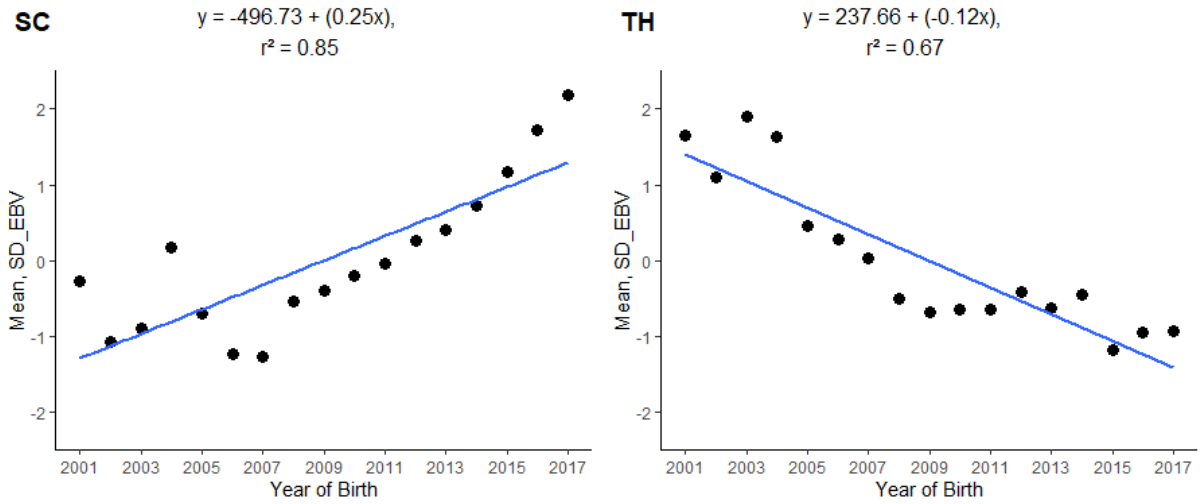

Figure 2. Genetic trends of posterior means of breeding values standardized by year of birth for SC and TH evaluated at yearling in a population of Nellore cattle, SC = scrotal circumference; TH = testicular hypoplasia; DP_VG = a posterior mean of the estimated breeding value expressed in units of genetic standard deviation.

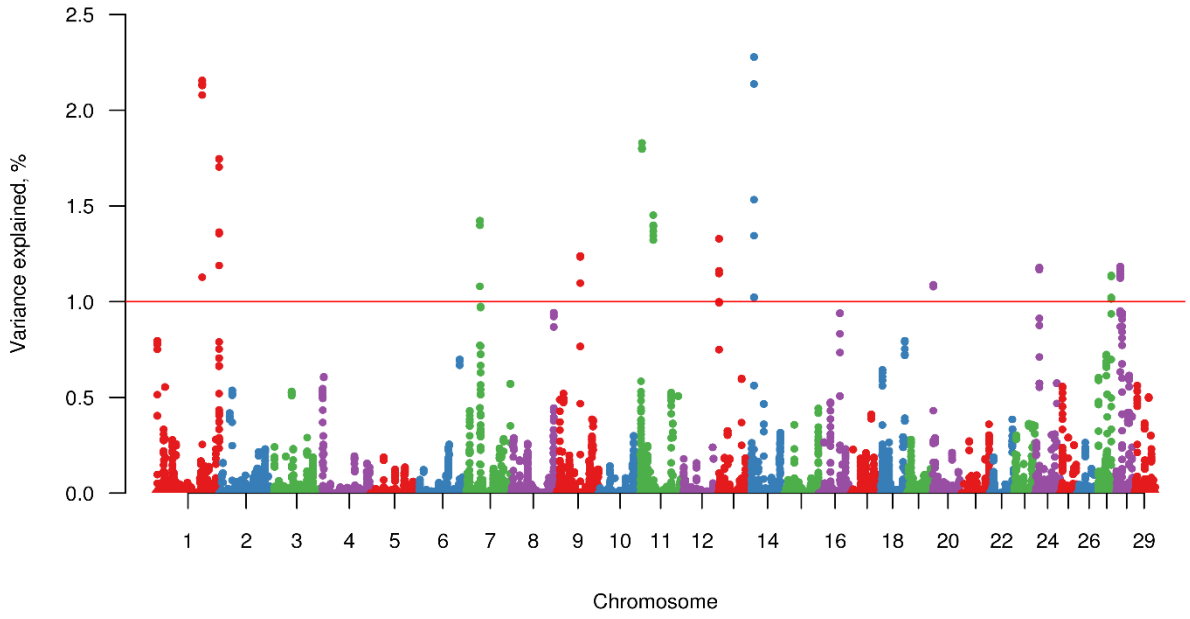

Figure 3. Manhattan plot for percentage of variance explained by SNPs within the top 12 non-overlapping windows of 20 adjacent SNPs for TH obtained by the WssGBLUP method. The red line indicates the threshold of 1% of the variance explained by the windows.

From the GWAS we identified twelve windows that explained more than 1% of the genetic variance for TH (Figure 3). Two of these windows were located on chromosome BTA1, two on chromosome BTA11, and one window on each of chromosomes BTA7, BTA9, BTA13, BTA14, BTA20, BTA24, BTA27 and BTA28

(Table 3). The sum of the variance explained by these windows corresponds to 18.04% of the total genetic variance for TH.

Expanding to 100 Kb upstream and downstream of the non-overlapping windows of 20 adjacent SNPs, a total of 49 positional candidate genes for TH were identified. The window that explained the highest proportion of variance (2.28%) was located in BTA14 (Table 3), but no candidate gene was identified within this region. The visual inspection in the Genome Data Viewer (NCBI, 2022) revealed that the KHDRBS3 gene, associated with the regulation of spermatogenesis, is relatively close (219 Kb) to this window.

Table 3. Annotated genes within the windows of 20 adjacent SNPs that explained more than 1% of the genetic variance for TH in Nellore cattle.

Window Region			Var (%)	Gene names
BTA[1]	Start	End		
14	6.120.069	6.173.536	2,28	
1	116.018.527	116.360.185	2,16	*SUCNR1, LOC613884, AADAC, LOC107132208, LOC782222, LOC782258, LOC100300210, LOC528748, LOC782298, LOC523130*
11	4.763.155	4.805.825	1,83	*AFF3, LOC112448742*
1	157.996.335	158.070.161	1,75	*LOC112448253, LOC112448339, ZBED9, GPX5*[2]*, GPX6, LOC112448270, LOC527195, ZSCAN23*
11	32.850.033	32.983.166	1,45	*NRXN1*
7	33.589.667	33.695.319	1,42	*LOC112447543*
13	2.599.092	2.669.483	1,33	*PLCB4, LAMP5, LOC112449364, PAK5, CRACD, LOC112449244*
9	59.279.324	59.329.161	1,24	*LOC782675*
28	10.049.254	10.108.921	1,18	*RYR2*
24	10.716.835	10.794.665	1,18	*CDH7, LOC112444222*
27	33.096.274	33.226.142	1,14	*LOC782543, ERLIN2*[2]*, PLPBP, LOC112444622, ADGRA2, BRF2, RAB11FIP1, GOT1L1, ADRB3, EIF4EBP1, LOC112444636, ASH2L, STAR*[3]*, LOC112444637*
20	1.930.640	1.972.651	1,09	*DOCK2, FOXI1, LOC112443064*

[1]BTA – *Bos taurus* autosome; [2]Prioritized candidate genes identified by window inspection in the Ensembl genome browser; [3]Prioritized candidate gene identified by GUILDify and ToppGene analysis, considering 100 Kb upstream and downstream of the window.

The window that housed the largest number of genes (fourteen genes) is located on BTA27 and explained 1.14% of the total genetic variance for the trait (Table 3). The STAR gene, located in this window, is related to the metabolism and synthesis of steroid hormones and was the only gene with a significant p-value (FDR-corrected p-value ≤ 0.05) in the prioritization analysis with ToppGene. Two other potentially important genes are GPX5 (BTA1), regulated by androgens, and ERLIN2 (BTA27),

involved in the process of spermatogenesis. Although the p-value of these two genes was not significant, according to the ToppGene prioritization analysis, both are expressed in the epididymis in the mammalian male reproductive tract.

All significant windows (explained variance > 1%) were visually inspected. Some of the candidate genes identified but not prioritized by the ToppGene analysis are involved in lipid metabolism (AADAC, PLCB4, ERLIN2, ADRB3), cell cycle regulation and cell proliferation (ZBED9, PAK5, CRACD, EIF4EBP1, ASH2L, DOCK2), and embryogenesis (FOXI1). Also, in the Ensembl genome browser it was possible to notice the existence of several genes without an associated gene symbol (starting with LOC) (Table 3) which may be relevant. However, they could not be prioritized by the ToppGene analysis.

In the Figure 4, a cluster plot represents the relationships between the genes prioritized in the functional annotation study and Biological Process (BP) GO terms. The complete table with all the results of the enrichment analysis can be found in the supplementary material (Appendix A).

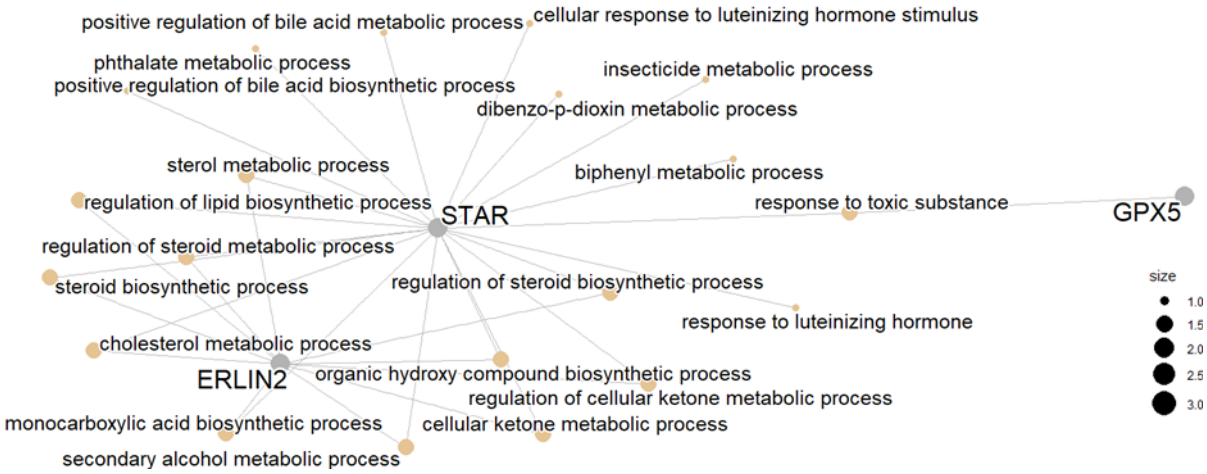

Figure 4. Gene Ontology enrichment analysis of genes GPX5, STAR and ERLIN2. Each node represents a gene, and each edge represents the overlap between two genes. For better visualization, only the top 20 significant (p value < 0.05) terms are shown.

A functional grouping tree diagram (Figure 5) shows the hierarchy of the top 20 Biological Process (BP) GO terms on the main pathway clusters: cholesterol steroid alcohol ketone, biphenyl dibenzo-p-dioxin insecticide phthalate, positive regulation bile acid, response to luteinizing hormone and response to toxic substance.

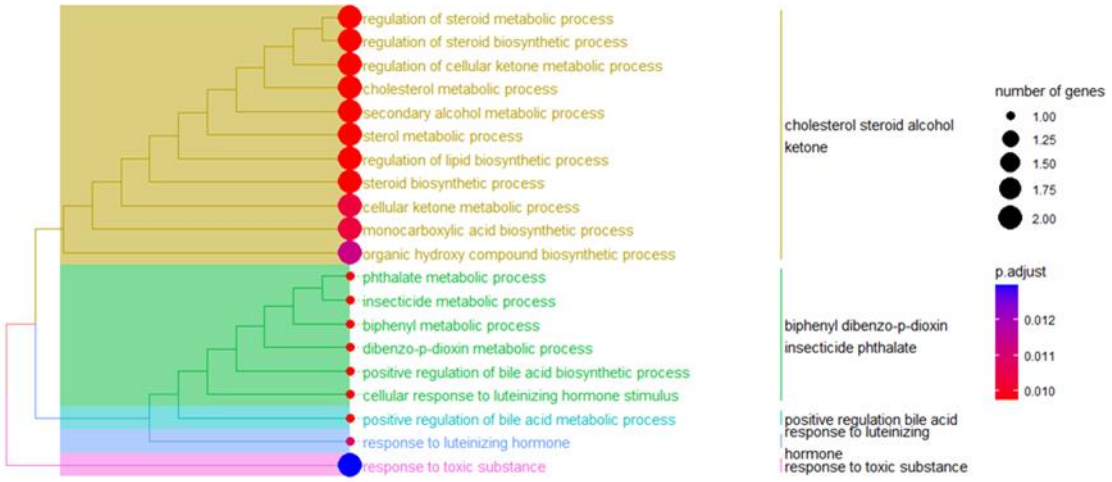

Figure 5. Tree plot for the terms associated with the prioritized genes.

2.4. Discussion

The total incidence of TH was the same reported by Neves et al. (2019) (4.61%), who evaluated phenotypic data from 47,563 animals born between 2005 and 2013, belonging to one of the genetic breeding programs that make up the Aliança Nellore database. This result suggests that the total incidence of this disorder in the population remains at a similar level, even considering a larger database. However, the yearly results suggest that there is an increasing trend in recent times. Heritability (Table 2) was also similar to that estimated by Neves et al. (2019) (0.16), indicating that the variability in this trait is influenced to some extent by additive genetic effects and that the incidence can be reduced by selection in breeding programs.

Whereas Neves et al. (2019) estimated a genetic correlation of -0.53 between TH and SC, we could not confirm that result because of our estimate was -0.0445 for the same parameter with almost five times more records.

The genetic trend of TH in Figure 2 is in the desirable direction, evidencing that there is ongoing genetic progress for the trait. It also suggests that the strategy of culling animals affected by TH has promoted favorable genetic changes in this population.

The presence of peaks in chromosomes BTA1, BTA7, BTA9, BTA11, BTA13, BTA14, BTA20, BTA24, BTA27 and BTA28 (Figure 3) identified by the GWAS are consistent with the postulation that reproductive traits are complex polygenic phenotypes. We found 86 SNPs in the region of BTA14 (between 6.12 and 6.17 Mb) that explained the highest proportion (2.28%) of the genetic variance for TH. This segment corresponds to an intergenic region that can carry unannotated regulatory

elements or coding genes and requires further investigation to clarify its influence on TH phenotypic variation. Maiorano et al. (2022) identified intronic regions with a possible regulatory role in traits under selection in subpopulations of pure-bred Nellore and CEIP cattle. It is known that animal and plant genomes can be transcribed much more extensively than current annotations predict, however, the biological functions of these unannotated transcribed regions are largely unknown (Manak et al., 2006).

In addition, the *KHDRBS3* gene, associated with fertility in bulls (Card et al., 2017; Gross et al., 2020), seems to be in LD with this BTA14 window since the distance between them (219 Kb) has a significant $r^2 > 0.1$ according to the LD decay standard for the Nellore breed (Pérez O'Brien et al., 2014). Thus, the window is either reflecting the effect of LD in this region or its potentially tagging a new region of epigenetic marks or regulatory elements. Epigenetic marks associated with the *KHDRBS3* gene are differentially present between bulls with high and low fertility (Gross et al., 2020). The *KHDRBS3* gene, found in the nuclei of Sertoli cells and germ cells at different stages of spermatogenesis, is associated with the RNA Binding Motif (RBM) (Elliott et al., 1998), which plays a role in pre-mRNA processing (Venables et al., 1999). Also, a similar protein to *KHDRBS3*, *SAM68*, is reduced in the testis of men with azoospermia (Li et al., 2014).

The *GPX5* (BTA1) gene, essentially present in the cytoplasmic compartment of the caput segment 2 epithelium of the bovine epididymis (Grignard et al., 2005), encodes a protein with predicted involvement in the protection of the sperm membrane from the harmful effects of lipid peroxidation during oxidative stress and potentially in preventing a premature acrosome reaction (Hall et al., 1998). Ferenčaković et al. (2017) estimated inbreeding depression under reproductive aspects in Simmental bulls and also found markers in BTA1 associated with semen quality traits close to the *GPX5* gene.

The only candidate gene prioritized by ToppGene analysis, *STAR* (position 33.312 - 33.320Mb on BTA27), is one of the most important factors associated with the steroidogenic pathway, including testosterone production in Leydig cells (Brown and Goldstein, 1997; Eacker et al., 2008; Rizzoto et al., 2020). The protein encoded by the *STAR* gene plays a key role in the acute regulation of steroid hormone synthesis, increasing the conversion/cleavage of cholesterol to pregnenolone mediating the transport of cholesterol from the outer mitochondrial membrane to the inner mitochondrial membrane (Stelzer et al., 2016). Its expression has been studied both in adult bovine testes and in fetal testes (Aspden et al., 1998; Pilon et al., 1997).

Bovine *STAR* gene expression and testicular development in fetuses have been shown to be reduced by maternal overnutrition during the latter half of pregnancy (Weller et al., 2016). Downregulation of gene expression was accompanied by a decrease in the diameter, length and volumetric proportion of seminiferous cords in male fetuses from cows fed an ad libitum diet. Rizzoto et al. (2020) also observed a profound reduction in testicular testosterone concentrations by heat stress-induced downregulation of the *STAR* gene in Nellore bulls. On the other hand, an increase in testicular growth rate in young bulls was associated with the observation of high levels of this protein in bulls treated with Deslorelin[2] (Aspden et al., 1998).

Testosterone, the male sex steroid, is synthesized from cholesterol in the testicular Leydig cell under the control of the pituitary gonadotropin LH. Unlike most cells that use cholesterol primarily for membrane synthesis, steroidogenic cells have additional requirements for cholesterol, which is the essential precursor of all steroid hormones (Eacker et al., 2008). The *ERLIN2* gene (BTA27) is involved in the regulation of cellular cholesterol homeostasis (NCBI, 2022) and the protein encoded by this gene was expressed in epididymal sperm, suggesting its potential involvement in maintaining the integrity of the spermatic neck (Kim et al., 2018).

Other candidate genes identified in this study are also known to be involved in lipid metabolism and, consequently, cholesterol levels. The protein encoded by the *ADRB3* gene (BTA27) belongs to the family of beta-adrenergic receptors, located mainly in adipose tissue, involved in the regulation of lipolysis and thermogenesis. The *AADAC* gene (BTA1) displays cellular triglyceride lipase activity in the liver, increases the levels of intracellular fatty acids derived from the hydrolysis of newly formed triglyceride stores and plays a role in very low-density lipoprotein assembly. Phospholipases, such as the one encoded by the *PLCB4* gene (BTA13), are a group of enzymes that hydrolyze phospholipids into fatty acids and other lipophilic molecules (Stelzer et al., 2016).

The functional candidate genes found in the analysis are involved in biological processes directly related to testis development and fertility in general. Furthermore, *STAR*, *ELRIN2* and *GPX5* seem to share important activities in biological processes related to fertility (Figures 4 and 5) such as regulation of steroid biosynthesis, cholesterol metabolic process and response to luteinizing hormone and to toxic substance stimulus

[2]Luteinizing hormone releasing hormone (LHRH) agonist.

(Brigelius-Flohé and Maiorino, 2013; Huber et al., 2013; Stocco, 2001). Consequently, suggesting that these genes are potential targets for prospecting molecular markers to reduce the frequency of reproductive disorders in the Nelore breed.

2.5. Conclusions

The estimated heritability of testicular hypoplasia in Nellore cattle suggests that a reduction in the incidence of this malformation can be achieved through selection. An increasing trend in this incidence highlights the importance of intensifying this selection, despite the favorable genetic gain observed during the evaluated period. The genetic correlation with scrotal circumference was not confirmed in this study, which may lead to changes in the selection criteria used in breeding programs and needs further confirmation. New peaks and chromosomal regions associated with testicular hypoplasia were identified in Nellore cattle with GWAS and a detailed comparison of these results with the top regions of future studies is suggested. Functional evidence indicates a potential role of positional candidate genes, identified manually and/or automatically, in the incidence of testicular hypoplasia. The genes and genomic regions identified in this study could be used in the future as part of a panel to predict testicular hypoplasia in Nellore cattle.

2.6. References

Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S., Lawlor, T.J., 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93, 743–752.

Aragão, A., Contini, E., 2021. O agro no Brasil e no Mundo: uma síntese do período de 2000 a 2020. Embrapa SIRE.

Arias, J.A., Keehan, M., Fisher, P., Coppieters, W., Spelman, R., 2009. A high density linkage map of the bovine genome. https://doi.org/10.1186/1471-2156-10-18

Aspden, W.J., Rodgers, R.J., Stocco, D.M., Scott, P.T., Wreford, N.G., Trigg, T.E., Walsh, J., D'Occhio, M.J., 1998. Changes in Testicular Steroidogenic Acute Regulatory (STAR) Protein, Steroidogenic Enzymes and Testicular Morphology Associated with Increased Testosterone Secretion in Bulls Receiving The Luteinizing Hormone Releasing Hormone Agonist Deslorelin. Domest. Anim. Endocrinol. 15, 227–238.

Bicudo, S., Siqueira, J., Meira, C., 2007. Patologias do sistema reprodutor de touros. Biológico 69, 43–48.

Brigelius-Flohé, R., Maiorino, M., 2013. Glutathione peroxidases. Biochim. Biophys. Acta - Gen. Subj. https://doi.org/10.1016/j.bbagen.2012.11.020

Brown, M.S., Goldstein, J.L., 1997. The SREBP pathway: Regulation of cholesterol metabolism by proteolysis of a membrane-bound transcription factor. Cell. https://doi.org/10.1016/S0092-8674(00)80213-5

Card, C.J., Krieger, K.E., Kaproth, M., Sartini, B.L., 2017. Oligo-dT selected spermatozoal transcript profiles differ among higher and lower fertility dairy sires. Anim. Reprod. Sci. 177, 105–123. https://doi.org/10.1016/j.anireprosci.2016.12.011

Carvalheiro, R., Boison, S.A., Neves, H.H.R., Sargolzaei, M., Schenkel, F.S., Utsunomiya, Y.T., O'Brien, A.M.P., Sölkner, J., McEwan, J.C., Van Tassell, C.P., Sonstegard, T.S., Garcia, J.F., 2014. Accuracy of genotype imputation in Nelore cattle. Genet. Sel. Evol. 46, 1–11. https://doi.org/10.1186/s12711-014-0069-1

Chen, J., Bardes, E.E., Aronow, B.J., Jegga, A.G., 2009. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. 37, 305–311. https://doi.org/10.1093/nar/gkp427

Eacker, S.M., Agrawal, N., Qian, K., Dichek, H.L., Gong, E.-Y., Lee, K., Braun, R.E., 2008. Hormonal Regulation of Testicular Steroid and Cholesterol Homeostasis. Mol. Endocrinol. 22, 623–635. https://doi.org/10.1210/me.2006-0534

Elliott, D.J., Oghene, K., Makarov, G., Makarova, O., Hargreave, T.B., Chandley, A.C., Eperon, I.C., Cooke, H.J., 1998. Dynamic changes in the subnuclear organisation of pre-mRNA splicing proteins and RBM during human germ cell development. J. Cell Sci. 111, 1255–1265. https://doi.org/10.1242/jcs.111.9.1255

Ferenčaković, M., Sölkner, J., Kapš, M., Curik, I., 2017. Genome-wide mapping and estimation of inbreeding depression of semen quality traits in a cattle population. J. Dairy Sci. 100, 4721–4730. https://doi.org/10.3168/jds.2016-12164

Ferraz, J.B.S., Wu, X.-L., Li, H., Xu, J., Ferretti, R., Simpson, B., Walker, J., Silva, L.R., Garcia, J.F., Tait, R.G., Bauck, S., 2020. Development and evaluation of a low-density single-nucleotide polymorphism chip specific to Bos indicus cattle. Anim. Prod. Sci. 60, 1769–1776. https://doi.org/10.1071/AN19396

Fonseca, P.A. de S., dos Santos, F.C., Lam, S., Suárez-Vega, A., Miglior, F., Schenkel, F.S., Diniz, L. de A.F., Id-Lahoucine, S., Carvalho, M.R.S., Cánovas, A., 2018. Genetic mechanisms underlying spermatic and testicular traits within and among cattle breeds: systematic review and prioritization of GWAS results. J. Anim. Sci. 96, 4978–4999. https://doi.org/10.1093/jas/sky382

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

GenSys, C.A.S.L., 2019. Aliança Nelore: Sumário de avaliação de reprodutores [WWW Document]. URL https://gensys.com.br/sumario/sumario-alianca-nelore-2019-2020/ (accessed 10.18.19).

Geweke, J., 1992. Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments, in: Bayesian Statistics. pp. 641–649.

Gianola, D., Foulley, J.L., 1983. Sire evaluation for ordered. Genet. Sel. Evol. 15, 201–224. https://doi.org/10.1016/j.nbt.2013.09.001

Gnoth, C., Godehardt, E., Frank-Herrmann, P., Friol, K., Tigges, J., Freundl, G., 2005. Definition and prevalence of subfertility and infertility. Hum. Reprod. 20, 1144–1147. https://doi.org/10.1093/humrep/deh870

Grignard, E., Morin, J., Vernet, P., Drevet, J.R., 2005. GPX5 orthologs of the mouse epididymis-restricted and sperm-bound selenium-independent glutathione peroxidase are not expressed with the same quantitative and spatial characteristics in large domestic animals. Theriogenology 64, 1016–1033. https://doi.org/10.1016/j.theriogenology.2005.01.008

Gross, N., Peñagaricano, F., Khatib, H., 2020. Integration of whole-genome DNA methylation data with RNA sequencing data to identify markers for bull fertility. Anim. Genet. 51, 502–510. https://doi.org/10.1111/age.12941

Guney, E., Garcia-Garcia, J., Oliva, B., 2014. GUILDify: A web server for phenotypic characterization of genes through biological data integration and network-based prioritization algorithms. Bioinformatics 30, 1789–1790. https://doi.org/10.1093/bioinformatics/btu092

Hall, L., Williams, K., Perry, A.C.F., Frayne, J., Jury, J.A., 1998. The majority of human glutathione peroxidase type 5 (GPX5) transcripts are incorrectly spliced: Implications for the role of GPX5 in the male reproductive tract. Biochem. J. 333, 5–9. https://doi.org/10.1042/bj3330005

Heidelberger, P., Welch, P.D., 1983. Simulation Run Length Control in the Presence of an Initial Transient. Oper. Res. 31, 1109–1144. https://doi.org/10.1287/opre.31.6.1109

Hu, Z.-L., Park, C.A., Reecy, J.M., 2022. Bringing the Animal QTLdb and CorrDB into the future: meeting new challenges and providing updated services. Nucleic Acids Res. 50, D956–D961. https://doi.org/10.1093/nar/gkab1116

Huber, M.D., Vesely, P.W., Datta, K., Gerace, L., 2013. Erlins restrict SREBP activation in the ER and regulate cellular cholesterol homeostasis. J. Cell Biol. 203, 427–436. https://doi.org/10.1083/jcb.201305076

Irano, N., De Camargo, G.M.F., Costa, R.B., Terakado, A.P.N., Magalhães, A.F.B., Silva, R.M.D.O., Dias, M.M., Bignardi, A.B., Baldi, F., Carvalheiro, R., De Oliveira, H.N., De Albuquerque, L.G., 2016. Genome-wide association study for indicator traits of sexual precocity in Nellore cattle. PLoS One 11, 159502.

Jaccard, P., 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull. la Société vaudoise des Sci. Nat. 37, 547–579.

Kastelic, J.P., Thundathil, J.C., 2008. Breeding Soundness Evaluation and Semen Analysis for Predicting Bull Fertility. Reprod. Domest. Anim. 43, 368–373.

Kim, Jihye, Kwon, J.T., Jeong, J., Kim, Jaehwan, Hong, S.H., Kim, Jinyoung, Park, Z.Y., Chung, K.H., Eddy, E.M., Cho, C., 2018. SPATC 1L maintains the integrity of the sperm head-tail junction. EMBO Rep. 19, e45991. https://doi.org/10.15252/embr.201845991

Li, L.J., Zhang, F. Bin, Liu, S.Y., Tian, Y.H., Le, F., Lou, H.Y., Huang, H.F., Jin, F., 2014. Decreased expression of SAM68 in human testes with spermatogenic defects. Fertil. Steril. 102, 61-67.e3. https://doi.org/10.1016/j.fertnstert.2014.03.036

Lourenco, D., Legarra, A., Tsuruta, S., Masuda, Y., Aguilar, I., Misztal, I., 2020. Single-Step Genomic Evaluations from Theory to Practice: Using SNP Chips and Sequence Data in BLUPF90. Genes (Basel). 11, 790. https://doi.org/10.3390/genes11070790

Maiorano, A.M., Cardoso, D.F., Carvalheiro, R., Júnior, G.A.F., de Albuquerque, L.G., de Oliveira, H.N., 2022. Signatures of selection in Nelore cattle revealed by whole-genome sequencing data. Genomics 114, 110304. https://doi.org/10.1016/j.ygeno.2022.110304

Manak, J.R., Dike, S., Sementchenko, V., Kapranov, P., Biemar, F., Long, J., Cheng, J., Bell, I., Ghosh, S., Piccolboni, A., Gingeras, T.R., 2006. Biological function of unannotated transcription during the early development of Drosophila melanogaster. Nat. Genet. 38, 1151–1158. https://doi.org/10.1038/ng1875

MAPA - MINISTÉRIO DA AGRICULTURA PECUÁRIA E ABASTECIMENTO, B., 2006. Instrução Normativa nº 53, de 27 de setembro de 2006 [WWW Document]. URL http://sistemasweb.agricultura.gov.br/sislegis/action/detalhaAto.do?method=visualizarAtoPortalMapa&chave=735332411 (accessed 10.29.19).

Masuda, Y., 2019. Introduction to BLUPF90 suite programs: Standard Edition. Univ. Georg. http//nce.ads.uga.edu/wiki/doku.php?id=documentation.

Masuda, Y., Legarra, A., Aguilar, I., Misztal, I., 2019. 331 Efficient quality control methods for genomic and pedigree data used in routine genomic evaluation. J. Anim. Sci. 97, 50–51. https://doi.org/10.1093/jas/skz258.101

McEntee, K., 1990. Scrotum, spermatic cord, and testis: proliferative lesions, in: Reproductive Pathology of Domestic Mammals. Academic Press, San Diego, CA, pp. 279–300.

Melo, T.P., Takada, L., Baldi, F., Oliveira, H.N., Dias, M.M., Neves, H.H.R., Schenkel, F.S., Albuquerque, L.G., Carvalheiro, R., 2016. Assessing the value of phenotypic information from non-genotyped animals for QTL mapping of complex traits in real and simulated populations. BMC Genet. 17, 1–9. https://doi.org/10.1186/s12863-016-0394-1

Misztal, I., 2020. BLUPF90 - a flexible mixed model program in Fortran 90. Available from: http://nce.ads.uga. edu/wiki/lib/exe/fetch.php?media=blupf90.pdf. Accessed 20 Abr 2020.

Moura, G.S., 2013. Estudo clínico-andrológico e genético-molecular da subfertilidade em touros da raça Gir Leiteiro (Bos taurus indicus) no Estado de Minas Gerais. UNIVERSIDADE FEDERAL DE MINAS GERAIS.

NCBI, N.C. for B.I., 2022. cattle (Bos taurus) - NCBI Genome Data Viewer [WWW Document]. Natl. Libr. Med. URL https://www.ncbi.nlm.nih.gov/genome/gdv/?org=bos-taurus (accessed 5.4.22).

Neves, H.H.R., Vargas, G., Brito, L.F., Schenkel, F.S., Albuquerque, L.G., Carvalheiro, R., 2019. Genetic and genomic analyses of testicular hypoplasia in Nellore cattle. PLoS One 14, e0211159. https://doi.org/10.1371/journal.pone.0211159

Pérez O'Brien, A.M., Mészáros, G., Utsunomiya, Y.T., Sonstegard, T.S., Garcia, J.F., Van Tassell, C.P., Carvalheiro, R., da Silva, M.V.B., Sölkner, J., 2014. Linkage disequilibrium levels in Bos indicus and Bos taurus cattle using medium and high density SNP chip data and different minor allele frequency distributions. Livest. Sci. 166, 121–132. https://doi.org/10.1016/j.livsci.2014.05.007

Pilon, N., Daneau, I., Brisson, C., Ethier, J.F., Lussier, J.G., Silversides, D.W., 1997. Porcine and bovine steroidogenic acute regulatory protein (StAR) gene expression during gestation. Endocrinology 138, 1085–1091. https://doi.org/10.1210/endo.138.3.5003

Plummer, M., Best, N., Cowles, K., Vines, K., 2006. CODA: Convergence Diagnosis and Output Analysis for MCMC. R News 6, 7–11.

R Development Core Team, 2017. R: A language and environment for statistical computing.

Rizzoto, G., Ferreira, J.C.P., Codognoto, V.M., Oliveira, K.C., Mogollón García, H.D., Pupulim, A.G.R., Teixeira-Neto, F.J., Castilho, A., Nunes, S.G., Thundathil, J.C., Kastelic, J.P., 2020. Testicular hyperthermia reduces testosterone concentrations and alters gene expression in testes of Nelore bulls. Theriogenology 152, 64–68. https://doi.org/10.1016/j.theriogenology.2020.04.029

Rosa, A. do N., Menezes, G.R. de O., Egito, A.A. do, Nogueira, E., 2021. Genética aplicada para pequenos e médios produtores de gado de corte.

Roso, V.M., Schenkel, F.S., 2006. AMC-a computer program to assess the degree of connectedness among contemporary groups, in: Proceedings of the 8th World Congress on Genetics Applied to Livestock Production, Belo Horizonte, Minas Gerais, Brazil, 13-18 August, 2006. Instituto Prociência, pp. 26–27.

Sargolzaei, M., Chesnais, J.P., Schenkel, F.S., 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15, 478. https://doi.org/10.1186/1471-2164-15-478

Smith, B.J., 2007. boa : An R Package for MCMC Output Convergence Assessment and Posterior Inference. J. Stat. Softw. 21, 1–37. https://doi.org/10.18637/jss.v021.i11

Sorensen, D., Gianola, D., 2002. Likelihood, Bayesian and MCMC methods in quantitative genetics, in: Dietz, K., Gail, M., Krickeberg, K., Samet, J., Tsiatis, A. (Eds.), Statistics for Biology and Health. Springer, New York, NY, p. 740.

Steffen, D., 1997. Genetic causes of bull infertility. Vet. Clin. North Am. Food Anim. Pract. 13, 243–253. https://doi.org/10.1016/S0749-0720(15)30338-8

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Iny Stein, T., Nudel, R., Lieder, I., Mazor, Y., Kaplan, S., Dahary, D., Warshawsky, D., Guan-Golan, Y., Kohn, A., Rappaport, N., Safran, M., Lancet, D., 2016. The GeneCards suite: From gene data mining to disease genome sequence analyses. Curr. Protoc. Bioinforma. 2016, 1.30.1-1.30.33. https://doi.org/10.1002/cpbi.5

Stocco, D.M., 2001. StAR protein and the regulation of steroid hormone biosynthesis. Annu. Rev. Physiol. https://doi.org/10.1146/annurev.physiol.63.1.193

Vallejo, R.L., Cheng, H., Fragomeni, B.O., Shewbridge, K.L., Gao, G., MacMillan, J.R., Towner, R., Palti, Y., 2019. Genome-wide association analysis and accuracy of genome-enabled breeding value predictions for resistance to infectious hematopoietic necrosis virus in a commercial rainbow trout breeding population. Genet. Sel. Evol. 51, 1–14. https://doi.org/10.1186/s12711-019-0489-z

Vallejo, R.L., Leeds, T.D., Fragomeni, B.O., Gao, G., Hernandez, A.G., Misztal, I., Welch, T.J., Wiens, G.D., Palti, Y., 2016. Evaluation of genome-enabled selection for bacterial cold water disease resistance using progeny performance data in rainbow trout: Insights on genotyping methods and genomic prediction models. Front. Genet. 7, 96. https://doi.org/10.3389/fgene.2016.00096

VanRaden, P.M., 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91, 4414–4423. https://doi.org/10.3168/jds.2007-0980

Venables, J.P., Vernet, C., Chew, S.L., Elliott, D.J., Cowmeadow, R.B., Wu, J., Cooke, H.J., Artzt, K., Eperon, I.C., 1999. T-STAR/ETOILE: A novel relative of SAM68 that interacts with an RNA-binding protein implicated in spermatogenesis. Hum. Mol. Genet. 8, 959–969. https://doi.org/10.1093/hmg/8.6.959

Wang, H., Misztal, I., Aguilar, I., Legarra, A., Muir, W.M., 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genet. Res. (Camb). 94, 73–83. https://doi.org/10.1017/S0016672312000274

Weller, M.M.D.C.A., Fortes, M.R.S., Marcondes, M.I., Rotta, P.P., Gionbeli, T.R.S., Valadares Filho, S.C., Campos, M.M., Silva, F.F., Silva, W., Moore, S., Guimarães, S.E.F., 2016. Effect of maternal nutrition and days of gestation on pituitary gland and gonadal gene expression in cattle. J. Dairy Sci. 99, 3056–3071.

Weng, Z.Q., Saatchi, M., Schnabel, R.D., Taylor, J.F., Garrick, D.J., 2014. Recombination locations and rates in beef cattle assessed from parent-offspring pairs. Genet. Sel. Evol. 46, 1–14. https://doi.org/10.1186/1297-9686-46-34

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., Yu, G., 2021. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innov. 2, 100141. https://doi.org/10.1016/j.xinn.2021.100141

Yu, G., Wang, L.G., Han, Y., He, Q.Y., 2012. ClusterProfiler: An R package for comparing biological themes among gene clusters. Omi. A J. Integr. Biol. 16, 284–287.

Zhang, Z., Liu, J., Ding, X., Bijma, P., de Koning, D.J., Zhang, Q., 2010. Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. PLoS One 5, 1–8. https://doi.org/10.1371/journal.pone.0012648

CAPÍTULO 3 - Feet and legs malformation in Nellore cattle: genetic analysis and prioritization of GWAS results[3]

Abstract – Beef cattle affected by feet and legs malformations (FLM) cannot perform their productive and reproductive functions satisfactorily, resulting in significant economic losses. Accelerated weight gain in young animals due to increased fat deposition can lead to ligaments, tendon and joint strain and promote gene expression patterns that lead to changes in the normal architecture of the feet and legs. The possible correlated response in the FLM due to yearling weight (YW) selection suggest that this second trait could be used as an indirect selection criterion. Therefore, FLM breeding values and the genetic correlation between FLM and yearling weight (YW) were estimated for 295,031 Nellore animals by fitting a linear-threshold model in a Bayesian approach. A genome-wide association study was performed to identify genomic windows and positional candidate genes associated with FLM. The effects of Single Nucleotide Polymorphisms (SNPs) on animal phenotypes (affected or unaffected), 12,537 genotypes, and 461,057 SNPs were estimated using by the weighted single-step genomic BLUP method. Twelve non-overlapping windows of 20 adjacent SNPs explaining more than 1% of the additive genetic variance were selected for candidate gene annotation. Functional and gene prioritization analysis of candidate genes identified six genes (ATG7, EXT1, ITGA1, PPARD, SCUBE3, SHOX) that may play a role in FLM expression due to their known role in skeletal muscle development, aberrant bone growth, lipid metabolism, intramuscular fat deposition and skeletogenesis.

Key words: *Bos taurus indicus*, candidate genes, functional trait, locomotion, welfare.

[3] Este capítulo corresponde ao artigo científico submetido à revista Frontiers in Genetics Livestock Genomics – 1118308, e encontra-se em avaliação para publicação.

3.1. Introduction

In Brazil, beef cattle breeding programs discard bulls affected by Feet and Legs Malformation (FLM), even though they have great genetic value for other traits of economic importance (e.g., reproductive traits, growth and carcass quality). This is a functional binary trait previously reported in the literature, has a heritability of 0.18 (SE=0.04) and an incidence of approximately 5% in Nellore cattle (Vargas et al., 2017). Animals affected by Feet and Legs Malformation cannot perform their productive and reproductive functions satisfactorily. As a consequence, a large portion of animals of with high genetic value and CEIP (Special Certificate of Identification and Production) candidates (MAPA, 2006) are disqualified, resulting in significant economic losses.

The slaughter of superior animals also results in a lack of replacements for improved bulls and leads to less efficient meat production systems. An effective way to reduce the occurrence of FLM is to study the degree of genetic variability, heritability, and possible genetic associations with other traits in the selection process. The genetic correlation between FLM and yearling weight (YW) (Vargas et al., 2017) suggests the possibility of using the second trait as an indirect selection criterion but needs to be recalculated due to the increased number of records available for FLM in the Nellore breed.

Identification of significant regions in the bovine genome and candidate genes associated with FLM can also help to understand the biology behind the phenotype (Vargas et al., 2018). In this way, Genome-wide Association Studies (GWAS), using phenotypic, genotype, and pedigree data - through the weighted single step genomic best linear unbiased predictor (WssGBLUP) method (Legarra, Aguilar and Misztal, 2009) enables identification of genomic regions that may be associated with the trait of interest.

In order to better understand the genetic structure of FLM in order to provide subsidies to help reduce the occurrence of this problem and thus improve the production and reproductive efficiency of Nellore cattle in selection programs, the goals of this study were to (1) estimate the genetic parameters of FLM, including genetic correlation between FLM and yearling weight (YW); (2) estimates of genetic trends for these traits in Nellore cattle; (3) use of WssGBLUP to find windows of adjacent SNPs significantly associated with FLM and (4) identification of candidate genes supported by functional evidence with a potential role in the incidence of FLM.

3.2. Materials and methods

Approval from the Welfare and Animal Use Committee was not required for this study because the data was obtained from an existing database of phenotypic and genotypic records of Nellore cattle. Previous studies using these data also did not need this approval.

3.2.1. Population structure and phenotypic data

Data of 295,031 Nellore cattle from the Nellore Alliance (GenSys, 2019), all with phenotypic information, collected between 2001 and 2017 were used in this study. Feet and legs malformation were evaluated by trained technicians which assigned binary scores (1 for affected animals and 0 for non-affected) to the overall structure of feet and legs at yearling (Vargas *et al.*, 2018). Contemporary groups (CG) were defined considering the effects of herd, year and season of birth, sex, management group at weaning and yearling, date of measurement at yearling. Only data from CG with more than 10 records, variability in the trait and at least 10 genetic links to other CG were kept for the analyses. Connectedness among CG was checked using AMC software (Roso and Schenkel, 2006) with default parameters. As a result of the use of lots of multiple sires in the breeding season on part of the farms that make up the database, 31% (90,500) of the animals had the father information as unknown.

3.2.2. Genetic parameter estimates

After checking consistency and editing data, estimates of variance components and estimated breeding values (EBV) were obtained for each animal by Bayesian inference using THRGIBBS1F90 software (Misztal et al., 2014). To evaluate a possible correlated selection response of YW in FLM, a linear-threshold animal model was fitted to the two-trait analysis. This statistical model was also used by Vargas et al. (2017) and can be written in matrix notation as:

$$l = X\beta + Za + e$$

where l is the vector of observations (or liabilities for the trait), β is the vector of fixed effects (including the classificatory effect of CG and the effect of age of the animal at

the time of evaluation as linear and quadratic covariates), a is the vector of direct additive genetic effects and e a vector of random residual effects, where X and Z are the incidence matrices relating the elements in l to the vectors β and a, respectively. We assumed that $a \sim N(0, A\sigma_a^2)$ and $e \sim N(0, I\sigma_e^2)$, where A is the numerator relationship matrix, I is an identity matrix, σ_a^2 and σ_e^2 are the additive and residual genetic variance components, respectively. As σ_e^2 is not estimable for binary models (Gianola and Foulley, 1983), the parameterization $\sigma_e^2 = 1$ was used (Sorensen and Gianola, 2002). A priori, the residual covariance was considered equal to zero, and remained at zero after being estimated by the software.

The relationship between the score for the i^{th} animal (y_i) and its liability (l_i) for FLM can be represented by:

$$\{y_i = 0, if\ l_i \leq t_1$$
$$y_i = 1, if\ l_i > t_1\}$$

in which: y_i is the score for the i^{th} animal, t_1 corresponds to threshold that define, on the underlying scale, the mutually exclusive categories of FLM (0 or 1) (Gianola and Foulley, 1983). A 500,000-cycle Markov chain was generated in the YW-FLM two-trait analysis with an initial discard period of 50,000 cycles and sampling interval of 50 cycles. The convergence of the chains was evaluated by Geweke (1992) and Heidelberger and Welch (1983) tests, using the packages CODA (Plummer *et al.*, 2006) and BOA (Smith, 2007) in language R version 4.1.2 (R Development Core Team, 2017), in addition to visual inspection.

3.2.3. Genotyping and quality control

Genotyping was completed from 6,624 animals using the Illumina BovineHD Genotyping BeadChip (HD; Illumina, Inc., San Diego, CA, USA), which included 612,174 SNP markers, and 28,155 animals genotyped with medium density panels (20k to 90k SNP). All animals belong to the Nellore Alliance population (GenSys, 2019) and had their own performance record and/or progeny evaluated for FLM. The FImpute v3 (Sargolzaei, Chesnais and Schenkel, 2014) software was used for genotype imputation from the lower density panels to the HD SNP chip and marker coordinates were mapped according to the genomic positions provided by new bovine genome assembly ARS-UCD1.2 (Hu, Park and Reecy, 2022).

Quality control (QC) of the genotypic and pedigree datasets was performed using QCF90 software (Masuda *et al.*, 2019) and the following criteria were used for the exclusion of SNP markers: p-value<10-5 for the Hardy-Weinberg equilibrium test; minor allele frequency (MAF) <0.02; and a call rate <90%.

3.2.4. Genome-Wide Association Study

The GWAS results were reported as the proportion of the variance explained by non-overlapping genomic windows of 20 adjacent SNP (Zhang *et al.*, 2016; Hay and Roberts, 2018; Oliveira *et al.*, 2019; Zhuang *et al.*, 2020). The phenotypes of FLM were used as dependent variables in a single-trait threshold animal model. SNP effects were estimated by the weighted single-step genomic BLUP (WssGBLUP) method with two iterations, as proposed by Wang et al. (2012), using the programs of BLUPF90 family (Misztal *et al.*, 2014). As in Vargas *et al.* (2018) the solutions (û) were obtained as a function of the genomic estimated breeding values (GEBVs) through the equation:

$$\hat{u} = DZ'[ZDZ']^{-1}\hat{a}_g$$

where D is a diagonal matrix of weights for SNP variances; Z is a matrix relating genotypes of each locus; and $\hat{a}_g$ is the vector of GEBVs of genotyped animals. The vector $\hat{u}$ and the matrix D were iteratively recalculated through the S2 method (looping for step 2) of the algorithm created by Wang et al. (2012). The results from the second iteration were used, because they generally provide greater accuracy for genomic predictions (Vallejo *et al.*, 2016; Lourenco *et al.*, 2020) and marker effects (Wang *et al.*, 2012; Irano *et al.*, 2016; Melo *et al.*, 2016).

The POSTGSF90 software was used to estimate the proportion of variance explained by SNP effects and the criterion used to identify potentially important genomic regions was the sum of the variance in non-overlapping windows of 20 adjacent SNPs, approximately 130 Kb (average density of one SNP per 6.5 kb). Windows that explained greater than 1% of the genetic variance were selected for gene annotation, conducted using R Version 4.1.2 (R Development Core Team, 2017) and the R package GALLO (Fonseca et al., 2020). The .gtf file Bos_taurus.ARS-UCD1.2.104.gtf, corresponding to the bovine reference map ARS-UCD1.2 (Hu, Park and Reecy, 2022) was used for gene annotation.

As LD values (r^2) > 0.08 are observed up to 400 Kb of distance between marker pairs for the Nellore breed, with higher linkage disequilibrium estimates (r^2 > 0.15) observed in the first 100 Kb (Pérez O'Brien et al., 2014), we evaluated the presence of genes in adjacent 100 Kb windows upstream and downstream of the main windows (average annotation interval of 330 Kb). We also visually inspected the windows with the Genome Data Viewer (NCBI, 2022), in case candidate gene annotation failed to detect related genes located near the coordinates.

3.2.5. Gene prioritization analysis

The candidate gene prioritization was conducted using GUILDify and ToppGene (Chen, Bardes, *et al.*, 2009; Guney, Garcia-Garcia and Oliva, 2014). A "trained list" of the top 100 genes associated with relevant trait keywords, including, "mobility aplomb", "feet", "foot angle", "leg", "rear legs", "leg conformation", "leg structure", "conformational structure", "feet quality", "leg quality", "articular cartilage", "osteogenic differentiation", "lameness resistance" and "hoof health" was created using GUILDify and a species-specific (Homo sapiens) interaction network. Then, the ToppGene Suite was used to perform a prioritization analysis comparing the functional information shared between the list of "trained" genes and the list of "test" genes (genes located within windows that explained more than 1% of the genetic variance).

Gene Ontology terms (molecular function, biological process, and cellular component), human and mouse phenotypes, metabolic pathways, PubMed publications, coexpression patterns and diseases were used to retrieve the functional information from genes in training and testing lists (Fonseca *et al.*, 2018). We selected, as prioritized, the genes on the "test" list with the same functional profile as the genes of "trained" list, based on a multiple correction false discovery rate of 5% (p-value ≤ 10-3) (Chen, Bardes, *et al.*, 2009).

3.3. Results

The incidence value (approximately 5%) for FLM after consistency checks and preliminary editing described above should be highlighted (Table 1). The pedigree file contained 468,302 animals.

Table 1. Descriptive structure of Nellore data used in the YW-FLM two-trait analysis.

Number of phenotypic records	295,031
Number of sires with progeny with phenotypic records	3,853
Mean age of animals at the time of evaluation (in days)	526
Number of contemporary groups	6,628
Number of males	177,130
Number of females	117,901
FLM incidence (%)	5.34

The incidence of FLM has increased in the most recent years for both sexes in the population studied (Figure 1). There was a statistically significant difference by Student's t test ($P < 0.05$) in the annual incidence averages between genders. Males (average incidence = 4.98%) were more affected than females (average incidence = 4.23%) by FLM in the period evaluated.

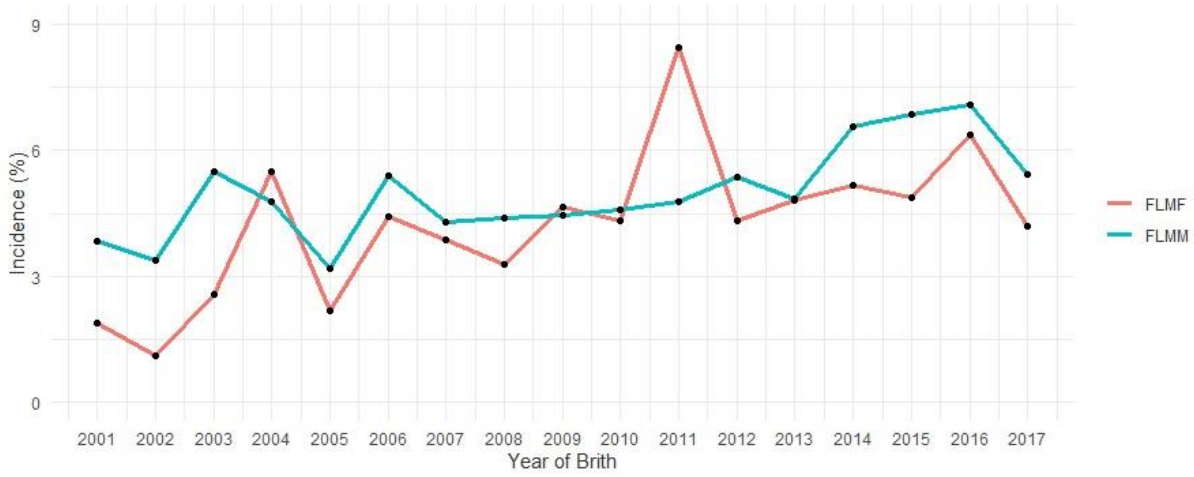

Figure 1. Incidence of FLM at yearling in males (FLMM) and females (FLMF) by year of birth.

After QC, 461,057 SNP remained. All genotyped samples had a call rate per individual greater than 0.9. Genotypes from 12,537 animals passed QC and were used in the GWAS analyses.

The posterior mean of the heritability estimates (L95%=0.1458; U95%=0.2250) for FLM was of moderate magnitude (0.1856). In single-trait analyses for comparison purposes, the mean heritability estimated for FLM (L95%=0.13; U95%=0.20) was 0.17. The linear trait (YW) had a high heritability value (0.4718) (Table 2), while the posterior

mean estimate of the genetic correlation between FLM and YW was -0.2311 (-0.25; -0.22).

Table 2. Estimates of additive (σ_a^2) and residual (σ_e^2) genetic variance components, covariance (σ), heritability (h^2) and genetic correlation (r) obtained in YW-FLM two-trait analysis in Nellore cattle.

Parameter*	σ_a^2(YW)	σ(YW,FLM)	σ_a^2(FLM)	σ_e^2(YW)	σ_e^2(FLM)	h^2(YW)	h^2(FLM)	r(YW,FLM)
Mean	341.64	-4.3829	0.2304	382.50	1.0074	0.4718	0.1856	-0.2311
L95%	331.10	-4.6750	0.1718	374.00	1.0000	0.4590	0.1458	-0.2460
U95%	352.40	-4.0940	0.2925	390.80	1.0150	0.4848	0.2250	-0.2164
Median	341.50	-4.3820	0.2298	382.50	1.0070	0.4717	0.1858	-0.2311
Mode	341.00	-4.3500	0.2350	383.00	1.0070	0.4725	0.1850	-0.2325
SD	5.5063	0.1457	0.0305	4.3181	0.0037	0.0067	0.0200	0.0076
Z-Geweke	0.8291	0.1150	0.5754	0.3386	0.0964	0.6027	0.5233	0.1478
ESS	1583.9	316.8	51.5	1571.5	9000.0	1574.1	68.0	360.7

*Posterior mean and respective lower (L95%) and upper (U95%) limits for the 95% credibility interval for yearling weight (YW) and feet and legs malformation (FLM), posterior Median, posterior Mode, Standard Deviation (SD), Z-Geweke (p-value), Effective Sample Size (ESS).

The annual genetic trends for YW and FLM were 0.25 and -0.25 units of genetic standard deviation, respectively (Figure 2).

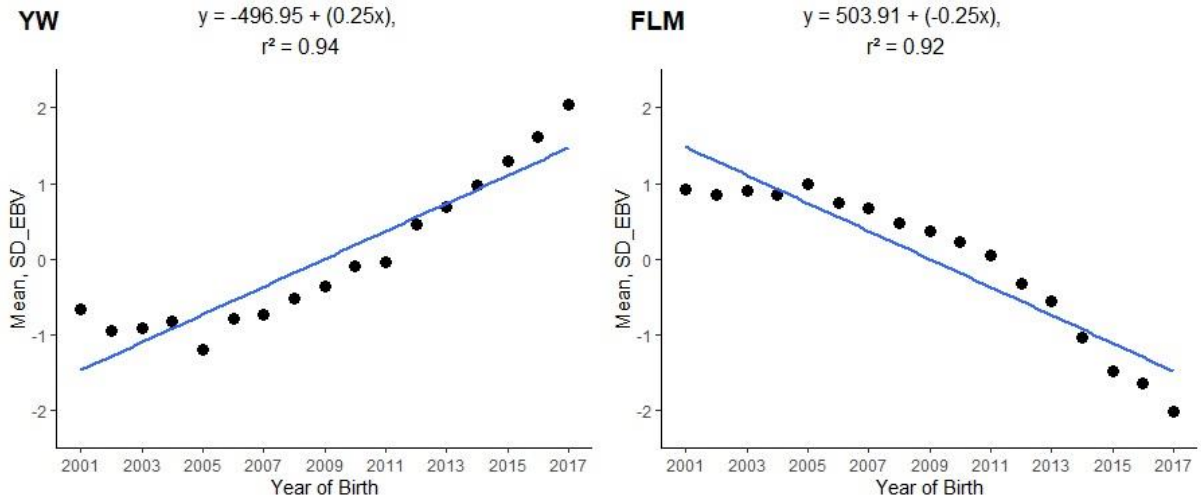

Figure 2. Genetic trends of posterior means of breeding values standardized by year of birth for YW and FLM evaluated at yearling in Nellore cattle, YW = yearling weight; FLM = feet and legs malformation; SD_EBV = a posterior mean of the estimated breeding value expressed in units of genetic standard deviation.

Twelve windows, two of these located on chromosome BTA23 and one on each of chromosomes BTA2, BTA3, BTA5, BTA6, BTA14, BTA18, BTA20, BTA22, BTA27 and BAT29, explained > 1% of the genetic variance of FLM. Together, these windows explained 16.37% of the total genetic variance of FLM (Figure 1).

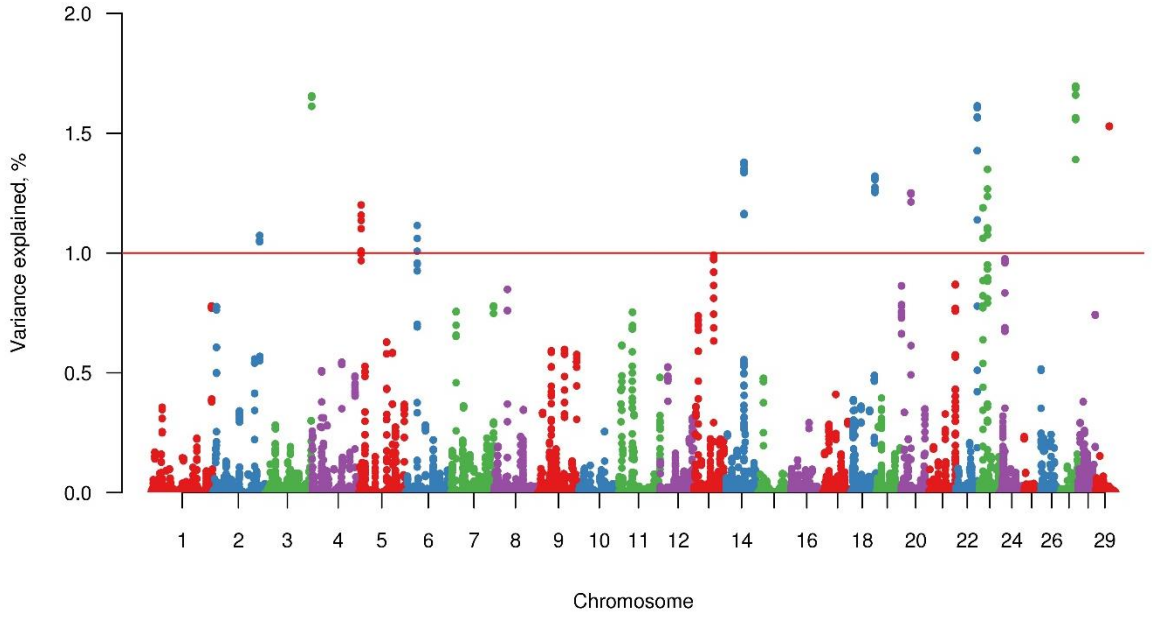

Figure 3. Manhattan plot for percentage of variance explained by non-overlapping windows of 20 adjacent SNP for FLM obtained by WssGBLUP method. The red line indicates the 1% variance threshold explained by the windows.

Table 3. Annotated genes within the windows of 20 adjacent SNP that explained more than 1% of genetic variance for FLM in Nellore cattle.

Window Region			Var (%)	Gene Names
BTA[1]	Start	End		
27	39,428,405	39,525,725	1.70	*LRRC3B*
3	119,525,250	119,674,756	1.66	*CSF2RA, LOC100336476, IL3RA, LOC112446045, LOC107131293, LOC112446046, P2RY8, LOC112445918*, *AKAP17A, LOC112445919, LOC112446047, LOC617467, LOC526047, COPS9, OTOS, LOC782275, LOC112446048*
22	55,269,065	55,321,257	1.61	*HRH1, ATG7*[2]
29	35,078,320	35,112,359	1.53	*NTM, LOC112444927*
14	46,405,542	46,506,899	1.38	*EXT1*[2]
23	20,946,123	21,032,829	1.35	*LOC107131722, PTCHD4*
18	62,415,280	62,498,670	1.32	*EPS8L1, RDH13, LOC100848752, GP6, NLRP2, LOC100336589, LOC100852077, LOC112442414, LOC100301263, NCR1, FCAR, KIR2DL5A, KIR3DL1, KIR3DS1*
20	26,181,870	26,332,917	1.25	*ITGA1*[2]*, PELO, LOC112443001*
5	6,318,396	6,420,053	1.20	*ZDHHC17, CSRP2, LOC112446899, E2F7*
23	9,228,154	9,312,557	1.19	*TCP11, SCUBE3*[3]*, LOC101907009*, *ZNF76, DEF6, PPARD*[2]
6	30,757,118	30,803,603	1.11	*ATOH1, GRID2, TRNAE-UUC, LOC112447051*
2	121,473,182	121,525,407	1.07	*LOC112442276*, *CLDN12, SHOX*[3]*, LOC112442278*, *MIR2887-1*

[1]BTA – *Bos taurus* autosome; [2]Prioritized candidate genes identified by GUILDify and ToppGene analysis, considering 100 Kb upstream and downstream of the window; [3]Prioritized candidate genes identified by window inspection in the Ensembl genome browser.

Considering 100 Kb upstream and downstream non-overlapping windows of 20 consecutive SNPs, 61 positional candidate genes for FLM were identified. The window containing the most FLM candidate genes (17 genes) was in BTA3, and the window explaining the highest proportion of variance (1.7%) was in BTA27 (Table 3). The four candidate genes (*EXT1*, *ITGA1*, *ATG7*, *PPARD*) prioritized by GUILDify and ToppGene analysis were in 4 windows, which together explained 5.43% of the total genetic variance. Additionally, skeletal development-related genes *SCUBE3* and *SHOX* that were not prioritized in the ToppGene analysis were identified by visual inspection in the Genome Data Viewer (NCBI, 2022). We also identified 24 genes (starting from *LOC*) for which no associated sign gene was present. There weren't any orthologous genes for these *LOC* that could be interesting to discuss in the FLM context.

3.4. Discussion

The new-born animal must participate in its own nutrition and locomotion to survive. Thus, he is equipped with a set of muscles wich have grown during ante-natal life in such a way that they play their role in ensuring survival. The new-born animal can walk and to stuck because “early developing” muscles of the distal parts of the limbs are well developed at birth (Berg and Butterfield, 1974). During their development, however, the musculoskeletal system is influenced by genetic factors that, if expressed, alter the animal's posture. In addition, environmental factors such as overfeeding, the availability and distribution of food and water, and the topography of grazing environments can also contribute to persistent postural issues (Vargas et al., 2018).

Differences in growth curves and body weight composition between the sexes can be explained by hormonal differences such as androgenic effects (e.g., testosterone) (Young and Bass, 1984; Mangwiro, Dhliwayo and Tayengwa, 2013). In the Nellore breed, there is little difference between male growth and female growth before weaning, but this difference becomes more pronounced with age (Malhado et al., 2009; Arruda et al., 2018). This could explain the observed differences in the incidence of FLM in males and females (Figure 1). The observed overall incidence rate (5.34%) was higher than that reported by Vargas et al. (2017) (4.5%), who used a subsample from a single Nellore Alliance program with approximately one-third of the observations used in this study. The increase in the incidence of FLM in recent years in the same

population may be influenced by the subjectivity of assessment and should be observed with caution. As field technicians gain experience over the years, they tend to place higher discerning demands on the evaluation. “Knock-kneed” (when the knee joints lie inside this line), “bow-legged” (when the knee joints lie outside this line) and “straight-legged” (high angulation of the tibiotarsal joint) (Vargas et al., 2018) are examples of malformation patterns that can be best rated by an experienced technician.

The magnitude of heritability estimated for the FLM depends on the rating assigned to the animal at the time of assessment and may be directly related to the quality of the recording set measured for the classification of the trait (Vargas et al., 2017). Furthermore, using YW as an “anchor”, in a two-trait analysis, may help capture a greater proportion of additional variability. Studies in dairy cows have shown heritability estimates for leg and foot conformation traits to range from 0.03 to 0.22 (Berry et al., 2004; Chapinal et al., 2013; Häggman and Juga, 2013). In the Nellore breed, Passafaro et al. (2013) and Vargas et al. (2018, 2017) obtained moderate to high heritability estimates of the FLM, as estimated in this study, which indicating the potential for improvement through selection.

The assessment of the genetic correlation between YW and FLM (-0.2311) was considered favourable, since in FLM, the lowest taxonomic rank (score 0) indicated that animals did not have feet and legs defects, whereas in YW animals with stronger expression of the trait are required. For example, in the population studied, animals that had difficulty walking due to congenital deformities of the legs and feet (score 1) tended to gain less weight. Consistent with the results of this study, Vargas et al. (2018) estimated the genetic correlation between YW and FLM to be 0.39, on a scale of 1 to 5 (ranging from least ideal to most ideal). This suggests that animals with better legs and feet tend to be heavier. In addition to YW-related responses, which are part of the selection index, estimated genetic trends from FLM suggest that strategies to cull problem animals promote favourable genetic changes and facilitate genetic progression of the trait in the study population.

The GWAS allowed associating genes reported in the literature with traits related to the feet and legs conformation in dairy and beef cattle populations. The literature review, used by (Vargas et al., 2018), is an indispensable step in the process of identifying functional candidate genes mapped around candidate markers. However, if a greater number of positional candidate genes are identified, the literature review may become unfeasible (Fonseca et al., 2018). The strategy of automating gene

prioritization analysis (Martins de Carvalho et al., 2020; Sweett et al., 2020) speeds up the identification of genes related to the regulation of biological processes associated with the phenotype. Fonseca et al. (2018) highlight the importance of combining tools and functional information from multiple sources (and species) to perform a better selection of functional candidate genes and further explore GWAS results.

The specific deletion in the muscle tissue of the *ATG7* gene (BTA22) in mice resulted in profound muscle atrophy and age-dependent decrease in strength (Masiero et al., 2009). The *ATG7* gene (BTA22) encodes an E1-like essential activating enzyme for autophagy and cytoplasmic transport to the vacuole (Gao et al., 2013). The autophagy flow is important to preserve muscle mass and maintain the integrity of the myofiber. In beef cattle, increased *ATG7* gene expression was correlated with skeletal muscle growth and body weight during the fattening period, regardless of the muscle evaluated (Nakanishi et al., 2019).

The *EXT1* gene (BTA14) encodes a glycosyltransferase responsible for heparan sulfate (HS) polymerization, related to fibroblast growth factor (Nadanaka et al., 2008). The interaction between growth factors and their receptors is regulated by the amount of HS. A defect in HS biosynthesis due to somatic mutations in *EXT1* can cause a localized break in the negative feedback loop that regulates chondrocyte proliferation and maturation, allowing premature differentiation and therefore aberrant bone growth (Duncan, McCormick and Tufaro, 2001).

Another prioritized candidate gene, *ITGA1* (BTA20), encodes a subunit of the cell-surface receptor integrin α1β1, which acts in the regulation of hepatic glucose and lipid metabolism under conditions of overnutrition in vivo. On a high fat diet, mice with the *ITGA1* gene inactivated show severe hepatic insulin resistance and decreased hepatic fat accumulation (Williams et al., 2015). In Korean Hanwoo cattle, the *ITGA1* gene was significantly associated with intramuscular fat deposition (Lee et al., 2013; de las Heras-Saldana et al., 2020). Weight gain accelerated by increased fat deposition in young animals can cause stress on ligaments, tendons and joints, promoting gene expression patterns that predispose to alterations in the normal conformation of feet and legs.

The same association can be made for the *PPARD* gene (BTA23), the most abundant form of Peroxisome Proliferator-Activated Receptor (*PPAR*) family in skeletal muscle (Brennan et al., 2009). A key role for the *PPARD* gene in controlling the biological processes that drive ruminal epithelial cell development in new-born calves

was suggested by Naeem et al. (2012). Its increased expression after improved diet suggests an important role in promoting the use of long-chain fatty acids as substrates for oxidation in cell membranes during differentiation in ruminal development. In dairy cows, the *PPARD* gene is involved in muscle fatty acid transport and oxidation during early lactation, and muscle fat replacement from days 3 to 30 of lactation (Schäff et al., 2013). Studies have shown that the gene is highly expressed in the intestinal epithelium, keratinocytes, and liver, consistent with an important biological role in these tissues and influenced by the onset lactation and the type of lipids that are supplied (Girroir et al., 2008; Akbar et al., 2013).

The *SCUBE3* (BTA23) and *SHOX* (BTA2) genes, identified by visual inspection of significant windows of the FLM, have relevant roles in bone formation. The *SCUBE3* gene acts as a BMP2/BMP4 (bone morphogenetic protein) coreceptor and positively regulates signalling possibly enhancing the specific interaction between BMP and BMP type I receptors (Lin et al., 2021). Malfunction of the *SCUBE3* gene has been linked to problems such as craniofacial and dental defects, reduced body size, and defective endochondral bone growth in mice, and osteosarcoma in humans (Liang et al., 2015). The *SCUBE3* and *PPARD* genes are in the same window of BTA23, 79 kb apart. Due to the presence of antagonistic causal mutations in these two genes, selection pressure for YW may affect the FLM, i.e., favourable mutations in YW in *PPARD* and unfavourable mutations in FLM in *SCUBE3*. This highlight elucidates a possible biological mechanism for the observed genetic correlations.

Rafati et al. (2016) found a complete association between the *SHOX* gene deletion and the skeletal atavism of the Shetland pony. It is a genetic disorder characterized by abnormal growth of the ulna and fibula, which elongate the carpal and tarsal joints, respectively. This can lead to abnormal bone structure and restricted movement. The *SHOX* gene is also associated with skeletal defects such as Léri-Weill dyschondrosteosis (Benito-Sanz et al., 2012), short stature, and limb deformities in humans (Jorge et al., 2007; Chen, Wildhardt, et al., 2009; Raudsepp et al., 2012).

3.5. Conclusions

Estimates of additive genetic variability and heritability for feet and legs malformation suggest that selection may reduce the incidence of this problem. The phenotypic incidence shows the importance of reinforcing this selection, although the

estimated trend curve shows a rather favourable genetic gain over the study period. The genetic association of yearling weight with feet and legs malformations may have implications for selection criteria used in breeding programs. The current GWAS was able to identify chromosomal regions associated with feet and legs malformation in Nellore cattle. The roles of prioritized candidate genes in skeletal muscle development (*ATG7*), aberrant bone growth (*EXT1*), lipid metabolism (*ITGA1*, *PPARD*), intramuscular fat deposition (*ITGA1*) and skeletogenesis (*SCUBE3*, *SHOX*) underscore the importance of these genes. Future studies on metabolic pathways and regulatory mechanisms involved in the expression of these candidate genes may contribute to a better understanding of feet and legs malformation and the characterization of causal variants may help improve genomic prediction of this trait in Nellore cattle.

3.6. References

Akbar, H. et al. (2013) 'Dietary lipid during late-pregnancy and early-lactation to manipulate metabolic and inflammatory gene network expression in dairy cattle liver with a focus on PPARs', Gene Regulation and Systems Biology, 2013(7), pp. 103–123. doi: 10.4137/GRSB.S12005.

Arruda, R. M. da S. de et al. (2018) 'Growth curves and nutritional requirements for maintenance of asymptotic weight of Nellore cattle', Revista Ciência Agronômica, 49(4), pp. 692–698. doi: 10.5935/1806-6690.20180078.

Benito-Sanz, S. et al. (2012) 'Identification of the first recurrent PAR1 deletion in Léri-Weill dyschondrosteosis and idiopathic short stature reveals the presence of a novel SHOX enhancer', Journal of Medical Genetics, 49(7), pp. 442–450. doi: 10.1136/jmedgenet-2011-100678.

Berg, R. T. and Butterfield, R. M. (1974) 'Growth of meat animals', in Cole, D. J. A. and Lawrie, R. A. (eds) Meat. Butterwort. London: Elsevier Science, pp. 22–29. Available at: https://books.google.com.br/books?id=TAbLBAAAQBAJ.

Berry, D. P. et al. (2004) 'Genetic relationships among linear type traits, milk yield, body weight, fertility and somatic cell count in primiparous dairy cows', Irish Journal of Agricultural and Food Research, 43(2), pp. 161–176. Available at: https://www.jstor.org/stable/25562515.

Brennan, K. M. et al. (2009) 'Body weight loss in beef cows: I. The effect of increased β-oxidation on messenger ribonucleic acid levels of uncoupling proteins two and three and peroxisome proliferator-activated receptor in skeletal muscle', Journal of Animal Science, 87(9), pp. 2860–2866. doi: 10.2527/jas.2008-1302.

Chapinal, N. et al. (2013) 'Genetic parameters for hoof lesions and their relationship with feet and leg traits in Canadian Holstein cows', Journal of Dairy Science, 96, pp. 2596–2604. doi: 10.3168/jds.2012-6071.

Chen, J., Wildhardt, G., et al. (2009) 'Enhancer deletions of the SHOX gene as a frequent cause of short stature: The essential role of a 250 kb downstream regulatory domain', Journal of Medical Genetics, 46(12), pp. 834–839. doi: 10.1136/jmg.2009.067785.

Chen, J., Bardes, E. E., et al. (2009) 'ToppGene Suite for gene list enrichment analysis and candidate gene prioritization', Nucleic Acids Research, 37(SUPPL. 2), pp. 305–311. doi: 10.1093/nar/gkp427.

Duncan, G., McCormick, C. and Tufaro, F. (2001) 'The link between heparan sulfate and hereditary bone disease: finding a function for the EXT family of putative tumor suppressor proteins', Journal of Clinical Investigation, 108(4), pp. 511–516. doi: 10.1172/JCI13737.

Fonseca, P. A. de S. et al. (2018) 'Genetic mechanisms underlying spermatic and testicular traits within and among cattle breeds: systematic review and prioritization of GWAS results', Journal of Animal Science, 96(12), pp. 4978–4999. doi: 10.1093/jas/sky382.

Fonseca, P. A. S. et al. (2020) 'GALLO: An R package for genomic annotation and integration of multiple data sources in livestock for positional candidate loci', GigaScience, 9(12). doi: 10.1093/gigascience/giaa149.

Gao, W. et al. (2013) 'E1-Like Activating Enzyme Atg7 Is Preferentially Sequestered into p62 Aggregates via Its Interaction with LC3-I', PLoS ONE, 8(9). doi: 10.1371/journal.pone.0073229.

GenSys, C. A. S. L. (2019) Aliança Nelore: Sumário de avaliação de reprodutores. Available at: https://gensys.com.br/sumario/sumario-alianca-nelore-2019-2020/ (Accessed: 18 October 2019).

Geweke, J. (1992) 'Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments', Bayesian Statistics, 4, pp. 641–649.

Gianola, D. and Foulley, J. L. (1983) 'Sire evaluation for ordered', Genetics Selection Evolution, 15(2), pp. 201–224. doi: 10.1016/j.nbt.2013.09.001.

Girroir, E. E. et al. (2008) 'Quantitative expression patterns of peroxisome proliferator-activated receptor-β/δ (PPARβ/δ) protein in mice', Biochemical and Biophysical Research Communications, 371(3), pp. 456–461. doi: 10.1016/j.bbrc.2008.04.086.

Guney, E., Garcia-Garcia, J. and Oliva, B. (2014) 'GUILDify: A web server for phenotypic characterization of genes through biological data integration and network-based prioritization algorithms', Bioinformatics, 30(12), pp. 1789–1790. doi: 10.1093/bioinformatics/btu092.

Häggman, J. and Juga, J. (2013) 'Genetic parameters for hoof disorders and feet and leg conformation traits in Finnish Holstein cows', Journal of Dairy Science, 96, pp. 3319–3325. doi: 10.3168/jds.2012-6334.

Hay, E. H. and Roberts, A. (2018) ‘Genome-wide association study for carcass traits in a composite beef cattle breed’, Livestock Science, 213, pp. 35–43. doi: 10.1016/j.livsci.2018.04.018.

Heidelberger, P. and Welch, P. D. (1983) ‘Simulation Run Length Control in the Presence of an Initial Transient’, Operations Research, 31(6), pp. 1109–1144. doi: 10.1287/opre.31.6.1109.

Hu, Z.-L., Park, C. A. and Reecy, J. M. (2022) ‘Bringing the Animal QTLdb and CorrDB into the future: meeting new challenges and providing updated services’, Nucleic Acids Research, 50(D1), pp. D956–D961. doi: 10.1093/nar/gkab1116.

Irano, N. et al. (2016) ‘Genome-wide association study for indicator traits of sexual precocity in Nellore cattle’, PLoS ONE, 11(8), p. 159502. doi: 10.1371/journal.pone.0159502.

Jorge, A. A. L. et al. (2007) ‘SHOX mutations in idiopathic short stature and Leri-Weill dyschondrosteosis: Frequency and phenotypic variability’, Clinical Endocrinology, 66(1), pp. 130–135. doi: 10.1111/j.1365-2265.2006.02698.x.

de las Heras-Saldana, S. et al. (2020) ‘Differential gene expression in longissimus Dorsi muscle of hanwoo steers—new insight in genes involved in marbling development at younger ages’, Genes, 11(11), pp. 1–17. doi: 10.3390/genes11111381.

Lee, H. J. et al. (2013) ‘Comparative Transcriptome Analysis of Adipose Tissues Reveals that ECM-Receptor Interaction Is Involved in the Depot-Specific Adipogenesis in Cattle’, PLoS ONE, 8(6). doi: 10.1371/journal.pone.0066267.

Legarra, A., Aguilar, I. and Misztal, I. (2009) ‘A relationship matrix including full pedigree and genomic information’, Journal of Dairy Science, 92(9), pp. 4656–4663. doi: 10.3168/jds.2009-2061.

Liang, W. et al. (2015) ‘The Expression of HSPD1, SCUBE3, CXCL14 and Its Relations with the Prognosis in Osteosarcoma’, Cell Biochemistry and Biophysics, 73(3), pp. 763–768. doi: 10.1007/s12013-015-0579-7.

Lin, Y. C. et al. (2021) ‘SCUBE3 loss-of-function causes a recognizable recessive developmental disorder due to defective bone morphogenetic protein signaling’, American Journal of Human Genetics, 108(1), pp. 115–133.

Lourenco, D. et al. (2020) ‘Single-Step Genomic Evaluations from Theory to Practice: Using SNP Chips and Sequence Data in BLUPF90’, Genes, 11(7), p. 790.

Malhado, C. H. M. et al. (2009) ‘Modelos não lineares utilizados para descrever o crescimento de bovinos da raça Nelore no estado da Bahia: 1. Efeito ambiental’, Revista Brasileira de Saúde e Produção Animal, 10(4), pp. 821–829. Available at: https://periodicos.ufba.br/index.php/rbspa/article/view/39913/22306.

Mangwiro, T. N., Dhliwayo, M. and Tayengwa, T. (2013) ‘Non genetic factors affecting calf growth traits in mashona cattle’, Scientific Journal of Animal Science, 2(2), pp. 222–227. doi: 10.14196/sjas.v2i8.854.

MAPA - MINISTÉRIO DA AGRICULTURA PECUÁRIA E ABASTECIMENTO, B. (2006) Instrução Normativa no 53, de 27 de setembro de 2006, Secretaria de Defesa Agropecuária. Available at: http://sistemasweb.agricultura.gov.br/sislegis/action/detalhaAto.do?method=visualizarAtoPortalMapa&chave=735332411 (Accessed: 29 October 2019).

Martins de Carvalho, L. et al. (2020) 'Identifying functionally relevant candidate genes for inflexible ethanol intake in mice and humans using a guilt-by-association approach', Brain and Behavior, 10(12), pp. 1–15. doi: 10.1002/brb3.1879.

Masiero, E. et al. (2009) 'Autophagy Is Required to Maintain Muscle Mass', Cell Metabolism, 10(6), pp. 507–515. doi: 10.1016/j.cmet.2009.10.008.

Masuda, Y. et al. (2019) '331 Efficient quality control methods for genomic and pedigree data used in routine genomic evaluation', Journal of Animal Science, 97(Supplement_3), pp. 50–51. doi: 10.1093/jas/skz258.101.

Melo, T. P. et al. (2016) 'Assessing the value of phenotypic information from non-genotyped animals for QTL mapping of complex traits in real and simulated populations', BMC Genetics, 17(1), pp. 1–9. doi: 10.1186/s12863-016-0394-1.

Misztal, I. et al. (2014) 'Manual for BLUPF90 family of programs'. Athens, GA: Animal and Dairy Science, University of Georgia. Available at: http://nce.ads.uga.edu/wiki/doku.php?id=start.

Nadanaka, S. et al. (2008) 'Chondroitin 4-O-sulfotransferase-1 modulates Wnt-3a signaling through control of E disaccharide expression of chondroitin sulfate', Journal of Biological Chemistry, 283(40), pp. 27333–27343. doi: 10.1074/jbc.M802997200.

Naeem, A. et al. (2012) 'Role of metabolic and cellular proliferation genes in ruminal development in response to enhanced plane of nutrition in neonatal Holstein calves', Journal of Dairy Science, 95(4), pp. 1807–1820. doi: 10.3168/jds.2011-4709.

Nakanishi, T. et al. (2019) 'Changes in expression of the autophagy-related genes microtubule-associated protein 1 light chain 3β and autophagy related 7 in skeletal muscle of fattening Japanese Black cattle: A pilot study', Asian-Australasian Journal of Animal Sciences, 32(4), pp. 592–598. doi: 10.5713/ajas.18.0370.

NCBI, N. C. for B. I. (2022) cattle (Bos taurus) - NCBI Genome Data Viewer, National Library of Medicine (US). Available at: https://www.ncbi.nlm.nih.gov/genome/gdv/?org=bos-taurus (Accessed: 4 May 2022).

Oliveira, H. R. et al. (2019) 'Genome-wide association for milk production traits and somatic cell score in different lactation stages of Ayrshire, Holstein, and Jersey dairy cattle', Journal of Dairy Science, 102(9), pp. 8159–8174. doi: 10.3168/jds.2019-16451.

Passafaro, T. L. et al. (2013) 'Parâmetros genéticos para escores visuais de tourinhos Nelore com modelos lineares e de limiar.', in X Simpósio Brasileiro de Melhoramento Animal Uberaba, MG – 2013. Available at: http://sbmaonline.org.br/anais/x/trabalhos/pdf/6NRQ.pdf.

Pérez O'Brien, A. M. et al. (2014) 'Linkage disequilibrium levels in Bos indicus and Bos taurus cattle using medium and high density SNP chip data and different minor allele frequency distributions', Livestock Science, 166(1), pp. 121–132. doi: 10.1016/j.livsci.2014.05.007.

Plummer, M. et al. (2006) 'CODA: Convergence Diagnosis and Output Analysis for MCMC', R News, 6(1), pp. 7–11. Available at: https://www.r-project.org/doc/Rnews/Rnews_2006-1.pdf (Accessed: 2 November 2020).

R Development Core Team (2017) 'R: A language and environment for statistical computing.' Vienna, Austria.: R Foundation for Statistical Computing. Available at: https://www.r-project.org/.

Rafati, N. et al. (2016) 'Large deletions at the SHOX locus in the pseudoautosomal region are associated with skeletal atavism in shetland ponies', G3: Genes, Genomes, Genetics, 6(7), pp. 2213–2223. doi: 10.1534/g3.116.029645.

Raudsepp, T. et al. (2012) 'The pseudoautosomal region and sex chromosome aneuploidies in domestic species', Sexual Development. Karger Publishers, pp. 72–83. doi: 10.1159/000330627.

Roso, V. M. and Schenkel, F. S. (2006) 'AMC-a computer program to assess the degree of connectedness among contemporary groups', in Proceedings of the 8th World Congress on Genetics Applied to Livestock Production, Belo Horizonte, Minas Gerais, Brazil, 13-18 August, 2006. Instituto Prociência, pp. 26–27. Available at: http://www.wcgalp8.org.br.

Sargolzaei, M., Chesnais, J. P. and Schenkel, F. S. (2014) 'A new approach for efficient genotype imputation using information from relatives', BMC Genomics, 15(1), p. 478. doi: 10.1186/1471-2164-15-478.

Schäff, C. et al. (2013) 'Increased muscle fatty acid oxidation in dairy cows with intensive body fat mobilization during early lactation', Journal of Dairy Science, 96(10), pp. 6449–6460. doi: 10.3168/jds.2013-6812.

Smith, B. J. (2007) 'boa : An R Package for MCMC Output Convergence Assessment and Posterior Inference', Journal of Statistical Software, 21(11), pp. 1–37. doi: 10.18637/jss.v021.i11.

Sorensen, D. and Gianola, D. (2002) 'Likelihood, Bayesian and MCMC methods in quantitative genetics', in Dietz, K. et al. (eds) Statistics for Biology and Health. 2nd edn. New York, NY: Springer, p. 740.

Sweett, H. et al. (2020) 'Genome-wide association study to identify genomic regions and positional candidate genes associated with male fertility in beef cattle', Scientific Reports, 10(1), pp. 1–14. doi: 10.1038/s41598-020-75758-3.

Vallejo, R. L. et al. (2016) 'Evaluation of genome-enabled selection for bacterial cold water disease resistance using progeny performance data in rainbow trout: Insights on genotyping methods and genomic prediction models', Frontiers in Genetics, 7(MAY), p. 96. doi: 10.3389/fgene.2016.00096.

Vargas, G. et al. (2017) ‘Genetic analysis of feet and leg conformation traits in Nelore cattle’, Journal of Animal Science, 95(6), p. 2379. doi: 10.2527/jas2016.1327.

Vargas, G. et al. (2018) ‘Genome-wide association study and functional analysis of feet and leg conformation traits in Nellore cattle’, Journal of Animal Science, 96(5), pp. 1617–1627. doi: 10.1093/jas/sky079.

Wang, H. et al. (2012) ‘Genome-wide association mapping including phenotypes from relatives without genotypes’, Genetics Research, 94(2), pp. 73–83. doi: 10.1017/S0016672312000274.

Williams, A. S. et al. (2015) ‘Integrin α1-null mice exhibit improved fatty liver when fed a high fat diet despite severe hepatic insulin resistance’, Journal of Biological Chemistry, 290(10), pp. 6546–6557. doi: 10.1074/jbc.M114.615716.

Young, O. A. and Bass, J. J. (1984) ‘Effect of castration on bovine muscle composition’, Meat Science, 11(2), pp. 139–156. doi: 10.1016/0309-1740(84)90011-1.

Zhang, X. et al. (2016) ‘Weighting strategies for single-step genomic BLUP: An iterative approach for accurate calculation of GEBV and GWAS’, Frontiers in Genetics, 7(AUG), p. 151. doi: 10.3389/fgene.2016.00151.

Zhuang, Z. et al. (2020) ‘Weighted single-step genome-wide association study for growth traits in chinese simmental beef cattle’, Genes, 11(2), p. 189. doi: 10.3390/genes11020189.

CAPÍTULO 4 - Unraveling genetic architecture of testicular hypoplasia and feet and legs malformation in Nellore cattle by genome wide association studies of imputed whole-genome sequence

Abstract – In order to find candidate genes and/or Quantitative Trait Loci (QTLs) significantly associated with Testicular Hypoplasia (TH) and Feet and Legs Malformation (FLM) in Nellore cattle, Estimated Breeding Values of 17,260 and 12,513 progeny-tested bulls, respectively, were used in Genome-wide Association Studies along with their imputed genotypes for Whole-genome sequence having as a reference population 151 bulls from the Alliance Nellore database. A SNP-BLUP model was fitted in a C/C++ routine to estimate marker effects and significant SNPs were determined using a genome-wide false discovery rate of 5%. A total of 262 SNPs were identified with p-value below the alternative significance threshold for TH and 80 SNPs for FLM. Fourteen candidate genes were prioritized by ToppGene analysis for TH and seventeen genes were identified for FLM. The genes identified as potentially associated are involved in reproductive aspects of bovines and short stature, respectively. Gene Ontology terms were used to perform an enrichment analysis and to detect functional information from the identified candidate genes and QTLs. Traits reported as associated to annotated QTLs for TH include ‘age at puberty’, ‘scrotal circumference’, ‘first service conception’, ‘calving interval’ and ‘sperm counts’. For FLM, traits associated with annotated QTLs include 'connective tissue amount' and 'loin strength’. Functional clustering of candidate genes by similar GO terms was also performed to identify potential related biological processes. Key roles of biological pathways and processes related to spermatogenesis, macrophage-mediated inflammation and control of tumor cell development were detected. We concluded that the risk of TH or FLM is driven by a large number of small-effect variants, potentially implicating most of the regulatory variants that are active in tissues relevant to these malformations. The observation that the heritability of this disqualifying traits is widely distributed across the genome seems inconsistent with the expectation that their causal variants are primarily in specific biologically relevant pathways and genes.

Key words: Candidate genes, functional clustering, imputed sequence, testis measures, limb deformities.

4.1. Introduction

The central objective of animal breeding is to promote the improvement of herds by understanding the association between genetic variations and phenotypes in traits of economic interest. It is known that many genes contribute to the expression of complex traits. Genome-wide Association Studies (GWAS) have been used to detect associations between molecular markers and these traits, contributing to a better understanding of the genetic and physiological mechanisms that regulate them, in addition to pointing out candidate genes for future investigations of causal mutations (Pereira et al., 2013). Several relatively common congenital abnormalities and familial disorders result from multifactorial inheritance, which raises questions about the relative importance of each candidate gene in the expression of complex traits. Given their relevance to beef cattle production systems, the disqualifying traits Testicular Hypoplasia (TH) and Feet and Legs Malformation (FLM) have been widely studied (Vargas et al., 2017; Vargas et al., 2018; Neves et al., 2019). Especially in the Nellore breed, due to its importance for the global beef industry.

Whole-genome sequence (WGS) data is expected to capture genetic variation more completely than common genotyping panels (Veerkamp et al., 2016). The pedigree structure of bovine populations makes it possible to impute a whole-genome sequence for each genotyped individual, thus generating a discovery population with thousands of animals, each with a genome sequence. Genome-wide sequencing data helped Veerkamp et al. (2016) to capture genetic variation linked to QTLs more completely compared to studies done with low- or high-density genotyping panels, in addition to resulting in clearer GWAS peaks across the genome. Thus, GWAS performed with WGS data could enable better identification of genomic regions associated with the traits of interest and, therefore, optimize the design of genotyping panels.

When the increased resolution in a window with a high pattern of linkage disequilibrium (LD) between different SNP (Single Nucleotide Polymorphism) markers shows several genes, it becomes difficult to infer a possible candidate gene. However, if there is only a single gene, it becomes a candidate to have a significant effect on the trait of interest. It is noteworthy that an association between an SNP marker and a trait does not necessarily mean that a gene in that region affects the trait, it is necessary to perform functional annotation studies to infer more precisely (Verardo et al., 2016).

Using WGS data in a GWAS of TH and FLM in Nelore cattle, the goals of this work were: i) to find candidate genes and/or QTLs significantly associated with these traits; ii) to do a Gene Ontology (GO) enrichment analysis to detect functional information and GO terms associated with the identified genes and QTLs; iii) to do a Functional clustering of candidate genes by similar GO terms to identify potential candidate biological processes and clusters of related genes.

4.2. Materials and methods

Approval from the Welfare and Animal Use Committee was not required for this study because the data was obtained from an existing database of phenotypic and genotypic records of Nellore cattle. Previous studies using these data also did not need this approval.

4.2.1. Phenotypes

Morphological and functional defects or disorders, like TH and FLM, are systematically evaluated as binary traits at weaning and yearling ages in the herds to which the animals belong. Phenotype and pedigree information of 368,948 and 468,302 Nellore cattle evaluated between 2001 and 2017 for TH and FLM, respectively, were collected from the Alliance Nellore database (GenSys, 2019) and considered for this study. The Estimated Breeding Values (EBV), obtained in chapters 2 (for TH) and 3 (for FLM) of this thesis, were used as phenotypes.

Contemporary groups (CG) were defined by concatenating sex (only for FLM), year and season of birth, farm and management group at weaning and yearling, evaluation date at yearling. Data from CG with fewer than 10 records and/or without variability in the traits were removed from further analyses. Connectedness among CG was checked using the AMC software (Roso and Schenkel, 2006) and only CG with at least 10 genetic links to other CG were kept for analysis.

4.2.2. Genotypes

In this study, 17,260 and 12,513 animals for TH and FLM, respectively, were genotyped using the Illumina BovineHD Genotyping BeadChip (HD; Illumina, Inc., San

Diego, CA, USA) (612,174 SNPs) or with a lower density panel (from 20K to 90K SNPs), and had their genotypes imputed to the sequence level. First, the animals genotyped with the lower density panels had their genotype imputed to the HD panel. Then, the HD genotypes were imputed to the sequence. All genotyped individuals were imputed to WGS using the software FImpute3 (Sargolzaei et al., 2014) and 151 sires from the Alliance Nellore database (GenSys, 2019) as reference population. The 151 Nellore sires were sequenced using the Illumina HiSeq X™ Ten and Illumina NovaSeq™ platforms (Illumina Inc., San Diego, USA). Sequencing of all samples was performed using the Illumina sequencing-by-synthesis technology at an overall average sequence coverage of 14.5×, ranging from 7.8 to 26.3×. The SAMtools program (Li et al., 2009) was used to index the sequences, and the alignment to the ARS-UCD1.2 reference genome was performed using HiSat (Kim et al., 2015). Non-biallelic markers and those with a minor allele frequency lower than 0.01 were excluded. Markers with a phred-scaled confidence (a genotype quality score) less than 15 were treated as missing and those SNPs with missing values for more than 40 individuals (26.5% of the total population) were removed. A total of 23,837,667 sequence SNPs in autosomal regions remained.

After imputation SNP markers were pruned using the indep-pairwise option in PLINK 2.0 (Chang et al., 2015). A window size of 130kb, a step of 50 markers, and a r^2 threshold of 0.9 resulted in 4,559,436 and 4,508,230 independent SNPs for subsequent analysis of TH and FLM, respectively.

4.2.3. GWAS using sequence information

A genome-wide association study (GWAS) for TH and FLM in Nellore cattle (*Bos primigenius indicus*) was performed using estimated breeding values (EBVs) of 17,260 and 12,513 progeny-tested bulls, respectively, genotyped and imputed for sequence data. The SNP-BLUP model was fitted in a C/C++ routine to estimate marker effects:

$$y = 1\mu + Zu + e,$$

where y is the vector of EBVs, μ is the intercept, u is the vector of random unknown marker effects, Z is the incidence matrix, and e is the vector of random residuals. The genomic relationship matrix (GRM) was calculated using the SNP panel with all the

animal samples. The marker effects u_i were assumed to be identically and independently distributed with mean zero and variance σ_g^2, that is $u \sim N(0, G\sigma_g^2)$, where G is the realized genomic relationship matrix calculated from all remaining SNPs, and $e \sim N(0, I\sigma_e^2)$. The percentage of genetic variances explained by each significant SNP was calculated according to the following formula:

$$\%V_g = \frac{2pq\alpha^2}{\sigma_g^2} * 100$$

Where p_i and q_i are the allele frequencies for the i_{th} SNP, α_i is the estimated additive genetic effect of the i_{th} SNP on TH or FLM, σ_g^2 is the estimated genetic variance.

Quantile-quantile plots were generated using the R package *qqman* (Turner, 2014). The genomic inflation factor (λ; Yang et al., 2011) was calculated to evaluate the extent of false positive signals by dividing the observed median value of the Chi-squared statistic for p-values (obtained from GWAS) by the expected median value of the Chi-squared statistic (approximately 0.456 for 1 df tests). Significant SNPs were determined using a genome-wide false discovery rate (FDR) of 5% (Benjamini and Hochberg, 1995). For further analysis, SNP markers with a −log10(P-value) which exceeded the threshold (i.e., 0.05/number of independent tests) were statistically significant and were further annotated into coding regions (genes) of the cattle genome. This approach was chosen as it provides a higher power while controlling false discovery rate. The negative logarithm of the P-value for each SNP was displayed in a Manhattan plot and a genome wide line for the 5% FDR was drawn to display significant SNPs.

4.2.4. Gene annotation and prioritization

Candidate genes were selected using the file Bos_taurus.ARS-UCD1.2.104.gtf, corresponding to the bovine gene annotation of the reference map ARS-UCD1.2 (Hu et al., 2022), and the Genomic Annotation in Livestock for positional candidate Loci (GALLO; Fonseca et al., 2020) package in R version 4.1.2 (R Development Core Team, 2017). Because of the linkage disequilibrium (LD) effect, we evaluated the presence of genes in adjacent 100 kb windows upstream and downstream of the significant SNP markers.

Using the same “trained lists” of the top 100 genes associated with relevant keywords for TH and FLM already used in Chapters 2 and 3, respectively, a functional candidate gene prioritization was conducted in the ToppGene Suite (Chen et al., 2009). Gene Ontology terms (molecular function, biological process, and cellular component), human and mouse phenotypes, metabolic pathways, PubMed publications, coexpression patterns and diseases were used to retrieve the functional information from genes in training and testing lists (Fonseca et al., 2018). The genes on the "test” list with the same functional profile as the genes of “trained” list, based on a multiple correction false discovery rate of 5% (p-value ≤ 10-3) (Chen et al., 2009) were selected as prioritized.

4.2.5. Enrichment analysis

A QTL enrichment analysis try to identify the potential effects of polymorphisms associated to any of the six trait classes (e.g., ‘Reproduction’, ‘Milk’, ‘Production’, ‘Exterior’, ‘Meat and Carcass’, and ‘Health’). To reduce the impact of bias of investigation for some related traits (such as milk production related traits) in the QTL Database for cattle, a QTL enrichment analysis was performed with GALLO package (Fonseca et al., 2020). This analysis consists of a hypergeometric test using the number of QTLs annotated within the candidate regions and the total number of the same QTL in the QTL_ARS_UCD1.gff file, from Animal QTL Database (Hu et al., 2013). QTL annotation uses the ARS-UCD1.2 assembly coordinates and the p-values for the QTL enrichment status of each annotated trait were calculated and corrected for multiple testing using FDR (5%).

The list of candidate genes was used to retrieve their functional information and to identify Gene Ontology (GO) terms (e.g., biological process, molecular function, and cellular component) metabolic pathways, PubMed publications, coexpression patterns and diseases (Fonseca et al., 2018). The R packages ClusterProfiler (Yu et al., 2012) and enrichplot (Wu et al., 2021) were used for enrichment analysis, graphic representation and functional clustering of GO terms of “test” genes (Silva et al., 2023). Genes and GO terms were considered enriched when the FDR was < 0.05.

4.3. Results

Deviations from the identity line to the left of the Q-Q plot suggest strong association of TH and FLM with the SNPs (Figure 1). The genomic inflation factors (lambda value) obtained from the QQ plots were 0.87 and 0.88, indicating a certain degree of deviation between the predicted and actual data of TH and FLM, respectively, therefore no adjustment was necessary.

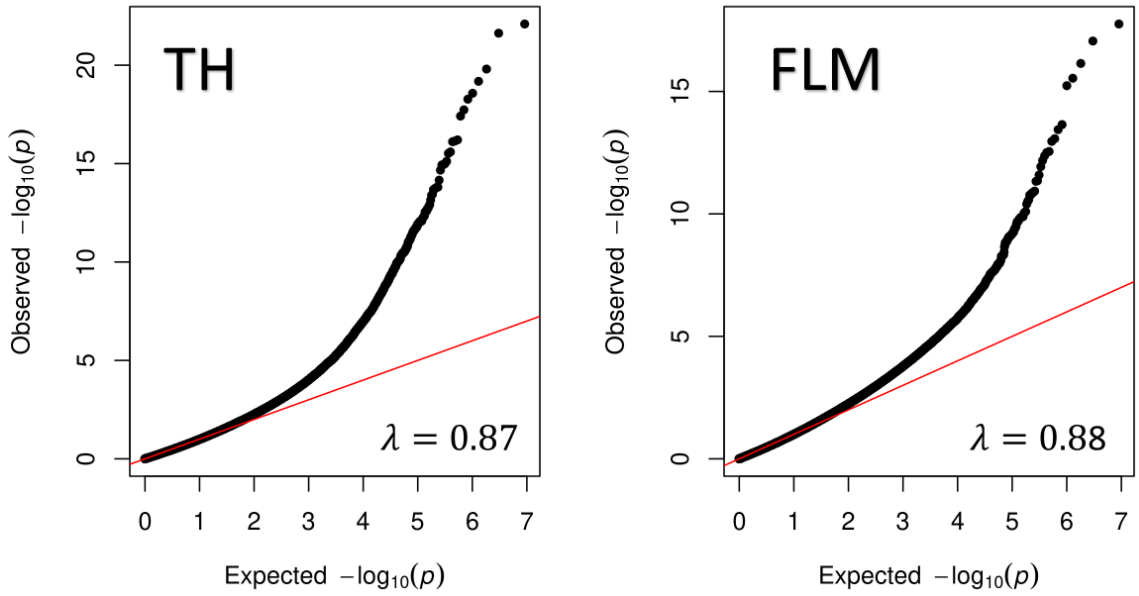

Figure 1. Quantile-quantile plot showing the observed vs. expected p-values outputted from the GWAS analyses for TH and FLM. The red line represents the 95% concentration band under the null hypothesis of no association among traits and SNPs.

There are many causal variants with small effects on both studied traits and they are widely spread throughout the genome, so that in most autosomal chromosome's windows of 200 kb around significant SNPs contributed to genetic variation. In the GWAS with sequence data for TH, 262 SNPs were identified with p-value below the alternative significance threshold ($p < 1\times10^{-8}$) in 25 of the 29 autosomal chromosomes (the exceptions were BTA5, BTA16, BTA19 and BTA27) (Figure 2), together they explained 0.29% of the total additive genetic variance. The chromosomes with the greatest number of significant SNPs found were BTA1 (n = 34) and BTA3 (n = 38). The most significant ($p = 8.09\times10^{-23}$) SNP on BTA18 explained 0.003% of the genetic variance and was at 62,975,859 bp position.

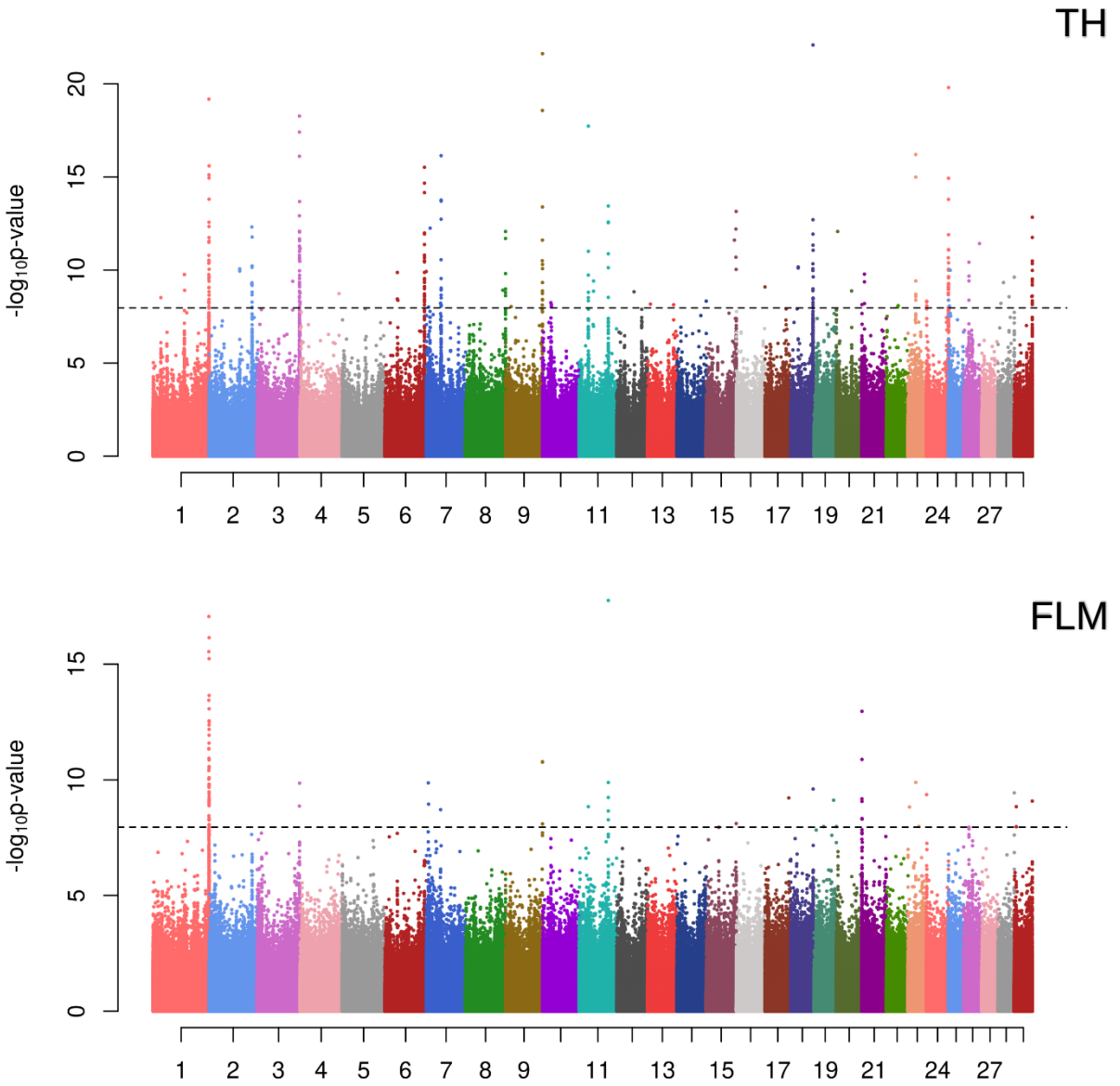

Figure 2. Manhattan plot from the genome-wide association analysis for TH and FLM. The dashed line indicates the Bonferroni statistical threshold.

For FLM, 80 SNPs on 14 autosomal chromosomes (BTA1, BTA3, BTA7, BTA9, BTA11, BTA15, BTA17, BTA18, BTA19, BTA21, BTA23, BTA24, BTA28 and BTA29) had a p-value below the threshold ($p < 1\times10^{-8}$) and together explain 0.09% of the total additive genetic variance. The most significant SNP on BTA11 was located at 82,842,082 bp with the p-value of 1.77×10^{-18}, and 0.002% of the genetic variance was explained by this SNP.

Fourteen candidate genes, based on ARS-UCD1.2 (Hu et al., 2022) and the 100 kb upstream and downstream distances, were prioritized by ToppGene analysis for TH and seventeen positional candidate genes were identified for FLM (Table 1).

Table 1. Potential candidate genes related to the identified SNPs significantly associated with TH and FLM in Nellore cattle.

Trait	BTA[1]	Gene Start (bp)	Gene End (bp)	p-value	Gene Name[2,3]
TH	28	45,757,544	45,766,377	0.00066	*AGT*
	18	63,213,519	63,227,723	0.00149	*MBOAT7*
	1	157,545,780	157,559,387	0.00914	*PRDM9*
	9	104,093,196	104,106,538	0.01268	*TBP*
	8	105,351,900	105,617,709	0.01293	*PAPPA*
	3	120,482,836	120,511,027	0.02116	*SEPTIN2*
	3	120,646,406	120,658,685	0.02161	*BOK*
	11	82,862,083	82,866,378	0.02177	*ODF3*
	2	121,225,598	121,257,595	0.02217	*HDAC1*
	28	8,099,433	8,188,439	0.02563	*TBCE*
	11	82,879,488	82,902,831	0.02908	*SIRT3*
	1	158,238,278	158,253,085	0.04191	*KPNA6*
	6	117,721,521	117,225,838	0.04457	*CTBP1*
	13	7,894,611	7,908,738	0.04508	*FLRT3*
FLM	28	45,757,544	45,766,377	0.00136	*AGT*
	19	27,312,041	27,314,130	0.00407	*CD68*
	19	27,283,274	27,291,150	0.01065	*TNFSF12*
	19	27,345,874	27,350,673	0.01859	*SHBG*
	19	55,350,120	55,358,160	0.02014	*SPHK1*
	7	5,260,712,	5,273,727	0.02091	*SLC5A5*
	11	82,789,488	82,902,831	0.02324	*SIRT3*
	1	158,326,443	158,359,986	0.02653	*KHDRBS1*
	11	82,739,293	83,207,344	0.0275	*NBAS*
	19	27,292,060	27,294,941	0.02808	*TNFSF13*
	1	157,545,780	157,559,387	0.02885	*PRDM9*
	7	6,175,404	6,177,278	0.02924	*F2RL3*
	19	55,249,924	55,275,793	0.03176	*RHBDF2*
	11	82,862,083	82,866,378	0.0335	*ODF3*
	3	119,509,842	119,555,458	0.03737	*P2RY8*
	29	50,933,708	50,934,839	0.04183	*IFITM1*
	18	63,260,500	63,271,290	0.04706	*PRPF31*

[1]BTA – *Bos taurus* autosome; [2]Prioritized candidate genes identified by GUILDify and ToppGene analysis, considering 100 Kb upstream and downstream of the window; [3]Official gene symbol (assembly ARS-UCD1.2).

The QTL enrichment analysis performed by GALLO package revealed 2,595 significant QTLs (FRD-corrected P-value ≤ 0.05) annotated for the positions of significant SNPs from TH and 604 from FLM. Of the TH QTLs, 111 were associated with ‘Reproduction’ phenotypic class and the traits reported as associated to annotated QTLs include ‘age at puberty’ (38%), ‘scrotal circumference’ (15%), ‘first service

conception’ (14%), ‘calving interval’ (28%) and ‘sperm counts’ (5%). For FLM, 'Meat and Carcass' and 'Exterior Conformation' traits associated with annotated QTLs include 'connective tissue amount' and 'loin strength', respectively.

Table 2. Significant enriched QTL identified annotated for TH and FLM in Nellore cattle.

Trait	QTL Type	QTL Name	CHR	Total Annotated QTLs	Adjusted p-value
TH	Reproduction	Age at puberty	21	15	0.0000066
		Scrotal circumference	9	16	0.0020452
		First service conception	9	16	0.0022895
		Calving interval	18	31	0.0173525
		Age at puberty	17	2	0.0227638
		Age at puberty	7	25	0.0236566
		sperm counts	8	6	0.0421388
FLM	Meat and Carcass	Connective tissue amount	18	1	0.0205202
	Exterior	Loin strength	28	2	0.0205202

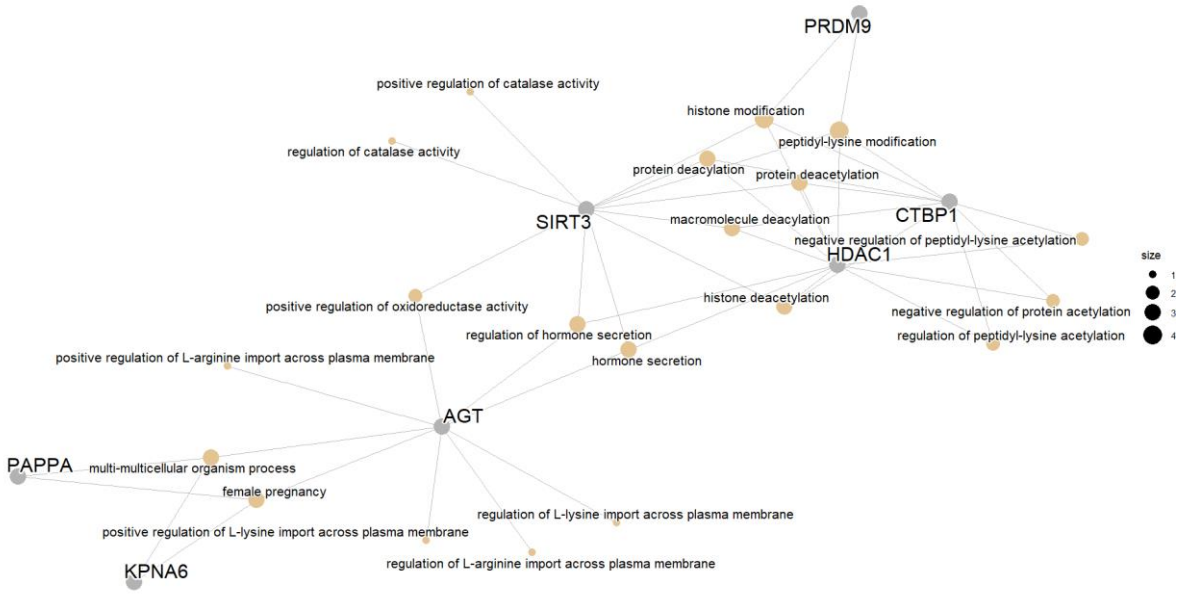

Figure 3. Gene Ontology enrichment analysis of the candidate genes for TH. Each node represents a gene, and each edge represents the overlap between two genes. For better visualization, only the top 20 significant (p value < 0.05) terms are shown.

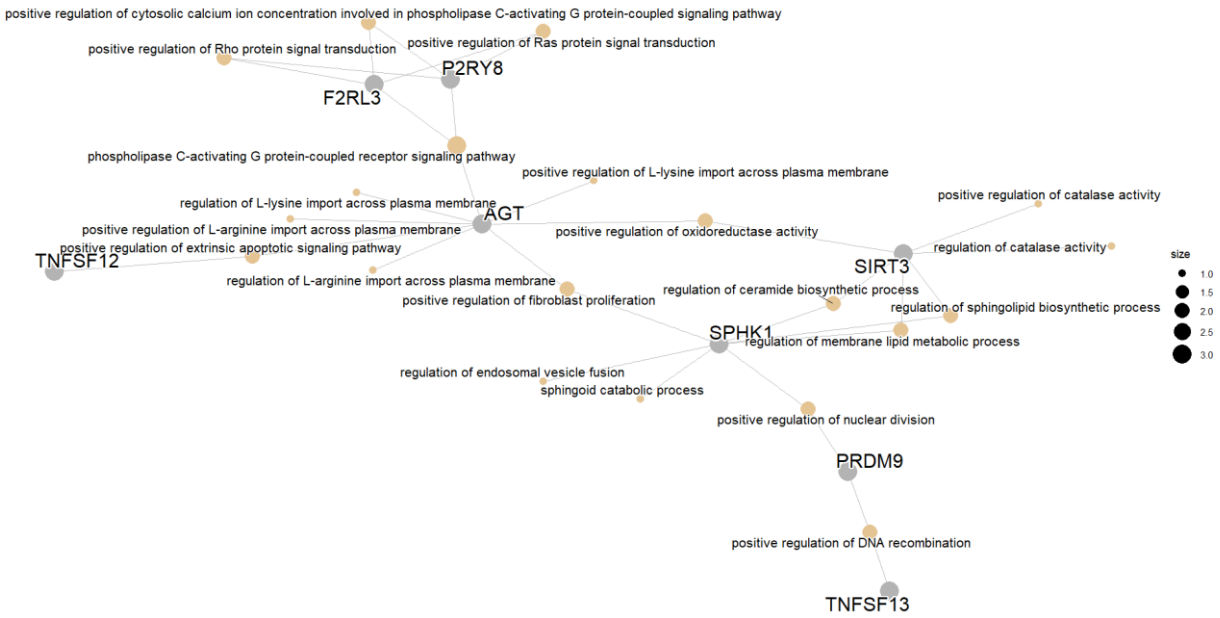

Figure 4. Gene Ontology enrichment analysis of the candidate genes for FLM. Each node represents a gene, and each edge represents the overlap between two genes. For better visualization, only the top 20 significant (p value < 0.05) terms are shown.

Figures 3 and 4 show cluster plots representing the relationships between the genes prioritized in the functional annotation study for TH and FLM, respectively, and Biological Process (BP) GO terms. The complete tables with all the results of the enrichment analysis can be found in the supplementary material (Appendix B and C). Gene ontology enrichment analyses performed for all prioritized genes for TH to identify genes involved in known metabolic pathways revealed that the identified potential candidate genes were involved in 113 biological process, 24 cellular component, and 27 molecular functions (Appendix B). The main pathway clusters of the Biological Process (BP) GO terms are hormone protein acetylation activity, gliogenesis, import across plasma membrane, entry female genome complex, cell migration involved morphogenesis (Appendix D).

The same Analyses for FLM revealed that the identified candidate genes were involved in 65 biological process, 3 cellular component, and 20 molecular functions (Appendix C). The main pathway clusters of the Biological Process (BP) GO terms include angiotensin-mediated apoptotic across plasma, ceramide catabolic biosynthetic process, catalase oxidoreductase peroxidase activity, protein calcium phospholipase C-activating, histone H3-K4 division trimethylation (Appendix E).

4.4. Discussion

The GWAS with imputed whole-genome sequence data may be effective to detect candidate genes for TH and FLM. However, even the most significant SNPs in the genome have small effect sizes and together explain only a modest fraction of the additive genetic variance. For both traits examined, significant SNP were found in most chromosomes. According to Boyle et al. (2017), each chromosome's contribution to the heritability of complex traits tends to be closely related to its physical length (Shi et al., 2016; Visscher et al., 2006), this suggests that causal variation may be distributed in a similar way, quite uniform.

4.4.1. Testicular Hypoplasia

For HT, in several scientific works, genes such as *AGT* (Gur et al., 1998), *PRDM9* (Sandor et al., 2012; Zhou et al., 2018), *PAPPA* (Giroto et al., 2019), *HDAC1* (Dang et al., 2022) appear to be associated with encoding proteins related with reproductive aspects of bovines and mice, although without any evidence of a direct effect on germinal epithelial formation.

Still for TH, terms of Gene Ontology (GO) show the involvement of genes such as *SEPTIN2* and *ODF3* in biological processes such as cell differentiation (GO:0030154) and spermatogenesis (GO:0007283). Furthermore, *SEPTIN2* appears as a cellular component in the sperm ring (GO:0097227) and *ODF3* is a component of sperm flagella outer dense fibers, which add stiffness, elastic recoil, and protection against shearing forces during sperm movement. The *MBOAT7* gene is involved in the regulation of the metabolic process of triglycerides (GO:0090207), essential in the synthesis of reproductive hormones such as testosterone. *TBC* cofactor *E* (*TBCE*) participates in biological processes such as developmental growth (GO:0048589), which is the increase in size or mass of an entire organism, a part of an organism or a cell, resulting in the progression of the organism from one condition to another, as well as in post-embryonic development (GO:0009791). Other genes identified are involved in various cellular metabolic processes and act as anti- or pro-apoptotic regulators, such as *SIRT3* or *BOK* (Jääskeläinen et al., 2010), involved in cell proliferation (*CTBP1*) and cell adhesion (*FLRT3*), nucleocytoplasmic transport (*KPNA6*) and RNA transcription (*TBP*) (Stelzer et al., 2016).

4.4.2. Feet and Legs Malformation

For FLM, the absence of evident peaks in the autosomal chromosomes verified in the Manhattan plot indicates the lack of a segment of the genome that explains a substantial part of the genetic variation. It is believed that this may be related to the architecture of this binary trait, which in its evaluation considered only the occurrence or not of a malformation, regardless of the affected limb, location, type of tissue, among other factors that would characterize it. The exception is the peak in BTA1, where the *KHDRBS1* (158,326,443-158,359,986 kb) and *PRDM9* (157,545,780-157,559,387 kb) genes are found. X-Linked Hereditary Ataxia and Spastic Paraplegia 4, Autosomal Dominant (SPG4) are diseases associated with these two genes, respectively (Stelzer et al., 2016). Both are characterized by X-linked inheritance where it is possible to observe gait changes due to neuromotor impairment in the limbs. What could lead to a positive diagnosis of FLM in the field given the subjectivity of the evaluation method.

Among the other prioritized candidate genes for FLM, mutations in the *NBAS* gene identified from the GWAS using the whole-genome sequence data and located between 82,739,293-83,207,344 bp on BTA11 are associated with short stature (Park and Lee, 2021). The *SIRT3* and *ODF3* genes that also appear with a significant p-value for FLM in BTA11, between 82,789,488-82,902,831 kb and 82,862,083-82,866,378 kb, respectively, may be in LD with significant SNPs associated with the *NBAS* gene, which would be promoting a bias in the automated prioritization analysis by ToppGene. The *PPHK1*, *F2RL3* and *P2RY8* genes encode receptors that, coupled to G protein, transmit signals from hormones and neurotransmitters, controlling the metabolism of cellular machinery, such as contraction, transcription and secretion (Stelzer et al., 2016).

The *AGT* candidate gene has as enriched GO terms positive regulation of macrophage-derived foam cell differentiation (GO:0010744) and regulation of extracellular matrix assembly (GO:1901201). Furthermore, the CD68 gene encodes a transmembrane glycoprotein that is highly expressed by tissue macrophages. Macrophages are associated with repair by releasing growth and angiogenic factors and by regulating T cell function, promoting remodeling of damaged tissue. The presence of complex macrophage phenotypes, mainly in the synovial membrane, may influence the progression of osteoarthritis (Topoluk et al., 2018).

The *TNFSF12* and *TNFSF13* genes encode proteins that are members of the tumor necrosis factor (TNF) ligand family, capable of inducing apoptosis through their interaction with other proteins of the same family, and are therapeutic targets for the management of complex diseases in cattle (Zhang et al., 2010). In addition to playing a role in regulating tumor cell growth, *TNFSF13* may be involved in macrophage-mediated immune processes (Stelzer et al., 2016). *TNFSF12* promotes endothelial cell proliferation and migration and therefore acts as a regulator of tumor angiogenesis (Tao et al., 2022). The *RHDBF2* gene also regulates a TNF-binding protease, thus playing a role in cell survival, proliferation, migration and inflammation. (Stelzer et al., 2016).

Among the other genes, *SHBG* and *IFITM1* are related to the bovine fertility (Fritz et al., 2013; Pru et al., 2001). The protein encoded by the *SLC5A5* gene is responsible the uptake of iodine in tissues such as the thyroid and lactating breast tissue, and mutations in this gene are associated with thyroid dyshormonogenesis. PRPF31 is one of several genes that cause retinitis pigmentosa (Stelzer et al., 2016).

4.4.3. Signal distribution and enrichment in GWAS

Most diseases can be directly affected by a modest number of genes or gene pathways with specific roles in their etiology (Chakravarti and Turner, 2016). These "core genes" will tend to have biologically interpretable roles in the disease. For a variety of complex traits such as TH and FLM, the highest effecting variants are modestly enriched in specific genes or pathways (Table 2) that may play direct roles in the trait. It was expected that malformation-causing variants clustered in key pathways leading to their etiology. However, as shown in figure 2, the SNPs that contribute the most heritability tend to be scattered throughout the genome and are not close to genes with trait-specific functions.

The enrichment of association signals in relevant genes were generally weak and tended to be spread over most of the genome, including many genes without an obvious connection to these traits. Thus, at the functional level of genomic organization, the various genes involved would be functionally linked through gene regulatory networks (GRN) within which one or more genes are commonly dysregulated in all cases of a disorder (Chakravarti and Turner, 2016).

When disease risk is mediated by multiple cell types or highly specialized cell types, we predict that cell networks will vary across cell types. The quantitative effect of any given variant would then be an average of its effect size on each cell type, weighted by the importance of the cell type (Price et al., 2011; Sonawane et al., 2017). Thus, most genes expressed in trait-relevant cell types may make non-zero contributions to heritability, whereas core genes apparently contribute only a small amount to total heritability. Although the degree of risk of disqualifying traits increased by larger effect size mutations varies between traits (Krumm et al., 2015; Marouli et al., 2017), it is expected that core genes, damaged by loss of function, or other mutations particularly harmful, still contribute more strongly to liability.

Future genomic prediction studies may investigate the set of significant genetic variants reported here, seeking to improve selection and mating decisions with the aim of reducing the incidence of Testicular Hypoplasia and Feet and Legs Malformation in Nelore cattle.

4.5. Conclusions

In summary, the risk of TH or FLM is driven by a large number of small-effect variants, potentially implicating most of the regulatory variants that are active in tissues relevant to these malformations. The observation that the heritability of these disqualifying traits is widely distributed across the genome seems to contradict the expectation that their causal variants are concentrated in specific biologically relevant pathways and specific genes. SNPs significantly associated with Testicular Hypoplasia and Feet and Legs Malformation are associated with QTL for reproductive traits, meat and carcass and exterior conformation. The genes identified as potentially associated are involved in reproductive aspects of bovines and short stature, respectively. Key roles of biological pathways and processes related to spermatogenesis, macrophage-mediated inflammation and control of tumor cell development were detected.

4.6. References

Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B 57, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Boyle, E.A., Li, Y.I., Pritchard, J.K., 2017. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. https://doi.org/10.1016/j.cell.2017.05.038

Chakravarti, A., Turner, T.N., 2016. Revealing rate-limiting steps in complex disease biology: The crucial importance of studying rare, extreme-phenotype families. BioEssays 38, 578–586. https://doi.org/10.1002/bies.201500203

Chang, C.C., Chow, C.C., Tellier, L.C.A.M., Vattikuti, S., Purcell, S.M., Lee, J.J., 2015. Second-generation PLINK: Rising to the challenge of larger and richer datasets. Gigascience 4, 7. https://doi.org/10.1186/s13742-015-0047-8

Chen, J., Bardes, E.E., Aronow, B.J., Jegga, A.G., 2009. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. 37, 305–311. https://doi.org/10.1093/nar/gkp427

Dang, Y., Li, S., Zhao, P., Xiao, L., Wang, L., Shi, Y., Luo, L., Wang, S., Wang, H., Zhang, K., 2022. The lysine deacetylase activity of histone deacetylases 1 and 2 is required to safeguard zygotic genome activation in mice and cattle. Development 149. https://doi.org/10.1242/dev.200854

Fonseca, P.A. de S., dos Santos, F.C., Lam, S., Suárez-Vega, A., Miglior, F., Schenkel, F.S., Diniz, L. de A.F., Id-Lahoucine, S., Carvalho, M.R.S., Cánovas, A., 2018. Genetic mechanisms underlying spermatic and testicular traits within and among cattle breeds: systematic review and prioritization of GWAS results. J. Anim. Sci. 96, 4978–4999. https://doi.org/10.1093/jas/sky382

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

Fritz, S., Capitan, A., Djari, A., Rodriguez, S.C., Barbat, A., Baur, A., Grohs, C., Weiss, B., Boussaha, M., Esquerré, D., Klopp, C., Rocha, D., Boichard, D., 2013. Detection of Haplotypes Associated with Prenatal Death in Dairy Cattle and Identification of Deleterious Mutations in GART, SHBG and SLC37A2. PLoS One 8, e65550. https://doi.org/10.1371/journal.pone.0065550

GenSys, C.A.S.L., 2019. Aliança Nelore: Sumário de avaliação de reprodutores [WWW Document]. URL https://gensys.com.br/sumario/sumario-alianca-nelore-2019-2020/ (accessed 10.18.19).

Giroto, A.B., Fontes, P.K., Franchi, F.F., dos Santos, P.H., Razza, E.M., Nogueira, M.F.G., Maioli, M.A., Nogueira, G.P., Nunes, G.B., Mingoti, G.Z., Mareco, E.A., Castilho, A.C.S., 2019. Use of pregnancy-associated plasma protein-A during oocyte in vitro maturation increases IGF-1 and affects the transcriptional profile of cumulus cells and embryos from Nelore cows. Mol. Reprod. Dev. 86, 1694–1704. https://doi.org/10.1002/mrd.23259

Gur, Y., Breitbart, H., Lax, Y., Rubinstein, S., Zamir, N., 1998. Angiotensin II induces acrosomal exocytosis in bovine spermatozoa. Am. J. Physiol. - Endocrinol. Metab. 275. https://doi.org/10.1152/ajpendo.1998.275.1.e87

Hu, Z.-L., Park, C.A., Reecy, J.M., 2022. Bringing the Animal QTLdb and CorrDB into the future: meeting new challenges and providing updated services. Nucleic Acids Res. 50, D956–D961. https://doi.org/10.1093/nar/gkab1116

Hu, Z.-L., Park, C.A., Wu, X.-L., Reecy, J.M., 2013. Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. Nucleic Acids Res. 41, D871–D879. https://doi.org/10.1093/nar/gks1150

Jääskeläinen, M., Nieminen, A., Pökkylä, R.M., Kauppinen, M., Liakka, A., Heikinheimo, M., Vaskivuo, T.E., Klefström, J., Tapanainen, J.S., 2010. Regulation of cell death in human fetal and adult ovaries-Role of Bok and Bcl-XL. Mol. Cell. Endocrinol. 330, 17–24. https://doi.org/10.1016/j.mce.2010.07.020

Kim, D., Langmead, B., Salzberg, S.L., 2015. HISAT: a fast spliced aligner with low memory requirements. Nat. Methods 2015 124 12, 357–360. https://doi.org/10.1038/nmeth.3317

Krumm, N., Turner, T.N., Baker, C., Vives, L., Mohajeri, K., Witherspoon, K., Raja, A., Coe, B.P., Stessman, H.A., He, Z.X., Leal, S.M., Bernier, R., Eichler, E.E., 2015. Excess of rare, inherited truncating mutations in autism. Nat. Genet. 47, 582–588. https://doi.org/10.1038/ng.3303

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/BIOINFORMATICS/BTP352

Marouli, E., Graff, M., Medina-Gomez, C. et al., 2017. Rare and low-frequency coding variants alter human adult height. Nat. 2017 5427640 542, 186–190. https://doi.org/10.1038/nature21039

Neves, H.H.R., Vargas, G., Brito, L.F., Schenkel, F.S., Albuquerque, L.G., Carvalheiro, R., 2019. Genetic and genomic analyses of testicular hypoplasia in Nellore cattle. PLoS One 14, e0211159. https://doi.org/10.1371/journal.pone.0211159

Park, J.W., Lee, S.J., 2021. Foveal hypoplasia in short stature with optic atrophy and Pelger-Huët anomaly syndrome with neuroblastoma-amplified sequence (NBAS) gene mutation, in: Journal of AAPOS. Mosby, pp. 257-259.e2. https://doi.org/10.1016/j.jaapos.2016.09.030

Pereira, G.L., Rosa, K.O., Curi, R.A., Reginato, L.C.A., Mota, M.D.S., 2013. Estado da arte do sequenciamento genômico na pecuária. Ars Veterinaria 29, 190–199. https://doi.org/10.15361/2175-0106.2013v29n3p190-199

Price, A.L., Helgason, A., Thorleifsson, G., McCarroll, S.A., Kong, A., Stefansson, K., 2011. Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. PLoS Genet. 7, e1001317. https://doi.org/10.1371/journal.pgen.1001317

Pru, J.K., Austin, K.J., Haas, A.L., Hansen, T.R., 2001. Pregnancy and interferon-τ upregulate gene expression of members of the 1-8 family in the bovine uterus. Biol. Reprod. 65, 1471–1480. https://doi.org/10.1095/biolreprod65.5.1471

R Development Core Team, 2017. R: A language and environment for statistical computing.

Roso, V.M., Schenkel, F.S., 2006. AMC-a computer program to assess the degree of connectedness among contemporary groups, in: Proceedings of the 8th World Congress on Genetics Applied to Livestock Production, Belo Horizonte, Minas Gerais, Brazil, 13-18 August, 2006. Instituto Prociência, pp. 26–27.

Sandor, C., Li, W., Coppieters, W., Druet, T., Charlier, C., Georges, M., 2012. Genetic variants in REC8, RNF212, and PRDM9 influence male recombination in cattle. PLoS Genet. 8. https://doi.org/10.1371/journal.pgen.1002854

Sargolzaei, M., Chesnais, J.P., Schenkel, F.S., 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15, 478. https://doi.org/10.1186/1471-2164-15-478

Shi, H., Kichaev, G., Pasaniuc, B., 2016. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. Am. J. Hum. Genet. 99, 139–153. https://doi.org/10.1016/j.ajhg.2016.05.013

Silva, T. de L., Gondro, C., Fonseca, P.A. de S., da Silva, D.A., Vargas, G., Neves, H.H. de R., Filho, I.C., Teixeira, C. de S., Albuquerque, L.G. de, Carvalheiro, R., 2023. Testicular hypoplasia in Nellore Cattle: Genetic analysis and functional analysis of genome-wide association study results. J. Anim. Breed. Genet. 140, 185–197. https://doi.org/10.1111/jbg.12747

Sonawane, A.R., Platig, J., Fagny, M., Chen, C.-Y., Paulson, J.N., Lopes-Ramos, C.M., DeMeo, D.L., Quackenbush, J., Glass, K., Kuijjer, M.L., 2017. Understanding Tissue-specific Gene Regulation. bioRxiv 110601. https://doi.org/10.1101/110601

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Iny Stein, T., Nudel, R., Lieder, I., Mazor, Y., Kaplan, S., Dahary, D., Warshawsky, D., Guan-Golan, Y., Kohn, A., Rappaport, N., Safran, M., Lancet, D., 2016. The GeneCards suite: From gene data mining to disease genome sequence analyses. Curr. Protoc. Bioinforma. 2016, 1.30.1-1.30.33. https://doi.org/10.1002/cpbi.5

Tao, D., Wang, Y., Zhang, X., Wang, C., Yang, D., Chen, J., Long, Y., Jiang, Y., Zhou, X., Zhang, N., 2022. Identification of Angiogenesis-Related Prognostic Biomarkers Associated With Immune Cell Infiltration in Breast Cancer. Front. Cell Dev. Biol. 10, 836. https://doi.org/10.3389/fcell.2022.853324

Topoluk, N., Steckbeck, K., Siatkowski, S., Burnikel, B., Tokish, J., Mercuri, J., 2018. Amniotic mesenchymal stem cells mitigate osteoarthritis progression in a synovial macrophage-mediated in vitro explant coculture model. J. Tissue Eng. Regen. Med. 12, 1097–1110. https://doi.org/10.1002/term.2610

Vargas, G., Neves, H.H.R., Camargo, G.M.F., Cardoso, V., Munari, D.P., Carvalheiro, R., 2018. Genome-wide association study and functional analysis of feet and leg conformation traits in Nellore cattle. J. Anim. Sci. 96, 1617–1627. https://doi.org/10.1093/jas/sky079

Vargas, G., Neves, H.H.R., Cardoso, V., Munari, D.P., Carvalheiro, R., 2017. Genetic analysis of feet and leg conformation traits in Nelore cattle. J. Anim. Sci. 95, 2379. https://doi.org/10.2527/jas2016.1327

Veerkamp, R.F., Bouwman, A.C., Schrooten, C., Calus, M.P.L., 2016. Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein-Friesian cattle. Genet. Sel. Evol. 48, 1–14. https://doi.org/10.1186/s12711-016-0274-1

Verardo, L.L., Silva, F.F., Lopes, M.S., Madsen, O., Bastiaansen, J.W.M., Knol, E.F., Kelly, M., Varona, L., Lopes, P.S., Guimarães, S.E.F., 2016. Revealing new candidate genes for reproductive traits in pigs: Combining Bayesian GWAS and functional pathways. Genetics Selection Evolution 48, 1–13. https://doi.org/10.1186/s12711-016-0189-x

Visscher, P.M., Medland, S.E., Ferreira, M.A.R., Morley, K.I., Zhu, G., Cornes, B.K., Montgomery, G.W., Martin, N.G., 2006. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS Genet. 2, 0316–0325. https://doi.org/10.1371/journal.pgen.0020041

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., Yu, G., 2021. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innov. 2, 100141. https://doi.org/10.1016/j.xinn.2021.100141

Yang, J., Weedon, M.N., Purcell, S., Lettre, G., Estrada, K., Willer, C.J., Smith, A. V., Ingelsson, E., O'Connell, J.R., Mangino, M., Mägi, R., Madden, P.A., Heath, A.C., Nyholt, D.R., Martin, N.G., Montgomery, G.W., Frayling, T.M., Hirschhorn, J.N., McCarthy, M.I., Goddard, M.E., Visscher, P.M., 2011. Genomic inflation factors under polygenic inheritance. Eur. J. Hum. Genet. 19, 807–812. https://doi.org/10.1038/ejhg.2011.39

Yu, G., Wang, L.G., Han, Y., He, Q.Y., 2012. ClusterProfiler: An R package for comparing biological themes among gene clusters. Omi. A J. Integr. Biol. 16, 284–287. https://doi.org/10.1089/omi.2011.0118

Zhang, J.X., Ma, H.W., Sang, M., Hu, Y.S., Liang, Z.N., Ai, H.X., Zhang, J., Cui, X.W., Zhang, S.Q., 2010. Molecular structure, expression, cell and tissue distribution, immune evolution and cell proliferation of the gene encoding bovine (Bos taurus) TNFSF13 (APRIL). Dev. Comp. Immunol. 34, 1199–1208. https://doi.org/10.1016/j.dci.2010.06.015

Zhou, Y., Shen, B., Jiang, J., Padhi, A., Park, K.E., Oswalt, A., Sattler, C.G., Telugu, B.P., Chen, H., Cole, J.B., Liu, G.E., Ma, L., 2018. Construction of PRDM9 allele-specific recombination maps in cattle using large-scale pedigree analysis and genome-wide single sperm genomics. DNA Res. 25, 183–194. https://doi.org/10.1093/dnares/dsx048

CAPÍTULO 5 - Considerações Finais

O GWAS usa dados genômicos de indivíduos para identificar associações entre variantes genéticas e características complexas, como TH e FLM. No entanto, essas associações não implicam causalidade direta entre uma variante e uma característica, mas sim sugerem a presença de genes candidatos que podem estar relacionados à característica em questão. Portanto, o GWAS faz uma inferência especulativa com base em evidências limitadas, sobre quais genes podem estar envolvidos na característica estudada. Essa inferência especulativa é necessária porque a relação entre genes e características complexas frequentemente envolve a interação de múltiplos genes e fatores ambientais.

A forma como cada uma das etapas envolvidas no GWAS é realizada pode afetar a capacidade do estudo de identificar corretamente as associações verdadeiras e excluir as falsas. Também existem fatores que explicam por que os resultados do GWAS com dados de HD e de sequência numa mesma população de bovinos podem ser diferentes, como: tamanho da amostra, qualidade dos dados, diferentes variantes genéticas e metodologias de análise.

O GWAS com dados de sequência pode exigir um maior número de amostras para ter poder estatístico suficiente para identificar variantes genéticas associadas a uma característica, uma vez que cobre todo o genoma. Assim, para resultados mais precisos, deve-se utilizar o maior número de animais sequenciados possível. Se a amostra utilizada no GWAS com dados de sequência for menor do que a usada no GWAS com dados de HD, isso pode afetar a capacidade do estudo de detectar associações significativas.

A qualidade dos dados é outro fator que pode afetar os resultados do GWAS. Dados de sequência têm maior cobertura do que dados de HD, o que pode aumentar a precisão na identificação de variantes genéticas associadas à característica de interesse. Além disso, os dados de HD podem ter mais ruído e erros de genotipagem. O uso de diferentes plataformas de genotipagem e métodos de imputação também podem levar a resultados diferentes. Economicamente a técnica de imputação é vantajosa para a seleção genômica, mas aspectos como o erro de imputação e a existência de halótipos que não estão presentes na população de referência utilizada na imputação devem ser considerados.

Enquanto o GWAS com dados de sequência cobre todo o genoma permitindo a identificação de variantes raras e desconhecidas, o GWAS com dados de HD se concentra em um conjunto limitado de variantes genéticas pré-selecionadas. Portanto, é possível que variantes genéticas não capturadas pelo GWAS com dados de HD estejam contribuindo para a variação da característica estudada.

Finalmente diferenças nas metodologias de análise podem levar a resultados diferentes em termos de regiões significativas e genes candidatos. O método WssGBLUP, aplicado na análise dos dados de HD usa informações sobre a estrutura de covariância entre os indivíduos para identificar regiões do genoma associadas a uma característica, enquanto o método SNP-BLUP, adotado no GWAS com dados de sequência, usa informações sobre as frequências alélicas dos SNPs para estimar os efeitos genéticos. É essencial a utilização de metodologias rigorosas e padronizadas no GWAS para minimizar o impacto dessas diferenças e aumentar a confiabilidade e a validade dos resultados.

Em resumo, o GWAS com dados de HD tende a ser mais barato, embora informações importantes presentes em variantes raras ou desconhecidas que não são cobertas pelos SNPs selecionados possam ser perdidas. Por outro lado, o GWAS com dados de sequência é mais completo e pode fornecer informações mais detalhadas sobre a relação entre variantes genéticas e características complexas, mas o custo de genotipagem é maior. Apesar da inferência especulativa do GWAS sobre genes candidatos não ser uma prova definitiva de causalidade, este é um primeiro passo importante na identificação de genes e vias biológicas que podem estar envolvidos em doenças complexas. Essa informação pode ser usada para identificar indivíduos com maior risco de desenvolver características desclassificantes, permitindo uma intervenção precoce e personalizada. A intensificação da seleção genômica pode aumentar a herdabilidade de características desclassificantes como TH e FLM. Os resultados obtidos neste trabalho podem ajudar em estudos futuros, como por exemplo numa meta-análise. O próximo passo seria sequenciar animais acometidos por TH e FLM e utilizar estes dados de sequência para fazer um estudo de validação.

APÊNDICE

Apêndice A. Results of the TH enrichment analysis from the genes prioritized in the functional annotation study.

ONTOLOGY	ID	Description	p-value	p-adjust	q-value	Gene ID
BP	GO:0050810	regulation of steroid biosynthetic process	7.99E-05	0.009740988	0.002023751	STAR/ERLIN2
BP	GO:0019218	regulation of steroid metabolic process	0.000168294	0.009740988	0.002023751	STAR/ERLIN2
BP	GO:0010565	regulation of cellular ketone metabolic process	0.000297738	0.009740988	0.002023751	STAR/ERLIN2
BP	GO:0008203	cholesterol metabolic process	0.000315896	0.009740988	0.002023751	STAR/ERLIN2
BP	GO:1902652	secondary alcohol metabolic process	0.000363617	0.009740988	0.002023751	STAR/ERLIN2
BP	GO:0016125	sterol metabolic process	0.000388722	0.009740988	0.002023751	STAR/ERLIN2
BP	GO:0017143	insecticide metabolic process	0.000427271	0.009740988	0.002023751	STAR
BP	GO:0018963	phthalate metabolic process	0.000427271	0.009740988	0.002023751	STAR
BP	GO:0046890	regulation of lipid biosynthetic process	0.00049167	0.009740988	0.002023751	STAR/ERLIN2
BP	GO:0006694	steroid biosynthetic process	0.0005032	0.009740988	0.002023751	STAR/ERLIN2
BP	GO:0018879	biphenyl metabolic process	0.000640854	0.009740988	0.002023751	STAR
BP	GO:0018894	dibenzo-p-dioxin metabolic process	0.000640854	0.009740988	0.002023751	STAR
BP	GO:0070859	positive regulation of bile acid biosynthetic process	0.000640854	0.009740988	0.002023751	STAR
BP	GO:0071373	cellular response to luteinizing hormone stimulus	0.000640854	0.009740988	0.002023751	STAR
BP	GO:1904253	positive regulation of bile acid metabolic process	0.000640854	0.009740988	0.002023751	STAR
BP	GO:0042180	cellular ketone metabolic process	0.000747284	0.010307888	0.002141528	STAR/ERLIN2
BP	GO:0072330	monocarboxylic acid biosynthetic process	0.000768571	0.010307888	0.002141528	STAR/ERLIN2
BP	GO:0034699	response to luteinizing hormone	0.000854404	0.010822452	0.002248432	STAR
BP	GO:1901617	organic hydroxy compound biosynthetic process	0.00094153	0.011298363	0.002347305	STAR/ERLIN2
BP	GO:0009636	response to toxic substance	0.001149043	0.012982781	0.002697254	GPX5/STAR
BP	GO:0042747	circadian sleep/wake cycle, REM sleep	0.001281401	0.012982781	0.002697254	STAR
BP	GO:0009635	response to herbicide	0.001494848	0.012982781	0.002697254	STAR
BP	GO:0060992	response to fungicide	0.001494848	0.012982781	0.002697254	STAR
BP	GO:0061370	testosterone biosynthetic process	0.001494848	0.012982781	0.002697254	STAR
BP	GO:0046394	carboxylic acid biosynthetic process	0.001645302	0.012982781	0.002697254	STAR/ERLIN2
BP	GO:0016053	organic acid biosynthetic process	0.001666122	0.012982781	0.002697254	STAR/ERLIN2
BP	GO:0008202	steroid metabolic process	0.001697592	0.012982781	0.002697254	STAR/ERLIN2

BP	GO:0045541	negative regulation of cholesterol biosynthetic process	0.001708261	0.012982781	0.002697254	ERLIN2
BP	GO:0090206	negative regulation of cholesterol metabolic process	0.001708261	0.012982781	0.002697254	ERLIN2
BP	GO:0106119	negative regulation of sterol biosynthetic process	0.001708261	0.012982781	0.002697254	ERLIN2
BP	GO:0019216	regulation of lipid metabolic process	0.001826343	0.013247144	0.002752177	STAR/ERLIN2
BP	GO:0062012	regulation of small molecule metabolic process	0.001859248	0.013247144	0.002752177	STAR/ERLIN2
BP	GO:0006066	alcohol metabolic process	0.002074293	0.013907893	0.002889452	STAR/ERLIN2
BP	GO:0017085	response to insecticide	0.002134984	0.013907893	0.002889452	STAR
BP	GO:0071372	cellular response to follicle-stimulating hormone stimulus	0.002134984	0.013907893	0.002889452	STAR
BP	GO:0035457	cellular response to interferon-alpha	0.002348294	0.014872527	0.00308986	STAR
BP	GO:0006703	estrogen biosynthetic process	0.00256157	0.014975331	0.003111218	STAR
BP	GO:0070857	regulation of bile acid biosynthetic process	0.00256157	0.014975331	0.003111218	STAR
BP	GO:0071872	cellular response to epinephrine stimulus	0.00256157	0.014975331	0.003111218	STAR
BP	GO:0071236	cellular response to antibiotic	0.002988019	0.016563141	0.003441096	STAR
BP	GO:0032354	response to follicle-stimulating hormone	0.003201193	0.016563141	0.003441096	STAR
BP	GO:0032933	SREBP signaling pathway	0.003201193	0.016563141	0.003441096	ERLIN2
BP	GO:0071871	response to epinephrine	0.003201193	0.016563141	0.003441096	STAR
BP	GO:1904251	regulation of bile acid metabolic process	0.003201193	0.016563141	0.003441096	STAR
BP	GO:0006979	response to oxidative stress	0.003291052	0.016563141	0.003441096	GPX5/STAR
BP	GO:0071501	cellular response to sterol depletion	0.003414332	0.016563141	0.003441096	ERLIN2
BP	GO:1902931	negative regulation of alcohol biosynthetic process	0.003414332	0.016563141	0.003441096	ERLIN2
BP	GO:0071371	cellular response to gonadotropin stimulus	0.003627437	0.017230325	0.003579707	STAR
BP	GO:0006991	response to sterol depletion	0.003840508	0.017512715	0.003638376	ERLIN2
BP	GO:0009404	toxin metabolic process	0.003840508	0.017512715	0.003638376	STAR
BP	GO:0010893	positive regulation of steroid biosynthetic process	0.004266547	0.01844618	0.003832309	STAR
BP	GO:0035455	response to interferon-alpha	0.004479515	0.01844618	0.003832309	STAR
BP	GO:0045717	negative regulation of fatty acid biosynthetic process	0.004479515	0.01844618	0.003832309	ERLIN2
BP	GO:0050802	circadian sleep/wake cycle, sleep	0.004479515	0.01844618	0.003832309	STAR
BP	GO:0008211	glucocorticoid metabolic process	0.004692449	0.01844618	0.003832309	STAR
BP	GO:0045540	regulation of cholesterol biosynthetic process	0.004692449	0.01844618	0.003832309	ERLIN2
BP	GO:0051412	response to corticosterone	0.004692449	0.01844618	0.003832309	STAR
BP	GO:0106118	regulation of sterol biosynthetic process	0.004692449	0.01844618	0.003832309	ERLIN2

BP	GO:0010288	response to lead ion	0.005118215	0.018523065	0.003848282	STAR
BP	GO:0022410	circadian sleep/wake cycle process	0.005118215	0.018523065	0.003848282	STAR
BP	GO:0044321	response to leptin	0.005118215	0.018523065	0.003848282	STAR
BP	GO:0071378	cellular response to growth hormone stimulus	0.005118215	0.018523065	0.003848282	STAR
BP	GO:0120255	olefinic compound biosynthetic process	0.005118215	0.018523065	0.003848282	STAR
BP	GO:0042537	benzene-containing compound metabolic process	0.005543844	0.019589651	0.004069872	STAR
BP	GO:0010894	negative regulation of steroid biosynthetic process	0.005756608	0.019589651	0.004069872	ERLIN2
BP	GO:0034698	response to gonadotropin	0.005756608	0.019589651	0.004069872	STAR
BP	GO:0042745	circadian sleep/wake cycle	0.005756608	0.019589651	0.004069872	STAR
BP	GO:0032366	intracellular sterol transport	0.006182032	0.020135762	0.00418333	STAR
BP	GO:0032367	intracellular cholesterol transport	0.006182032	0.020135762	0.00418333	STAR
BP	GO:0071549	cellular response to dexamethasone stimulus	0.006182032	0.020135762	0.00418333	STAR
BP	GO:0045939	negative regulation of steroid metabolic process	0.006394693	0.020249862	0.004207035	ERLIN2
BP	GO:0045940	positive regulation of steroid metabolic process	0.006394693	0.020249862	0.004207035	STAR
BP	GO:0030431	sleep	0.006819913	0.02130055	0.004425322	STAR
BP	GO:0045922	negative regulation of fatty acid metabolic process	0.007457486	0.022670758	0.004709991	ERLIN2
BP	GO:0090181	regulation of cholesterol metabolic process	0.007457486	0.022670758	0.004709991	ERLIN2
BP	GO:0006699	bile acid biosynthetic process	0.007669943	0.022710999	0.004718352	STAR
BP	GO:0051385	response to mineralocorticoid	0.007669943	0.022710999	0.004718352	STAR
BP	GO:0060416	response to growth hormone	0.007882365	0.023040758	0.004786861	STAR
BP	GO:0008210	estrogen metabolic process	0.008094753	0.023362071	0.004853616	STAR
BP	GO:0071548	response to dexamethasone	0.008307107	0.023675254	0.004918682	STAR
BP	GO:0071276	cellular response to cadmium ion	0.008519427	0.023980608	0.004982121	STAR
BP	GO:0071312	cellular response to alkaloid	0.008943964	0.02456011	0.005102516	STAR
BP	GO:0032365	intracellular lipid transport	0.009156181	0.02456011	0.005102516	STAR
BP	GO:0042181	ketone biosynthetic process	0.009156181	0.02456011	0.005102516	STAR
BP	GO:0120178	steroid hormone biosynthetic process	0.009156181	0.02456011	0.005102516	STAR
BP	GO:0006984	ER-nucleus signaling pathway	0.00979263	0.025295317	0.00525526	ERLIN2
BP	GO:0035094	response to nicotine	0.00979263	0.025295317	0.00525526	STAR
BP	GO:0046677	response to antibiotic	0.010004711	0.025295317	0.00525526	STAR
BP	GO:0048512	circadian behavior	0.010004711	0.025295317	0.00525526	STAR

BP	GO:0120009	intermembrane lipid transfer	0.010216758	0.025295317	0.00525526	STAR
BP	GO:0007622	rhythmic behavior	0.010428771	0.025295317	0.00525526	STAR
BP	GO:0008206	bile acid metabolic process	0.010428771	0.025295317	0.00525526	STAR
BP	GO:0042304	regulation of fatty acid biosynthetic process	0.010428771	0.025295317	0.00525526	ERLIN2
BP	GO:0048168	regulation of neuronal synaptic plasticity	0.010428771	0.025295317	0.00525526	STAR
BP	GO:1902930	regulation of alcohol biosynthetic process	0.01064075	0.025537801	0.005305637	ERLIN2
BP	GO:0071320	cellular response to cAMP	0.011276483	0.026781647	0.005564054	STAR
BP	GO:0071385	cellular response to glucocorticoid stimulus	0.01191191	0.027921134	0.00580079	STAR
BP	GO:0006695	cholesterol biosynthetic process	0.01212365	0.027921134	0.00580079	ERLIN2
BP	GO:1902653	secondary alcohol biosynthetic process	0.01212365	0.027921134	0.00580079	ERLIN2
BP	GO:0051055	negative regulation of lipid biosynthetic process	0.01254703	0.028607228	0.00594333	ERLIN2
BP	GO:0000381	regulation of alternative mRNA splicing, via spliceosome	0.012758668	0.028710895	0.005964867	KHDRBS3
BP	GO:0019748	secondary metabolic process	0.012970273	0.028710895	0.005964867	STAR
BP	GO:0071384	cellular response to corticosteroid stimulus	0.012970273	0.028710895	0.005964867	STAR
BP	GO:0016126	sterol biosynthetic process	0.013604883	0.029826089	0.006196556	ERLIN2
BP	GO:0042446	hormone biosynthetic process	0.014027786	0.030460335	0.006328324	STAR
BP	GO:0046686	response to cadmium ion	0.014450553	0.031082321	0.006457546	STAR
BP	GO:0043627	response to estrogen	0.015929167	0.033942525	0.007051771	STAR
BP	GO:0014823	response to activity	0.016140262	0.034073886	0.007079062	STAR
BP	GO:0000380	alternative mRNA splicing, via spliceosome	0.016351323	0.034202767	0.007105838	KHDRBS3
BP	GO:0046889	positive regulation of lipid biosynthetic process	0.017406117	0.036078134	0.007495457	STAR
BP	GO:0030433	ubiquitin-dependent ERAD pathway	0.018038587	0.037052232	0.007697832	ERLIN2
BP	GO:0016101	diterpenoid metabolic process	0.018460063	0.037579414	0.007807358	STAR
BP	GO:0019217	regulation of fatty acid metabolic process	0.019723679	0.039447358	0.008195434	ERLIN2
BP	GO:0051591	response to cAMP	0.019723679	0.039447358	0.008195434	STAR
BP	GO:1901655	cellular response to ketone	0.020355029	0.040356057	0.008384222	STAR
BP	GO:0006721	terpenoid metabolic process	0.020565411	0.04042167	0.008397854	STAR
BP	GO:0048024	regulation of mRNA splicing, via spliceosome	0.021406601	0.041071947	0.008532953	KHDRBS3
BP	GO:0098869	cellular oxidant detoxification	0.021406601	0.041071947	0.008532953	GPX5
BP	GO:0045833	negative regulation of lipid metabolic process	0.021616814	0.041071947	0.008532953	ERLIN2
BP	GO:0062014	negative regulation of small molecule metabolic process	0.021616814	0.041071947	0.008532953	ERLIN2

BP	GO:0036503	ERAD pathway	0.022667371	0.042017565	0.008729411	ERLIN2
BP	GO:0071868	cellular response to monoamine stimulus	0.022667371	0.042017565	0.008729411	STAR
BP	GO:0071870	cellular response to catecholamine stimulus	0.022667371	0.042017565	0.008729411	STAR
BP	GO:0018958	phenol-containing compound metabolic process	0.023297298	0.042201914	0.008767711	STAR
BP	GO:0006805	xenobiotic metabolic process	0.023507206	0.042201914	0.008767711	STAR
BP	GO:0071867	response to monoamine	0.023507206	0.042201914	0.008767711	STAR
BP	GO:0071869	response to catecholamine	0.023507206	0.042201914	0.008767711	STAR
BP	GO:0044344	cellular response to fibroblast growth factor stimulus	0.023926921	0.042619828	0.008854535	STAR
BP	GO:0043279	response to alkaloid	0.024346501	0.042739104	0.008879315	STAR
BP	GO:0006720	isoprenoid metabolic process	0.02455624	0.042739104	0.008879315	STAR
BP	GO:1990748	cellular detoxification	0.02455624	0.042739104	0.008879315	GPX5
BP	GO:0030301	cholesterol transport	0.024765945	0.042777541	0.008887301	STAR
BP	GO:0071346	cellular response to interferon-gamma	0.024975616	0.042815342	0.008895154	STAR
BP	GO:0071774	response to fibroblast growth factor	0.025394858	0.043209161	0.008976972	STAR
BP	GO:0097237	cellular response to toxic substance	0.026232935	0.044304512	0.009204538	GPX5
BP	GO:0015918	sterol transport	0.027489037	0.046084562	0.009574355	STAR
BP	GO:0046683	response to organophosphorus	0.027698269	0.046096389	0.009576812	STAR
BP	GO:0034754	cellular hormone metabolic process	0.028952953	0.047151952	0.009796112	STAR
BP	GO:0045471	response to ethanol	0.028952953	0.047151952	0.009796112	STAR
BP	GO:0050684	regulation of mRNA processing	0.028952953	0.047151952	0.009796112	KHDRBS3
BP	GO:0046165	alcohol biosynthetic process	0.029579839	0.047185509	0.009803084	ERLIN2
BP	GO:0008584	male gonad development	0.029788734	0.047185509	0.009803084	STAR
BP	GO:0034341	response to interferon-gamma	0.029788734	0.047185509	0.009803084	STAR
BP	GO:0046546	development of primary male sexual characteristics	0.029997595	0.047185509	0.009803084	STAR
BP	GO:0062013	positive regulation of small molecule metabolic process	0.030206422	0.047185509	0.009803084	STAR
BP	GO:0043524	negative regulation of neuron apoptotic process	0.030623975	0.047185509	0.009803084	STAR
BP	GO:0042542	response to hydrogen peroxide	0.030832701	0.047185509	0.009803084	STAR
BP	GO:0010212	response to ionizing radiation	0.031250052	0.047185509	0.009803084	STAR
BP	GO:0014074	response to purine-containing compound	0.031250052	0.047185509	0.009803084	STAR
BP	GO:0043484	regulation of RNA splicing	0.031250052	0.047185509	0.009803084	KHDRBS3
BP	GO:0051384	response to glucocorticoid	0.031250052	0.047185509	0.009803084	STAR

BP	GO:0045834	positive regulation of lipid metabolic process	0.031458677	0.047188016	0.009803604	STAR
BP	GO:0071333	cellular response to glucose stimulus	0.031875826	0.047197228	0.009805518	STAR
BP	GO:0098754	detoxification	0.03208435	0.047197228	0.009805518	GPX5
BP	GO:0071331	cellular response to hexose stimulus	0.03229284	0.047197228	0.009805518	STAR
BP	GO:0120254	olefinic compound metabolic process	0.03229284	0.047197228	0.009805518	STAR
BP	GO:0071326	cellular response to monosaccharide stimulus	0.032501297	0.047199335	0.009805956	STAR
BP	GO:0006633	fatty acid biosynthetic process	0.03437589	0.049293729	0.010241079	ERLIN2
BP	GO:0071322	cellular response to carbohydrate stimulus	0.03437589	0.049293729	0.010241079	STAR
BP	GO:0046661	male sex differentiation	0.034792096	0.049578737	0.010300292	STAR
BP	GO:0031960	response to corticosteroid	0.035208167	0.049860014	0.010358729	STAR
CC	GO:0030061	mitochondrial crista	0.001840396	0.011042377	0.005811778	STAR
CC	GO:0005758	mitochondrial intermembrane space	0.017482363	0.03900043	0.020526542	STAR
CC	GO:0031970	organelle envelope lumen	0.019500215	0.03900043	0.020526542	STAR
MF	GO:0015485	cholesterol binding	4.34E-05	0.000501604	6.60E-05	STAR/ERLIN2
MF	GO:0032934	sterol binding	6.27E-05	0.000501604	6.60E-05	STAR/ERLIN2
MF	GO:0043178	alcohol binding	0.000138428	0.000699357	9.20E-05	STAR/ERLIN2
MF	GO:0005496	steroid binding	0.000174839	0.000699357	9.20E-05	STAR/ERLIN2
MF	GO:0004602	glutathione peroxidase activity	0.00478299	0.011428549	0.001503756	GPX5
MF	GO:0120020	cholesterol transfer activity	0.00478299	0.011428549	0.001503756	STAR
MF	GO:0120015	sterol transfer activity	0.00499999	0.011428549	0.001503756	STAR
MF	GO:0015248	sterol transporter activity	0.007817764	0.015635529	0.002057306	STAR
MF	GO:0120013	lipid transfer activity	0.010413471	0.01734586	0.00228235	STAR
MF	GO:0004601	peroxidase activity	0.011061601	0.01734586	0.00228235	GPX5
MF	GO:0016684	oxidoreductase activity, acting on peroxide as acceptor	0.011925279	0.01734586	0.00228235	GPX5
MF	GO:0016209	antioxidant activity	0.018599617	0.024799489	0.003263091	GPX5
MF	GO:0017124	SH3 domain binding	0.028014744	0.034479685	0.004536801	KHDRBS3
MF	GO:0005319	lipid transporter activity	0.033121626	0.037853287	0.004980696	STAR

Apêndice B. Results of the TH enrichment analysis from the genes prioritized in the functional annotation study with whole-genome sequence data.

ONTOLOGY	ID	Description	p-value	p-adjust	q-value	Gene ID
BP	GO:0016575	histone deacetylation	2.85E-05	0.018053589	0.00923394	HDAC1/SIRT3/CTBP1
BP	GO:0006476	protein deacetylation	5.31E-05	0.018053589	0.00923394	HDAC1/SIRT3/CTBP1
BP	GO:0035601	protein deacylation	7.23E-05	0.018053589	0.00923394	HDAC1/SIRT3/CTBP1
BP	GO:0098732	macromolecule deacylation	8.03E-05	0.018053589	0.00923394	HDAC1/SIRT3/CTBP1
BP	GO:2000757	negative regulation of peptidyl-lysine acetylation	9.79E-05	0.018053589	0.00923394	HDAC1/CTBP1
BP	GO:0018205	peptidyl-lysine modification	0.000136635	0.018700304	0.009564718	PRDM9/HDAC1/SIRT3/CTBP1
BP	GO:1901984	negative regulation of protein acetylation	0.000141976	0.018700304	0.009564718	HDAC1/CTBP1
BP	GO:0016570	histone modification	0.000303469	0.034974795	0.017888695	PRDM9/HDAC1/SIRT3/CTBP1
BP	GO:0007565	female pregnancy	0.000361066	0.036989182	0.018919001	AGT/PAPPA/KPNA6
BP	GO:0044706	multi-multicellular organism process	0.000529462	0.038882125	0.019887192	AGT/PAPPA/KPNA6
BP	GO:0003345	proepicardium cell migration involved in pericardium morphogenesis	0.000747783	0.038882125	0.019887192	FLRT3
BP	GO:1902551	regulation of catalase activity	0.000747783	0.038882125	0.019887192	SIRT3
BP	GO:1902553	positive regulation of catalase activity	0.000747783	0.038882125	0.019887192	SIRT3
BP	GO:1905008	regulation of L-lysine import across plasma membrane	0.000747783	0.038882125	0.019887192	AGT
BP	GO:1905010	positive regulation of L-lysine import across plasma membrane	0.000747783	0.038882125	0.019887192	AGT
BP	GO:1905541	regulation of L-arginine import across plasma membrane	0.000747783	0.038882125	0.019887192	AGT
BP	GO:1905589	positive regulation of L-arginine import across plasma membrane	0.000747783	0.038882125	0.019887192	AGT
BP	GO:0046883	regulation of hormone secretion	0.000759087	0.038882125	0.019887192	AGT/HDAC1/SIRT3
BP	GO:0051353	positive regulation of oxidoreductase activity	0.000867087	0.039383821	0.020143797	AGT/SIRT3
BP	GO:2000756	regulation of peptidyl-lysine acetylation	0.000988037	0.039383821	0.020143797	HDAC1/CTBP1
BP	GO:0046879	hormone secretion	0.001239208	0.039383821	0.020143797	AGT/HDAC1/SIRT3
BP	GO:0042063	gliogenesis	0.001313147	0.039383821	0.020143797	MBOAT7/BOK/HDAC1
BP	GO:0009914	hormone transport	0.001376855	0.039383821	0.020143797	AGT/HDAC1/SIRT3

BP	GO:1901983	regulation of protein acetylation	0.001471461	0.039383821	0.020143797	HDAC1/CTBP1
BP	GO:0003051	angiotensin-mediated drinking behavior	0.001495047	0.039383821	0.020143797	AGT
BP	GO:0007199	G protein-coupled receptor signaling pathway coupled to cGMP nucleotide second messenger	0.001495047	0.039383821	0.020143797	AGT
BP	GO:0010963	regulation of L-arginine import	0.001495047	0.039383821	0.020143797	AGT
BP	GO:0035106	operant conditioning	0.001495047	0.039383821	0.020143797	AGT
BP	GO:0075506	entry of viral genome into host nucleus through nuclear pore complex via importin	0.001495047	0.039383821	0.020143797	KPNA6
BP	GO:0097639	L-lysine import across plasma membrane	0.001495047	0.039383821	0.020143797	AGT
BP	GO:1904700	granulosa cell apoptotic process	0.001495047	0.039383821	0.020143797	BOK
BP	GO:1904708	regulation of granulosa cell apoptotic process	0.001495047	0.039383821	0.020143797	BOK
BP	GO:1905435	regulation of histone H3-K4 trimethylation	0.001495047	0.039383821	0.020143797	PRDM9
BP	GO:1905437	positive regulation of histone H3-K4 trimethylation	0.001495047	0.039383821	0.020143797	PRDM9
BP	GO:2000470	positive regulation of peroxidase activity	0.001495047	0.039383821	0.020143797	SIRT3
BP	GO:0019216	regulation of lipid metabolic process	0.00172475	0.043729423	0.022366459	AGT/MBOAT7/SIRT3
BP	GO:0031058	positive regulation of histone modification	0.002091666	0.043729423	0.022366459	PRDM9/CTBP1
BP	GO:2001233	regulation of apoptotic signaling pathway	0.002123556	0.043729423	0.022366459	AGT/BOK/HDAC1
BP	GO:0048709	oligodendrocyte differentiation	0.002228222	0.043729423	0.022366459	BOK/HDAC1
BP	GO:0002019	regulation of renal output by angiotensin	0.002241793	0.043729423	0.022366459	AGT
BP	GO:0002035	brain renin-angiotensin system	0.002241793	0.043729423	0.022366459	AGT
BP	GO:0006311	meiotic gene conversion	0.002241793	0.043729423	0.022366459	PRDM9
BP	GO:0061198	fungiform papilla formation	0.002241793	0.043729423	0.022366459	HDAC1
BP	GO:2000468	regulation of peroxidase activity	0.002241793	0.043729423	0.022366459	SIRT3
BP	GO:2001255	positive regulation of histone H3-K36 trimethylation	0.002241793	0.043729423	0.022366459	PRDM9
BP	GO:1904035	regulation of epithelial cell apoptotic process	0.002274659	0.043729423	0.022366459	BOK/HDAC1
BP	GO:1901890	positive regulation of cell junction assembly	0.002662619	0.043729423	0.022366459	AGT/FLRT3
BP	GO:0051341	regulation of oxidoreductase activity	0.002815607	0.043729423	0.022366459	AGT/SIRT3
BP	GO:0000416	positive regulation of histone H3-K36 methylation	0.002988019	0.043729423	0.022366459	PRDM9

BP	GO:0002018	renin-angiotensin regulation of aldosterone production	0.002988019	0.043729423	0.022366459	AGT
BP	GO:0002034	maintenance of blood vessel diameter homeostasis by renin-angiotensin	0.002988019	0.043729423	0.022366459	AGT
BP	GO:0003072	renal control of peripheral vascular resistance involved in regulation of systemic arterial blood pressure	0.002988019	0.043729423	0.022366459	AGT
BP	GO:0010520	regulation of reciprocal meiotic recombination	0.002988019	0.043729423	0.022366459	PRDM9
BP	GO:0010845	positive regulation of reciprocal meiotic recombination	0.002988019	0.043729423	0.022366459	PRDM9
BP	GO:0014873	response to muscle activity involved in regulation of muscle adaptation	0.002988019	0.043729423	0.022366459	AGT
BP	GO:0035822	gene conversion	0.002988019	0.043729423	0.022366459	PRDM9
BP	GO:0051387	negative regulation of neurotrophin TRK receptor signaling pathway	0.002988019	0.043729423	0.022366459	AGT
BP	GO:0075732	viral penetration into host nucleus	0.002988019	0.043729423	0.022366459	KPNA6
BP	GO:0097676	histone H3-K36 dimethylation	0.002988019	0.043729423	0.022366459	PRDM9
BP	GO:1901671	positive regulation of superoxide dismutase activity	0.002988019	0.043729423	0.022366459	SIRT3
BP	GO:1903899	positive regulation of PERK-mediated unfolded protein response	0.002988019	0.043729423	0.022366459	BOK
BP	GO:2000676	positive regulation of type B pancreatic cell apoptotic process	0.002988019	0.043729423	0.022366459	HDAC1
BP	GO:2001253	regulation of histone H3-K36 trimethylation	0.002988019	0.043729423	0.022366459	PRDM9
BP	GO:0016032	viral process	0.003281196	0.045379805	0.023210586	HDAC1/KPNA6/CTBP1
BP	GO:1904019	epithelial cell apoptotic process	0.00358306	0.045379805	0.023210586	BOK/HDAC1
BP	GO:0052548	regulation of endopeptidase activity	0.003674446	0.045379805	0.023210586	AGT/BOK/HDAC1
BP	GO:0003331	positive regulation of extracellular matrix constituent secretion	0.003733727	0.045379805	0.023210586	AGT
BP	GO:0007023	post-chaperonin tubulin folding pathway	0.003733727	0.045379805	0.023210586	TBCE
BP	GO:0036149	phosphatidylinositol acyl-chain remodeling	0.003733727	0.045379805	0.023210586	MBOAT7
BP	GO:0060789	hair follicle placode formation	0.003733727	0.045379805	0.023210586	HDAC1
BP	GO:0061197	fungiform papilla morphogenesis	0.003733727	0.045379805	0.023210586	HDAC1
BP	GO:0097360	chorionic trophoblast cell proliferation	0.003733727	0.045379805	0.023210586	BOK

BP	GO:1901382	regulation of chorionic trophoblast cell proliferation	0.003733727	0.045379805	0.023210586	BOK
BP	GO:1905516	positive regulation of fertilization	0.003733727	0.045379805	0.023210586	PRDM9
BP	GO:2001235	positive regulation of apoptotic signaling pathway	0.003878301	0.045379805	0.023210586	AGT/BOK
BP	GO:0019079	viral genome replication	0.004184553	0.045379805	0.023210586	KPNA6/CTBP1
BP	GO:0052547	regulation of peptidase activity	0.004410137	0.045379805	0.023210586	AGT/BOK/HDAC1
BP	GO:0023061	signal release	0.004463952	0.045379805	0.023210586	AGT/HDAC1/SIRT3
BP	GO:0035931	mineralocorticoid secretion	0.004478918	0.045379805	0.023210586	AGT
BP	GO:0035932	aldosterone secretion	0.004478918	0.045379805	0.023210586	AGT
BP	GO:0051902	negative regulation of mitochondrial depolarization	0.004478918	0.045379805	0.023210586	BOK
BP	GO:0061196	fungiform papilla development	0.004478918	0.045379805	0.023210586	HDAC1
BP	GO:0090154	positive regulation of sphingolipid biosynthetic process	0.004478918	0.045379805	0.023210586	SIRT3
BP	GO:0090241	negative regulation of histone H4 acetylation	0.004478918	0.045379805	0.023210586	CTBP1
BP	GO:0097198	histone H3-K36 trimethylation	0.004478918	0.045379805	0.023210586	PRDM9
BP	GO:0097638	L-arginine import across plasma membrane	0.004478918	0.045379805	0.023210586	AGT
BP	GO:1903598	positive regulation of gap junction assembly	0.004478918	0.045379805	0.023210586	AGT
BP	GO:1990918	double-strand break repair involved in meiotic recombination	0.004478918	0.045379805	0.023210586	PRDM9
BP	GO:2000304	positive regulation of ceramide biosynthetic process	0.004478918	0.045379805	0.023210586	SIRT3
BP	GO:2000855	regulation of mineralocorticoid secretion	0.004478918	0.045379805	0.023210586	AGT
BP	GO:2000858	regulation of aldosterone secretion	0.004478918	0.045379805	0.023210586	AGT
BP	GO:0001819	positive regulation of cytokine production	0.004572791	0.045827317	0.023439477	AGT/HDAC1/KPNA6
BP	GO:0043524	negative regulation of neuron apoptotic process	0.005100009	0.047684658	0.024389459	BOK/HDAC1
BP	GO:0000414	regulation of histone H3-K36 methylation	0.005223591	0.047684658	0.024389459	PRDM9
BP	GO:0008635	activation of cysteine-type endopeptidase activity involved in apoptotic process by cytochrome c	0.005223591	0.047684658	0.024389459	BOK
BP	GO:0010536	positive regulation of activation of Janus kinase activity	0.005223591	0.047684658	0.024389459	AGT

BP	GO:0033864	positive regulation of NAD(P)H oxidase activity	0.005223591	0.047684658	0.024389459	AGT
BP	GO:0044648	histone H3-K4 dimethylation	0.005223591	0.047684658	0.024389459	PRDM9
BP	GO:1902022	L-lysine transport	0.005223591	0.047684658	0.024389459	AGT
BP	GO:1903401	L-lysine transmembrane transport	0.005223591	0.047684658	0.024389459	AGT
BP	GO:1904180	negative regulation of membrane depolarization	0.005223591	0.047684658	0.024389459	BOK
BP	GO:0045834	positive regulation of lipid metabolic process	0.005377092	0.048604691	0.024860031	AGT/SIRT3
BP	GO:0031056	regulation of histone modification	0.005589388	0.049127338	0.025127352	PRDM9/CTBP1
BP	GO:0006606	protein import into nucleus	0.005805511	0.049127338	0.025127352	AGT/KPNA6
BP	GO:2000377	regulation of reactive oxygen species metabolic process	0.005951711	0.049127338	0.025127352	AGT/SIRT3
BP	GO:0002016	regulation of blood volume by renin-angiotensin	0.005967746	0.049127338	0.025127352	AGT
BP	GO:0003344	pericardium morphogenesis	0.005967746	0.049127338	0.025127352	FLRT3
BP	GO:0007021	tubulin complex assembly	0.005967746	0.049127338	0.025127352	TBCE
BP	GO:0034983	peptidyl-lysine deacetylation	0.005967746	0.049127338	0.025127352	SIRT3
BP	GO:1900119	positive regulation of execution phase of apoptosis	0.005967746	0.049127338	0.025127352	BOK
BP	GO:1901668	regulation of superoxide dismutase activity	0.005967746	0.049127338	0.025127352	SIRT3
BP	GO:1903400	L-arginine transmembrane transport	0.005967746	0.049127338	0.025127352	AGT
BP	GO:0051170	import into nucleus	0.006099599	0.049768413	0.025455245	AGT/KPNA6
CC	GO:0017053	transcription repressor complex	0.000878766	0.047645444	0.030569508	HDAC1/CTBP1
CC	GO:0005667	transcription regulator complex	0.002865007	0.047645444	0.030569508	TBP/HDAC1/CTBP1
CC	GO:0036126	sperm flagellum	0.003346901	0.047645444	0.030569508	SEPTIN2/ODF3
CC	GO:0005672	transcription factor TFIIA complex	0.003575986	0.047645444	0.030569508	TBP
CC	GO:0097729	9+2 motile cilium	0.004079708	0.047645444	0.030569508	SEPTIN2/ODF3
CC	GO:0001940	male pronucleus	0.004289757	0.047645444	0.030569508	TBP
CC	GO:0097227	sperm annulus	0.004289757	0.047645444	0.030569508	SEPTIN2
CC	GO:0001520	outer dense fiber	0.005003053	0.047645444	0.030569508	ODF3
CC	GO:0033106	cis-Golgi network membrane	0.005003053	0.047645444	0.030569508	BOK
CC	GO:0032584	growth cone membrane	0.005715874	0.047645444	0.030569508	FLRT3
CC	GO:0001939	female pronucleus	0.006428222	0.047645444	0.030569508	TBP

CC	GO:0097550	transcription preinitiation complex	0.007140095	0.047645444	0.030569508	TBP
CC	GO:0016580	Sin3 complex	0.007851495	0.047645444	0.030569508	HDAC1
CC	GO:0018995	host cellular component	0.008562421	0.047645444	0.030569508	KPNA6
CC	GO:0043657	host cell	0.008562421	0.047645444	0.030569508	KPNA6
CC	GO:0045120	pronucleus	0.009272874	0.047645444	0.030569508	TBP
CC	GO:0031514	motile cilium	0.009953422	0.047645444	0.030569508	SEPTIN2/ODF3
CC	GO:0005940	septin ring	0.009982855	0.047645444	0.030569508	SEPTIN2
CC	GO:0016581	NuRD complex	0.009982855	0.047645444	0.030569508	HDAC1
CC	GO:0031105	septin complex	0.009982855	0.047645444	0.030569508	SEPTIN2
CC	GO:0070822	Sin3-type complex	0.009982855	0.047645444	0.030569508	HDAC1
CC	GO:0090545	CHD-type complex	0.009982855	0.047645444	0.030569508	HDAC1
CC	GO:0032156	septin cytoskeleton	0.010692363	0.048812961	0.031318592	SEPTIN2
CC	GO:0098831	presynaptic active zone cytoplasmic component	0.011401399	0.04988112	0.032003926	CTBP1
MF	GO:0017136	NAD-dependent histone deacetylase activity	5.63E-05	0.002495732	0.001196495	HDAC1/SIRT3
MF	GO:0034979	NAD-dependent protein deacetylase activity	6.44E-05	0.002495732	0.001196495	HDAC1/SIRT3
MF	GO:0000979	RNA polymerase II core promoter sequence-specific DNA binding	0.000101711	0.002495732	0.001196495	TBP/HDAC1
MF	GO:0004407	histone deacetylase activity	0.000112368	0.002495732	0.001196495	HDAC1/SIRT3
MF	GO:0033558	protein deacetylase activity	0.000123551	0.002495732	0.001196495	HDAC1/SIRT3
MF	GO:0019213	deacetylase activity	0.000373391	0.00628541	0.003013328	HDAC1/SIRT3
MF	GO:0001046	core promoter sequence-specific DNA binding	0.000456515	0.006586855	0.003157846	TBP/HDAC1
MF	GO:0031703	type 2 angiotensin receptor binding	0.000762237	0.009107596	0.004366331	AGT
MF	GO:0051287	NAD binding	0.000811568	0.009107596	0.004366331	SIRT3/CTBP1
MF	GO:0016811	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amides	0.001301052	0.012826442	0.006149206	HDAC1/SIRT3
MF	GO:0010844	recombination hotspot binding	0.001523934	0.012826442	0.006149206	PRDM9
MF	GO:0035851	Krueppel-associated box domain binding	0.001523934	0.012826442	0.006149206	HDAC1
MF	GO:0140297	DNA-binding transcription factor binding	0.002990457	0.021972627	0.010534037	TBP/HDAC1/CTBP1
MF	GO:0001093	TFIIB-class transcription factor binding	0.003045711	0.021972627	0.010534037	TBP
MF	GO:0016810	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds	0.00396338	0.026686758	0.012794068	HDAC1/SIRT3

MF	GO:0000995	RNA polymerase III general transcription initiation factor activity	0.004565334	0.028818669	0.013816142	TBP
MF	GO:0031702	type 1 angiotensin receptor binding	0.005324338	0.031632834	0.015165298	AGT
MF	GO:0031701	angiotensin receptor binding	0.006082806	0.034131299	0.016363103	AGT
MF	GO:0017162	aryl hydrocarbon receptor binding	0.006840736	0.034545717	0.016561782	TBP
MF	GO:0047144	2-acylglycerol-3-phosphate O-acyltransferase activity	0.006840736	0.034545717	0.016561782	MBOAT7
MF	GO:0031078	histone deacetylase activity (H3-K14 specific)	0.007598129	0.034882322	0.016723156	HDAC1
MF	GO:0032041	NAD-dependent histone deacetylase activity (H3-K14 specific)	0.007598129	0.034882322	0.016723156	HDAC1
MF	GO:0046975	histone methyltransferase activity (H3-K36 specific)	0.008354987	0.034975523	0.016767838	PRDM9
MF	GO:0051400	BH domain binding	0.008354987	0.034975523	0.016767838	BOK
MF	GO:0003714	transcription corepressor activity	0.008657308	0.034975523	0.016767838	HDAC1/CTBP1
MF	GO:0070403	NAD+ binding	0.012131239	0.047125197	0.022592591	SIRT3
MF	GO:0042800	histone methyltransferase activity (H3-K4 specific)	0.012884885	0.048199015	0.023107396	PRDM9

Apêndice C. Results of the FLM enrichment analysis from the genes prioritized in the functional annotation study with whole-genome sequence data.

ONTOLOGY	ID	Description	p-value	p-adjust	q-value	Gene ID
BP	GO:2000303	regulation of ceramide biosynthetic process	4.50E-05	0.016147898	0.00952398	SPHK1/SIRT3
BP	GO:0090153	regulation of sphingolipid biosynthetic process	5.31E-05	0.016147898	0.00952398	SPHK1/SIRT3
BP	GO:1905038	regulation of membrane lipid metabolic process	5.31E-05	0.016147898	0.00952398	SPHK1/SIRT3
BP	GO:0007200	phospholipase C-activating G protein-coupled receptor signaling pathway	8.85E-05	0.020171805	0.011897267	AGT/F2RL3/P2RY8
BP	GO:0035025	positive regulation of Rho protein signal transduction	0.000356002	0.045847289	0.027040587	F2RL3/P2RY8
BP	GO:0051482	positive regulation of cytosolic calcium ion concentration involved in phospholipase C-activating G protein-coupled signaling pathway	0.000378063	0.045847289	0.027040587	F2RL3/P2RY8
BP	GO:0045911	positive regulation of DNA recombination	0.000723793	0.045847289	0.027040587	TNFSF13/PRDM9
BP	GO:0048146	positive regulation of fibroblast proliferation	0.000754886	0.045847289	0.027040587	AGT/SPHK1
BP	GO:2001238	positive regulation of extrinsic apoptotic signaling pathway	0.000754886	0.045847289	0.027040587	AGT/TNFSF12
BP	GO:0046521	sphingoid catabolic process	0.00085461	0.045847289	0.027040587	SPHK1
BP	GO:1902551	regulation of catalase activity	0.00085461	0.045847289	0.027040587	SIRT3
BP	GO:1902553	positive regulation of catalase activity	0.00085461	0.045847289	0.027040587	SIRT3
BP	GO:1905008	regulation of L-lysine import across plasma membrane	0.00085461	0.045847289	0.027040587	AGT
BP	GO:1905010	positive regulation of L-lysine import across plasma membrane	0.00085461	0.045847289	0.027040587	AGT
BP	GO:1905364	regulation of endosomal vesicle fusion	0.00085461	0.045847289	0.027040587	SPHK1
BP	GO:1905541	regulation of L-arginine import across plasma membrane	0.00085461	0.045847289	0.027040587	AGT
BP	GO:1905589	positive regulation of L-arginine import across plasma membrane	0.00085461	0.045847289	0.027040587	AGT
BP	GO:0051785	positive regulation of nuclear division	0.001063305	0.047992009	0.028305536	SPHK1/PRDM9
BP	GO:0051353	positive regulation of oxidoreductase activity	0.001138787	0.047992009	0.028305536	AGT/SIRT3
BP	GO:0046579	positive regulation of Ras protein signal transduction	0.001338448	0.047992009	0.028305536	F2RL3/P2RY8

BP	GO:0046513	ceramide biosynthetic process	0.001380251	0.047992009	0.028305536	SPHK1/SIRT3
BP	GO:0051057	positive regulation of small GTPase mediated signal transduction	0.001690215	0.047992009	0.028305536	F2RL3/P2RY8
BP	GO:0003051	angiotensin-mediated drinking behavior	0.001708534	0.047992009	0.028305536	AGT
BP	GO:0007199	G protein-coupled receptor signaling pathway coupled to cGMP nucleotide second messenger	0.001708534	0.047992009	0.028305536	AGT
BP	GO:0010963	regulation of L-arginine import	0.001708534	0.047992009	0.028305536	AGT
BP	GO:0035106	operant conditioning	0.001708534	0.047992009	0.028305536	AGT
BP	GO:0097639	L-lysine import across plasma membrane	0.001708534	0.047992009	0.028305536	AGT
BP	GO:1905435	regulation of histone H3-K4 trimethylation	0.001708534	0.047992009	0.028305536	PRDM9
BP	GO:1905437	positive regulation of histone H3-K4 trimethylation	0.001708534	0.047992009	0.028305536	PRDM9
BP	GO:2000470	positive regulation of peroxidase activity	0.001708534	0.047992009	0.028305536	SIRT3
BP	GO:0042058	regulation of epidermal growth factor receptor signaling pathway	0.001736961	0.047992009	0.028305536	AGT/RHBDF2
BP	GO:0002437	inflammatory response to antigenic stimulus	0.001784322	0.047992009	0.028305536	CD68/RHBDF2
BP	GO:1901184	regulation of ERBB signaling pathway	0.002030295	0.047992009	0.028305536	AGT/RHBDF2
BP	GO:0048145	regulation of fibroblast proliferation	0.002081318	0.047992009	0.028305536	AGT/SPHK1
BP	GO:0048144	fibroblast proliferation	0.002132947	0.047992009	0.028305536	AGT/SPHK1
BP	GO:0007204	positive regulation of cytosolic calcium ion concentration	0.00232754	0.047992009	0.028305536	AGT/F2RL3/P2RY8
BP	GO:0035023	regulation of Rho protein signal transduction	0.002400167	0.047992009	0.028305536	F2RL3/P2RY8
BP	GO:0002019	regulation of renal output by angiotensin	0.002561775	0.047992009	0.028305536	AGT
BP	GO:0002035	brain renin-angiotensin system	0.002561775	0.047992009	0.028305536	AGT
BP	GO:0006311	meiotic gene conversion	0.002561775	0.047992009	0.028305536	PRDM9
BP	GO:1904200	iodide transmembrane transport	0.002561775	0.047992009	0.028305536	SLC5A5
BP	GO:2000468	regulation of peroxidase activity	0.002561775	0.047992009	0.028305536	SIRT3
BP	GO:2001255	positive regulation of histone H3-K36 trimethylation	0.002561775	0.047992009	0.028305536	PRDM9
BP	GO:0019216	regulation of lipid metabolic process	0.002584859	0.047992009	0.028305536	AGT/SPHK1/SIRT3
BP	GO:0007568	aging	0.002765866	0.047992009	0.028305536	AGT/CD68/SIRT3
BP	GO:0045765	regulation of angiogenesis	0.002835724	0.047992009	0.028305536	AGT/TNFSF12/SPHK1
BP	GO:1901342	regulation of vasculature development	0.002978713	0.047992009	0.028305536	AGT/TNFSF12/SPHK1

BP	GO:0051480	regulation of cytosolic calcium ion concentration	0.003101233	0.047992009	0.028305536	AGT/F2RL3/P2RY8
BP	GO:0006672	ceramide metabolic process	0.003355735	0.047992009	0.028305536	SPHK1/SIRT3
BP	GO:0000416	positive regulation of histone H3-K36 methylation	0.003414332	0.047992009	0.028305536	PRDM9
BP	GO:0002018	renin-angiotensin regulation of aldosterone production	0.003414332	0.047992009	0.028305536	AGT
BP	GO:0002034	maintenance of blood vessel diameter homeostasis by renin-angiotensin	0.003414332	0.047992009	0.028305536	AGT
BP	GO:0002605	negative regulation of dendritic cell antigen processing and presentation	0.003414332	0.047992009	0.028305536	CD68
BP	GO:0003072	renal control of peripheral vascular resistance involved in regulation of systemic arterial blood pressure	0.003414332	0.047992009	0.028305536	AGT
BP	GO:0010520	regulation of reciprocal meiotic recombination	0.003414332	0.047992009	0.028305536	PRDM9
BP	GO:0010845	positive regulation of reciprocal meiotic recombination	0.003414332	0.047992009	0.028305536	PRDM9
BP	GO:0014873	response to muscle activity involved in regulation of muscle adaptation	0.003414332	0.047992009	0.028305536	AGT
BP	GO:0035822	gene conversion	0.003414332	0.047992009	0.028305536	PRDM9
BP	GO:0051387	negative regulation of neurotrophin TRK receptor signaling pathway	0.003414332	0.047992009	0.028305536	AGT
BP	GO:0097676	histone H3-K36 dimethylation	0.003414332	0.047992009	0.028305536	PRDM9
BP	GO:1901671	positive regulation of superoxide dismutase activity	0.003414332	0.047992009	0.028305536	SIRT3
BP	GO:2000622	regulation of nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	0.003414332	0.047992009	0.028305536	NBAS
BP	GO:2000623	negative regulation of nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	0.003414332	0.047992009	0.028305536	NBAS
BP	GO:2001253	regulation of histone H3-K36 trimethylation	0.003414332	0.047992009	0.028305536	PRDM9
BP	GO:0030148	sphingolipid biosynthetic process	0.003420483	0.047992009	0.028305536	SPHK1/SIRT3
CC	GO:0070618	Grb2-Sos complex	0.00086961	0.03649372	0.025609628	KHDRBS1
CC	GO:0005690	U4atac snRNP	0.002606694	0.03649372	0.025609628	PRPF31
CC	GO:0070939	Dsl1/NZR complex	0.002606694	0.03649372	0.025609628	NBAS
MF	GO:0005164	tumor necrosis factor receptor binding	0.000369073	0.021750966	0.011935021	TNFSF12/TNFSF13

MF	GO:0032813	tumor necrosis factor receptor superfamily binding	0.0009243	0.021750966	0.011935021	TNFSF12/TNFSF13
MF	GO:0008507	sodium:iodide symporter activity	0.000925573	0.021750966	0.011935021	SLC5A5
MF	GO:0031703	type 2 angiotensin receptor binding	0.000925573	0.021750966	0.011935021	AGT
MF	GO:0008481	sphinganine kinase activity	0.00185034	0.024847419	0.013634082	SPHK1
MF	GO:0010844	recombination hotspot binding	0.00185034	0.024847419	0.013634082	PRDM9
MF	GO:0017050	D-erythro-sphingosine kinase activity	0.00185034	0.024847419	0.013634082	SPHK1
MF	GO:0015111	iodide transmembrane transporter activity	0.002774301	0.032598034	0.017886939	SLC5A5
MF	GO:0005497	androgen binding	0.003697457	0.034756094	0.019071093	SHBG
MF	GO:0030622	U4atac snRNA binding	0.003697457	0.034756094	0.019071093	PRPF31
MF	GO:0001727	lipid kinase activity	0.004619808	0.036188499	0.019857071	SPHK1
MF	GO:0015057	thrombin-activated receptor activity	0.004619808	0.036188499	0.019857071	F2RL3
MF	GO:0031702	type 1 angiotensin receptor binding	0.006462101	0.043369507	0.023797378	AGT
MF	GO:0031701	angiotensin receptor binding	0.007382044	0.043369507	0.023797378	AGT
MF	GO:0038036	sphingosine-1-phosphate receptor activity	0.007382044	0.043369507	0.023797378	SPHK1
MF	GO:0070990	snRNP binding	0.007382044	0.043369507	0.023797378	PRPF31
MF	GO:0030621	U4 snRNA binding	0.008301185	0.045900667	0.025186256	PRPF31
MF	GO:0048018	receptor ligand activity	0.0095508	0.047644198	0.026142953	AGT/TNFSF12/TNFSF13
MF	GO:0030546	signaling receptor activator activity	0.009984395	0.047644198	0.026142953	AGT/TNFSF12/TNFSF13
MF	GO:0046975	histone methyltransferase activity (H3-K36 specific)	0.010137063	0.047644198	0.026142953	PRDM9

Apêndice D. Tree plot for the terms associated with the TH prioritized genes using WGS data.

Apêndice E. Tree plot for the terms associated with the FLM prioritized genes using WGS data.