SÃO PAULO STATE UNIVERSITY
SCHOOL OF AGRICULTURAL AND VETERINARIAN SCIENCES
CAMPUS OF JABOTICABAL

ASSESSMENT OF SELECTION SIGNALS IN BRAZILIAN HORSE POPULATIONS

Wellington Bizarria dos Santos
Animal Scientist

2024

SÃO PAULO STATE UNIVERSITY
SCHOOL OF AGRICULTURAL AND VETERINARIAN SCIENCES
CAMPUS OF JABOTICABAL

ASSESSMENT OF SELECTION SIGNALS IN BRAZILIAN HORSE POPULATIONS

Wellington Bizarria dos Santos

Advisor: Prof. Dr. Henrique Nunes de Oliveira
Co-Advisor: Prof. Dr. Rogério Abdallah Curi

Thesis presented to the School of Agricultural and Veterinary Sciences at São Paulo State University, Campus of Jaboticabal, in partial fulfillment of the requirements for a Ph.D. degree in Genetics and Animal Breeding.

2024

S237a

Santos, Wellington Bizarria dos
Assessment of selection signals in Brazilian horse populations / Wellington Bizarria dos Santos. -- Jaboticabal, 2024
75 p. : il., tabs.

Tese (doutorado) - Universidade Estadual Paulista (UNESP), Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal
Orientador: Henrique Nunes de Oliveira
Coorientador: Rogerio Abdallah Curi

1. Equus caballus. 2. Marchas de cavalos. 3. Cavalos de corrida. 4. Genomics. I. Título.

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

Essa ficha não pode ser modificada.

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Jaboticabal

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: ASSESSMENT OF SELECTION SIGNALS IN BRAZILIAN HORSE POPULATIONS

AUTOR: WELLINGTON BIZARRIA DOS SANTOS
ORIENTADOR: HENRIQUE NUNES DE OLIVEIRA
COORIENTADOR: ROGERIO ABDALLAH CURI

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

Documento assinado digitalmente
gov.br
HENRIQUE NUNES DE OLIVEIRA
Data: 02/02/2024 09:22:57-0300
Verifique em https://validar.iti.gov.br

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

Documento assinado digitalmente
gov.br
DANISIO PRADO MUNARI
Data: 31/01/2024 19:11:39-0300
Verifique em https://validar.iti.gov.br

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

Documento assinado digitalmente
gov.br
LUIS ARTUR LOYOLA CHARDULO
Data: 01/02/2024 12:06:48-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. LUIS ARTUR LOYOLA CHARDULO (Participaçao Virtual)
Departamento de Melhoramento e Nutrição Animal / FMVZ UNESP Botucatu

Documento assinado digitalmente
gov.br
AMANDA MARCHI MAIORANO
Data: 01/02/2024 08:48:18-0300
Verifique em https://validar.iti.gov.br

Profa. Dra. AMANDA MARCHI MAIORANO (Participaçao Virtual)
Departamento de Zootecnia / Universidade Federal de Uberlândia - Uberlândia/MG

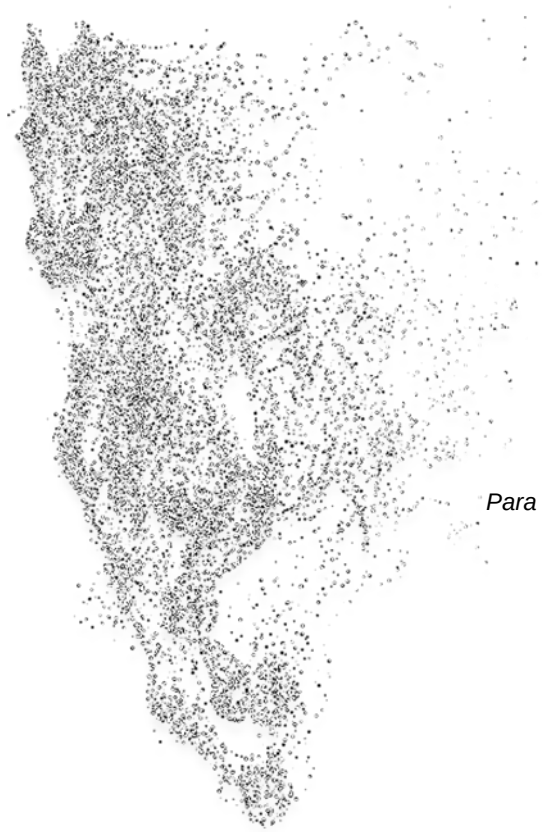
Documento assinado digitalmente
gov.br
FRANCISCO RIBEIRO DE ARAUJO NETO
Data: 02/02/2024 08:46:14-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. FRANCISCO RIBEIRO DE ARAUJO NETO (Participaçao Virtual)
Departamento de Zootecnia / Instituto Federal Goiano (IFG) - Rio Verde/GO

Jaboticabal, 31 de janeiro de 2024

AUTHOR CURRICULUM INFORMATION

Wellington Bizarria dos Santos - born on May 3, 1993, in Garanhuns – PE. In 2017, he earned his Bachelor's degree in Animal Science from the Federal Rural University of Pernambuco (UFRPE). In February 2020, he completed his master's degree under the guidance of Prof. Dr. Rogério Abdallah Curi. The thesis, titled "Genome-wide Scan for Selection Signatures and Estimates of Autozygosity Levels in Mangalarga Marchador Horses," was published in the Journal of Animal Breeding and Genetics and BMC Genomics. The studies were financed by CAPES scholarship, granted by the "Coordenação de Aperfeiçoamento de Pessoal de Nível Superior", and the project itself secured funding from the "Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP)". During the doctorate, he was supported by a CNPq scholarship from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (141564/2020-2).



“Somos assim: sonhamos o voo, mas tememos a altura.
Para voar é preciso ter coragem para enfrentar o terror do vazio.
Porque é só no vazio que o voo acontece.
O vazio é o espaço da liberdade, a ausência de certezas.
Por isso trocamos o voo por gaiolas.
As gaiolas é o lugar onde as certezas moram.”
Fiodor Dostoïevski

Portanto,

“A realidade
Sempre é mais ou menos
Do que nós queremos!
Segue o teu destino,
Rega as tuas plantas,
Ama as tuas rosas.
O resto é a sombra
De árvores alheias.”
Fernando Pessoa

ACKNOWLEDGMENTS

I express my gratitude to the Great Architect of the Universe for the perfect algorithm. There are things that do not need a scientific explanation to be what they are.

I would like to extend special thanks to my advisor, Prof. Henrique Nunes de Oliveira, for his invaluable guidance, advice, and friendship. Particularly noteworthy is his support and understanding during the most challenging stages of my life while pursuing this Ph.D. He demonstrated that strength would guide me, reminiscent of the teachings of the Jedi. I would like to one day be at least the average of this great researcher.

I am profoundly grateful to my co-advisor, Prof. Rogério Abdallah, for the invaluable opportunity he has provided, as well as for his unwavering commitment to my academic development since my master's degree. His wise counsel, continuous guidance, devoted time, and friendship.

I would like to express my sincere gratitude to Prof. Amanda Maiorano for her significant contributions to my professional education and the invaluable friendship that we developed during the postgraduate studies. Amanda followed my work, read my constant mistakes and was always present in situations that made my difficulties solvable.

I extend my heartfelt appreciation to Prof. Dr. Danísio Munari for his invaluable contributions to my education, his wise guidance, friendship, and the dedication he has invested in mentoring his students. I am very grateful for the opportunity to have been one of them.

I would like to express my gratitude to Prof. Arthur Chardulo for his invaluable contributions to my education since my master's degree. His consistent efforts in providing constructive feedback and enhancing the quality of our work have been greatly appreciated.

I would also like to express my gratitude to Prof. Francisco Neto for his contributions to the correction of this thesis. His insightful comments were essential in enhancing the quality of this work.

I express my gratitude to the Postgraduate Program in Genetics and Animal Breeding and the professors who taught the topics during my PhD, including Prof. Maria Eugênia Mercadante, Prof. Lucia Galvão, Prof. Roberto Carvalheiro, Prof. Mauricio Barbanti, Prof. Mateus Paranhos, Prof. Otavio Neto, Prof. Fernando Baldi, Prof. Danilo Millen, and Prof. Garret Suen.

Throughout this journey, I received support from individuals not directly involved in the project, and I am immensely grateful for their contributions. Special mention goes to Prof. Antonio Xisto, Jorge Lucena, Prof. Geane Dias, Prof. Josabete Carvalho, Prof. Marcelo Milfont, Prof. Omer Cavalcanti, Kleber Regis, and Prof. Elizabel Oliveira. I extend my heartfelt thanks to my family (Bizarria and Vilela) for their unwavering support and affection during the most challenging moments. Gratitude also goes to my three mothers (Cristina, Lília, and Iraci), grandfathers (Manoel and Elizeu), uncle (Edenildo), aunts (Josenilda and Joseilda), nephews (Lucas, Larissa, and João Victor), and brothers (IBD/IBS) – Wedja & Wedson, Josiclecia & Josicleide.

I also want to express my appreciation to my friends, those who have been part of my life since my High school and undergraduate degree (Marconi, Ially, Thamyllys, Sany, Matheus, Edu, Jacy, Monique, Luh, Gabriel, Will, Almir, Diana, Rita, Bruno, Mário, Dany, Dania, and Júlia), as well as those I met here in São Paulo City (Edgar, Elena, Davi, Cassinha, Erick, Miller, Zafia, Eduardo & Matheus, Luís, Gustavo, Alejandra, Bruna, Anita, Cherlyn, Thiago, and Sabrina). Thank you for your support, attention, and motivation.

This study was financed by the *CNPq* scholarship from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (n141564/2020-2).

CONTENTS

ASSESSMENT OF SELECTION SIGNALS IN BRAZILIAN HORSE POPULATIONS

ABSTRACT – Genetic improvement in horses demands an incessant search for the selection and improvement of desirable attributes over successive generations. This study involved identifying certain genomic regions associated with economic traits in the Quarter Horse racing line and Mangalarga Marchador breed. For the Quarter Horse, individual genomic inbreeding coefficients were estimated by analyzing runs of homozygosity (FROH). With a sample of 336 registered animals, genomic inbreeding with moderate- to high-level variations was identified, resulting in 46,594 runs of homozygosity (ROH) and 16,101 regions of heterozygosity (ROHet). The genomic islands associated with ROH and ROHet presented candidate genes that influence essential biological processes, such as cell differentiation, metabolic regulation of glucose, heme transport proteins, and import of calcium ions. In the analyses concerning the Mangalarga Marchador, genomic regions associated with different gait patterns were identified, considering the genetic bases of the "Picada" and "Batida" phenotypes. The sample used comprises 192 animals, 62 males and 130 females. The results pointed to eight notable candidate genes (*PIP5K1B*, *PABIR1*, *PGM5*, *DOCK8*, *KANK1*, *DMRT1*, *DMRT3*, and *DMRT2*) based on genome-wide association studies (GWAS). Complementary analyses included extended haplotype homozygosity (EHH), including providing new evidence about the *DMRT3* gene, highlighting the single nucleotide polymorphism (SNP) AX-103466344, which consistently segregated the "Batida" and "Picada" gaits in homozygous conditions. Thus, notable insights and genomic regions relevant to the selection of Quarter Horses with greater regenerative capacity and potential treatments for muscular disorders were identified. Furthermore, new discoveries about the possible causal relationship between mutations in the gait modalities of the Mangalarga Marchador may improve selection within the breed.

Keywords: Equus caballus, genomics, horse gaits, racing horses

AVALIAÇÃO DE SINAIS DE SELEÇÃO EM POPULAÇÕES DE CAVALOS BRASILEIROS

RESUMO - O melhoramento genético em equinos demanda uma incessante busca pela seleção e aprimoramento de atributos desejáveis ao longo das sucessivas gerações. O estudo consistiu em identificar determinadas regiões genômicas associadas a características econômicas no Quarto de Milha da linhagem corrida e Mangalarga Marchador. No Quarto de Milha foram estimados os coeficientes de endogamia genômica individual por meio da análise de corridas de homozigose (FROH). Com uma amostra de 336 animais registrados foram identificadas endogamias genômicas com variações de nível moderado a elevado, resultando em 46.594 corridas de homozigose (ROH) e 16.101 regiões em heterozigosidade (ROHet). As ilhas genômicas associadas a ROH e ROHet apresentaram genes candidatos que exercem influência em processos biológicos essências, tais como diferenciação celular, regulação metabólica da glicose, proteínas de transporte de heme e importação de íons de cálcio. Nas análises que concernem ao Mangalarga Marchador foram identificadas regiões genômicas associadas aos distintos padrões de marcha, considerando as bases genéticas dos fenótipos "Picada" e "Batida". A amostra utilizada incluiu 192 animais, sendo 62 machos e 130 fêmeas. Os resultados apontaram para oito genes candidatos notáveis (*PIP5K1B, PABIR1, PGM5, DOCK8, KANK1, DMRT1, DMRT3* e *DMRT2*) por estudos de Associação Genômica Ampla (GWAS). Análises complementares foram incluídas como a homozigose do haplótipo estendido (EHH), inclusive fornecendo novas evidências acerca do gene *DMRT3*, destacando o SNP (*AX-103466344*), que segregou consistentemente a marcha nas duas modalidades "Batida" e "Picada" em condições homozigóticas. Destarte, foram identificadas regiões genômicas notáveis relevantes para a seleção de animais Quarto de Milha com maior capacidade regenerativa e potenciais tratamentos para distúrbios musculares. Além disso, novas descobertas sobre a possível relação causal entre mutações nas modalidades de marcha do Mangalarga Marchador podem aprimorar a seleção dentro da raça.

Palavras-chave: Equus caballus, genômica, marchas de cavalos, cavalos de corrida

CHAPTER I – GENERAL CONSIDERATIONS

1. INTRODUCTION

Equine genetics has evolved in studies to understand the distinct traits present in the most varied breeds of horses, and there is a constant search in production animals for the selection and improvement of desirable attributes over generations. This study focuses on two prominent horse breeds in Brazil: Quarter Horse and the Mangalarga Marchador, each with unique characteristics and specific applications.

Originating in the United States around the year 1600, the Quarter Horse breed stands out for its versatility and remarkable performance in various disciplines, such as rural equestrianism, flat racing, and rodeos. Its anatomy, with robust rear muscles, gives it remarkable starting capacity and speed over short distances. In addition to its success in sporting competitions, the Quarter Horse plays a significant role in agricultural activities, highlighting its adaptability and usefulness in everyday farming (Michele, 2021; ABQM, 2023).

On the other hand, the Mangalarga Marchador, a genuine Brazilian breed, originated approximately 200 years ago, resulting from the crossing between Alter horses from Portugal and animals selected in southern Minas. It is characterized by its “Batida” and “Picada” gaits, giving it unique natural movements and intermediate speeds. In the “Batida” gait, diagonal supports predominate in relation to lateral and triple supports, establishing a distinction relative to the “Picada” gait (Beck, 1992; USMMA, 2018). In addition to their applications in sports, riding, and transport, the Mangalarga Marchador horses are recognized for their intelligence and docility, facilitating their training, and highlighting their versatility in various activities (Andrade, 2016; ABCCMM, 2023).

Given the economic and cultural relevance of these breeds, it is imperative to deepen the understanding of their genetic bases. Efficacy in identifying linkage disequilibrium between single nucleotide polymorphism (SNP) markers and causal alterations has been demonstrated by the adoption of high-density chips in recent genetic studies in domestic animals, marking substantial advances in the genetic

understanding of these traits. Genome-wide association studies (GWAS) have contributed to dissection of the genetic mechanisms of diseases and complex phenotypes in populations of economic interest.

An alternative approach to investigating significant regions in the genome of a specific species or breed is the assessment of inbreeding levels. Specific segments of these regions can be used for more efficient management of areas with low genetic diversity, called homozygous regions. This management is crucial because homozygosity in these regions is associated with decreased performance in traits of economic importance. Furthermore, regions with high heterozygosity have been explored in the study of equilibrium or negative selection, introgression, and admixture, or in hypervariable regions.

Therefore, the central objective of this study is to identify the genomic regions associated with specific economic traits in Quarter Horses and Mangalarga Marchador horses. Specifically, this entails estimating genomic endogamy, identifying runs of homozygosity (ROH) and heterozygosity-rich regions (ROHet) within the Quarter Horse racing line, and discerning both rare and common variants, as well as DNA polymorphisms, linked to the gait phenotype in Mangalarga Marchador horses.

2. STATE OF THE ART

2.1 Horse Farming

Regarding the global equine population, the United States leads the way with approximately 10.26 million horses. Mexico follows in second place with a substantial population of around 6.36 million, while China boasts a similar number, totaling around 6.03 million horses. In this context, it is pertinent to note that Brazil holds the fourth position globally in terms of the equine population, with a total of approximately 5.46 million horses (GLIPHA, 2017).

In addition to the quantitative aspects of the equine population, it is essential to emphasize the diversity of Brazil’s equine population, which has evolved over time, adapting to the needs of breeders and the unique characteristics of various regions

within the country (Gurgul et al., 2019). Noteworthy examples include the Mangalarga, the Mangalarga Marchador, the Quarter Horse, the Campolina, the Crioulo, the Pampa, the Nordestino, and the Brazilian Equestrian.

Within the Brazilian context, horses serve a multitude of roles and purposes. This encompasses leisure and sport horses, totaling 1.1 million individuals, as well as working horses, which comprise a contingent of 3.9 million (MAPA, 2016). In terms of equestrian sports, there is a wide range of competitions, primarily including classic show jumping, dressage, eventing, cross-country, endurance riding, horse racing, barrel racing, pollo, vaulting, and reining. It is worth noting that these animals display diverse locomotion patterns, as categorized by Ensminger (1990), which may be natural (untrained) or artificial (requiring a high degree of training for the horse–rider partnership).

Furthermore, horses can be classified in non-conventional roles as companions, for military use, and as production animals. For many years, the perspective of viewing horses as production animals was not common in Brazil. Thus, historically, there was a ban on horse slaughter for various reasons related to culture, animal welfare, and changes in legislation. However, in 2016, this ban was lifted, allowing horse slaughter to resume under stringent regulations. This market primarily serves a specific niche, with most products being exported. At present, Brazil ranks fifth in horse meat production globally (FAO, 2017). Horses also play a crucial role in the production of antibodies for the pharmaceutical industry (Lönker et al., 2020) and contribute to equine therapy, which offers therapeutic interventions for patients with specific neurological pathologies, focusing on rehabilitation and mental and motor re-education (Telles & Moço, 2022).

The diverse contribution of horses in the Brazilian context exerts a fundamental influence in the areas of economic development, agricultural sustainability, and the country’s rich cultural heritage. The data from the "Study of the Equine Agro-industrial Complex," conducted by the Ministry of Agriculture, Livestock, and Supply (MAPA) in 2016, highlights the substantial economic impact of the equine production chain, estimated at over R$16.15 billion annually. It is important to note that ongoing efforts are aimed at updating these figures, with Esalq/USP (Luiz de Queiroz College of Agriculture

– USP) leading the initiative to assess the sector's gross domestic product (GDP). Current estimates indicate that the equine agro-industrial complex already generates over R$30 billion annually (ABQM, 2022). According to Araújo (2013), there are approximately 30 segments within the horse agribusiness, from input production to the destination, generating more than 3 million direct and indirect jobs.

2.2 Molecular Markers

Molecular markers have been characterized as discernible variations in the expression of a specific gene or DNA sequence, each with a well-defined position on the chromosome. These markers are characterized by their ability to be precisely quantified and traced within a population and can be linked to particular genes or traits of interest (Hayward et al., 2015). Currently, there are various types of molecular markers, including SNPs, restriction fragment length polymorphisms (RFLPs), DNA fingerprinting (DFP), random amplified length polymorphic DNAs (RAPDs), microsatellites or single sequence repeats (SSRs), amplified fragment length polymorphisms (AFLPs), and many more (Raj & Das, 2021).

Remarkably, there have been few studies on designing reliable markers, and there are almost no criteria for assessing the qualities that distinguish a marker as either "good" or "poor" (Platten et al., 2019). Each technique has its advantages and disadvantages because there is no perfect genetic marker. It is important to note that to be considered ideal, a marker should exhibit polymorphism and an even distribution across the genome; the ability to finely resolve genetic variations; the capacity to yield multiple, independent, and dependable markers; simplicity; speed; cost-effectiveness; minimal tissue and DNA sample requirements; clear associations with distinct phenotypes; and independence from prior knowledge about an organism’s genome (Mondini et al., 2009).

Although there is no ideal marker, the one that is currently most widely used is the SNP-type marker due to their large numbers in virtually all populations of individuals (Kumar et al., 2012). However, there are some exceptions to consider. For example, when testing the association between complex phenotypic traits and genetic loci, single-

locus SNP analyses present a loss of information due to the bi-allelic nature of the markers, compared with multi-allelic microsatellites. However, this limitation can be overcome by performing haplotype frequency estimates at multiple SNPs within a locus (Fallin et al., 2001) and even possibly improved, because SNPs will be closer to the site responsible for variation than microsatellites. Therefore, SNPs can assess the variations in single nucleotides (Vallejos-Vidal et al., 2019), and due to their distribution, they have proved to be highly effective in population contexts (Peñaloza et al., 2020). Currently, they are extensively employed in genetic improvement studies in farm animal populations, as well as in biodiversity conservation and health promotion.

2.3 Sequencing and SNP Genotyping

Over the past two decades, there has been a notable evolutionary trajectory of the genomics field since the publication of the first human genome sequence in prestigious journals such as *Nature* and *Science* (Lander et al., 2001; Venter et al., 2001). These pioneering contributions marked the genesis of the genomic era, albeit leaving conspicuous lacunae in comprehension. A milestone materialized with the culmination of the first complete human genome in 2003, covering approximately 92% of the entire human genome sequence (NIH, 2022), thereby establishing a foundational basis for subsequent transformative advancements.

The advent of Sanger sequencing by Frederick Sanger in the 1970s played a pivotal role in catalyzing the Human Genome Project, thereby facilitating the elucidation of genetic information (Sanger et al., 1977). Concurrently, the proposition by Allan Maxam and Walter Gilbert of chemical degradation presented an alternative avenue for decoding DNA sequences through controlled chemical modifications (Maxam and Gilbert, 1977). The mid-2000s witnessed the initiation of the next-generation sequencing (NGS) era, promising a genomic revolution by enabling the parallel analysis of millions, if not billions, of DNA fragments (ASM, 2019). This commitment materialized with third-generation sequencing, heralding extended reads and the capability to explore complex genomic regions, including RNA sequencing and diverse omics studies (Ozsolak et al.,

2012; Sheynkman et al., 2016; Moldován et al., 2018; Meslier et al., 2022; Perez et al., 2023).

Termed the “Genomic Era,” this nomenclature underscores the extraordinary strides made in biological information production. The acquisition of longer reads, particularly with the advent of third-generation sequencing synthesis (TGS) in 2020, has notably augmented the precision and quality of genetic sequences. However, the computational infrastructure has yet to fully align with the exigencies of this epoch (Athanasopoulou et al., 2022). The envisaged integration of quantum computers and more robust data processing languages stands poised to play a pivotal role in the future exploration of this wealth of genetic information.

Oligonucleotide microarrays, initially conceived for sequencing via hybridization, have undergone substantial evolution. Solid support-based oligonucleotide hybridization methods, characterized by extensive parallelization, have emerged as the predominant platform for genetic analysis. This platform, oriented toward genotyping arrays reliant on single nucleotide variants (SNVs), has assumed a pivotal role in comprehending the phenotypic ramifications of genetic variations such as SNPs (Verlouw et al., 2021).

Technological advancements in SNP genotyping have ushered in innovative methodologies predicated on direct sequencing of polymerase chain reaction (PCR) products (Ruff et al., 2020). This methodology facilitates the analysis of over 5 million genetic markers in a singular specimen at the nucleotide level. In the arena of array-based SNP genotyping, Affymetrix and Illumina include diverse systems suitable for a variety of projects, from the small to large scale, with applications spanning the entire genome. It is also possible to use predesigned genotyping matrices that are adapted to specific needs.

Affymetrix provides genotyping matrices tailored for numerous species, encompassing livestock and crops, while Illumina furnishes genotyping bead arrays designed for both human and non-human species, distinguished by varying densities between chips. High-density arrays have, in numerous contexts, become the preferred modality owing to their capacity to afford a more intricate analysis of the genome. Notably, high-density genotyping, propelled by Illumina’s Infinium and Golden Gate®

arrays, facilitates robust, large-scale GWAS and the ability to detect mutations and copy number variants.

The simultaneous scrutiny of myriad polymorphisms dispersed across genomes has assumed a pivotal role in elucidating the genetic structure of heterogeneous populations. This comprehensive approach has efficaciously enabled the estimation of intra-population diversity, inter-population genetic divergence, genomic selection, GWAS, detection of selection signatures, analysis of inbreeding, and identification of genomic regions subject to selective processes (Zenger et al., 2006; Hayes et al., 2006; Sabeti et al., 2007; Hancock and Rienzo, 2008; McKay et al., 2008; Muro et al., 2009; Gibbs et al., 2009; Maceachern et al., 2009; Cantor et al., 2010; Howard et al., 2017; Budhlakoti et al., 2022). This holistic methodology has wrought a paradigm shift in the understanding of genetics, imparting unprecedented insights into the nuances of the genetic complexity across diverse species.

2.4 Recent Runs of Homozygosity and Heterozygosity Studies in Horses

Runs of homozygosity (ROH) and runs of heterozygosity (ROHet) have been utilized to understand the demographic history as well as adaptive evolution of populations (Li et al., 2022). ROH are described as long, unbroken stretches of homozygosity in the genome. Due to their long length, these stretches are assumed to have a common ancestor (Ceballos et al., 2018). In contrast, ROHet are not random runs, but rather heterozygosity-rich regions (Williams et al., 2016).

There are several studies that have evaluated ROH in the most diverse equine breeds to assess genomic inbreeding and selection signatures. Szmatoła et al. (2022) reported a higher level of inbreeding for populations of light and primitive horses compared with draft horses. These islands include many genes linked to important breed traits, including *WFIKNN2*, *CACNA1G*, *STXBP4*, *NOG*, *FAM184B*, *QDPR*, and *LCORL*, as well as the zinc finger protein family. Mousavi et al. (2023) investigated 169 horses belonging to the Caspian, Turkmen, Kurdish, and Persian Arabian populations and found different effective population sizes. Phylogeographic analysis clustered breeds according to their geographical origin. Selection signal statistics revealed

candidate SNPs highly significant for morphological traits, with *HMGA2* and *LLPH* as the genes associated with height variation in Caspian horses.

Ablondi et al. (2019) studied homozygous regions in the Swedish Warmblood (SWB) and Exmoor pony populations. The authors observed ROH in SWB on ECA4, ECA6, ECA7, ECA10, and ECA17 that were shared by more than 85% of the genotyped individuals. On the other hand, Exmoor ponies showed uniform distribution of long ROH across 77% of chromosomes. The authors also identified selection signatures on ECA1, ECA4, and ECA6 in SWB based on population differentiation tests FST (The fixation index) and XP-EHH (the cross population extended haplotype homozygosity), indicating that genes associated with behavior, physical abilities, and fertility were targeted for selection.

Santos et al. (2021a) used Tajima's D (TD), the integrated haplotype score (iHS), and ROH to assess the presence of selection signals in Mangalarga Marchador horses. Accordingly, analyses of TD, iHS, and ROH revealed 27, 104, and 38 candidate genes for selection. These genes are involved in important biological processes that concern anterior/posterior patterns, limb morphogenesis skeletal, and muscle stretch response. Laseca et al. (2022) studied the application of increased inbreeding on the fertility of Pura Raza Española mares. When using a genomic approach, an inverse relation ($\rho = -0.144$) emerged between whole-genome homozygosity and fertility estimated breeding values (EBVs). The authors observed different correlations at the chromosome level. Some of the chromosomes exhibited high correlation, while others displayed low correlation. However, among a subset of the least fertile horses, where inbreeding increased by about 30%, the correlation was weakened to −0.241.

Ardestani et al. (2020) investigated the genetic principles of variation in horse types, especially the pony and light horses breeds. The authors aimed to analyze whole-genome sequencing of 14 ponies and 32 light horses as these breeds have their own selective forces and evolutionary dynamics unique to them. They detected new genomic signatures of selective sweeps near the important genes, including *C4ORF33*, *CRB1*, *CPN1*, *FAM13A*, and *FGF12*, that had been associated with body measurements in

previous studies. These results led to highly positive conclusions that may be used to improve the present selection schemes utilized in breeding equines.

The latest discoveries show that ancient and modern situations of inbreeding have varied for different horse breeds. However, it is worth mentioning that recent inbreeding events have also been rare, especially in the indigenous horse breeds (Chen et al., 2023). Colpitts et al. (2022) recently investigated the use of ROH in conservation strategies among natural horse populations. They compared the ROH metrics in domestic versus recently feral horse populations to identify genetic health and differences in ROH between managed and unmanaged settings.

Little is known about ROHet in equine populations due to the small number of published studies on this subject. According to Santos et al. (2020b), the genetic variability on chromosome 11 is highly prevalent in Mangalarga Marchador horses. The authors only reported observations in the 0–2 M ROHet category. They found three heterozygosity-rich regions containing genes with known functions: *TRIM37*, *PPM1E*, and *CA10*. Santos et al. (2023) reported gene regions encoding respiratory capacity (*OR7D19*, *OR7D4G*, *OR7D4E*, and *OR7D4J*) and muscle repair (*EGFR* and *BCL9*) for the Quarter Horse racing line.

The ROHet studies have the capability of studying the population structure as well as demographic history (size problems of populations, genetic bottlenecks, dynamics of metapopulations, genetic variations, and mixing of two formerly isolated populations). As a relatively new approach compared with ROH, it is assumed that the future research will pay much attention to ROHet in other horse breeds.

2.5 GWAS

Numerous traits, such as growth, reproduction, feed efficiency, production, as well as the response to specific medications and susceptibility to certain diseases, are presumably co-determined by genetic factors (Mozzi et al., 2018; Zepeda-Batista et al., 2021; Guo et al., 2022; Sun et al., 2023). Specifically in the context of animal production, there is great interest in identifying genetic peculiarities in populations because selection objectives often vary (Van and Visser, 2021). Therefore, the GWAS approach requires

the availability of comprehensive, high-quality genetic and non-genetic datasets. Technological advances and the decreasing cost of genotyping and sequencing over the last decade have exponentially improved the performance and success of GWAS aimed at finding statistical dependence or associations between a trait of interest and point mutations in DNA (Gumpinger et al., 2018; McCombie et al., 2019).

GWAS research offers diverse scenarios, encompassing classical approaches that investigate the relationship between genetic variants and measurable phenotypic traits (Shaffer et al., 2012), as well as methods that incorporate gene expression data, for example, eQTL (Expression quantitative trait loci) GWAS (Varathan et al., 2022). The nature of the phenotype under study also sheds light on issues of pleiotropy and its interaction with complex diseases and phenotypes (Solovieff et al., 2013). Studies can vary in scale, with extensive datasets involving millions or even billions of associations, or on smaller scales with more limited samples and fewer dispersed markers (Bahram and Inoko, 2007; Ballesta et al., 2020; Hammond et al., 2021). The choice between exploratory and targeted approaches will depend on the specific objectives of the research, as well as the nature of the populations studied, whether diverse or homogeneous. Below, some GWAS contexts are explored in more detail.

- ***Case-Control Study -*** A conventional case-control study operates under the premise that a shared trait among individuals is concomitant with a common genetic variation within this cohort. Nevertheless, there are challenges in replicating and comprehending the effects and interactions of multiple SNPs (Wu et al., 2010), particularly in the exploration of complex diseases and traits through GWAS. These conditions frequently demonstrate a polygenic nature, where multiple SNPs collectively contribute small effect sizes (Banerjee et al., 2018).

Comparisons of statistical power and sample size equivalence are pertinent, particularly when assessing binary and continuous characteristics under a threshold model and considering the similarity in effect sizes (Yang et al., 2010). Banerjee et al. (2018) showcased the significant enhancements offered by the Bayesian multiple logistic regression (B-LORE) method, especially in scenarios involving multiple variants

at loci with heritability ≥ 0.4 and unbalanced case-control distributions. Additionally, B-LORE facilitates meta-analyses by approximating likelihood functions using multivariate normal distributions, incorporating means and covariance matrices as summary statistics.

Fundamental association analysis in case-control GWAS relies on established hypothesis testing methods, primarily employing contingency tables and logistic regression models (Chatterjee, 2017). The extensive genetic data in GWAS can correct for potential biases, including those stemming from hidden population structures and batch-related effects (Chatterjee, 2017).

The growing interest in identifying interactions between genetic and environmental factors arises from the acknowledgment that genetic factors alone cannot fully account for phenotypic variation. Environmental and behavioral changes likely exert a significant influence, and disregarding environmental modifiers may obscure crucial genetic associations, leading to inaccurate effect estimates. Identifying such modifiers is pivotal for managing genetic variant–associated risks (Manolio, 2009).

Yang et al. (2010) undertook a comparison between binary and continuous traits, assuming a threshold model for diseases. They posited that the effect size for disease susceptibility shares similarities with quantitative traits and established an approximate ratio for the non-centrality parameter (NCP) between case-control and quantitative trait association studies. This ratio is influenced by factors such as sample size, disease prevalence (K), and the proportion of cases (v) in the case-control study. Their findings revealed that for rare diseases (prevalence < 0.1), conducting a case-control study with an equal number of cases and controls (v = 0.5) requires a smaller sample size than a study focusing on quantitative traits while maintaining an equivalent level of detection capability. However, when v = K and the sample size remains the same, the power of the case-control association study significantly lags behind that of the quantitative trait association study, primarily due to the loss of information during the transition from a continuous quantitative trait to a binary one.

- ***Regression Analysis -*** Statistical correlation serves as a pivotal tool for elucidating the associations between variables, particularly when investigating potential linear relationships among continuous variables (Mukaka et al., 2012). The assumption underlying such analyses is typically one of linearity, wherein it is implied that a consistent increment or decrement occurs in one variable for every unit change in the other. An alternative yet commonly employed approach in such contexts is regression analysis, which involves the identification of an optimal straight line to succinctly encapsulate the observed association (Bewick et al., 2003).

In the animal science context, the use of regression models has become practically indispensable. Various regression models can be applied, including linear regression, logistic regression, polynomial regression, ridge regression, and lasso regression, among others (Sharma, 2023). In the context of GWAS, Wang et al. (2018) articulated two notable limitations inherent in basic regression models. First, the observed trait association may be a consequence of an indirect influence exerted by the SNP through a covariate. Second, the simplicity of a basic regression model precludes the utilization of additional information pertaining to correlated phenotypes, potentially compromising its statistical power.

Conventional practice involves the use of linear regression models for continuous phenotypes and logistic models for binary phenotypes (Uffelmann et al., 2021). However, owing to the intricate nature of the relationships between genetic variations and phenotypic characteristics, more robust methodologies are essential. Lee et al. (2016) advocated for a methodology grounded in nonlinear regression, facilitating the detection of marginal and pairwise interaction effects on traits. This approach incorporates cutting-edge techniques such as screening, randomization, and stability selection. Consequently, the GWAS literature is replete with a diverse array of applications, ranging from comparative analyses to the adaptation or proposition of novel methodologies.

- ***Family-Based and Structured Populations Analysis -*** When considering genome-wide association studies, significant progress has been made due to

advancements in genotyping technology (Price et al., 2010). These innovations have ushered in a new era of research, allowing for the generation of robust and reproducible findings (Marigorta et al., 2018). Nevertheless, certain challenges persist as fundamental issues in the field of statistical analysis (Sul et al., 2018).

Over the last decade, novel approaches utilizing mixed models have emerged to counter the adverse effects of population structure and genetic relatedness in association studies. These models aim to provide more accurate assessments of genetic associations while addressing these concerns (Sul et al., 2018). However, it is important to note that developing GWAS techniques that effectively correct for population structure remains a complex computational and statistical task (Zeng et al., 2015).

While commonly used methods for GWAS do consider population structure, their application is typically limited to individuals with genotyped data and corresponding phenotypes. Recently, single-step genome-wide association studies (ssGWAS) have gained attention as they enable the utilization of phenotypic information from ungenotyped relatives, expanding the scope of genetic association research (Mancin et al., 2021).

Population stratification is a potential confounder in GWAS research. It arises from systematic differences in ancestry between case and control groups. Although methods have been developed to estimate genetic ancestry and to mitigate this issue, they perform best when population stratification is the sole source of variation in the dataset. These methods may not be suitable when the data also encompass kinship differences or involve individuals with unknown familial relationships (Price et al., 2010).

Two primary sources of relatedness that can lead to elevated rates of false-positive associations are ancestry differences and cryptic relatedness. Ancestry differences refer to variations in ancestry among individuals in a study, with shared ancestry indicating a higher level of relatedness (Sul et al., 2018).

Addressing these challenges becomes particularly crucial in the context of family-based designs in genetic association analysis. In such designs, when both parents are genotyped, employing regression models that account for parental genotype has proved

to be an effective strategy (Guan et al., 2022). However, it is noteworthy that parental genotypes are frequently missing, and thus, alternative approaches are sought. The family-based design offers a significant avenue for genetic association research. Nonetheless, estimating parameters for family data can be numerically intricate, and few alternative methods are available for family-based samples, apart from the linear mixed model, which is typically applied to quantitative phenotypes (Won et al., 2015).

Linear mixed-effects regression models play a central role in these analyses, assuming the family effect to be a random effect. Within this framework, the covariance structure for the random effect is assumed to align with that proposed by a polygenic model, encompassing the genetic relationships or kinship between every pair of individuals (Hsu et al., 2018).

- ***Haplotype-Based Methods -*** The utilization of haplotype models in examining candidate genes linked to complex diseases has proved to be effective in identifying significant haplotypes (Hauser et al., 2009). However, accurately estimating these underlying haplotypes is inherently challenging (Delaneau et al., 2013), necessitating the use of haplotype analyses, which are considered superior in representing parental chromosomes and encapsulating allele-specific biological functions (Hauser et al., 2009).

Genomics presents numerous challenges, which become even more prominent when dealing with a high number of homologous chromosomes, as is common in polyploid plants like sugar cane, with variable ploidy levels and a chromosome count ranging from 2n = 20 to approximately 200 (Premachandran et al., 2011). The complexity is further compounded by increased heterozygosity, indicating the frequent presence of variants among haplotypes (Moeinzadeh et al., 2020).

Jacobson et al. (2022) suggested that “haplotype-based” methods can serve as a viable alternative to alignment-based techniques in specific scenarios. These methods assess genetic similarity based on shared haplotypes, as opposed to nucleotide composition within a multiple sequence alignment (MSA). In haplotype-based

comparisons, selecting an appropriate distance metric is crucial, and there are multiple statistical options available (Jacobson et al., 2022; Li and Jiang, 2005).

Bu et al. (2021) presented a method for reconstructing haplotypes, useful in a haplotype-based association inference algorithm, to enhance the ability to identify target individuals in a group of cases compared with existing methods relying on SNVs. Moreover, the study highlights the uniqueness of Haplotag in providing genotypic inferences for tag-level (TL) haplotypes, also known as “Haplotag Loci.” These TL haplotypes may span multiple SNPs, and Haplotag generates an alternative set of genotypes based on the underlying SNP calls (GBS-SNPs). More recently, Chen et al. (2023) introduced tests based on haplotype frequencies rather than SNP frequencies, and illustrated their potential advantages. Therefore, when precise haplotype frequencies are unavailable, haplotype reconstruction methods offer the potential to infer these data by leveraging allele frequency information (Pelizzola et al., 2020).

3. REFERENCES

ABQM (2022) Notícias Available at: <https://abqm.com.br/noticias/30-bilhoes-e-o-valor-estimado-pela-esalqusp-para-o-pib-do-setor-na-atualizacao-de-estudo-sobre-o-setor-de-equideocultura> Accessed: 7th Jul. 2023.

Ablondi M, Viklund Å, Lindgren G et al. (2019) Signatures of selection in the genome of Swedish warmblood horses selected for sport performance. BMC Genomics 20:717 (2019). https://doi.org/10.1186/s12864-019-6079-1.

Araújo N (2013) Cavalos marchadores brasileiros - Globo Rural. Available at: <http://globotv.globo.com/rede-globo/globo-rural/>. Accessed: 25th Feb. 2023.

ASM (2019) Metagenomic Next Generation Sequencing: How Does It Work and Is It Coming to Your Clinical Microbiol Available at: < https://asm.org/> Accessed: 19th Jul. 2023.

Ardestani SS, Aminafshar M, Maryam MBZB et al. (2020) Signatures of selection analysis using whole-genome sequence data reveals novel candidate genes for pony and light horse types. *Genome* 63(8):387-396. https://doi.org/10.1139/gen-2020-0001.

Athanasopoulou K, Boti MA, Adamopoulos PG, Skourou PC, Scorilas A (2022) Third-Generation Sequencing: The Spearhead towards the Radical Transformation of Modern Genomics. Life *12*:30. doi: https://doi.org/10.3390/life12010030.

Azevêdo DM (2007) A pecuária de corte no Brasil: características morfológicas e produtivas da raça Nelore. Embrapa. Available at: <https://www.infoteca.cnptia.embrapa.br/infoteca/bitstream/doc/69545/1/APecuariaCorteBrasil-S-27-08.pdf> Accessed: 27th Jan. 2023.

Baba T, Morota G, Kawakami J, Gotoh Y el al. (2023) Longitudinal genome-wide association analysis using a single-step random regression model for height in Japanese Holstein cattle. JDS Communications 4:363-368.

Bahram S, Inoko H (2007) Microsatellite markers for genome-wide association studies. Nat Rev Genet 8:164. doi: https://doi.org/10.1038/nrg1962-c1.

Ballesta P, Bush D, Silva FF, Mora F (2020) Genomic Predictions Using Low-Density SNP Markers, Pedigree and GWAS Information: A Case Study with the Non-Model Species *Eucalyptus cladocalyx*. Plants 9(1):99. doi: 10.3390/plants9010099.

Banerjee S, Zeng L, Schunkert H, Söding J (2018) Bayesian multiple logistic regression for case-control GWAS. PLoS Genet 14(12): e1007856. https://doi.org/10.1371/journal.pgen.1007856.

Bein T, Karagiannidis C, Quintel M (2020) Climate change, global warming, and intensive care. Intensive Care Med. 46:485-487. doi: 10.1007/s00134-019-05888-4.

Bewick V, Cheek L, Ball J (2003) Statistics review 7: Correlation and regression. Crit Care 7(6):451-9. doi: 10.1186/cc2401.

Bu D, Wang X, Tang H (2021) Haplotype-based membership inference from summary genomic data. Bioinformatics 37:i161-i168. doi: 10.1093/bioinformatics/btab305.

Budhlakoti N, Kushwaha AK, Rai A, Chaturvedi KK et al. (2022) Genomic Selection: A Tool for Accelerating the Efficiency of Molecular Breeding for Development of Climate-Resilient Crops. Front Genet. 13:832153. doi: 10.3389/fgene.2022.832153.

Cantor RM, Lange K, Sinsheimer JS (2010) Prioritizing GWAS results: A review of statistical methods and recommendations for their application. Am J Hum Genet. 86(1):6-22. doi: 10.1016/j.ajhg.2009.11.017.

Chapman J, Clayton D (2007) Detecting association using epistatic information. Genet Epidemiol. 31:894-909.

Chatterjee N (2017) Case-Control Designs for Modern Genome-Wide Association Studies: Basic Principles and Overview. Book - Handbook of Statistical Methods for Case-Control Studies, Edition 1st Edition. eBook ISBN 9781315154084.

Ceballos FC, Joshi PK, Clark DW, Ramsay M, Wilson JF (2018) Runs of homozygosity: windows into population history and trait architecture. Nature reviews. Genetics, 19(4):220-234. doi: 10.1038/nrg.2017.109.

Chen H, Pelizzola M, Futschik A (2023) Haplotype based testing for a better understanding of the selective architecture. BMC Bioinformatics 24:322. doi:10.1186/s12859-023-05437-3.

Colpitts J, McLoughlin PD & Poissant J (2022) Runs of homozygosity in Sable Island feral horses reveal the genomic consequences of inbreeding and divergence from domestic breeds. BMC Genomics 23:501. https://doi.org/10.1186/s12864-022-08729-9.

Curi RA, Chardulo LAL, Giusti J, Silveira AC et al. (2010) Assessment of *GH1, CAPN1* and *CAST* polymorphisms as markers of carcass and meat traits in *Bos indicus* and *Bos taurus–Bos indicus* cross beef cattle. Meat Science 86:915-920.

Daetwyler HD, Villanueva B, Woolliams JA (2008) Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS ONE 3: e3395.

Daetwyler HD, Villanueva B, Woolliams JA (2008) Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach. PLoS ONE 3(10):e3395. Doi: 10.1371/journal.pone.0003395.

Delaneau O, Howie B, Cox AJ, Zagury JF, Marchini J (2013) Haplotype estimation using sequencing reads. Am J Hum Genet. 93(4):687-96. doi: 10.1016/j.ajhg.2013.09.002.

Drmanac R, Drmanac S, Strezoska Z, Paunesku T et al. (1993) DNA sequence determination by hybridization: a strategy for efficient large-scale sequencing. Science 260(5114):1649-52.

Ensminger ME (1990) Horses and Horsemanship. 6th ed Danville, I: Interstate Publishers p. 692.

Fallin D, Cohen A, Essioux L, ChumakovI et al. (2001) Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. Genome Res. 11:143-151.

FAO (2018) FAOSTAT. Available at: <http://www.fao.org/faostat/en/#home>. Accessed: 7th Jul. 2023.

Ferraz JBS, Eler JP (2007) Seleção de *Bos indicus* para precocidade sexual. Rev. Bras. Reprod. Anim 31:167-171.

Gibbs RA, Taylor JF, Van TCP (2009) Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. Science 324:528-532.

GLIPHA (2017) Global Livestock Production and Health Atlas Available at: <http://www.fao.org/ag/againfo/home/en/news_archive/AGA_in_action/glipha.html>. Accessed: 27th Jun. 2023.

Guan J, Nehzati SM, Benjamin DJ, Young AI (2022) Novel estimators for family-based genome-wide association studies increase power and robustness. bioRxiv doi: https://doi.org/10.1101/2022.10.24.513611.

Gumpinger AC, Roqueiro D, Grimm DG, Borgwardt KM (2018) Methods and Tools in Genome-wide Association Studies. Methods Mol Biol. 1819:93-136. doi: 10.1007/978-1-4939-8618-7_5.

Guo Q, Huang L, Jiang Y, Wang Z et al. (2022) Genome-Wide Association Study of Feed Efficiency Related Traits in Ducks. Animals 12(12):1532. doi: 10.3390/ani12121532.

Gurgul A, Jasielczuk I, Semik-Gurgul E, et al. (2019) A genome-wide scan for diversifying selection signatures in selected horse breeds. PLoS ONE 14:e0210751.

Guttmacher AE, Collins FS (2003) Welcome to the Genomic Era. New England Journal of Medicine 349:996-998. doi: https://www.nejm.org/doi/full/10.1056/NEJMe038132.

Hadlich JC, Curi RA, Silva MGB, Factori MA et al. (2013) Maciez da carne bovina e sua relação com o crescimento e os tipos de fibra musculares. Ciências Agrárias e Ambientais 11:421-430. doi: https://doi.org/10.7213/academico.011.004.AO09.

Hammond RK, Pahl MC, Su C, Cousminer DL et al. (2021) Biological constraints on GWAS SNPs at suggestive significance thresholds reveal additional BMI loci. Elife 10:e62206. doi: 10.7554/eLife.62206.

Hancock AM, Rienzo AD (2008) Detecting the Genetic Signature of Natural Selection in Human Populations: Models, Methods, and Data. Annu Rev Anthropol. 37:197-217. doi: 10.1146/annurev.anthro.37.081407.085141.

Hauser E, Cremer N, Hein R, Deshmukh H (2009) Haplotype-based analysis: a summary of GAW16 Group 4 analysis. Genet Epidemiol. 33:S24-8. doi: 10.1002/gepi.20468.
Hayes BJ, Lien S, Nielsen H, Olsen HG et al. (2006) The origin of selection signatures on bovine chromosome 6. Animal Genetics 39:101-105.

Hayward AC, Tollenaere R, et al. (2015) Molecular Marker Applications in Plants. In: Batley J (eds.) Plant Genotyping. Methods in Molecular Biology (Methods and Protocols) New York: Humana Press p. 1-312.

Howard JT, Pryce JE, Baes C, Maltecca C (2017) Invited review: Inbreeding in the genomics era: Inbreeding, inbreeding depression, and management of genomic variability. Journal of Dairy Science 100:6009-6024.

Hsu Y, Auerbach J, Zheng T, Lo SH. Coping with family structure in genome-wide association studies: a comparative evaluation. BMC Proc. 12:42. doi: 10.1186/s12919-018-0151-8.

Jacobson D, Zheng Y, Plucinski MM, Qvarnstrom Y, Barratt YLN (2022) Evaluation of various distance computation methods for construction of haplotype-based phylogenies from large MLST datasets. Molecular Phylogenetics and Evolution 177:107608.

Kumar S, Banks TW, Cloutier S (2012) SNP Discovery through Next-Generation Sequencing and Its Applications. Int J Plant Genomics 2012;2012:831460. doi: 10.1155/2012/831460.

Lander ES, Linton LM, Birren B, Nusbaum C et al. (2001) Initial sequencing and analysis of the human genome. Nature 409:860-921.

Lau M, Kress S, Schikowski T, Schwender H (2023) Efficient gene-environment interaction testing through bootstrap aggregating. Sci Rep. 13(1):937. doi: 10.1038/s41598-023-28172-4.

Lee S, Lozano A, Kambadur P, Xing EP (2016) An Efficient Nonlinear Regression Approach for Genome-wide Detection of Marginal and Interacting Genetic Variations. J Comput Biol. 23(5):372-89. doi: 10.1089/cmb.2015.0202.

Lee SH, Wray NR (2013) Correction: Novel Genetic Analysis for Case-Control Genome-Wide Association Studies: Quantification of Power and Genomic Prediction Accuracy. PLOS ONE 8(8). Doi: 10.1371/annotation/0b29c9c7-a86d-4e0f-bbb8-5c29b16e2884.

Li J and Jiang T (2005) Haplotype-based linkage disequilibrium mapping via direct data mining. Bioinformatics 21:4384-4393. doi:10.1093/bioinformatics/bti732.

Lin WY, Huang CC, Liu YL, Tsai SJ, Kuo PH (2019) Genome-Wide Gene-Environment Interaction Analysis Using Set-Based Association Tests. Front. Genet. 9:715. doi: 10.3389/fgene.2018.00715.

Li G, Tang J, Huang J, Jiang Y, Fan Y, Wang X, Ren J (2022) Genome-Wide Estimates of Runs of Homozygosity, Heterozygosity, and Genetic Load in Two Chinese Indigenous Goat Breeds. Front Genet. 13:774196. doi: 10.3389/fgene.2022.774196.

Lönker NS, Fechner K, Wahed AAE (2020) Horses as a Crucial Part of One Health. Vet Sci. 29:28. doi: 10.3390/vetsci7010028.

Maceachern S, Hayes B, MC Ewan J, Goddard M (2009) An examination of positive selection and changing effective population size in Angus and Holstein cattle populations (Bos taurus) using a high density SNP genotyping platform and the contribution of ancient polymorphism to genomic diversity in Domestic cattle. BMC Genomics 10:181.

Mancin E, Lourenco D, Bermann M, Mantovani R, Misztal I (2021) Accounting for Population Structure and Phenotypes from Relatives in Association Mapping for Farm Animals: A Simulation Study. Front. Genet. 12. doi: doi.org/10.3389/fgene.2021.642065.

Manolio Teri (2009) Case-control and cohort studies in the age of genome-wide associations. In Muin Khoury, and others (eds), Human Genome Epidemiology: Building the evidence for using genetic information to improve health and prevent disease. Oxford Academic p. 100–119. doi: doi.org/10.1093/acprof:oso/9780195398441.003.0006.

MAPA (2016) Estudo do Complexo do Agronegócio do Cavalo Available at: <http://www.agricultura.gov.br>. Accessed: 16th Jul. 2023.

Marcondes MI, Chizzotti ML, Filho SCV, Gionbelli MP et al. (2010) Energy requirements of Zebu beef cattle. In: Filho SCV, Marcondes MI, Chizzotti ML, Paulino PVR. Nutrient Requirements of Zebu Beef Cattle. Viçosa: Suprema Gráfica e Editora, p. 81-106.

Marigorta UM, Rodríguez JA, Gibson G, Navarro A (2018) Replicability and Prediction: Lessons and Challenges from GWAS. Trends Genet. 34(7):504-517. doi: 10.1016/j.tig.2018.03.005.

Mascher M, Wicker T, Jenkins J et al. (2021) Long-read sequence assembly: a technical evaluation in barley. The Plant Cell 33:1888-1906. doi: https://doi.org/10.1093/plcell/koab077.

Maxam AM, Gilbert W (1977) A new method for sequencing DNA. Proc Natl Acad Sci 74:560-564. doi:10.1073/pnas.74.2.560.

McCombie WR, McPherson JD (2019) Future Promises and Concerns of Ubiquitous Next-Generation Sequencing. Cold Spring Harb Perspect Med. 9(9):a025783. doi: 10.1101/cshperspect.a025783.

Mckay SD, Schnabel RD, Murdoch BM et al. (2008) An assessment of population structure in eight breeds of cattle using a whole genome SNP panel. BMC Genetics 9 doi: 10.1186/1471-2156-9-37.

Meslier V, Quinquis B, Da Silva K. et al. (2022) Benchmarking second and third-generation sequencing platforms for microbial metagenomics. Sci Data 9:694. doi: https://doi.org/10.1038/s41597-022-01762-z.

Miao J, Wu J, Lu Q (2003) Statistical methods for gene-environment interaction analysis. Wiley Interdisciplinary Reviews e1635. doi: 10.1002/wics.1635.

Miller RL and Ho SM (2008) Environmental epigenetics and asthma: current concepts and call for studies. Am J Resp Crit Care Med. 177:567–73.

Moeinzadeh M, Yang J, Muzychenko E, Gallone G, Heller D, Reinert K, et al. (2020) Ranbow: A fast and accurate method for polyploid haplotype reconstruction. PLoS Comput Biol 16(5):e1007843. doi: doi.org/10.1371/journal.pcbi.1007843.

Moldován N, Tombácz D, Szűcs A et al. (2018) Third-generation Sequencing Reveals Extensive Polycistronism and Transcriptional Overlapping in a Baculovirus. Sci Rep 8:8604. doi: https://doi.org/10.1038/s41598-018-26955-8.

Mondini L, Noorani A, Pagnotta MA (2009) Assessing Plant Genetic Diversity by Molecular Tools. Diversity 1:19-35. doi: https://doi.org/10.3390/d1010019.

Mozzi A, Pontremoli C, Sironi M (2018) Genetic susceptibility to infectious diseases: Current status and future perspectives from genome-wide approaches. Infect Genet Evol. 66:286-307. doi: 10.1016/j.meegid.2017.09.028.

Mousavi SF, Razmkabir, M, Rostamzadeh J. et al. (2023) Genetic diversity and signatures of selection in four indigenous horse breeds of Iran. Heredity 131:96–108. https://doi.org/10.1038/s41437-023-00624-7

Mukaka MM (2012) Statistics corner: A guide to appropriate use of correlation coefficient in medical research. Malawi Med J. 24(3):69-71.

Muro EM, McCann JA, Rudnicki MA, Andrade-Navarro MA (2009) Use of SNP-arrays for ChIP assays: computational aspects. Methods Mol Biol. 567:145-54. doi: 10.1007/978-1-60327-414-2_10.

NIH (2007) Understanding Human Genetic Variation. Available at: <https://www.ncbi.nlm.nih.gov/books/NBK20363/> Accessed: 23th Aug. 2023.
NIH (2022) Matérias de pesquisa do NIH Available at: <https://www.nih.gov> Accessed: 23th Jul. 2023.

NUALPEC (2015) Anuário da Pecuária Brasileira 21 ed. Instituto FNP, 81-82 p.
Oliveira JHF, Magnabosco CU, Borges AMSM (2022) Nelore: Base Genética e Evolução Seletiva no Brasil. Available at:

<https://www.infoteca.cnptia.embrapa.br/bitstream/doc/566499/1/doc49.pdf> Accessed: 27th Jan. 2023.

Oliveira MC, Alencar MM, Giglioti R, Beraldo MC et al. (2013) Resistance of beef cattle of two genetic groups to ectoparasites and gastrointestinal nematodes in the state of São Paulo. Vet Parasitol. 197:168-75. doi: 10.1016/j.vetpar.2013.06.021.

Ottman R (1996) Gene-environment interaction: definitions and study designs. Prev Med. 25(6):764-70. doi: 10.1006/pmed.1996.0117.

Ozsolak F (2012) Third-generation sequencing techniques and applications to drug discovery. Expert Opin Drug Discov. 7(3):231-43. doi: 10.1517/17460441.2012.660145.

Pelizzola M, Behr M, Li H, Munk A, Futschik A (2021) Multiple haplotype reconstruction from allele frequency data. Nat Comput Sci. 1:1-10. doi: 10.1038/s43588-021-00056-5.

Peñaloza C, Robledo D, Barría A, Trịnh TQ et al. (2020) Development and Validation of an Open Access SNP Array for Nile Tilapia (Oreochromis niloticus). G3 10(8):2777-2785. doi: 10.1534/g3.120.401343.

Perez M, Aroh O, Sun Y, Lan Y et al. (2023) Third-Generation Sequencing Reveals the Adaptive Role of the Epigenome in Three Deep-Sea Polychaetes. Mol Biol Evol. 40(8):msad172. doi: 10.1093/molbev/msad172.

Platten JD, Cobb JN, Zantua RE (2019) Criteria for evaluating molecular markers: Comprehensive quality metrics to improve marker-assisted selection. PLoS One 14(1):e0210529. doi: 10.1371/journal.pone.0210529.

Premachandran MN, Prathima PT, Lekshmi M (2011) SUGARCANE AND POLYPLOIDY – AREVIEW. Journal of Sugarcane Research 1(2):1-15.

Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. Nat Rev Genet. 11(7):459-63. doi: 10.1038/nrg2813.

Raj RS, Das ND (2021) Molecular markers and its application in animal breeding. Advances in Animal Genomics 123-140. doi: https://doi.org/10.1016/B978-0-12-820595-2.00009-6.

Ruff TM, Marston EJ, Eagle JD, Sthapit SR et al. (2020) Genotyping by multiplexed sequencing (GMS): A customizable platform for genomic selection. PLoS One 15(5):e0229207. doi: 10.1371/journal.pone.0229207.

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

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America. 74:5463-7.

Santos WB, Schettini GP, Maiorano AM et al. (2021a) Genome-wide scans for signatures of selection in Mangalarga Marchador horses using high-throughput SNP genotyping. BMC Genomics 22:737. doi: 10.1186/s12864-021-08053-8.

Santos WB, Schettini GP et al. (2021b) Fine-scale estimation of inbreeding rates, runs of homozygosity and genome-wide heterozygosity levels in the Mangalarga Marchador horse breed. J Anim Breed Genet. 138(2):161-173. doi: 10.1111/jbg.12508.

Santos WB, Pereira CB, Maiorano AM, Arce CDS, Baldassini WA, Pereira GL, Chardulo LAL, Neto ORM, Oliveira HN, Curi RA (2023) Genomic inbreeding estimation, runs of homozygosity, and heterozygosity-enriched regions uncover signals of selection in the Quarter Horse racing line. J Anim Breed Genet. 140(6):583-595. doi: 10.1111/jbg.12812.

Shaffer JR, Feingold E, Marazita ML (2012) Genome-wide association studies: prospects and challenges for oral health. J Dent Res. 91(7):637-41. doi: 10.1177/0022034512446968.

Sharma P (2023) Different Types of Regression Models Available at: <https://www.analyticsvidhya.com/blog/2022/01/different-types-of-regression-models/> Accessed: 9th Oct. 2023.

Sheynkman GM, Shortreed MR, Cesnik AJ, Smith LM (2016) Proteogenomics: Integrating Next-Generation Sequencing and Mass Spectrometry to Characterize Human Proteomic Variation. Annu Rev Anal Chem 12(1):521-45. doi: 10.1146/annurev-anchem-071015-041722.

Silveira MV, Souza JC, Filho PBF, Silva LOC et al. (2019) Adaptability and phenotypic stability for productive and reproductive traits in Nelore cattle. Ciência Rural 49:3. doi http://dx.doi.org/10.1590/0103-8478cr20180327.

Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW (2013) Pleiotropy in complex traits: challenges and strategies. Nat Rev Genet. 14(7):483-95. doi: 10.1038/nrg3461.

Southern EM, Maskos U, Elder JK (1992). Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models. Genomics 13(4):1008-17.

Sul JH, Martin LS, Eskin E (2018) Population structure in genetic studies: Confounding factors and mixed models. PLoS Genet. 14(12):e1007309. doi: 10.1371/journal.pgen.1007309.

Sun J, Xiao J, Jiang Y, Wang Y et al. (2023) Genome-Wide Association Study on Reproductive Traits Using Imputation-Based Whole-Genome Sequence Data in Yorkshire Pigs. Genes 14(4):861. doi: 10.3390/genes14040861.

Szmatoła T, Gurgul A, Jasielczuk I, Oclon E, Ropka-Molik K, Stefaniuk-Szmukier M, Polak G, Tomczyk-Wrona I, Bugno-Poniewierska M (2022) Assessment and Distribution of Runs of Homozygosity in Horse Breeds Representing Different Utility Types. Animals (Basel). 12(23):3293. doi: 10.3390/ani12233293.

Telles AE, Moço CMN (2022) Possíveis efeitos da equoterapia como auxílio na reabilitação psicológica de crianças com transtorno do espectro autista (TEA). Revista Ibero-Americana de Humanidades, Ciências e Educação. 8:10.

Thomas D (2010) Gene-environment-wide association studies: Emerging approaches. Nature Reviews Genetics 11(4):259–272.

Thomas D (2010) Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. Annu Rev Public Health 31:21-36. doi: 10.1146/annurev.publhealth.012809.103619.

Uffelmann E, Huang QQ, Munung NS et al. Genome-wide association studies. Nat Rev Methods Primers 1:59. https://doi.org/10.1038/s43586-021-00056-.

Vallejos-Vidal E, Reyes-Cerpa S, Rivas-Pardo JA, Maisey K et al. (2020) Single-Nucleotide Polymorphisms (SNP) Mining and Their Effect on the Tridimensional Protein Structure Prediction in a Set of Immunity-Related Expressed Sequence Tags (EST) in Atlantic Salmon (*Salmo salar*). Front Genet. 27:10-1406. doi: 10.3389/fgene.2019.01406.

Van ME, Visser C (2021) Unintended consequences of selection for increased production on the health and welfare of livestock. Arch Anim Breed. 64(1):177-185. doi: 10.5194/aab-64-177-2021.

Varathan P, Gorijala P, Jacobson T et al. (2022) Integrative analysis of eQTL and GWAS summary statistics reveals transcriptomic alteration in Alzheimer brains. BMC Med Genomics 15:93. doi: https://doi.org/10.1186/s12920-022-01245-5.

Venter JC, Adams MD, Myers EW et al. (2001) The sequence of the hu- man genome. Science 291:1304-51.

Verlouw JAM, Clemens E, Vries JH et al. (2021) A comparison of genotyping arrays. Eur J Hum Genet 29:1611–1624. doi: https://doi.org/10.1038/s41431-021-00917-7.

Wang T, Xue X, Xie X, Ye K, Zhu X, Elston RC (2018) Adjustment for covariates using summary statistics of genome-wide association studies. Genet Epidemiol. 42(8):812-825. doi: 10.1002/gepi.22148.

Westerman KE, Pham DT, Hong L, Chen Y, Sevilla-González M et al. (2021) Scalable and flexible gene-environment interaction analysis in millions of samples. Bioinformatics 37:3514–3520. doi: 10.1093/bioinformatics/btab223.

Williams JL, Stephen JGH, Marcello DC, Ballingall KT et al. (2016) Inbreeding and purging at the genomic Level: the Chillingham cattle reveal extensive, non-random SNP heterozygosity. Animal genetics 47:19-27.

Won S, Wilk JB, Mathias RA, O'Donnell CJ, Silverman EK, Barnes K, et al. (2009) On the Analysis of Genome-Wide Association Studies in Family-Based Designs: A Universal, Robust Analysis Approach and an Application to Four Genome-Wide Association Studies. PLoS Genet 5(11):e1000741. https://doi.org/10.1371/journal.pgen.1000741.

Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X (2010) Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet. 86(6):929-42. doi: 10.1016/j.ajhg.2010.05.002.

Yang J, Wray NR, Visscher PM (2010) Comparing apples and oranges: equating the power of case-control and quantitative trait association studies. Genet Epidemiol. 34(3):254-7. doi: 10.1002/gepi.20456.

Zeng P, Zhao Y, Qian C, Zhang L, Zhang R, Gou J, Liu J, Liu L, Chen F (2015) Statistical analysis for genome-wide association study. J Biomed Res. 2015 Jul;29(4):285-97. doi: 10.7555/JBR.29.20140007.

Zenger KR, Khatkar MS, Cavanagh JAL et al. (2006) Genome-wide genetic diversity of Holstein Friesian cattle reveals new insights into Australian and global population variability, including impact of selection. Animal Genetics 38:7-14.

Zepeda-Batista JL, Núñez-Domínguez R, Ramírez-Valverde R, Jahuey-Martínez FJ, Herrera-Ojeda JB, Parra-Bracamonte GM (2021) Discovering of Genomic Variations Associated to Growth Traits by GWAS in Braunvieh Cattle. Genes 12(11):1666. doi: 10.3390/genes12111666.

CHAPTER II – Genomic inbreeding estimation, runs of homozygosity, and heterozygosity-enriched regions uncover signals of selection in the Quarter Horse racing line [a]

Abstract - With the advent of genomics, significant progress has been made in the genetic improvement of livestock species, particularly through increased accuracy in predicting breeding values for selecting superior animals and the possibility of performing a high-resolution genetic scan throughout the genome of an individual. The main objectives of this study were to estimate the individual genomic inbreeding coefficient based on runs of homozygosity (F_{ROH}), to identify and characterize runs of homozygosity and heterozygosity (ROH and ROHet, respectively; length and distribution) throughout the genome, and to map selection signatures in relevant chromosomal regions in the Quarter Horse racing line. A total of 336 animals registered with the Brazilian Association of Quarter Horse Breeders (ABQM) were genotyped. One hundred and twelve animals were genotyped using the Equine SNP50 BeadChip (Illumina, USA), with 54,602 single nucleotide polymorphisms (SNPs) (54K). The remaining 224 samples were genotyped using the Equine SNP70 BeadChip (Illumina, USA) with 65,157 SNPs (65K). To ensure data quality, we excluded animals with a call rate below 0.9. We also excluded SNPs located on non-autosomal chromosomes, as well as those with a call rate below 0.9 or a p-value below 1×10^{-5} for Hardy-Weinberg equilibrium. The results indicate moderate to high genomic inbreeding, with 46,594 ROH and 16,101 ROHet detected. In total, 30 and 14 candidate genes overlap with ROH and ROHet regions, respectively. The ROH islands showed genes linked to crucial biological processes, such as cell differentiation (*CTBP1*, *WNT5B*, and *TMEM120B*), regulation of glucose metabolic process (*MAEA* and *NKX1-1*), heme transport (*PGRMC2*), and negative regulation of calcium ion import (*VDAC1*). In ROHet, the islands showed genes related to respiratory capacity (*OR7D19*, *OR7D4G*, *OR7D4E*, and *OR7D4J*) and muscle repair (*EGFR* and *BCL9)*. These findings could aid in selecting animals with greater regenerative capacity and developing treatments for muscle disorders in the QH breed. This study serves as a foundation for future research on equine breeds. It can contribute to developing reproductive strategies in animal breeding programs to improve and preserve the Quarter Horse breed.

Keywords: autozygosity*, Equus caballus*, SNP array, signatures of selection

[a]*Article published in the "The Journal of Animal Breeding and Genetics" Journal: https://doi.org/10.1111/jbg.12812.*

1. INTRODUCTION

Selection based on different criteria in the Quarter Horse (QH) breed has resulted in the development of lineages, including the racing line (Evans 1996). According to Meira et al. (2013), the QH breed consists of different lineages with distinct abilities that result from varying selection objectives. The cutting line is designed to test functional agility and obedience, which are crucial for managing cattle in the field. Conversely, the racing lineage is characterized by remarkable speed over short distances on straight tracks. Horses belonging to the QH racing line exhibit superior performance in runs over short distances compared to other lineages (ABQM, 2019). Furthermore, high costs are incurred in the maintenance of competitive horses in this sport, as they require special care, medication, and veterinary procedures (Lima et al. 2006).

Genetic improvement strategies are crucial to enhance production system outcomes. As the same animals are repeatedly used for reproduction, an increase in individual and population inbreeding coefficients is expected over generations, making it essential to monitor population parameters. A high rate of inbreeding reduces genetic variability within populations, resulting in a decline in performance through inbreeding depression (Keller et al. 2011). Additionally, inbred individuals are more likely to exhibit recessive diseases and anomalies in their offspring due to the combination of rare deleterious alleles present in homozygosity (Szpiech et al. 2013).

The classical and most commonly used method to evaluate inbreeding is through the use of pedigree information (F_{PED}). However, estimating the inbreeding coefficient based solely on pedigree information has several flaws. These include biases introduced in inbreeding depression estimates due to the variance properties of F_{PED}, pedigree errors, and incomplete pedigrees, among other factors. Curik et al. (2017) provides an elegant review of these issues. Advancements in molecular tools have provided more accurate ways to estimate the proportions of autozygosity in an individual’s genome without relying solely on pedigree information. One such method is based on homozygosity runs (ROH) (Keller et al. 2011). ROH is a genomic indicator of autozygosity, inherited when an individual has contiguous stretches of homozygous

genotypes/haplotypes from both parents that are identical by descent (IBD) from a common ancestor. It reflects the degree of relatedness in a population. (Gibson et al. 2006). The estimated coefficient of an individual's genomic inbreeding is abbreviated as F_{ROH}.

Changes resulting from recent positive selection along the genome can be detected by analyzing signatures of selection (Pemberton et al. 2012; Zhang et al. 2015). Additionally, signatures of selection analysis enable the identification of a reduction, elimination, or change in genetic variation in genomic regions adjacent to the causative variants in response to natural and/or artificial selection. Therefore, searching for signatures of selection provides an opportunity to reveal chromosomal segments that control traits important for a specific population, species, or breed/lineage (Bertolini et al. 2018; Meira et al. 2014). Because most individuals within a population share ROH islands, this method has been extensively used to identify selection signatures (Zhang et al. 2015).

The genomic distribution of heterozygous loci and the correlation between regions of heterozygosity, also known as runs of heterozygosity (ROHet), have not been sufficiently explored in economically important populations (Ruan et al. 2022). Genetic diversity in populations is typically assessed by using neutral molecular markers, which measure both expected heterozygosity and the number of alleles per locus (López-Cortegano et al. 2019). According to López-Cortegano et al. (2019), these two measures complement each other: While the former is directly linked to genetic variance for quantitative traits (the short-term response to selection and adaptation), the latter is more sensitive to population bottlenecks (the long-term capacity of populations to adapt).

Heterozygosity is an important measure of genetic diversity that indicates the amount of genetic variation present within populations (Harris & DeGiorgio 2017). These regions can be found in different parts of the genome, including protein-coding and non-coding regions (Zappala & Montgomery, 2017; Zheng et al. 2013), and are the result of several processes such as mutation, recombination, genetic drift, gene flow, selection, and other genetic events (Llaurens et al. 2017; Sakamoto & Innan 2019; Santos et al.

2021a). Regions rich in heterozygosity are important for population genetics and biodiversity conservation because they can provide valuable information about genomic regions that undergo gene introgression or admixture (Marras et al. 2017; Samuels et al. 2016), offering insights into the evolutionary history and genetic diversity of organisms. Regions rich in heterozygosity have the potential to uncover population demographic events, such as evolutionary aspects, and issues related to size, composition, bottlenecks, metapopulation dynamics, genetic variability, and isolate-breaking effect (Llaurens et al. 2017; Santos et al. 2021a). These regions, which are commonly found at a high frequency in the population, can also be considered signatures of selection.

The objectives of this study were to estimate the genomic inbreeding coefficient based on ROH, to identify and characterize ROH and ROHet throughout the genome by exploring their length and distribution, and to map the main signatures of selection for ROH and ROHet islands in chromosomal regions identifying potential candidate genes and pathways relevant to the Quarter Horse racing line.

2. MATERIALS AND METHODS

2.1 Data availability statement

The data that support the findings of this study are available upon request (email rogerio.curi@unesp.br). The data are not publicly available due to privacy or ethical restrictions.

2.2 Ethical statement

All experimental procedures involving horses in this study were performed by following the relevant guidelines of animal welfare. The project was approved by the Ethics Committee on Animal Use of the College of Veterinary and Animal Science (FMVZ), UNESP, Botucatu/SP (Approval No. 110/2016-CEUA).

2.3 Competing interests

The author(s) certify that they have no conflict of interest.

2.4 Animals

A total of 336 horses of the QH racing line registered with the Brazilian Association of Breeders (ABQM) were used in the study. The animals were raised at 25 properties and Jockey Clubs in the state of São Paulo. There were 70 males and 266 females, the offspring of 83 stallions and 249 mares, resulting in an average of 4.1 progeny/stallion and 1.3 progeny/mare and belonging to 156 breeders. The samples were chosen to better represent the genetic diversity in the QH racing line of the state of São Paulo and, consequently, Brazil. Animals with many offspring and descendants of influential individuals in the line formation were included in the dataset. Closely related individuals such as full sibs were avoided in the sampling.

2.5 Blood collection and DNA extraction

A 5-mL blood sample was collected from each animal by venipuncture of the jugular neck region, and the material was stored in vacuum tubes containing 7.5 mg of ethylenediaminetetraacetic acid (EDTA). DNA was extracted with the Illustra Blood GenomicPrep Mini Spin Kit (GE Healthcare, USA), following the manufacturer's instructions. At the end of the extraction procedure, the DNA was quantified using the Qubit® 3.0 Fluorometer (Invitrogen, USA) and analyzed for purity (260/280 and 260/230 ratios) and integrity with the NanoDrop Lite small-volume spectrophotometer (Thermo Scientific, USA) and a 0.8% agarose gel, respectively. After dilution to the concentration required for single nucleotide polymorphism (SNP) genotyping (approximately 50 ng/uL), DNA samples were stored in 1.7-mL mini-tubes at -20°C until genotyping.

2.6 Genotyping, quality control, and genotype imputation

One hundred and twelve animals were genotyped using the Equine SNP50 BeadChip (Illumina, USA) with 54,602 SNPs (54K). The remaining 224 samples were genotyped with the Equine SNP70 BeadChip (Illumina) with 65,157 SNPs (65K). Both chips were scanned and analyzed using the HiScan system with standard settings (Illumina).

The genotype imputation procedure was performed using the FImpute v.2.2 program (Sargolzaei et al. 2014). Imputation was carried out in two steps, from the 65k to the 54k chip and vice versa, because the dataset did not contain animals genotyped with both panels. The accuracy of imputation was evaluated using the concordance rate (CR) and allelic r^2. The accuracy of imputation was estimated through simulations involving different scenarios, with each panel being used as the reference and the imputed set in turn. The accuracy values were obtained per SNP and per sample. To obtain additional information on genotype imputation and accuracy assessment, Pereira et al. (2017) had previously published a comprehensive analysis with detailed statistics for this dataset.

An additional quality control of genotypes was verified at the individual and SNP levels using the snpStats R package (Clayton 2015). To ensure data quality, we excluded animals with a call rate below 0.9. We also excluded SNPs located on non-autosomal chromosomes, as well as those with a call rate below 0.9 or a P-value below 1×10^{-5} for Hardy-Weinberg equilibrium. The final genotyping file contained information from 55,130 SNPs located on the 31 autosomal chromosomes.

2.7 Population structure

The population structure in the QH racing line was assessed with principal component analysis (PCA) and discriminant analysis of principal components (DAPC) using the Adegenet package in the R software (Jombart et al. 2010). Additional details can be found in the publication by Marchiori et al. (2019); they used a portion of the same database for two QH lineages.

2.8 Genomic inbreeding

The genomic inbreeding (F_{ROH}) coefficient were estimated as the sum of the individual's ROH lengths, divided by the total length of the autosome chromosomes (in number of SNPs), according to McQuillan et al. (2008):

$$F_{ROH} \frac{\sum_k Length\,(ROH_k)}{L},$$

where: k is the number of runs of ROH of each individual multiplied by the average length ROH; and L is the total genome length. The population inbreeding was estimated by using the average of all individuals belonging to the representative sample of the population.

2.9 Characterization of ROH and ROHet

To calculate the ROH and ROHet segments for each individual, we utilized the detectRUNS R package (Biscarini et al. 2018), applying the windows-free consecutive runs method for both analyses. This technique involves scanning the genome SNP by SNP, as described by Marras et al. (2015). The parameters in the detectRUNS package (https://cran.r-project.org/web/packages/detectRUNS/detectRUNS.pdf) were:

(a) ROH – ROHet = FALSE; minSNP = 40; maxGap = 10^6; minLengthBps = 250 kb; maxOppRun = 1; and maxMissRun = 1; and (b) ROHet – ROHet = TRUE; minSNP = 20; maxGap = 10^6; minLengthBps = 10 Kb; maxOppRun = 2; and maxMissRun = 1. The ROH and ROHet segments were classified into five classes according to length: 1–2, 2–4, 4–8, 8–16, and > 16 mega base pairs (Mb).

2.10 Identification of islands of homozygosity and heterozygosity

Shared segments with frequencies ≥ 50% for ROH and ≥ 40% for ROHet were considered signals of positive selection. Gene prospecting was performed with coordinate remapping for the EquCab3.0 version (NCBI; https://www.ncbi.nlm.nih.gov/genome/tools/remap) and gene annotation was performed with the use of BioMart R package (Smedley et al. 2009).

The focus was on the most representative regions within the ROH regions based on frequencies, which included the top three SNPs with a high frequency of ROH occurrence. This analysis considered only the top ROH island. Annotation windows for the top SNPs were amplified by 125 kb for both directions in the EquCab3.0 coordinates.

Enrichment analysis was conducted on the candidate genes using an over-representation test ($p < 0.05$) with the clusterProfiler package in R (Yu et al., 2012). This

test comprises three orthogonal Gene Ontology (GO) terms: molecular function, biological process, and cellular component. In addition, network analysis served as a complementary approach to visualize and predict the function of gene sets (https://genemania.org).

3. RESULTS

3.1 Population structure

The dataset presented subpopulations that are the result of the formation of families descended from important sires (Figure 1). Marchiori et al. (2019) analyzed the same dataset, including the QH cutting line. The same substructures reported by Marchiori et al. (2019) were identified in the racing line because the two studies shared data. In this sense, Marchiori et. al (2019) provided detail on the genetic structure observerd in the QH racing line.

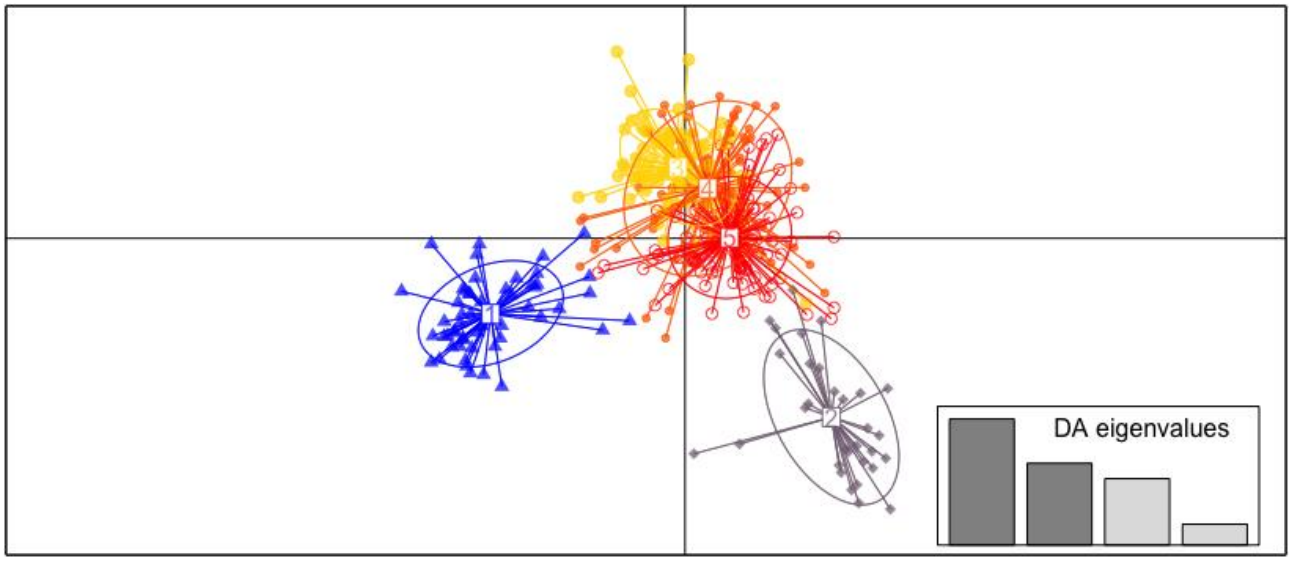


FIGURE 1 Discriminant analysis (DA) of principal components on the genomic data in the Quarter Horse racing line.

3.2 Genomic inbreeding coefficient

Individual estimates of the genomic inbreeding coefficient (F_{ROH}) ranged from 0.10 to 0.35, with an average of 0.19 for the entire genome (Figure 2). Regarding the classification by chromosomes: ECA 18 (0.26), ECA 4 (0.26), and ECA 1 (0.23) had the highest average F_{ROH}, while ECA 16 (0.15) had the lowest average F_{ROH}.

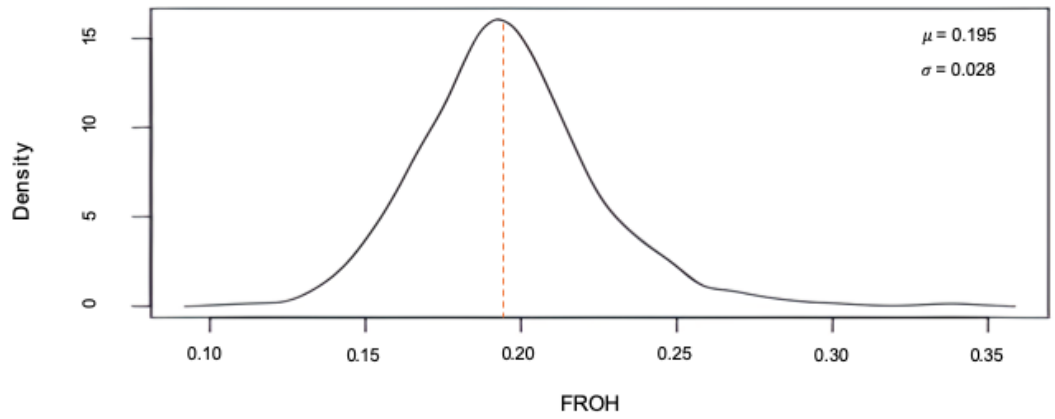


FIGURE 2 Distribution of genomic inbreeding coefficient in the Quarter Horse racing line genome, with the mean indicated by the dashed orange line.

3.3 Characterization of ROH and ROHet

In the evaluated individuals, there were 46,594 ROH and 16,101 ROHet (Table 1). Five classes (1–2, 2–4, 4–8, 8–16, and >16 Mbp) were formed based on the run lengths for the two approaches.

The ROH class with the highest percentage of observations was 1–2 Mbp with 50.71%, and the ROH class with the lowest percentage was > 16 Mbp with 1.79%. In the 1–2 Mbp class, there were 23,630 ROH with an average size of 1.41 Mbp. On the other hand, in the > 16 Mbp class, the average size for the 837 ROH was 23.12 Mbp. Overall, the representative QH racing line samples carried a greater proportion of smaller segments in autozygosity in the genome when analyzed with a high-density SNP panel.

The ECA 1 had the most ROH per chromosome, followed by ECA 2 and 3. ECA 31 had the lowest number of ROH. Thus, there was a trend toward a higher number of ROH in the larger chromosomes and vice versa. The opposite happened with ROHet analysis: Most ROHet were found in ECA 1, 4, and 18. ECA 29 had the fewest ROHet runs (11). Furthermore, there were no ROHet with lengths of > 2–4 Mbp (Table 1).

TABLE 1 Number, percentage, and mean for runs of homozygosity (ROH) and runs of heterozygosity (ROHet) per chromosome and length classes.

	ROH			ROHet		
Per chromosome						
	No of ROH	Percentage	Mean (Mbp)	No of ROHet	Percentage	Mean (Mbp)
ECA 1	3872	0.0831	3.94	1124	0.0698	0.87
ECA 2	2632	0.0564	2.74	643	0.0399	0.74
ECA 3	2439	0.0523	3.18	747	0.0464	0.76
ECA 4	2427	0.0520	4.05	2153	0.1337	1.13
ECA 5	1747	0.0374	3.72	480	0.0298	0.91
ECA 6	1955	0.0419	3.03	715	0.0444	0.78
ECA 7	2379	0.0510	3.15	898	0.0558	1.02
ECA 8	2011	0.0431	3.35	780	0.0484	0.86
ECA 9	1705	0.0365	3.48	639	0.0397	1.02
ECA 10	1459	0.0313	3.84	993	0.0617	0.90
ECA 11	1330	0.0285	3.27	541	0.0336	0.85
ECA 12	517	0.0110	2.77	109	0.0068	0.90
ECA 13	699	0.0150	3.10	116	0.0072	0.86
ECA 14	2106	0.0451	3.32	487	0.0302	0.85
ECA 15	1697	0.0364	3.30	557	0.0346	0.80
ECA 16	1682	0.0360	2.70	556	0.0345	0.78
ECA 17	1658	0.0355	3.11	519	0.0322	0.91
ECA 18	1708	0.0366	4.43	1027	0.0638	0.71
ECA 19	1435	0.0307	2.50	378	0.0235	0.68
ECA 20	1272	0.0272	3.10	488	0.0303	0.78
ECA 21	1256	0.0269	2.94	112	0.0070	0.79
ECA 22	1036	0.0222	3.21	378	0.0235	0.84
ECA 23	1245	0.0267	2.81	259	0.0161	0.80
ECA 24	875	0.0187	2.98	144	0.0089	0.83
ECA 25	891	0.0191	3.38	462	0.0287	0.74
ECA 26	1089	0.0233	2.49	148	0.0092	0.77
ECA 27	775	0.0166	3.05	257	0.0160	0.91
ECA 28	892	0.0191	2.80	96	0.0060	0.79
ECA 29	607	0.0130	2.66	11	0.0007	1.17
ECA 30	574	0.0123	3.09	142	0.0088	0.78
ECA 31	624	0.0133	2.27	142	0.0088	0.72
Total	46,594	1	97.76	16,101	1	26.25

Per classes						
1-2	23630	0.5071	1.42	15831	0.9832	0.84
2-4	13062	0.2803	2.78	270	0.0168	2.90
4-8	6587	0.1414	5.50	-	-	-
8-16	2478	0.0532	10.84	-	-	-
>16	837	0.0180	23.12	-	-	-
Total	46,594	1	43.66	16,101	1	3.74

3.4 Identification of ROH and ROHet islands

Altogether, there were 34 representative regions in ROH islands with frequencies > 50% in the sampled individuals. They were in ECA 1, 2, 3, 4, 6, 7, 8, 14,17, 18, and 23 (Figure 3). For ROHet, seven regions presented frequencies of > 40% of the individuals. They were in ECA 4, 5, and 7; however, there were other heterozygosity-rich regions with frequencies of > 30% in ECA 8, 10, 11, 20, and 22 (Figure 4).

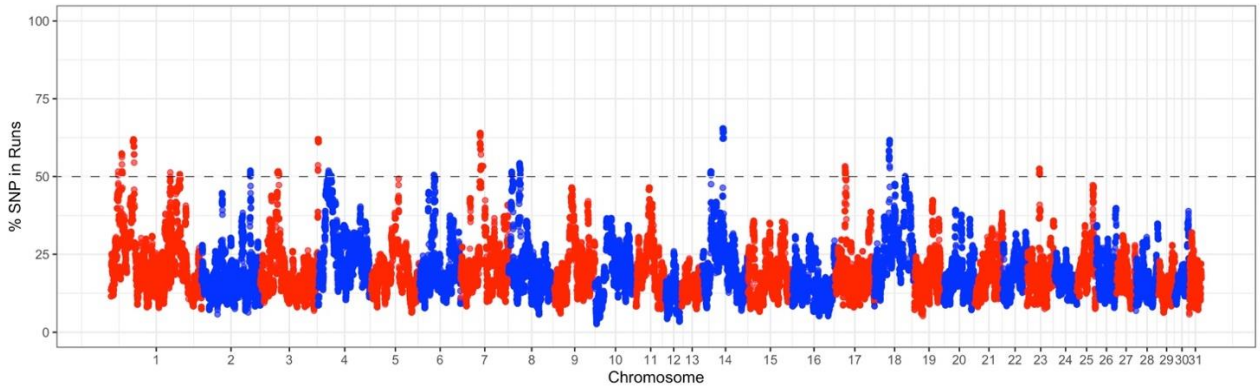


FIGURE 3 Distribution of runs of homozygosity (ROH) islands per chromosome in the Quarter Horse racing line genome, with the dashed line marking the 50% frequency.

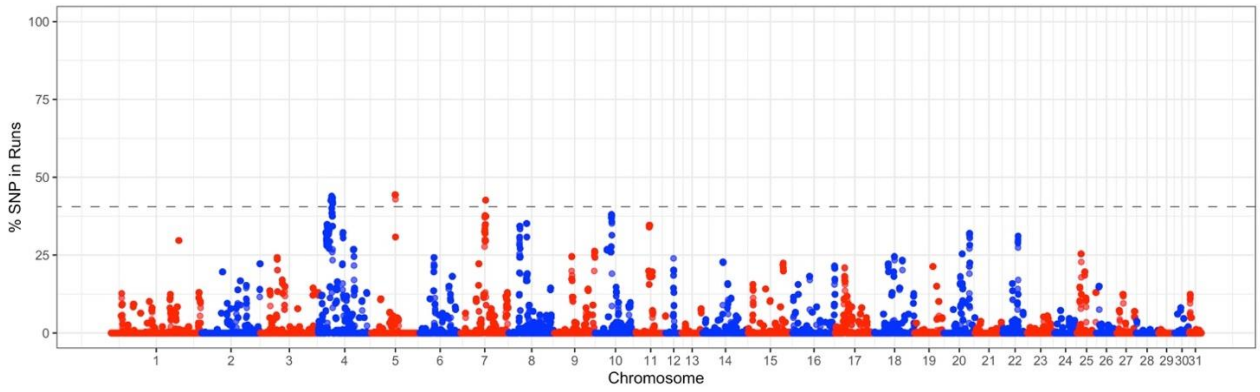


FIGURE 4 Distribution of runs of heterozygosity (ROHet) islands per chromosome in the Quarter Horse racing line genome, with the dashed line marking the 40% frequency.

3.5 Signatures of selection

The top three SNPs in the runs that showed the highest frequency on each chromosome were considered signatures of selection. Each candidate gene with its chromosome, start and end nucleotide positions, and Ensembl description for ROH and ROHet results is presented in Supplementary Material 1. The ROH gene annotation list includes 48 Ensembl gene IDs, but only 30 have known functions. They are in ECA 2, 3, 4, 6, 7, 8, 14, 17, 18, and 23 (Table S1). The ROHet gene list includes 20 Ensembl gene IDs, 14 of which have known functions and are distributed in ECA 4, 5, and 7 (Table S2). Ensembl gene IDs without known functions were removed from the list and were not explored further.

In total, there were 44 candidate genes within the signatures of selection identified in ROH and ROH analyses. They are in linkage disequilibrium, linked, or even causal of the phenotypic changes that possibly provide performance advantages for an athlete horse. Functional enrichment was performed individually for each gene set (ROH and ROHet). This analysis yielded 10 main biological processes (Figure 5), which are explicated in the Discussion. Details on other significant GO terms for molecular functions, biological processes, and cellular components are shown in Supplementary Material 2.

The network analysis revealed that *EGFR* has a high physical interaction percentage (77.64%) with a gene set that is associated with *BCL9*, *LANCL2*, and *VOPP1*, in addition to other interactions depicted in Figure 6.

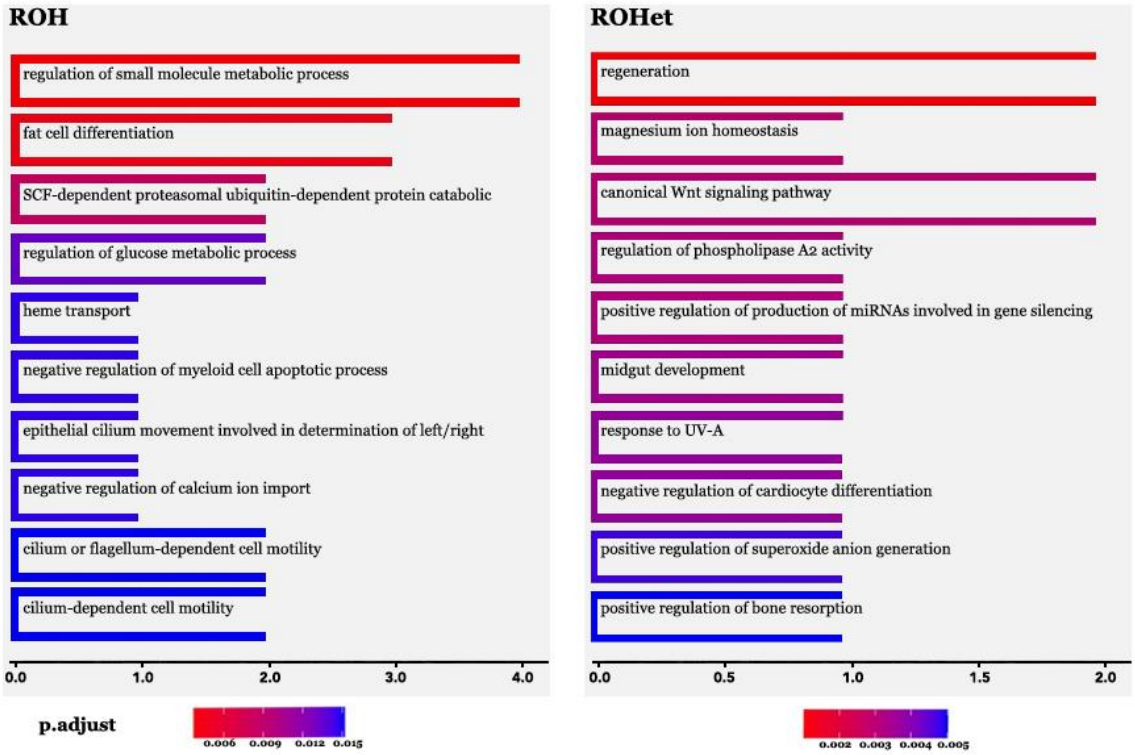


FIGURE 5 Functional annotation for the top ten significant biological functions ($p < 0.01$).

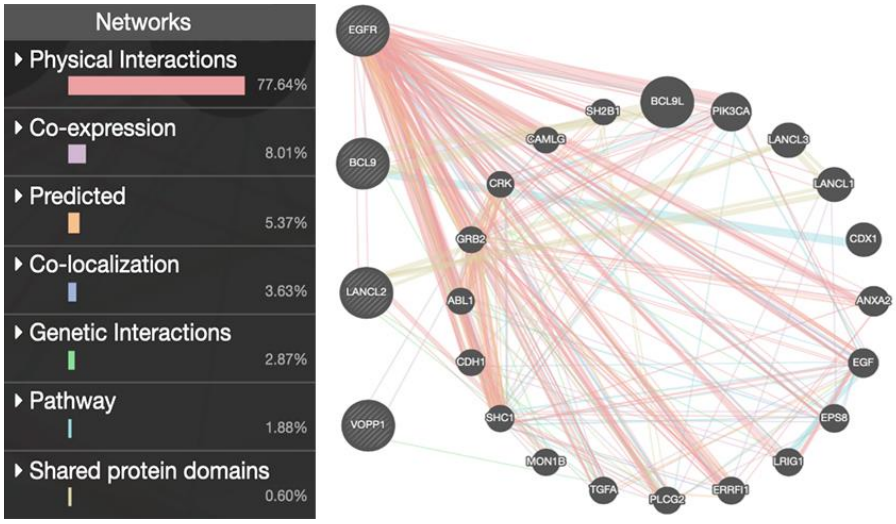


FIGURE 6 Cluster of interacting genes identified by GeneMANI.

4. DISCUSSION

The QH has been bred for 67 years in Brazil, with the inclusion of outside genetics with the use of different animals raised in other countries. The modern QH presents high performance with athletic functions in the most varied modalities (ABQM 2019). Nevertheless, an important point is the inbreeding in the racing line; individual estimates of F_{ROH} in this study revealed moderate to high inbreeding in this line equivalent to 0.19.

Meira et al. (2013) estimated that the average inbreeding coefficient based on pedigree information (F_{ped}) for the QH racing line in Brazil at 0.36, pointing to a higher incidence of inbred mating in this line when compared with the cutting line. However, those authors considered a small number of individuals and there were limitations due to the number of generations they considered in the pedigree file, leading to overestimated results. Other problems such as incomplete pedigree or errors could bias the inbreeding estimation when considering only pedigree information.

Estimates of autozygosity using genomic information are more accurate than those calculated using pedigree data (F_{ped}) (Rebelato & Caetano 2018). The autozygosity calculated by the pedigree disregards the sampling effects resulting from selection (Curik et al. 2014; Zhang et al. 2015). The F_{ROH} coefficient is recommended as one of the most accurate methods, even when compared with other estimates based on genomic data (Marras et al. 2015).

The ROH characterization results showed that the racing line had a higher proportion of the genome covered by smaller ROH, as determined by the ROH classes. Shorter segments in autozygosity arise from loss of diversity, the drift/bottleneck effect, the founder effect (Al-Mamun et al. 2015), or effects of more ancestral inbreeding in the population (Howrigan et al. 2011). On the other hand, longer ROH fragments indicate intense selection pressures (Metzger et al. 2015) and recent inbreeding events (Al-Mamun et al. 2015).

The longest shared homozygous segment was detected in ECA 7, with a length of around 25 Mbp. Studies have already reported the same findings in ECA 7, and it is probably an important region conserved for most equine breeds (Ablondi et al. 2019;

Grilz-Seger et al. 2019; Santos et al. 2021b). Therefore, selection events and genetic drift initially contribute to the formation of long ROHs (Pemberton et al., 2012). However, over time, these long ROHs undergo recombination events and genetic mutations, causing their size to decrease over successive generations (Curik et al. 2014).

There are several ROHet regions in the genome of the QH racing line. According to Williams et al. (2016), genes located in these regions are relevant to elucidate the fitness and survival of individuals, suggesting that some loci are under balancing selection or maybe harbor recessive lethal mutation sites. Szpiech et al. (2013) concluded that regions with high levels of heterozygosity may contain lower risks of rare deleterious variants, which in a way increases protection against diseases.

In our study, we did not observe any ROHet > 4 Mbp. However, no literature is available to provide insights into the biological events associated with varying lengths of ROHet in the genome. Mc Parland et al. (2009) reported that genomes with higher heterozygosity might exhibit better fitness, a phenomenon visible in the QH breed. This breed comprises diverse lineages with distinct selection objectives and criteria (Meira et al., 2013), and selective pressures have increased the inbreeding value. Therefore, genetic diversity based on ROHet is substantially high in the QH, as we identified several ROHet < 4 Mbp regions and some ROHet islands. In contrast, the Mangalarga Marchador horse breed genome exhibited only one ROHet island in ECA 11 (Santos et al., 2021a).

According to Williams et al. (2016), maintaining genomic heterozygosity in economically important species is essential for the sustainability of breeding programs. Therefore, ROHet can serve as a measure for assessing the genetic diversity of populations and assisting in animal breeding. Additionally, Samuels et al. (2016) suggest exploring patterns and phenotypic effects of different levels of genetic variation in populations based on heterozygosity ratio.

Identification of ROH islands helps visualize and identify haplotypic patterns of breeds or species being fixed due to selective pressures. In addition, ROH islands usually encompass genomic regions large enough to contain genes or a set of genes that may be under selection across generations (Rebelato & Caetano 2018). The

highlighted genes and their respective protein products play relevant roles in determining essential characteristics in high-performance horses.

Most homozygosity studies use complete annotation of the segments, an approach that makes it difficult to understand which regions are the most relevant indicators of this selection. Therefore, we only consider selection signatures with a High Frequency of ROH Occurrence. Therefore, one could expect to find many ROH and genes in these regions because species of economic interest have been subjected to artificial selection for countless generations. Thus, we focused on the most representative regions within the ROH, including the top three SNPs with a High Frequency of ROH Occurrence, noting only the top ROH island on each chromosome.

Among the 10 biological processes highlighted in ROH islands, those that are most associated with the performance of an athlete horse refer to fat cell differentiation (with the genes *CTBP1*, *WNT5B*, and *TMEM120B*), regulation of glucose metabolic process (*MAEA* and *NKX1-1*), heme transport (*PGRMC2*), and negative regulation of calcium ion import (*VDAC1*). Other genes in ROH islands such as *NTM*, *RFX3*, *BCL7A*, *MLXIP*, *LRTM2*, and *COBL* are related to body mass index; *RFX3* is related to weight; and *HPD*, *BCL7A*, *MLXIP*, *RNASEH2B*, *GLIS3*, and *NKX1-1* are related to height, essential functions for morphological evaluation of athlete horses. According to Jetten (2018), *GLIS3* plays a critical role as a key regulator for pancreatic β-cell generation and maturation, insulin gene expression, thyroid hormone biosynthesis, spermatogenesis, and the maintenance of normal kidney functions, showing a relationship, among others, with energy metabolism.

The gene *SETD1B* found within the ROH island in ECA 8 for the QH racing line has also been pointed out in Lipizzan horses with a frequency of > 50% (Grilz-Seger et al. 2019). Some ROH islands are likely signatures of selection in several equine breeds, and admixture probably justifies this phenomenon because some breeds have been responsible for the formation of others. Furthermore, Gurgul et al. (2019), studying a genome-wide scan for diversifying selection signatures in selected horse breeds, observed the association between *NTM* variation and cognitive function in humans. Thus, the assumptions suggest that diverse selection at this locus between draft horses

and light horses can potentially contribute to their different temperaments and the ability to develop different gaits as a function of brain-managed motor coordination.

Based on information available from the GeneCards (2022) and Ensembl (2021) websites, there are several possible associations for some of the above-mentioned genes with traits of interest in the QH racing line. The genes *BCL7A*, *UVSSA*, *WNT5B*, *FBXL14*, *NKX1-1*, *UVSSA*, *WNT5B*, *CACNA2D4*, *FBXL14*, and *PGRMC2* found in the QH racing line have already been related to bone density and calcium channels in different species. *RFX3* and *CFAP251* are linked to reproductive function, particularly sperm motility and testosterone levels. Genetic diversity investigated in the Italian Heavy Draught horse (IHDH) using both traditional and a genomic-based approaches revealed ROH islands with 60% frequency on ECA 6 (*FBXL14*) (Mancin et al. 2020).

LARP1B, *DLEV7*, *EPC2*, *RFX3*, *UVSSA*, *MAEA*, *KIF5C*, *CTBP1*, *SPON2*, *GLIS3*, and *PGRMC2* are related to type II diabetes mellitus in humans (GeneCards 2022; Sun et al. 2019) and, therefore, to energy metabolism involving glucose. *RFX3* and *GLIS3* are linked to insomnia (GeneCards 2022; Rouillard et al. 2016; Stein et al. 2018;); *UVSSA* and *MAEA* to Parkinson's disease; and *EPC2*, *MLXIP*, *KIF5C*, and *GLIS3* to schizophrenia (GeneCards 2022). It is notable that all these disorders are related to the nervous system. In turn, *ADIPOR2* is associated with an essential hormone that regulates glucose and lipid metabolism, acting in fat mass and glucose homeostasis (Ji et al. 2021).

The ROHet analysis identified eight olfactory receptor genes in ECA 7: family 7, subfamily D and E (*OR7D19*, *OR7D4G*, *OR7D4E*, and *OR7D4J*). The olfactory receptors likely constitute the most prominent gene superfamily in the vertebrate genome (Glusman et al. 2001). According to Olender et al. (2008), olfactory receptor genes undergo allelic deletion; each sensory neuron usually expresses only one odorant receptor allele. Moreover, those authors calculated a pseudogene ratio for humans and other species, but the ratio was higher in rats and mice, reflecting the necessity of olfactory function for the survival of rodents but not humans. Therefore, it is tempting to speculate that the receptors encoded by olfactory receptor genes found in the QH racing line might be the signature of selection associated with a respiratory capacity that

became required with genetic improvement of the breed, mainly due to the increase in respiratory demands required for a high-performance animal.

Highly heterozygous regions may harbor genes associated with immunity and resistance to specific pathogens (Penn et al. 2002). The candidate gene *EGFR* in a ROHet (ECA 4) is induced in peptic-injured equine gastric squamous epithelium, a factor in healing gastric squamous mucosal ulcers in horses (Jeffrey et al. 2001). Furthermore, the EGF-mediated response might be involved in the hyperproliferative response characteristic of chronic laminitis (Grosenbaugh et al. 1991). After functional annotation, the results of the current study indicate that *EGFR*, *BCL9*, *LANCL2*, and *VOPP1* probably act together, regulating some biological processes, cellular components, or molecular functions. *VOPP1* probably act together, regulating some biological processes, cellular components, or molecular functions. The results of the current study indicate that *EGFR*, *BCL9*, *LANCL2*, and *VOPP1* probably act together, regulating some biological processes, cellular components, or molecular functions. According to Zhao et al. (2021), in a human glioblastoma study, the *EGFR* and *LANCL2* genes are located in the same amplicon, with the *VOPP1* gene having one of the highest co-amplified frequencies with *LANCL2 or EGFR*.

Upon a closer examination of the network, intriguing genetic interactions (2.87%) were uncovered among the three genes, revealing a shared interaction with TGFA - a gene that was included in the analysis. TGFA encodes a growth factor that acts as a ligand for the epidermal growth factor receptor, thereby activating a signaling pathway that plays a pivotal role in regulating cell proliferation, differentiation, and development (Calaf et al., 2022).

The Wnt signaling pathway (including *EGFR* and *BCL9*) is an essential component of canonical Wnt signaling for differentiation of myogenic progenitors during muscle regeneration (Brack et al. 2009), a process of fundamental importance for individual athletes. The current work presented regeneration as the main biological process among the 10 highlighted. The other ROHet candidate genes found act in DNA repair (*CHD1L*) and as a transcriptional regulator (*ZNF699*).

5. CONCLUSIONS

The QH racing line population in Brazil shows moderate to high genomic inbreeding. ROH islands showed genes related to cell differentiation, regulation of glucose metabolic process, heme transport, and negative regulation of calcium ion import. This study is the first to characterize ROHet in the breed/line. Candidate genes identified in ROHet point to signatures of selection with biological importance for genes linked to respiratory capacity and muscle repair. Understanding genes related to muscle repair may lead to selecting animals with greater regenerative capacity and developing treatments for muscle disorders in the QH and other breeds. The identification of ROHet lies in contributing to enhancing and preserving the Quarter Horse breed through the development of reproductive strategies in animal breeding programs. Furthermore, these findings can serve as a foundational framework for further research on other equine breeds.

7. REFERENCES

Ablondi, M., Viklund, Å., Lindgren, G., Eriksson, S. *et al*. (2019) Signatures of selection in the genome of Swedish warmblood horses selected for sport performance. *BMC Genomics* 20, 1–12. https://doi.org/10.1186/s12 864-019-6079-1

ABQM (2019) Associação brasileira dos criadores de cavalos Quarto de Milha [webpage]. Retrieved from https://www.abqm.com.br

Al-Mamun, H.A., Clark, S.A., Kwan, P., Gondro, C. (2015) Genome-wide linkage disequilibrium and genetic diversity in five populations of Australian domestic sheep. *Genetics Selection Evolution* 47, 90. https://doi.org/10.1186/s12711-015-0169-6

Brack, A.S., Murphy-Seiler, F., Hanifi, J. et al. (2009) BCL9 is an essential component of canonical Wnt signaling that mediates the differentiation of myogenic progenitors during muscle regeneration. *Dev Biol.* 335, 93-105. https://doi.org/10.1016/j.ydbio.2009.08.014

Bertolini F., Servin B., Talenti A., Rochat, E. *et al*. (2018) Signatures of selection and environmental adaptation across the goat genome post- domestication. *Genetics, Selection, Evolution* 50(1), 57. https://doi.org/10.1186/s12711-018-0421-y

Biscarini, F., Cozzi, P., Gaspa, G., & Marras, G. (2018) detectRUNS: Detect runs of homozygosity and runs of heterozygosity in diploid genomes.

Calaf, G.M., Crispin, L.A., Muñoz, J.P., Aguayo, F., Roy, D., Narayan, G. (2022) Ionizing Radiation and Estrogen Affecting Growth Factor Genes in an Experimental Breast Cancer Model. *Int J Mol Sci.* 23, 14284. https://doi.org/10.3390/ijms232214284

Clayton, D. (2015) snpStats: SnpMatrix and XSnpMatrix classes and methods, R package version 1.24.0.

Curik, I., Ferencakovic, M., Solkner, J. (2014) Inbreeding and runs of homozygosity: a possible solution to an old problem. *Livestock Science* 166, 26–34. https://doi.org/10.1016/j.livsci.2014.05.034

Curik, I., Ferencakovic, M., Solkner, J. (2017) Genomic dissection of inbreeding depression: a gate to new opportunities. *Brazilian Journal of Animal Science* 46, 773–782. https://doi.org/10.1590/S1806-92902017000900010

Ensembl (2021) Genome browser [webpage]. Retrieved from https://www.ensembl.org

Evans, J.W. (1996) Horses: a guide to selection, care and enjoyment. Freeman and Company, New York. p. 797.

GeneCards® (2022) The Human Gene Database [webpage]. Retrieved from https://www.genecards.org

Gibson, J., Newton, E.M., Collins, A. (2006) Extended tracts of homozygosity in outbred human populations. *Human Molecular Genetics* 15, 789–795. https://doi.org/10.1093/hmg/ddi493

Grilz-Seger, G., Neuditschko, M., Ricard, A., Velie, B. *et al*. (2019) Genome-wide homozygosity patterns and evidence for selection in a set of European and near Eastern horse breeds. *Genes* 10, 491. https://doir.org/10.3390/genes10070491

Gurgul, A., Jasielczuk, I., Semik-Gurgul, E., Pawlina-Tyszko, K. *et al*. (2019) A genome-wide scan for diversifying selection signatures in selected horse breeds. *PLoS One* 14, e0210751. https://doi.org/10.1371/journal.pone.0210751

Glusman, G., Yanai, I., Rubin, I., Lancet, D. (2001) The complete human olfactory subgenome. *Genome Res* 11, 685-702. https://doi.org/10.1101/gr.171001

Grosenbaugh, D.A., Hood, D.M., Amoss, M.S.J., Williams, J.D. (1991) Characterisation and distribution of epidermal growth factor receptors in equine hoof wall laminar tissue: comparison of normal horses and horses affected with chronic laminitis. *Equine Vet J.* 23, 201-6. https://doi.org/10.1111/j.2042-3306.1991.tb02755.x

Harris, A.M., DeGiorgio, M. (2017) An unbiased estimator of gene diversity with improved variance for samples containing related and inbred individuals of any ploidy. *G3* 9, 671–691. https://doi.org/10.1534/g3.116.037168

Howrigan, D., Simonson, M., Keller, M. (2011) Detecting autozygosity through runs of homozygosity: a comparison of three autozygosity detection algorithms. *BMC Genomics* 12, 460. https://doi.org/10.1186/1471-2164-12-460

Jeffrey, S.C., Murray, M.J., Eichorn, E.S. (2001) Distribution of epidermal growth factor receptor (EGFr) in normal and acute peptic-injured equine gastric squamous epithelium. *Equine Vet J.* 33, 562-9. https://doi.org/10.2746/042516401776563481

Jetten, A.M. (2018) GLIS1-3 transcription factors: critical roles in the regulation of multiple physiological processes and diseases. *Cell Mol Life Sci.* 75, 3473–3494. https://doi.org/10.1007/s00018-018-2841-9

Ji, R., Xu, X., Turchini, M.G., Mai, K., Ai, Qi. (2021) Adiponectin's roles in lipid and glucose metabolism modulation in fish: Mechanisms and perspectives. *Reviews in Aquaculture* 13, 2305–2321. https://doi.org/10.1111/raq.12569

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

Keller, M.C., Visscher, P.M., Goddard, M.E. (2011) Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. *Genetics* 189, 237–249. https://doi.org/10.1534/genetics.111.130922

Lima, R.A.S., Shirota, R., Barros, G.S.C. (2006) Estudo do Complexo do Agronegócio Cavalo no Brasil. Centro de Estudos Avançados em Economia Aplicada – CEPEA.

Llaurens, V., Whibley, A. Joron, M. (2017) Genetic architecture and balancing selection: the life and death of differentiated variants. *Molecular Ecology* 26, 2430–2448. https://doi.org/10.1111/mec.14051

López-Cortegano, E., Pouso, R., Labrador, A., Pérez-Figueroa, A. (2019) Optimal Management of Genetic Diversity in Subdivided Populations. *Frontiers in Genetics* 10, 843. https://doi.org/10.3389/fgene.2019.00843

Mancin, E., Ablondi, M., Mantovani, R., Pigozzi, G. *et al.* (2020) Genetic Variability in the Italian Heavy Draught Horse from Pedigree Data and Genomic Information. *Animals* 10, 1310. https://doi.org/10.3390/ani10081310

Marchiori, C., Pereira, G.L., Maiorano, A.M., Rogatto, G.M. *et al*. (2019) Linkage disequilibrium and population structure characterization in the cutting and racing lines of Quarter Horses bred in Brazil. *Livestock Science* 219, 45–51. https://doi.org/10.1016/j.livsci.2018.11.013

Marras, G., Gaspa, G., Sorbolini, S., Dimauro, C. *et al*. (2015) Analysis of runs of homozygosity and their relationship with inbreeding in five cattle breeds farmed in Italy. *Animal Genetics* 46, 110–121. https://doi.org/10.1111/age.12259

Marras, G., Wood, B.J., Makanjuola B., Malchiodi F. *et al.* (2017) Characterization of runs of homozygosity and heterozygosity-rich regions in a commercial turkey (Meleagris gallopavo) population. *Proceedings of the World Congress on Genetics Applied to Livestock Production* 11.763.

McQuillan, R., Leutenegger, A.L., Abdel-Rahman, R., Franklin, C.S. *et al*. (2008) Runs of homozygosity in European populations. *The American Journal of Human Genetics* 83, 359–372. https://doi.org/10.1016/j.ajhg.2008.08.007

Meira, C.T., Curi, R.A., Farah, M.M., Oliveira, H.N. *et al*. (2014) Prospection of genomic regions divergently selected in racing line of Quarter Horses in relation to cutting line. *Animal* 8, 1754-1764. https://doi.org/10.1017/S1751731114001761

Meira, C.T., Curi, R.A., Silva, J.A.II.V., Correa, M.J.M. *et al*. (2013) Morphological and genomic differences between cutting and racing lines of Quarter Horses. *Journal of Equine Veterinary Science* 33, 244–249. https://doi.org/10.1016/j.jevs.2012.07.001

Metzger, J., Karwath M., Tonda R., Beltran S. *et al*. (2015) Runs of homozygosity reveal signatures of positive selection for reproduction traits in breed and non-breed horses. *BMC Genomics* 16, 764. https://doi.org/10.1186/s12864-015-1977-3

Olender, T., Lancet, D., Nebert, D.W. (2008) Update on the olfactory receptor (OR) gene superfamily. *Hum Genomics* 3, 87-97. doi:10.1186/1479-7364-3-1-87

Mc Parland, S., Kearney, F., Berry, D.P. (2009) Purging of Inbreeding Depression within the Irish Holstein-Friesian Population. *Genet. Sel Evol.* 41, 16. https://doi.org/10.1186/1297-9686-41-16

Pemberton, T.J., Absher, D., Feldman, M.W., Myers, R.M. *et al*. (2012) Genomic patterns of homozygosity in worldwide human populations. *The American Journal of Human Genetics* 91, 275–292. https://doi.org/10.1016/j.ajhg.2012.06.014

Pereira, G.L, Chud, T.C.S., Bernardes, P.A, Venturini, G.V *et al.* (2017) Genotype Imputation and Accuracy Evaluation in Racing Quarter Horses Genotyped Using Different Commercial SNP Panels. *Journal of Equine Veterinary Science* 58, 89–96. https://doi.org/10.1016/j.jevs.2017.07.012

Penn, D.J., Damjanovich, K., Potts, W.K. (2002) MHC heterozygosity confers a selective advantage against multiple-strain infections. *Proceedings of the National Academy of Sciences of the United States of America* 99, 11260–11264. https://doi.org/10.1073/pnas.162006499

Rebelato, A.B., Caetano, A.R. (2018) Runs of homozygosity for autozygosity estimation and genomic analysis in production animals. *Pesquisa Agropecuária Brasileira*, 53, 975–984. https://doi.org/10.1590/S0100-204X2018000900001

Rouillard, A.D., Gundersen, G.W., Fernandez, N.F., Wang, Z. *et al.* (2016) The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* 2016, baw100. https://doi.org/10.1093/database/baw100

Ruan, D., Yang, J., Zhuang, Z., Ding, R. *et al*. (2022) Assessment of heterozygosity and genome-wide analysis of heterozygosity regions in two duroc pig populations. *Frontiers in Genetics* 12, 812456. https://doi.org/10.3389/fgene.2021.812456

Sakamoto, T., Innan, H. (2019) The evolutionary dynamics of a genetic barrier to gene flow: from the establishment to the emergence of a peak of divergence. *Genetics* 212(4), 1383–1398. https://doi.org/10.1534/genetics.119.302311

Samuels, D.C., Wang, J., Ye, F., He, J. *et al*. (2016) Heterozygosity ratio, a robust global genomic measure of autozygosity and its association with height and disease risk. *Genetics* 204, 893–904. https://doi.org/10.1534/genetics.116.189936

Santos, B.W., Pimenta Schettini, G., Fonseca, M.G., Pereira, G.L. et al. (2021a) Fine-scale estimation of inbreeding rates, runs of homozygosity and genome-wide heterozygosity levels in the Mangalarga Marchador horse breed. *Journal of Animal Breeding and Genetics* 138, 161–73. https://doi.org/10.1111/jbg.12508

Santos, W.B., Schettini, G.P., Maiorano, A.M., Bussiman, F.O. *et al*. (2021b) Genome-wide scans for signatures of selection in Mangalarga Marchador horses using high-throughput SNP genotyping. *BMC Genomics* 22, 737. https://doi.org/10.1186/s12864-021-08053-8

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

Smedley, D., Haider, S., Ballester, B., Holland, R. *et al*. (2009) BioMart – biological queries made easy. *BMC Genomics* 10, 22. https://doi.org/10.1186/1471-2164-10-22

Stein, M.B., McCarthy, M.J., Chen, C.Y., Jain, S. *et al.* (2018) Genome-wide analysis of insomnia disorder. *Molecular Psychiatry* 23, 2238–2250. https://doi.org/10.1038/s41380-018-0033-5

Sun, J., Have, C.T., Hollensted, M., Grarup, N. *et al*. (2019) Sequencing reveals protective and pathogenic effects on development of diabetes of rare GLIS3 variants. *PLoS One* 14, e0220805. https://doi.org/10.1371/journal.pone.0220805

Szpiech, Z.A., XU, J., Pemberton, T.J., Peng, W. *et al*. (2013) Long Runs of Homozigosity are Enriched for Deleteriuos Variation. *The American Journal of Human Genetics* 93,90–102. https://doi.org/10.1016/j.ajhg.2013.05.003

Williams, J.L., Hall, S.J., Del Corvo, M., Ballingall, K.T. *et al*. (2016) Inbreeding and purging at the genomic Level: The Chillingham cattle reveal extensive, non-random SNP heterozygosity. *Animal Genetics* 47, 19–27. https://doi.org/10.1111/age.12376

Yu, G., Wang, L., Han, Y., He, Q. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287. https://doi.org/10.1089/omi.2011.0118

Zappala, Z., Montgomery, S.B. (2016) Non-coding loss-of-function variation in human genomes. *Human Heredity* 81(2), 78–87. https://doi.org/10.1159/000447453

Zhang, Q., Guldbrandtsen, B., Bosse, M., Lund, M.S. *et al*. (2015) Runs of homozygosity and distribution of functional variants in the cattle genome. *BMC Genomics* 16, Wu 542. https://doi.org/10.1186/s12864-015-1715-x.

Zhao, H.F., Zhou, X.M., Wang, J. et al. (2021) Identification of prognostic values defined by copy number variation, mRNA and protein expression of LANCL2 and EGFR in glioblastoma patients. *J Transl Med* 19, 372. https://doi.org/10.1186/s12967-021-02979-z

Zheng, W., Huang, L., Huang, J., Wang, X. *et al.* (2013) High genome heterozygosity and endemic genetic recombination in the wheat stripe rust fungus. *Nature Communications* 4, 2673. https://doi.org/10.1038/ncomms3673

CHAPTER III - New Genomic Insights into Gait Variations in Mangalarga Marchador Horses [b]

Abstract – Continuing genetic advancements and systematic investigations with higher genomic resolution have revealed the interconnectedness of the *DMRT3* gene with other genes carrying specific mutations, collectively influencing the intricate regulation of equine gait patterns. Our study involved a cohort of 62 male and 130 female Mangalarga Marchador horses displaying distinct gait phenotypes: “Picada” (n = 86) and “Batida” (n = 106). We performed genome-wide association studies (GWAS) and assessed extended haplotype homozygosity (EHH) to identify variants associated with candidate genes and thus to obtain novel insights regarding selection signals for the “Batida” and “Picada” gaits in Mangalarga Marchador horses. Our findings have unveiled essential insights into the genetic foundations potentially underlying gait type differentiation in Mangalarga Marchador horses, as well as similar phenotypes in other equine breeds. We successfully pinpointed specific candidate genes or gene groups that may complement the role of *DMRT3* in shaping gait phenotypes. Of the 11 single nucleotide polymorphisms (SNPs) we identified, eight genes were notable (*PIP5K1B*, *PABIR1*, *PGM5*, *DOCK8*, *KANK1*, *DMRT1*, *DMRT3*, and *DMRT2).* The AX-103466344 SNP displayed distinct behavior when in homozygous conditions: Except for two horses, the AA genotype consistently resulted in the “Batida” gait, while the GG genotype consistently resulted in the “Picada” gait. We speculate that animals heterozygous for this genotype may be subject to some type of dominance deviation. Our research has provided robust evidence indicating that the “Batida” and “Picada” gait phenotypes in the Mangalarga Marchador breed may be linked to causative mutations in the eight identified genes. Additionally, these SNPs located in non-coding regions of the genome could potentially serve as enhancers.

Keywords: Batida gait, DMRT3, equine genotyping array, Picada gait

[b] *Article prepared for submission to "Heredity - Springer Nature".*

1. INTRODUCTION

Initially, the doublesex and Mab-3 related transcription factor 3 (*DMRT3*) gene was primarily associated with sex, as it encodes an extensive family of transcription factors, and its function was extensively studied in the context of sexual development (Zarkower 2001; Volff et al. 2003; Hong et al. 2007). However, Hong et al. (2007) expanded this concept by demonstrating that the gene's function in vertebrates extends beyond its role in sex. It regulates gene expression in developing gonads, suggesting that this gene may play a fundamental role in other developmental processes.

Locomotion in mammals is regulated by a central circuit that replicates established spinal interneuron patterns. This circuit coordinates limb movements, ensures alternation between the left and right sides, and activates flexor and extensor muscles (Kullander 2005; Grillner 2006). As highlighted by Robilliard et al. (2007), quadrupeds employ various step patterns during locomotion. Gaits can be classified into discrete step patterns and subdivided into symmetric and asymmetric categories. Regarding horses, there is notable diversity in locomotion patterns, including three natural gaits in increasing order of speed: walk, trot, and gallop. Some horses also exhibit alternating gaits at intermediate speeds, leading to the development of specialized breeds. These alternative gaits, such as the running walk, rack, broken pace, hard pace, and broken trot, have had historical importance and are popular today among equestrians for the trail or pleasure gait and at horse shows (Vincelette 2023).

Since the remarkable discovery made by Andersson et al. (2012), revealing the significant impact of a premature stop codon mutation in the *DMRT3* gene on the locomotion pattern in horses, it has been understood that this mutation exerts a permissive influence on the ability to perform alternating gaits. It also has beneficial effects on the performance of horses, particularly in sports. The mutant A allele of the g.22999655C>A single nucleotide polymorphism (SNP) in the *DMRT3* gene is prevalent in horses exhibiting the gait phenotype (Anderson et al. 2012; Promerová et al. 2014). Initially, it was believed that the homozygous AA condition was strictly associated with gait due to the almost total presence of this allele in these horses. However, further investigation has revealed that breeds that do not manifest the gait phenotype may also

possess the mutant (A) allele (Promerová et al. 2014). Furthermore, the A allele has also been associated with a higher quality of rhythm and gait compared with horses that have the CC genotype (Andersson et al. 2012; Jäderkvist et al. 2015). This underscores the complexity of the genetics behind locomotion patterns in horses and how the presence of the mutant A allele can influence different aspects of gait.

For an untrained individual, clearly and objectively identifying the two gait modalities in Mangalarga Marchador horses can be an overwhelming task that involves observing numerous details. However, the primary distinction between the “Batida” and “Picada” gaits essentially lies in the way the movement is executed. In the “Batida” gait, diagonal support is more common compared with triple support. Conversely, in the “Picada” gait, there is an overlap of lateral and triple supports, resulting in a smoother execution of the movement (Santos et al. 2021).

As genetic improvements have advanced and studies have been replicated in populations of different equine breeds, it has become evident that the *DMRT3* mutation is not sufficient to explain the diversity of gait patterns, which are distinctive characteristics of certain equine breeds. Therefore, we employed two genome-wide association study (GWAS) methodologies to investigate the two gait modalities in the Mangalarga Marchador breed. Our goal was to identify variants associated with candidate genes, shedding light on new findings regarding selection signals in each of these modalities. Additionally, we examined the pathways in which these candidate genes are involved by drawing on the existing literature and extended haplotype homozygosity (EHH). We have delved into aspects of significance for the Mangalarga Marchador breed, particularly those pertaining to the gait type and quality.

2. METHODS

2.1 Sample collection, Animals, Phenotypes, and DNA Extraction

The animals were meticulously selected from the 36th Brazilian National Exhibition of the Mangalarga Marchador breed in Brazil and from various stud farms located in the states of São Paulo and Minas Gerais. The cohort comprised 62 males and 130 females with distinct gait phenotypes: “Picada” (n = 86) and “Batida” (n = 106).

To ensure the integrity of the study, steps were taken to include animals from different lineages, avoiding any potential inclusion of close relatives. Jugular blood samples (5 mL) were collected from each horse and preserved with 7.5 mg of ethylenediaminetetraacetic acid (EDTA). Genomic DNA was extracted using the Illustra Blood Genomic PrepMini Spin Kit (GE Healthcare, USA), following the manufacturer's instructions. The DNA concentration was quantified with a Qubit® 3.0 Fluorometer (Invitrogen, USA) and its quality was assessed with a NanoDrop™ Lite Spectrophotometer (NanoDrop Lite, Thermo Scientific, USA), and 0.8% agarose gel electrophoresis. The resulting sample dilutions averaged approximately 10 ng/μL.

2.2 Genotype and Quality Control

Equine samples were genotyped by using the 670k Axiom® Equine Genotyping Array, referred to as Axiom MNEC670 (Thermo Scientific, USA). The genotyping was performed with default settings for diploid organisms, and specific configurations were applied using the Axiom™ Analysis Suite Software version 4. To ensure high-quality data, the SNP quality control criteria included a data quality control (DQC) score of ≥ 0.82, a calling rate of ≥ 97%, and a passing sample percentage of ≥ 95%. Furthermore, a comprehensive set of 26 parameters tailored specifically for diploid organisms was meticulously employed, with in-depth descriptions provided in a previous publication (Additional file 4: Methods S1 in Santos et al. 2021).

2.3 LD and Population Structure Analyses

The LD and population structure analyses have been described previously (Santos et al. 2021). In summary, the marker density was reduced from 422,656 to 347,935 SNPs using LD decay analysis, following the default settings of the PopLDdecay pipeline. Subsequent analyses were carried out in R, utilizing tools such as pegas (Paradis 2010), ape (Paradis and Schliep 2018), and ggplot2 (Wickham et al. 2016). Plink 1.9 (Purcell et al. 2007) was used for principal component analysis (PCA). The procedure included the removal of SNP pairs with high base correlations and optimization of computational operations using the “--indep-pairwise” option.

Additionally, genome-wide estimates of IBD (identity by descent) were calculated to assess relatedness. In the PCA, one sample from each closely related group was excluded, which was identified through high pairwise PI_HAT statistic sums in R. For more detailed information about methods, please refer to a previous publication (Santos et al. 2021).

2.4 GWAS Analysis

The GWAS analysis was performed by two different mixed model association methods: Mixed linear model association (MLMA) of the GCTA and Weighted Single-Step Genomic Best Linear Unbiased Prediction (WssGBLUP).

2.5 GCTA

Mixed linear model association (MLMA) analysis was executed within the GCTA software (Yang et al. 2011), utilizing the “'--mlma” option specifically for autosomal regions. This method offers numerous advantages compared with traditional genetic association techniques. MLMA serves as a robust safeguard against spurious associations arising from population structure and inherent kinship, both common features in GWAS (Yang et al. 2014). The association between genotyped SNPs and gait trait used the following mixed linear model equation:

$$y = a + b * x + g + e$$

where y is the phenotype; a is the mean term; b is the additive effect (fixed the genetic variance estimator var(g) in each population is computed from the null model (y = a + g + e) and fixed later as we test each SNP for association with the trait; x is the genotyping indicator variable, coded as 0, 1, and/or 2; g is the sum of random effects of all SNPs in the genome-wide array; and the residual.

2.6 Weighted Single-Step Genomic Best Linear Unbiased Prediction (WssGBLUP)

The genomic estimated breeding values (GEBVs) for all animals were determined through a single-step genomic best linear unbiased prediction (ssGBLUP) method, followed by transformation of these values into SNP effects. The percentage of genetic

variance attributed to specific SNP windows was ascertained based on consecutive SNP effects sets in the analysis. The investigation was carried out using the BLUPF90 software family developed by Misztal et al. (2002). The linear mixed effects model was:

$$y = Xb + Zu + e$$

where y is the phenotypic observations; X is design matrix relating fixed effects to the phenotypes; b is a vector of fixed effects; Z is a design matrix relating the breeding values to the phenotypes; u is a vector of random animal breeding values with u ~ N (0, $H\sigma^2_u$); and e is the residual vector with e ~ N (0, $H\sigma^2_e$).

Initially, WssGBLUP was conducted by utilizing the BLUPF90 software by Misztal et al. (2014). The relationship between the variables a and e has zero covariance, with a having a variance of H times $H\sigma a^2$ and e having a variance of I times $I\sigma e^2$. Here, $H\sigma a^2$ represents the direct additive variance, while $I\sigma e^2$ represents the residual variance. The H matrix combined genomic information with pedigree information, and the H^{-1} matrix provided the inverse of a joint relationship (Aguilar et al. 2010):

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

where A22 denotes the component within matrix A that pertains to the genotyped animals, and G represents the genomic relationship matrix (VanRaden 2008). Subsequently, these ssGBLUP results were employed to predict the GEBV. The SNP effects were calculated by using the postGSf90 module within the BLUPF90 software suite. The steps for calculating the SNP effects and weights in the WssGWAS were carried out as described by Wang et al. (2012):

$$\hat{u}_{(t)} = \lambda D_{(t)} Z' G^{*-1}_{(t)} \hat{a}_g$$

where û represents the vector of SNP effects; λ is the variance ratio computed according to VanRaden et al. (2009); $\hat{a_g}$ denotes the GEBVs; Z is the matrix relating the genotypes for each locus; G is the genomic relationship matrix; and D is the diagonal matrix of SNP variance weights (initially set as D = 1). The calculation of D is derived through the steps below, where t signifies the iteration number, and i pertains to the i-th SNP.

1. $t = 0$; $D = I$; $G^* = ZDZ'\lambda$.

2. Calculate $\hat{a}_g$ by ssGBLUP.

3. Compute $\hat{u}_{(t)} = \lambda D_{(t)} Z' G^{*-1}_{(t)} \hat{a}_g$.

4. Compute $G^{x}_{(t+1)} = u^2 2p(1-p_i)$ for all *i* as described by Zhang et al. (2010).

5. Normalize by $D_{(t+1)} = \frac{tr(D_{(0)})}{tr(D^{*}_{(t+1)})} D^{*}_{(t+1)}$.

6. Calculate $G^{*}_{(t+1)} = ZD_{(t+1)}Z'$.

7. $t = (t + 1)$.

8. Return to step 2.

The nonlinear A method, described by VanRaden (2008), was used to estimate the variances of SNPs (used as weights). Iterations progressively amplify the influence of SNPs with substantial effects while attenuating the impact of those with minor effects. The procedure underwent five iterations, guided by accuracy considerations as outlined by Wang et al. (2014), and one specific iteration was chosen for the present study.

2.7 EHH Analysis

The R package rehh was employed to compute EHH and, subsequently, to construct bifurcation structures illustrating the decline in haplotype homozygosity around a central marker, as described by Sabeti et al. (2002). For biallelic markers without any missing data, only bifurcations are possible. Consequently, the bifurcation structure for a particular allele and a specific side forms a tree, with its origin situated at the central marker. EHH was calculated by following the approach outlined by Tang et al. (2007):

$$EHH^{a}_{s,t} = \frac{1}{na(na-1)} \sum_{k=1}^{k^{a}_{s,t}} nk(nk-1)$$

where a is the core allele (ancestral and derived); s is the focal SNP; t is the chromosome interval comprised between the core allele as and the SNP; $k^{a}_{s,t}$ represents the number of distinct haplotypes (extending from SNP s to SNP t) carrying the core allele as; nk is the observed count for the kth haplotype; and na is the total number of haplotypes carrying the core allele as.

2.8 Gene Annotation and Enrichment Analysis

Gene annotation was conducted for genomic regions identified as indicative of selection. The selection of window sizes was established at 125 kb in both directions around each significant region or SNP (with a significance level of $p < 0.01$). The determination of the window size was guided by LD information and approximate values drawn from the existing literature. To identify genes within these designated windows, the latest equine genome assembly (EquCab3.0) was employed in conjunction with the BioMart R package (Smedley et al. 2009). Enrichment analysis utilized the R package clusterProfiler to infer biological processes, molecular functions, and cellular components through GO Enrichment Analysis of a gene set (Wu et al. 2021). The database employed was "org.Hs.eg.db." To ensure an in-depth discussions, only GO terms with a q-value < 0.05 were considered.

3. RESULTS

In both analyses, we detected selection signals on chromosome 23 (ECA 23), involving the same set of SNPs (Fig. 1a and 1b). It is worth noting that weighted single-step genome-wide association study (WssGWAS) and genome-wide complex trait analysis (GCTA) results were conveyed as SNP effects through an iterative approach and p-values, respectively, and thus differed in their extensions.

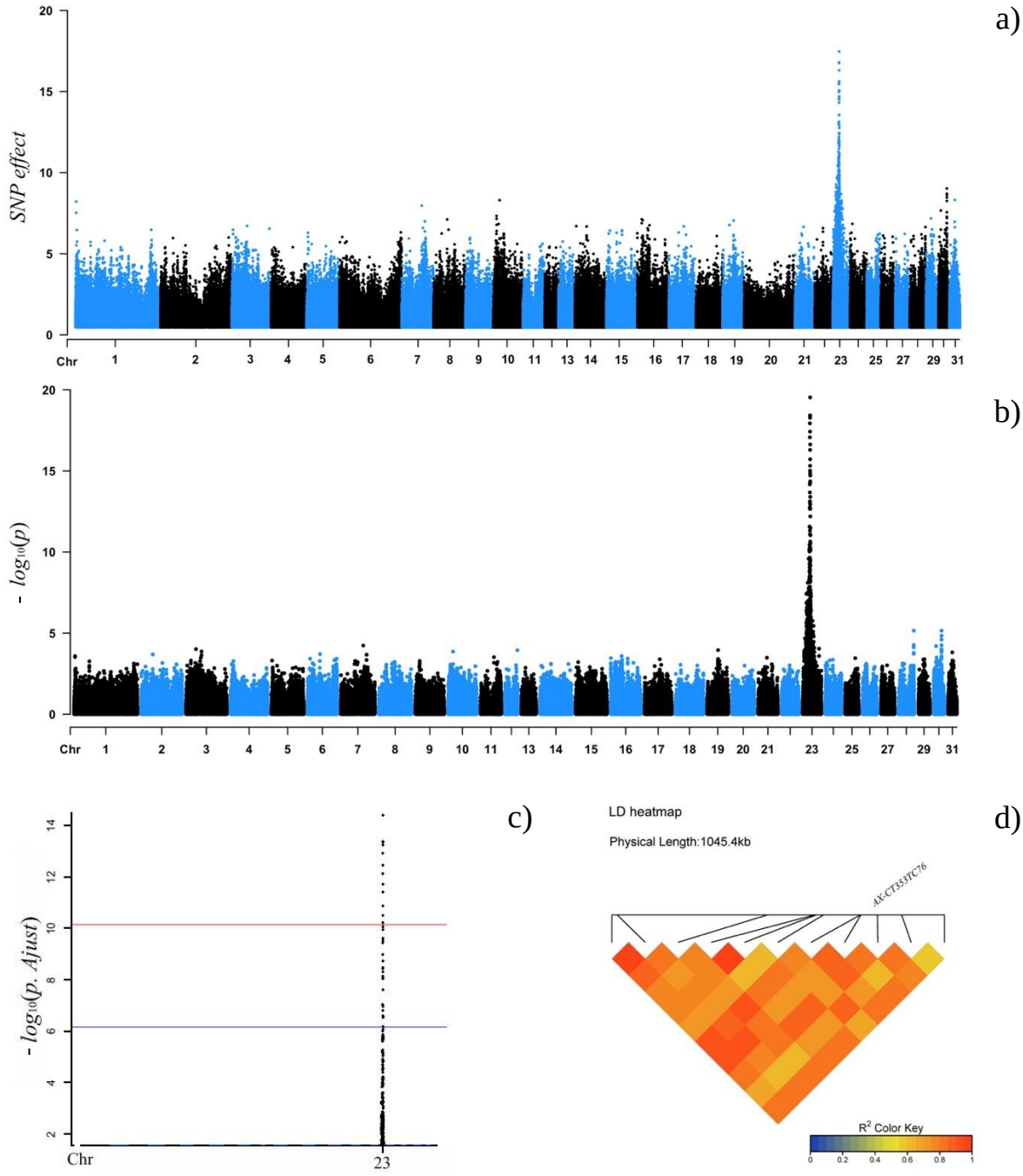


Fig. 1 Manhattan plot for the values of single nucleotide polymorphism (SNP) effects [*abs(snp_i/sd(snp)*] in the equine autosomal chromosomes. (a) Weighted single-step genomic best linear unbiased prediction (WssGBLUP) for the values of SNP effects [*abs(snp_i/sd(snp)*]. (b) Genome-wide complex trait analysis (GCTA) for p-values. (c) GCTA for Bonferroni-adjusted p-values (p.adjusted). The blue line represents p.adjusted

< 1e-05 and red line represents p.adjusted < 5e-10. (d) Linkage disequilibrium (LD) heatmap for 11 SNP (p.adjusted < 5e-10).

We investigated nucleotides on positions 21366827–22412229 pb on ECA 23 regarding linkage disequilibrium (LD) that is persistent in this region. We found numerous highly significant SNPs (Bonferroni-adjusted p < 1e-10) (Fig. 1c) with a high correlation. Even though the LD heatmap only displays 11 SNPs, it still clearly reveals the strong correlation present within this specific region of ECA 23 (Fig. 2d). Consequently, we opted to focus on a specific subset of 11 SNPs (Bonferroni-adjusted p < 2e-10) identified within the peak region of the GWAS (Fig. 2c). With this focused investigation, we explored potential associations between the genotype combinations at these 11 SNPs and the observed patterns of gait modality segregation, as meticulously outlined in Table 1.

Table 1. Specific subset of 11 single nucleotide polymorphisms (SNPs) identified in the genome-wide association study (GWAS) peak region.

Chr	SNP	bp	A1	A2	Freq	b	se	p. Adjusted*
23	AX-103466344	22202053	A	G	0.458333	0.402735	0.0436736	1.238581e-14
23	AX-103774706	22004513	A	G	0.432292	0.390965	0.0437116	1.583599e-13
23	AX-103042226	22148643	G	A	0.432292	0.402094	0.0449871	1.675197e-13
23	AX-103362470	22009395	C	T	0.434896	0.387770	0.0435336	2.208492e-13
23	AX-103007532	21366827	C	T	0.414062	0.408592	0.0463356	4.920223e-13
23	AX-103530176	21856570	T	C	0.421875	0.396263	0.0456170	1.576321e-12
23	AX-103450841	21383394	G	A	0.408854	0.399777	0.0465198	3.559681e-12
23	AX-103194688	22412229	T	G	0.393229	0.386499	0.0455967	9.810902e-12
23	AX-104633558	22149737	C	T	0.406250	0.389082	0.0463920	2.110761e-11
23	AX-103949722	22277970	T	C	0.406250	0.362848	0.0440990	8.043609e-11
23	AX-103674961	22032558	A	G	0.377604	0.377618	0.0465251	2.029104e-10

Chr: chromosome; SNP, single nucleotide polymorphism; bp, physical position; A1, reference allele; A2, other allele; Freq, frequency of the reference allele; b, SNP effect; se, standard error; * p. Adjust for Bonferroni test.

We found that the peak SNP (AX-103466344) plays a significant role in explaining the variation. All but one horse with the AA genotype consistently exhibited the “Batida” gait, while all but one horse with the GG genotype displayed the “Picada” gait. Among the 11 analyzed SNPs, we identified a total of 36 distinct genotype

combinations, considering the two types of heterozygotes as separate genotypes. However, only five of these genotypes were associated with ambiguity in terms of phenotype, where the same genotype was linked to two distinct phenotypes. The relationship between genotypes and the mentioned gait types is detailed in Table 2 and Supplementary Material 1.

The SNP annotation revealed the presence of eight genes (*PIP5K1B*, *PABIR1*, *PGM5*, *DOCK8*, *KANK1*, *DMRT1*, *DMRT3*, and *DMRT2*). *DMRT3* is known for its association with gait characteristics in horses and several other animal species (Table 3). The peak SNP (AX-103466344) is approximately 176.3 kilobases (kb) away from *DMRT3*, while the closest SNP (AX-103194688) is located 33.8 kb from *DMRT3*. Overall, these 11 SNPs are in a window of approximately 1,045 megabases (Mb) (Fig. 2).

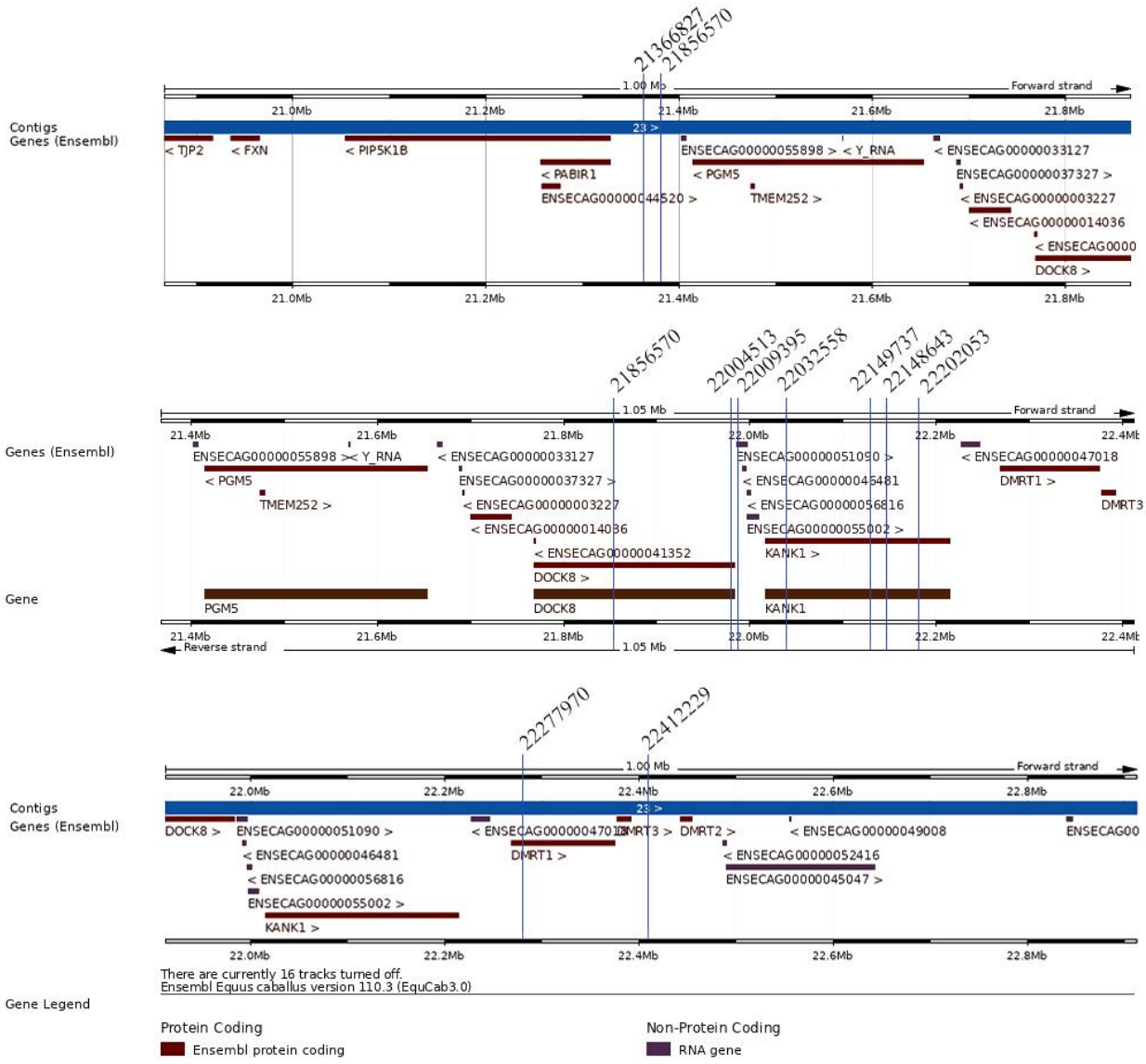


Fig. 2 The location of single nucleotide polymorphisms (SNPs) in the *Equus caballus* ensemble v 3.0 is marked by the blue lines, which delineate the positions of each of the 11 SNPs.

Table 2. Group of genotypes derived from the 11 single nucleotide polymorphisms (SNPs) identified in the genome-wide association study (GWAS) peak region.

Genotypes	Batida gait	Picada gait
AAAAAGCCCTCTAGGTCTTTAG	5	0
AAAAGGCCCCCTAATTTTTTAA	2	0
AAAAGGCCCCTTAATTTTCTAG	10	0
AAAAGGCCCCTTAATTTTTTAA	31	0
AAAAGGCCCCTTAATTTTTTAG	1	0
AAAAGGCCCCTTAGTTCTTTAA	1	0
AAAAGGCCCTTTAGGTCTCTAG	0	1
AAAAGGCCCTTTAGGTCTTTAA	1	0
AAAAGGCCCTTTAGTTCTTTAA	1	0
AAAAGGCCCTTTAGTTTTCTAG	1	0
AAAGAGCTCTCTAGGGCTTTAG	1	0
AAAGGGCTCCCTAAGTTTTTAA	1	0
AAAGGGCTCCTTAAGTTTCTAG	1	0
AAAGGGCTCTTTAGGGCTTTAA	1	0
AAAGGGCTCTTTAGGTTTTTAA	1	0
AAGGGGTTCCTTAAGGTTTTAA	1	0
AGAGAACTTTCCGGGGCCCTGG	2	2
AGAGAGCTCTCCAGGTCTCTAG	1	1
AGAGAGCTCTCTAGGGCCCTAG	2	0
AGAGAGCTCTCTAGGTCTCCGG	5	2
AGAGAGCTCTCTAGGTCTCTAG	32	3
AGAGAGCTCTCTGGGTCCCTAG	1	0
AGAGAGCTTTCTGGGTCCCTAG	2	0
AGGGAGTTCTCTAGGGCTCTAG	1	2
GGGGAACTCTCTAGGGCCCCGG	0	1
GGGGAATTTTCCGGGGCCCCGG	1	74

For the EHH analysis, we employed AX-103466344 as the focal marker. EHH begins at 1 and diminishes to 0 as the distance “t” (the chromosomal interval between the central allele “as” and the SNP) from the focal marker “s” (focal SNP) increases. We observed high EHH (Fig. 3a), along with the segregation of two haplotypes based on bifurcation diagrams for the derived and ancestral alleles of the marker in the region, for both chromosomes and individual trees (Fig. 3b and 3c).

The functional annotation revealed 89 biological processes, 13 molecular functions, and 29 cellular components (q-value < 0.05) (Supplementary Material 2). Fig. 4 shows the top 10 GO terms for each category.

Table 3. Candidate genes Associated with the 11 single nucleotide polymorphisms (SNPs) detected from the genome-wide association study (GWAS) of Mangalarga Marchador horses.

Ensembl_gene_id	Symbol	Chr	Start	End	Description
ENSECAG00000019992	*PIP5K1B*	23	21053966	21329204	phosphatidylinositol-4-phosphate 5-kinase type 1 beta [Source:VGNC Symbol;Acc:VGNC:21467]
ENSECAG00000003116	*PABIR1*	23	21256305	21328786	PP2A Aalpha (PPP2R1A) and B55A (PPP2R2A) interacting phosphatase regulator 1 [Source:VGNC Symbol;Acc:VGNC:17793]
ENSECAG00000010805	*PGM5*	23	21414523	21652879	phosphoglucomutase 5 [Source:VGNC Symbol;Acc:VGNC:21361]
ENSECAG00000011917	*DOCK8*	23	21768845	21983585	dedicator of cytokinesis 8 [Source:VGNC Symbol;Acc:VGNC:17269]
ENSECAG00000016219	*KANK1*	23	22016369	22214469	KN motif and ankyrin repeat domains 1 [Source:VGNC Symbol;Acc:VGNC:57186]
ENSECAG00000060160	*DMRT1*	23	22268768	22375685	doublesex and mab-3 related transcription factor 1 [Source:HGNC Symbol;Acc:HGNC:2934]
ENSECAG00000023412	*DMRT3*	23	22378399	22392510	doublesex and mab-3 related transcription factor 3 [Source:NCBI gene (formerly Entrezgene);Acc:100147177]
ENSECAG00000035701	*DMRT2*	23	22443097	22454628	doublesex and mab-3 related transcription factor 2 [Source:HGNC Symbol;Acc:HGNC:2935]

Chr: chromosome.

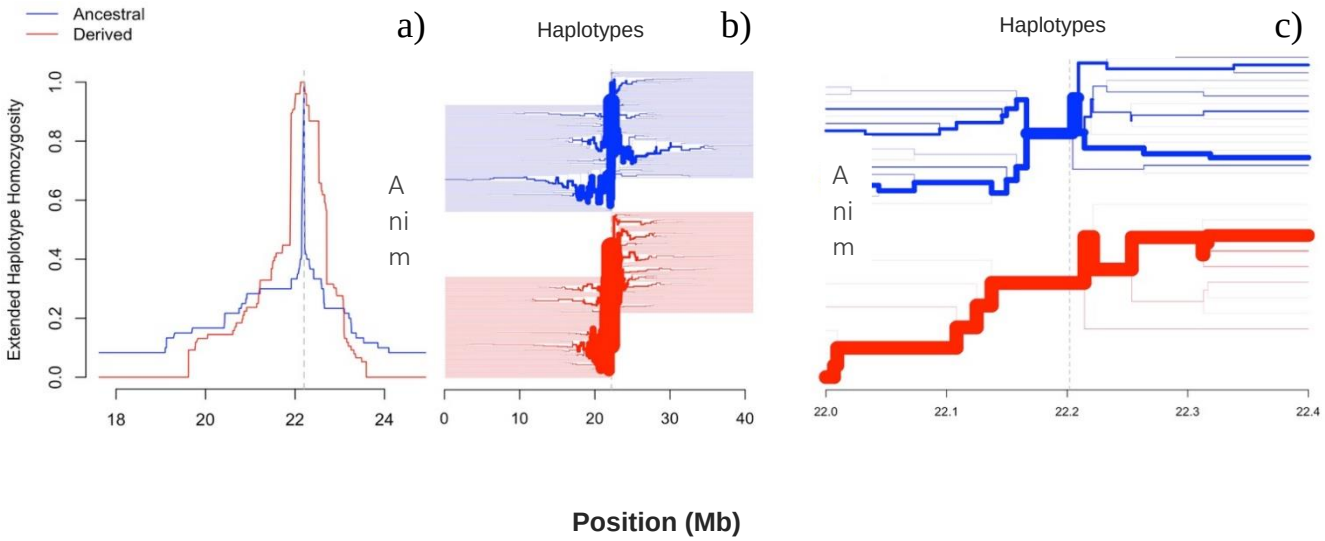


Fig. 3 Extended haplotype homozygosity (EHH) plots and bifurcation diagrams for the AX-103466344 single nucleotide polymorphism (SNP) in Mangalarga Marchador horses. (a) EHH for the focal AX-103466344 SNP. (b) Bifurcation diagrams showing the breakdown of these extended haplotypes from increasing distances by chromosome. (c) Bifurcation diagrams showing the breakdown of these extended haplotypes from increasing distances by chromosome and individual tree. The thickness of each branch in the furcation plot is determined by the haplotype frequency.

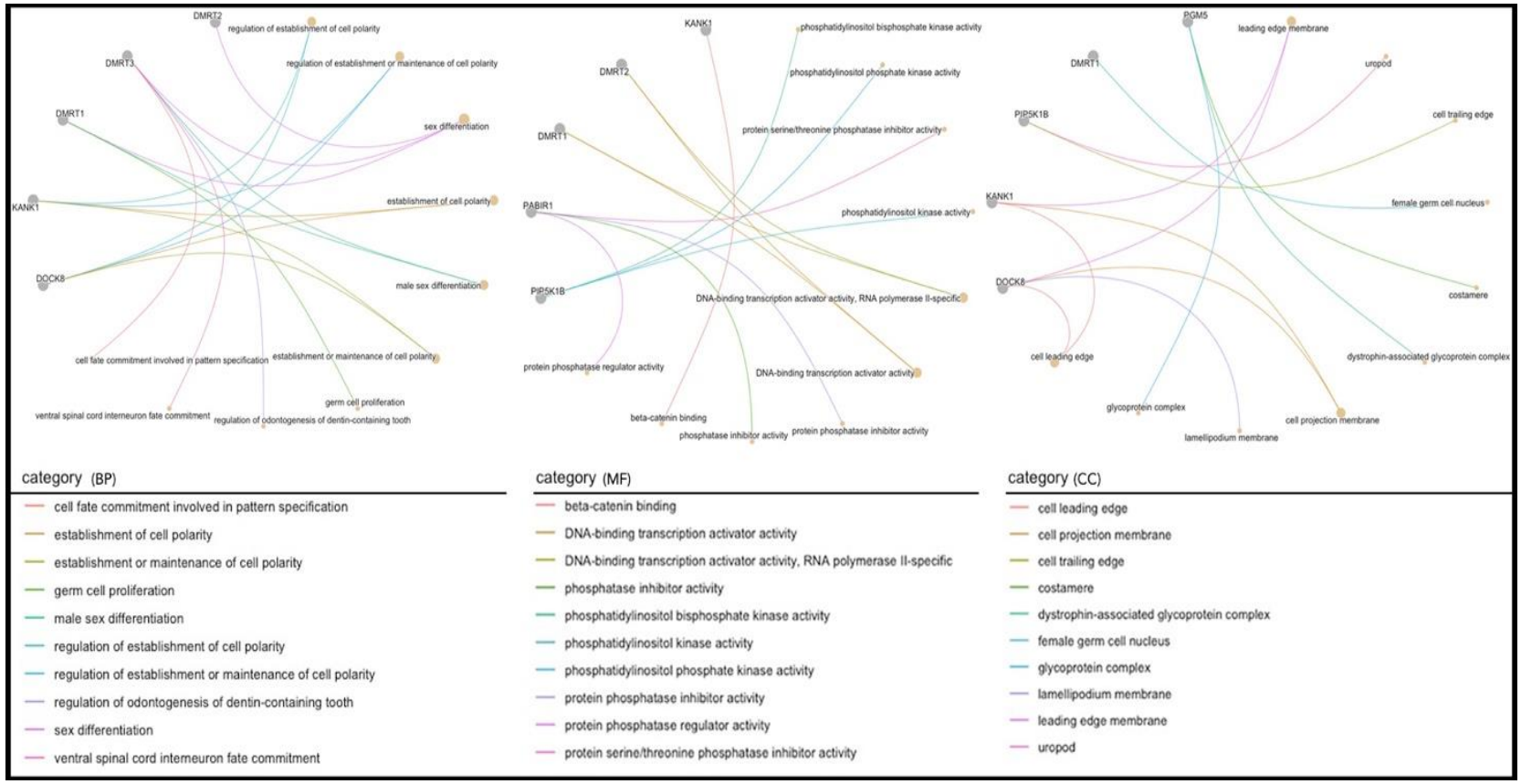


Fig. 4 Enrichment analysis for the top 10 Gene Ontology (GO) terms in biological process (BP), molecular function (MF), and cellular component (CC).

4. DISCUSSION

4.1 Considerations Regarding Population Structure and LD

The findings related to population structure and LD mirror a previous study by Santos et al. (2021), who utilized the same dataset. Their study revealed that the top five eigenvectors accounted for 54.98% of the cumulative genetic variance, with cluster 1 contributing 40.33% of this variance. Furthermore, Santos et al. (2021) observed a decrease in LD as the physical marker distance increased, with r^2 values falling below 0.20 within a 15 kb range. Therefore, the observed gait variations do not appear to be significantly associated with population substructure. This assertion is supported by the dispersion of the dataset and segregations (substructures) was attributed to the importance of sires from different families in the breed formation.

4.2 GWAS, EHH, Annotation, and Gene Enrichment Analysis

We have unveiled new insights into the genetic underpinnings that could potentially underlie gait type differentiations in Mangalarga Marchador horses, as well as in other equine breeds exhibiting similar phenotypes. Candidate genes linked in this specific region of ECA 23 may complement the function of *DMRT3* in shaping the gait phenotype in this breed.

Of the 11 identified SNPs, we noted eight genes (*PIP5K1B*, *PABIR1*, *PGM5*, *DOCK8*, *KANK1*, *DMRT1*, *DMRT3*, and *DMRT2*) in a highly correlated 1.045 Mb window. Among these genes, we expected to find *DMRT3* as it has been identified in several studies associated with the gait phenotype in horses (Andersson et al. 2012; Petersen et al. 2013; Promerová et al. 2013; Filho et al. 2015; Jäderkvist et al. 2015). Promerová et al. (2013) genotyped more than 4000 horses belonging to 141 breeds to investigate the stop codon mutation in the *DMRT3* gene. However, half of these horses were subsequently genotyped for a SNP located 32 kb upstream of the *DMRT3* nonsense mutation, due to its strong association. In our study of Mangalarga Marchador horses, of the 11 SNPs selected, one of them is found 33.8 kb away from *DMRT3*,

revealing a certain similarity. The evidence points to selection signatures in this region, most likely driven by the hitchhiking effect.

Due to the strong correlation and likely hitchhiking effect, we were unable to analyze these regions in a specific manner. Genotyping data, even at a high density in GWAS, has certain limitations. Additionally, it is important to note that we did not include non-gaited breeds in this study due to the lack of available high-density data. Our findings are consistent with earlier studies investigating *DMRT3*, which included several horse breeds. However, it is crucial to emphasize that despite this evidence, we cannot draw a definitive conclusion, as the true causal variant may not precisely correspond to the identified SNPs. These SNPs may only indicate a close relationship and/or correlation with the true causal variants (Schaid et al. 2018).

In total, we identified four SNPs located within the KN motif and ankyrin repeat domains 1 (*KANK1*) gene. We also observed specific SNPs in the double-sex and mab-3 related transcription factor 1 (*DMRT1*) and dedicator of cytokinesis 8 (*DOCK8*) genes. The remaining SNPs were mapped to adjacent regions. The SNP with the greatest impact was in the *KANK1* gene. The KANK1 protein plays a crucial role in regulating cell adhesion, actin polymerization, and cell proliferation in several cell lineages. Furthermore, it has a well-established function in the negative regulation of actin polymerization in the cytoplasm and thus in the disassembly of F-actin, which controls the formation of the cytoskeleton. Therefore, it is a critical regulator of the proliferation and differentiation of muscle progenitor cells (Kakinuma et al. 2009; Nguyen and Lee 2022). Disruption of KANK1 function may have significant implications for tendon/ligament mechanotransduction and collagen elasticity (Paradžik et al. 2020). However, much remains to be discovered about its specific function in horses. Recent studies, such as that by Momen et al. (2022), have suggested that KANK1 plays a vital role in tendon/ligament mechanotransduction in horses.

Kubota et al. (2018) aimed to investigate enhancer factors in the pathogenesis of cerebral palsy; they identified a deletion involving the *KANK1* gene locus in the 9p24.3 region. The chromosomal conformation of mouse neural progenitor and embryonic stem cells suggests a spatial interaction of this element with the *DMRT3* gene promoter,

specific to mab-3, located approximately 120 kb downstream of the *RAR/RXR* gene—a site-important connection. Deletion of the cis-regulatory element of the *DMRT3* gene in humans may have significant implications for forebrain development and the manifestation of locomotion abnormalities, resulting in spastic cerebral palsy. Our findings provide a substantial contribution, as they indicate the possibility of a genetic relationship between *DMRT3* and *KANK1*, providing a new perspective on a candidate enhancer and mechanisms associated with *DMRT3*.

Considering the results from the haplotype-based approach, there is a distinct pattern of extension, with a more pronounced reach toward the derived variant compared with the ancestral one in the region surrounding the peak (22202053 pb). Interestingly, this pattern undergoes a reversal at the peak's base. This finding suggests that the allele reached its current population frequency more quickly than under neutrality, characterizing positive signals of selection in the upper region (Vargas et al. 2020). Moreover, the ancestral allele exhibited a higher concentration of animals with the "Batida" gait, whereas the derived animals displayed the "Picada" gait. Santos et al. (2021) used the runs of homozygosity (ROH) method and obtained evidence of selection signatures on 12 chromosomes, including ECA 23, containing some ROH regions in Mangalarga Marchador horses. The authors included some selection signature methods that revealed genes associated with athletic performance, limb development, and the energy required for muscular activity, including the notable involvement of HOX family genes. This discovery underscores the embryonic stages as the critical period for the formation of the "Batida" and "Picada" gait phenotypes.

In biological processes, the *DMRT1*, double-sex and mab-3 related transcription factor 2 (*DMRT2*), and *DMRT3* genes have an obvious biological relationship in sexual regulation, including sex differentiation (GO:0007548), male sex differentiation (GO:0046661), regionalization (GO:0003002), and pattern specification process (GO:0007389). Even though there is a consensus on the role of these genes in other biological processes, they are not yet included in the functional enrichment. DMRT1 plays a crucial role in differentiation of the testis, including germ cells and somatic cells (Zarkower and Murphy 2022). Researchers recently discovered that *DMRT2* promotes

the endochondral bone formation transition by binding Sox9 and Runx2 (Ono et al. 2021). According to Ono et al. (2021) endochondral bone formation is fundamental for skeletal development as chondrocytes undergo several phases of differentiation and transition, progressing from a proliferative stage to a hypertrophic stage.

KANK1 and *DOCK8* also show a close biological relationship. The most prominent biological process involves the establishment, regulation, and maintenance of cell polarity (GO:0030010; GO:2000114; GO:0032878). While we did not identify links between these two genes in terms of molecular function, we observed enrichment in several cellular components, specifically the leading-edge membrane (GO:0031256), cell projection membrane (GO:0031253), and cell leading edge (GO:0031252). The establishment and maintenance of cell polarity are fundamental for the proper functioning of various cell types (Flasse et al. 2020). KANK1 plays a significant role in shaping the cytoskeleton through its control of actin polymerization (Gee et al. 2015). Furthermore, Capkova et al. (2021) reported findings of duplications in the chromosome 9p24.3 region related to these two genes. This evidence underscores the pathogenic impact of such duplications in patients with certain diseases associated with brain development.

Phosphatidylinositol-4-phosphate 5-kinase type-1 beta (*PIP5K1B*), along with *DMRT1*, demonstrated negative regulation of the cell cycle process (GO:0010948). Furthermore, *PIP5K1B* has been linked to racing performance and success, with evidence suggesting that isoforms of this gene can affect the ability and speed of horses (Velie et al. 2018). On the other hand, the genes PPP2R1A-PPP2R2A-interacting phosphatase regulator 1 (*PABIR1*) and phosphoglucomutase-like protein 5 (*PGM5*) did not demonstrate connections with any biological, molecular, or cellular pathway that interacts with the other six genes identified in the study. However, they are associated with several important pathways and genes linked to the performance and development of horses. These pathways include myofibril assembly (GO:0030239), negative regulation of cell cycle G2/M transition (GO:1902750), negative regulation of mitotic cell cycle G2/M transition (GO:0010972), development of striated muscle cells (GO:0055002), serine/threonine protein phosphatase inhibitory activity (GO:0004865),

phosphatase regulatory activity (GO:0019208; GO:0019888), and complex glycoproteins (GO:0090665), among others. It is important to highlight that these genes are in a conserved region in the horse genome, which leads us to consider the hitchhiking effect (Liu et al. 2014; Bruce and Lynch 2018).

The biological signals we identified pointed to strong evidence that these regions are likely enhancer genes, especially those that showed some relationship with *DMRT3*. The AX-103466344 SNP showed distinct behavior in homozygous conditions: except for two horses, the AA genotype resulted in the “Batida” gait, while the GG genotype resulted in the “Picada” gait. The two horses showing divergent phenotypes should be genotyped again for this locus. We also speculate that the portion of animals heterozygous for this genotype is probably subject to some type of dominance deviation.

In conclusion, there is strong evidence that the “Batida” and “Picada” gait phenotypes in the Mangalarga Marchador may be associated with causal mutations in genes around *DMRT3.* Therefore, new studies that investigate these particularities will be fundamental for the implementation of a genetic improvement program in the Mangalarga Marchador breed, as well as in other gaited horse breeds.

5. REFERENCES

Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ (2010) Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. Journal of Dairy Science 93(2):743-752.

Andersson L, Larhammar M, Memic F. et al. (2012) Mutations in DMRT3 affect locomotion in horses and spinal circuit function in mice. Nature 488, 642–646. doi: 10.1038/nature11399.

Bruce W and Lynch M (2018) Hitchhiking and Selective Sweeps. Evolution and Selection of Quantitative Traits. Oxford, online edn. https://doi.org/10.1093/oso/9780198830870.003.0008.

Capkova Z, Capkova P, Srovnal J, Adamova K, Prochazka M, Hajduch M (2021) Duplication of 9p24.3 in three unrelated patients and their phenotypes, considering

affected genes, and similar recurrent variants. Mol Genet Genomic Med. 9(3):e1592. doi: 10.1002/mgg3.1592.

Filho HCM, Cothran EG, Juras R, Filho MAG, et al. (2015) Alelo DMRT3 mutante em equinos de marcha batida e picada das raças Campolina e Mangalarga Marchador. Ciência Veterinária nos Trópicos 18: 6-11.

Flasse L, Yennek S, Cortijo C, Barandiaran IS, Kraus MR, Grapin-Botton A (2020) Apical Restriction of the Planar Cell Polarity Component VANGL in Pancreatic Ducts Is Required to Maintain Epithelial Integrity. Cell Rep. 31(8):107677. doi: 10.1016/j.celrep.2020.107677.

Gee HY, Zhang F, Ashraf S, Kohl S et al. (2015) KANK deficiency leads to podocyte dysfunction and nephrotic syndrome. J Clin Invest. 125(6):2375-84. doi: 10.1172/JCI79504.

Grillner S (2006) Biological pattern generation: the cellular and computational logic of networks in motion. Neuron. 52(5):751-66. doi: 10.1016/j.neuron.2006.11.008.

Hong CS, Park BY, Saint-Jeannet JP (2007) The function of Dmrt genes in vertebrate development: it is not just about sex. Dev Biol. 310(1):1-9. doi: 10.1016/j.ydbio.2007.07.035.

Jäderkvist K, Johansson L, Mykkänen A, Mäenpää M, et al. (2015) The DMRT3 “Gait keeper” mutation affects harness racing performance and riding traits in Finnhorses. J. Equine Vet. Sci. 35:399.

Kakinuma N, Zhu Y, Wang Y, Roy BC, Kiyama R (2009) Kank proteins: Structure, functions and diseases. Cell Mol. Life Sci. 66:2651-2659.

Kubota N, Yokoyama T, Hoshi N, Suyama M (2018) Identification of a candidate enhancer for DMRT3 involved in spastic cerebral palsy pathogenesis. Biochem Biophys Res Commun. 496(1):133-139. doi: 10.1016/j.bbrc.2018.01.011.

Kullander K (2005) Genetics moving to neuronal networks. Trends Neurosci. 28(5):239-47. doi: 10.1016/j.tins.2005.03.001.

Li YR, Keating BJ (2014) Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. Genome Med 6-91. doi: 10.1186/s13073-014-0091-5.

Liu H, Sørensen AC, Meuwissen TH, Berg P (2014) Allele frequency changes due to hitch-hiking in genomic selection programs. Genet Sel Evol. 46(1):8. doi: 10.1186/1297-9686-46-8.

Misztal S, Tsuruta T, Strabel B, Auvray T, Druet et al. (2002) BLUPF90 and related programs (BGF90). 7th World Congress on Genetics Applied to Livestock Production 19-23.

Misztal I, Tsuruta S, Lourenco D, Aguilar I, Legarra A, Vitezica Z. (2014) Manual for BLUPF90 family of programs. Athens: University of Georgia.

Momen M, Brounts SH, Binversie EE, Sample SJ et al. (2022) Selection signature analyses and genome-wide association reveal genomic hotspot regions that reflect differences between breeds of horse with contrasting risk of degenerative suspensory ligament desmitis. G3 12(10):jkac179. doi: 10.1093/g3journal/jkac179.

Nguyen MT, Lee W (2022) Kank1 Is Essential for Myogenic Differentiation by Regulating Actin Remodeling and Cell Proliferation in C2C12 Progenitor Cells. Cells. 11(13):2030. doi: 10.3390/cells11132030.

Ono K, Hata K, Nakamura E, Ishihara S et al. (2021) Dmrt2 promotes transition of endochondral bone formation by linking Sox9 and Runx2. Commun Biol. 4(1):326. doi: 10.1038/s42003-021-01848-1.

Paradis E (2010) pegas: an R package for population genetics with an integrated–modular approach. Bioinformatics 26:419–420.

Paradis E, Schliep K (2019) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35:526-528.

Paradžik M, Humphries JD, Stojanović N, Nestić D et al. (2020) KANK2 Links αVβ5 Focal Adhesions to Microtubules and Regulates Sensitivity to Microtubule Poisons and Cell Migration. Front Cell Dev Biol. 8:125. doi: 10.3389/fcell.2020.00125.

Petersen JL, Mickelson JR, Rendahl AK et al. (2013) Genome-wide analysis reveals selection for important traits in domestic horse breeds. PLoS Genet. 9(1):e1003211. doi: 10.1371/journal.pgen.1003211.

Promerová M, Andersson LS, Juras R, Penedo MC et al. (2014) Worldwide frequency distribution of the 'Gait keeper' mutation in the DMRT3 gene. Anim Genet. 45(2):274-82. doi: 10.1111/age.12120.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Sham PC (2007). PLINK: a tool set for whole-genome association and population-basedlinkage analyses. American journal of human genetics 81:559–575.

Robilliard JJ, Pfau T, Wilson AM (2007) Gait characterisation and classification in horses. Journal of Experimental Biology 210:187–197.

Sabeti PC, Reich DE, Higgins JM, et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832-837.

Santos WB, Schettini GP, Maiorano AM et al. (2021) Genome-wide scans for signatures of selection in Mangalarga Marchador horses using high-throughput SNP genotyping. BMC Genomics 22:737. doi: 10.1186/s12864-021-08053-8.

Schaid DJ, Chen W, Larson NB (2018) From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat Rev Genet. 19(8):491-504. doi: 10.1038/s41576-018-0016-z.

Smedley D, Haider S, Ballester B, et al. (2009) BioMart – biological queries made easy. BMC Genomics 10:22.

Tang K, Thornton KR, Stoneking M (2007) A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome. Plos Biology 7:e171.

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

Vargas LB, Dourado RM, Amorim LM, Ho B et al. (2020) Single Nucleotide Polymorphism in KIR2DL1 Is Associated With HLA-C Expression in Global Populations. Front Immunol. 11:1881. doi: 10.3389/fimmu.2020.01881.

Velie BD, Fegraeus KJ, Solé M. et al. (2018) A genome-wide association study for harness racing success in the Norwegian-Swedish coldblooded trotter reveals genes for learning and energy metabolism. BMC Genet 19:80. doi: 10.1186/s12863-018-0670-3.

Vincelette A (2023) The Characteristics, Distribution, Function, and Origin of Alternative Lateral Horse Gaits. Animals 13(16):2557. doi: 10.3390/ani13162557.

Volff JN, Zarkower D, Bardwell VJ, Schartl M (2003) Evolutionary dynamics of the DM domain gene family in metazoans. J. Mol. Evol. 57:S241–S249.

Wang H, Misztal I, Aguilar I, Legarra A, Fernando RL et al. (2014) Genome-wide association mapping including phenotypes from relatives without genotypes in a single-step (ssGWAS) for 6-week body weight in broiler chickens. Front. Genet. 5:134. doi: 10.3389/fgene.2014.00134.

Wang H, Misztal I, Aguilar I, Legarra A, Muir WM (2012) Genome-wide association mapping including phenotypes from relatives without genotypes. Genetic Research 94:73-83.

Wang J (2012) On the measurements of genetic differentiation among populations. Genet. Res., 94:275–289.

Wickham H (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag 77:1-3.

Wu T, Hu E, Xu S, Chen M et al. (2021) clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation. 2(3):100141.

Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 88(1):76-82. doi: 10.1016/j.ajhg.2010.11.011.

Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. (2014) Advantages and pitfalls in the application of mixed-model association methods. Nat Genet. 46(2):100-6. doi: 10.1038/ng.2876.

Zarkower D (2001) Establishing sexual dimorphism: conservation amidst diversity?. Nat. Ver. Genet. 2 175–185.

Zhang Z, Ersoz E, Lai CQ. et al. (2010) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42, 355–360. https://doi.org/10.1038/ng.546.

Zarkower D, Murphy MW (2022) DMRT1: An Ancient Sexual Regulator Required for Human Gonadogenesis. Sex Dev. 16(2-3):112-125. doi: 10.1159/000518272.

CHAPTER IV – IMPLICATIONS

Horse genetic improvement programs in Brazil have not yet implemented genomics as an auxiliary resource to identify the best stallions for future generations in their databases. The selection methods currently used in Brazil are based on the performance of animals considered superior that have won competitions in certain categories, as well as their descendants. Factors involving the environment, level of training, judgment and other variables can influence the decision of which animals would be the best. Thus, Breed registries employ only genetic analysis for genealogical recording and to identify genotypes associated with inherited genetic diseases.

In this study, we carried out different analyzes for two breeds of horses in Brazil: Quarter Horse and the Mangalarga Marchador. The investigation of the genomic structure of the Brazilian Quarter Horse (QH) racing line population revealed significant implications for the development and improvement of the breed. Pointing to a moderate-to high-level of genomic inbreeding, regions of homozygosity (ROH) containing crucial genes involved in heme transport proteins, glucose metabolism, cell differentiation and regulation of calcium ions were also identified. This is the first study that investigates heterozygosity-rich regions (ROHet) in the QH breed. The results point to genes associated with respiratory capacity and muscle repair. This can be used for breeding practices that aim to increase regenerative capacity and treat muscular diseases, not limited to the QH breed, but to other breeds of horses.

For the study of the Mangalarga Marchador, candidate genes were located around *DMRT3* in ECA 23. These results provide new information about how gait pattern is possibly influenced in this breed through its genetic mechanisms. The paper looks at the possible roles of genes such as *KANK1, DMRT1* and *DMRT3* in regulating muscle cell differentiation, as well as cell adhesion and actin polymerization. These factors can also alter tendon/ligament mechanotransduction and collagen elasticity, essential for the control of locomotion. Furthermore, the results shed light on a potential genetic association between *DMRT3* and *KANK1*, opening the door for future studies on enhancer genes involved in gait variation in different equine breeds.