UNIVERSIDADE ESTADUAL PAULISTA – UNESP

CAMPUS OF JABOTICABAL

GENOME-WIDE SCAN FOR SELECTION SIGNATURE AND ESTIMATES OF LEVELS OF AUTOZYGOSITY IN MANGALARGA MARCHADOR HORSES

Wellington Bizarria dos Santos
Animal scientist

2020

UNIVERSIDADE ESTADUAL PAULISTA – UNESP

CAMPUS OF JABOTICABAL

GENOME-WIDE SCAN FOR SELECTION SIGNATURE AND ESTIMATES OF LEVELS OF AUTOZYGOSITY IN MANGALARGA MARCHADOR HORSE

Wellington Bizarria dos Santos

Advisor: Prof. Dr. Rogério Abdallah Curi
Co- Advisors: Prof. Dr. Henrique Nunes de Oliveira
Prof. Dr. Guilherme Luis Pereira

Dissertation presented to the Faculdade de Ciências Agrárias e Veterinárias – Unesp, Campus of Jaboticabal, UNESP, to obtain the title of Master in the Postgraduate Program in Genetics and Animal Breeding.

2020

S237g

Santos, Wellington Bizarria dos

Genome-wide scan for selection signature and estimates of levels of autozygosity in Mangalarga Marchador horses / Wellington Bizarria dos Santos. -- Jaboticabal, 2020

88 p. : il., tabs.

Dissertação (mestrado) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal

Orientador: Rogério Abdallah Curi

Coorientador: Henrique Nunes de Oliveira e Guilherme Luis Pereira

1. iHS. 2. ROHet. 3. ROH. 4. FPED. 5. Tajima’s D. I. Título.

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

unesp

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Jaboticabal

Jaboticabal

CERTIFICADO DE APROVAÇÃO

TÍTULO DA DISSERTAÇÃO: GENOME-WIDE SCAN FOR SELECTION SIGNATURE AND ESTIMATES OF LEVELS OF AUTOZYGOSITY IN MANGALARGA MARCHADOR HORSES

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

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

Prof. Dr. ROGERIO ABDALLAH CURI
Departamento de Melhoramento e Nutrição Animal / FMVZ / UNESP - Botucatu

Prof. Dr. LUIS ARTUR LOYOLA CHARDULO
Departamento de Melhoramento e Nutrição Animal / FMVZ/UNESP - Botucatu

Jéssica moraes malheiros
Pós-Doutoranda JÉSSICA MORAES MALHEIROS
EMBRAPA / São Carlos/SP

Jaboticabal, 28 de fevereiro de 2020

AUTHOR CURRICULUM INFORMATION

Wellington Bizarria dos Santos – In 2017 he became a Bachelor of Animal Science from Federal Rural University of Pernambuco (UFRPE). In 2018, he started a Master's degree in Genetics and Animal Breeding at Sao Paulo State University, under the guidance of Prof. Dr. Rogério Abdllah Curi. During his master's degree, the author was a CAPES scholarship (Coordination for the Improvement of Higher Education Personnel), with funding of the respective project from São Paulo Research Foundation – FAPESP.

Not everything that counts can be counted

End not everything that's counted truly counts

Unplug yourself, life does not have proofreading

At least, not yet.

(*Albert Einstein & W.B.S*)

ACKNOWLEDGMENTS

I thank the *Great Architect of the Universe* by the perfect algorithm. Just observe, the math are not so exact, and life is not so random either.

I thank *Dr. Rogério Abdallah* for the opportunity, advice, guidance, dedicated time, and friendship. Again, thanks for not demanding less than my duty, especially for contributing in a unique way to my professional training.

I would like to thank *Dr. Henrique Nunes de Oliveira* for the guidance, advice, friendship, especially for help me in the most critical stages of the learning. He showed me that the force will guide me (Jedi teachings).

I would like to thank *Dr. Luis Artur, Dra. Amanda, Dra. Jessica*, and *Dr. Guilherme* for providing valuable corrections and contributions to this work.

During this journey, some people even not involved with the project helped me. I was extremely grateful with their contributions, and for these reasons, I could not fail to mention it (*Dr. Augusto, Dra. Lígia, Dra. Adriana, Dr. Maurício, Dr. Paulo, Marcus, Dr. Omer*, and *Dra. Josabete*).

I thank my family (Bizarria and Vilela) for support and affection in the most difficult moments (My three mothers *(Cristina, Lília,* end *Iraci),* grandfathers *(Manoel* and *Elizeu), uncle (Edenildo), aunts (Josenilda* and *Joseilda),* nephews *(Larissa* and *João Victor),* and brothers *(IBD/IBS) – Wedja, Wedson, Josiclecia,* and Josicleide). This is also to my friends, those who have been part of my life since graduation degree (Marconi, Jads, *Jacy, Monique, Luh, Luciana, Gabriel, Will, Almir, Diana, Rita, Bruno, Mário, Albério, Dany, Dania, Edy*, *Bia*, and *Julia*), as well as those I did here in São Paulo City (*Zafia, Eduardo & Matheus, Luís, Gustavo, Alejandra, Bruna, Anita, Bia, Cherlyn, Pati, Sabrina, Ivan, Walber,* and *Nedenia*) - Thanks for support, attention, and motivation.

Finally, I would like to thank the *Postgraduate Program in Genetics and Animal Breeding* - Each professor, person, who directly or indirectly made all of this possible.

***

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) - Financial code 2016/19081-9.

CONTENTS

Page

ABSTRACT iii
RESUMO iv
1. CHAPTER 1 – Overall considerations 1
1.1 Introduction 1
1.2 Objectives 2
1.2.1 General objective 2
1.2.2 Specific objectives 2
1.3 Literature Review 2
1.3.1 Mangalarga Marchador horses overview and their status overseas 2
1.3.2 Sequencing and genotyping in horses 4
1.3.3 Linkage disequilibrium in horse genome 6
1.3.4 Inbreeding on population 6
1.3.5 Principles of signature of selection 7
1.3.6 Genomic methodologies for signatures of selection and inbreeding 9
1.3.7 Signatures of selection in horses 13
1.1 References 14
CHAPTER 2 - GENOME-WIDE SCANS FOR SIGNATURES OF SELECTION IN MANGALARGA MARCHADOR HORSES USING HIGH-THROUGHPUT SNP GENOTYPING 21
2.1 Introduction 22
2.2 Methods 24
2.2.1 Ethical statement 24
2.2.2 Sample collection, gait patterns, and DNA extraction 24
2.2.3 Genotype, quality control, filter and phase genotypes 25
2.2.4 Population structure and linkage disequilibrium analyses 26
2.2.5 Genome-Wide scan for signals of positive selection 27
2.2.6 Gene-annotation enrichment analysis 28
2.2.7 Gene network analysis 28
2.3 Results 29
2.3.1 Breed structure 29
2.3.2 Candidate genes identifications 30
2.3.3 Genes and enriched pathways 37

2.3.4 Candidates related to gait, and locomotor system. 38
2.3.5 Candidates related to energy, exercise, and athletic performance. 39
2.3.6 Integrative gene networks. 40
2.4 Discussion 42
2.5 Acknowledgements 45
2.6 References 45
APPENDIX 1 51
CHAPTER 3 - FINE-SCALE ESTIMATION OF INBREEDING RATES, RUNS OF HOMOZYGOSITY AND GENOME-WIDE HETEROZYGOSITY LEVELS IN THE MANGALARGA MARCHADOR HORSE BREED 55
3.1 Introduction 56
3.2 Materials and Methods 58
3.2.1 Ethical statement 58
3.2.2 Sample collection, SNP genotyping and quality control 58
3.2.3 Complementary genotyping quality control 59
3.2.4 Pedigree-based estimates of inbreeding 59
3.2.5 Genomic estimates of inbreeding 59
3.2.6 Candidate gene, pathway and functional analysis 61
3.3 Results 61
3.3.1 Inbreeding levels 61
3.3.2 Assessment of runs of homozygosity and heterozygosity 63
3.3.3 Gene annotation in ROHet islands 66
3.4 Discussion 67
3.5 Conclusion 72
3.6 Acknowledgements 72
3.7 References 72

GENOME-WIDE SCAN FOR SELECTION SIGNATURE AND ESTIMATES OF LEVELS OF AUTOZYGOSITY IN MANGALARGA MARCHADOR HORSE

ABSTRACT - The Brazilian Mangalarga Marchador horse has dominated the attention of many horse breeders for their gaited phenotype over decades. A particularity of this breed is its intermediate-speed gait known as "marcha" in Brazil, which is subdivided into “marcha batida” and “marcha picada”. Considering the few studies focused on the genetic improvement of the breed, with so many potentials for the economy of the country, the objective of this study was to analyze through molecular/genomic information, the occurrences of signatures of selection using different statistics, featuring runs of homozygosity and heterozygosity, as well as accessing the inbreeding levels by measurements of genomics coefficients, and based on pedigree. To achieve these objectives, 192 animals were genotyped using the platform: *Axiom ® Equine Genotyping Array - 670.796 SNP* (Thermo Fisher, USA). To study the selection signatures, three methods already established in the literature were used: Tajima’s D, Runs of Homozygosity (ROH) islands, and Integrated Haplotype Score (iHS). For the evaluation of inbreeding were investigated the characterization of Runs of Homozygosity (ROH) and Heterozygosity (ROHet), and genomics coefficients (F_{HOM}, and F_{ROH}), and those based on Pedigree (F_{PED}) were calculated. As a final result, our findings reveal evidence of signals of selection associated with athletic performance, gait type, and energy muscle activity. The other potential signatures were associated with energy metabolism, bronchodilator response, *NADH* regeneration, reproduction, keratinization, and immunological system. The observed inbreeding coefficients were considered low to moderate. For F_{ROH} the results were considered moderate (0.16). However, its levels were low in the method based on pedigree information (F_{PED}) (0.008), as well as for the genomic method based on the differences between the observed and expected number of homozygous genotypes (F_{HOM}) (0.010). Besides, the correlations between the inbreeding coefficients were also of low to moderate. The availability of high-density SNP chips made it possible to improve estimates of inbreeding coefficients. The calculation of F_{ROH} allowed access to information on the demographic history and genetic relationships in the population based on molecular information, and therefore estimates were higher than those observed in the classical approach.

Keywords: iHS, ROHet, ROH, FPED, Tajima’s D

GENOME-WIDE SCAN PARA ASSINATURA DE SELEÇÃO E ESTIMATIVAS DE NÍVEIS DE AUTOZIGOSIDADE EM CAVALOS MANGALARGA MARCHADOR

RESUMO - Ao longo de décadas o cavalo brasileiro Mangalarga Marchador tem dominado a atenção de inúmeros criadores de cavalos pelo fenótipo marchador. Uma particularidade dessa raça é a marcha de velocidade intermediária, e que se subdivide em marcha batida e marcha picada. Considerando os poucos estudos voltados para o melhoramento genético da raça, com tantos potenciais para a economia do país, o objetivo deste trabalho foi analisar, por meio de informações moleculares/genômicas, as ocorrências de assinaturas de seleção com o uso de diferentes estatísticas, corridas de homozigosidade e heterozigosidade, bem como acessando os níveis de endogamia por medidas de coeficientes genômicos, e com base no pedigree. Para atingir os objetivos, 192 animais foram genotipados com o uso da plataforma: *Axiom ® Equine Genotyping Array* - 670.796 SNP (Thermo Fisher, EUA). Para estudar as assinaturas de seleção, três métodos estabelecidos na literatura foram utilizados: Tajima’ D, corridas de homozigose (“Runs of Homozygosity” - ROH) e escore de integração dos haplótipos (“Integrated Haplotype Score” - iHS). Para a avaliação da endogamia foi investigado as ROH e corridas de heterozigosidade (“Runs of Heterozygosity” - ROHet), além de coeficientes genômicos (F_{HOM}, e F_{ROH}), bem como aqueles baseados no Pedigree (F_{PED}). Como resultado final, revelamos evidências de assinaturas de seleção associadas ao desempenho atlético, tipo de marcha e energia para atividade muscular. As outras assinaturas potenciais foram associadas ao metabolismo energético, resposta ao broncodilatador, regeneração do *NADH*, reprodução, queratinização e sistema imunológico. Os coeficientes de endogamia calculados foram classificados de baixos a moderados. Para F_{ROH} os resultados foram considerados moderados (0,16). No entanto, seus níveis foram baixos no método baseado em informações de pedigree (F_{PED}) (0,008), bem como no método genômico baseado nas diferenças entre o número observado e esperado de genótipos homozigotos (F_{HOM}) (0,010). Além disso, as correlações entre os coeficientes de endogamia também foram de baixas a moderadas. A disponibilidade de chips SNP de alta densidade tornou possível melhorar as estimativas dos coeficientes de endogamia. O cálculo do FROH possibilitou o acesso a informações sobre a história demográfica e as relações genéticas da população com base em informações moleculares e, portanto, as estimativas foram superiores às observadas na abordagem clássica.

Palavras-chave: iHS, ROHet, ROH, FPED, Tajima’s D

1. CHAPTER 1 – Overall considerations

1.1 Introduction

In over two centuries of history, the Mangalarga Marchador (MM) is one of the most populous equine breeds in the Brazilian territory (IBGE, 2017; MAPA, 2016). The National Exhibition of the breed that occurs every year in Belo Horizonte/MG, celebrated the 70th anniversary of the Brazilian Association of Mangalarga Marchador Horse Breeders (ABCCMM) in 2019, which represents the largest exhibition of horses in Latin America. Besides, the Nacional Exhibition is also the largest private event in Belo Horizonte. Currently, 16,000 breeders are associated, and over 600,000 horses were registered, comprising 70 centers in several states of Brazil and abroad, being the US market the most promising (ABCCMM, 2019).

Batida and picada gait type are the main trait of the breed that represents the unique natural movements allowed in the MM for intermediate speeds (Andrade, 2016). In the batida gait, the diagonal supports are more frequent than the lateral and triple, differing from the picada gait (Beck, 1992; USMMA, 2018). Briefly, during the execution of the gait in each modality, the judgment must evaluate the animal's potential as a gaited horse, and its differential in terms of comfort, regularity, performance, and training.

Since the publication of Andersson et al. (2012) explaining the *DMRT3* gene (first gene mapped and associated with gait), which some studies sought to classify and elucidate discoveries about equine locomotor performance. For example, signature studies promise to provide the necessary resolution to identify important genomic regions without the need for phenotypic measurement, and even finer discoveries about traits of importance in several horse's breeds. According to Gurgul et al. (2018), detection of the signature of selections in genomic regions support a direct insight into the mechanism of artificial selection and allow further disclosure of the candidate genes related to the animals' phenotypic variation.

Another way to study important regions in the genome of a given species/breed is to access inbreeding levels. Region-specific stretches can be used to more effectively manage areas of low genetic diversity (homozygous regions), that results in the reduction of the performance across economically important traits (Howard et al., 2017). Also, there are heterozygosity-rich regions that have been used to study balancing or negative selection, introgression, admixture or hypervariable regions (Marras et al., 2018), which is a complementary approach to understanding the genetic basis of many complex traits. With the sequencing of the equine genome, completed in 2009, and consequently the availability of large-scale DNA marker genotyping platforms, these types of genetic studies mentioned above have become more frequent and more accurate.

1.2 Objectives

1.2.1 General objective

Assessment of genomic regions of economic importance for the Mangalarga Marchador breed, especially those associated with recent positive selection, also accessing the inbreeding levels.

1.2.2 Specific objectives

i. Analyze the occurrence of recent signatures of selection (natural or artificial), through different and established technics;

ii. Assessment and characterization of inbreeding levels, as well as homozygous and heterozygous segments present in the Mangalarga Marchador breed;

iii. Identify important genomic regions for gait type and quality, such as muscular and skeletal structure (locomotor system), conformation, and temperament.

1.3 Literature Review

1.3.1 Mangalarga Marchador horses overview and their status overseas

It was in southern Minas Gerais State, which the first Mangalarga Marchador horses (MM) originated, specifically in the city of Cruzília (ABCCMM, 2017). Composing one of the four gaited breeds in Brazil, the MM has become popular throughout the national territory and more recently has been standing out in the world (IBGE, 2017; MAPA, 2016).

During the breed formation is known that major bases contributions descended from the genetic infusion of Portuguese horses, as well as those originating from Spain, Holland, France, and Germany; being the founding breeds of Latin American troops (Araújo, 2013; Edwards, 1994). However, Beck (1992) confirmed a deeper and complex discussion on the breed formation. The author showed that Alter and Andalusian horse contributions are purely aesthetics, and that possibly the Berber horses would have promoted the main bases of the naturally gaited breed. Moreover, there is evidences that other breeds are yet unknown in the historical issues and also by lack of accurate information.

Concerning conformation, MM horses are easier for identification by some prominent traits. They have a trapezoidal head with a straight profile and roundness in the nose, a softly convex bevel profile and vivid eyes, deep thorax, medium to small ears, medium length neck, short loin-back, muscular back, and thin and diverse coat (IMH, 2018). All animals are inspected by the official ABCCMM visual assessment technicians, and only after approval of registration are given permanent breed registration in the book (ABCCMM, 2017). Besides that, the animals have calm and docile temperament.

Up to date (December 2019), there are over 16.000 breeders associated, and over 600,000 registered horses. The Brazilian Association of Mangalarga Marchador Horse Breeders (ABCCMM) comprises 70 centers in several states of Brazil and abroad, being the US market the most promising. Nevertheless, MM horses are present in Europe and some South American countries. For all its expression, on May 19, 2014 it was officially declared a National Brazilian Breed by Law nº 12,975.

Already described in its name, the gait is the most outstanding natural movement in the breed, which is divided into two modalities: "batida" and "picada" gait. In batida gait, the diagonal supports are more frequent than the lateral and triple (determining factor for segregation of the two modalities), with eight supports, four tripodal (anterior left, posterior right, anterior right, and posterior left), two diagonal (left and right) and two lateral (left and right). During execution, the speed ranges from 7-14 km/h, showing softness due low-friction in the vertical direction and virtually none in the lateral direction (Beck, 1992; USMMA, 2018).

1.3.2 Sequencing and genotyping in horses

Amid the genomic rise, molecular markers were defined as variations in the expression of a gene or DNA sequence with a known location on the chromosome, with quantification and traceability in the population and which could be associated with a particular gene or trait of interest (Hayward et al., 2015). Thus, studies have been elucidated along with advances in sequencing and markers genotyping (Eggen, 2012).

As genome sequences were unraveled, a large number of polymorphisms/markers units were found in the comparison of the corresponding segments, with approximately 600-1000 bp (Dunston et al., 2014). With the sequencing of the domestic horse genome (*Equus caballus*), scientists were able to conduct the first step in the search for genomic resolution of the species, which is the main tool for investigating diseases in horses today. The study was conducted from a Thoroughbred mare, with a map of horse genetic variation using DNA samples with a variety of modern and ancestral breeds, including the Andalusian, Arabic, Icelandic, Quarter Horse, Standardbred, and Akel-Teke. This first analysis of the horse genome identified one million SNPs, generating a broad view of genetic variability with the potential to identify contributions to physical and behavioral differences as well as disease susceptibility (NIH, 2007).

The horse genome is relatively repetitive with little segmental duplication, it has 64 chromosomes (2n = 64; 31 pairs of autosomes and one pair of sex chromosomes), of which 13 pairs are meta and subcentric autosomes, and 18 are acrocentric, both sex chromosomes (X and Y) are subcentric (Chowdhary and Raudsepp, 2006). By complete *E. caballus* genome database assembling and depositing, it was possible to conduct new approaches that made it possible to build the version 2.0 with 2.33 Gb. Thus, the reference genome (EquCab2) was published in 2009 and, since 2014, the scientific community has refined the reference sequence, drawing on the basis of EquCab2 and incorporating new reading data (short and long), being the ideal condition for genome assembly (Michael et al., 2018).

EquCab3.0 was launched in 2018 by the University of Louisville/USA, consisting of 10,987 contigs mounted on 4,701 N50 length scaffolds, among 1,502,753 contigs and 87,230,776 scaffolds. At the completion of the genome, 2,506,966,135 bp, 21,559 coding genes, 9,383 non-coding genes, 273 pseudogenes, 56,546 transcripts, 21,198,236 small variants, 193,747 structural variants (EMBL-EBI, 2019) were read. Thus, EquCab3 is currently more accurate by reducing gaps, resulting in the smallest drop in alignment coverage as well as contiguity, which has also been improved by almost 40 times (Kalbfleisch et al., 2018). Jagannathan et al. (2018) using the latest version of the reference genome in a population of 88 horses of several breeds identified approximately 23.5 million SNPs and 2.3 million InDels variants, and on average each genome carried 5.7 million SNPs and 0.8 million InDels relative to the reference genome assembly.

As new technologies were implemented, the feasibility and search for sequencing in the studies increased mainly due to significant cost savings. However, even with decreasing values, sequencing is not a low-cost technology when it comes to populations with thousands of individuals. Thus, genotyping, product of SNP chip automation, has been widely used. Such panels can scan and capture selection-modified genomic regions without the need for phenotyping (selection signatures) (Hancock et al., 2008), identify genes and SNPs associated with traits of economic interest (GWAS) (Koellinger et al., 2010), and many other parallel studies.

One of the first horse genotyping chips, the Equine SNP50 BeadChip from Illumina (Illumina Inc., USA), had 54,602 SNPs evenly distributed throughout the genome. By the second generation, there were already 65,157 SNPs, of which 19,000 are new markers. And finally the highest density panel, the Axiom® Equine Genotyping Array (Thermo Fisher, USA), which proposed something well beyond Illumina standards with a density of 670,796 SNPs. For the different chips mentioned, there is the possibility of imputation, with continuity of existing projects.

Estimates suggest that ~ 100,000 SNPs are sufficient for mapping and horse genomic association studies for all breeds (Wade et al., 2009), but these conditions prove to be a limitation for signatures of selection studies. In other words, high SNP densities are required regardless of the methodology used.

The populational history formation of the horses led to haplotype sharing with increasing in the mapping viability. Also, the various horse genome mapping projects will still provide more contributions in the coming years to identify QTLs, genes, and causal variants related to morphology, immunology, and metabolism. Thus, it will benefit the Animal Breeding Programs, nutrition, and health, as well as human studies by the genomic synteny.

1.3.3 Linkage disequilibrium in horse genome

Linkage disequilibrium (LD) is statistically defined as the non-random association of alleles in two or more loci (Slatkin, 2008). The most commonly used measures to evaluate DL among biallelic markers are D' (LEWONTIN, 1964) and r^2 (HILL and ROBERTSON, 1968), each with different statistical properties (BOHMANOVA et al., 2010). The r^2 represents the correlation between two loci and proved to be the most appropriate measure for LD estimation between biallelic markers (ZHU and ZHAO, 2007; BOHMANOVA et al., 2010). Sargolzaei et al. (2008) pointed out that to be successful in their applications, it is necessary to be based on the relationship between the extent of this imbalance and the density (coverage) of the markers used. Thus, LD maps are fundamental tools for exploring the genetic basis of economically important traits in population studies.

According to Wade et al. (2009), the equine genome has intermediate LD when compared to other species. Besides, characteristics for LD is a long-range haplotype sharing among equine breeds, and of these LDs, the longest was found in the Thoroughbred horses, whose LD resembling the canine genome, being 5x larger than in the human genome. The shorter LDs found were in ancestral horse breeds, while the other breeds presented average values.

1.3.4 Inbreeding on population

The inbreeding rates are accelerating in most species of economic interest. Inbreeding is quantified in many ways, defined as the probability of autozygosity (the expectation that a random individual from the population is autozygous at a random *locus*) (Aulchenko, 2011). The coefficient of inbreeding is closely related to the coefficient of kinship, defined earlier for a pair of individuals as the probability that two alleles sampled at random from these individuals are identical by descent (IBD). The fast-genetic progress has accumulated inbreeding through strong impacts on some selected individuals or families. Economic losses caused by inbreeding depression in production, growth, health, and fertility that are a serious concern (Weigel, 2001).

Falconer and Mackay (1996) reported that increased inbreeding promotes reduced genetic variability, which consequently reduces heterozygosity over many *loci*.

Currently, inbreeding coefficients have been studied basically in two distinct ways. The classical approach that is based on pedigree data (F_{PED}), representing an infinitesimal model with distributed autozygosity across the genome (Wray et al., 1990). This approach neglects the inbreeding stochastic variation, as well as the recombination rates, reliability, limited pedigree knowledge, etc. According to Leory (2011) the recent developments of genomic, and other 'omics' approaches provided the estimation for understanding and managing of inbreeding depression in populations. First, because genomic estimators of inbreeding do not suffer from drawbacks inherent to genealogical tools (reliability, limited pedigree knowledge, assumption that founders are unrelated, among others). Second, the genealogical approaches generally do not take into account the stochastic nature of recombination. The genomic inbreeding coefficient based on runs of homozygosity (F_{ROH}) reflects the realized autozygosity. Also, it is possible to make partitions by chromosomes and under chromosomal segments (McQuillan et al., 2008; Curik et al., 2014).

Most articles published in the last five years has been included different approaches for calculating inbreeding coefficients, and their correlations, seeking to extract as much information as possible. Those most often approaches compared are F_{PED}, F_{ROH}, based on differences between the observed and the expected number of homozygous genotypes (F_{HOM}) (Purcell et al., 2007), SNP-by-SNP inbreeding is based on the increasing frequency of homozygous genotypes, that including IBD and identical by state (IBS) alleles (F_{SNP})(Leutenegger et al., 2003), and genomic inbreeding calculated from a Genomic relationship matrix (G) (F_{GRM}) (VanRaden, 2011).

1.3.5 Principles of signature of selection

Adaptive evolution in domestic animals has been extensively studied either by contributions associated with evolutionary success, or by the improvement, maintenance, and prospecting of these species. During evolution, a series of demographic events increased the complexity of detecting modified genomic regions due to different selective pressures (Ma et al., 2015). These regions are formed from adjacent *loci* (haplotypes) that are in binding imbalance with a particularly favorable mutation (Goddard et al., 2009; Kim et al., 2004). Based on this, different statistical methods were developed to interrogate large data sets, in which each method has particularities (Horscroft et al., 2018; Vitti et al., 2013).

Different perspectives are used to study signals of selection into macro and microevolutionary context (Vitti et al. 2013). However, in order to find signatures within each population, it was decided to emphasize only the microevolutionary scenario with methods based in **(a)** allele frequency, **(b)** linkage imbalance, and **(c)** population differentiation:

a) Uses informational evidence based on the allelic frequency and segregation sites to represent selective sweeps. Mutations should reach high prevalence with close variants, as alleles are derived with a high-frequency corresponding to a sudden loss of genetic variation in neutral loci when a new favorable allele is fixed or initially still in low-frequency (rare allele excess) (Vitti et al., 2013; Smith and Haigh, 2009). Thus, frequency distributions can be derived analytically, so it is possible to obtain statistical simulations that have desirable properties with zero expectation and known variance (McVean, 2002).

b) Detects selective sweeps in genomic regions of high prevalence in the population (mapped by markers and adjacent *loci* associations). It is noteworthy that the combination of a group of alleles with adjacent *loci* forms the haplotypic blocks, which are chromosomal regions in high binding imbalance. Such allelic associations in contrast to individual polymorphisms show how important haplotypes represent a majority of population studies (Crawford and Nickerson, 2005). Remember that structurally haplotypes are influenced by several genetic factors (recombination, selective forces, demographic, etc.). Therefore, even before testing a null model, it is recommended to capture accurate estimates of recombination rates in the population (McVean, 2002). Thus, rather than relying only on allelic frequencies, the method in question uses haplotype data and associations along each chromosome.

c) Selection acts on one particular allele in the population, but not another, forming marked divergences in allele frequency to the population level. This segregational effect concerning neutral alleles (those not selected) stands out against the differentiation between populations (Vitti et al., 2013). According to Nielsen (2005) when a *locus* shows extraordinary levels of genetic differentiation on population compared to other *loci*, signals can be captured and interpreted as evidence for positive selection.

Evolutionary adaptation is a surprising process to conduce genomic patterns to selective sweeps (soft sweeps and hard sweeps), multiple adaptive alleles are carried across populations either because the alleles were present as standing variation or even by the formation of recurrent mutations (*de novo* mutations) (Messer e Petrov, 2013). Thus, soft sweeps are scans characterized by adaptive allelic multiplicity inthe *locus*, the result of genetic additive variation resulting from the positive selection or multiple mutations driven simultaneously during selective scanning. On the other hand, hard sweeps, are mutations that rapidly increase in frequency to fixation, eliminating variation in interconnected sites as they propagate, that is, a single adaptive allele characterizes the sweep on population (Pritchard et al., 2010). Additionally, it is important to highlight that the formation of these signatures also may be intermediate or incomplete processes, defined as partial sweeps, which initially increases the allelic frequency, but which has not yet reached fixation. Those are possibly attributed to genomic signatures that are still underway or that have some selective advantage at first, although later its favorability decays (Pritchard et al., 2010).

1.3.6 Genomic methodologies for signatures of selection and inbreeding

The identification of genomic regions modified by recent positive selection has provided current information on the adaptive course of the species, constituting the main theoretical and applied evolutionary studies (Gouveia et al., 2014). There are currently around 20 methods for studying selection signatures, where about half of them are only within one population approach. Thus, we decided to highlight only those with an emphasis on the latest signs, as well as those with a connection to upcoming chapters.

Tajima’s D – The statistics presented by Tajima (1989), estimates the comparison of nucleotide diversity from observing polymorphic sites in a given set of chromosomes against nucleotide diversity estimated from the allele frequency of polymorphic sites (Carlson et al., 2005). Thus, the comparison of Tajima's D estimates is based on two parameters: the number of segregation sites (*s*) and nucleotide diversity by the mean parameter difference (π). Recalling that both parameters are equal to θ under the hypothesis of neutrality (*θ = 4Nμ*). The *4Ne* represents the effective population size (diploid), and *μ* the mutation rate per generation (Berwick, 2005). Thus, the expected value of Tajima's D under neutrality is zero, positive values bring heterozygous advantages with a dynamic retraction (balancing), while negative values indicate selection of specific alleles concerning alternative alleles with an increase in the population size (positive selection) (Nei and Kumar, 2000; Carlson et al., 2005; Omori et al., 2017). The formula that calculates Tajima's D was presented below:

$$Tajima's\ D = \frac{\pi - s/a1}{\sqrt{V}}$$

Assuming the respective conditions of the neutral theory model (diploid DNA) - the population is in constant equilibrium size.

$$E[\pi] = \theta = E\left[\frac{s}{\sum_{i=1}^{n-1}\frac{1}{i}}\right] = 4N\mu$$

The π and S/a1 both estimate θ (under the null hypothesis), being roughly equal in value.

$$a1 = \sum_{i=1}^{n-1}\frac{1}{i}$$

The Tajima's D sampling variance is calculated for this case as $V = Var[\pi - S/a1]$.

Wherein:

S = number of segregating sites;

n = number of samples;

N = effective population size;

μ = mutation rate at the examined genomic locus;
i = index of summation.

Extended Haplotype Homozygosity (EHH) – Defined as the probability that two random homologs chromosomes carry the same core SNP variation (used to define allelic classes) in identical by descent (IBD) status around under a certain distance (Gautier et al, 2012). The approach is capable of detecting positive selection signatures in the genome without any prior knowledge of specific variants, or selective advantages. Common alleles (haplotypes) are ancestral and have short-range LD; on the other hand, rare alleles (ancestor or derivative) may have short or long-range LD. Given this, a recent selection signature can be characterized in this method by suddenly increasing allelic frequencies in such a short time that the recombination will not be able to break the haplotype with the mutation (Sabeti et al., 2002). Following is an adaptation of Gautier et al. (2017) for the calculation of EHH in the R package (rehh v.2.0).

$$\mathrm{EHH}a_s, t = \frac{1}{n_{a_s}\left(n_{a_s} - 1\right)} \sum_{k=1}^{k_{a_s,t}} n_k(n_k - 1)$$

Wherein:
a = core allele (ancestral and derived);
s = focal SNP;
t = 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 a_S;
n_k = observed count for the kth haplotype.
n_a = gives the total number of haplotypes carrying the core allele a_S.

Integrated Haplotype Score (iHS) – The iHS method represents an improvement of EHH in recent selection signature studies, providing more accurate values by reducing the biases attributed to the influence of demographic history on population (Gautier and Vitalis, 2012). Calculation is constructed by integrating the EHH and is truncated if the EHH value reaches a certain limit (p <0.05). In the package R rehh v.2.04 the integration of *iHH* is calculated by the trapezoidal method, adding both the SNP core offset directions. The classification for *iHH*$_A$ or *iHH*$_D$ will depend on the relationship computed to the ancestral or derived core allele, respectively (Voight et al., 2006). For an extreme negative iHS score (iHS <-2), the derived allele haplotypes should be longer compared to the ancestral allele haplotypes. Already an extreme positive iHS score (iHS> 2) the ancestral allele may be associated with the hitchhiking effect along with the selected allele, or even the ancestral allele itself is the target of selection.

$$Standardized\ iHS = ln\left(\frac{iHH_A}{iHH_D}\right)$$

The iHS is standardized using the mean and standard deviation values in all SNPs with similar allele frequencies, as such measurements within the population have low power reliability when the selected allele frequency is high (Gautier and Vitalis, 2012; Tang et al. , 2007).

$$iHS = \left(\frac{ln\left(\frac{iHH_A}{iHH_D}\right) - E_p\left[ln\left(\frac{iHH_A}{iHH_D}\right)\right]}{SD_p\left[ln\left(\frac{iHH_A}{iHH_D}\right)\right]}\right)$$

Runs of Homozygosity (ROH) – The ROH are continuous homozygous segments of the DNA sequence. Such genomic regions arise when identical haplotypes are inherited from each parent and thus a long-range of genotypes is homozygous (Ceballos et al., 2018). The number and extent of ROH reflect evolutionary history as well as individual and population demographic history, while the homozygous load can be used to investigate the genetic architecture of several traits, mapping selection signatures, helping to minimize the inbreeding rate, and exposing deleterious variants in the genome. (Ceballos et al., 2018; Peripolli et al., 2017; Peripolli et al., 2018). Howrigan et al. (2011) classified segments of ROH in approximately 10 cM (~ 10000kb) as evidence of recent inbreeding (~ 5 generations), while shorter segments 1cM (~ 1000kb), indicate ancestral effect (50 generations). Studies have been calculating molecular homozygosity (FHOM) and inbreeding from runs of homozygosity (FROH) by the following formulas:

$$\mathrm{FHOM} = \frac{\text{Observed Homozygosis} - \text{Expected Homozygosis}}{\text{Observations} - \text{Expected Homozygosis}}$$

$$\mathrm{FROH} = \frac{\sum_{\mathrm{k}} \text{Length } (\mathrm{ROH}_{\mathrm{k}})}{\mathrm{L}}$$

Wherein:

k = Number of each individual's ROH multiplied by the average length of ROHs;
L = Total length of the genome.

Although this method is used to calculate genomic inbreeding coefficients, it is also a selection signature method. Where top ROH islands are strong candidate regions for signals of recent selection (Ceballos et al., 2018; Peripolli et al., 2018).

1.3.7 Signatures of selection in horses

The scarcity of information over selection response in certain populations is a challenge, especially when it comes to populations formed from a broader genetic base or that have suffered specific environmental conditions. Thus, with the identification of genomic regions under positive selection, and genome annotation methods, several QTLs were corroborated.

In horses, since domestication, selective pressures on the genome have been directed to work in agriculture, transport, and war. Only recently, traits as morphology and performance have been introduced more properly (Metzger et al., 2015; Petersen et al., 2013). According to Gurgul et al. (2019), genetic differentiation of the present horse population was evolutionarily created by natural and artificial selection, shaping genomes individually over time with unique traits.

Given this, we have a variety of studies that sought to identify such signatures of selection in horses, and in the most varied breeds. Comparative analyzes involving homozygous extensions were explored in the Sorraia, Dülmen Horse, Arabian, Saxon-

Thuringian Heavy Warmblood, Thoroughbred and Hanoverian, allowing the detection of selected genomic regions within and outbreeds. Where three consensus excerpts were mapped in the *KITLG* region (*KIT ligand*) known to act in processes such as melanogenesis, hematopoiesis, and gametogenesis (Metzger et al., 2015); to the German warmblood breeds through iHS signals, and ROH was identified as potentials candidate and pathways genes for muscle functionality (*TPM1, TMOD2-3, MYO5A,* and *MYO5C*), energy metabolism and growth (*AEBP1, RALGAPA2, IGFBP1, IGFBP3-4*), embryonic development (HOXB-complex) and fertility (*THEGL, ZPBP1-2, TEX14, ZP1, SUN3,* and *CFAP61*) (Nolte et al., 2019). Already in Quarter Horse population, genes associated with muscle and skeletal growth, energy muscle metabolism, as well as cardiovascular and nervous system (*FKTN, INSR, GYS1, CLCN1, MYLK, SYK, ANG,* and *HTR2B*); positive selection signals in more than 30 horse breeds indicated hereditary mutation in the glycogen synthase gene (*GYS1*) by skeletal muscle glycogen excess and polysaccharide storage myopathy (McCoy et al., 2014); and in Swedish Warmblood horses, breed with excellent gaits and/or jumping ability, several genes were identified related to behaviour, physical abilities and fertility, which appear to be targets of selection, located on ECA4, ECA6, ECA7, ECA10 and ECA17 (Ablondi et al., 2019).

1.1 References

Ablondi M, Viklund Å, Lindgren G, Eriksson S, Mikko S (2019) Signatures of selection in the genome of Swedish warmblood horses selected for sport performance. **BMC Genomics** 20:2-12.

Araújo N (2013) Cavalos marchadores brasileiros - Globo Rural. Available at: <http://globotv.globo.com/rede-globo/globo-rural/>. Accessed: 25th Feb. 2016.

Associação Brasileira dos Criadores do Cavalo Mangalarga Marchador - ABCCMM (2017) Available at: <http://www.abccmm.org.br>. Accessed: 23th Nov. 2017.

Associação Brasileira dos Criadores do Cavalo Mangalarga Marchador - ABCCMM (2019) Available at: <http://www.abccmm.org.br/projeto/nacional2019>. Accessed: 02th Dec. 2019.

Aulchenko YS (2011) **Effects of Population Structure in Genome-wide Association Studies.** Netherlands: Elsevier Inc., p.129.

Beck SL (1992) **Mangalarga Marchador, caracterização, história, seleção**. Brasília: edição dos autores, p. 333.

Beeson SK, Schaefer RJ, Mason VC, McCue ME (2019) Robust remapping of equine SNP array coordinates to EquCab3.0. **Animal Genetics** 50:114–5.

Berwick R (2005) Calculating Tajima's D. **Division of Health Sciences and Technology** 47-49.

Bohmanova J, Sargolzaei M, Schenkel FS (2010) Characteristics of link age disequilibrium in North American Holsteins. **BMC Genomics** 11:421.

Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, Rieder MJ, Nickerson DA (2005) Genomic regions exhibiting positive selection identified from dense genotype data. **Genome Res** 15:1553–1565.

Ceballos FC, Joshi PK, Clark DW, Ramsay M, Wilson JF (2018) Runs of homozygosity: windows into population history and trait architecture. **Nature Reviews Genetics** 19:220–234.

Chowdhary BP, Raudsepp T (2006) The Horse Genome. **Genome Dyn.** 2:97-110. Crawford DC, Nickerson DA (2005) Definition and clinical importance of haplotypes. **Annu Rev Med** 56:303-20.

Curik I, Ferenčaković M, Sölkner J (2014) Inbreeding and runs of homozygosity: a possible solution to an old problem. **Livestock Science** 166:26-34.

Dunston GM, Mason TE, et al. (2014) Single Nucleotide Polymorphisms: A Window into the Informatics of the Living Genome. **Advances in bioscience and biotechnology** 5:623-626.

Eggen A (2012) The development and application of genomic selection as a new breeding paradigm. **Animal Frontiers** 2:10-15.

Edwards EH (1994) **Cavalos: um guia ilustrado com mais de 100 raças de cavalos de todo o mundo**. Brasília: Ediouro S.A. p. 256.

Falconer DS, and Mackay TFC (1996) **Introduction to quantitative genetics. 4th edition.** Longman Scientific and Technical, Harlow, UK.

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

Gautier M, Vitalis R (2012) rehh: An R package to detect footprints of selection in genome-wide SNP data from haplotype structure. **Bioinformatics** 28:1176-1177.

Goddard ME, Wray NR, Verbyla K, Visscher PM (2009) Estimating effects and making predictions from genome-wide marker data. **Statistical Science** 24:517-529.

Gouveia JJS, Silva MVGB, Paiva SR, Oliveira SMP (2014) Identification of selection signatures in livestock species. **Genet. Mol. Biol.** 37:330-342.

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.

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.

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 1-312.

Hill WG, Robertson A (1968) Link age disequilibrium in finite populations. **Theoretical and Applied Genetics** 38:226-231.

Horscroft C, Ennis S, Pengelly RJ, Sluckin TJ, Collins A (2018) Sequencing era methods for identifying signatures of selection in the genome. **Briefings in Bioinformatics** 1-12.

Howard JT, Pryce JE, Baes C, Maltecca C (2017) Inbreeding in the genomics era: inbreaeding, inbreeding depression, and management of genomic variability. **J Dairy Sci.** 100:6009–24.

Howrigan DP, Simonson MA, Keller MC (2011) Detecting autozygosity through runs of homozygosity: a comparison of three autozygosity detection algorithms. **BMC Genomics** 12:460.

Instituto Brasileiro de Geografia e Estatística - IBGE (2017) Resultado dos Dados Preliminares do Censo – 2017. Available at: <https://www.ibge.gov.br>. Accessed: 3th Jan. 2019.

International Museum of the Horse – IMH (2018). Available at: <http://imh.org/index-2.html>. Accessed: 15th Jan. 2019.

Jagannathan V, Gerber V, Rieder S, Tetens J, Thaller G, Drögemüller C, Leeb T (2018) Comprehensive characterization of horse genome variation by whole-genome sequencing of 88 horses. **Animal Genetics** EarlyView. DOI 10.1111/age.12753.

Kalbfleisch TS, Rice E, et al. (2018) EquCab3, an updated reference genome for the domestic horse. **bioRxiv** doi: https://doi.org/10.1101/306928.

Kim Y, Nielsen R (2004) Linkage disequilibrium as a signature selection sweeps. **Genetics** 167:1513-1524.

Leutenegger A, Prum B, Génin E, Verny C, Lemainque A, Clerget-Darpoux F, Thompson EA (2003) Estimation of the Inbreeding Coefficient through Use of Genomic Data. **Am. J. Hum. Genet.** 73:516–523.

Lewontin RC (1964) The interaction of selection and link age. I. General consideration sheterotic models. **Genetics** 49:49-67.

Marras G, Wood B, Makanjuola B, Malchiodi F,Peeters K, As P, Baes C, Biscarini F (2018) Characterization of runs of homozygosity and heterozygosity-rich regions in a commercial turkey (Meleagris gallopavo) population. **Conference: 11th World Congress on Genetics Applied to Livestock Production**, New Zealand: WCGALP, p. 1-5.

Ma Y, Ding X, Qanbari S, Weigend S, Zhang, Simianer H (2015) Properties of different selection signature statistics and a new strategy for combining them. **Heredity** 115:426-436.

McCoy AM, Schaefer R, Petersen JL, Morrell PL, Slamka MA, Mickelson JR, Valberg SJ, McCue ME (2013) Evidence of Positive Selection for a Glycogen Synthase (GYS1) Mutation in Domestic Horse Populations. **Journal of Heredity** 105:163–172.

McQuillan R, Leutenegger A, Abdel-Rahman R, et al. (2008) Runs of homozygosity in European populations. **American Journal of Human Genetics** 83:359-372.

McVean G (2009) Natural selection: Department of Statistics at University of Oxford. Available at: <http://www.stats.ox.ac.uk/~mcvean/L4notes.pdf>. Accessed: 28 Feb. 2019.

Meira CT, Curi RA, Farah MM, Oliveira HN, Béltran NA, Silva JAIIV, Mota MDS (2014) Prospection of genomic regions divergently selected in racing line of Quarter Horses in relation to cutting line. **Animal** 8:1754-64.

Messer WP, Petrov DA (2013) Population genomics of rapid adaptation by soft selective sweeps. **Trends Ecol Evol.** 28:1-22.

Metzger J, Karwath M, Tonda R, Beltran S, Águeda L, Gut M, Gut IG, Distl O (2015) Runs of homozygosity reveal signatures of positive selection for reproduction traits in breed and non-breed horses. **BMC Genomics** 16:764.

Michael TP, Florian J, et al. (2018) High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. **Nature Communications** 9:1-8.

Ministério da Agricultura, Pecuária e Abastecimento - MAPA (2016) Estudo do Complexo do Agronegócio do Cavalo. Available at: <http://www.agricultura.gov.br>. Accessed: 16th Jan. 2019.

National Institutes of Health – NIH (2007). Available at: <https://www.nih.gov/news-events/news-releases/horse-genome-assembled>. Accessed: 23th Jan. 2019.

Nei M, Kumar S (2000) Molecular evolution and phylogenetics. **Oxford University Press. Inc.** 2000.

Nielsen R (2005) Molecular Signatures of Natural Selection. **Annu. Rev. Genet.** 39:197-218.

Nolte W, Thaller G, Kuehn C (2019) Selection signatures in four German warmblood horse breeds: Tracing breeding history in the modern sport horse. **PLoS One** 14:e0215913.

Omori R, Wu J (2017) Tajima's D and Site-Specific Nucleotide Frequency in a Population during an Infectious Disease Outbreak. **Appl. Math** 77:2156–2171.

Peripolli E, et al. (2018) Assessment of runs of homozygosity islands and estimates of genomic inbreeding in Gyr (Bos indicus) dairy cattle. **BMC Genomics** 19:34.

Peripolli E, Munari DP, Silva MVGB, Lima ALF, Irgang R, Baldi F (2017) Runs of homozygosity: current knowledge and applications in livestock. **Animal Genetics** 48:255-271.

Petersen JL, Mickelson JR, et al. (2013) Genetic Diversity in the Modern Horse Illustrated from Genome-Wide SNP Data. **PLoS ONE** 8:e54997.

Koellinger PD, van der Loos MJHM, Groenen PJF, et al. (2010) Genome-wide association studies in economics and entrepreneurship research: promises and limitations. **Small Bus Econ** 35:1–18.

Pritchard JK, Pickrell JK, Coop G (2010) The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. **Curr Biol.** 20:R208–R21.

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 BA, Pereirab GL, Bussiman FO (2019) Genomic analysis of the population structure in horses of the Brazilian T Mangalarga Marchador breed. **Livestock Science** 229: 49-55.

Sargolzaei M, Chesnais JP, Schenkel FS (2014). A new approach for efficient genotype imputation using information from relatives. **BMC Genomics** 15:478.

Slatkin M (2008) Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. **Nat Rev Genet.** 9:477–485.

Smith JM, Haigh J (2007) The hitch-hiking effect of a favourable gene. **Genet Res** 89:391-403.

Tajima F (1989) The effect of change in population size on DNA polymorphism. **Genetics** 123:597–601.

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.

The European Bioinformatics Institute - EMBL-EBI (2019) Available at: <https://www.ensembl.org/Equus_caballus/Info/Annotation>. Accessed: 26th Jan. 2019.

U.S. Mangalarga Marchador Association - USMMA (2017). Available at: <http://www.namarchador.org>. Accessed: 27th Dec. 2018.

VanRaden PM, Olson KM, Wiggans GR, Cole JB, Tooker ME (2011) Genomic inbreeding and relationships among Holsteins, jerseys, and Brown Swiss. **J Dairy Sci.** 94:5673–82.

Vitti JJ, Grossman SR, Sabeti PC (2013) Detecting natural selection in genomic data. **Ann Rev Genet** 47: 97-120.

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. **Plos Biology** 4:e72.

Wade CM, Giulotto E, et al. (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. **Science** 326:865-867.

Weigel KA (2001) Controlling Inbreeding in Modern Breeding Programs. **J. Dairy Sci.** 84:E177-E184.

Wray N, Woolliams JA, Thompson R (1990) Methods for predicting rates of inbreeding in selected populations. **Theoretical and Applied Genetics** 80:503-512.

Zhu M, Zhao S (2007) Candidate Gene Identification Approach: Progress and Challenges. **International Journal of Biological Sciences** 3:420-427.

CHAPTER 2 - GENOME-WIDE SCANS FOR SIGNATURES OF SELECTION IN MANGALARGA MARCHADOR HORSES USING HIGH-THROUGHPUT SNP GENOTYPING

ABSTRACT - Brazilian Mangalarga Marchador horse is one of the breeds shaped by generations from local adaptations and specific preferences of breeders to morphology, functionality, and locomotion. Natural gait is a highlighted trait for this breed; their stability during the execution promotes comfort and softness during the ride. Detection of selection signals in genomic regions provide insights over the evolutionary process to understand recent discoveries into complex phenotypic traits, major genes, and metabolic pathways. Collectively, our findings reveal some pieces of evidence for signatures signals associated with athletic performance, gait type, and energy muscle activity, catching the goals expected. It were 169 pruned candidate genes harboring important biological processes, highlighting: anterior/posterior pattern (*GLI3, HOXC9, HOXC6, HOXC5, HOXC4, HOXC13, HOXC11* and *HOXC10*); limb morphogenesis, skeletal system, proximal/distal pattern formation, JUN kinase activity (*CCL19* and *MAP3K6*); and muscle stretch response (*MAPK14*). The others potential signatures were associated with energy metabolism, bronchodilator response, *NADH* regeneration, reproduction, keratinization, and immunological system.

Keywords: iHS, limb morphogenesis, ROH, Tajima’s D

2.1 Introduction

The Mangalarga Marchador horse (MM) is an equine breed relevant to Brazil and the world agribusiness. It is among the four breeds gaited horses in the country, being the batida or picada gait one of the expected traits. There are over 16.000 breeders associated, and over 600,000 registered horses. The Brazilian Association of Mangalarga Marchador Horse Breeders (ABCCMM) comprises 70 centers in various states of Brazil and abroad, being the US market the most promising, however, the MM is present in Europe and some South American countries (ABCCMM, 2018).

Classified's into saddle group, the MM are mediolinear or rectilinear with an eumetric format (ranging from 408 to 500 kg), but there are also curvilinear animals (ABCCMM, 1998). The withers height in the idealized males is 59,85 in, although definitive registration with an interval between 57,87 to 61,81 in is allowed. In females, howsoever, the idealized withers height changes slightly, being 57,48 in, with a tolerance range between 55,12 to 60,63 in (ABCCMM, 1998). In the saddle horse classification, it is possible to categorize it in three biotypes according to the international horse standards: sport that requires speed and/or jumping; livestock service; leisure and/or sports that do not require jumping and speed (Martinho, 2016).

According to Andrade (2016), batida and picada gait type are the main trait of the breed that represents the unique natural movements allowed in the MM for intermediate speeds. The gait is composed of four-beat, with alternative lateral and diagonal support interspersed by triple support moments. In the batida gait, the diagonal is more frequently than triple support, with eight supports, being four of them tripedal (left anterior, right posterior, right anterior and posterior left), two diagonals (left and right), and two laterals (left and right). This whole explanation is organized in four-beat “two by two” (two major and minor, respectively). Already the picada gait is considered softer, given the above definitions, thus that lateral and triple supports overlap when compared to the batida gait, and this particularity represents the major phenotypic trait on gait type segregation (ABCCMM, 2018; USMMA, 2019).

Besides the phenotypic aspects of gait type, Andersson et al. (2012) explained with more detail a bit portion of this phenomenon, howsoever, from a genetic perspective by mapping the *DMRT3* gene, and its allelic patterns associated with equine locomotion. It was the first genetic study about locomotion dissociation patterns in horses, being capable to stimulate discoveries on gait types and the genomic aspects associated with the transcription factors involved in the coordination of limb movement. In it, the frequency for *DMRT3* allele A (mutant) was nearly 100% in the gaited horse, thus the AA homozygous condition was more than proven to be associated at gait. However, further investigations on the *DMRT3* allelic patterns have shown that breed without the gait trait could also have the mutant allele (A), as well as gait horses the C (wild) allele (Promerov et al., 2014). The *DMRT3* gene, also entitled 'Gait Keeper', can explain a major part of the phenotypic variation for gait horses. Although, in MM the gait has different genotypes associated (AA and AC), even so, do not capable to explain that segregation entirely.

The selection signatures represent a strategy for observing the behavior of genes over the artificial/natural selection imposed on gait segregations. Besides that, important complex phenotypes discoveries under genetic aspects might be accessible to the genetic improvement in the breed. Variation studies underlying the hitchhiking effects on genomic scanning, and search recent adaptive fixations were first inspired by Lewontin & Krakauer (Lewontin e Krakauer, 1973; Nei e Maruyama, 1975). Thus, past and current studies bring to us the concept of selection signatures, that are particular patterns of DNA identified in regions of the genome with mutation and/or have been under natural/artificial selection pressure on population (Nielsen et al., 2007; Bertolini e Servin, 2018; Bamshad e Wooding, 2003). The exploitation from these signals helps to find important genomic regions that have been under selective pressure and might host genes and variants that modulate important phenotypes in horses (Avila et al., 2018; Srikanth et al., 2019).

Over the past years, the detection of selection signatures has resulted in the publication of many studies involving livestock species, marked genetically by selection, domestication process, and artificial selection which aim to increase herd performance and productivity (Qanbari e Simianer, 2014; Gouveia et al., 2017). There are several approaches to identifying signatures of selection (Weigand e Leese, 2018; Nielsen, 2005; Sabeti et al., 2002; Purfield et al., 2017; Fariello et al., 2013; Pérez O'Brien et al., 2014; Tajima, 1989; Purcell et al., 2007; Gautier e Vitalis, 2012), and Weigand et al. (2018) gather most available technics in a review study, which addressed the particularities of each method in a non-model species perspectivity. In this research, we used three different approaches: Tajima's D (TD) (Tajima, 1989), Integrated Haplotype Score (iHS) (Gautier e Vitalis, 2012), and Runs of Homozygosity (ROH) (Purcell et al., 2007). The choices for these tests were based on the population genetic structure. As the horses of both gait types presented only one population structure, signals of selection were scanned only within population. Therefore, we aims identification of indirectly modified genomic regions due to recent selection pressures (natural and/or artificial), as well as candidate genes associated with traits of importance in the breed, especially genes related to type and gait quality, temperament, conformation, and locomotor system (muscular and skeletal structure).

2.2 Methods

2.2.1 Ethical statement

All experimental procedures involving horses in this study were performed in accordance with 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. 0029/2017).

2.2.2 Sample collection, gait patterns, and DNA extraction

Horse samples were collected in Brazil during the 36th Brazilian National Exhibition of the Mangalarga Marchador breed, and also in stud-farm from São Paulo, and the Minas Gerais State. It was regarded both sexes in the sampling - males (n=62), and females (n=130) selected in two gait patterns with well-defined traits: picada (n=86) and batida gait (n=106). Besides, the presence of animals from unrelated lineages was considered in the composition of the sample, avoiding the presence of full-sibs. Jugular blood of 5mL was collected with the immersion in 7.5 mg EDTA. We extract the genomic DNA from each horse using an Illustra Blood Genomic PrepMini Spin Kit (GE Healthcare, USA), according to the manufacturer's instructions. The DNA was quantified using a Qubit® 3.0 Fluorometer (Invitrogen, USA) and quality assessment DNA by NanoDrop™ Lite Spectrophotometer (NanoDrop Lite, Thermo Scientific, USA), and 0.8% agarose gel electrophoresis. The final dilutions per sample were ~10 ng/μL.

2.2.3 Genotype, quality control, filter and phase genotypes

All horses were genotyped from 670k *Axiom ® Equine Genotyping Array* (Axiom MNEC670). The assessments were done by Axiom™ Analysis Suite Software using the default configurations for diploid organisms in the version 4 with best/recommended SNPs (sample QC: DQC ≥ 0.82, call rate ≥ 97, percent of passing samples ≥ 95, average call rate for passing samples ≥ 98.5; and SNP QC thresholds: call rate ≥ 97, plus twenty-six others parameters for diploid organisms (standard protocol) can be consulted with more details (Supplementary Methods 1). With the recently updated reference assembly of the equine genome, the SNP array coordinates were remapped to EquCab3.0 (Beeson et al., 2019), being excluded non-autosomal chromosomes in the remapping, except ECA X. The raw reports with EquCab3.0 SNP coordinates for the MNEc670k array are hosted at https://www.animalgenome.org/repository/pub/UMN2018.1003/. Furthermore, coordinates between the two assemblies can be easily converted now from NCBI (https://www.ncbi.nlm.nih.gov/genome/tools/remap). The final density was 545,219 SNPs with 32 chromosomes analyzed, including the ECA X (Fig. 1).

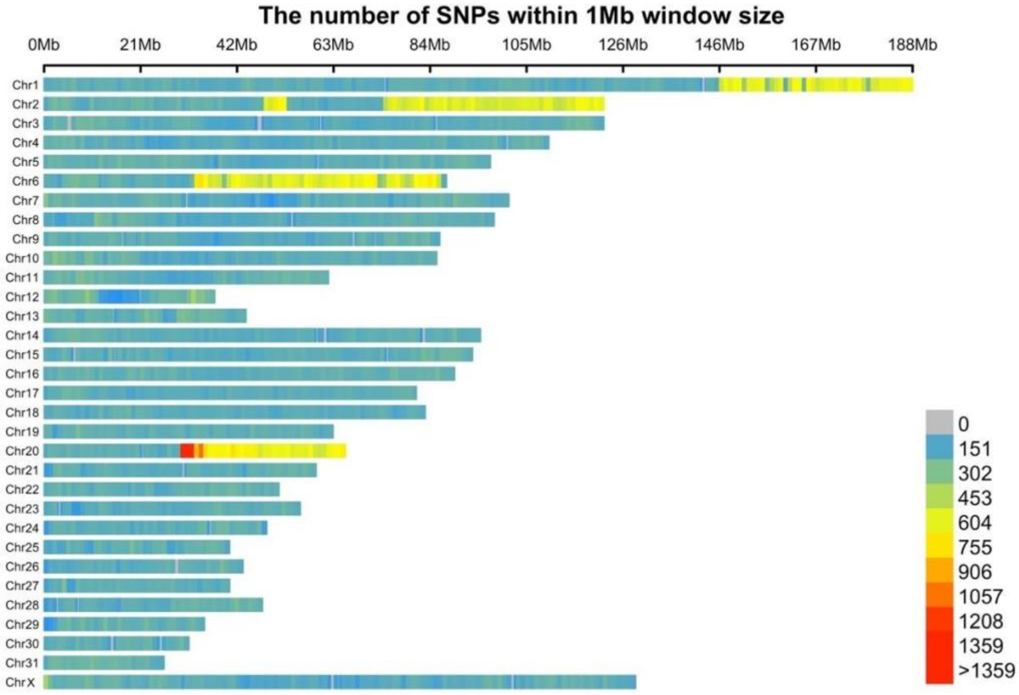


Figure 1. Final density of 545,219 SNP in the Mangalarga Marchador horse genome after Axiom™ Analysis Suite pruning.

Complementary quality control was performed in VCFtools and R software to pruning the data into the standards required by each test (Hardy-Weinberg (P<1e-8) --hwe; and MAF=0.01 to the iHS/Tajima's D and 0.005 for ROH analyses. To both situations were used the --maf parameter). Besides, through R function used, non-autosomal chromosomes were removed, SNPs for the same position have been removed, and the database was ordered by chromosome and position. Thereby, the differences for distinct minor allele frequency parameters: MAF 0.01=422656 SNP (iHS/Tajima's D), and MAF 0.005=444929 SNP (ROH) resulted in two databases. The MAF parameter to selection signatures analyses is not yet well established in the literature. However, in the ROH studies, we chose to adopt an extreme lower parameter due studies recommending not to use the MAF threshold (possible underestimation) (Ferencakovic et al., 2013). For iHS, and Tajima's D analyses, the database was computed on Beagle version 5.0 that provides faster and accurate algorithms for genotypes haplotyping/phasing (Pook et al., 2019).

2.2.4 Population structure and linkage disequilibrium analyses

The Principal Component Analysis (PCAs) were performed in Plink 1.9 (Purcell et al., 2007) using linkage disequilibrium pruning to remove the SNP pair base correlations to remaining approximately independent SNPs, and faster subsequent computations --indep-pairwise. After was used the relatedness between samples for the computation of genome-wide IBD estimates --genome. To compute PCAs, we choose a sample of each closely-related pair and exclude in R software (based on high-values for pairwise PI_HAT statistic sum). The goals were to refine the analysis for the entire genome, and also exclude samples with greater inaccuracy.

The Linkage disequilibrium (LD) level was calculated for the entire panel using phased data. To conduct the LD Decay analysis was used the PopLDdecay pipeline - OutStat on default prunning (Chen et al., 2017) with markers density reduction to 347,935 SNPs, where the plotter and complementary analysis was conducted in R using pegas (Paradis, 2010), ape (Paradis e Schliep, 2018), and ggplot2 (Wickham et al., 2016) packages.

2.2.5 Genome-Wide scan for signals of positive selection

We used three distinct approaches to capture as much as possible the evolutionary aspects of the selection in the MM. Each approach has some strengths and disadvantages, and the combination/integration and reproducibility of the results add better accuracy to the analyses.

Tajima's D (TD) – Older feature of selection were investigated using the traditional frequency-based neutrality test, this neutrality test uses the site's frequency spectrum to capture selective scans occurring up to ~ 250,000 years ago (He et al., 2008). Tajima's D statistics estimate the comparison of nucleotide diversity by observing polymorphic sites in a given set of chromosomes against nucleotide diversity estimated from the allele frequency of polymorphic sites (Akhunov et al., 2010). The VCFtools (http://vcftools.sourceforge.net/) calculated 20 kb sliding windows across all autosomal regions --TajimaD. Windows containing missing variants were ignored. Implementations were conducted in R software to provide graphics and to sort-windows based in ascending order of the Tajima's D values using empirical p-values (Yu et al., 2009) of less than 0.01.

Integrated Haplotype Score (iHS) – It is the highest power method nowadays to catch signals of ongoing selection, when fixation of the selected allele is not reached. Developed based on extensive computational simulations to determine the best statistics among several, iHS complements the extended haplotype homozygosity (EHH) in recent selection signature studies by providing more accurate values by reducing biases attributed to the influence of demographic history on the population (Gautier et al., 2012). The package used for this analysis was R rehh v.3.01 (Gautier e Vitalis, 2012; Gautier et al., 2017) which were discard focal markers with Minor Allele Frequency (MAF) equal to or below 0.01. Due to the absence of representative studies in horses and most non-model species for designation of alleles as 'ancestral' or 'derived', iHS analysis was conducted using unpolarized alleles (new feature of the latest version of the package). This version allows defining the function as FALSE, ideal for many domestic animal studies as well as non-model organisms. Thus, iHH (integrated EHH) values were computed for the major (most frequent) and minor (second-most frequent) alleles. Where iHS values >3.5 and/or <-3.5, wich piHS (p-value for iHS) ≥ 3 were considered statistically significant rejecting the null hypothesis

(P<0.001). The piHS values are products of iHS transformation to assign a p-value, piHS = [- log10 [1 - 2 | ΦiHS - 0.5 |], wherein Φ iHS is the Gaussian cumulative distribution function of iHS.

Runs of Homozygosity (ROH) – It has been described as a powerful approach to study selection signatures in the genome. The ROH are continuous homozygous segments of the DNA sequence that arise when identical haplotypes are inherited from each parent and thus a long-range of genotypes is homozygous (Ceballos et al., 2018). The analysis was conducted with Plink 2.0 –homozyg (-density 50, -gap 1000, -kb 250, - snp 50, -window-het 2, -window-missing 2, -window-snp 50, -window-threshold 0.05) (Ceballos et al., 2018). The posterior analyses were performed in R software with the script developed by the Boison (https://github.com/soloboan/ROHs) to generate binary runs of homozygosity. SNPs with ROH proportion lower than 0.01 were discarded. The signals for signature of selections were defined as ROH islands regions (hotspot mean) with frequencies in the population ≥ 0.5.

2.2.6 Gene-annotation enrichment analysis

Gene annotation was carried out from genomic regions identified as signals of selection considering the different approach used. Window sizes were set at 125 kb for both directions for each region/SNP (P<0.01), based on LD information and approximate values from literature. Genes mapped in these windows were identified based on the most recent assembly of the equine genome sequence (EquCab3.0) (Beeson et al., 2019) through the BioMart R package (Smedley et al.,, 2009) and PANTHER Classification System (www.pantherdb.org). Only genes identified with known functions were annotated/enriched for biological processes, molecular functions, and cellular components analysis.

2.2.7 Gene network analysis.

Gene networks can shed light on the complex behavior of horse genes related to gait patterns, diseases, performance, and physiology. The networks were constructed considering common genes results for the aforementioned methods, and also by the biological know about the equine sector in a wider aspect. The interactions were calculated by GeneMANIA (Franz et al., 2018), and STRINGdb R package

(Franceschini et al., 2013); interactions include direct (physical) and indirect (functional) associations between genes (Szklarczyk et al., 2015).

2.3 Results

2.3.1 Breed structure

Only one population structure was found by PCA. The top five-seteigenvectors explained 54.98% of the cumulative variance, of which 40.33% belongs to the cluster 1 for PCA 1 x PCA 2 (Fig. 2). In the LD analysis, the sudden decrease in LD was observed, as the physical distance between the markers increased, being quite low. In the Fig. 3, r^2 values already show up below 0.20 in distances smaller than 15 kb.

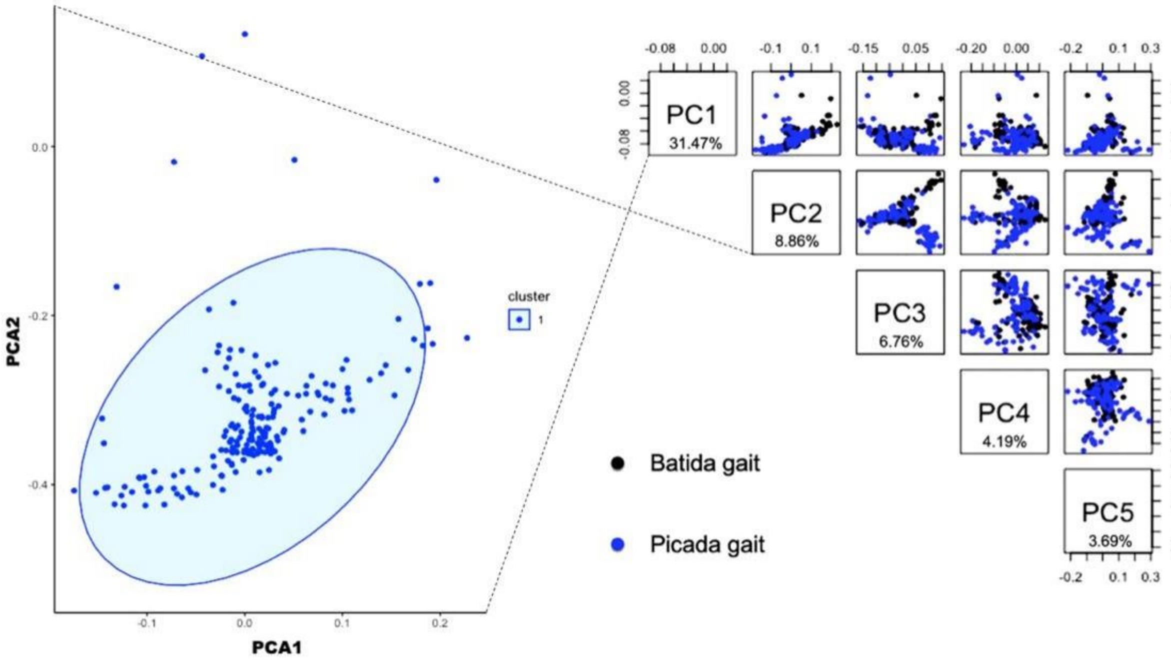


Figure 2. Principal Component Analysis (PCA) based on genotype data for top five-set eigenvectors in two Mangalarga Marchador horse gait type. The core PCAs were highlighted in cluster 1.

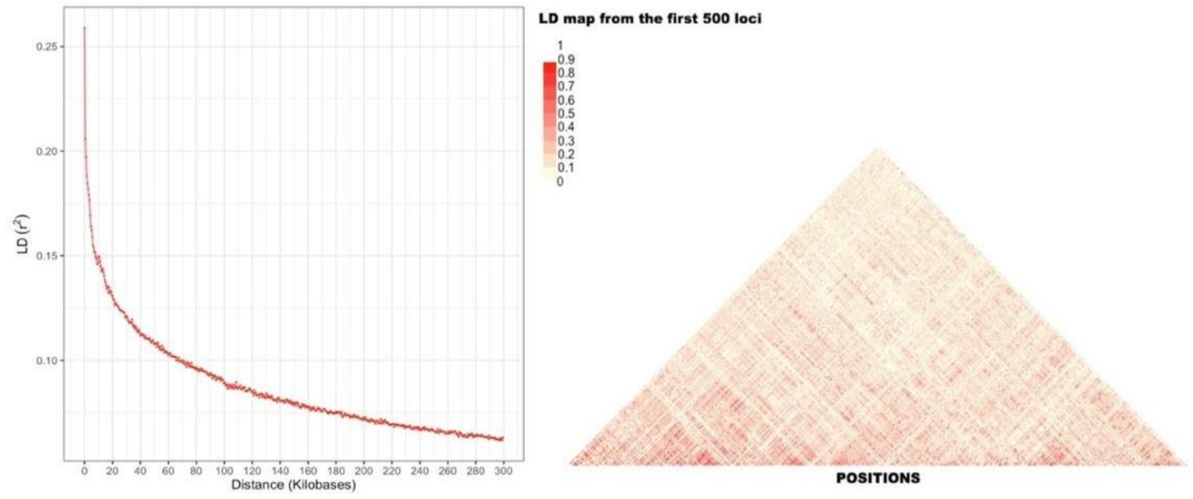


Figure 3. Genome-wide linkage disequilibrium (LD) decay plot for 192 Mangalarga Marchador based on 347,935 SNP markers.

Additional aspects of population structure and LD have been reported by Santos et al. (2019) using 240 animals, the same of which a portion belongs to the studied population. As we preferred not to conduct the study with imputed data from two different platforms, the analyzes were conducted only for 192 animals genotyped on Axiom MNEC670. The results were practically reproducible for population structure and linkage disequilibrium analyses. Slight changes were noticeable possibly attributed to the different methods used in both studies, as well as, the reduction of the number of animals, however, the conclusions on the population structure and LD remained the same.

2.3.2 Candidate genes identifications

Higher Tajima’s D values were identified under balanced selection in a wide aspect, and almost all chromosomes demonstrated at least one significant as signal of selection (Supplementary Data 1). In general, the higher proportion of SNPs noticeable under balance selection or sudden population contraction scenarios has some subjectivity in interpretations.

Tajima's D values in the negative tail were slightly below those already reported in the literature. We used the values of -log10(p-value)≥2 from empirical p-values to do the selection of significant regions. In the total, 147 significant genomic regions where observed in both tails (P<0.01). Often, however, the most representative regions are selected as positive selection in the extreme negative tail. Due to some limitations/biases inherent to the method (Zhang et al., 2015), as well as to clarify the results. We considered in this statistic only the negative tail, which is associated with signals of positive selection (Carlson, 2005). Thus, only 10 most representative genomic regions (negative tail) were selected on ECA 1, 6, 7, 8, 20, and 26 (Fig. 4).

In the annotation results 27 candidate genes were targeted as evidences of positive selection (Table 1).

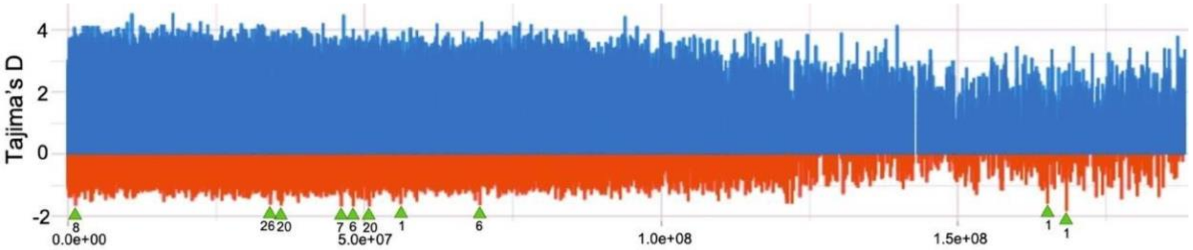


Figure 4. Patterns of genome-wide polymorphism for Tajima's D statistics were calculated in 20 kb windows across the genome. Top 10 genomic regions in the negative tail were marked with small green arrowheads, and correspond chromosomes numbers.

Table 1. Candidate genes harboring Tajima's D under evidence of positive signature of selection in the Brazilian Mangalarga Marchador horses.

Ensembl Gene ID	Chr	Start Position	End Position	Genes	Description
ENSECAG00000002972	1	168151718	168251697	*SCFD1*	sec1 family domain containing 1
ENSECAG00000010464	1	168366363	168459590	*STRN3*	striatin 3
ENSECAG00000021944	1	168350423	168362177	*COCH*	cochlin
ENSECAG00000001908	6	69477881	69485306	*KRT84*	keratin 84
ENSECAG00000002542	6	69388943	69394050	*KRT81*	keratin, type II cuticular Hb1
ENSECAG00000007842	6	69494571	69506248	*KRT82*	keratin 82
ENSECAG00000008097	6	48143741	48163481	*CMAS*	cytidine monophosphate N-acetylneuraminic acid synthetase
ENSECAG00000009201	6	69402662	69409182	*KRT86*	keratin 86
ENSECAG00000009991	6	69523789	69533483	*KRT75*	keratin 75
ENSECAG00000013512	6	69553390	69558280	*KRT6C*	keratin 6C
ENSECAG00000015478	6	69416432	69422664	*KRT83*	keratin 83
ENSECAG00000017378	6	47951001	48065838	*ABCC9*	ATP binding cassette subfamily C member 9
ENSECAG00000020216	6	69340116	69353237	*KRT7*	keratin 7
ENSECAG00000006093	8	1325765	1462223	*CABIN1*	calcineurin binding protein 1
ENSECAG00000017804	8	1142374	1169168	*UPB1*	beta-ureidopropionase 1
ENSECAG00000020031	8	1187365	1197777	*GUCD1*	guanylyl cyclase domain containing 1
ENSECAG00000021670	8	1273135	1293683	*GGT5*	gamma-glutamyltransferase 5
ENSECAG00000023316	8	1198613	1218433	*SNRPD3*	small nuclear ribonucleoprotein D3 polypeptide
ENSECAG00000023404	8	1239052	1245534	*LRRC75B*	leucine rich repeat containing 75B
ENSECAG00000025078	8	1316427	1322900	*SUSD2*	sushi domain containing 2
ENSECAG00000000493	20	35958531	36021014	*SLC26A8*	solute carrier family 26 member 8
ENSECAG00000012160	20	35818020	35820026	*CLPS*	Equus caballus colipase (CLPS), mRNA
ENSECAG00000014034	20	35831559	35837234	*LHFPL5*	LHFPL tetraspan subfamily member 5

ENSECAG00000014175	20	36052316	36094294	*MAPK14*	mitogen-activated protein kinase 14
ENSECAG00000014213	20	35848788	35881283	*SRPK1*	SRSF protein kinase 1
ENSECAG00000014228	20	50724469	50742569	*GCM1*	glial cells missing homolog 1
ENSECAG00000014755	20	50814846	50837355	*ELOVL5*	ELOVL fatty acid elongase 5

Chr: Chromosomes

The iHS statistic was analyzed for both tails with the goals of capture recent positive natural and/or artificial selection. The ancestral allele's state were of 251 genomic regions and 41 in derived state – positive and negative tails, respectively. Thus, 292 genomic regions were catching on almost all chromosomes (except ECA 21, 22, 26, 28 and 31) (Supplementary Data 2) (Fig. 5). Genomic annotations were applied in the two allelic states, that corresponds both tails, being annotated all significant genomic regions (P<0.001). Three hundred thirty two genes were identified as signals of selection, being 20 of them non-protein-coding RNA (miRNAs, U6 spliceosomal RNA, small nucleolar RNA).

Due to a large number of genes in iHS results, we do not follow the commonly used method of choosing only the top genomic regions – this is a good strategy for highlighting the most representative areas in the genome. However, by the particulars of each chromosome, as well as by the limitation of computational methods, other parameters were used: (I) it was decided to paid special attention in top highlighted genes, (II) genes related to locomotion, athletic performance, growth, fertility, conformation, pigmentation, and metabolism were overlaid, (III) as well as common genes between the statistical methods.

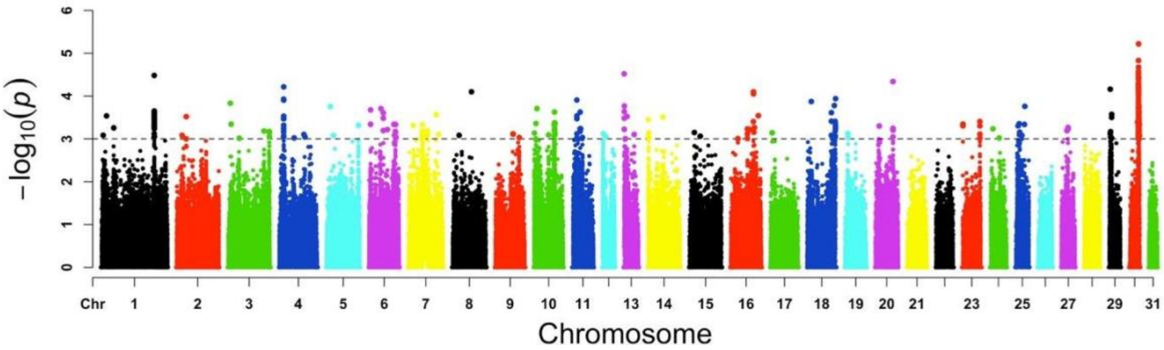


Figure 5. Genome-wide distribution of selection signatures detected by Integrated Haplotype Score (iHS) on autosomal chromosomes. Significant marker values were demarcated by the dashed line (P <0.001).

The list of 104 genes highlighted from the annotation of iHS results is shown in Table 2. Results for ROH found 340 SNPs within ROH islands (Fig. 6).

We used the same principle of annotation as the iHS test with 250k windows (part of the SNPs on each chromosome shared the same window). With the final annotation, there were 38 genes (Table 3), and nine of them were common between ROH and iHS (ECA 1: *RASGRP1* and ECA 23: *C9orf24, CNTFR, DCTN3, DNAI1, ENHO, FAM219A, RPP25L,* and *SIGMAR1*).

Table 2. Candidate genes harboring Integrated haplotype Score (iHS) under evidence of recent positive signature of selection in the Brazilian Mangalarga Marchador horse.

Ensembl Gene ID	Chr	Start Position	End Position	Genes	Description
ENSECAG00000008623	1	149907955	150022286	*SPRED1*	sprouty related EVH1 domain containing 1
ENSECAG00000010114	1	149706774	149775059	*RASGRP1*	RAS guanyl releasing protein 1
ENSECAG00000005510	2	28397066	28398058	*GPR3*	G protein-coupled receptor 3
ENSECAG00000010268	2	28323886	28385023	*WASF2*	WAS protein family member 2
ENSECAG00000011296	2	28562401	28609929	*SLC9A1*	solute carrier family 9 member A1
ENSECAG00000014444	2	28406073	28409277	*CD164L2*	CD164 molecule like 2
ENSECAG00000014857	2	28412906	28416935	*FCN3*	ficolin 3
ENSECAG00000015410	2	28420416	28429784	*MAP3K6*	mitogen-activated protein kinase kinase kinase 6
ENSECAG00000020672	2	28430246	28438417	*SYTL1*	synaptotagmin like 1
ENSECAG00000023706	2	28443577	28453634	*TMEM222*	transmembrane protein 222
ENSECAG00000024411	2	28463993	28508594	*WDTC1*	WD and tetratricopeptide repeats 1
ENSECAG00000009649	3	7693420	7744860	*LPCAT2*	lysophosphatidylcholine acyltransferase 2
ENSECAG00000011520	3	7794621	7835751	*SLC6A2*	solute carrier family 6 member 2
ENSECAG00000009281	4	13120953	13294999	*GLI3*	GLI family zinc finger 3
ENSECAG00000007481	5	12015652	12304265	*ASTN1*	astrotactin 1
ENSECAG00000024570	5	12310453	12412709	*BRINP2*	BMP/retinoic acid inducible neural specific 2
ENSECAG00000025428	5	12172407	12172489		eca-mir-488
ENSECAG00000000386	6	34369455	34374801	*LRRC23*	leucine rich repeat containing 23
ENSECAG00000000465	6	34410281	34420057	*PTPN6*	protein tyrosine phosphatase, non-receptor type 6
ENSECAG00000000701	6	5486218	5551290	*FN1*	fibronectin 1
ENSECAG00000000726	6	70865117	70867507	*HOXC9*	homeobox C9
ENSECAG00000003682	6	70892992	70894488	*HOXC6*	homeobox C6
ENSECAG00000004151	6	70897601	70899132	*HOXC5*	homeobox C5
ENSECAG00000004202	6	70917898	70919290	*HOXC4*	homeobox C4
ENSECAG00000007386	6	34377361	34383187	*ENO2*	enolase 2
ENSECAG00000009049	6	34274460	34301295	*CD4*	CD4 molecule
ENSECAG00000009519	6	34515391	34524075	*C1S*	complement C1s
ENSECAG00000012522	6	34321532	34326725	*GNB3*	G protein subunit beta 3
ENSECAG00000014517	6	34328207	34330205	*CDCA3*	cell division cycle associated 3
ENSECAG00000014532	6	34331414	34344976	*USP5*	ubiquitin specific peptidase 5
ENSECAG00000014653	6	5446142	5472875	*ATIC*	5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase
ENSECAG00000015581	6	34346419	34349728	*TPI1*	triosephosphate isomerase 1
ENSECAG00000016937	6	34425844	34429448	*PHB2*	prohibitin 2
ENSECAG00000019250	6	34304988	34308833	*GPR162*	G protein-coupled receptor 162
ENSECAG00000021403	6	34393931	34400776	*ATN1*	atrophin 1
ENSECAG00000021815	6	34310310	34319714	*P3H3*	prolyl 3-hydroxylase 3
ENSECAG00000022412	6	34429726	34434811	*EMG1*	EMG1, N1-specific pseudouridine methyltransferase
ENSECAG00000023202	6	34435377	34471395	*LPCAT3*	lysophosphatidylcholine acyltransferase 3
ENSECAG00000024867	6	70802998	70809716	*HOXC13*	homeobox C13
ENSECAG00000024869	6	34402198	34404001	*C6H12orf57*	chromosome 6 C12orf57 homolog
ENSECAG00000024893	6	70819239	70820860	*HOXC12*	homeobox C12
ENSECAG00000024900	6	70837383	70840203	*HOXC11*	homeobox C11
ENSECAG00000024985	6	70850147	70854018	*HOXC10*	homeobox C10
ENSECAG00000025389	6	34423082	34423146		eca-mir-200c
ENSECAG00000025607	6	70898503	70898599		eca-mir-615
ENSECAG00000026310	6	34423490	34423561		eca-mir-141
ENSECAG00000027042	6	34402169	34402230		U7 small nuclear RNA
ENSECAG00000027594	6	34426452	34426715		small nucleolar RNA U89
ENSECAG00000003757	10	6624595	6634234	*GAPDHS*	glyceraldehyde-3-phosphate dehydrogenase, spermatogenic
ENSECAG00000005226	10	6561153	6562124	*FFAR2*	free fatty acid receptor 2
ENSECAG00000011198	10	60335470	60340309	*AMD1*	adenosylmethionine decarboxylase 1
ENSECAG00000011975	10	6634647	6636230	*TMEM147*	transmembrane protein 147

ENSECAG00000012822	10	9635035	9645344	*EIF3K*	eukaryotic translation initiation factor 3 subunit K
ENSECAG00000013121	10	6639494	6652250	*ATP4A*	ATPase H+/K+ transporting subunit alpha
ENSECAG00000014214	10	60375200	60382281	*GTF3C6*	general transcription factor IIIC subunit 6
ENSECAG00000015344	10	9510873	9616030	*RYR1*	ryanodine receptor 1
ENSECAG00000017061	10	9616257	9633779	*MAP4K1*	mitogen-activated protein kinase kinase kinase kinase 1
ENSECAG00000017121	10	60395300	60425982	*RPF2*	ribosome production factor 2 homolog
ENSECAG00000020313	10	60557764	60601940	*SLC16A10*	solute carrier family 16 member 10
ENSECAG00000021777	10	9692742	9718476	*ACTN4*	actinin alpha 4
ENSECAG00000025001	10	6589049	6591925	*KRTDAP*	keratinocyte differentiation associated protein
ENSECAG00000006771	11	13417359	13812648	*PRKCA*	protein kinase C alpha
ENSECAG00000007214	11	13765651	14005312	*CACNG4*	calcium voltage-gated channel auxiliary subunit gamma 4
ENSECAG00000000176	13	1935848	1947933	*ZDHHC4*	zinc finger DHHC-type containing 4
ENSECAG00000008056	13	2414177	2422727	*FSCN1*	fascin actin-bundling protein 1
ENSECAG00000009724	13	2153427	2160374	*RBAK*	RB associated KRAB zinc finger
ENSECAG00000010225	13	1882012	1913837	*GRID2IP*	Grid2 interacting protein
ENSECAG00000011713	13	1949573	1958047	*C7orf26*	chromosome 7 open reading frame 26
ENSECAG00000013171	13	2265413	2398956	*RNF216*	ring finger protein 216
ENSECAG00000015935	13	2463585	2465463	*ACTB*	Equus caballus actin beta (ACTB), mRNA
ENSECAG00000016420	13	2086916	2092792	*ZNF12*	zinc finger protein 12
ENSECAG00000018678	13	2472540	2510292	*FBXL18*	F-box and leucine rich repeat protein 18
ENSECAG00000022114	13	2711477	2738292	*WIPI2*	WD repeat domain, phosphoinositide interacting 2
ENSECAG00000013897	16	65160454	65270909	*RFTN1*	raftlin, lipid raft linker 1
ENSECAG00000008768	18	79106315	80034010	*PARD3B*	par-3 family cell polarity regulator beta
ENSECAG00000012151	18	12086034	12116622	*MARCO*	macrophage receptor with collagenous structure
ENSECAG00000016824	18	80076435	80186347	*NRP2*	neuropilin 2
ENSECAG00000018298	18	76437419	76456956	*STRADB*	STE20-related kinase adaptor beta
ENSECAG00000019645	18	76634162	76650235	*TMEM237*	transmembrane protein 237
ENSECAG00000022800	18	76653422	76689588	*MPP4*	membrane palmitoylated protein 4
ENSECAG00000010916	20	50162197	50233536	*TRAM2*	translocation associated membrane protein 2
ENSECAG00000015579	20	50310519	50323621	*TMEM14A*	transmembrane protein 14A
ENSECAG00000016221	20	50347190	50357534	*GSTA1*	Equus caballus glutathione S-transferase alpha 1 (GSTA1), mRNA
ENSECAG00000019567	20	50425513	50435370	*LOC100271875*	glutathionine S-transferase alpha 3
ENSECAG00000004463	23	50231564	50255138	*UBAP1*	ubiquitin associated protein 1
ENSECAG00000004776	23	50338512	50340656	*MYORG*	myogenesis regulating glycosidase (putative)
ENSECAG00000004839	23	50465243	50465473	*ENHO*	Equus caballus energy homeostasis associated (ENHO), mRNA
ENSECAG00000006176	23	50484877	50502709	*CNTFR*	ciliary neurotrophic factor receptor
ENSECAG00000010758	23	50257759	50299304	*KIF24*	kinesin family member 24
ENSECAG00000011552	23	50328495	50331173	*NUDT2*	nudix hydrolase 2
ENSECAG00000011566	23	50345688	50359034	*C9orf24*	chromosome 9 open reading frame 24
ENSECAG00000012578	23	50362111	50367137	*FAM219A*	family with sequence similarity 219 member A
ENSECAG00000016961	23	50426532	50464571	*DNAI1*	dynein axonemal intermediate chain 1
ENSECAG00000027205	23	50423793	50424056		RNA, 7SK small nuclear pseudogene 24
ENSECAG00000002357	23	50540041	50540532	*RPP25L*	ribonuclease P/MRP subunit p25 like
ENSECAG00000013178	23	50543087	50549476	*DCTN3*	dynactin subunit 3
ENSECAG00000019783	23	50562602	50564385	*SIGMAR1*	sigma non-opioid intracellular receptor 1
ENSECAG00000001054	25	27004948	27005868	*LOC100071212*	olfactory receptor 1L6-like
ENSECAG00000001330	25	27025906	27026868	*OR5C1*	olfactory receptor 5C1
ENSECAG00000002169	25	27033670	27034620	*OR1K1*	olfactory receptor family 1 subfamily K member 1
ENSECAG00000002222	25	27136728	27138002	*ZBTB6*	zinc finger and BTB domain containing 6
ENSECAG00000006897	25	26957307	26958330	*LOC100071227*	olfactory receptor 1L4-like
ENSECAG00000006946	25	26979321	26980244	*LOC100071218*	olfactory receptor 1L4-like
ENSECAG00000017397	25	27143414	27153522	*ZBTB26*	zinc finger and BTB domain containing 26
ENSECAG00000017729	25	27161291	27312547	*RABGAP1*	RAB GTPase activating protein 1
ENSECAG00000021253	25	26896324	27056065	*PDCL*	phosducin like
ENSECAG00000022176	25	27085189	27132323	*RC3H2*	ring finger and CCCH-type domains 2
ENSECAG00000025393	25	27106545	27106655		small nucleolar RNA SNORD90
ENSECAG00000007192	30	26241146	26299185	*PTPRC*	protein tyrosine phosphatase, receptor type C
ENSECAG00000023881	30	26077245	26096063	*ATP6V1G3*	ATPase H+ transporting V1 subunit G3
ENSECAG00000025552	30	26398918	26399027		eca-mir-181a-2

Chr: Chromosomes

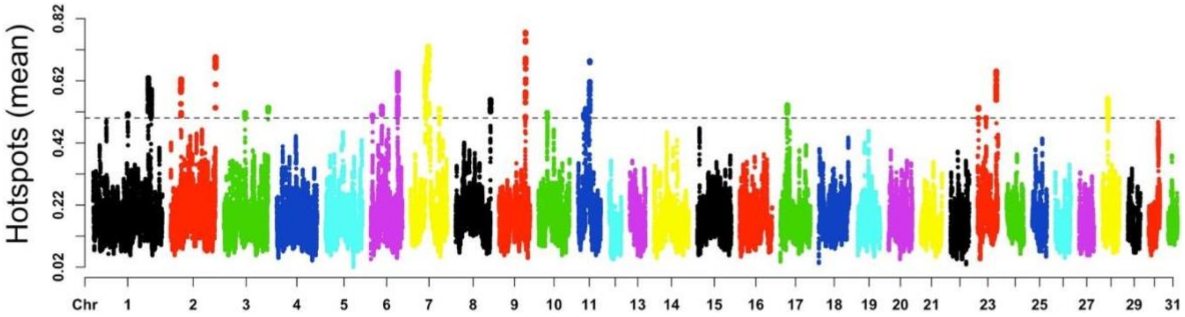


Figure 6. ROH islands regions on 31 autosomal chromosomes. The dashed line represent a threshold for significant ROH hotspot mean frequencies ≥ 0.50.

Table 3. Candidate genes harboring Runs of Homozygosity (ROH) under evidence of positive signature of selection in the Brazilian Mangalarga Marchador horses.

Ensembl Gene ID	Chr	Start position	End Position	Genes	Description
ENSECAG00000002212	17	18615804	18617704	*FOXO1*	forkhead box O1
ENSECAG00000002357	23	50540041	50540532	*RPP25L*	ribonuclease P/MRP subunit p25 like
ENSECAG00000002945	11	32087533	32087985	*CCDC182*	coiled-coil domain containing 182
ENSECAG00000003551	9	73341857	73423501	*LRRC6*	leucine rich repeat containing 6
ENSECAG00000003600	17	18742806	18778545	*MRPS31*	mitochondrial ribosomal protein S31
ENSECAG00000003634	6	30832832	30834614	*RHNO1*	RAD9-HUS1-RAD1 interacting nuclear orphan 1
ENSECAG00000004839	23	50465243	50465473	*ENHO*	Equus caballus energy homeostasis associated (ENHO), mRNA
ENSECAG00000005017	7	45641390	45646041	*FBXW9*	F-box and WD repeat domain containing 9
ENSECAG00000005303	6	30883302	30896118	*TULP3*	tubby like protein 3
ENSECAG00000006176	23	50484877	50502709	*CNTFR*	ciliary neurotrophic factor receptor
ENSECAG00000008176	23	50568433	50571634	*GALT*	galactose-1-phosphate uridylyltransferase
ENSECAG00000008886	7	45647009	45647307	*GNG14*	G protein subunit gamma 14
ENSECAG00000009177	7	45651702	45655097	*DHPS*	deoxyhypusine synthase
ENSECAG00000009337	6	31002891	31197638	*TSPAN9*	tetraspanin 9
ENSECAG00000010114	1	149706774	149775059	*RASGRP1*	RAS guanyl releasing protein 1
ENSECAG00000010144	6	30609983	30638725	*DDX11*	DEAD/H-box helicase 11
ENSECAG00000010693	6	30781746	30790420	*ITFG2*	integrin alpha FG-GAP repeat containing 2
ENSECAG00000011303	6	30931802	30968566	*TEAD4*	TEA domain transcription factor 4
ENSECAG00000011435	11	31647130	32031445	*MSI2*	musashi RNA binding protein 2
ENSECAG00000011566	23	50345688	50359034	*C9orf24*	chromosome 9 open reading frame 24
ENSECAG00000012154	7	45617437	45620601	*TRIR*	telomerase RNA component interacting RNase
ENSECAG00000012578	23	50362111	50367137	*FAM219A*	family with sequence similarity 219 member A
ENSECAG00000012611	9	73453208	73478427	*TMEM71*	transmembrane protein 71
ENSECAG00000013178	23	50543087	50549476	*DCTN3*	dynactin subunit 3
ENSECAG00000013410	6	30360399	30398526	*SLC6A13*	solute carrier family 6 member 13
ENSECAG00000013412	23	50605846	50607075	*CCL19*	C-C motif chemokine ligand 19
ENSECAG00000013673	7	45626599	45637845	*TNPO2*	transportin 2
ENSECAG00000016961	23	50426532	50464571	*DNAI1*	dynein axonemal intermediate chain 1
ENSECAG00000017442	23	50576370	50582294	*IL11RA*	interleukin 11 receptor subunit alpha
ENSECAG00000017467	9	72950611	72999460	*KCNQ3*	potassium voltage-gated channel subfamily Q member 3
ENSECAG00000018082	6	30792657	30799176	*NRIP2*	nuclear receptor interacting protein 2
ENSECAG00000018777	6	30810694	30816253	*TEX52*	testis expressed 52
ENSECAG00000019129	6	30817902	30826537	*FOXM1*	forkhead box M1
ENSECAG00000019283	6	30595381	30608852	*WASHC1*	WASH complex subunit 1
ENSECAG00000019783	23	50562602	50564385	*SIGMAR1*	sigma non-opioid intracellular receptor 1
ENSECAG00000019788	7	45655144	45658885	*WDR83*	WD repeat domain 83
ENSECAG00000020465	6	30769282	30775457	*FKBP4*	FK506 binding protein 4
ENSECAG00000021981	7	45659262	45660408	*WDR83OS*	WD repeat domain 83 opposite strand
Chr: Chromosomes					

Tajima's D did not show common genes with the other two statistics. It was noticeable that the pairwise differences were more pronounced than segregating sites, which may have influenced and/or limited the results of positive selection with indications of being positive selection but not so recent or even indicating that is a population in expansion. Conversely, even with common results between the ROH and iHS statistics, only two chromosomes with a total of nine genes were reproducible between methods. Thereby, we seek to broaden the understanding of these genes through enrichment analysis, and gene networks.

ROH arises when two copies of an ancestral haplotype are brought together in an individual (Ceballos et al., 2018). According to Hillestad et al. (2017), a homozygous segment originating from a more recent ancestor is expected to be longer, as there were fewer opportunities for recombination to reduce its length. ECA 7 was the longest shared homozygosity chromosome, while the other chromosomes had short ROH, which corresponds to the evidence that the recombination has already caused its reduction (Fig. 7).

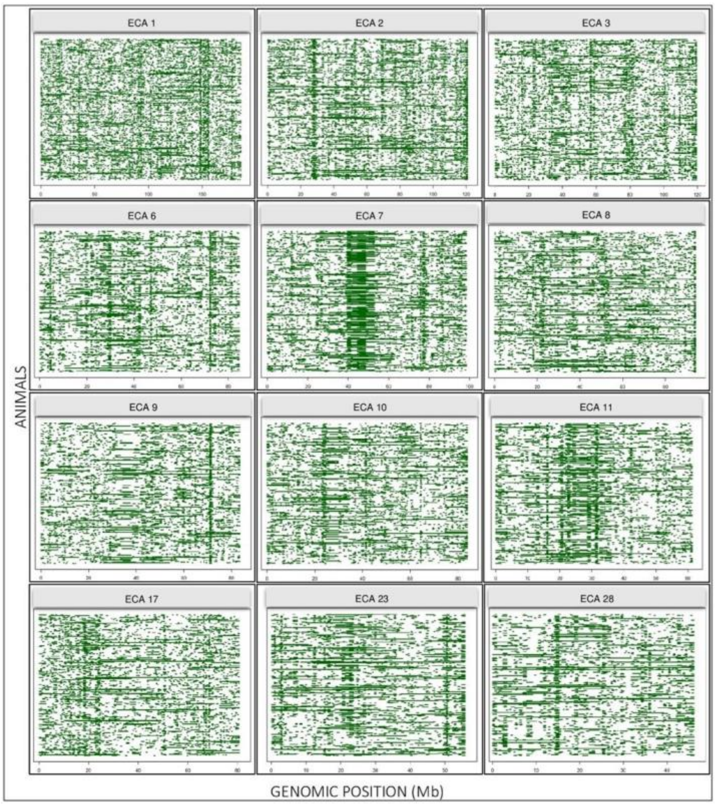


Figure 7. Shared homozygosity Interval for the most representative 12 chromosomes in ROH. Green horizontal lines represent the length of ROH. Based on footprints one can observe regions shared between individuals on population.

2.3.3 Genes and enriched pathways

The enrichment analyses were performed separately for the three methods. Genes with biological processes relevant to the horse were analyzed for pathways, molecular functions, and cellular components. The p-values were adjusted to Benjamini–Hochberg (BH) (P<0.05), which implements methods to analyze and visualize functional profiles of genes and gene clusters (Hu et al., 2010). To the visualization of gene enrichment results, the focus was the biological process that is most relevant to aspects of study.

Complimentary enrichment was performed on PANTHER GO-Slim using functional classification view in gene list, most genes enrichment GO (Gene Ontology) values related to biological process were attributed to cellular and metabolic processes (Fig. 8).

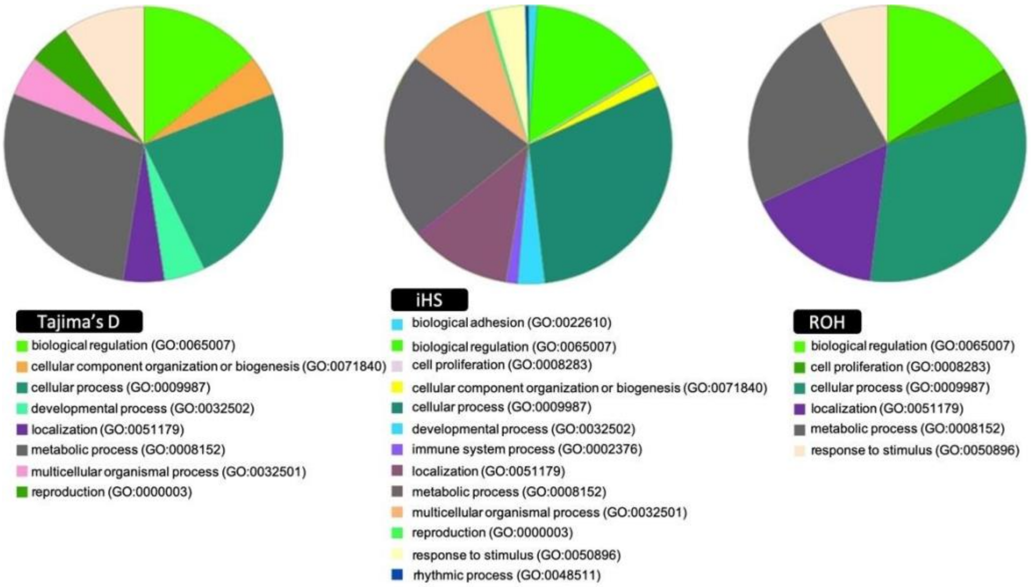


Figure 8. PANTHER GO-Slim pie chart analysis for biological processes for the three methods used for identification of selection signatures.

The results for Tajima's D were 27 genes with hits for 21 biological processes, 12 molecular functions, and 15 cellular components. In iHS were used 316 genes with 299 hits for biological processes, 239 molecular functions, and 210 cellular components. Finally, ROH for 38 genes with 25 hits for biological processes, 33 molecular functions, and 38 cellular components.

2.3.4 Candidates related to gait and locomotor system

In general, there are many foot-fall patterns that quadrupeds could use during locomotion. The gaits are generally considered to be discrete patterns of foot-falls and are divided into symmetrical and asymmetrical (Robilliard et al., 2007). The *DMRT3* gene in MM was associated with gait type, but does not control their gait ability, because it is not the only locus responsible for the lateral gait pattern. As a result, both batida and picada gait needs more studies and discoveries. Some signals of selections were identified in the MM horse genome (P>0.05) involving the genes: *GLI3*; *HOXC9*; *HOXC6*; *HOXC5*; *HOXC4*; *HOXC13*; *HOXC11*; *HOXC10* – "anterior/posterior pattern specification" (GO:0009952), and *GLI3; HOXC13; HOXC11; HOXC10; RC3H* – "limb development" (GO:0060173); *CCL19* and *MAP3K6* – "embryonic limb morphogenesis" (GO:0030326), "embryonic skeletal system development" (GO:0048706), "proximal/distal pattern formation" (GO:0009954), "activation of JUN kinase (JNK) activity" (GO:0007257), "regionalization" (GO:0003002), and "pattern specification process" (GO:0007389).

HOX genes define the axial position of the limb-forming fields, directly activating transcription of the forelimb initiation gene (Tanaka, 2016), *GLI3* is a major regulator of Hedgehog signaling during limb development (Hayashi et al., 2016). It gives us evidence that these genes are regulating the limbs formation and others process associated with locomotor system. Besides that, studies have found that JNK activity increased only in leg exercises (Thompson, 2013), which may be associated with the performance of the two gaits types (batida and picada) performed by MM. Also, the JNK activity, for being composed of a group of mitogen-activated protein, participates in several signal transduction events mediating specific cellular functions (Cuevas et al., 2007). The Fig. 9 improves the visualization of these genes, as well as for others not commented above.

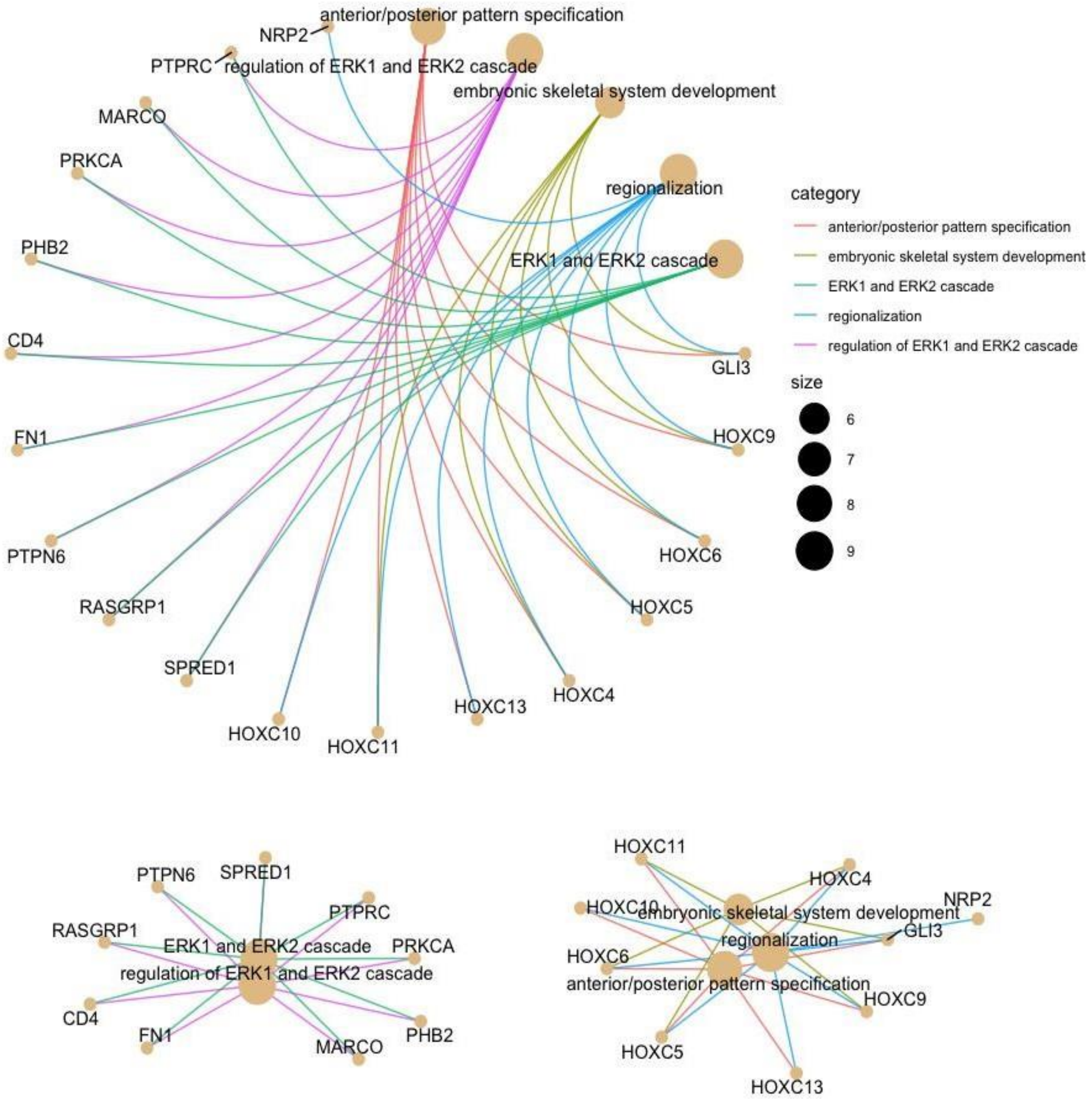

anterior/posterior pattern specification
NRP2
PTPRC
regulation of ERK1 and ERK2 cascade
embryonic skeletal system development
MARCO
PRKCA
regionalization
PHB2
ERK1 and ERK2 cascade
CD4
GLI3
FN1
HOXC9
PTPN6
HOXC6
RASGRP1
HOXC5
SPRED1
HOXC4
HOXC13
HOXC10
HOXC11
category
anterior/posterior pattern specification
embryonic skeletal system development
ERK1 and ERK2 cascade
regionalization
regulation of ERK1 and ERK2 cascade
size
6
7
8
9
PTPN6
SPRED1
PTPRC
RASGRP1
ERK1 and ERK2 cascade
PRKCA
regulation of ERK1 and ERK2 cascade
CD4
FN1
MARCO
PHB2
HOXC11
HOXC4
HOXC10
embryonic skeletal system development
NRP2
GLI3
regionalization
HOXC6
anterior/posterior pattern specification
HOXC9
HOXC5
HOXC13

Figure 9. Functional annotation for top 5 significant biological functions for candidates related to gait, and locomotor system (P<0.05).

2.3.5 Candidates related to energy, exercise, and athletic performance.

It was expected that signatures of selection were flanking these genes, as these animals are extremely dependent on energetic functions for full performance, especially the athletic horse: *ENO2*; *TPI1* and *GAPDHS* – “NADH regeneration” (GO:0006735), “canonical glycolysis” (GO:0061621), “glucose catabolic process to pyruvate” (GO:0061718), “glycolytic process through fructose-6-phosphate”

(GO:0061615), “glycolytic process through glucose-6-phosphate” (GO:0061620), and “glucose catabolic process” (GO:0006007); *MAPK14* – “response to muscle stretch” (GO:0035994), “positive regulation of myoblast differentiation” (GO:0045663) and “skeletal system morphogenesis” (GO:0048705); *GGT5*, *MAPK14* and *ELOVL5* – “fatty acid metabolic process” (GO:0006631); *RYR1* and *MYORG* – “skeletal muscle fiber development” (GO:0048741); *SLC9A1* and *CD4* – “positive regulation of calcium-mediated signaling” (GO:0050850); *FOXO1* – “regulation of cardiac muscle hypertrophy in response to stress” (GO:1903242); *FOXO1* and *CCL19* – “response to bronchodilator” (GO:0097366); *CCL19* and *WASHC1* – “regulation of lipid kinase activity” (GO:0043550).

The *ELOVL5* gene has been classified into many functions associated with energy production from fatty acids (GO:1901570, GO:0030497, GO:0042761, GO:1901568, GO:0035338, GO:0045723, GO:0035336, GO:0000038, GO:0046949, GO:0045923). Besides that, several other genes with major importance were identified and can be highlighted: *COCH* – "bone and cartilage morphogenesis" (GO: 0003433 and GO: 0003429, respectively); *COCH* and *MAPK14* – "skeletal system morphogenesis" (GO: 0048705); *SLC26A8* – “sperm training” (GO: 0048240);*LRRC6* and *DNAI1* – "sperm motility" (GO: 0003341, GO: 0097722, GO: 0030317), and others functions associated with the immune system (GO: 0001771, GO: 0002313, GO: 0002827, GO: 0002285, GO: 0002825) for many genes.

2.3.6 Integrative gene networks

We merge 169 above-pruned genes for an Integrative gene networks (TD=27, iHS=104, and ROH=38 genes) from all methods to conduct the network analysis (known and predicted protein-protein interactions). Each gene annotated in the previous stages of gene enrichment was used, except for iHS, which due to the high number of signals, only contemplated the pruned genes according to the three criteria established into iHS results. Where the interactions correspond at direct (physical) and indirect (functional) associations. During the analysis, the STRING identifiers could not map 16 genes, and nine of them were common among the approaches, so totalizing 144 genes. Thus, the network analysis used 144 genes, which were noticeable high relationships between most of the genes under association for two clusters of genes (Fig. 10).

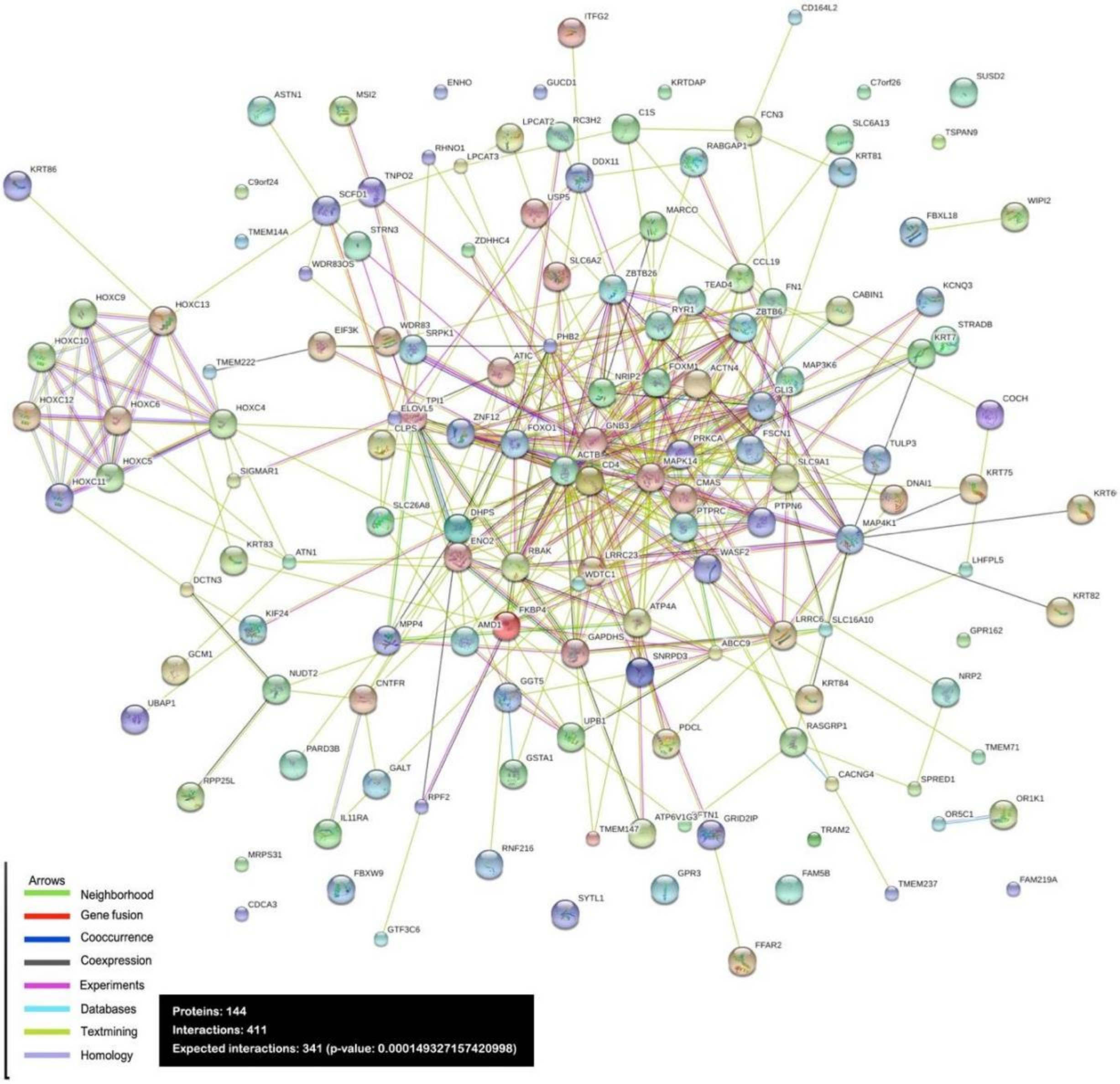


Figure 10. Interaction networks of candidate genes identified from signatures of selection. Different colored arrows indicate the types of evidence used in predicting the associations.

2.4 Discussion

This study provided evidence regarding the genetic background stored on the most current selection in the MM genome, being the first study in the literature related to signals of selection in the breed, able to shed evidence to an original knowledge about the possible candidate gene/gene groups for those regions undergoing selection.

In general, Tajima's D data suggested that MM is under strong balancing selection. However, many hitchhiking effects were highlighted into other statistics, based on extended haplotype homozygosity and/or footprints on homozygous regions. The pronounced balancing status in the population supported by Tajima's D result was an interesting consequence, possibly explained by the nonexistence of any genetic improvement program in the breed.

The artificial selection is based exclusively on competition where the record of gaited performance is always evaluated relative to that of competitors, being often an empirical selection. Thus, we presume that is necessary strong artificial selection to farfetched a possible gait type segregation to well-defined lineages, but first of all, it is important to understand which genes are most relevant to accomplish such goals. According to Arnason et al. (Arnason, 2000), the Thoroughbred carried out a long history of artificial selection for galloping speed, while being ridden by a jockey, and maybe for MM will not be so distinctive to reach well-defined lineages.

At first glance, the *DMRT3* was defined as the only one gene capable to elucidate the gait phenotype variation (Andersson, 2011). Nevertheless, other discoveries reported alleles related to the type of gait were differently fixed within each gait (Promerov et al, 2014). In MM no significant results of the DMRT3 were obtained as selection signature ($P<0.01$). Further, this is not the first study discussing the genetic complexity of locomotion in horses, it was verified in Icelandic horses that no SNP demonstrated genome-wide significance, implying that the ability to pace goes beyond the presence of a single gene variant (Fegraeus et al., 2017).

Identification of genomic regions modified by positive selection has provided discoveries on the adaptive directions of species, being today one of the main theoretical and applied evolutionary studies (Nielsen et al., 2007). In the present study, the iHS statistic was chosen by powerful identification of the recent selection signatures. In Tajima’s D, some limitations were considered, addressing only the aspects relevant to the objectives of this study. And finally, ROH, which proves to be an interesting method to apply with others that aims to browse recent signals of signatures. However, due to the density and complexity of biological information, still unable to exploit the full potential of each method.

In our results, at first, we seek to verify the reproducibility between the methods employed. Common results were found between iHS and ROH to nine genes – *C9orf24, CNTFR*, *DNAI1*, *ENHO*, *DCTN3*, *FAM219A*, *RPP25L*, *SIGMAR1* and *RASGRP1,* being eight of them located on ECA23, and the last of the above sequence belonging from ECA1. These genes are at ~ 28 Mb away from *DMRT3*, which does not rule out the possibility of some genetic relationship. In the network analysis apart, including the *DMRT3*, only one low co-expression was found between *DCTN3* and *DMRT3*. Thus, we rule out a possible physiological relationship of the eight genes with *DMRT3*. However, we recognize that the limitations for non-model species may have interfered. Besides that, according to Ma et al. (2015), during evolution, a series of unknown demographic events further increased the difficulty in detecting modified genomic regions due to different selective pressures.

In ROH, the longest shared homozygosity was identified on ECA7. On the other hand, short intervals were more abundant possibly by the recombination that has already caused its reduction (Santos et al., 2019; Stapley et al., 2017). We observed two main pieces of evidence for this long ROH on ECA7. The first evidence consider the quest for high sports traits performance in the horse, thus, being a recent positive selection based on the intense artificial selection. In a second view, strong bottlenecks occurred in this region during the breed formation. Ablondi et al. (2019) found similar results on ECA7 in Swedish Warmblood horses and Exmoor ponies. Thus, due to the reproducibility of the similar results in this region, we speculate indeed that is possibly a previous bottleneck and not a recent positive selection, corroborating with Ablondi's argument about the Exmoor ponies in a possible intense bottleneck, but generalizing as a common moment to the horse evolution process. Four genes were highlighted in the longest shared homozygosity (*TRIR*, *TNPO2*, *WDR83* and *WDR83OS*) identified under biological functions for localization (GO:0051179) and metabolic process (GO:0008152).

A pruned candidates genes group has been identified as the potential to gait, locomotor system, exercise, athletic performance, energy metabolism, skeletal system, reproduction, morphogenesis, keratinization, and immunological system. We will cover their respective aspects in sequence. A total of 11 candidate genes were identified and associated with aspects of the gait and locomotor system, being eight of them regulating anterior/posterior pattern specification. The HOX genes encode homeodomain transcription factors in the development of many embryonic structures in vertebrates and invertebrates (Akam et al., 1995). According to Capdevila et al. (2001) as development progresses, tight spatial and temporal control of gene expression and cellular behavior sculpts the developing embryo, adding specific morphological and functional characteristics that determine the adult animal's lifestyle and functionality.

The *GLI3* gene was identified under the same HOX gene group to the regulation anterior/posterior pattern specification. Exploring this information, we find that *GLI3* is a transcriptional activator and a repressor to the sonic hedgehog pathway, and also plays an import role in limb development, being described in the literature as an embryonic patterning of human limbs and other structures (Wang et al., 2000). In addition, it is already known, the relationship between the HOX genes and limb musculoskeletal development. Pineault et al. (2014) suggested that integration of the musculoskeletal system is regulated, at least in part, by HOX function in the stromal connective tissue, and play critical roles in skeletal patterning throughout the axial and appendicular skeleton. Grilz-Seger et al. (2019) studying a set of European and Near Eastern Horse Breeds found several GO terms were shared by more than one breed. Where high significance levels were reached for the GO terms anterior/posterior pattern specification (GO:0009952), embryonic skeletal system morphogenesis (GO:0048704) and sequence-specific DNA binding (GO:0043565), mainly based upon the HOXB-cluster in the breeds Gidran, Lipizzan, Posavina, and Noriker.

Another major signal was found for *CCL19* and *MAP3K6* genes with the activation of JUN kinase (JNK) activity. The exercise stimulates c-Jun NH2 Kinase Activity and c-Jun transcriptional activity in human skeletal muscle, shown that the JNK pathway may serve as a link between contractile activity and transcriptional responses in skeletal muscle (Aronson et al., 1998). Exercises cause selective changes over gene expression, leading to a differentiation in skeletal muscle structure and function, which provides strong evidence that this regulation may be associated with gait type segregation in the skeletal muscle on limbs. The effect of activity during exercise in the c-jun mRNA expression is by phosphorylation of two serine residues through the JNKs in c-Jun transactivation domain, leading to an increase in transcriptional activity (Aronson et al., 1998).

As is well known in the modern horse, the athletic performance has been the target of selection in the recent years in many breeds. Increasingly, a perfect horse is idealized in the most countless sports modalities. Indeed, 17 candidate genes were highlighted under important biological functions to the exercise physiology, energy mechanism, catabolic process, morphogenesis (bone, skeletal system and cartilage) and fertility. However, these genes/functions were not the only one associates at MM performance, as can saw in networks analysis, where genes functions have dependencies for the major part of them, being regulated in sets.

Thus, our results confirmed previously described evidence that the segregation of type of gait in horses MM is a polygenic trait corroborating with many other studies, being that its particularities are already defined since the embryogenesis of limbs. In addition, some candidate genes for signals of selection were highlighted and related to the gait, and we speculate that these events play an anatomical or tissue differentiation difficult to be measured in the limbs and/or some alternative mechanism of differential gene expression for both lineages. Finally, many other important genes were found underlying various biological processes that have association with the MM horse performance, development, health, and reproduction.

2.5 Acknowledgements

This work was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo – Fapesp (Grant 2016/19081-9). Wellington Bizarria dos Santos was the recipient of a Master's fellowship from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Capes (Finance code 001).

2.6 References

Ablondi M, Viklund Å, et al. (2019) Signatures of selection in the genome of Swedish warmblood horses selected for sport performance. **BMC Genomics** 20:717.

Akam M (1995) Hox Genes and the Evolution of Diverse Body Plans. **Philosophical Transactions Biological Sciences** 349:313-319.

Akhunov ED, Akhunova AR, Jan D (2010) Nucleotide diversity maps reveal variation in diversity among wheat genomes and chromosomes. **BMC Genomics** 11:702.

Andersson L, Larhammar M, et al. (2012) Mutations in DMRT3 affect locomotion in horses and spinal circuit function in mice. **Nature** 488:642-646.

Andrade LSA (2011) Herança genética da marcha: matéria técnica. Available at: http://hytalobretas.com.br/site/videos.php. Accessed: 7th March 2019.

Arnason T, Van Vleck L (2000) **Genetic Improvement of the Horse**. In: A. Bowling and A. Ruvinsky, ed., The genetics of the horse. New York: CABI Publishing p. 473–497.

Aronson D, Boppart MD, Dufresne SD, Fielding RA, Goodyear LJ (1998) Exercise stimulates c-Jun NH2 kinase activity and c-Jun transcriptional activity in human skeletal muscle. Biochem. **Biophys. Res. Commun.** 251:106–110.

Associação Brasileira de Criadores de Cavalos da Raça Mangalarga - ABCCMM (2018). Available at: http://leia.abccmm.org.br/revistas/revista89a/. Accessed: 3th July 2019.

Associação Brasileira de Criadores de Cavalos da Raça Mangalarga - ABCCMM (1998). Available at: http://leia.abccmm.org.br/portal/regulamentos/padraodaraca/. Accessed: 29th Jan. 2019.

Avila F, Mickelson JR, Schaefer RJ, Mccue ME (2018) Genome-Wide Signatures of Selection Reveal Genes Associated With Performance in American Quarter Horse Subpopulations. **Front. Genet.** 9:1–13.

Bamshad M, Wooding S (2003) Signatures of natural selection in the human genome. **Nat Rev Genet** 4:99-110.

Beeson SK, Schaefer RJ, Mason VC, McCue ME (2019) Robust remapping of equine SNP array coordinates to EquCab3.0. **Animal Genetics** 50:114–5.

Bertolini F, Servin B, Talenti A. et al. (2018) Signatures of selection and environmental adaptation across the goat genome post-domestication**. Genet Sel Evol** 50:57.

Capdevila J, Belmonte JCI (2001) Patterning Mechanisms Controlling Vertebrate Limb Development. **Annu. Rev. Cell Dev. Biol.** 17:87–132.

Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, Rieder MJ, Nickerson DA (2005) Genomic regions exhibiting positive selection identified from dense genotype data. **Genome Research** 15:1553–1565.

Ceballos FC, Hazelhurst S, Ramsay M (2018) Assessing runs of Homozygosity: a comparison of SNP Array and whole genome sequence low coverage data. **BMC Genomics** 19:106.

Ceballos FC, Joshi PK, Clark DW, Ramsay M, Wilson JF (2018) Runs of homozygosity: Windows into population history and trait architecture. **Nat. Rev. Genet.** 19:220–234.

Chen S, Gan M, Lv H, Jiang R (2017) DeepCAPE: a deep convolutional neural network for the accurate prediction of enhancers. **Bioinformatics** 1-26.

Cuevas B, Abell A, Johnson G (2007) Role of mitogen-activated protein kinasekinase kinases in signal integration. **Oncogene** 26:3159–3171.

Detecting natural selection by empirical comparison to random regions of the genome. **Hum. Mol. Genet.** 18:4853–4867.

Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B (2013) Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. **Genetics** 193:929–941.

Fegraeus KJ, Hirschberg I, et al (2017) To pace or not to pace: a pilot study of four- and five-gaited Icelandic horses homozygous for the DMRT3 'Gait Keeper' mutation. **Animal Genetics** 48:694–697.

Ferencakovic M, Solkner J, Curik I (2013) Estimating autozygosity from high-throughput information: effects of SNP density and genotyping errors. **Genet Sel Evol.** 45:42.

Franceschini A, Szklarczyk D, et al. (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. **Nucleic acids research** 41:D808–D815.

Franz M, Rodriguez H, Lopes C, Zuberi K, Montojo J, Bader GD, Morris Q (2018) GeneMANIA update 2018. **Nucleic acids research** 46:W60–W64.

Gautier M, Klassmann A, Vitalis R (2017) rehh 2:0. A reimplementation of the R package rehh to detect positive selection from haplotype structure. **Molecular Ecology Resources** 17:78-90.

Gautier M, Vitalis R (2012) rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. **Bioinformatics** 28:1176–1177.

Gautier M, Vitalis R (2012) Rehh An R package to detect footprints of selection in genome-wide SNP data from haplotype structure. **Bioinformatics** 28:1176-1177.

Gouveia JJS, Paiva SR, et al. (2017) Genome-wide search for signatures of selection in three major Brazilian locally adapted sheep breeds. **Livestock Science** 197:36-45.

Grilz-seger G, Neuditschko M, et al. (2019) Genome-Wide Homozygosity Patterns and Evidence Horse Breeds. **Genes** 10:491.

Hayashi S, Akiyama R, et al. (2016) Gata6-Dependent GLI3 Repressor Function is Essential in Anterior Limb Progenitor Cells for Proper Limb Development. **PLoS Genet.** 12:1–18.

He F, Wu DD, Kong QP, Zhang YP (2008) Intriguing balancing selection on the intron 5 region of LMBR1 in human population. **PLoS One** 3:3–7.

Hillestad B, Woolliams JA, et al. (2017) Detection of runs of homozygosity in Norwegian Red: Density, criteria and genotyping quality control. **Acta Agric. Scand. Sect. A - Anim. Sci.** 67:107–116.

Hu JX, Zhao H, Zhou HH (2010) False Discovery Rate Control With Groups. **Journal of the American Statistical Association** 105:1215–1227.

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. **Genetics** 74:175-195.

Ma Y, Ding X, et al. Properties of different selection signature statistics and a new strategy for combining them. **Heredity (Edinb)** 115:426–436.

Martinho T (2016) **Mangalarga Marchador do Brasil: a história da raça e suas cavalgadas pelo mundo***.* Comg Editora: p. 228.

Nei M, Maruyama T (1975) Lewontin-Krakauer test for neutral genes. **Genetics** 80:395-395.

Nielsen R (2005) Molecular Signatures of Natural Selection. **J Annual Review of Genetics** 39:197-218.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG (2007) Recent and ongoing selection in the human genome. **Nature reviews** 8:857-868.

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.

Pérez O’Brien AM, Utsunomiya YT, et al. (2014) Assessing signatures of selection through variation in linkage disequilibrium between taurine and indicine cattle. **Genet. Sel. Evol.** 46:1–14.

Pineault KM, Wellik DM (2014) Hox genes and limb musculoskeletal development. **Current osteoporosis reports** 12:420–427.

Pook T, Mayer M, et al. (2019) Improving Imputation Quality in BEAGLE for Crop and Livestock Data. **G3: Genes| Genomes| Genetics** 1-13.

Promerov M, Andersson LS, et al. (2014) Worldwide frequency distribution of the ‘ Gait keeper ’ mutation in the DMRT3 gene. **Animal Genetics** 45:274-282.

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-based linkage analyses. **American journal of human genetics** 81:559–575.

Purfield DC, McParland S, Wall E, Berry DP (2017) The distribution of runs of homozygosity and selection signatures in six commercial meat sheep breeds. **PLoS One** 12:1–23.

Qanbari S, Simianer H (2014) Mapping signatures of positive selection in the genome of livestock. **Livestock Science** 166:1-11.

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

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

Santo BA, Pereira GL, et al. (2019) Genomic analysis of the population structure in horses of the Brazilian Mangalarga Marchador breed. **Livest. Sci.** 229:49–55.

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

Srikanth K, Kim N, et al. (2019) Comprehensive genome and transcriptome analyses reveal genetic relationship, selection signature, and transcriptome landscape of small-sized Korean native Jeju horse. **Sci Rep** 9:16672.

Stapley J, Feulner PGD, Johnston SE, Santure AW, Smadja CM (2017) Variation in recombination frequency and distribution across eukaryotes: patterns and processes. **Phil. Trans. R. Soc. B** 372:20160455.

Szklarczyk, D. et al. (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. **Nucleic acids research** 43:D447–D452.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. **Genetics** 123:585–595.

Tanaka M (2016) Developmental Mechanism of Limb Field Specification along the Anterior-Posterior Axis during Vertebrate Evolution. **Journal of developmental biology** 4:18.

Thompson EA (2013) Identity by descent: variation in meiosis, across genomes, and in populations. **Genetics** 194:301–326 (2013).

U.S. Mangalarga Marchador Association - USMMA (2019) Available at: http://www.namarchador.org. Accessed: 27th November 2018.

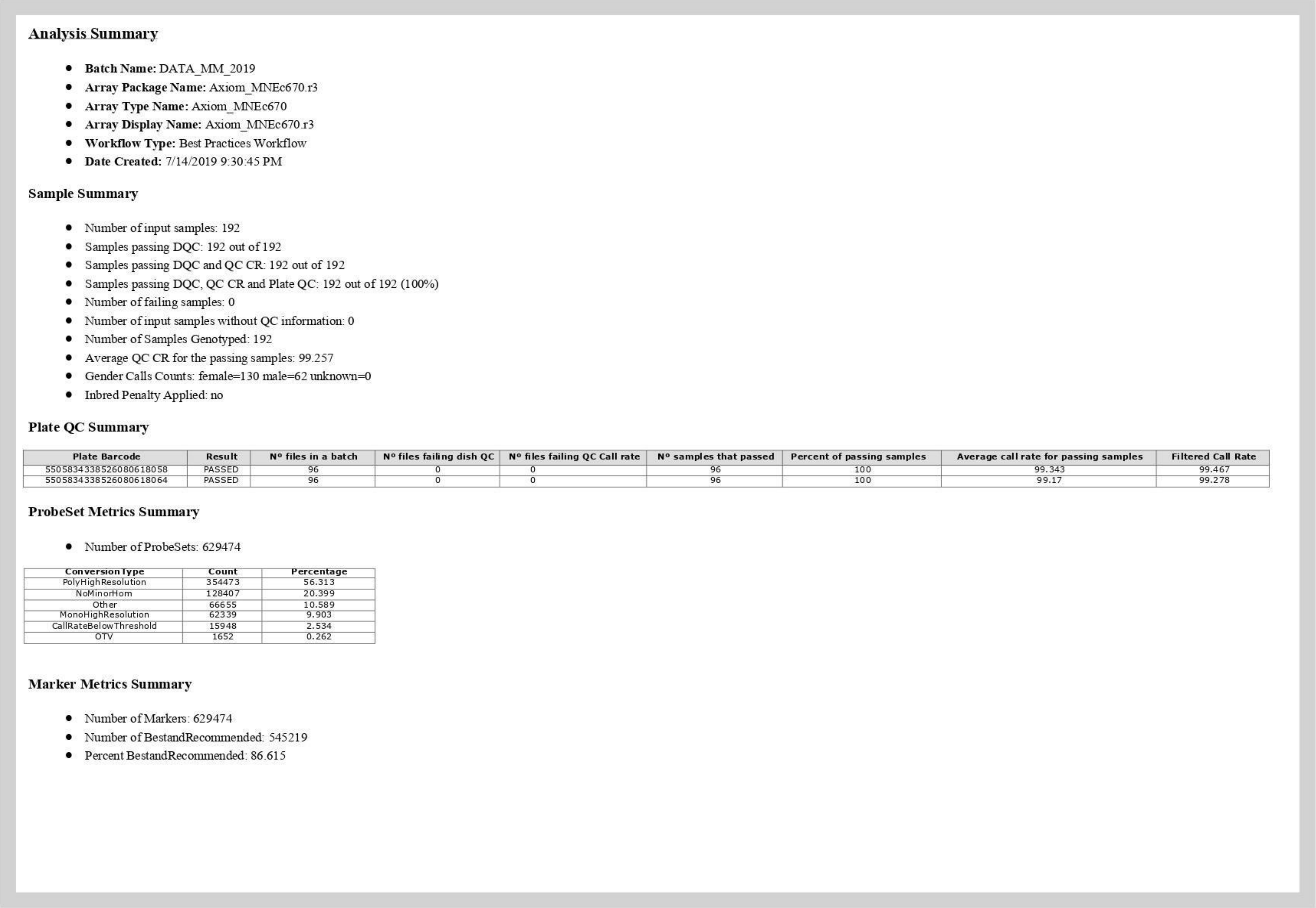
Wang B, Fallon JF, Beachy PA (2000) Hedgehog-Regulated Processing of Gli3 Produces an Anterior/Posterior Repressor Gradient in the Developing Vertebrate Limb. **Cell Press** 100:423–434.

Weigand H, Leese F (2018) Detecting signatures of positive selection in non-model species using genomic data. **Zoological Journal of the Linnean Society** 184:528–583.

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

Yu F, Keinan A, Chen H, Ferland RJ, Hill RS, Mignault AA, Walsh CA, Reich D (2009) Zhang Q, Tyler-Smith C, Long Q (2015) An extended Tajima's D neutrality test incorporating SNP calling and imputation uncertainties. **Statistics and its interface** 8:447–456.

APPENDIX 1



Analysis Summary

- **Batch Name:** DATA_MM_2019
- **Array Package Name:** Axiom_MNEc670.r3
- **Array Type Name:** Axiom_MNEc670
- **Array Display Name:** Axiom_MNEc670.r3
- **Workflow Type:** Best Practices Workflow
- **Date Created:** 7/14/2019 9:30:45 PM

Sample Summary

- Number of input samples: 192
- Samples passing DQC: 192 out of 192
- Samples passing DQC and QC CR: 192 out of 192
- Samples passing DQC, QC CR and Plate QC: 192 out of 192 (100%)
- Number of failing samples: 0
- Number of input samples without QC information: 0
- Number of Samples Genotyped: 192
- Average QC CR for the passing samples: 99.257
- Gender Calls Counts: female=130 male=62 unknown=0
- Inbred Penalty Applied: no

Plate QC Summary

Plate Barcode	Result	Nº files in a batch	Nº files failing dish QC	Nº files failing QC Call rate	Nº samples that passed	Percent of passing samples	Average call rate for passing samples	Filtered Call Rate
5505834338526080618058	PASSED	96	0	0	96	100	99.343	99.467
5505834338526080618064	PASSED	96	0	0	96	100	99.17	99.278

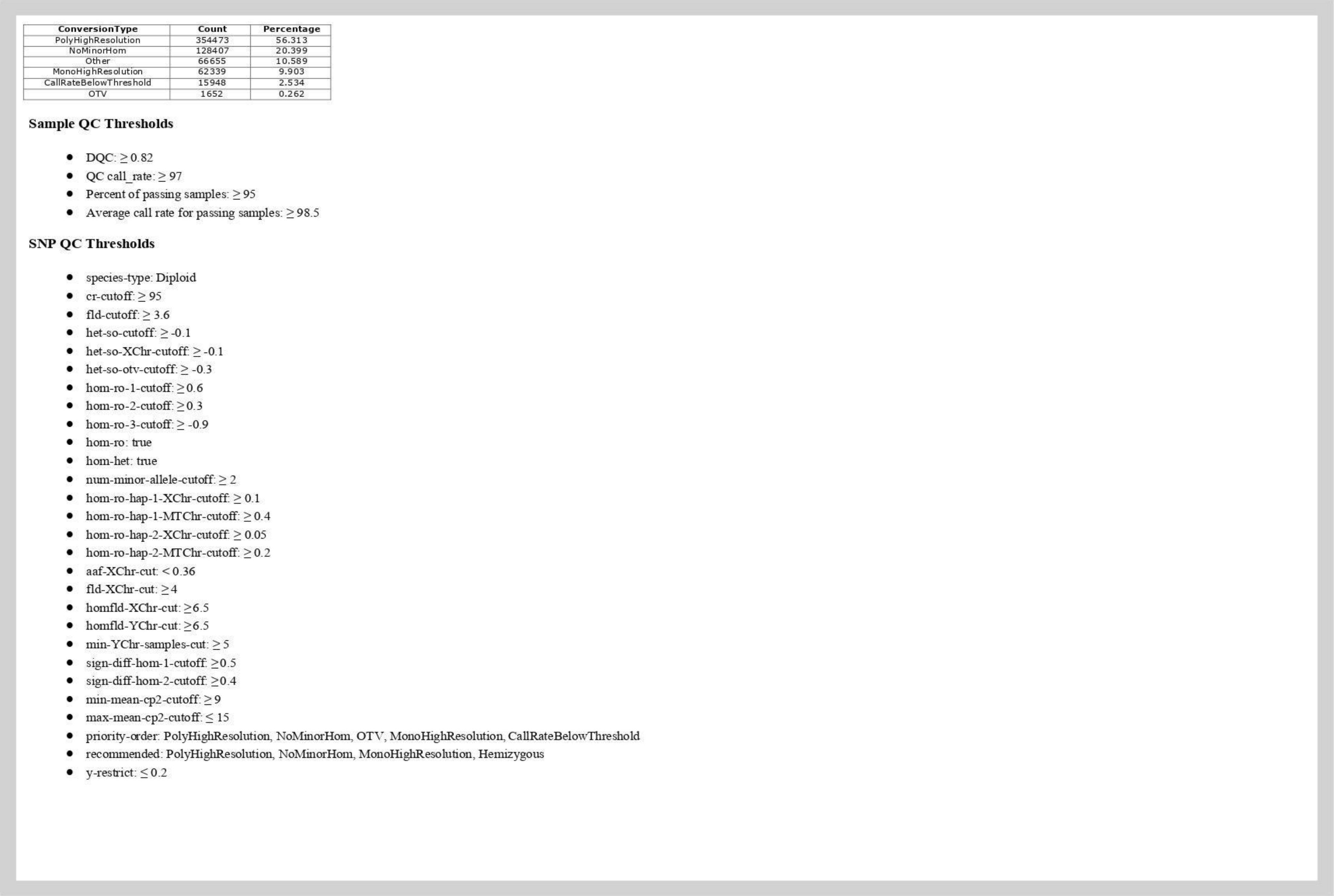
ProbeSet Metrics Summary

- Number of ProbeSets: 629474

ConversionType	Count	Percentage
PolyHighResolution	354473	56.313
NoMinorHom	128407	20.399
Other	66655	10.589
MonoHighResolution	62339	9.903
CallRateBelowThreshold	15948	2.534
OTV	1652	0.262

Marker Metrics Summary

- Number of Markers: 629474
- Number of BestandRecommended: 545219
- Percent BestandRecommended: 86.615



ConversionType	Count	Percentage
PolyHighResolution	354473	56.313
NoMinorHom	128407	20.399
Other	66655	10.589
MonoHighResolution	62339	9.903
CallRateBelowThreshold	15948	2.534
OTV	1652	0.262

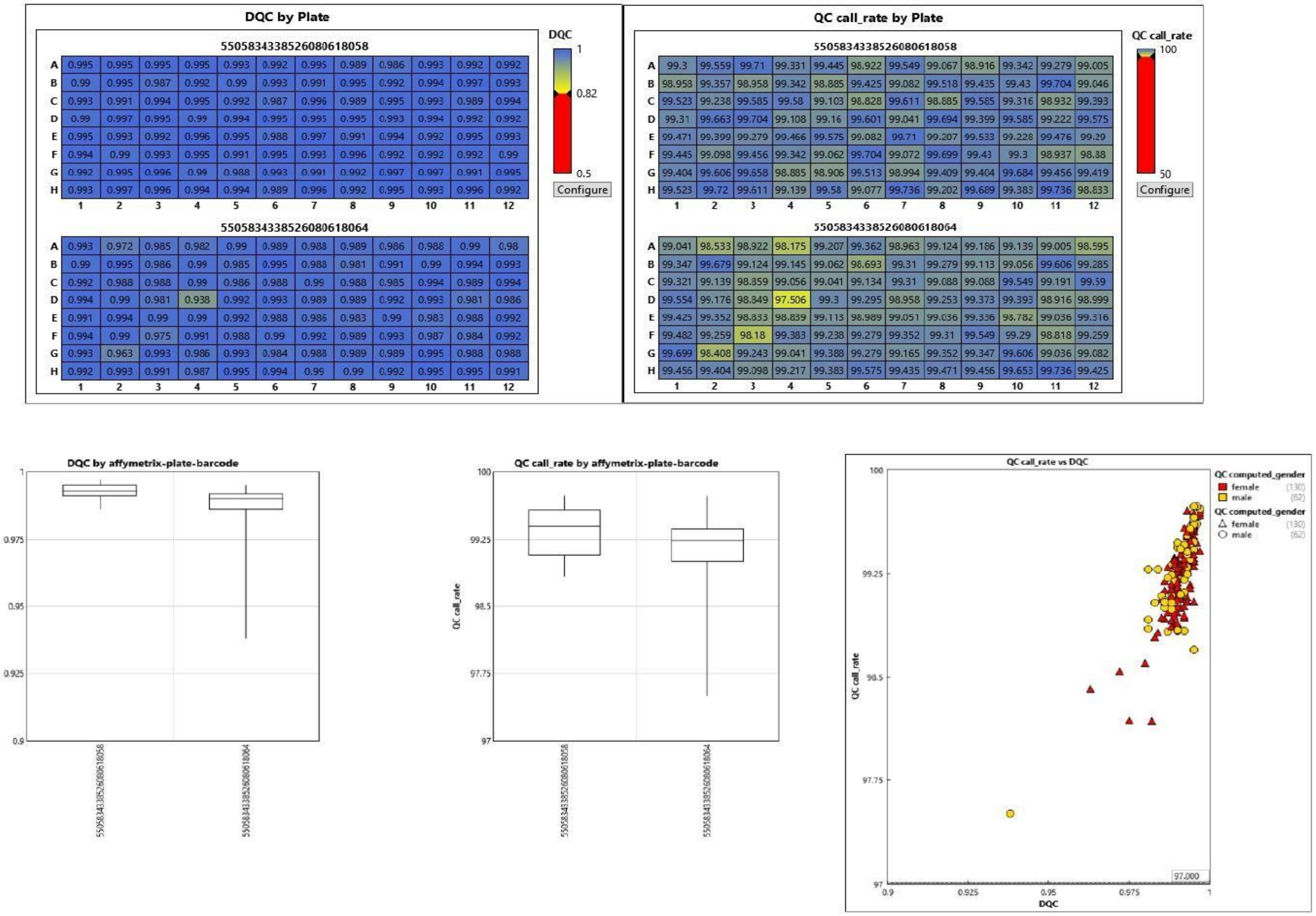
Sample QC Thresholds

- DQC: ≥ 0.82
- QC call_rate: ≥ 97
- Percent of passing samples: ≥ 95
- Average call rate for passing samples: ≥ 98.5

SNP QC Thresholds

- species-type: Diploid
- cr-cutoff: ≥ 95
- fld-cutoff: ≥ 3.6
- het-so-cutoff: ≥ -0.1
- het-so-XChr-cutoff: ≥ -0.1
- het-so-otv-cutoff: ≥ -0.3
- hom-ro-1-cutoff: ≥ 0.6
- hom-ro-2-cutoff: ≥ 0.3
- hom-ro-3-cutoff: ≥ -0.9
- hom-ro: true
- hom-het: true
- num-minor-allele-cutoff: ≥ 2
- hom-ro-hap-1-XChr-cutoff: ≥ 0.1
- hom-ro-hap-1-MTChr-cutoff: ≥ 0.4
- hom-ro-hap-2-XChr-cutoff: ≥ 0.05
- hom-ro-hap-2-MTChr-cutoff: ≥ 0.2
- aaf-XChr-cut: < 0.36
- fld-XChr-cut: ≥ 4
- homfld-XChr-cut: ≥ 6.5
- homfld-YChr-cut: ≥ 6.5
- min-YChr-samples-cut: ≥ 5
- sign-diff-hom-1-cutoff: ≥ 0.5
- sign-diff-hom-2-cutoff: ≥ 0.4
- min-mean-cp2-cutoff: ≥ 9
- max-mean-cp2-cutoff: ≤ 15
- priority-order: PolyHighResolution, NoMinorHom, OTV, MonoHighResolution, CallRateBelowThreshold
- recommended: PolyHighResolution, NoMinorHom, MonoHighResolution, Hemizygous
- y-restrict: ≤ 0.2

Axiom™ Analysis Suite final report



DQC by Plate

5505834338526080618058

	1	2	3	4	5	6	7	8	9	10	11	12
A	0.995	0.995	0.995	0.995	0.993	0.992	0.995	0.989	0.986	0.993	0.992	0.992
B	0.99	0.995	0.987	0.992	0.99	0.993	0.991	0.995	0.992	0.994	0.997	0.993
C	0.993	0.991	0.994	0.995	0.992	0.987	0.996	0.989	0.995	0.993	0.989	0.994
D	0.99	0.997	0.995	0.99	0.994	0.995	0.995	0.995	0.993	0.994	0.992	0.992
E	0.995	0.993	0.992	0.996	0.995	0.988	0.997	0.991	0.994	0.992	0.995	0.993
F	0.994	0.99	0.993	0.995	0.991	0.995	0.993	0.996	0.992	0.992	0.992	0.99
G	0.992	0.995	0.996	0.99	0.988	0.993	0.991	0.992	0.997	0.997	0.991	0.995
H	0.993	0.997	0.996	0.994	0.994	0.989	0.996	0.992	0.995	0.993	0.996	0.992

5505834338526080618064

	1	2	3	4	5	6	7	8	9	10	11	12
A	0.993	0.972	0.985	0.982	0.99	0.989	0.988	0.989	0.986	0.988	0.99	0.98
B	0.99	0.995	0.986	0.99	0.985	0.995	0.988	0.981	0.991	0.99	0.994	0.993
C	0.992	0.988	0.988	0.99	0.986	0.988	0.99	0.988	0.985	0.994	0.989	0.994
D	0.994	0.99	0.981	0.938	0.992	0.993	0.989	0.989	0.992	0.993	0.981	0.986
E	0.991	0.994	0.99	0.99	0.992	0.988	0.986	0.983	0.99	0.983	0.988	0.992
F	0.994	0.99	0.975	0.991	0.988	0.99	0.992	0.989	0.993	0.987	0.984	0.992
G	0.993	0.963	0.993	0.986	0.993	0.984	0.988	0.989	0.989	0.995	0.988	0.988
H	0.992	0.993	0.991	0.987	0.995	0.994	0.99	0.99	0.992	0.995	0.995	0.991

QC call_rate by Plate

5505834338526080618058

	1	2	3	4	5	6	7	8	9	10	11	12
A	99.3	99.559	99.71	99.331	99.445	98.922	99.549	99.067	98.916	99.342	99.279	99.005
B	98.958	99.357	98.958	99.342	98.885	99.425	99.082	99.518	99.435	99.43	99.704	99.046
C	99.523	99.238	99.585	99.58	99.103	98.828	99.611	98.885	99.585	99.316	98.932	99.393
D	99.31	99.663	99.704	99.108	99.16	99.601	99.041	99.694	99.399	99.585	99.222	99.575
E	99.471	99.399	99.279	99.466	99.575	99.082	99.71	99.207	99.533	99.228	99.476	99.29
F	99.445	99.098	99.456	99.342	99.062	99.704	99.072	99.699	99.43	99.3	98.937	98.88
G	99.404	99.606	99.658	98.885	98.906	99.513	98.994	99.409	99.404	99.684	99.456	99.419
H	99.523	99.72	99.611	99.139	99.58	99.077	99.736	99.202	99.689	99.383	99.736	98.833

5505834338526080618064

	1	2	3	4	5	6	7	8	9	10	11	12
A	99.041	98.533	98.922	98.175	99.207	99.362	98.963	99.124	99.186	99.139	99.005	98.595
B	99.347	99.679	99.124	99.145	99.062	98.693	99.31	99.279	99.113	99.056	99.606	99.285
C	99.321	99.139	98.359	99.056	99.041	99.134	99.31	99.088	99.088	99.549	99.191	99.59
D	99.554	99.176	98.849	97.506	99.3	99.295	98.958	99.253	99.373	99.393	98.916	98.999
E	99.425	99.352	98.833	98.839	99.113	98.989	99.051	99.036	99.336	98.782	99.036	99.316
F	99.482	99.259	98.18	99.383	99.238	99.279	99.352	99.31	99.549	99.29	98.818	99.259
G	99.699	98.408	99.243	99.041	99.388	99.279	99.165	99.352	99.347	99.606	99.036	99.082
H	99.456	99.404	99.098	99.217	99.383	99.575	99.435	99.471	99.456	99.653	99.736	99.425

DQC by affymetrix-plate-barcode

QC call_rate by affymetrix-plate-barcode

QC call_rate vs DQC

CHAPTER 3 - FINE-SCALE ESTIMATION OF INBREEDING RATES, RUNS OF HOMOZYGOSITY AND GENOME-WIDE HETEROZYGOSITY LEVELS IN THE MANGALARGA MARCHADOR BREED

ABSTRACT – With the availability of high-density SNP panels and the establishment of approaches for characterizing homozygosity and heterozygosity sites, it is possible to access fine-scale information regarding genomes, providing more than just comparisons of different inbreeding coefficients. This is the first study that seeks to access such information for the Mangalarga Marchador (MM) horse breed on a genomic scale. To this end, we aimed to assess inbreeding levels using different coefficients, as well as to characterize homozygous and heterozygous runs in the population. Using Axiom ® Equine Genotyping Array – 670k SNP (Thermo Fisher, USA), 192 horses were genotyped. Our results showed different estimates: inbreeding from genomic coefficients (FROH) = 0.16; pedigree-based (FPED) = 0.008; and a method based on excess homozygosity (FHOM) = 0.010. The correlations between the inbreeding coefficients were low to moderate, and some comparisons showed negative correlations, being practically null. In total, 85,295 runs of homozygosity (ROH) and 10,016 runs of heterozygosity (ROHet) were characterized for the 31 horse autosomal chromosomes. The class with the highest percentage of ROH was 0–2 Mbps, with 92.78% of the observations. In the ROHet results, only the 0–2 class presented observations, with chromosome 11 highlighted in a region with high genetic variability. Three regions from the ROHet analyses showed genes with known functions: tripartite motif-containing 37 (TRIM37), protein phosphatase, Mg2 + / Mn2 + dependent 1E (PPM1E), and carbonic anhydrase 10 (CA10). Therefore, our findings suggest moderate inbreeding, possibly attributed to breed formation, annulling possible recent inbreeding. Furthermore, regions with high variability in the MM genome were identified (ROHet), associated with the recent selection and important events in the development and performance of MM horses over generations.

Keywords: Equus caballus, FHOM, FPED, FROH, ROH, ROHet

3.1 Introduction

Artificial selection, which has been practiced for many years and has seen new paths develop since the formation of the Brazilian Association of Mangalarga Marchador Horse Breeders (ABCCMM), has led to the improvement of many phenotypes in the Mangalarga Marchador (MM) horse breed. Such horses have dominated the attention of horse breeders in recent decades due to their gaited phenotype. A particularity of this breed is its intermediate-speed gait, which differs from trotting. This gait is known as "marcha," subdivided into "marcha batida" and "marcha picada."

Recently, in the MM was suggested polygenic control over gait types, corroborating with the findings of Patterson et al. (2015) and Fonseca et al. (2017) (see Bussiman et al., 2019). In addition, several candidate genes have been identified and associated with signatures of selection for both gait types (Chapter 2). Nevertheless, and despite the relevance of the breed to Brazil, no study has yet evaluated genomic inbreeding involving one of the most common horse breeds in the country.

Inbreeding occurred by the mating between close relatives which increases offspring homozygosity and usually results in reduced fitness (Pekkala et al., 2014). The consequences of inbreeding are numerous, but the most prominent ones are associated with genetic variation (Hedrick and Kalinowski, 2000; Nowak et al., 2007), fertility (Robert et al., 2009), and the accumulation of recessive lethal genetic mutations (Bull, 2017), leading to significant, broad aspect important impacts on the species/breeds. According to Marras et al. (2015) inbreeding is inevitable in populations under selection, as only a subset of individuals is used for breeding, and strategies to restrict inbreeding are an essential requirement.

Studies have shown the superiority of the genomic approaches to access inbreeding in a given population (Kardos et al., 2015; Forutan et al., 2018; Ablondi et al., 2019), and pedigree-based inbreeding has increasingly gained a comparative role because of its limitations. The Runs of homozygosity (ROH) have been explored for two main purposes, consanguinity estimation (Trevor et al., 2015; LI et al., 2011) and detection of genomic signatures of selection (Ablondi et al., 2019; Xu et al., 2019, Chapter 2). ROH are continuous homozygous segments of the DNA sequence (Peripolli et al., 2018). This event arises when two copies of an ancestral haplotype are brought together in an individual, forming a probable autozygous, i.e. homozygous haplotype by descent (Ceballos et al., 2018).

The ROH size are inversely correlated with its age: longer ROH originates from recent common ancestors while shorter ROH comes from distant ancestors, being the genetic recombination one of the main breakdown factors over the generations (Keller et al., 2011; Gomez-Raya et al., 2015). The genomic inbreeding based on ROH (FROH) provides a range of advantages, and represents the quotient of autozygous regions by total genome length. The derived genomic coefficients from animals/populations can be calculated without pedigree records or incomplete pedigree information. Many horse breeds like the Arabian (Al Abri et al., 2017), Lipizzan (Grilz-Seger et al., 2019), Thoroughbred (Fawcett et al., 2019), Saxon-Thuringian (Metzger et al., 2015), Norik of Muran (Kasarda et al., 2019) have been their inbreeding coefficients studied in a genomic-wide view.

In addition to homozygosity, we have the inferences of heterozygosity runs that in diploid organisms are single nucleotide differences observed between paternal and maternal chromosomes called heterozygous sites (Renaud et al., 2019). According to Marras et al. (2018), these are not actual “runs”, but rather heterozygosity-rich regions. Evidences suggest that increased heterozygosity over time may be attributed to selection (Kaeuffer et al., 2007). In addition, heterozygosity can reveal much about the population structure and demographic history (population size problems, bottlenecks, metapopulation dynamics, genetic variability, mixing of two previously isolated populations, etc.). Renaud et al. (2019) while developing a Bayesian framework to estimate local and global rates of heterozygosity (ROHan) to jointly estimate for heterozygosity and long ROH, tested the method on several horse samples (modern and ancient). One of them, endangered Przewalski’s horse brings forward a large fraction of ROH and low heterozygosity, what according to the author is a contrast to their Eneolithic direct ancestors, which showed larger genetic diversity and were not found to be inbred.

Considering that almost nothing is known about MM on a genomic scale, the objective of this study was to obtain (using high-density SNP genotype information of the 670k Affymetrix Axiom Equine Genotyping Array) the breed's inbreeding coefficients. This was achieved through the use of traditional and genomic-based approaches, as well as through correlations between the coefficients and assessment of the runs of homozygosity and heterozygosity.

3.2 Materials and Methods

3.2.1 Ethical statement

The Management and treatment of the animals during blood cell extraction were approved by the Ethics Committee on the Use of Animals (CEUA) at Sao Paulo State University (UNESP) - FMVZ (approval number: 0029/2017).

3.2.2 Sample collection, SNP genotyping and quality control

A total of 192 animals of both sexes - males (n=62) and females (n=130) - were analyzed. All the samples were obtained from Brazil during the 36th Brazilian National Exhibition of the Mangalarga Marchador breed, as well as from stud farms in the states of São Paulo and Minas Gerais. Besides, we sought to include the two gait types present in the breed with 86 picada and 106 batida gait animals.

Samples were genotyped using the 670k Axiom ® Equine Genotyping Array (Thermo Fisher, USA) from DNA extracted from blood samples. The parameters used for the data pruning in the platform Axiom™ Analysis Suite (Thermo Fisher, USA) is default for diploid organisms in the version 4 sample QC: DQC ≥ 0.82, call rate ≥ 97, percent of passing samples ≥ 95, average call rate for passing samples ≥ 98.5; and SNP QC thresholds: call rate ≥ 97, and plus twenty-six other parameters for diploid organisms (standard protocol) that can be consulted with more details in a previous study already reported for the same database with the same quality control (Chapter 2). We also include in this study, the recent equine genome assembly updated with SNP array coordinates remapped to EquCab3.0 (Beeson et al., 2019). The coordinates can be easily accessed from NCBI (https://www.ncbi.nlm.nih.gov/genome/tools/remap).

3.2.3 Complementary genotyping quality control

Complementary quality control was performed after Axiom™ Analysis Suite pruning in VCFtools software to Hardy-Weinberg ≤ 1e-8, and MAF of 0.005. Additionally in R software, non-autosomal chromosomes were removed, the database being sorted by chromosomes and positions. We decided not to work with sex chromosomes in this research. Allosomes present a different effective population size (Sayres et al., 2018), and specific problems with respect to an efficient analysis of mixed-sex population studies (Clayton et al., 2009).

There is no consensus in the literature about the best parameters for MAF in quality control for estimation of heterozygosity, runs of homozygosity, and inbreeding, but most publications have been reported under the range of 0.01 to 0.05 (Kim et al., 2015; Purfield et al., 2017; Ablond et al., 2019; Peripolli et al., 2018), or even without the use of any MAF parameter (Ferencakovic et al., 2013b). Lencz et al. (2007) reported some ascertainment biases by the inclusion of SNPs with high minor allele frequencies. Thus, we decide to apply a low parameter for MAF (0.005), but not absent. The final markers density was of 444.929 SNPs.

3.2.4 Pedigree-based estimates of inbreeding

Pedigree information was collected from the website of the Brazilian Association of Mangalarga Marchador Horse Breeders (ABCCMM) for the 192 animals. The pedigree consisted of 1397 individuals with a depth of four generations. The traditional inbreeding coefficients (F_{PED}) were calculated using the Inbupgf90 software (http://nce.ads.uga.edu/wiki/doku.php?id=readme.inbupgf90), which uses a recursive algorithm assuming non-zero inbreeding for unknown parents as presented in Aguilar & Misztal (2008), based on VanRaden (1992).

3.2.5 Genomic estimates of inbreeding

Two genomic inbreeding coefficients were calculated. The first based on runs of homozygosity (F_{ROH}), and the second by the excess of homozygosity inbreeding coefficient - differences between the observed and expected number of homozygous genotypes (F_{HOM}). Thus, the genomic inbreeding coefficients , $F_{ROH,}$ and $F_{HOM,}$ were computed by the following equations:

$$\mathrm{FHOM} = \frac{\text{Observed Homozygosis} - \text{Expected Homozygosis}}{\text{Observations} - \text{Expected Homozygosis}}$$

$$\mathrm{FROH} = \frac{\sum_{k} \text{Length}\ (\mathrm{ROH}_{k})}{\mathrm{L}}$$

Where:

k = Number of each individual's ROH multiplied by the average length of ROHs;

L = Total length of the genome.

Two different approaches were used for conducting the ROH: the consecutive-runs (Marras et al., 2015), and sliding-window-based run detection (Purcell et al., 2007). F_{ROH} also was calculated based into five length classes: 0-2, 2-4, 4-8, 8-16, and >16 Mbps. The objective to use two different methodologies is to observe the reproducibility of the results, as well as to discuss the possible advantages and disadvantages of each one. Additionally, Runs of heterozygosity (ROHet) were calculated only in consecutive stretches of heterozygous SNP genotypes. Thus, the consecutive-runs for *(i)* homozygosity and *(ii)* heterozygosity analyses were performed with the R detectRUNS package - recent methodology (Biscarini et al., 2019), and *(iii)* homozygosity sliding-runs in the Plink software - commonly used methodology (Purcell et al., 2007). All respective parameters used are described as follows: *(i)* min number of SNP in a RUN (*minSNP*) = 50; max distance between consecutive SNP in a window to be still considered a potential run (*maxGap*) = 10^6; min length of run in bps (*minLengthBps*) = 250 kb; max number of opposite genotype SNPs in the run (*maxOppRun*) = 1; and max number of missing SNPs in the run (*maxMissRun*) = 1. *(ii) minSNP* = 15; *ROHet* = TRUE; *maxGap* = 10^6; *minLengthBps* = 10 Kb;*maxOppRun* = 2; and *maxMissRun* = 1.*(iii)* min number of SNP in a RUN (*homozyg-snp*) = 50; length of the sliding window (*homozyg-kb*) = 250 kb; min density in a RUN (*hmozyg-density*) = 50 (1 SNP/50 Kb); length between two SNPs to be considered two different segments (*homozyg-gap*) = 1000 kb; number of SNPs for sliding window (*homozyg-window-snp*) = 50; max number of missing SNPs in the run (*homozyg-window-het*) = 2; max number of missing SNPs in the run (*homozyg-window-missing*) = 2; and the proportion of overlapping windows that must be called homozygous to define a given SNP as in a homozygous segment (*homozyg-window-threshold*) = 0.05. Besides, the observed and expected number of homozygous genotypes for F_{HOM} were obtained using the -het command (Purcell et al., 2007).

3.2.6 Candidate gene, pathway and functional analysis

A previous study has reported signals of selection in ROH islands to the same database, focusing exclusively on signatures of selection (Chapter 2). On the other hand, runs of heterozygosity (ROHet) were not included in this study. For these reasons, only genomic regions that shared a frequency ≥ 0.30 were annotated. We do not yet have an established threshold in the literature, not even for ROH islands, which generally uses frequencies ≥ 0.50. Thus, the windows for annotation were performed from the start to the end of ROHet on the BioMart R package (Smedley et al., 2009) using EquCab3.0, and PANTHER Classification System (www.pantherdb.org), using default parameters. The genes identified with known functions were enriched for Biological Processes (BP), Molecular Functions (MF), and Cellular Components (CC) analysis.

3.3 Results

3.3.1 Inbreeding levels

The FROH estimates revealed moderate coefficients (0.16), while the other inbreeding estimates were low: FPED = 0.008, and FHOM = 0.010. In FPED, of the 1,397 pedigree animals, 1,228 individuals (87.90%) had inbreeding levels equal to 0, while only 11 of them (0.79%) were within the range of 18.75–25.00%. FHOM of the 97 individuals (50.52%) showed negative values (-0.0003 to -0.0003), while the remaining 95 animals with positive values were in the range of 0 to 0.246 (49.48%) (Figure 1).

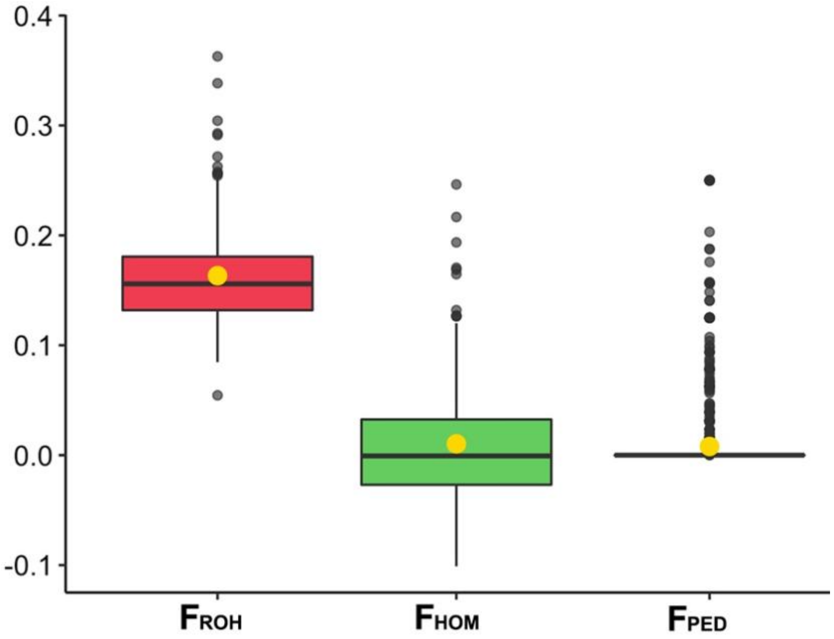


Figure 1. Inbreeding levels for three distinct methods. The yellow circle represents the mean, and the black dash the median.

The correlations between FPED and the other coefficients were low (FPED x FROH = 0.02, and FPED x FHOM = 0.02). The same was true for the comparisons between FROH and FHOM (0.16). Additionally, correlations were found between the FROH classes and FPED and FHOM, with the exception of the 16 Mbps FROH class, which yielded a single observation, preventing us from finding correlations with the other coefficients. Therefore, we decided to remove the FROH class > 16 Mbps, as can be observed in Figure 2. Negative correlations were present in FROH classes 4–8 and 8–16 Mbps for the FPED coefficients; the class 8–16 Mbps also showed the same behavior with FHOM. As expected, strong correlations were present between the FROH 0-2 and FROH 2-4 (0.74), FROH 2-4 and FROH 4-8 (0.87), and FROH 4-8 and FROH 8-16 Mbps classes.

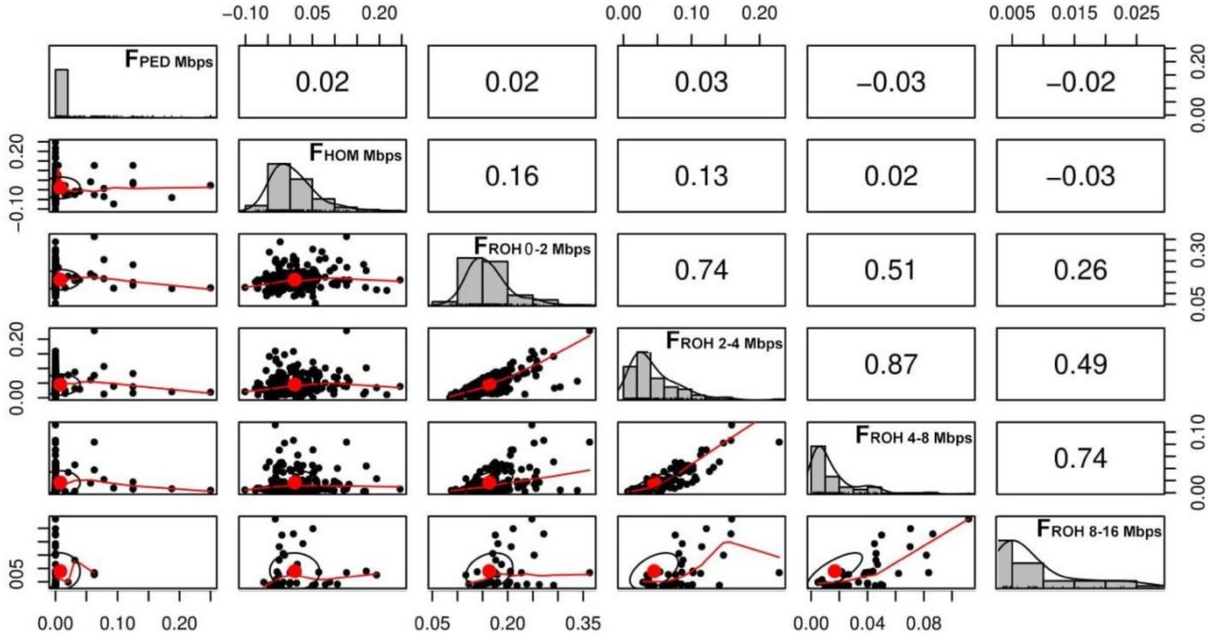
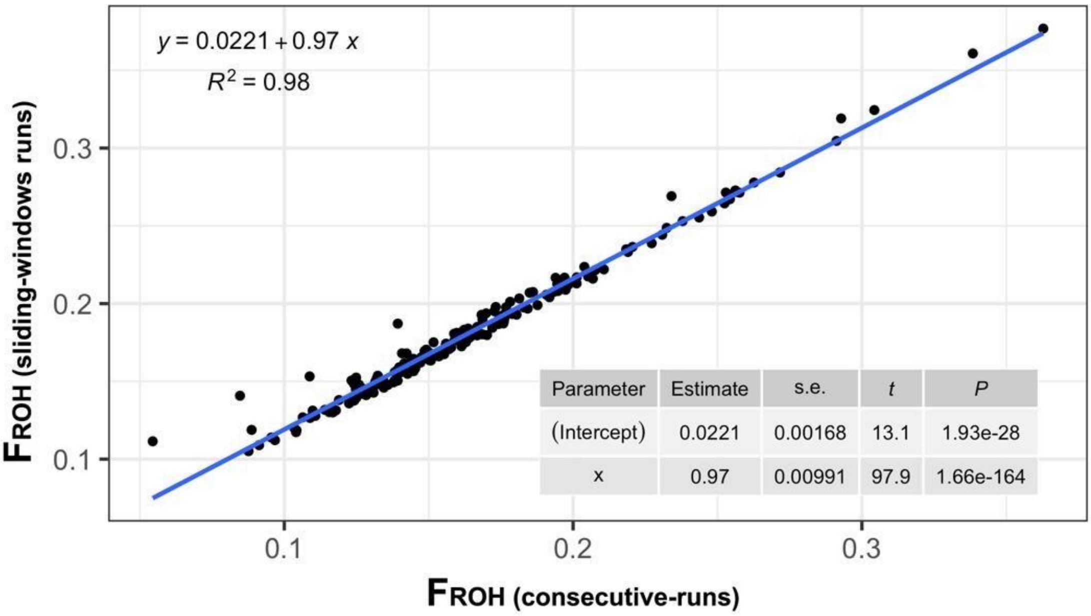


Figure 2. Scatterplots in the bottom-left diagonal, and correlations in the top-right diagonal for inbreeding coefficients F_{ROH} (F_{ROH} 0-2, F_{ROH} 2-4, F_{ROH} 4-8, and F_{ROH} 8-16 Mbps), F_{HOM}, and F_{PED}. The diagonals boxes describe the types of inbreeding coefficients with their corresponding histograms.

3.3.2 Assessment of runs of homozygosity and heterozygosity

Both approaches for calculating genomic inbreeding identified a large number of ROH. The consecutive runs had a total of 85,295, and the sliding windows 67,478. The difference was 17,817 ROH, indicating a probable underestimation of the number of observations. However, due to the large amount of information in ROH, we decided to compare these methods using a simple linear regression analysis for the genome-wide FROH (Figure 3). The axes presented a correlation of high magnitude = 0.99, with an R2 = 0.98, showing the model's good fit.

It is already known that computational approaches using windows or sliding-windows have some analytical bias. The purpose of this comparison was to verify the reproducibility of the methods, reduce information density, as well as to contrast the new statistics with those commonly used. Despite having some different parameters, the results were very similar, and for these reasons, we do not see the necessity to maintain many comparisons. Thus, the study will be conducted only with the consecutive-run test, which proves to be more accurate.



Parameter	Estimate	s.e.	*t*	*P*
(Intercept)	0.0221	0.00168	13.1	1.93e-28
x	0.97	0.00991	97.9	1.66e-164

Figure 3. Simple linear regression analysis for genome-wide F_{ROH} calculated in two distinct methodologies.

The ROHet was less frequent when compared to ROH, with all percentages of SNPs per chromosome below 0.50 (Figure 4). A total of 10,016 ROHet were found, alongside the clear formation of a ROHet island in ECA 11. Regarding ROH, several islands were reported in the Chapter 2, whose exclusive objective was to identify the signature of selection. Therefore, this information was not included in this Chapter. Table 1 shows some descriptive information about ROH and ROHet, with two distinct sections: (A) and (B). The first section (A) correspond to estimates within chromosomes (ECA1 to ECA 31), and in the section (B) the estimates of the different classes of ROH and ROHet (0–2, 2–4, 4–8, 8–16).

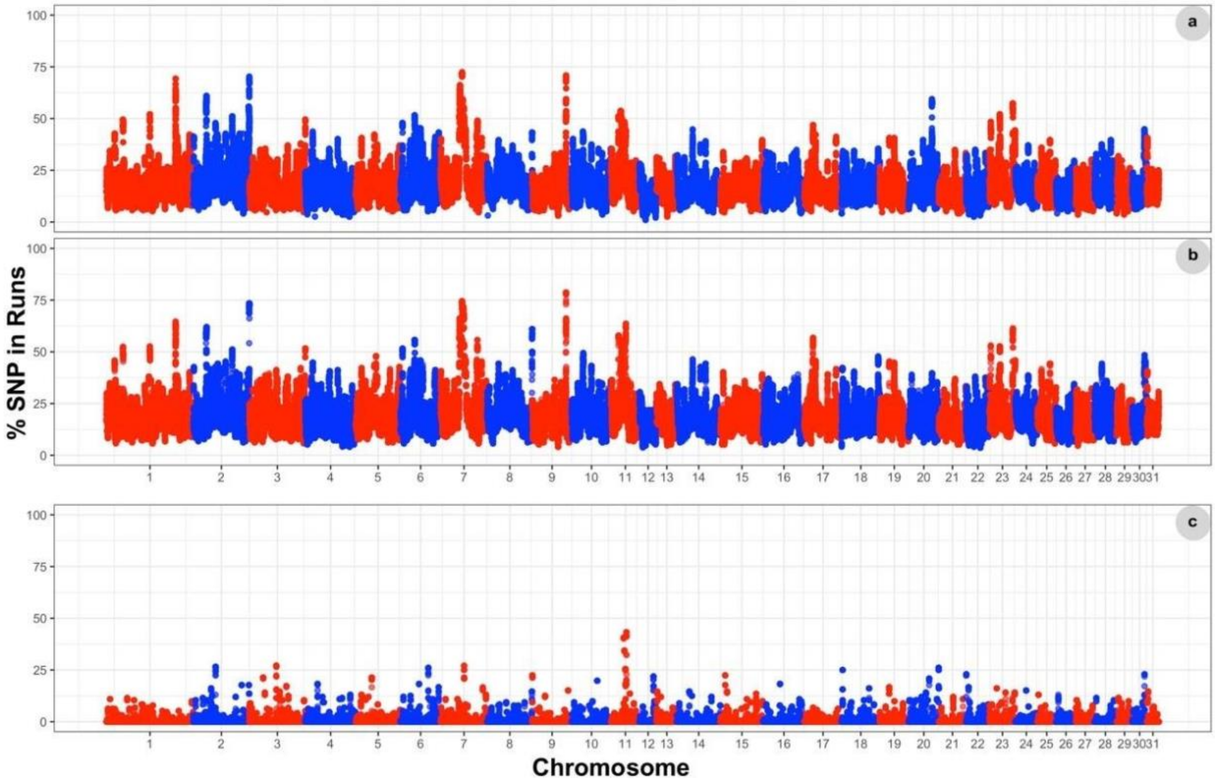


Figure 4. Manhattan plot for ROH in consecutive (a) and sliding-windows runs (b), and ROHet in consecutive-runs (c).

Table 1. Number, percentage and mean for ROH and ROHet per chromosome and length classes.

[A] Chr	ROH			ROHet		
	N. of ROH	*Percentage*	*Mean* (Mbps)	*N. of ROHhet*	*Percentage*	*Mean* (Mbps)
ECA 1	7,339	0.08604256	0.7735226	576	0.05750799	0.06545312
ECA 2	5,699	0.06681517	0.7382070	620	0.06190096	0.05719704
ECA 3	4,367	0.05119878	0.8319847	587	0.05860623	0.09789695
ECA 4	3,906	0.04579401	0.8000352	440	0.04392971	0.09345371
ECA 5	3,315	0.03886512	0.9132927	376	0.03753994	0.07685551
ECA 6	4,051	0.04749399	0.6721837	586	0.05850639	0.03886768
ECA 7	3,864	0.0453016	1.1130141	444	0.04432907	0.10951935
ECA 8	3,558	0.04171405	0.9558703	358	0.03574281	0.09531366
ECA 9	2,917	0.03419896	0.8808950	259	0.02585863	0.09475750
ECA 10	3,137	0.03677824	0.8740523	375	0.0374401	0.07378232
ECA 11	2,619	0.0307052	0.9051766	609	0.06080272	0.11168146
ECA 12	1,136	0.01331848	0.7480483	209	0.02086661	0.06497044
ECA 13	1,518	0.01779706	0.8972032	235	0.02346246	0.06843448
ECA 14	3,076	0.03606308	0.9065744	305	0.03045128	0.09777316
ECA 15	3,405	0.03992028	0.8483815	358	0.03574281	0.08755156
ECA 16	3,128	0.03667272	0.8553078	355	0.03544329	0.08581551
ECA 17	3,150	0.03693065	0.7857312	193	0.01926917	0.08779241
ECA 18	3,122	0.03660238	0.8986634	295	0.02945288	0.08576266
ECA 19	2,355	0.02761006	0.7671079	220	0.02196486	0.09530306
ECA 20	2,548	0.02987279	0.6939706	524	0.05231629	0.04652243
ECA 21	1,905	0.02233425	0.8226458	160	0.01597444	0.07453349
ECA 22	1,983	0.02324873	0.7120910	310	0.03095048	0.08335272
ECA 23	2,122	0.02487836	0.9652836	345	0.03444489	0.08542557
ECA 24	1,759	0.02062255	0.8945340	164	0.0163738	0.08334885

ECA 25	1,419	0.01663638	0.9238189	143	0.01427716	0.07112820
ECA 26	1,376	0.01613225	0.7965494	125	0.01248003	0.07991423
ECA 27	1,463	0.01715224	0.7343431	173	0.01727236	0.09254995
ECA 28	1,841	0.02158391	0.8346904	172	0.01717252	0.07184931
ECA 29	1,133	0.01328331	0.7947807	184	0.01837061	0.07769311
ECA 30	1,119	0.01311917	0.9162765	185	0.01847045	0.07839914
ECA 31	965	0.01131368	0.7865315	131	0.01307907	0.06970279
Total	**85,295**	**1**	**26.04077**	**10,016**	**1**	**2.50260137**
[B] Classes						
0 - 2	79,145	9.278973e-01	0.65712230	10,016	1	0.08013578
2 - 4	5,086	5.962835e-02	2.68665830	NA	NA	NA
4 - 8	982	1.151298e-02	5.12194350	NA	NA	NA
8 - 16	81	9.496453e-04	9.64592370	NA	NA	NA
>16	1	1.172402e-05	16.9259420	NA	NA	NA
Total	**85,295**	**1**	**35.0375898**	**10,016**	**1**	**0.08013578**

3.3.3 Gene annotation in ROHet islands

It was found three genomic regions for ROHet signals on ECA11. The first region with 16 SNPs, starts at AX-103415704 and ends at AX-103146230 SNP (from 26940094 to 27035288); the second with 15 SNPs, starts at AX-104772304 and ends at AX-103144928 (from 28649201 to 28819686); and the third 15 SNPs, starts at AX-103559365 and ends at AX-104150897 (from 33050441 to 33198699).

Annotation for six regions were identified (ENSECAG00000013061, ENSECAG00000004799, ENSECAG00000004853, ENSECAG00000008191, ENSECAG00000009225, and ENSECAG00000009239), but only three of them have a known function (ENSECAG00000013061, ENSECAG00000008191, and ENSECAG00000009239). These regions correspond to the tripartite motif-containing 37 gene (*TRIM37*), protein phosphatase, Mg2 + / Mn2 + dependent 1E (*PPM1E*), and carbonic anhydrase 10 (*CA10*), respectively. The enrichment analysis found six Biological Processes, ten Molecular Functions, and three Cellular Components (Table 2).

Table 2. Functional enrichment analysis for genes identified within ROHet.

Chr	Ensembl	Genes	Molecular Functions	Biological Processes	Cellular Components
11	ENSECAG00000009239	*TRIM37*	zinc ion binding metal ion binding	protein ubiquitination	
11	ENSECAG00000013061	*CA10*	carbonate dehydratase activity zinc ion binding		
11	ENSECAG00000008191	*PPM1E*	catalytic activity	negative regulation of protein kinase activity	nucleus
			phosphoprotein phosphatase activity	protein dephosphorylation	nucleolus
			protein serine/threonine phosphatase activity	cellular response to drug	protein-containing complex
			hydrolase activity	peptidyl-threonine dephosphorylation	

	cation binding	positive regulation of stress fiber assembly
	metal ion binding	

3.4 Discussion

Two previous studies have already described population structure and linkage disequilibrium analyses in MM using the same database. The first, which used a principal components analysis (PCA), found only one population structure, where the top five-set eigenvectors explained 54.98% of the cumulative variance, 40.33% of which belonged to cluster 1 for PCA 1 x PCA 2 (Santos et al., 2020). The second was conducted with imputed population data (Santos et al., 2019), in which some segregations (substructures) were visualized within the breed related to important stallions, but not linked to gait type (marcha batida and marcha picada).

Santos et al. (2019) have also described the current effective population size (Ne) of the MM from the genomic data of 99 animals. The analysis estimated the Ne of 16 generations, demonstrating a marked reduction in recent generations. Furthermore, in both studies described above, low linkage disequilibrium (LD) was found in the MM genome. Thus, Santos et al. (2019) concluded that the formation of the breed may be linked to a broad and partially open genetic base and increased selection pressures. Santos et al. (2019) concluded that the formation of the breed may be linked to a broad and partially open genetic base and increased selection pressures. However, we must consider the countless cross-relatives over the generations (very common in equine breeds), as well as the wide use of a single animal within each breed to promote the improvement of animals over the generations. Studying pedigree representation in a genome-wide association study (GWAS) in MM, Bussiman et al. (2019) reported that familiar structures are very common in horses, probably associated with a mating system based on the intense use of specific animals that achieve more awards in competitions.

When analyzing autozygosity in the population, we found that the FROH performed on the MM population was higher than the FPED estimates. Thanks to previous studies in cattle and supported by computer simulations (Howard et al., 2017; Kardos et al., 2015; Pryce et al., 2014), it is well known that these results correspond to a common event already noted in the literature. However, for the results of the three calculated coefficients, widely discrepant values were found. Such results may owe to limitations inherent to the techniques applied, the incompleteness of pedigree information (four generations), as well as the advantages present when evaluating inbreeding by genomic coefficients. Using computer simulations, Kardos et al. (2015) reported that the proportion of the genome that is identical by descent (IBDG) is more strongly correlated with genomic measures of inbreeding (marker-based, e.g., FROH) than FPED. For the tested scenarios, all genomic measures of inbreeding explained > 90% of the variation using at least 30k of SNPs, while FPED explained < 80% of the variation in IBDG on average, even when the pedigrees included 20 generations.

Regarding our results, the values of FPED and FHOM calculated were practically null, while the mean FROH was considered moderate. These classifications are based on a broad scenario taking into account livestock species; however, Grilz-Seger et al. (2019) state that 16% is a high inbreeding value for the equine genome. Druml et al. (2017) used high-density genotype information of 531 horses originating in seven populations involved in the formation of Haflinger horses. While studying the breeding history, it was found that the mean FROH ranged from 10.1% (Noriker) to 17.7% (Purebred Arabian), with the Shagya Arabian in this study being the breed that came closest to the values found in MM horses, presenting mean FROH values of 15.8%.

According to Cassel et al. (2003), the incompleteness of pedigree information constrains estimations of the real value of inbreeding. In some published studies comparing the correlations between inbreeding coefficients, FPED-FROH values were found to be very similar only when a large number of generations was introduced (Gurgul et al., 2016; Marras et al., 2014). In general, all FPED correlations with the other coefficients were very close to zero. We believe that some more distant generations were extremely inbreeding, and for these reasons, the four generations used in this study were unable to access such information, differing from some works that have already demonstrated proximity between FPED and FROH (Ferenčaković et al., 2011; Ferenčaković et al., 2013b). Additionally, the low correlations between FROH and FHOM necessitate care when studying horse genome inbreeding depression and the genomic proportion of identical by descent (IBD) based only on information from FHOM. Yengo et al. (2017) proved that the consistency of inbreeding depression (ID) estimates obtained with FHOM was also determined by LD differences between SNPs and causal variants. Furthermore, the bias was verified where the FHOM could not simply be predicted by the ratio of the mean LD score in causal variants over the mean LD score in SNPs. Moreover, the possible directional effects of minor alleles confounded FHOM because of the correlation between minor allele counts and FHOM. The authors concluded that directional effects may have arisen as a consequence of directional selection (when the minor allele is also the derived allele) or simply because of population stratification.

When evaluating FROH, we also sought to compare different methodologies (consecutive runs and sliding windows-based runs). It is known that the numerous parameters and fundamentals used in genetic analyses can abruptly influence the results attained. Thus, we contrasted the approach that has been used in the majority of studies (classic) (Purcell et al., 2007) with a more recent approach (modern) (Marras et al., 2015), proposing to calculate with precision and accuracy the coefficients of genomic inbreeding and to correct the computational limitations of previous methods (our analysis eliminated the use of windows when computing the runs). Comparing the assessments of the two approaches, we noticed few differences regarding the ROH islands. However, the sliding windows based-run managed to capture 17,817 more runs than the windowless approach. Thus, we do not discard possible evidence in the underestimation of ROH values, their classes, as well as FROH in the classical approach.

All chromosomes in the MM genome presented ROH. The ROH frequencies across the genome were correlated with local genomic variables such as a recombination rate with a higher probability of accumulation of similar haplotypes, as well as with signals of recent positive selection, resulting in increased homozygosity around the target site (selective sweep) (Pemberton et al., 2012). ECA 1 had the highest proportion with 7,339 ROH (8.60%), whereas the lowest proportion was found in ECA 31, with only 965 ROH (1.13%). This was as expected, because these results are possibly related to the size of the chromosomes. We also found that the different densities of the 670k Axiom ® Equine Genotyping Array (Thermo Fisher, USA) for some regions of the chromosomes did not influence these proportions. The 0–2 Mbps class had the highest proportion, with 79,145 ROH (92.78%). By observing such a proportion, we possibly elucidated the moderate or even high inbreeding found in the

MM genome, this being attributed to a non-recent origin. Furthermore, we corroborated the findings of Grilz-Seger et al. (2018), who when studying the ROH and the population history of three horse breeds found different ROH length classes for Posavje horses, these showing the lowest proportion of ROHs with lengths greater than 6 Mb, thereby also indicating a relative lack of recent inbreeding level within the breed. The other classes of ROH (2–4, 4–8, 8–16, >16 Mbps) corresponded to a percentage of 7.22%, in which the class with >16 Mbps identified only a single ROH in ECA 7. According to Santos et al. (2020), these signatures may represent a previous bottleneck and not a recent positive selection, that is, generalizing this event as a common moment in the horses' evolution process. Grilz-Seger et al. (2018) have also verified in the Bosnian mountain horse the highest genome length covered by ROH (SROH)/FROH values and simultaneously the longest-ranging ROHs >10 Mb, attributed to a possible indication of bottleneck effects due to the Bosnian War in the 1990s, connected with ongoing consanguineous mating in a small population.

Following on from the runs analysis, it was possible to access the ROHet sites through consecutive runs, which were characterized by high rates of recombination. Variant-enriched regions prevent homozygosis due to possible serious negative impacts for a given trait, or even preclude a deleterious event from occurring. Studying a bovine breed, Williams et al. (2015) have demonstrated the importance of local conservation and its associations with the components of global biodiversity, a reservoir of genetic variation relevant to future generations. In their study, a large proportion of the Chillingham individuals examined were heterozygous at many of these polymorphic loci, suggesting that some loci imbalance selection.

Biologically, it is predictable that ROHet hotspots are less frequent when compared to ROH hotspots, and for the MM genome this was no different. According to Hedrick (2012), recent genomic data indicate that many genes show the signal of selection, this being their heterozygote advantage. However, only a small proportion of loci have polymorphisms maintained by heterozygote advantage. Even though some sites of heterozygous advantage have important adaptive functions, their role in general evolutionary change may be more an unusual phenomenon than an important participant of the adaptation, justifying their lower proportion.

In the present study, 10,016 ROHet were identified in the 0–2 Mbps class, whereas other classes did not show results. In ECA 11, a ROHet hotspot with frequency over 0.30 was found, this being a highlighted region with high genetic variability. The ROHet annotations showed genes with known functions: tripartite motif-containing 37 (TRIM37), protein phosphatase, Mg2 + / Mn2 + dependent 1E (PPM1E), and carbonic anhydrase 10 (CA10). ECA 11 has been a prominent chromosome in the equine genome. Avila et al. (2018) have identified regions that are potentially important for athletic racing ability in American Quarter Horse subpopulations, while Velie et al. (2019) exploring the genetics of trotting of three Nordic horse populations have found results identical for the selection signatures of TRIM37 and PPM1E genes. According to Velie et al. (2019), mutations in these regions can affect the underlying mechanisms of muscle, ligament and tendon development, which would certainly influence trotting racing ability, limiting it in some instances while enhancing it in others. Thus, as this is a hypervariable region, we can support evidence that the same may be happening for important traits in the MM breed, whether related to its gait, development or performance as an athlete horse.

The PPM1E gene has already been reported as a negative regulator of the p21-activated protein kinase and the 5-AMP-activated protein kinase (Koh et al., 2002; Voss et al., 2011). Kinases are important regulators of the actin cytoskeleton (Larsson, 2006). Jessen (2010) has hypothesized that the negative regulation of these kinases can cause disorders in the actin cytoskeleton of neurons in the brain, thereby affecting in some way Alzheimer’s disease. The author observed that PPM1E had a degenerative effect on the number of dendritic mushroom spines and the dendritic arbor, indicating that phosphatase may play a role in the development of Alzheimer’s disease. Regarding our findings, it was hypothesized that these effects on horses would be different, possibly related to the central nervous system, but associated with impacts on the health and performance of these animals, potentially ranging from metabolic factors to behavioral disorders.

Finally, the CA10 gene encodes a protein that belongs to the carbonic anhydrase family of zinc metalloenzymes, catalyzing the reversible hydration of carbon dioxide in various biological processes. Furthermore, in the literature, multiple transcript variants encoding the same protein have been found for this gene (NCBI, 2019). In horses, carbonic anhydrase (CA) is therefore of interest because it catalyzes the reaction CO2 + H2O<-->HCO3- + H+. CA is consequently important in counteracting the alkalosis developed after exercise by delivering HCO3- for the generation of the alkaline pH in sweat (Dahlborn et al., 1999).

3.5 Conclusion

Supported by FROH results and homozygosity runs that showed a high proportion of short ROH, our findings suggest a moderate inbreeding in the MM genome, which is attributed to some more distant generations and not the recent inbreeding. We also found in order to obtain accurate results in population inbreeding studies in the MM breed, that is necessary to access the complete pedigree to avoid underestimating the actual values of inbreeding. In addition, regions with high variability in the MM genome were identified (ROHet), where genes in these regions are possibly associated with recent selection, acting in important events for the development and performance of the MM horse over generations.

3.6 Acknowledgements

This work was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo – Fapesp (Grant 2016/19081-9). Wellington Bizarria dos Santos was the recipient of a Master's fellowship from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Capes (Finance code 001).

3.7 References

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.

Abri MAAl, Borstel UKV, VS, Brooks SA (2017) Application of Genomic Estimation Methods of Inbreeding and Population Structure in an Arabian Horse Herd. **Journal of Heredity** 108:361-368.

Aguilar I, Misztal I (2008) Technical note: recursive algorithm for inbreeding coefficients assuming nonzero inbreeding of unknown parents. **J Dairy Sci** 91:1669–1672.

Avila F, Mickelson JR, Schaefer RJ, ME (2018) Genome-Wide Signatures of Selection Reveal Genes Associated With Performance in American Quarter Horse Subpopulations. **Frontiers in Genetics** 9:249.

Beeson SK, Schaefer RJ, Mason VC, McCue ME (2019) Robust remapping of equine SNP array coordinates to EquCab3.0. **Animal Genetics** 50:114–5.

Biscarini F, Cozzi P, Gaspa G, Marras G (2019) detectRUNS: Detect Runs of Homozygosity and Runs of Heterozygosity in Diploid Genomes. R package version 0.9.6. https://CRAN.R-project.org/package=detectRUNS.

Bull JJ (2017) Lethal gene drive selects inbreeding. **Evol Med Public Health** 1: 1–16.

Bussiman FO, Santos BA, Silva BCA, et al. (2019) Allelic and genotypic frequencies of the DMRT3 gene in the Brazilian horse breed Mangalarga Marchador and their association with types of gait. **Genetics and Molecular Research** 18:gmr18217.

Cassell BG, Adamec V, Pearson RE (2003) Effect of incomplete pedigrees on estimates of inbreeding and inbreeding depression for days to first service and summit milk yield in Holsteins and Jerseys. **Journal of Dairy Science** 86:2967–2976.

Ceballos FC, Joshi PK, Clark DW, Ramsay M, Wilson JF (2018) Runs of homozygosity: windows into population history and trait architecture. **Nature Reviews Genetics** 19:220–234.

Clayton DG (2009) Sex chromosomes and genetic association studies. **Genome Med.** 1:110.

Dahlborn K, Jansson A, Nyman S, Morgan K, Holm L, Ridderstråle Y (1999) Sweat production and localisation of carbonic anhydrase in the equine sweat gland during exercise at two ambient temperatures. **Equine Vet J** 30:398-403.

Druml T, Neuditschko M, Grilz-Seger G, Horna G, Ricard A, Mesarič M, Cotman M, Pausch H, Brem G (2018) Population Networks Associated with Runs of Homozygosity Reveal New Insights into the Breeding History of the Haflinger Horse. **Journal of Heredity** 109:384–392.

Fawcett JA, Sato F, Sakamoto T, Iwasaki WM, Tozaki T, Innan H (2019) Genome-wide SNP analysis of Japanese Thoroughbred racehorses. **PLoS ONE** 14: e021840.

Ferenčaković M, Hamzic E, Gredler B, Curik I, Sölkner J (2011) Runs of homozygosity reveal genome-wide autozygosity in the Austrian Fleckvieh cattle. **Agriculturae Conspectus Scientificus** 76:325–8.

Ferenčaković M, Hamzic E, Gredler B, Solberg TR, Klemetsdal G, Curik I, Sölkner J (2013a) Estimates of autozygosity derived from runs of homozygosity: empirical evidence from selected cattle populations. **Journal of Animal Breeding and Genetics** 130:286–93.

Ferenčaković M, Sölkner J, Curik I (2013b) Estimating autozygosity from high-throughput information: effects of SNP density and genotyping errors. **Genetics Selection Evolution** 45:42.

Fonseca MG, Ferraz GC, Lage J, Pereira GL, et al. (2017). A Genome-Wide Association Study Reveals Differences in the Genetic Mechanism of Control of the Two Gait Patterns of the Brazilian Mangalarga Marchador Breed. **J. Equine Vet. Sci.** 53: 64–67.

Forutan M, Ansari MS, Baes C, et al. (2018) Inbreeding and runs of homozygosity before and after genomic selection in North American Holstein cattle. **BMC Genomics** 98:2-12.

Gomez-Raya L, Rodríguez C, Barragán C, Silió L (2015) Genomic inbreeding coefficients based on the distribution of the length of runs of homozygosity in a closed line of Iberian pigs. ***Genet Sel Evol*.** 47:81.

Grilz-Seger G, Druml T, Neuditschko M, Dobretsberger M, Horna M, Brem G (2019) High-resolution population structure and runs of homozygosity reveal the genetic architecture of complex traits in the Lipizzan horse. **BMC Genomics** 20:174.

Grilz-Seger G, Mesaric M, Cotman M, Neuditschko M, Druml T, Brem G (2018) Runs of Homozygosity and Population History of Three Horse Breeds With Small Population Size. **Journal of Equine Veterinary Science** 71:27e34.

Gurgul A, Szmatoła T, Topolski P, Jasielczuk I, Żukowski K, BUGNO-PONIEWIERSKA M (2016) The use of runs of homozygosity for estimation of recent inbreeding in Holstein cattle. **Journal of Applied Genetics** 57:527-530.

Hedrick PW (2012) What is the evidence for heterozygote advantage selection?. **Trends in Ecology & Evolution** 27:P698-704.

Hedrick PW, Kalinowski ST (2000) Inbreeding depression and conservation biology. **Ann. Rev. Ecol. Syst**. 31:139–162.

Howard JT, Pryce JE, Baes C, Maltecca C (2017) Inbreeding in the genomics era: inbreaeding, inbreeding depression, and management of genomic variability. **J Dairy Sci.** 100:6009–24.

Jessen AL (2010) **Localization and truncation in brain tissue and effects on neuronal morphology in primary neuronal culture**. Dissertation (Doctor rerum naturalium) – Georg August University Göttingen, Göttingen.

Kaeuffer R, Coltman DW, Chapuis JL, Pontier D, Réale D (2007) Unexpected heterozygosity in an island mouflon population founded by a single pair of individuals. ***Proc Biol Sci*.** 274:527-33.

Kardos M, Luikart G, Allendorf FW (2015) Measuring individual inbreeding in the age of genomics: marker-based measures are better than pedigrees. **Heredity (Edinb).** 115:63–72.

Kasarda R, Moravčíková N, Kadlečík O, Trakovická A, Halo M, Candrák J (2019) Level of inbreeding in Norik of Muran horse: pedigree vs. genomic data. **ACTA UNIVERSITATIS AGRICULTURAE ET SILVICULTURAE MENDELIANAE BRUNENSIS** 67: 1457- 1463.

Keller MC, Visscher PM, Goddard ME (2011) Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. **Genetics** 189:237-49.

Kim E, Sonstegard TS, Tassell CPV, Wiggans G, Rothschild MF (2015) The Relationship between Runs of Homozygosity and Inbreeding in Jersey Cattle under Selection. **PLoS ONE** 10: e0129967.

Koh C, Tan E, Manser E, Lim L (2002) The p21-activated kinase PAK is negatively regulated by POPX1 and POPX2, a pair of serine/threonine phosphatases of the PP2C family. **Current Biology** 12:317–321.

Larsson C (2006) Protein kinase C and the regulation of the actin cytoskeleton. **Cellular Signalling** 18:276–284.

Lencz T, Lambert C, DeRosse P, Burdick KE, Morgan TV, Kane JM, Kucherlapati R, Malhotra AK (2007) Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. **Proc Natl Acad Sci** 104:19942e7.

Li M, Strandén I, Tiirikka T, Sevón-Aimonen M, Kantanen JA (2011) Comparison of Approaches to Estimate the Inbreeding Coefficient and Pairwise Relatedness Using Genomic and Pedigree Data in a Sheep Population. **PLoS ONE** 6: e26256.

Marras G, Gaspa G, Sorbolini S, Dimauro C, Ajmone-Marsan P, Valentini A, Williams JL, Macciotta NP (2014) Analysis of runs of homozygosity and their relationship with inbreeding in five cattle breeds farmed in Italy. **Animal Genetics** 46:110–21.

Marras G, Wood B, Makanjuola B, Malchiodi F,Peeters K, As P, Baes C, Biscarini F (2018) Characterization of runs of homozygosity and heterozygosity-rich regions in a commercial turkey (Meleagris gallopavo) population. **Conference: 11th World Congress on Genetics Applied to Livestock Production**, New Zealand: WCGALP, p. 1-5.

Metzger J, Karwath M, Tonda R, Beltran S, Águeda L, Gut M, Gut IG, Distl O (2015) Runs of homozygosity reveal signatures of positive selection for reproduction traits in breed and non-breed horses. **BMC Genomics** 16:764.

National Center for Biotechnology Information – NCBI (2019). Available at: https://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=ShowDetailView&TermToSearch=56934. Accessed February 14th of 2020.

Nowak C, Jost D, Vogt C, Oetken M, Schwenk K, OehlmannJ (2007) Consequences of inbreeding and reduced genetic variation on tolerance to cadmium stress in the midge Chironomus ripariu. **Aquatic Toxicology** 84:278-284.

Patterson L, Staiger EA, Brooks SA (2015) DMRT3 is associated with gait type in Mangalarga Marchador horses , but does not control gait ability. **Animal Genetics** 46: 213–215.

Pekkala N, Knott KE, Kotiaho JS, Nissinen K, Puurtinen M (2014) The effect of inbreeding rate on fitness, inbreeding depression and heterosis over a range of inbreeding coefficients. **Evol Appl.** 7:1107-19.

Pemberton TJ, Absher D,Feldman MW, Myers RM, Rosenberg NA, Li JZ (2012) Genomic Patterns of Homozygosity in Worldwide Human Populations. **The American Journal of Human Genetics** 91:275–292.

Peripolli E, Stafuzza NB, Munari DP, Lima ALF, Irgang R, Machado MA, Panetto JCDC, Ventura RV, Baldi F, da Silva MVGB (2018) Assessment of runs of homozygosity islands and estimates of genomic inbreeding in Gyr (Bos indicus) dairy cattle. **BMC Genomics** 19:34.

Pryce JE, Haile-Mariam M, Goddard ME, Hayes BJ (2014) Identification of genomic regions associated with inbreeding depression in Holstein and Jersey dairy cattle. **Genet Sel Evol.** 46:71.

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-based linkage analyses. **American journal of human genetics** 81:559–575.

Purfield DC, McParland S, Wall E, Berry DP (2017) The distribution of runs of homozygosity and selection signatures in six commercial meat sheep breeds. **PLoS ONE** 12: e0176780.

Renaud G, Hanghøj K, Korneliussen TS, Willerslev E, Orlando L (2019) Joint Estimates of Heterozygosity and Runs of Homozygosity for Modern and Ancient Samples. **Genetics** 212:587-614.

Robert A, Toupance B, Tremblay M, Heyer E (2009) Impact of inbreeding on fertility in a pre-industrial population. **European Journal of Human Genetics** 17:673–681.

Santos BA, Pereirab GL, Bussiman FO (2019) Genomic analysis of the population structure in horses of the Brazilian T Mangalarga Marchador breed. **Livestock Science** 229: 49-55.

Sayres MAW (2018) Genetic Diversity on the Sex Chromosomes. **Genome Biol Evol. 10:1064-1078.**

Smedley D, Haider S, Ballester B, et al. (2009) BioMart – biological queries made easy. Trevor J, Pembertona, Noah A (2015) Rosenbergb Population-genetic influences on genomic estimates of the inbreeding coefficient: a global perspective. **Hum Hered.** 77: 37–48.

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

Velie BD, Solé M, Fegraeus KJ, Rosengren MK, Røed KH (2019) Genomic measures of inbreeding in the Norwegian–Swedish Coldblooded Trotter and their associations with known QTL for reproduction and health traits. **Genet Sel Evol** 51:22.

Voss M, Paterson J, Kelsall et al. (2011) Ppm1E is an in cellulo AMP-activated protein kinase phosphatase. **Cellular Signalling** 23:114–124.

Williams JL, Hall SJG, Del Corvo M, Ballingall KT, Colli L, Ajmone Marsan P & Biscarini F (2016) Inbreeding and purging at the genomic Level: the Chillingham cattle reveal extensive, non-random SNP heterozygosity. **Animal genetics** 47:19-27. 10.1111/age.12376

Xu L, Zhao G, Yang L (2019) Genomic patterns of Homozygosity in chinese Local cattle. **Scientific Reports** 9:16977.

Yengo L, Zhu Z, Wray NR, WeirBS, Yang J, Robinson MR, Visscher PM (2017) Detection and quantification of inbreeding depression for complex traits from SNP data. **Proc Natl Acad Sci** 114:8602–8607.