Exploring the genetic cause of myotonic dystrophy in horses

Thomas Simon

Independent project • 30 credits
Swedish University of Agricultural Sciences, SLU
Department of Animal Breeding and Genetics
Master's programme in Bioinformatics
Uppsala 2023

Exploring the genetic cause of myotonic dystrophy in horses

Thomas Simon

Supervisor:	**Gabriella Lindgren, Swedish University of Agricultural Sciences, Department of Animal breeding and genetics**
Assistant supervisor:	Erik Bongcam-Rudloff, Swedish University of Agricultural Sciences, Department of Animal breeding and genetics
Assistant supervisor:	Rakan Naboulsi, Swedish University of Agricultural Sciences, Department of Animal breeding and genetics
Examiner:	Martin Johnsson, Swedish University of Agricultural Sciences, Department of Animal breeding and genetics

Credits:	30
Level:	A2E
Course title:	**I**ndependent project in Bioinformatics
Course code:	EX1002
Programme/education:	Master in Bioinformatics
Course coordinating dept:	Department of animal breeding and genetics
Place of publication:	SLU, Uppsala, Sweden
Year of publication:	2023

Keywords: genetics, bioinformatics, myotonic dystrophy, horses, variant analysis

Swedish University of Agricultural Sciences
Faculty of veterinary medicine
Department of Animal Breeding and Genetics

Abstract

Myotonic Dystrophy (MD) is an inherited neuromuscular disorder that leads to a wide range of clinical signs including stiffness, abnormal muscle relaxation, and muscle atrophy or hypertrophy. Affected muscles develop a prolonged contracture in response to percussion of the muscle and show waxing and waning complex repetitive discharges in electromyography (EMG). Two genetic mutations underlie the genetic basis for MD in humans that cause abnormal RNA processing. There are a limited number of case reports of MD in horses in the veterinary literature and its genetic cause is unknown. By analysing whole genome sequencing (WGS), RNAseq, and proteomics data from horses in our MD dataset, we aim to identify the molecular mechanisms contributing to the condition. Understanding the molecular mechanisms behind MD may lead to non-invasive diagnostic tests and potentially therapeutic strategies.

MD horses and seven age, sex and breed-matched controls were low pass Whole genome sequenced to approximately 4X coverage. Fst values were calculated and 6 chromosome regions on ECA 1, 3, 6, 7, 16 and 18 appeared higher than our threshold of 0.0005% and conserved across multiple window sizes. On chromosome 3 was found a candidate gene responsible for the Facioscapulohumeral Muscular Dystrophy 1 in human. SNPs were investigated and one splice variant has been identified. Sequencing depth will be increased on the same horses to confirm the preliminary findings. The obtained results of this project may also provide insights into the related human MD disease, which affects an average of 1 out of 8000 individuals.

As a complementary study within the main investigation, an allele frequency analysis was conducted on a specific variant located on ECA 3. Whole-genome sequencing data of horses were retrieved from NCBI, and alignment of the exon harbouring the variant was performed followed by genotyping of each horse. A distinction was observed between Endomorph (muscular) and Ectomorph (lean) horses, with different alleles linked to the horses' morphological characteristics. However, further analysis is warranted due to the limited dataset and time constraints involved in this study.

Table of Contents

List of Figures

List of Tables

Abbreviations

A	Adenine
BRCA1	BReast CAncer gene 1
C	Cytosine
DMPK	Dystrophia myotonica protein-kinase
DNA	Deoxyribonucleic acid
EMG	Electromyography
G	Guanine
GATK	Genome analysis toolkit
MD	Myotonic Dystrohpy
MIT	Massachusetts Institute of Technology
PCR	Polymerase Chain Reaction
QH	Quarter Horse
RNA	Ribonucleic acid
SLU	Sveriges lantbruksuniversitet
SNP	Single Nucleotide Polymorphism
T	Tyrosine
THEGL	Testicular Haploid Expressed Gene Protein-Like
US	United States
WGS	Whole Genome Sequencing

Introduction

Myotonic dystrophy (MD) in human is an autosomal dominant (Bird, 1993) inherited neuromuscular disorder characterized by a wide range of clinical signs and symptoms, including muscle stiffness, abnormal muscle relaxation, and muscle atrophy or hypertrophy . It is one of the most common adult-onset muscular dystrophies, with a prevalence of approximately 1 in 8,000 individuals worldwide (Spiro, 2002).

The pathophysiology of myotonic dystrophy in humans involves an abnormal expansion of CTG trinucleotide repeats in the dystrophia myotonica protein-kinase (DMPK) gene on chromosome 19 (Mahadevan et al., 1992). This repeat expansion results in the sequestration of RNA-binding proteins, such as Muscleblind-like 1 (MBNL1), leading to disruption of RNA processing and subsequent dysregulation of alternative splicing in affected tissues (Day and Ranum, 2005). The aberrant RNA splicing is thought to contribute to the development of various clinical features observed in myotonic dystrophy.

Clinical manifestations of myotonic dystrophy vary widely and can affect multiple organ systems, including skeletal muscle, heart, central nervous system, and endocrine glands. Muscular involvement is prominent, characterized by prolonged muscle contractures following percussion and a phenomenon known as "myotonia." Myotonia refers to the delayed relaxation of the muscle after contraction, leading to a characteristic stiffness or difficulty in initiating movement. Electromyography (EMG) typically reveals waxing and waning complex repetitive discharges, which further support the diagnosis (Spiro, 2002).

Diagnosis of myotonic dystrophy is primarily based on clinical evaluation, family history, genetic testing, and electromyographic findings (Spiro, 2002). Early detection of myotonic dystrophy is crucial as there is currently no known cure for the disease. Early diagnosis offers several benefits, including the opportunity for genetic counselling and family planning. Understanding the genetic basis of the disease allows individuals to make informed decisions about family planning and assess the risk of passing the condition to future generations (Mahadevan et al., 1992).

Additionally, early diagnosis enables healthcare professionals to proactively manage the disease by closely monitoring individuals and implementing appropriate medical interventions. Regular medical evaluations and proactive management of symptoms and complications can improve the quality of life for individuals with myotonic dystrophy (Hilbert et al., 2017).

Prompt identification of myotonic dystrophy also facilitates timely access to support and perform interventions. Individuals can benefit from specialized care provided by a multidisciplinary team, as well as physical and occupational therapy. Assistive devices and resources for managing specific symptoms, such as cardiac or respiratory complications, can be implemented to enhance overall well-being (Heatwole et al., 2015).

Furthermore, early diagnosis plays a crucial role in the participation of individuals in clinical trials and research studies. By being diagnosed early, individuals have the opportunity to

contribute to the advancement of knowledge and the development of potential treatments for myotonic dystrophy (Modoni et al., 2004).

Early diagnosis of myotonic dystrophy through molecular genetic testing of the DMPK gene (Mahadevan et al., 1992) enables informed decision-making, proactive medical management, support access, and participation in research, positively impacting affected individuals and their families. Prenatal testing and genetic counselling are important considerations for those with a family history. Supportive management involving a multidisciplinary team and regular surveillance for specific complications (cardiac, respiratory, cognitive, and endocrine) are crucial for optimizing patient outcomes (Spiro, 2002).

Understanding the underlying pathophysiology, clinical features, and diagnostic approaches is vital for accurate diagnosis and appropriate management of individuals affected by myotonic dystrophy.

This project will focus on the myotonic dystrophy in horses. Knowledge of the disease in horse is very limited with no scientific articles published. The diagnosis of myotonic dystrophy in horses involves a comprehensive approach that includes clinical evaluation, electromyography (EMG), and, in some cases, muscle biopsies. The diagnostic process aims to accurately identify the presence of myotonic dystrophy based on clinical signs and muscle function.

During the clinical evaluation, a veterinarian conducts a thorough physical examination of the horse, carefully assessing specific clinical signs associated with myotonic dystrophy. These signs may include muscle stiffness, delayed muscle relaxation (myotonia), abnormal muscle size (hypertrophy or atrophy), abnormal gait, and exercise intolerance.

Electromyography (EMG) is a diagnostic procedure used to evaluate muscle activity and detect any abnormalities in muscle function. In horses suspected of having myotonic dystrophy, EMG can reveal distinctive repetitive discharges or electrical activity patterns in affected muscles. This helps in establishing a preliminary diagnosis.

In some cases, a muscle biopsy may be performed to further support the diagnosis of myotonic dystrophy. A small sample of muscle tissue is extracted, typically from the gluteal or semimembranosus muscles, and examined under a microscope. Muscle biopsies allow for the assessment of characteristic histopathological changes associated with myotonic dystrophy, including fiber size variation, fibrosis, and accumulation of abnormal material within muscle fibers.

Combining the findings from clinical evaluation, EMG, and muscle biopsy (if performed) helps in establishing a definitive diagnosis of myotonic dystrophy in horses. Genetic testing is currently unavailable for horses due to the unidentified cause of the condition. By identifying the genetic cause of the condition, this research aims to fill this gap and pave the way for future advancements in genetic testing for horses. Understanding the genetic cause of myotonic dystrophy in horses is crucial for breeders. It allows them to selectively breed horses and reduce the occurrence of the disease. By avoiding mating pairs that carry the genetic mutation responsible for myotonic dystrophy, breeders can lower the prevalence of the

condition within the population. This, in turn, contributes to the overall health and well-being of the breed.

Additionally, eliminating myotonic dystrophy from the breeding population has broader benefits. It improves the overall quality and reputation of the affected breed by producing healthy horses. These horses are likely to exhibit better characteristics, performance, and market value, making them more desirable to potential buyers.

Furthermore, uncovering the genetic cause of myotonic dystrophy contributes to the economic considerations of horse breeding. The disease imposes financial burdens on horse owners due to the costs associated with diagnosis, treatment, and long-term management. By reducing the prevalence of the disease through informed breeding practices, breeders can contribute to the economic sustainability of the breed and minimize expenses related to the condition.

In conclusion, understanding the genetic cause of myotonic dystrophy in horses empowers breeders to make informed decisions, prevent the transmission of the disease, improve animal welfare, enhance breed quality, and alleviate economic burdens. It is a significant step towards producing healthier horses and ensuring the long-term viability and success of the breed.

In our case, 12 horses have been diagnosed positive to the myotonic dystrophy with help of either clinical evaluation, EMG, muscle biopsy or both. They have been Whole Genome Sequenced and the data were sent to SLU from the US Michigan State University. These samples were all taken from Quarter Horses.

The Quarter Horse is a versatile and popular breed known for its athleticism, strength, and docile temperament. Originating in the United States about 1660s, it has become one of the most prominent and widely recognized horse breeds worldwide. This breed's unique characteristics make it suitable for a wide range of activities, including racing, ranch work, rodeo events, and recreational riding.

Quarter Horses are generally muscular and compact, with a well-developed chest, strong hindquarters, and a broad, expressive head. They have a short, fine coat and can come in various solid colours or patterns such as roan, palomino, and buckskin. The average height of Quarter Horses ranges from 14.3 to 16 hands (145 to 163 inches) at the withers, and they typically weigh between 950 to 1,200 pounds (“American Quarter Horse | breed of horse | Britannica,” n.d.).

Known for their willing and cooperative nature, Quarter Horses are often described as intelligent and easy to train. They possess a calm and sensible disposition, making them well-suited for riders of all levels, including novice and youth riders. This breed's versatility and trainability have made it successful in various disciplines, including Western pleasure, reining, cutting, and barrel racing.

Quarter Horses excel in short-distance sprinting, making them highly competitive in racing events. They are capable of reaching impressive speeds, with some individuals clocking

speeds of up to 55 miles per hour in quarter-mile races. Their agility, quick turns, and powerful bursts of speed make them ideal for working with livestock, as they can swiftly change direction and respond to the movements of cattle.

To study the horses previously mentioned, the use of the current horse reference genome EquCab 3.0 is needed. The horse reference genome EquCab 3.0 is a valuable genomic resource that provides a comprehensive blueprint of the horse genome. It serves as a crucial reference point for studying the genetic composition, variation, and functional elements of the horse genome. In particular, the reference will enable us to systematically examine potential variations on a chromosome-by-chromosome basis for example. EquCab 3.0 was generated through a collaborative effort involving researchers from the Equine Genome Project and the Broad Institute of MIT and Harvard (Kalbfleisch et al., 2018).

EquCab 3.0 incorporates advancements in sequencing technologies and assembly methods, resulting in improved accuracy, contiguity, and annotation of the horse genome. It consists of 31 autosomes, the X chromosome, the mitochondrial genome, and unplaced scaffolds that are not annotated, providing a detailed representation of the horse's genetic information. The reference genome enables researchers to map and analyse genetic variations, study gene expression patterns, identify disease-associated genes, and explore the evolutionary history of horses (Wade et al., 2009).

The aim of this study is to elucidate the underlying genetic basis of myotonic dystrophy in horses, with the ultimate goal of developing effective strategies for disease prevention and eradication. Specifically, we aim to identify the specific genetic mutations or variants that contribute to the development and progression of myotonic dystrophy in horses. By understanding the genetic mechanisms involved, we can potentially implement selective breeding programs to identify and breed animals that are genetically resistant to the disease.

Materials and Methods

Sample collection

The 12 horses were samples during the last 30 years by Dr. Stephanie Valberg.
Muscle samples had previously been obtained from five MD horses of Quarter Horse (3 female, 2 male) 2 to 3 years-of-age. Signs of stiffness, muscle contractures, muscle hypertrophy were present shortly after birth. EMG performed in 4 horses revealed complex repetitive discharge within hindlimb muscles and few abnormal discharges in forelimb muscles. Histopathologic analysis identified fibre size variation, internalized myonuclei, a preponderance of slow twitch type 1 fibers and fiber type grouping in gluteal and semimembranosus (SM) muscles of the MD foals but few abnormalities in triceps muscles consistent with EMG findings.

Sequencing

Libraries were prepared using the Illumina ("DNA Sequencing | Understanding the genetic code," n.d.) TruSeq Nano DNA Library Preparation Kit with IDT for Illumina Unique Dual Index adapters following manufacturer's recommendations. Libraries were QC'd and quantified using a combination of Qubit dsDNA HS and Agilent 4200 TapeStation HS DNA1000. Libraries were divided into two equal pools, and libraries within each were combined in equimolar amounts for multiplexed sequencing. The pools were quantified using Kapa Biosystems Illumina Library Quantification qPCR assay. Each was loaded on one lane of an Illumina HiSeq 4000 flow cell and sequencing was performed in a 2x150bp paired end format using HiSeq 4000 SBS reagents. Base calling was done by Illumina Real Time Analysis (RTA) v2.7.7 and output of RTA was demultiplexed and converted to FastQ format with Illumina Bcl2fastq v2.19.1. The sequencing was carried out by the Plant Biology Laboratories of Michigan State University.

Quality control

The quality of the raw data was assessed using FastQC v0.11.8 ("Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data," n.d.), which provided an initial evaluation of the reads' quality. MultiQC v1.14 (Ewels et al., 2016) was then employed. This tool takes the previously created reports from FastQC as input and allow us to gain a comprehensible and graphical view of the quality of all the samples. The files were then processed using fastp v0.19.5 (Chen et al., 2018) using the default parameters, which removed duplicate reads as well as any remaining adapter sequences. Subsequently, FastQC and MultiQC were applied again to verify the quality after cleaning.

Mapping to the reference genome

Bowtie2 v2.5.1 (Langmead et al., 2019), a robust tool for read-to-reference genome mapping, was utilized to map the reads to the EquCab3.0 reference horse genome (Wade et al., 2009). This mapping process involved all 26 files and required approximately 36 hours to complete. As a result, 12 bam files were prepared for SNP calling.

SNP calling

The initial preparation step involved creating a dictionary of the reference genome using the CreateSequenceDictionary tool from the picard toolbox v1.137 ("Picard Tools - By Broad Institute," n.d.). Subsequently, an index of the reference genome was constructed using samtools faidx. With these preparations complete, the reference genome was now ready for further analysis.

To enable the usage of HaplotypeCaller, a specific bam file needed to be sorted. This sorting process was accomplished using picard SortSam v1.137, followed by the identification and marking of duplicates with MarkDuplicates v1.137, which identified duplicates from the previous bowtie mapping step. Additional preparation steps included modifying the headers and creating an index. The bamaddrg tool from the bamtools suite v2.4.0 was employed to modify the headers of each mapped read, followed by the samtools index command to create an index file for the modified bam. With these steps concluded, the files were now prepared for analysis by HaplotypeCaller from the GATK pipeline v3.4.0.0 ("GATK," n.d.).

The SNP calling procedure employed HaplotypeCaller from the GATK pipeline with default options, as detailed in the HaplotypeCaller manual page and supplementary material (Appendix 2). This step took approximately 28 hours per file which makes a total of 14 days.

It has been decided to separate insertion/deletion from SNPs. To do so, the tool SelectVariant from GATK has been used with -select-type SNP option. Afterwards, it is recommended by the GATK team to do hard filtering step which is described in the HaplotypeCaller manual. No additional filtering measures, such as removing low coverage SNPs, were implemented since the data already had low coverage. The whole pipeline of the analysis is available in Appendix 2.

Fst

Calculation and plotting

For calculating Fst values, the program vcftools v0.1.13 ("VCFtools," n.d.) was employed. Given that whole-genome sequencing (WGS) data was used, the program's options allowed for selecting a specific window size to include multiple SNPs and compute a weighted mean Fst. Window sizes of 40kb, 50kb, 100kb, and 150kb were chosen for the Fst calculation and subsequent plotting.

The plotting step was conducted using the R programming language v4.2.1 ("R: The R Project for Statistical Computing," n.d.), with libraries such as tidyverse, to read the specific files, and qqman utilized to generate visual representations of the Fst files (Appendix 3).

Pathway analysis/Gene identification

After employing Fst analysis, a series of conserved genomic regions, referred to as windows, were identified. Subsequently, the genes residing within these windows were annotated using the Ensembl database by retrieving each gene name and function. To identify potential genes expressed in muscle that may contribute to muscular diseases, each annotated gene was manually searched on the GeneCards database (Fishilevich et al., 2017; Safran et al., 2010; Stelzer et al., 2016).

The genes of interest, chosen for further analysis, was then explored using GeneHancers within the GeneCards database. This investigation led to the identification of three genes that could exhibited interactions with THEGL as enhancers and promoters.

To broaden our understanding and gain a comprehensive perspective, we utilised the Pathway Commons database (Rodchenkov et al., 2020) to examine the interaction of the previously identified gene of interest with BRCA1. Moreover, we expanded our investigation by employing the String database (Szklarczyk et al., 2021; von Mering et al., 2003).

SNPs of interest

The subsequent step involved retrieving all single nucleotide polymorphisms (SNPs) within these genes and annotating them by specifying the mutation consequence to understand their potential impact on protein function/structure or splicing. To accomplish this, we utilized the Ensembl database (Martin et al., 2023), obtaining variant tables for each gene.

Using a basic bash script, the relevant column (Consequence type) from the variant table was merged with the SNP file generated from the GATK pipeline. Essentially, if a position identified in the SNP file matched a position in the Ensembl variant table, the corresponding consequence type was appended to a new column.

Sanger Sequencing

Due to the limited data coverage, we conducted Sanger sequencing.

The submission of a DNA template for sequence analysis involves several essential steps, including identification, isolation, purification, and quantification of the target DNA template. In this study, the targeted DNA was an amplicon located in chromosome 3, specifically containing the THEGL gene. To obtain the desired product, a primer pair was designed using the Primer3 website and was based on the sequence of interest. The M13 universal primer sequence was incorporated into the primer pair as per the instructions provided in the Big Dye sequencing manual.

To ensure optimal results during sequencing, it was important for the target DNA to exhibit a single product. This was assessed by performing gel electrophoresis procedures to visualize the presence of a specific band. Once the desired band was confirmed, purification of the DNA was carried out to isolate the product from other contaminants. Subsequently, the isolated DNA underwent quantification to determine its concentration and ensure it was within the appropriate range.

Following the purification and quantification steps, the original DNA template was subjected to PCR amplification. For the Sanger sequencing process, the BigDye sequencing method was employed, with slight modifications. During the first PCR, the annealing temperature was set at 60°C to match the requirements of the primers used. Standard procedures were followed during the second PCR, as outlined in the provided table (Table 1).

Overall, these steps were essential in preparing the DNA template for Sanger sequencing, ensuring the presence of a specific target product, purifying the DNA, quantifying its concentration, and performing the necessary PCR amplification steps for subsequent sequencing analysis.

PCR			**Sequencing PCR**		
Temp (C)	**Time**		**Temp (C)**	**Time**	
96	5 min	35 Cycles	37	15 min	25 Cycles
94	30 sec		80	2 min	
60	45 sec		96	1 min	
68	45 sec		96	10 sec	
72	2 min		50	5 sec	
4	indef		60	4 min	
			4	indef	

Table 1: PCR times and temperatures cycles.

To facilitate this process, primers were designed using the Primer3 website (Chuang et al., 2013) by entering the sequence of interest. Subsequently, two primers (one forward and one reverse) were prepared for use after undergoing a final verification using the in-silico PCR tool provided by the University of California, Santa Cruz (Appendix 1). Following this verification, all necessary materials were sent to the laboratory.

The resulting DNA files were analysed and aligned to the horse reference genome using CodonCodeAligner v10.0.2, which aligns with the mapping step described earlier in the thesis.

Allele frequency

In this particular phase of the study, we performed a BLAST search of exon 6 from the THEGL gene against various whole-genome sequences (WGS) of different horse breeds, which were freely accessible on NCBI. Subsequently, we calculated the allele frequency at a specific position, namely the last nucleotide of exon 6, by genotyping the horses. The dataset comprised nine breeds, consisting of four Ectomorph breeds and five Endomorph breeds. The selection of these breeds was based on data availability in NCBI, aiming to include extreme examples of both muscular and lean phenotypes, as well as breeds with varying degrees of classification difficulty, such as the Quarter Horse, where muscularity is influenced by specific breeding purposes (racing, show, etc.). The next table will show the detailed dataset:

	Breed	Number of horses
Ectomorph	Arabian	9
	Thoroughbred	10
	Akhal Teké	4
	Standardbred	5
Endomorph	Shetland	20
	Quarter horse	9
	Criollo	1
	Icelandic	2
	Appaloosa	2

Table 2: Dataset used for the Allele frequency study.

By looking at the alignment output of NCBI blast, we could count the number of times the nucleotide was found in the reads and thus assume the genotyping. Here is an example:

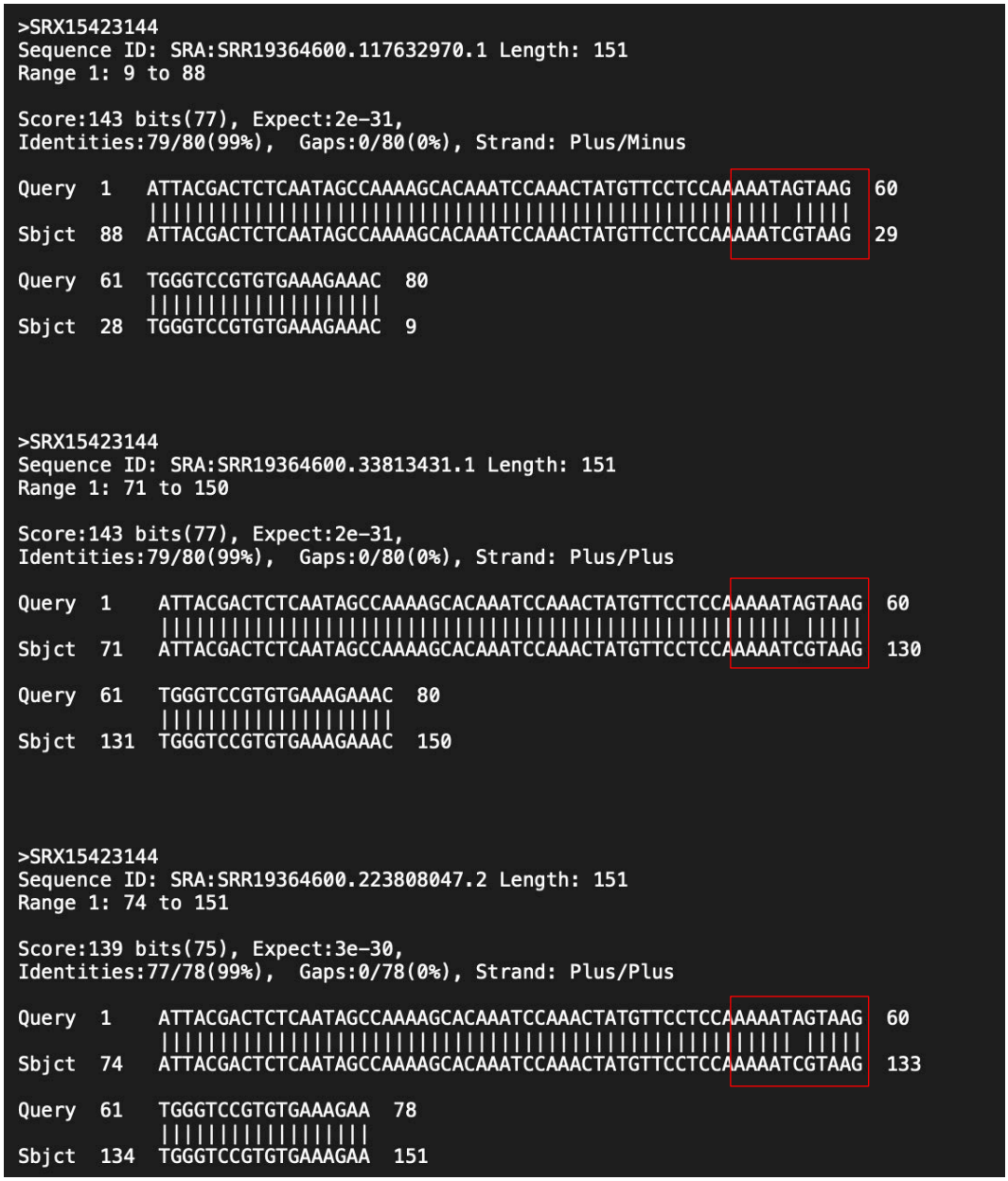


Figure 1: Alignment output from NCBI blast. The query is the Whole genome sequenced horse retrieved from NCBI and the subject is the exon 6 of THEGL. In the red boxes are the differences at the splice variant site.

As an example to illustrate the theoretical concept, consider a hypothetical scenario where the whole-genome sequencing (WGS) coverage is 30x. Following the alignment process between exon 6 and the horse WGS data obtained from NCBI, we theoretically identified 27 C and 3 T nucleotides at the variant position. From this information, we could assume that this horse is most probably homozygous C. This example is purely illustrative and does not represent actual experimental results. The genotyping is only based on assumptions and no statistical test has been made. As a reminder, the goal of this study is just to have an overview of the genotyping of horses to analyse the allele frequency. Every data was stored and plotted in excel.

Results and Discussion

Quality Control

After the FastQC, MultiQC and fastp to clean the files, here are the differences between the raw data and the filtered data (Table 3): As we can see from the "% Dups" column, we have an average of 14% of duplicates. The number of duplicates is not very high but needed to be reduced a little bit to save some computational time during the next steps, have a more accurate representation of genomic regions, improve further statistical analysis or avoid over representation of a specific allele for example. The number of sequences varies between 39 million and 47 million which already shows a quite low coverage before filtering.

Sample Name	% Dups	% GC	M Seqs
13161_S10_L008_R1_001	12.7%	42%	38.9
13161_S10_L008_R2_001	11.1%	42%	38.9
13375_S11_L008_R1_001	13.2%	42%	43.8
13375_S11_L008_R2_001	11.8%	42%	43.8
13394_S12_L008_R1_001	14.7%	42%	45.7
13394_S12_L008_R2_001	13.3%	42%	45.7
13523_S13_L008_R1_001	13.9%	42%	39.0
13523_S13_L008_R2_001	12.4%	42%	39.0
2603_S1_L007_R1_001	15.8%	42%	40.2
2603_S1_L007_R2_001	13.9%	42%	40.2
2607_S2_L007_R1_001	15.5%	42%	44.0
2607_S2_L007_R2_001	13.4%	42%	44.0
5674_S3_L007_R1_001	17.0%	42%	47.0
5674_S3_L007_R2_001	15.0%	42%	47.0
5677_S4_L007_R1_001	16.3%	42%	46.4

Table 3: MultiQC on the raw data.

The table (Table 4) shows an improvement in the rate of duplicates which went down to an average of 9%. the number of sequences also decreased to an interval of 34 to 40 million reads.
There are still enough to pursue the analysis.

Sample Name	% Dups	% GC	M Seqs
13161_S10_L008_R1_001_MD_filtered	9.1%	42%	34.9
13161_S10_L008_R2_001_MD_filtered	8.1%	42%	34.9
13375_S11_L008_R1_001_MD_filtered	9.1%	42%	39.2
13375_S11_L008_R2_001_MD_filtered	8.2%	42%	39.2
13394_S12_L008_R1_001_filtered	9.6%	42%	39.9
13394_S12_L008_R2_001_filtered	8.6%	42%	39.9
13523_S13_L008_R1_001_CONTROL_filtered	9.2%	42%	34.5
13523_S13_L008_R2_001_CONTROL_filtered	8.2%	42%	34.5
2603_S1_L007_R1_001_MD_filtered	10.9%	42%	34.2
2603_S1_L007_R2_001_MD_filtered	9.5%	42%	34.2
2607_S2_L007_R1_001_MD_filtered	11.0%	42%	37.5
2607_S2_L007_R2_001_MD_filtered	9.5%	42%	37.5
5674_S3_L007_R1_001_CONTROL_filtered	11.7%	42%	39.1
5674_S3_L007_R2_001_CONTROL_filtered	10.2%	42%	39.1
5677_S4_L007_R1_001_CONTROL_filtered	11.3%	42%	38.8

Table 4: MultiQC on the post fastp filtered data.

Mapping to the reference genome

The mapping to the horse reference genome of these reads showed a very high mapping percentage of approximately 99% for each file. As a result, 13 alignment files in bam format were generated. To facilitate the subsequent variant calling step and considering that two files originated from the same horse, it was decided to merge them. From that, GATK pipeline has been used to retrieve SNPs on 12 files.

SNP calling

Before segregating the SNPs from the INDELs, a total of 12,393,850 variants were initially identified. Subsequently, during the filtering process, no variants were discarded due to low coverage data. The rationale behind this approach was to manually examine the quality of any potentially significant variant. As a result, the final VCF file contained a total of 11,164,360 SNPs.

Fst analysis

Calculation and plotting

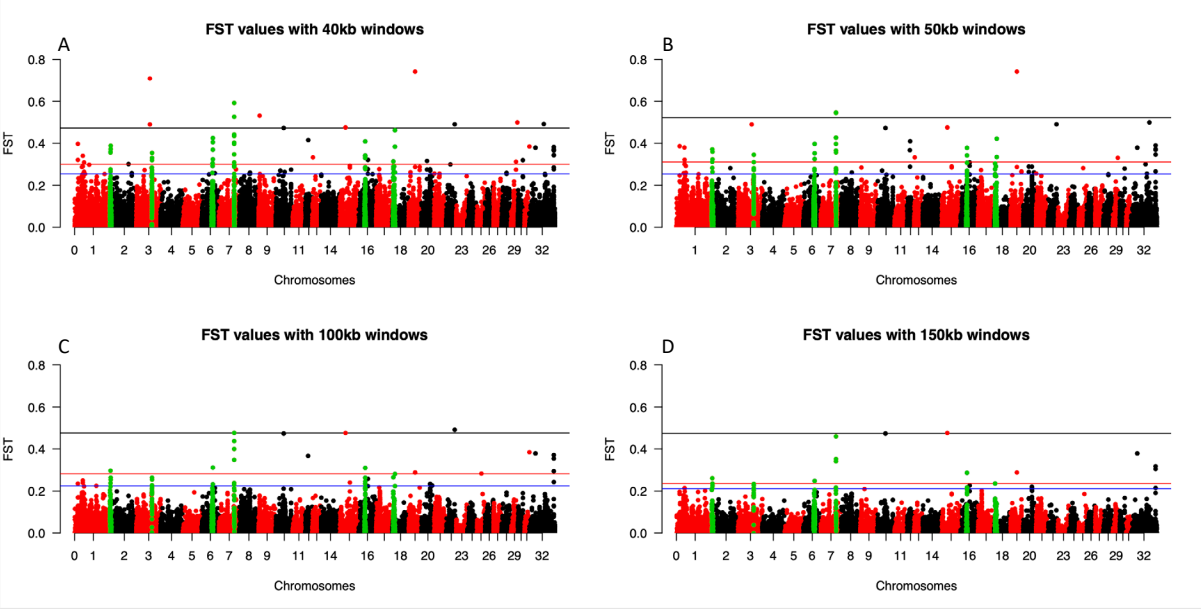


Figure 2: A) Fst plot with 40kb windows and 20kb steps. B) Fst plot with 50kb windows and 25kb steps. C) Fst plot with 100kb windows and 50kb steps. D) Fst plot with 200kb windows and 100kb steps. The black line corresponds to the top 0.0001% windows, the red line for 0.0005% and the blue line for 0.001%.

The plots generated in our analysis reveal multiple regions that surpass the established threshold, indicating conservation across various window sizes (Figure 2, Appendix 4, Appendix 5, Appendix 6, Appendix 7). Each dot in the plot corresponds to a specific window, with the size of each window indicated above the plot. Notably, several dots are observed above the black line, signifying the top 0.001% windows. Additionally, the dots between the red and black lines, representing the top 0.005% windows, were considered for subsequent analysis. Furthermore, certain regions, such as those highlighted by the green peaks in chromosomes 1, 3, 6, 7, 16, and 18, were specifically retained for further investigation.

Pathway analysis/Gene identification

Genes residing within these windows were retrieved and investigated, as detailed in the methodology section. Among them, a gene named THEGL (Testicular Haploid Expressed Gene Protein-Like) caught our attention. According to the GeneCards database, this gene has been implicated as the primary cause of Facial Muscular Dystrophy in humans. Given the rarity of this disease and the lack of related muscular proteins among the other genes within the Fst windows, THEGL emerged as a highly promising candidate.

Initially, in the context of the horse, we discovered four proteins that interacted with our primary gene of interest with String database. Then, in the mouse, we retrieved a set of ten proteins, with four of them being common to the horse dataset.

Lastly, we investigated the paralogous gene of our primary gene, as paralog genes often exhibit protein-level interactions. At this stage, our analysis yielded one main gene of interest and an additional set of 15 genes to explore, encompassing paralogs, enhancers/promoters, as well as protein and gene interactions.

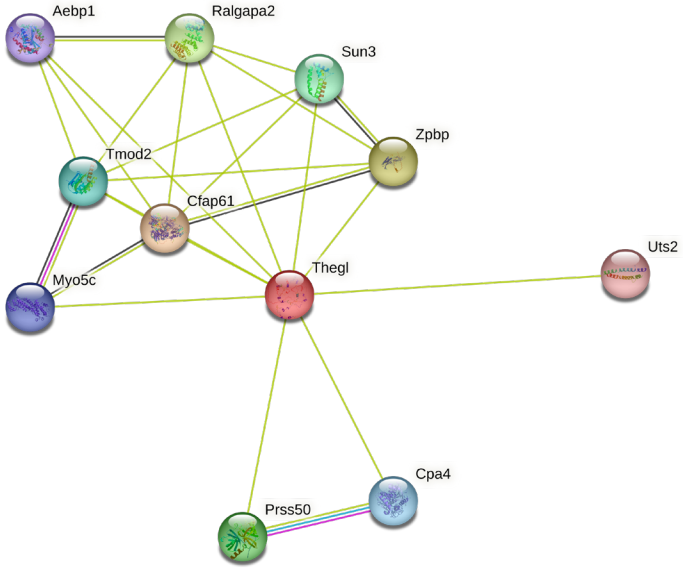


Figure 3: String network plot of the proteins interacting with THEGL. This network has been generated on the basis of THEGL in the mouse.

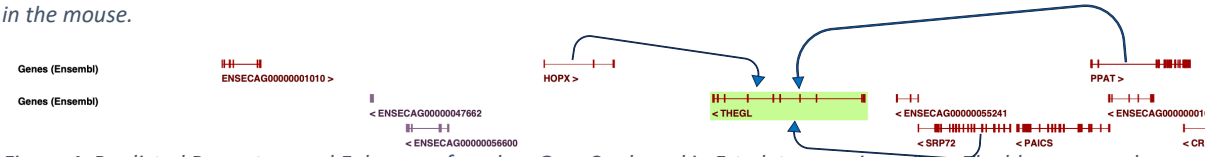


Figure 4: Predicted Promoters and Enhancers found on GeneCards and in Fst plot genomic regions. The blue arrows show the predicted genes interacting with THEGL.

Table 5 shows the summary of all the predicted genes found according to different databases detailed in the method. All these genes could have a link with THEGL:

	GENES
Gene/protein interactions	BRCA1, AEBP1, RALGAPA2, SUN3, ZPBP, TMOD2, CFAP61, MYO5C, UTS2, PRSS50, CPA4
Enhancers/promoters	HOPX, SRP72, PPAT
Paralog	THEG

Table 5: Genes related to THEGL that will be investigated in further analysis.

Only 5 genes (BRCA1, HOPX, SRP72, PPAT and THEG) were analysed due to time issue. This process resulted in the production of an Excel file, comprising the 16 genes of interest, with approximately 200 annotated SNPs for each gene.

SNPs of interest

Building upon THEGL as our starting point, we further explored other genes associated with it, as described in the methodology section. To organize the information obtained regarding the genes and the identified SNPs, we compiled a comprehensive excel file that encompasses

the position of each SNP and its corresponding consequence type. Initially, our focus was on investigating stop gained, start lost, missense, and splice variants, while intron variants and UTR variants were initially excluded from analysis.

GENE	ch	Location	POS	REF	ALT	QUAL	FILTER	Conseq. Type
THEGL	3	exon_6	78025224	T	G	1184.90	PASS	splice region variant

Table 6: Extract of the vcf file. POS stand for the position on the chromosome, REF stands for the reference allele from EquCab 3.0 ALT is the variant nucleotide, FILTER stands for the snps filtration step result.

Within THEGL which is composed of 9 exons, only one splice variant was discovered (Table 6), with no findings of start loss, stop gain, or missense variants. Notably, this splice variant is located at the terminal position of exon 6. Splice variants, such as this one, have the potential to significantly impact splicing events, potentially leading to exon skipping or intron retention. In our case, the presence of a mutation at the very last nucleotide of exon 6 may influence the transcription process of this gene, potentially involving the adjacent intron 7. Specifically, the mutation involves a change from Tyrosine to Guanine (Table 6).
To confirm the presence of the variant, we examined the genotyping of each horse in the excel file. Interestingly, most of the control horses were genotyped as homozygous with a G allele 71% (corresponding to the variant), while the 80% cases were predominantly genotyped as homozygous with a T allele (Table 7). This divergence in genotype between cases and controls strongly indicates a distinct pattern at this specific position. Here is a summary of the genotyping:

	Horses	Genotyping
Cases	Kasse	T|G
	Azuls Tiny Bubble	T|T
	Peggy	T|T
	Zekke	T|T
	Woody	T|T
Controls	None	T|G
	Olive	G|G
	Rita	G|G
	Whittle Wed Wagon	T|T
	She Came Undone	G|G
	Cat House Trix	G|G
	Shesa Smart Cat	G|G

Table 7: Genotype of the 12 horses at the last position of exon 6 in THEGL gene.

Sanger Sequencing

As the coverage was pretty low (around 3x), sanger sequencing has been carried out on those horses (except one case because of a lack of DNA). The quality of the Sanger sequencing was assessed by a laboratory expert, and it has demonstrated a very high overall quality (Figure 4).

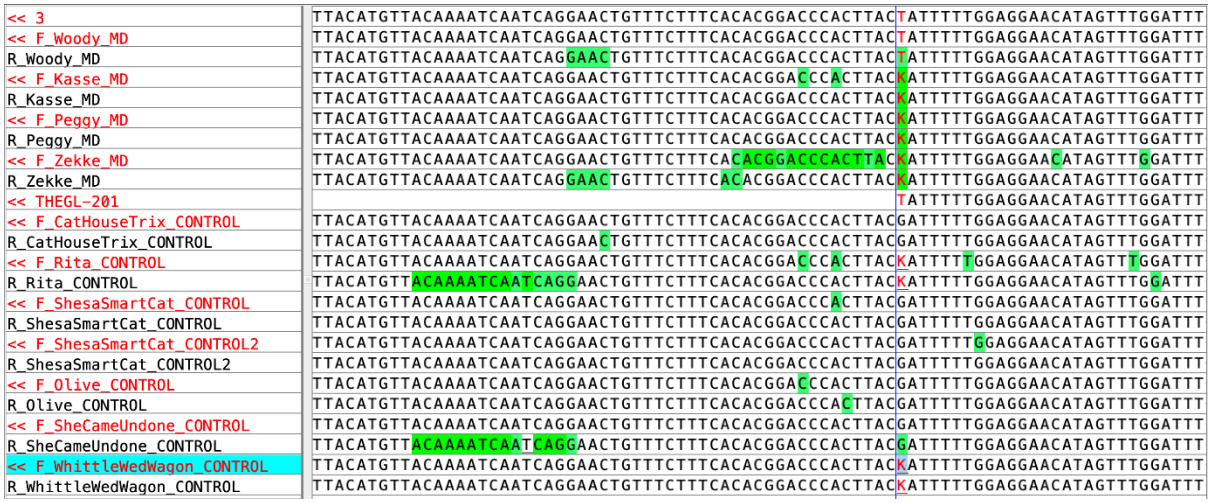


Figure 5: CodonCode Aligner. The first line represents the horse reference genome EquCab 3.0. All the cases are above the white line annotated as the exon_6 of THEGL. Everything below this line are the controls. K stands for heterozygous T|G.

Here is a more comprehensible table translating the image from CodonCode Aligner:

	Horses	Genotyping
Cases	Kasse	T|G
	Azuls Tiny Bubble	-
	Peggy	T|G
	Zekke	T|G
	Woody	T|T
Controls	None	T|G
	Olive	G|G
	Rita	G|G
	Whittle Wed Wagon	T|G
	She Came Undone	G|G
	Cat House Trix	G|G
	Shesa Smart Cat	G|G

Table 8: Genotype of 11 horses after Sanger sequencing on the exon 6 of THEGL gene.

Sanger sequencing being more reliable than our low pass WGS, we will take this second genotyping in account. Important thing to notice, the genotyping of the low coverage WGS being discarded, the Fst is not. We in fact decide to do Fst analysis as it is a very common analysis that could be done with low pass data. The genomic regions observed in the plot are trustful enough, for prove, a gene related to Muscular dystrophy in human has be found.
It is then evident from the observed genotype patterns that a distinct variation is occurring at this specific position, as the genotypes of the cases and controls exhibit clear differentiation.

Moreover, we observed that the controls (Quarter Horses - QH) possess an allele different from the reference (Thoroughbred). Given that QH breeds are known for their muscularity (Endomorph), in contrast to the Thoroughbred (Ectomorph), and considering the association of THEGL with muscle-related functions, we hypothesized that Endomorph breeds bear a G allele at this position, while Ectomorph breeds exhibit a T allele similar to the reference genome.

In this part of the study we will use Endomorph and Ectomorph terms to define two different "body shape" and more precisely two different level of muscularity in the horse. Endomorph will define more muscular horses including breeds such as Appaloosa, Quarter horses or Shetland pony for example. Ectomorph will define a lean horse which can be considered as not vey muscular, including breeds such as Arabian, Akhal Teké or Thoroughbred.

Taking the available data into account and considering the controls as the reference group in our study, it is evident that the cases possess a different nucleotide at this position.

Allele frequency (preliminary results)

To test our hypothesis saying that Ectomorph and Endomorph breeds may have a different allele at the last position at exon 6 and thus explain the strange genotyping of our study's horses, an analysis of allele frequencies at the last position of exon 6 of THEGL was conducted across various horse breeds where the number of alleles has been counted one by one from WGS data. From this counting step, genotyping has been assumed and our findings indicate that there is a discernible difference between endomorph and ectomorph breeds, as illustrated in the following plots:

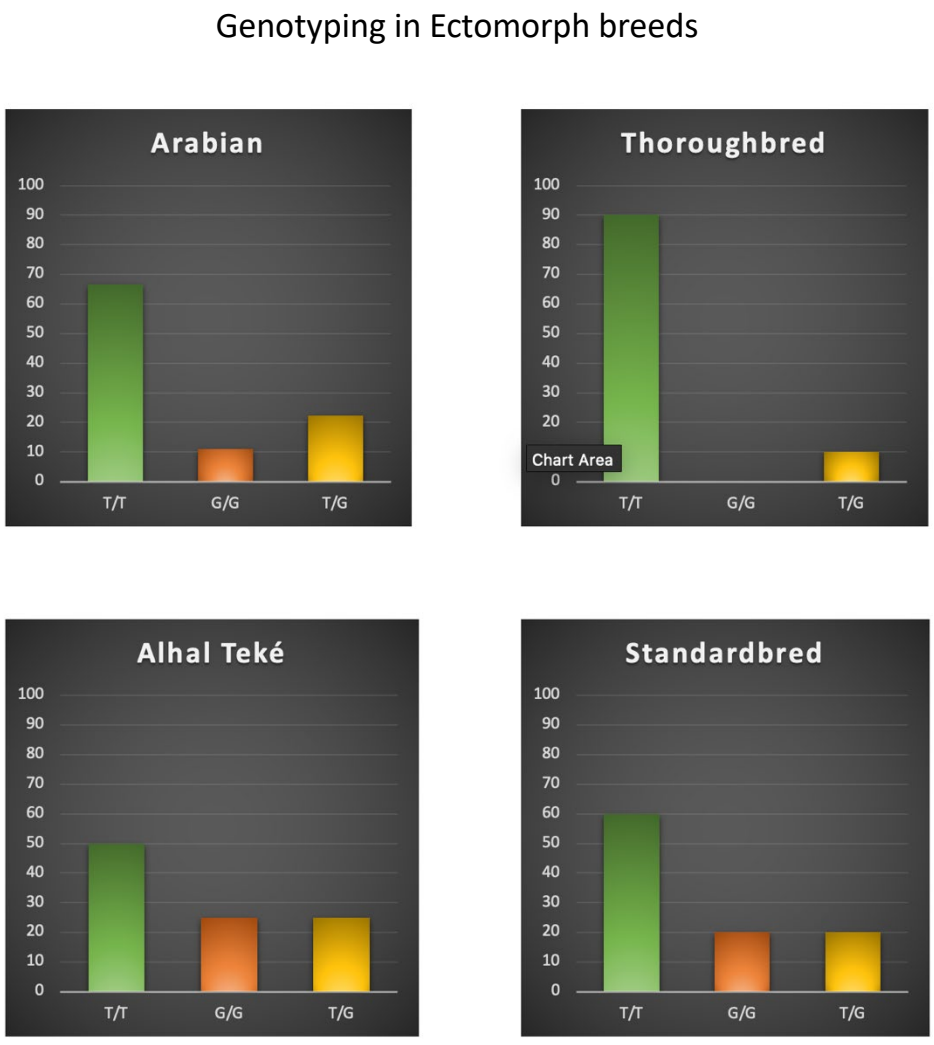


Figure 6: Genotyping of different ectomorph breeds. the Y axis represents the proportion in percent.

Genotyping in Endomorph breeds

Figure 7: Genotyping of different endomorph breeds. the Y axis represents the proportion in percent.

Here is an easier visualisation of the results:

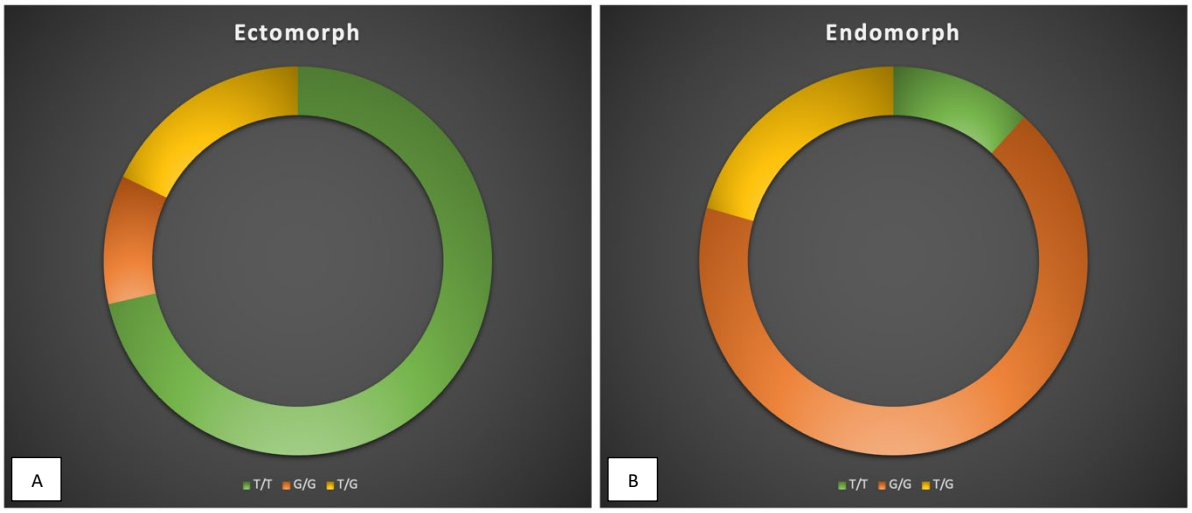


Figure 8: Diagram visualisation of different horse breed morphology genotyping. A) Genotyping of ectomorph breeds. B) Genotyping of endomorph breeds.

A notable distinction is evident between the two breed groups, indicating a potential explanation for the prevalence of the G allele in the majority of our control Quarter Horses. However, it is important to refrain from assuming that these control samples can be considered as reference individuals for this study. In fact, Shetland pony being the major breed in this study, it shows an extremely high G homozygosity across endomorph breeds. We definitely cannot assume that endomorph breeds have this genotype. This could induce that taking the horse reference genome is not the best for our study. In fact, taking the Quarter horses control as reference would be more efficient because of the morphology of the breed.

This allele frequency study will need to be redone to accurately determine the presence of a genetic variant in the cases, and to validate these findings, by sequencing a larger cohort of horses since the current study had a limited sample size. Genotyping will then be properly assumed.

Conclusion

The initial application of Fst analysis on SNPs provided valuable insights into intriguing regions of the equine genome. Among these regions, genes were identified and annotated for further investigation. In particular, our focus turned to a gene associated with muscular dystrophy in humans, namely THEGL. Subsequently, we scrutinized SNPs within this gene to assess potential protein-level consequences. Notably, a splice variant situated at the terminal position of the 6th exon of THEGL captured our attention.

The genotyping of the majority of cases were homozygous with a T allele, aligning with the allele in the horse genome reference, whereas the controls were homozygous with a G allele, representing the variant allele associated with splicing issues. To eliminate any uncertainty regarding potential sequencing artifacts, Sanger sequencing was conducted on this genomic region. The results confirmed that controls indeed displayed a homozygous G genotype for 71% of them, while the cases exhibited a heterozygous T|G genotype, except for one case that remained homozygous T|T. The clear differentiation in genotypes between cases and controls emphasized the importance of further investigating this genomic region.

However, a question arose as to why the controls possessed a different allele than the reference genome. Could this discrepancy be attributed to breed-related factors? Consequently, a new hypothesis was formulated, suggesting that endomorph and ectomorph horses may show distinct alleles at this specific position. To validate this hypothesis and establish the reference allele for this population, an investigation of allele frequencies was conducted across multiple horse breeds. We anticipated observing endomorph horses with an homozygous G genotypes at this position, while ectomorph horses were expected to be homozygous T. Although our study revealed a significant difference, it is premature to draw any definitive conclusions at this stage. Additional analysis is necessary to provide a more comprehensive understanding of the observed findings.

With the new sequencing and higher coverage, additional analysis methods could be explored. Runs of Homozygosity may serve as a valuable complementary study to Fst analysis. Improved coverage will enhance the reliability of variant calling and lead to refined filtering steps. Enrichment and pathway analysis could also be carried out with a better reliability. In the future, advancements in protein structure analysis and the inclusion of RNAseq data analysis will complement the existing DNA data analysis.

Popular science summary

In this study, we investigated a neuromuscular disorder called Myotonic Dystrophy (MD) in horses. MD leads to various muscle problems, like stiffness and abnormal relaxation. We wanted to understand its genetic cause and how it affects the horse's well-being.

To do this, we used advanced genetic techniques to analyse the DNA of affected horses and healthy ones. We found several regions in the horse genome that may be related to MD. By knowing the genetic cause, we can prevent the disease in future generations through selective breeding.

Out of these regions, a gene called THEGL has been found, which has a link to muscle problems in humans. We discovered some variations in this gene in the horses affected by MD. However, further research is needed to confirm our findings.

Our study is a steppingstone towards developing better ways to diagnose and treat MD in horses. By understanding the genetic basis of this disease, we can improve the health and welfare of these animals.

References

American Quarter Horse | breed of horse | Britannica [WWW Document], n.d. URL https://www.britannica.com/animal/American-Quarter-Horse (accessed 5.20.23).

Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data [WWW Document], n.d. URL https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed 5.25.23).

Bird, T.D., 1993. Myotonic Dystrophy Type 1, in: Adam, M.P., Mirzaa, G.M., Pagon, R.A., Wallace, S.E., Bean, L.J., Gripp, K.W., Amemiya, A. (Eds.), GeneReviews®. University of Washington, Seattle, Seattle (WA).

Chen, S., Zhou, Y., Chen, Y., Gu, J., 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34, i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Chuang, L.-Y., Cheng, Y.-H., Yang, C.-H., 2013. Specific primer design for the polymerase chain reaction. Biotechnol. Lett. 35, 1541–1549. https://doi.org/10.1007/s10529-013-1249-8

Day, J.W., Ranum, L.P.W., 2005. Genetics and molecular pathogenesis of the myotonic dystrophies. Curr. Neurol. Neurosci. Rep. 5, 55–59. https://doi.org/10.1007/s11910-005-0024-1

DNA Sequencing | Understanding the genetic code [WWW Document], n.d. URL https://emea.illumina.com/techniques/sequencing/dna-sequencing.html (accessed 5.25.23).

Ewels, P., Magnusson, M., Lundin, S., Käller, M., 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32, 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Iny Stein, T., Rosen, N., Kohn, A., Twik, M., Safran, M., Lancet, D., Cohen, D., 2017. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database J. Biol. Databases Curation 2017, bax028. https://doi.org/10.1093/database/bax028

GATK [WWW Document], n.d. URL https://gatk.broadinstitute.org/hc/en-us (accessed 5.25.23).

Heatwole, C., Johnson, N., Bode, R., Dekdebrun, J., Dilek, N., Hilbert, J.E., Luebbe, E., Martens, W., McDermott, M.P., Quinn, C., Rothrock, N., Thornton, C., Vickrey, B.G., Victorson, D., Moxley, R.T., 2015. Patient-Reported Impact of Symptoms in Myotonic Dystrophy Type 2 (PRISM-2). Neurology 85, 2136–2146. https://doi.org/10.1212/WNL.0000000000002225

Hilbert, J.E., Barohn, R.J., Clemens, P.R., Luebbe, E.A., Martens, W.B., McDermott, M.P., Parkhill, A.L., Tawil, R., Thornton, C.A., Moxley, R.T., 2017. High frequency of gastrointestinal manifestations in myotonic dystrophy type 1 and type 2. Neurology 89, 1348–1354. https://doi.org/10.1212/WNL.0000000000004420

Kalbfleisch, T.S., Rice, E.S., DePriest, M.S., Walenz, B.P., Hestand, M.S., Vermeesch, J.R., O'Connell, B.L., Fiddes, I.T., Vershinina, A.O., Saremi, N.F., Petersen, J.L., Finno, C.J., Bellone, R.R., McCue, M.E., Brooks, S.A., Bailey, E., Orlando, L., Green, R.E., Miller, D.C., Antczak, D.F., MacLeod, J.N., 2018. Improved reference genome for the domestic horse increases assembly contiguity and composition. Commun. Biol. 1, 1–8. https://doi.org/10.1038/s42003-018-0199-z

Langmead, B., Wilks, C., Antonescu, V., Charles, R., 2019. Scaling read aligners to hundreds of threads on general-purpose processors. Bioinformatics 35, 421–432. https://doi.org/10.1093/bioinformatics/bty648

Mahadevan, M., Tsilfidis, C., Sabourin, L., Shutler, G., Amemiya, C., Jansen, G., Neville, C., Narang, M., Barceló, J., O'Hoy, K., 1992. Myotonic dystrophy mutation: an unstable CTG repeat in the 3' untranslated region of the gene. Science 255, 1253–1255. https://doi.org/10.1126/science.1546325

Martin, F.J., Amode, M.R., Aneja, A., Austine-Orimoloye, O., Azov, A.G., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., Bhurji, S.K., Bignell, A., Boddu, S., Branco Lins, P.R., Brooks, L., Ramaraju, S.B., Charkhchi, M., Cockburn, A., Da Rin Fiorretto, L., Davidson, C., Dodiya, K., Donaldson, S., El Houdaigui, B., El Naboulsi, T., Fatima, R., Giron, C.G., Genez, T., Ghattaoraya, G.S., Martinez, J.G., Guijarro, C., Hardy, M., Hollis, Z., Hourlier, T., Hunt, T., Kay, M., Kaykala, V., Le, T., Lemos, D., Marques-Coelho, D., Marugán, J.C., Merino, G.A., Mirabueno, L.P., Mushtaq, A., Hossain, S.N., Ogeh, D.N., Sakthivel, M.P., Parker, A., Perry, M., Piližota, I., Prosovetskaia, I., Pérez-Silva, J.G., Salam, A.I.A., Saraiva-Agostinho, N., Schuilenburg, H., Sheppard, D., Sinha, S., Sipos, B., Stark, W., Steed, E., Sukumaran, R., Sumathipala, D., Suner, M.-M., Surapaneni, L., Sutinen, K., Szpak, M., Tricomi, F.F., Urbina-Gómez, D., Veidenberg, A., Walsh, T.A., Walts, B., Wass, E., Willhoft, N., Allen, J., Alvarez-Jarreta, J., Chakiachvili, M., Flint, B., Giorgetti, S., Haggerty, L., Ilsley, G.R., Loveland, J.E., Moore, B., Mudge, J.M., Tate, J., Thybert, D., Trevanion, S.J., Winterbottom, A., Frankish, A., Hunt, S.E., Ruffier, M., Cunningham, F., Dyer, S., Finn, R.D., Howe, K.L., Harrison, P.W., Yates, A.D., Flicek, P., 2023. Ensembl 2023. Nucleic Acids Res. 51, D933–D941. https://doi.org/10.1093/nar/gkac958

Modoni, A., Silvestri, G., Grazia Pomponi, M., Mangiola, F., Tonali, P.A., Marra, C., 2004. Characterization of the Pattern of Cognitive Impairment in Myotonic Dystrophy Type 1. Arch. Neurol. 61, 1943–1947. https://doi.org/10.1001/archneur.61.12.1943

Picard Tools - By Broad Institute [WWW Document], n.d. URL https://broadinstitute.github.io/picard/ (accessed 5.25.23).

R: The R Project for Statistical Computing [WWW Document], n.d. URL https://www.r-project.org/ (accessed 3.28.23).

Rodchenkov, I., Babur, O., Luna, A., Aksoy, B.A., Wong, J.V., Fong, D., Franz, M., Siper, M.C., Cheung, M., Wrana, M., Mistry, H., Mosier, L., Dlin, J., Wen, Q., O'Callaghan, C., Li, W., Elder, G., Smith, P.T., Dallago, C., Cerami, E., Gross, B., Dogrusoz, U., Demir, E., Bader, G.D., Sander, C., 2020. Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. Nucleic Acids Res. 48, D489–D497. https://doi.org/10.1093/nar/gkz946

Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., Nativ, N., Bahir, I., Doniger, T., Krug, H., Sirota-Madi, A., Olender, T., Golan, Y., Stelzer, G., Harel, A., Lancet, D., 2010. GeneCards Version 3: the human gene integrator. Database J. Biol. Databases Curation 2010, baq020. https://doi.org/10.1093/database/baq020

Spiro, A.J., 2002. Myotonic dystrophy, 3rd edition: By P.S. Harper. 448 pp. Philadelphia: WB Saunders, 2001. $85.00. ISBN 0-7020-2152-0. Pediatr. Neurol. 27, 76. https://doi.org/10.1016/S0887-8994(02)00407-1

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T.I., Nudel, R., Lieder, I., Mazor, Y., Kaplan, S., Dahary, D., Warshawsky, D., Guan-Golan, Y., Kohn, A., Rappaport, N., Safran, M., Lancet, D., 2016. The GeneCards Suite: From

Gene Data Mining to Disease Genome Sequence Analyses. Curr. Protoc. Bioinforma. 54, 1.30.1-1.30.33. https://doi.org/10.1002/cpbi.5

Szklarczyk, D., Gable, A.L., Nastou, K.C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N.T., Legeay, M., Fang, T., Bork, P., Jensen, L.J., von Mering, C., 2021. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. 49, D605–D612. https://doi.org/10.1093/nar/gkaa1074

VCFtools [WWW Document], n.d. URL https://vcftools.sourceforge.net/ (accessed 5.25.23).

von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., Snel, B., 2003. STRING: a database of predicted functional associations between proteins. Nucleic Acids Res. 31, 258–261. https://doi.org/10.1093/nar/gkg034

Wade, C.M., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., Lear, T.L., Adelson, D.L., Bailey, E., Bellone, R.R., Blöcker, H., Distl, O., Edgar, R.C., Garber, M., Leeb, T., Mauceli, E., MacLeod, J.N., Penedo, M.C.T., Raison, J.M., Sharpe, T., Vogel, J., Andersson, L., Antczak, D.F., Biagi, T., Binns, M.M., Chowdhary, B.P., Coleman, S.J., Della Valle, G., Fryc, S., Guérin, G., Hasegawa, T., Hill, E.W., Jurka, J., Kiialainen, A., Lindgren, G., Liu, J., Magnani, E., Mickelson, J.R., Murray, J., Nergadze, S.G., Onofrio, R., Pedroni, S., Piras, M.F., Raudsepp, T., Rocchi, M., Røed, K.H., Ryder, O.A., Searle, S., Skow, L., Swinburne, J.E., Syvänen, A.C., Tozaki, T., Valberg, S.J., Vaudin, M., White, J.R., Zody, M.C., Broad Institute Genome Sequencing Platform, Broad Institute Whole Genome Assembly Team, Lander, E.S., Lindblad-Toh, K., 2009. Genome sequence, comparative analysis, and population genetics of the domestic horse. Science 326, 865–867. https://doi.org/10.1126/science.1178158

Appendix

	Sequence	Annealing temperature
Forward primer	cgagtctctgatccttcccc	61.1 °C
Reverse primer	gcgcatgggaggtagttaga	60.2 °C

Appendix 1: Sanger sequencing primers details for exon 6 of THEGL.

```
### Whole pipeline for SNP calling for myotonia project ###


# Mapping of all the WGS horses to EquCab3.0 as reference #

bowtie2-build -f
../ref_genome/ncbi_dataset/data/GCF_002863925.1/GCF_002863925.1_EquCab3.0_
genomic.fna EquCab_index

bowtie2 -p 24 -x EquCab_index -1
../filtered_data/MD/HORSE_filtered.fastq.gz -2
../filtered_data/MD/HORSE_filtered.fastq.gz -S 13161_Kasse_MD.sam
samtools view -b HORSE.sam > HORSE.bam
rm HORSE.sam


# INDEX creation #

picard CreateSequenceDictionary R=reference.fasta O=reference.dict
samtools faidx GCF_002863925.1_EquCab3.0_genomic.fasta
picard SortSam SO=coordinate INPUT=../mapped_data/HORSE.bam OUTPUT=
HORSE_sorted.bam VALIDATION_STRINGENCY=LENIENT CREATE_INDEX=true

## Merging of 2 samples from the same horse ##
#samtools merge HORSE_CONTROL.bam HORSE_CONTROL_sorted1.bam
HORSE_CONTROL_sorted2.bam
#picard SortSam SO=coordinate INPUT=../mapped_data/HORSE_CONTROL.bam
OUTPUT=HORSE_CONTROL_sorted.bam VALIDATION_STRINGENCY=LENIENT
CREATE_INDEX=true



# Marking the duplicates following by bamaddrg to add read group to bam
files #

picard MarkDuplicates I= HORSE_sorted.bam O= HORSE_sorted_marked_dupes.bam
M=marked_dup_metrics_HORSE.txt
bamaddrg -b HORSE_sorted_marked_dupes.bam >
HORSE_sorted_marked_dupes_headers.bam
```

```
# Indexing again the new bam files with RG #

samtools index HORSE_sorted_marked_dupes_headers.bam



# SNP calling #

gatk -T HaplotypeCaller -R
../ref_genome/ncbi_dataset/data/GCF_002863925.1/GCF_002863925.1_EquCab3.0_
genomic.fasta -I HORSE_sorted_marked_dupes_headers.bam -o HORSE.g.vcf.gz
-ERC GVCF



# Genotyping all the gvcf files previously created #

gatk GenotypeGVCFs -R
../../ref_genome/ncbi_dataset/data/GCF_002863925.1/GCF_002863925.1_EquCab3
.0_genomic.fasta -V HORSE.g.vcf.gz -O HORSE.vcf.gz



# Combining all the vcf files #

bcftools merge HORSE1.vcf.gz HORSE2.vcf.gz HORSE3.vcf.gz HORSE4.vcf.gz
HORSE5.vcf.gz HORSE6.vcf.gz HORSE7.vcf.gz HORSE8.vcf.gz HORSE9.vcf.gz
HORSE10.vcf.gz HORSE11.vcf.gz HORSE12.vcf.gz -o all.vcf.gz



# Separating the SNPs from the INDELs #

gatk SelectVariants -V all.vcf.gz -select-type SNP -O all_SNPs.vcf.gz
gatk SelectVariants -V all.vcf.gz -select-type INDEL -O all_INDELs.vcf.gz



# Filtration step based on quality score of the SNPs (basic filtration
from GATK manual) #

gatk VariantFiltration -V all_SNPs.vcf.gz -filter "QD < 2.0" --filter-name
"QD2" -filter "QUAL < 30.0" --filter-name "QUAL30" -filter "SOR > 3.0" --
filter-name "SOR3" -filter "FS > 60.0" --filter-name "FS60" -filter "MQ <
40.0" --filter-name "MQ40" -filter "MQRankSum < -12.5" --filter-name
"MQRankSum-12.5" -filter "ReadPosRankSum < -8.0" --filter-name
"ReadPosRankSum-8" -O snps_filtered.vcf.gz
```

```
# Fst calculation using vcftools #

vcftools --gzvcf snps_filtered.vcf.gz --weir-fst-pop MD --weir-fst-pop
CONTROL --fst-window-size 100000 --fst-window-step 50000 --out
./MD_CONTROL_100kb
```

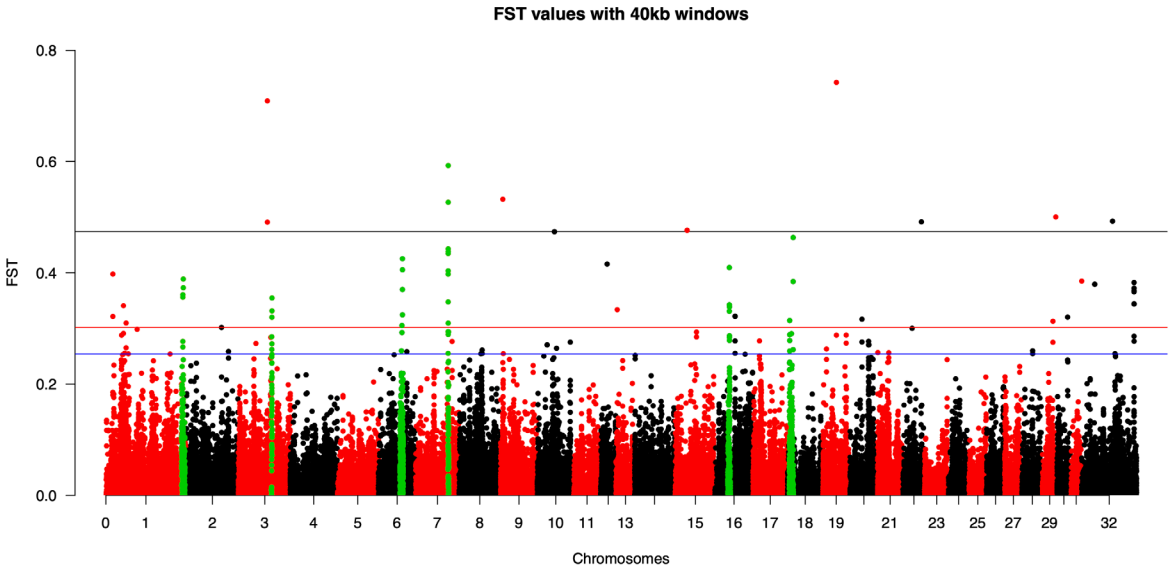
Appendix 2: Whole analysis from quality control to Fst calculation bash script.

```
library(tidyverse)
library(qqman)
attach(mtcars)
par(mfrow=c(2,2))



Fst_40kb <- read_tsv("./Desktop/SLU/Master
thesis/DNA/FST/Files/FST_final_40kb.txt")
Fst_50kb <- read_tsv("./Desktop/SLU/Master
thesis/DNA/FST/Files/new_Fst_50kb.txt")
Fst_100kb <- read_tsv("./Desktop/SLU/Master
thesis/DNA/FST/Files/FST_final_100kb.txt")
Fst_150kb <- read_tsv("./Desktop/SLU/Master
thesis/DNA/FST/Files/FST_final_150kb.txt")

length_40 <- dim(Fst_40kb)[1]
length_50 <- dim(Fst_50kb)[1]
length_100 <- dim(Fst_100kb)[1]
length_150 <- dim(Fst_150kb)[1]

Fst_40kb$SNP <- paste('SNP',1:length_40)
Fst_50kb$SNP <- paste('SNP',1:length_50)
Fst_100kb$SNP <- paste('SNP',1:length_100)
Fst_150kb$SNP <- paste('SNP',1:length_150)

Fst_40kb %>%
  filter(CHROM==7 & BIN_START>75110001 & BIN_END<77160000
         | CHROM == 3 & BIN_START>77013311 & BIN_END<78865858
         | CHROM == 1 & BIN_START>177013311 & BIN_END<184865858
         | CHROM == 18 & BIN_START>4000000 & BIN_END<14000000
         | CHROM == 6 &BIN_START>50013311 & BIN_END<57865858
         | CHROM == 16 &BIN_START>30013311 & BIN_END<35865858
  )->importantSNPS_40kb
conserved_regions_40kb<-importantSNPS_40kb$SNP

Fst_50kb %>%
  filter(CHROM==7 & BIN_START>75110001 & BIN_END<77160000
         | CHROM == 3 & BIN_START>77013311 & BIN_END<78865858
         | CHROM == 1 & BIN_START>177013311 & BIN_END<184865858
         | CHROM == 18 & BIN_START>4000000 & BIN_END<14000000
         | CHROM == 6 &BIN_START>50013311 & BIN_END<57865858
         | CHROM == 16 &BIN_START>30013311 & BIN_END<35865858
  )->importantSNPS_50kb
conserved_regions_50kb<-importantSNPS_50kb$SNP
```

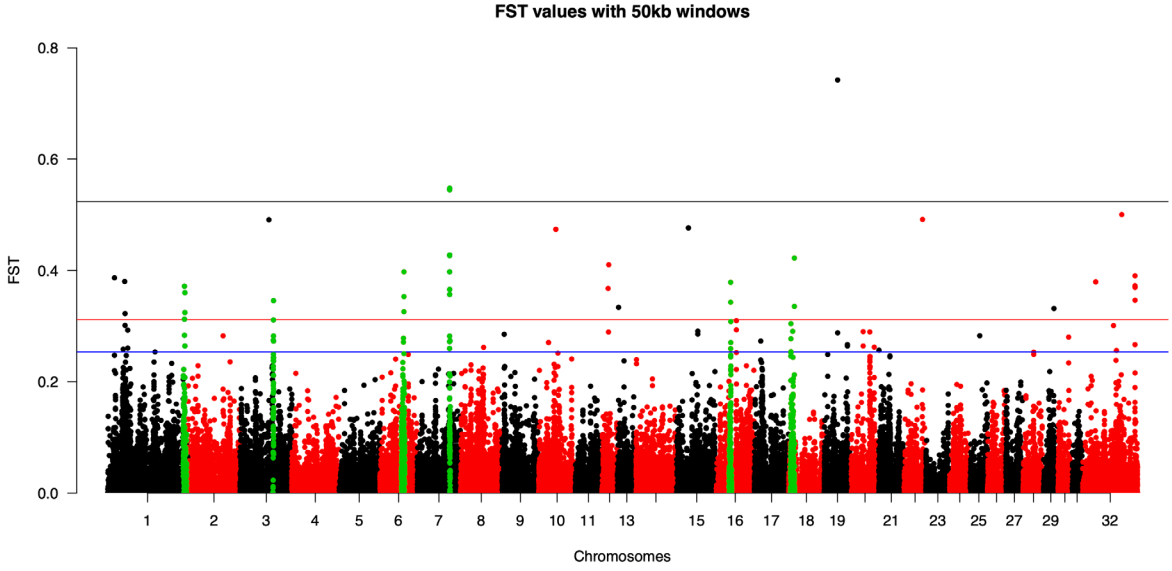
```
Fst_100kb %>%
  filter(CHROM==7 & BIN_START>75110001 & BIN_END<77160000
         | CHROM == 3 & BIN_START>77013311 & BIN_END<78865858
         | CHROM == 1 & BIN_START>177013311 & BIN_END<184865858
         | CHROM == 18 & BIN_START>4000000 & BIN_END<14000000
         | CHROM == 6 &BIN_START>50013311 & BIN_END<57865858
         | CHROM == 16 &BIN_START>30013311 & BIN_END<35865858
  )->importantSNPS_100kb
conserved_regions_100kb<-importantSNPS_100kb$SNP

Fst_150kb %>%
  filter(CHROM==7 & BIN_START>75110001 & BIN_END<77160000
         | CHROM == 3 & BIN_START>77013311 & BIN_END<78865858
         | CHROM == 1 & BIN_START>177013311 & BIN_END<184865858
         | CHROM == 18 & BIN_START>4000000 & BIN_END<14000000
         | CHROM == 6 &BIN_START>50013311 & BIN_END<57865858
         | CHROM == 16 &BIN_START>30013311 & BIN_END<35865858
  )->importantSNPS_150kb
conserved_regions_150kb<-importantSNPS_150kb$SNP

q40<-quantile(Fst_40kb$MEAN_FST,0.9999)
q50<-quantile(Fst_50kb$MEAN_FST,0.9999)
q100<-quantile(Fst_100kb$MEAN_FST,0.9999)
q150<-quantile(Fst_150kb$MEAN_FST,0.9999)

r40<-quantile(Fst_40kb$MEAN_FST,0.9995)
r50<-quantile(Fst_50kb$MEAN_FST,0.9995)
r100<-quantile(Fst_100kb$MEAN_FST,0.9995)
r150<-quantile(Fst_150kb$MEAN_FST,0.9995)

s40<-quantile(Fst_40kb$MEAN_FST,0.9990)
s50<-quantile(Fst_50kb$MEAN_FST,0.9990)
s100<-quantile(Fst_100kb$MEAN_FST,0.9990)
s150<-quantile(Fst_150kb$MEAN_FST,0.9990)

q40[[1]]
q50[[1]]
q100[[1]]
q150[[1]]

r40[[1]]
r50[[1]]
r100[[1]]
r150[[1]]

s40[[1]]
s50[[1]]
s100[[1]]
s150[[1]]

p<-manhattan(Fst_40kb,chr='CHROM', bp='BIN_START',
             p='MEAN_FST',snp='SNP',
             logp=FALSE,ylab='FST',xlab='Chromosomes',
             col=c("black", "red"),
```

```
            abline(h = q40[[1]], col='black', lwd=1),
            genomewideline = r40,
            suggestiveline = s40,
            ylim = c(0, 0.8),
            main = 'FST values with 40kb windows',
            highlight = conserved_regions_40kb)
t<-manhattan(Fst_50kb,chr='CHROM', bp='BIN_START',
            p='MEAN_FST',snp='SNP',
            logp=FALSE,ylab='FST',xlab='Chromosomes',
            col=c("red", "black"),
            abline(h = q50[[1]], col='black', lwd=1),
            genomewideline = r50,
            suggestiveline = s50,
            ylim = c(0, 0.8),
            main = 'FST values with 50kb windows',
            highlight = conserved_regions_50kb)
u<-manhattan(Fst_100kb,chr='CHROM', bp='BIN_START',
            p='MEAN_FST',snp='SNP',
            logp=FALSE,ylab='FST',xlab='Chromosomes',
            col=c("black", "red"),
            abline(h = q100[[1]], col='black', lwd=1),
            genomewideline = r100,
            suggestiveline = s100,
            ylim = c(0, 0.8),
            main = 'FST values with 100kb windows',
            highlight = conserved_regions_100kb)
v<-manhattan(Fst_150kb,chr='CHROM', bp='BIN_START',
            p='MEAN_FST',snp='SNP',
            logp=FALSE,ylab='FST',xlab='Chromosomes',
            col=c("black", "red"),
            abline(h = q150[[1]], col='black', lwd=1),
            genomewideline = r150,
            suggestiveline = s150,
            ylim = c(0, 0.8),
            main = 'FST values with 150kb windows',
            highlight = conserved_regions_150kb)
```

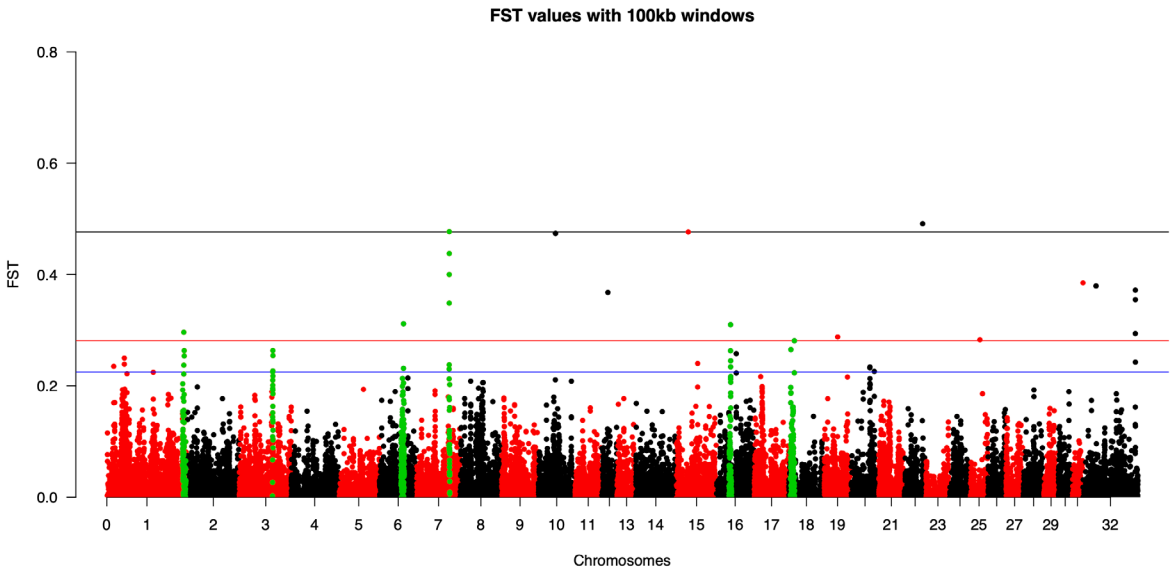
Appendix 3: R code to plot Fst on four different windows: 40kb, 50kb, 100kb and 150kb.



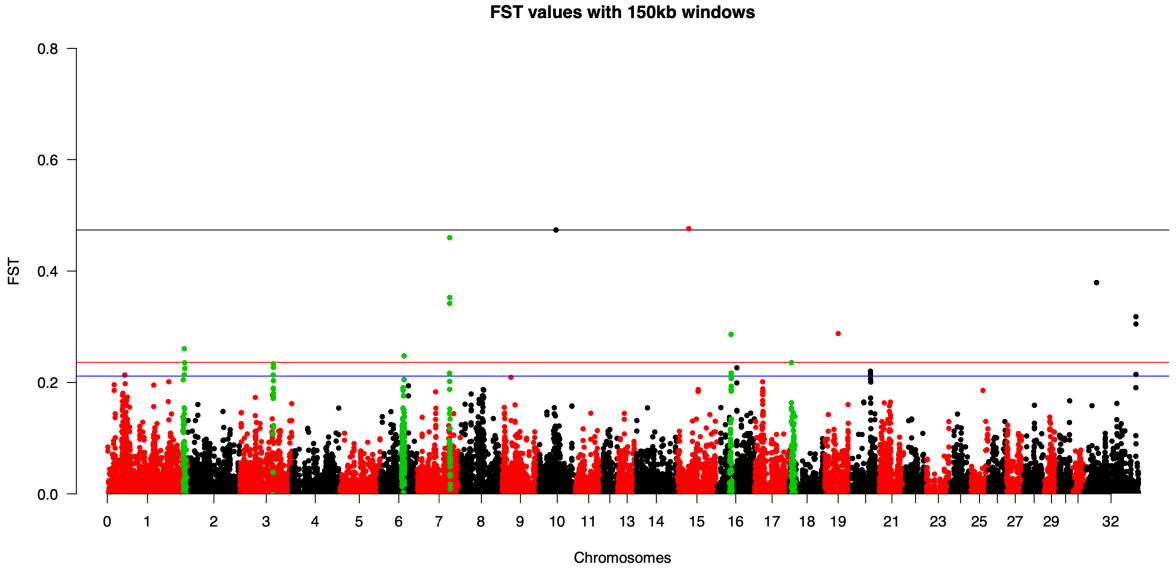
Appendix 4: Fst plot with 40kb windows and 20kb steps. The black line corresponds to the top 0.0001% windows, the red line for 0.0005% and the blue line for 0.001%.



Appendix 5: Fst plot with 50kb windows and 25kb steps. The black line corresponds to the top 0.0001% windows, the red line for 0.0005% and the blue line for 0.001%.



Appendix 6: Fst plot with 100kb windows and 50kb steps. The black line corresponds to the top 0.0001% windows, the red line for 0.0005% and the blue line for 0.001%.



Appendix 7: Fst plot with 200kb windows and 100kb steps. The black line corresponds to the top 0.0001% windows, the red line for 0.0005% and the blue line for 0.001%.

Table of Appendix:

Publishing and archiving

Approved students' theses at SLU are published electronically. As a student, you have the copyright to your own work and need to approve the electronic publishing. If you check the box for **YES**, the full text (pdf file) and metadata will be visible and searchable online. If you check the box for **NO**, only the metadata and the abstract will be visible and searchable online. Nevertheless, when the document is uploaded it will still be archived as a digital file.
If you are more than one author you all need to agree on a decision. Read about SLU's publishing agreement here: https://www.slu.se/en/subweb/library/publish-and-analyse/register-and-publish/agreement-for-publishing/.

☒ YES, I/we hereby give permission to publish the present thesis in accordance with the SLU agreement regarding the transfer of the right to publish a work.

☐ NO, I/we do not give permission to publish the present work. The work will still be archived and its metadata and abstract will be visible and searchable.