Norwegian University
of Life Sciences

Master's Thesis 2022 30 ECTS
Faculty of Biosciences
Department of Animal and Aquacultural Sciences

Fine mapping of a QTL on chromosome 23 for white markings in Scandanavian Fjord horses.

Charlotte Cufffe
European Master in Animal Breeding and Genetics

EUROPEAN MASTER'S IN ANIMAL BREEDING AND GENETICS

Fine mapping of a QTL for white markings on chromosome 23 in Scandinavian Fjord horses.

Charlotte Cuffe

May 2022

Main supervisor: Dag Inge Våge (NMBU)

Co-supervisor(s): Richard Crooijmans (WUR)

Matthew Peter Kent (NMBU)

Gunnar Klemetsdal (NMBU)

Hanne Fjerdingby Olsen (NMBU)

Thu-Hien To (NMBU)

Co-funded by the Erasmus+ Programme of the European Union

Acknowledgements

Thanks must be given firstly to my supervisors Dag Inge, Gunnar, Hanne, Matthew, Hien and Richard for all their help given throughout this thesis. Especially Dag Inge for always having the door open to discuss when I wasn't sure and to Hien for always being at the end of a Teams message to answer my endless questions about the programmes I was using. Thanks as well to my family for their unwavering support throughout and to my Dad for always listening to me talk through problems. Mentions must be given as well to Tina for the coffee and company in the study room and to Izzy for always being there with words of encouragement.

Abstract

In Fjord horses, it is undesirable to have extensive white markings on the head or legs. White markings tended to be multifactorial pleitropic traits with phenotypic effects not solely restricted to coat colour but a variety of other syndromes which could have a negative impact on the horse. In this thesis, 328 Fjord horses underwent a GWAS for white markings based on 67k SNPchip data. Of these 328 horses, 19 had white markings on the head or body. This GWAS identified 2 regions on chromosome 23 with significant association to white markings. Following this, 16 horses, evenly split between cases and controls, underwent resequencing culminating in variant calling using the GATK and FreeBayes. Further analysis was carried out on the peaks identified in the GWAS, as well as regions containing known white marking genes. This further analysis comprised of differences in allele frequency, Weir and Cockerham's Fst and variant annotation. No clear mutations were identified in a protein coding gene, however an association with U6 spliceosomal RNA were indicated following filtering on variant annotation.

Contents

Background

Colour Genetics

In horses, coat colours were among the first characteristics to be analysed at the genetic level as they tend to follow simple mendelian inheritance patterns (Rieder, 2009). Coat colour can also be seen as a feature of domestication in the modern horse. In the wild a limited variety of coat colours are seen, while a greater range of colours is visible in domesticated populations (Klungland & Våge, 2000). This can be regarded as one of the more obvious signatures of selection in domesticated populations, with a greater variety of colours occurring due to selective breeding strategies during domestication. Nowadays there are several horse breeds that are based around specific coat colours, such as Fjord horses all having a similar dun coat colour, or the American Paint horse differentiated from the Quarter horse primarily from the presence of large white marks over the entire body.

Often animals with white markings or spotting were actively selected for as their striking appearance made them highly valuable. This resulted in many spontaneous mutations for white markings being maintained in domesticated populations. White markings occur due to a lack of melanocytes in the skin and hair follicles caused by several different mutations to a few different genes. Due to these markings and therefore also the mutations that caused them being maintained in populations, there has been ample opportunity to study the effect of these mutations on the migration, proliferation, differentiation and survival of melanocytes (Hauswirth et al., 2012). Mutations acting on melanocytes are an important focus for research as genes that act on melanocyte development and distribution also influence a variety of other cells such as primordial germ cells, haematopoietic cells and neurones (Rieder, 2009).

Equine coat colour is determined, at its most basic level, by the presence or absence of the Agouti and Extension genes. Extension controls the base coat colour, red or black, while agouti is concerned with the distribution of red and black pigmentation to the points or across the entire body (Cone et al., 1996; Searle, 1968). The extension locus encodes Melanocortin 1 Receptor (MC1R) while the Agouti signalling protein encoded by the Agouti locus acts as an indirect antagonist on MC1R (Cone et al., 1996; Rieder, 2009). Pheomelanin is produced due to a losof function mutation of MC1R. While eumelanin is the result of a gain of function mutation in MC1R or a loss of function in ASIP (Rieder, 2009; Rieder et al., 2001). At the extension locus the presence of one or two wild type E alleles will result in a black coat colour (depending on the alleles present at the Agouti locus), however two recessive e alleles will result in a chestnut coat. Then the Agouti locus will determine whether the black is across the entire body, if two copies of the recessive a allele are present, or with the presence

of one or two of the wild type A allele, black will be restricted to the points causing the bay coat colour. This results in the three basic colours, red (chestnut), black or bay.

While genetically these three basic colours exist in all breeds, many other coat colours can be seen. This is the result of different genes acting on these base colours. An example of this would be dilution genes, such as cream, champagne, dun and silver which lighten the base colour. These genes also often have a dosage effect where two copies results in a lighter colour, a double dilute. As an illustration, on a chestnut base colour one copy of the cream gene would give you a palomino while two copies would give you a cremello. With the cream gene the wild type is denoted by C and indicates no change to the base colour, while Cr is the alternate allele that results in a lightening effect on the base colour.

While most of the dilution genes will have a dosage effect, dun will not. The dun allele is completely dominant over all basic coat colours and one cannot distinguish between heterozygotes and homozygotes (Rieder, 2009). The dun coat colour is considered to be the wild type as it is seen in Przewalski's horse, a species of horse that diverged from the modern horse between 13,300 and 11,400 years ago (Kvist & Niskanen, 2021), and other close equid relatives (Imsland et al., 2016). Imsland recorded two non-dun alleles, non-dun1 and non-dun 2, where non-dun1 and dun are both ancient mutations occurring pre domestication. A founder effect or other evolutionary bottleneck can also be seen here as the Przewalski's horse dun allele has more nucleotide diversity as compared to modern dun which has as little diversity as non-dun (Imsland et al., 2016).

Fjord Colour Genetics

In Fjord horses coat colours are traditionally limited to variations of dun, with primitive markings. There are 5 recognised coat colours including brown dun (brunblakk), red dun (rødblakk), grey dun (grå), white dun (ulsblakk) and yellow dun (gulblakk). Brown dun is the most common colour, with 85-90% having this colour and currently yellow dun is the least common, with 0.5% of the population registered as this colour (Nestaas, 2014). Primitive markings are small brown markings over the eyes, cheeks and thighs, zebra stripes on the legs, and dark stripes over the withers. A dark dorsal stripe as well as dark sections in the forelock, mane and tail is also typical. Some Fjords may have small white markings on the legs or a small star on the face, however excessive white markings is undesirable by studbook standards (NFHR, n.d.) and are not permitted in breeding males (Nestaas, 2002). It can often be difficult to determine the coat colours in Fjord horses accurately without genetic tests to ascertain the underlying genotype.

Genetic testing for coat colour has been in practice for many years and is often used to determine ambiguous phenotypes. All Fjord horses carry the dominant dun gene, therefore as seen in Table 1

below, variation in colour comes from differences in base coat colour as well as the presence of different dilution factors. In the Fjord horse, genetic tests can be used to determine whether animals carry genes for creme coat colours. This knowledge can then be used in breeding to prevent the birth of double dilute fjords, called white (Kvit), resulting from breeding two fjords carrying at least one creme allele. These horses are nearly white with one or both eyes being blue eyes. This phenotype is highly undesirable in the Fjord horse studbook and foals with this appearance will not be accepted for breeding (Nestaas, 2002). Anecdotally, these white horses are to be avoided as their eyes and skin do not tolerate sunlight well.

Table 1: Genetics of Fjord coat colours. All Fjords are homozygous for the Dun gene, with Extension (E or e), Agouti (A or a) and the cream dilution (C or Cr) dictating other variations. A single creme allele will not alter the grey dun phenotype and so can be present 'silently'. A double dilute (Cr/Cr), white, will never be accepted in the studbook. The bottom row indicates the coat colour that would be seen without the presence of the dun gene.

Gene	**Brown dun**	**White dun**	**Grey dun**	**Red dun**	**Yellow dun**	**White**
Extension	E/E or E/e	E/E or E/e	E/E or E/e	e/e	e/e	E/E or E/e or e/e
Agouti	A/A or A/a	A/A or A/a	a/a	A/A or A/a or a/a	A/A or A/a or a/a	A/A or A/a or a/a
Cream	C/C	Cr/C	Cr/C or C/C	C/C	Cr/C	Cr/Cr
Colour without Dun gene	Bay	Buckskin	Black	Chestnut	Palomino	Perlino or Cremello

Genetics of White Markings

In comparison to the mendelian inheritance patterns observed for coat colours, white markings and spotting is considered to be a complex trait with large phenotypic variance (Hauswirth et al., 2012). Initially thought to be inherited in an autosomal recessive mode (Rieder, 2009). Nowadays, many dominant white alleles that result in a completely white phenotype are monogenic autosomal dominant traits (Haase et al., 2009), while white markings do not often show a monogenic mode of inheritance (Hauswirth et al., 2012). White markings can cover a wide range of phenotypes, from a small white mark, or unpigmented mark, on the forehead of the animals to almost the entire face being unpigmented. White markings occur as a result of changes to how melanoblasts move across the body from the neural crest and differentiate into melanocytes (Sponenberg 2009 as reviewed in Rieder 2009, Hauswirth et al., 2012). A variety of different genes are previously known to relate to different white spotting phenotypes. These include Proto-Oncogene Receptor Tyrosine Kinase (KIT),

Paired Box 3 (PAX3), Melanocyte Inducing Transcription Factor (MITF), Endothelian Receptor Type B (EDNRB), and Transient Receptor Potential Cation Channel, Subfamily M, Member 1 (TRPM1).

Many of these genes are pleiotropic in their effect and cause not only white spotting but are connected to different syndromes or illnesses. EDNRB and TRPM1 cause white spotting over the entire body. EDNRB is associated with the frame overo phenotype when present in a heterozygous state, however, the homozygous version causes overo lethal white syndrome. Foals with this syndrome are born all white and die within days due to complications resulting from a lack of nerve cells regulating colon function.(Santschi et al., 1998). In contrast to this, TRPM-1 is associated not only with leopard complex spotting, but also congenital stationary night blindness, as seen when horses are homozygous for the leopard complex spotting gene (Bellone et al., 2013; Sandmeyer et al., 2012). MITF has been shown to cause white markings on the face in several different populations of unrelated animals, and when found in combination with PAX3 in Quarter Horses several were found to be deaf (Hauswirth et al., 2012, 2013). Hauswirth (2012) also suggested that MITF mutations causing white markings have existed for hundreds years as it can also be found in both Thoroughbreds and Icelandics, therefore the mutation occurred before distinct breeds were developed.

Additionally, in the French-Montagne breed it has been shown that animals with white markings were at higher risk of sunburn, poorer quality hoof horn as well as pastern dermatitis (Federici et al., 2015). Furthermore, much research has also gone into the cause for the higher rate of melanomas recorded in grey horses, regardless of breed, when compared to the general horse population. This increased incidence of melanoma has been linked to mutations in syntaxin-17 (STX17) and a loss of function mutation in the agouti signalling protein (ASIP) causing a higher incidence of melanomas (Rosengren Pielberg et al., 2008).

Finally, KIT is the most prolific gene when it comes to white markings. Over 30 polymorphisms have been associated with a range of white spotting phenotypes. As recently as 2021, two new mutations were categorised that were associated with white spotting in the American Paint horse and Quarter horse (Patterson Rosa et al., 2021). The KIT gene has been previously recognised as causing a white coat colour in pigs, mice and humans with several mutations causing a similar dominant white phenotype in horses. The KIT receptor is crucial for the development of haematopoietic, gonadal and pigment stem cells, in addition to acting as an essential survival factor for migrating and proliferating melanoblasts, thereby explaining the extensive pleiotropic effects observed as a result of KIT mutations (Haase et al., 2009).

White markings have a wide range of phenotypes, and several different genes are known to be associated with them. It is a unique phenotype due to its extensive connections with other diseases and other disorders. Thereby providing a rich base of research for this investigation in Fjord horses.

Potential Genes

An association analysis was done using genotyped Fjord horses and their phenotypes for coat colour and markings (Høiseth, 2017). This revealed an association between white markings and genetic variants on chromosome 23, however, chromosome 23 does not contain any of the known genes associated with white markings that were mentioned above.

One gene on chromosome 23 with a potential connection to the observed phenotype is tyrosinase related protein 1 (TYRP1), a gene shown to affect brown coat colour in mice (Kobayashi et al., 1998). TYRP1 has also been found to be expressed in lower quantities in the skin of grey horses, as compared to horses with a darker coat colour (Rieder et al., 2000). Another potential gene is MLANA (melan-a), this is involved in melanosome biogenesis, that was shown to be part of a transcriptional pathway regulated by MITF (Du et al., 2003), a gene highly associated with white spotting phenotypes as discussed above. However, neither of these genes have previously been linked to white markings.

Aim Statement

This study seeks to explore the association between white markings in fjord horses and variation on chromosome 23. This will be done by performing a genome wide association study (GWAS) which looks for associations between the phenotype, white markings, and 67K single nucleotide polymorphisms (SNPs) recorded in 328 horses. Subsequently, short-read re-sequencing data from 16 Fjord horses (8 cases and 8 controls) will be used to fine map any QTLs.

Materials and Methods

Genotyping Data

Data available for this project includes genotype data describing 328 Scandinavian (predominantly Norwegian and Swedish) Fjord horses collected for a previous research project at NMBU. All animals have a known pedigree, with the data set including both full and half sib relationships from 187 sires and 279 dams. Phenotype data is available detailing coat colour for each animal as well as the presence or absence of white markings on the body or a white star marking on the face. The phenotype used for association in this study is white markings located anywhere on the body, this includes markings on both the head and legs. Phenotype data related to markings was retrieved manually as it is not automatically recorded by the studbook (Høiseth, 2017). Of the 328 animals recorded in the dataset, 19 have white markings and/or a white star. Genotypes were generated using data that was collected on the animals through genotyping with the Affymetrix MNEc670k SNP-chip. From this collection of Fjord horses 16 were selected for whole genome sequencing.

Sequencing

Genomic DNA was isolated from blood collected from 16 of the genotyped individuals representing 8 "cases", with white markings, and 8 "controls", without white markings. The specific individuals were selected to not be closely related. Extraction was done with Qiagen Blood-and-tissue kit (Qiagen, Germany). DNA integrity was assessed using agarose gel-electrophoresis, purity was assessed with spectrophotometric measurements (nanodrop) and quantity measured using fluorescence (Qubit). DNA was prepared for sequencing and sequenced by a commercial provider (Novogene UK) using NEBNext UltraII DNA library prep kit and a Novoseq6000 using an S4 flowcell, with a request for 80Gb (approx 25x coverage) raw read data (PE150).

GWAS

To begin with, a GWAS for white markings was carried out, this was done using the programme genome-wide complex trait analysis (GCTA) (Yang et al., 2011). Utilising the option mlma-loco will carry out a mixed linear model-based association analysis (mlma) while excluding the chromosome where the candidate SNP is located from calculating the genetic relationship matrix, referred to as leaving one chromosome out (loco). GCTA was developed to address the 'missing heritability' problem and works by estimating the variance explained by all the SNPs on a chromosome or whole genome for a complex trait (Yang et al., 2011). The effects of all the SNPs, except on the chromosome of interest, are fitted as random effects by a mlm. This is done alternatively to the usual method of testing the association of any SNP with the trait of interest. The GWAS done here used GCTA in contrast to the GWAS previously carried out which used the software GEMMA (Zhou & Stephens, 2012), which

uses a similar mlm method, only without the loco aspect. These results were then plotted using R version 4.0.5.

Pre-processing

The sequence reads underwent quality control looking at factors such as read length, GC content and quality score before assembly. FastQC version 0.11.9 was used for this in addition to MultiQC, (Ewels et al., 2016) to summarise across FastQC reports, as well as across alignment and duplication metrics resulting from the following steps.

Reads were aligned to the most recent equine reference genome, EquCab 3.0 (Kalbfleisch et al., 2018) GCA_002863925.1, using the Burrows-Wheeler aligner (BWA) version 0.7.17-GCC-9.3.0, (Li & Durbin, 2009). Following alignment, reads were sorted and converted to BAM files using both Picard tools version 2.26.10 (*Picard Tools - By Broad Institute*, 2022) and SAMtools version 1.11, (Li et al., 2009) updated in (Danecek et al., 2021). The next step included marking of PCR duplicates with Picard tools (*Picard Tools - By Broad Institute*, 2022) before indexing the sequences and references as well as creating a sequence dictionary for the reference. Following this preparation, the files underwent base quality score recalibration (BQSR) using the GATK (McKenna et al., 2010). This was the final pre-processing step before variant calling.

Variant Calling

Following base quality score recalibration, the sequences underwent variant calling. Variant calling was done using two callers, GATK HaplotypeCaller (McKenna et al., 2010) and Freebayes version 0.9.21 (Garrison & Marth, 2012). GATK HaplotypeCaller (HC) was run in GVCF mode, which calls haplotypes per sample to create an intermediate file, genomic variant call format (GVCF). Then the GVCF files were consolidated using GATK GenomicsDBImport before joint genotyping using GATK's GenotypeGCF. This results in a set of variants which are separated out into SNPs and Indels before undergoing filtration and being recombined in one file to form a set of analysis ready variants. Freebayes variant calling was run using default settings following base quality score recalibration and variants with a quality score over 40 were used in for further analysis. GATK variant calling was done over chromosomes 3,6,14,16 and 23 in the interest of time, while Freebayes was run over the entire genome. Once the two different variant callers had completed the variant calling, the intersect of the two files was taken as recommended in Field et al., 2015 to increase the specificity of the variant calling, resulting in a final filtered genome-wide variant call format (VCF) file.

A single sample also underwent the steps from pre-processing to variant calling as outlined above a second time. However, in this instance the reference was a repeat masked version of EquCab 3.0 (Kalbfleisch et al., 2018).

Table 2: This table shows the regions of interest subjected to further analysis and their reason for inclusion

Chromosome	Region	Reason for Inclusion
3	79,500,000-79,700,000bp	Location of KIT gene
6	11,100,000-11,250,000bp	Location of PAX3 gene
16	21,500,000-21,760,000bp	Location of MITF gene
23	25,000,000-32,000,000bp	Peak 1 in GWAS
23	47,000,000-52,000,000bp	Peak 2 in GWAS

A few specific areas of interest (see Table 2) were subjected to further analysis because of their significance following GWAS. For these regions, allele frequencies, as well as a Weir and Cockerham's Fst, was carried out between the cases and controls. Allele frequency is the rate at which a particular allele occurs at a locus divided by the total number of alleles in the population. While Weir and Cockerham's Fst is a measure of genomic diversity within and between populations (Holsinger & Weir, 2009; Weir & Cockerham, 1984). Weir and Cockerham's Fst is calculated using a method of moments estimate which essentially results in an ANOVA of allele frequencies within and between subpopulations being carried out (Holsinger & Weir, 2009).

Plotting of these results was done using R version 4.1.1 (R Core Team, 105 C.E.) and Rstudio version 1.4.1717. Variant annotation was carried out using snpEff (Cingolani et al., 2012), with SnpSift used to filter once annotation had been carried out. Variants were filtered utilising SnpSift's Case Control utility, which carries out Fisher exact tests based on genotypes between case and controls while accounting for different modes of inheritance. A codominant/genotypic model was used for this data, which creates a contingency table for cases and controls and the 3 possible genotypes at each variant location.

In carrying out the variant calling process as described above, direction was received from three main sources. Firstly the GATK best practices as outlined in Van der Auwera (2013), and regularly updated on the GATK website, as well as tool specific descriptions. Secondly an overview of the entire workflow and extensive explanations was provided as a tutorial in a blog post written for the Genomics Core at NYU (Khalfan, 2020). However, as the specific tutorial from NYU was made in relation to specifics for their dataset, when our data deviated, guidance and tool specific arguments to use as well as help on typical GATK pitfalls was found on the web version of OVarFlow (Bathke & Lühken, 2021). OVarFlow is a nextflow package comprising of the complete variant calling and annotation pipeline.

Results

GWAS

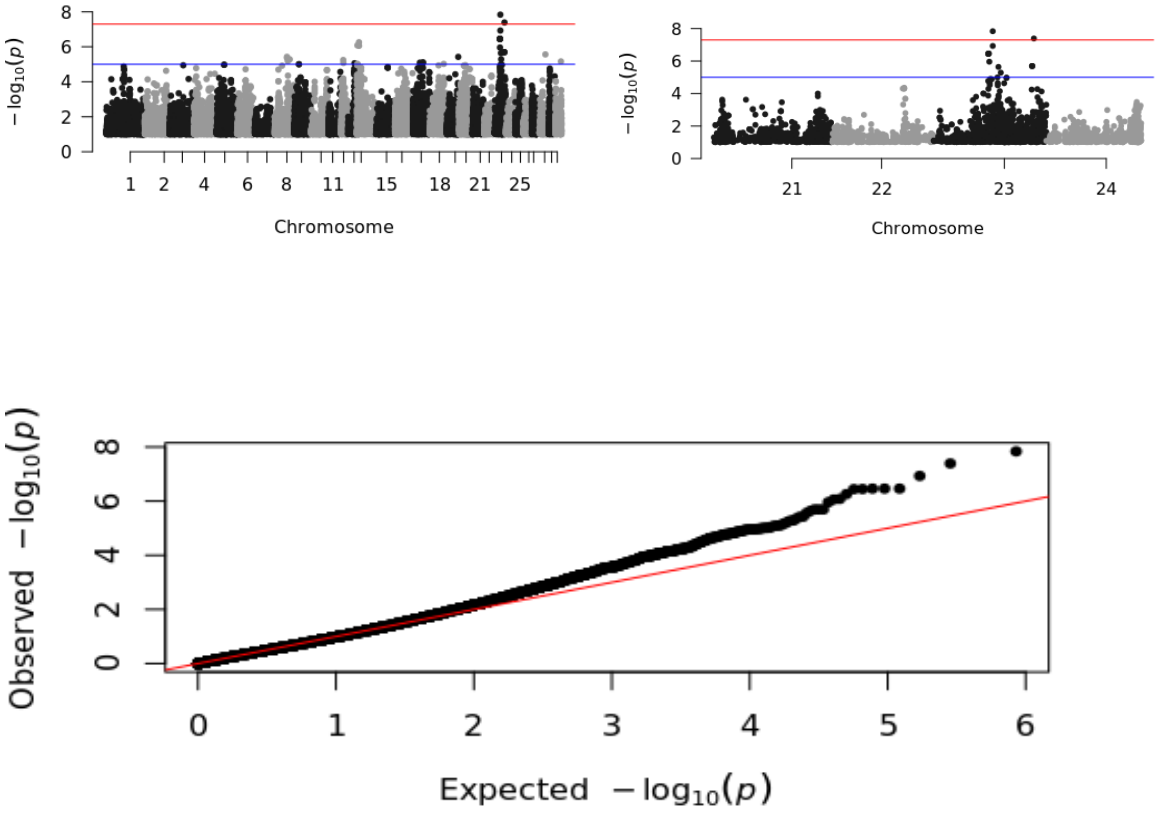


Figure 1: Clockwise from the top left: A. Manhattan plot of GWAS data done using GCTA where the red line is the Bonferroni correction, and the blue is a standard p value of 0.05. A peak is seen here at chromosome 23. B. Is a close-up view of the previous plot to highlight the two peaks evident on chromosome 23. C. A qqplot of expected vs observed log transformed p values for the same data

Here you can see the results of the GWAS done with GCTA. The GWAS was done with all 328 Fjord horses and included a pedigree so that GCTA could account for relationships between the animals used in this analysis. While a 670k SNPchip was used, 431,950 SNPs were used for the association analysis following GCTA's own QC. As seen in

Figure 1A, a peak was found on chromosome 23, on closer inspection as shown in

Figure 1B, this peak actually contains two peaks on chromosome 23 associated with white markings. These peaks are considered significant as they cross the red line which marks the Bonferroni correction, a significance threshold that accounts for the multiple tests that are being carried out in a GWAS. The blue line denotes a significance level of p= 0.005 and as can be seen here is less stringent,

with several other peaks crossing this level of significance. The top two most significant SNPs as seen in Figure 1B, were located at 29,746,436 bp and 49,678,502 bp.

Finally, in

Figure 1C a qqplot of expected versus observed log transformed p values can be seen. Here the red line denotes what would be seen if the expected and observed values correlated perfectly. While the black line shows what was actually observed from this data. Here the black line begins to drift from the red one shortly after 2 on the x-axis, and gradually increases in distance from the red line.

Table 3: The top 8 most associated SNPs and their significance level as identified during the GWAS

Position	29,746,436	49,678,502	29,755,166	27,661,913	27,760,289	27,897,074	27,628,802	27,544,921
P value	1.36E-08	3.84E-08	1.11E-07	3.27E-07	3.27E-07	3.27E-07	3.39E-07	3.42E-07

Additionally, in Table 3, the location of the top 8 SNPs as indicated following the GWAS, as well as their associate p-values are shown. These 8 SNPs are all found on chromosome 23. While only the top 2 SNPs are above the threshold set by the Bonferroni correction, all other SNPs are still considered to have some significance, as they have a p value above 0.005.

Re-sequencing data and pre-processing

These results look at whole genome sequencing data on 16 of the Fjord horses included in the GWAS done above. These 16 animals included 8 cases and 8 controls. Pre-processing is a vital step to ensure that variants called at later stages, and any analysis done on these variants is as accurate as possible.

Raw file QC

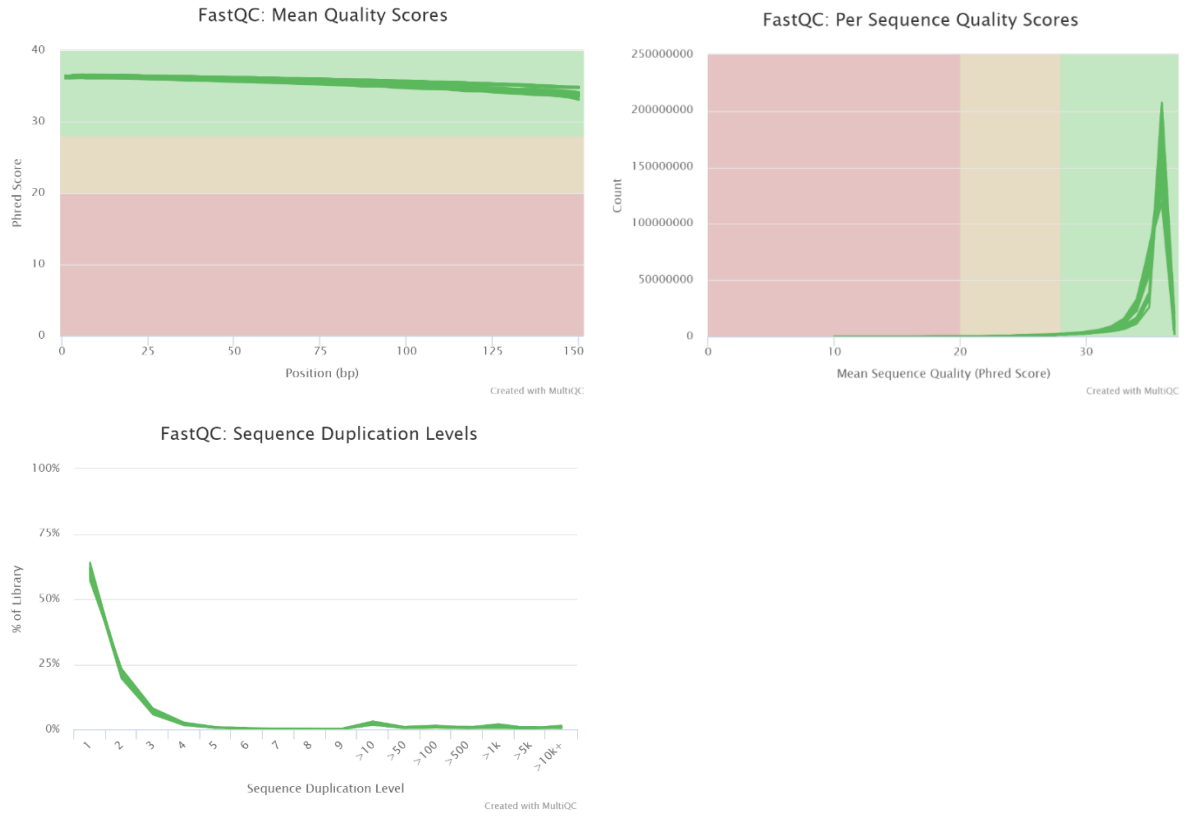


Figure 2: A. Indicates mean quality scores are of sufficient quality. B. Looks at per sequence quality scores which are all to a good standard. C. Highlights the level of duplication present in the samples.

FastQC creates a comprehensive set of statistics with which to judge the quality of the raw sequences files that have been received. MultiQC compiled graphs based on these statistics, 3 of which are shown here. Figure 2 A and B both look at sequence quality in Phred score. Figure 2A looks at mean quality scores over the entire sequence length while B looks at overall means of the quality score. Both graphs indicate that the sequences were of very high quality with quality scores remaining above 30 for the entire sequence as seen in Fig2A. Additionally, all samples remain in a tight band in both plots, indicating that all samples had relatively similar quality scores. The colour of the lines in all graphs indicate whether that sequence passed that specific quality check., with green denoting passed, orange showed a warning and red lines failed that QC check. Figure 2C showed the level of duplication in the sequences, where roughly 25% of the sequences are duplicates. In this graph the x axis represents the number of times a sequence occurs, the majority occur only once, although some sequences repeat more than 10, 100 or 1000 times. Overall, the data was of good quality with no reads failing quality control and so all were used for mapping in the next step.

Alignment

Table 4: This table details some metrics for assessing the alignment

Sample	Total Reads	% Mapped	% Paired	% Proper Paired	Mate & Self mapped
FH9019	5,670,669	99.78	99.72	97.48	99.41
FH9022	5,549,161	99.75	99.72	97.45	99.35
FH9071	5,939,938	99.83	99.73	97.63	99.48
FH9082	5,441,704	99.81	99.68	97.7	99.41
FH9097	6,336,438	99.8	99.71	97.73	99.44
FH9119	5,954,046	99.83	99.71	97.94	99.45
FH9125	5,625,782	99.79	99.71	97.53	99.37
FH9143	5,467,072	99.7	99.69	97.66	99.24
FH9149	5,464,665	99.83	99.75	98.06	99.51
FH9177	5,605,597	99.82	99.7	97.88	99.44
FH9185	5,493,766	99.82	99.74	97.98	99.47
FH9196	5,692,307	99.82	99.74	97.91	99.47
FH9199	5,548,190	99.81	99.76	98.05	99.49
FH9207	5,409,647	99.77	99.74	98.06	99.41
FH9274	5,476,586	99.77	99.75	98	99.44
FH9324	5,954,052	99.77	99.64	96.87	99.31

Table 4 looks at some basic alignment metrics gathered from samtools Flagstat. Total reads indicates the reads that passed quality control and were used in the alignment. As visible from Table 4, only a very small portion of the reads were not mapped, with similarly high percentages paired, properly paired and recorded with both mate and itself mapped. Properly paired when using the BWA aligner indicates that a read and its mate were paired correctly on the same chromosome. Using the unmapped genome resulted in a very high percent of the reads being mapped, in general however, the samples were of high quality and they aligned well.

Duplication Statistics

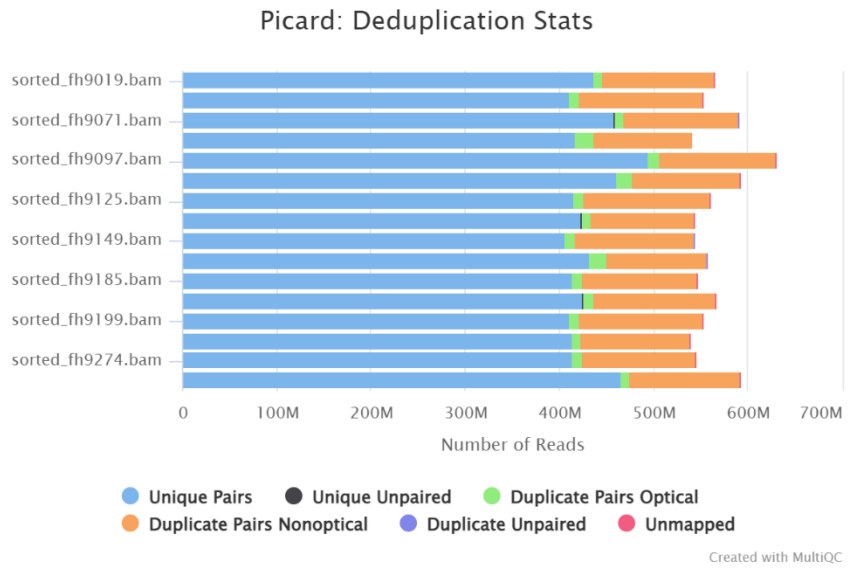


Figure 3: Statistics examining the rate of different kinds of duplicates in the reads

Figure 3 shows the duplicate reads in the samples. While not every sample is represented here, this gives an impression of the duplicate rates. Deduplication is done to remove duplicates that occur as part of the PCR process. Optical duplicates occur when a single amplification cluster is incorrectly called as multiple clusters during sequencing, while non-optical duplicates can occur during the amplification step during library prep. Here the percentage of duplicates seen comprised between 21.3% and 25.6% of the total reads.

Base Quality Score Recalibration

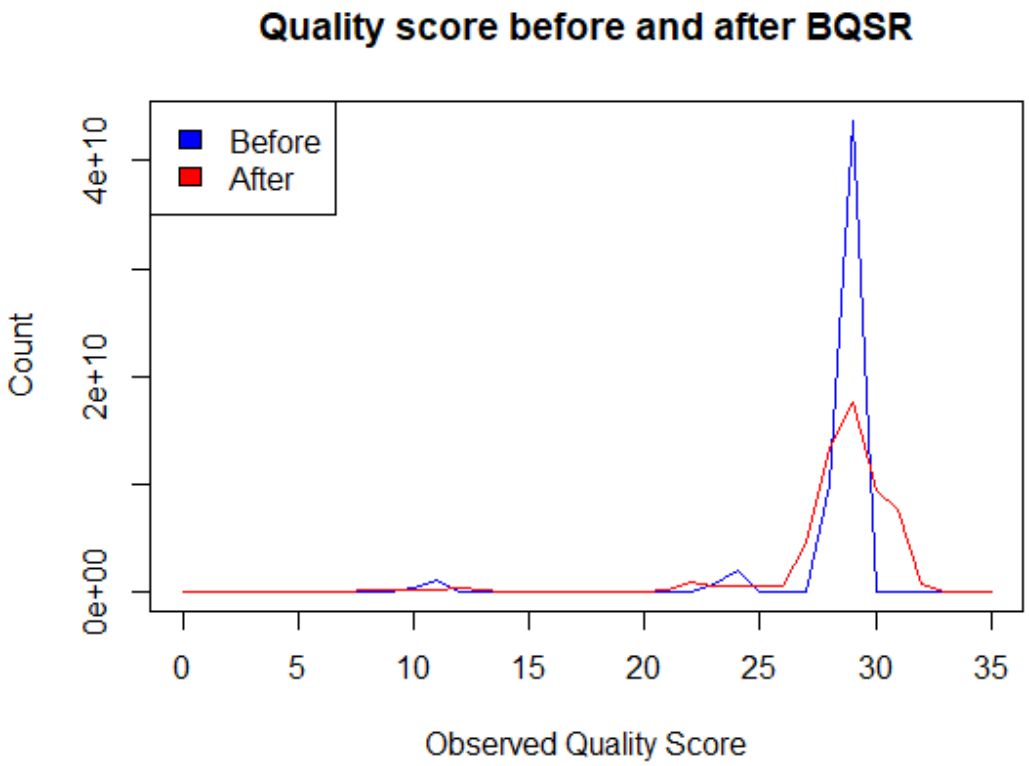


Figure 4: Comparison of the observed quality scores before and after processing with BQSR.

Base quality score recalibration is done as sequencing machines often make systematic errors while calling bases. GATK's BQSR uses a machine learning model to learn how to empirically adjust for errors made in determining the quality score. This will result in some scores remaining the same, some increasing and some deceasing. It results in a greater number of quality scores as seen above and so the graph has become more spread out as some of the quality scores are adjusted higher than the threshold of 30 seen in Figure 4 above.

Close Examination of Candidate Genes

In this section of the results, a closer examination is carried out on the genes and regions of interest as mentioned in Table 2. This includes three genes associated with white markings not located on chromosome 23, KIT, MITF and PAX3 to assess the involvement of these genes. In addition, two regions on chromosome 23 highlighted in the GWAS are also subjected to further analysis.

KIT

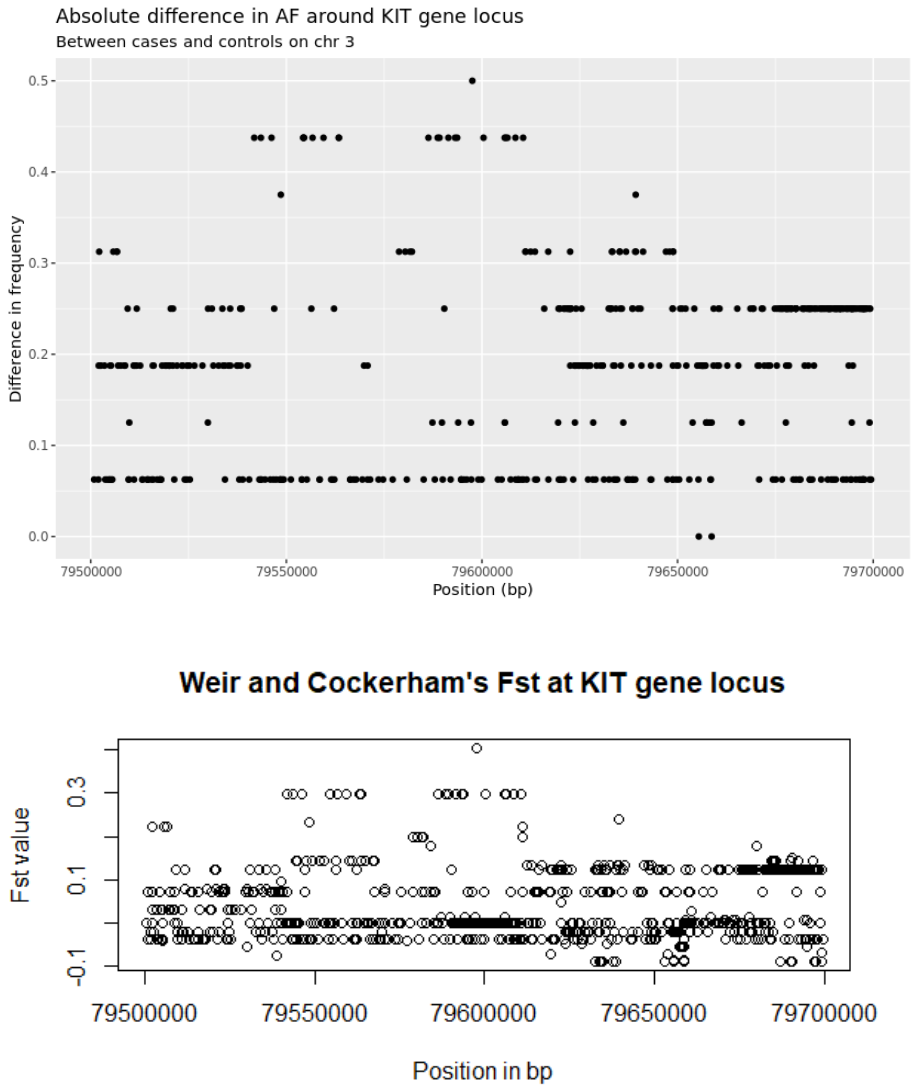


Figure 5:A The top is a scatterplot showing the difference in allele frequency between cases and controls in the region of KIT. B. The bottom plot, plots Weir and Cockerhams Fst over the same region.

KIT is associated with many white spotting phenotypes (Haase et al., 2009; Hauswirth et al., 2012, 2013). ENSEMBL lists the gene as spanning the region from 79,504,108 – 79,618,886 bp on chromosome 3. As seen in Figure 5A, all allele frequencies above 0.4 directly overlap with this region.

Similarly, in Figure 5B, Weir and Cockerham's F statistic also shows a similar trend in the difference between cases and controls over this region.

PAX3

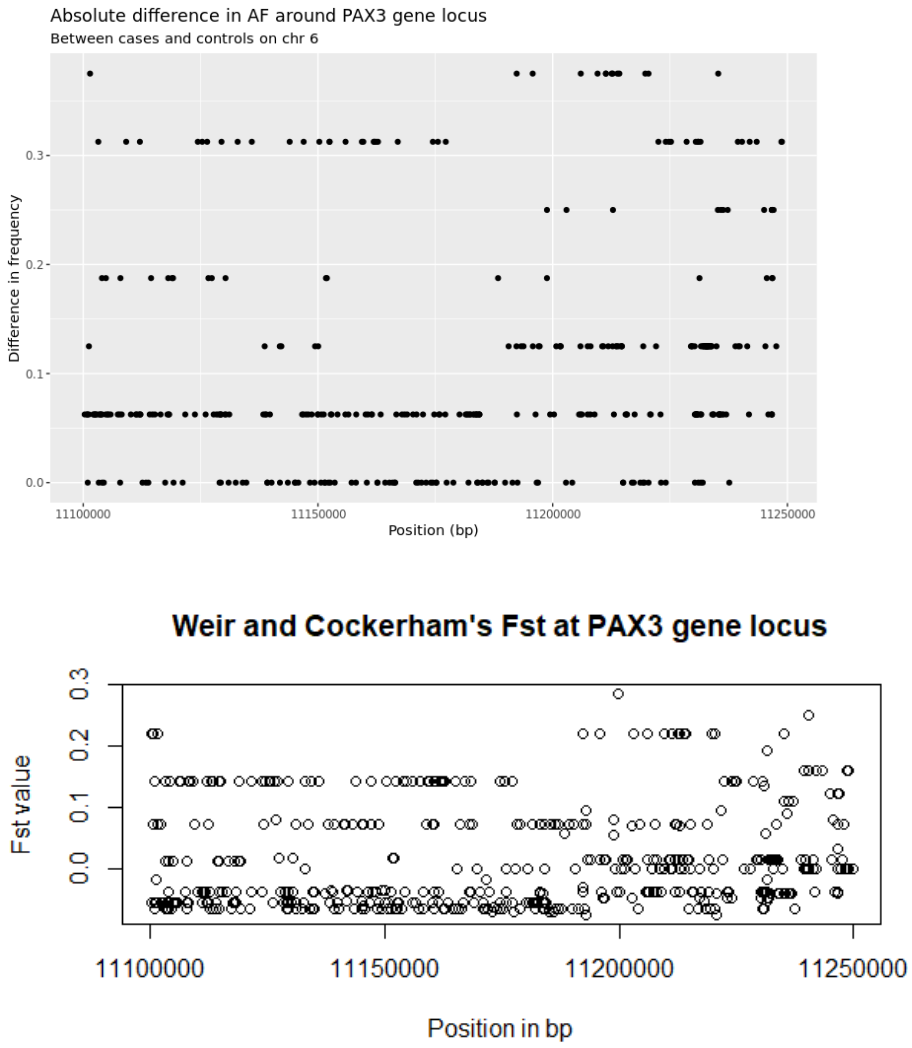


Figure 3: A. Shows scatterplot indicating the difference in allele frequency between cases and controls in the region of PAX3. In addition, B. shows a scatterplot of Weir and Cockerhams Fst for the same region.

PAX3 is associated with the splashed white phenotype and leopard complex spotting patterns (Hauswirth et al., 2012, 2013). ENSEMBL lists the gene as spanning the region 11,109,734-11,200,548 bp on chromosome 6. This graph has the difference in allele frequencies between the cases and controls on the y axis and the position on the chromosome on the x axis. As seen here, there is some difference between allele frequencies in this region. However, the difference is not very great and is spread over the entire region plotted, with no obvious segregation where PAX3 is located.

Interestingly the Weir's Fst has a maximum value similar to that seen for the KIT gene locus, although the differences in allele frequency are more striking for KIT.

MITF

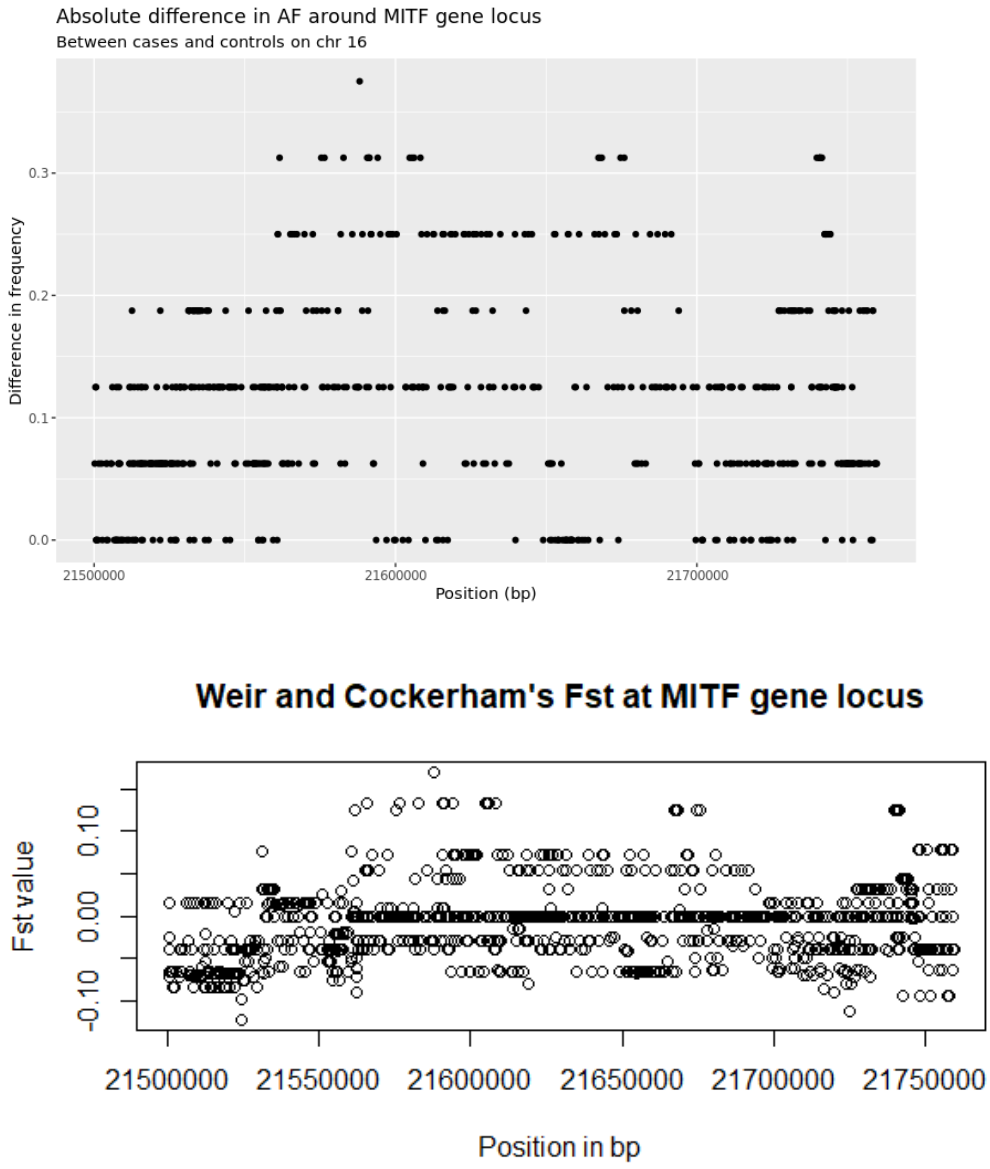


Figure 7: A. Shows a scatterplot of the difference in allele frequency between cases and controls in the region of MITF, while B. indicates Weir's Fst over this same region

Figure 7 shows the differences in allele frequencies around the region of the MITF gene. ENSEMBL indicates the position of the gene to fall in the region of 215,480,000 – 21,757,591 bp on chromosome 16. Again, areas with highest allele frequencies differences coincide with the gene region. However,

Weir's Fst shows very little difference between cases and controls with a max value of 0.17. This max value coincides with the location of the SNP with greatest allele differences.

Peak 1

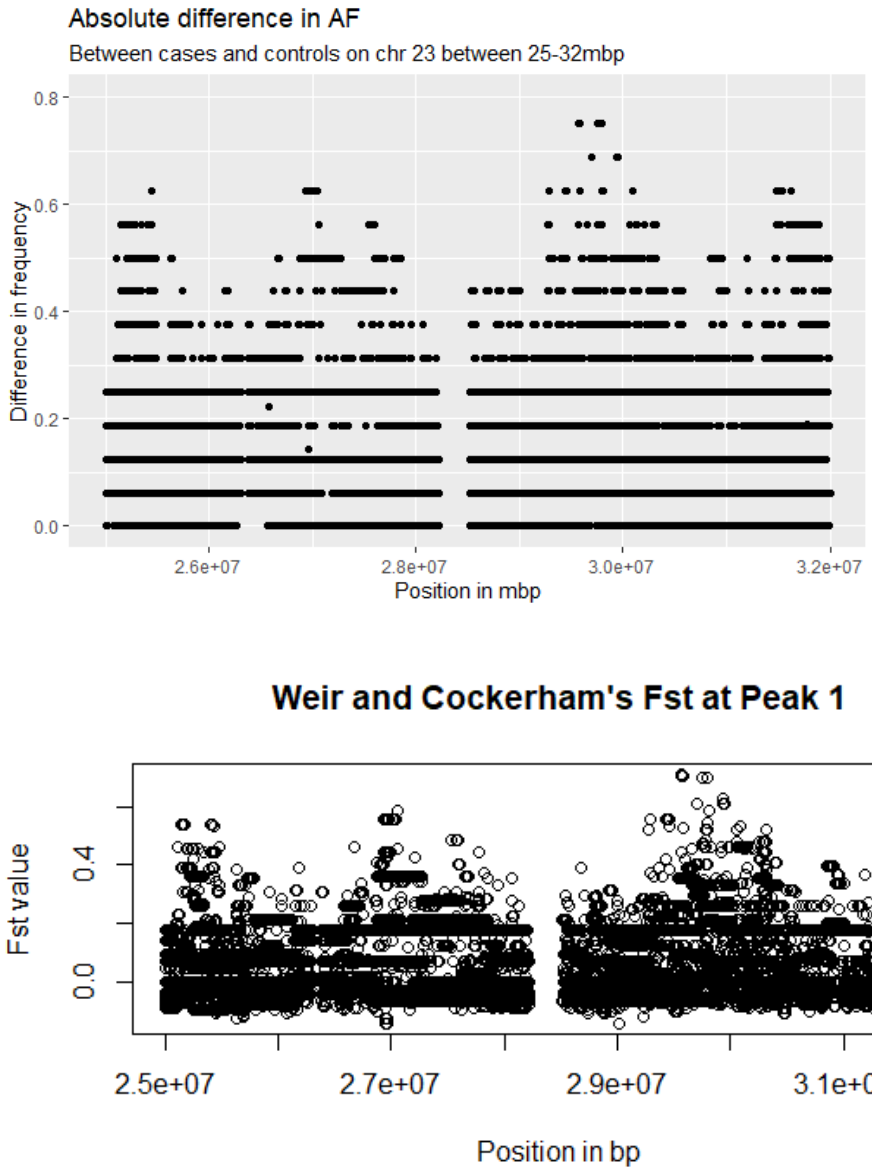


Figure 8: A Shows a scatterplot highlighting the difference in allele frequency between cases and controls in the region of 25-32m bp on chr 23. B. Shows Weir's Fst over this same region of the chromosome.

These graphs examine the region surrounding the first peak on chr 23 picked up by the GWAS. In Figure 8 the greatest difference in allele frequencies is observed in accordance with the location of greatest association in the GWAS. There were 26 SNPs called which had a difference in allele frequency of 0.75. They can be clustered into approximately 2 peaks as seen in Figure 8A above. The first includes 9 SNPs that are found from 29,566,014 - 29,584,198 bp and the second region includes 17 SNPs going from 29,749,364 – 29,797,765 bp. Interestingly the SNPs in these regions do not directly correspond

to the most significant SNPs from the GWAS as seen in Table 3, with the most significant SNP located just before this region and the third most significant SNP located within the second region.

The gap in allele frequencies, is a gap of approximately 200,000bp where no variants were called. This is a highly repetitive region, so SNPs called here were likely discarded during filtering as mapping quality would have been too low to pass quality control.

Then Figure 8B. shows Weirs Fst over the same region on chromosome 23. The max values peak in accordance with Figure 8A, with a max value of 0.71 recorded at 29,566,046bp.

Peak 2

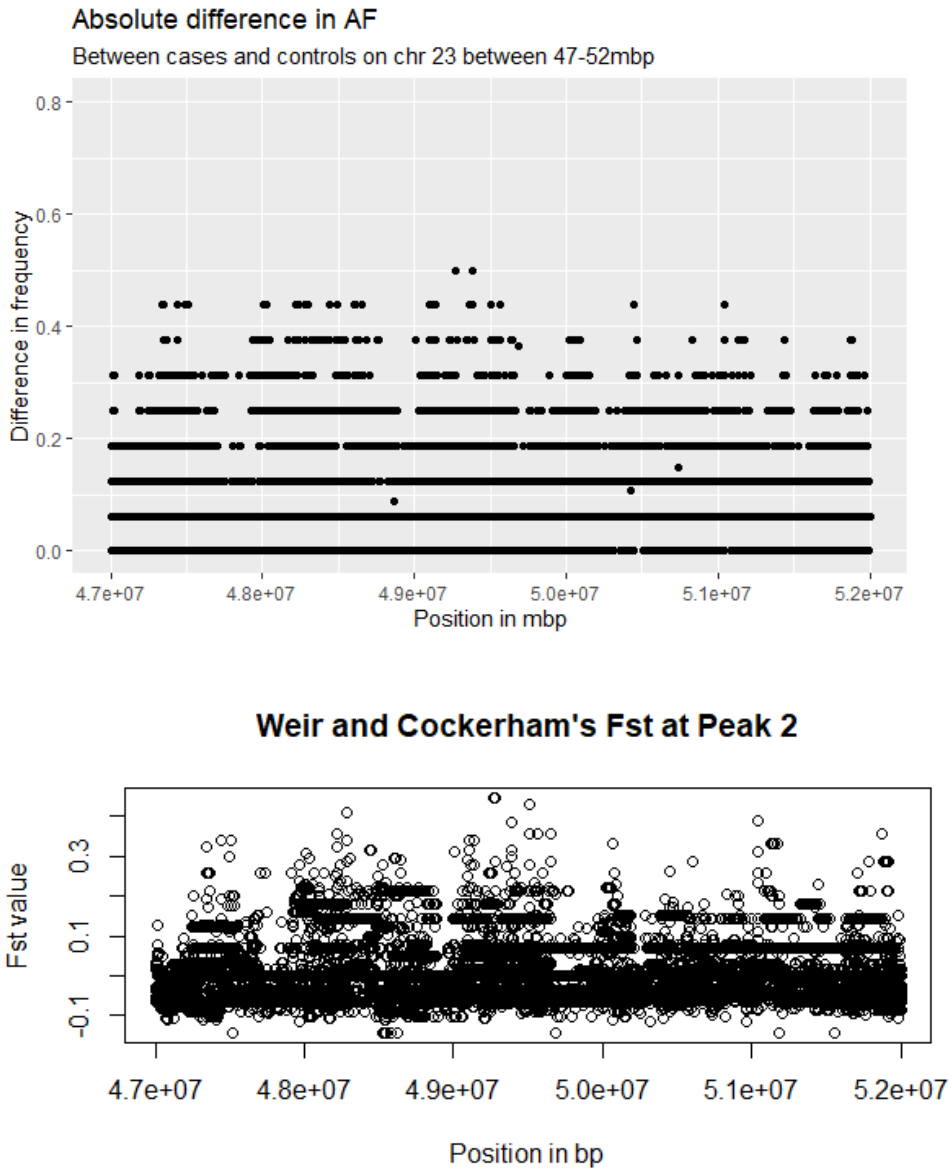


Figure 9: A scatterplot showing the difference in allele frequency between cases and controls in the region of 47-52m bp on chr 23. B. Shows Weir's Fst over the same region on chr 23.

This is the final area undergoing close examination. Similarly, to Figure 8, this plot is the difference in allele frequencies around the region of the second most associated SNP from the GWAS. The greatest difference in allele frequency is in the approximate location of the second most associated SNP in the GWAS, however, again this SNP was not called during variant calling. Figure 9B again peaks in accordance with Figure 9A, with a max value recorded here of 0.45.

Variants Called

Table 5: DA= results from GATK Haplptype Caller from Durward-Akhurst et al., 2021, Chr23= Results from GATK's Haplotype caller (GATK) and Freebayes (FB) for chr 23, as well as for FB using a repeat masked genome (RM), Ts/Tv = Transition/transversion, MA= multialle

Source	Ts/Tv ratio	SNPs	Indels	MA sites	MA SNP sites
DA	1.87	38,205,667	4,694,627	2,974,935	2,127,391
Chr 23-GATK	1.97	313,984	37,660	8,928	902
Chr 23- FB	1.58	334,088 + 14,803 MNPs	29,368	12,096	1,921
Chr 23- RM	1.96	71,595 + 4,259 MNPs	3,770	572	55

The results for DA are based off whole genome sequencing and so the raw values are not directly comparable to what we called. GATK's HaplotypeCaller (HC) and the calling done with a repeat masked genome had a comparable TS/TV ratio as that found by Durward-Akhurst et al., 2021. There were no multiple nucleotide polymorphisms (MNPs) called as GenomicsDBimport does not support MNPs. A single sample was used to carry out variant calling for the repeat masked genome as compared to the other instances where all 16 samples were used. While Freebayes called more SNPs and multi-allelic sites, GATK called more indels.

SnpEff annotated a total of 74,670 variants found in the regions of interest on chromosomes 3,6,16 and 23 as shown in Table 5. Of these 1,223 were multi-allelic variants and a total of 149,685 effects were annotated. SnpEff categorised 99.14% of the impact of these effects as modifiers, 0.3003% had a moderate impact, 0.542% a low impact and just 0.01% had a high impact. The most common type of effect was a silent mutation, 60.61%, then missense, 39.22% and finally nonsense, 0.173%.

Using the case control function of snpSift, variants were filtered using CC_Geno, which looks at differences between the cases and controls across all 3 genotypes, homozygous for the reference, homozygous for the alternate allele or heterozygous. Using a stringent p value of 0.005 to filter the variants resulted in 44 variants remaining. From these 44 variants, 9 are variants that correspond to SNPs with the greatest differences in allele frequencies, and highest weir's Fst on chromosome 23.

Repeat Masked Genome

Image 1: Screengrab from IGV showing an area at Chr23: 28,217,000-28,225,000bp. Both tracks are from the same sample just aligned to an unmasked or a repeat masked genome.

Image 1 shows two tracks from the same sample, fh9149. The reference genome used here in IGV is a repeat masked genome. The top track was aligned to an unmasked genome, while the bottom track was aligned to a repeat masked genome. Colours seen at the top where the sequence depth is shown, indicate deviations from the genome used in IGV. As seen at the top there is far more colour seen, this is due to the unmasked genome calling bases at places that are repeat masked in the reference. This image is taken from the beginning of the gap seen in Figure X. The boxes all represent different reads sequenced to this location, with the unshaded boxes being ones with a very low mapping quality. In this case they all have mapping qualities of 0. This is how BWA indicates bases that have been assigned to at least two locations with equal probabilities (Yu et al., 2012), such as when trying to align reads to repetitive regions.

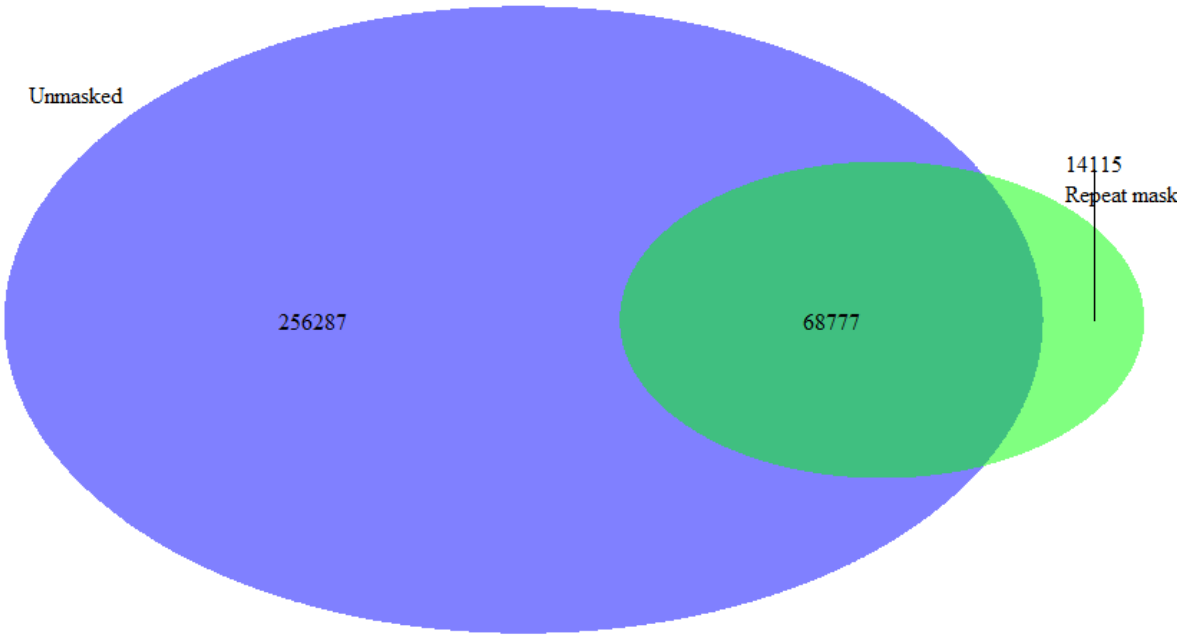


Figure 10: Venn diagram of the variants called on chr 23 using an unmasked versus repeat masked reference genomes

Figure 10 shows a venn diagram to give a basic comparison between the number of variants called on chromosome 23 when using a repeat masked reference genome versus an unmasked reference genome. This is only a rough estimate as the variant calling done with the unmasked genome had the added benefit of joint variant calling, to result in far more variants called. While the variant calling using the repeat masked genome was only carried out using one fjord horse, so variants seen will only be those present in this horse.

Discussion

GWAS

GCTA estimates genetic variance explained by all SNPs on a chromosome or whole genome for a complex trait in comparison to other methods which test the association for any particular SNP to the trait. In this way GCTA addresses the problem of missing or hiding heritability resulting from many SNPs with small effects (Yang et al., 2011). As seen in Figure X this GWAS once again found two peaks on chromosome 23 to be associated with white markings, previously recorded by Høiseth (2017). This confirms that these regions on chromosome 23 play a role in white markings for the fjord horse.

In Figure 1C, the qqplot shows the black line deviating from the red baseline relatively early. This could be indicative that this finding is a spurious association. Spurious associations can occur due to underlying family structure or other causes of population stratification, such as non-random mating that has not been fully accounted for by the program. However, the most obvious source of population stratification, family structure, is already accounted for in GCTA through the use of a genomic relationship matrix (Yang et al., 2011).

Pre-processing & Variant Calling

Pre-processing is done to prepare the reads got from sequencing for analysis. There are a few basic steps that are always done to ensure the variant calling is as accurate as it can be. These steps include aligning reads, marking duplicates and indexing. These steps are standard across different variant calling pipelines. When carrying out variant calling with GATK, it is recommended that base quality score recalibration is also carried out to account for biases in base score calling by the sequencing machines (Van der Auwera et al., 2013). This step was carried out before variant calling with Freebayes as well, although not necessary, to maintain a consistent input across both variant calling programmes. Carrying out this pre-processing step will have had a negligible effect on the variants called. Previous research showed no effect is seen on the number of Indels called, although for Freebayes, reduced sensitivity when calling SNPs is noted in regions of low divergence, while in regions with low coverage running BQSR results in increased precision (Tian et al., 2016). Therefore, carrying out this additional step should not have unduly influenced the variants called by Freebayes and ensured the input was consistent with what was used by GATK for variant calling.

When carrying out variant calling with GATK, there are several intermediate steps, as explained below, before arriving at a final VCF file. Initially, in this analysis, GATK HaplotypeCaller (HC) was used in GVCF mode as this runs the HC per-sample to generate an intermediary file, GVCF, which is then used for joint genotyping. HC can call both SNPs and Indels simultaneously, as when regions of variation are encountered, existing mapping information is discarded, and the reads are reassembled in this region

(GATK team, 2022a). HC locates regions of significant variation, which then undergo reassembly with a de Bruijn type graph indicating all possible haplotypes. Next read quality, and other factors, are taken into account to calculate the probability of a read at each position given all possible haplotypes, including the reference. These probabilities are subsequently used to calculate the evidence present for individual alleles at each locus resulting in allele likelihoods. Culminating in these per read allele likelihoods being used to calculate genotype likelihoods. Bayes theorem is used to work out the likelihoods of each possible genotype, with the most probable finally selected (GATK team, 2022b). These genotypes are recorded per sample in a GVCF file before being merged into a single VCF file using GenomicsDBImport. Joint genotyping is carried out in the next step with GenotypeGVFs. The final VCF file is a squared off matrix where SNP's and Indels are called jointly, resulting in genotypes for all sites of interest in all samples considered.

In comparison Freebayes, also a haplotype based variant caller, works by looking at haplotypes based on the literal sequence rather than the alignment and in this way partially avoids the issues caused by repetitive sequences (Garrison & Marth, 2012).

Filtering was done on each VCF file separately. Additionally, GATK recommends that SNPs and Indels be filtered separately, before being recombined into a VCF file containing all variants called. The samples here were filtered using the base recommendations from GATK, these tend to be quite lenient to avoid dismissing true variants too soon. The intersection of the VCF files from GATK and FreeBayes were taken as this has previously been shown to improve the specificity of the variant calls (Field et al., 2015). Specificity is a measure of the true negatives called, where a true negative in this instance means that variants not called are definitively the reference allele and not a variant allele. The combined VCF file undergoes filtering based on a minor allele frequency of 0.01, and hardy-weinberg frequency of 0.00001, as well as a minimum quality of 30. The aim of filtering is to reduce the number of false positives that are still present in the data before analysis on the data is begun.

Repeat Masked vs masked genome

When looking at Figure 10, more variants are called when using the unmasked genome as compared to using the repeat masked genome. This is due to a combination of factors, such as the variant calling process, the effect of repeat masking and the variant callers used. The variant calling done with the unmasked genome will result in more variants as it was done across all samples and utilised joint variant calling. While variant calling done with the repeat masked genome used only a single sample and so the only variants called were those present within this sample, thereby excluding locations where the sample was homozygous for the reference but others within the population were heterozygous or homozygous for the alternate allele. Next there will be a reduction in variants called by using a repeat masked genome, as repetitive sequences account for 46% of the equine genome

(Wade et al., 2009), there is far less of the genome for reads to be aligned to and fewer variants will thus be called. However, of the variants called, there is likely to be more confidence in them, as they are unlikely to have resulted from an improperly aligned read in a repetitive region.

Different aligners and variant callers deal with the issues of repetitive reads differently. These reads were aligned using BWA-MEM. With this program, if a read matches equally to at least 2 locations then it will assign a mapping quality of 0. This information is accounted for when calculating the quality of the variant and so can often result in variants that were called in repetitive regions being filtered out. This can be clearly seen in Figure 8 and Image 1, where a gap of 200,000 bp is seen in the plotted allele frequencies and Weir and Cockerham's Fst, while in the image from IGV reads are still mapping to much of this region. However, they are unshaded as they have a mapping quality of 0. Thereby indicating no confidence in the nucleotides called at these locations. This is one of the methods used to reduce false positives when mapping to repetitive regions. In addition, FreeBayes, as mentioned above, avoids some of the issues caused by repetitive regions by using the literal sequence rather than the precise alignment.

In this thesis an unmasked reference genome was used for the entire process up to and including variant calling. The greatest risk when doing this is an increase in false positives, variants that have been called but are not actually variants, caused by the difficulties in mapping to repetitive regions. This risk has been reduced through the aligners and variant callers used here, as well as filtering the results and taking the intersect of the two VCF files produced from the different variant callers.

Genes

This thesis was undertaken to investigate the association between white markings and the peaks seen in the GWAS on chromosome 23. This was of particular interest as white markings had not been associated with a gene in this region. While there are substantial differences in allele frequency over the areas where the top SNPs were located, there are no protein coding genes in these areas. The closest gene is MLANA which is part of a transcription pathway that is regulated by MITF, a gene known to be involved with white markings.

As mentioned before the KIT gene is connected with many white spotting phenotypes, in one instance even considered to be responsible for 80% of the inheritance we see in white markings (Rieder, 2009). The GWAS did not indicate associations between this gene and the white markings seen in the Fjord horses. This however does not preclude the possibility that KIT-alleles may play a role in some but not all individuals with white markings (making the association hard to detect). Many of the KIT mutations can cause white markings even when only present in the heterozygous form. For example, W20 is one

allele that causes white markings of the head and legs, similar to those investigated here in the fjord horse. When one W20 allele is present a white marking is visible on the head as well as white leg markings, however when homozygous for the W20 allele, a more pronounced white marking, covering more of the face is seen (Hauswirth et al., 2013). There is also striking variability within W alleles regarding the amount of white in the coat colour as horses with W1, W5 and W10 alleles can be completely white or still be significantly pigmented (Haase et al., 2009).

The differences seen in the allele frequencies around the KIT gene as seen in Figure 5, were all below 0.5, indicating a less than 50% difference in the frequency at which an allele occurred in the cases as opposed to the controls. Which might not seem to indicate anything of significance in the region. However, as seen in Figure 5, all differences above 0.4 were found overlaying the region of the KIT gene. Therefore, it is possible that there is a heterozygous mutation here, or that only some animals have a mutation in this gene. This is because if all the controls were homozygous for the wild type reference allele and all or most of the cases were heterozygous for the mutation, this would give a max difference in allele frequency of 0.5. This might not have been significant enough to be picked up in the GWAS, however, there is some difference to be seen here. In addition, Weir and Cockerham's Fst follows a similar pattern, indicating that in this region 40% of the diversity seen is due to between population differences and not within population differences. The populations in this situation being animals with or without white markings. There is some variants that have an Fst below 0, this is due to how Weir and Cockerham's Fst is calculated, The negative values can be interpreted as 0, where there is no diversity between the populations being analysed.

The MITF gene has been implicated with white markings, not to the extent of the KIT gene however all three important melanin antigens are found within common transcriptional pathways that are regulated by MITF (Du et al., 2003). MITF encodes a transcription factor that has an important role in the normal development of melanocytes and regulates expression of several pigmentation genes (Du et al., 2003). Mutations in MITF have in combination with KIT been observed to result in more extreme depigmentation than if mutations in either gene are present solely (Hauswirth et al., 2013). The differences in allele frequency in this area is also elevated around the gene, Figure 7, however the difference is not very striking, additionally a max difference in Fst of 0.17 was recorded, which can be interpreted as 17% of the variation seen is due to between population differences, in addition the pattern of high and low Fst values mirrors that of the allele frequencies.

Finally, PAX3 is another gene that has been associated with white markings. PAX3 has been associated with the splashed white phenotype, with mutations recorded in Appaloosas with leopard spotting complex and white splashed markings, as well as in several unrelated horses with splashed white

markings (Hauswirth et al., 2012). However when mutations are present in both PAX3 and MITF some animals with the splashed white phenotype were found to also be deaf (Hauswirth et al., 2013). PAX3 is not thought to be the sole gene responsible for the splashed white phenotype, especially as a large variance in phenotypes is seen. The allele frequencies seen in this gene, Figure 7, reach a max difference of 0.3, with no clear differentiation between the region the gene is found in and that around it. The max Fst value mirrors the allele frequencies here, however there is nearly twice as much between population variation recorded in this region, than in the region around MITF, which has similar allele frequencies.

The possibility of multiple mutations causing the white markings seen here must also be considered. In many instances white markings have been found to be a polygenic multifactorial trait and so there is no single causal mutation responsible for the phenotype seen (Rieder et al., 2008). Many of the KIT dominant white, W, alleles have come about in the last 20 years, all starting in one founding member (Haase et al., 2013), and so several mutations in different genes could have caused the white markings we see. Especially as the samples were roughly selected so as to not be closely related, this could have an impact on the strength of the association seen. In closely related animals with similar white marking phenotypes, one would expect to see the same mutations in the same genes to be responsible for the phenotype, meaning associations are more easily recognised.

In addition to examining the known white marking genes the regions surrounding the peaks seen in the GWAS were inspected for differences in allele frequency as well as Weir and Cockerham's Fst to examine differences between cases and controls. There were large differences seen in the allele frequencies surrounding peak 1, with the greatest difference corresponding to the location of the SNP with the greatest significance in the GWAS. There was a max difference in allele frequency seen of 0.75, while the Fst value had a max of 0.71, indicating that 71% of variation seen at this location is between population variation. This strongly supports the results from the GWAS. The second peak also had large differences in allele frequency, with a max difference in allele frequency of 0.5. This is more similar to the differences in allele frequency seen around the KIT gene. Weir and Cockerham's Fst followed the pattern of the allele frequencies, and had a max value of 0.5. This is a larger difference than seen for any of the known genes and again reinforces the peaks identified through the GWAS.

Variants Called

As noted in the results for Figures 8 and 9 none of the SNPs highlighted by the GWAS were found in the sequenced data. This is likely due to differences in reference genome. The genotyping done for the GWAS was done in 2017 and so the locations of SNPs on the SNPchip came from EquCab2.0,

however for this data the reference sequence was EquCab3.0 (Kalbfleisch et al., 2018). The significant SNPs found following WGS were all in similar but not identical locations and this is likely due to improvements made to the reference genome.

The regions of interest were all annotated using SnpEff and then filtered using SnpSifts case control function. This generated a contingency table of the occurrence of each possible genotype at a variant location. Here a fishers exact test was used to compare between cases and controls and a p-value of 0.005 was used as the cut off for significant differences. This resulted in 44 variants remaining of which 9 were in the regions with the greatest allele frequency differences on chromosome 23 as highlighted in Figure 8. These genes were not protein coding and all located in an intergenic region. They were all associated with U6 spliceosomal RNA, ENSECAG00000027196, with no obvious connection to white markings. U6 spliceosomal RNA is the most highly conserved spliceosomal RNA, found unchanged from yeast to mammals (Brow & Guthrie, 1988). However, our understanding of the involvement of non-coding and regulatory regions in a variety of phenotypes is constantly evolving and these variants could have an as yet unknown part to play in the phenotype we see. The filtering may have also been too stringent as filtering was carried out on all the regions of interest but only variants on chromosome 23 passed filtering by SnpSift. While only variants on chromosome 23 passed the filtering by SnpSift, it is likely that mutations in the other white marking genes have some involvement.

Phenotypes

There is a distinct weakness in the phenotypes gathered as they rely on owners self-declaring white head and leg markings. In an ideal situation there would have been photos of all the horses to accurately document the white markings, or white markings would have been recorded on a diagram as blood samples were being taken. This would have allowed for differences relating to the size or location of the white markings to have been investigated and accounted for. This has been previously shown to affect the genes involved depending on the size of the head marking and whether leg markings are found on the front or hind legs (Haase et al., 2013, p. 201; Hauswirth et al., 2013; Rieder et al., 2008). With only limited number of cases sequenced this could have affected further analysis by clouding associations between cases and potential causal SNPs.

Future research

Of interest for future research would be to use long read sequencing to sequence in particular the highly repetitive regions on chromosome 23 to allow for more accuracy in variant calling. In addition, investigating if the fjord horses with white markings have any of the known mutations in KIT or MITF as these were potentially indicated as sources of variation. Examination of the SNPs that were highly

significant in the GWAS for SNPs that are in LD with them could be another area of further research. In addition having more animals sequenced could increase the power of the analysis and provide clearer associations to the multiple variants that are likely involved.

Conclusion

The aim of this thesis was to fine map a QTL on chromosome 23 connected with white markings in fjord horses, with the hope of locating a causal variant. While a single causal variant was not uncovered this is not surprising as white markings are often a polygenic trait and therefore many different genes have a role in producing the phenotype that is seen. However, through close examination of regions of interest on chromosome 23, these regions have been further narrowed down to a region spanning 29,566,014- 29,797,765bp. Within this region there are no protein coding genes, however variants identified using SnpSift CaseControl in the region are connected with U6 spliceosomal RNA. These may have an as yet unknown role to play in the white marking phenotypes seen in Fjord horses.

References

Bathke, J., & Lühken, G. (2021). OVarFlow: A resource optimized GATK 4 based Open source Variant calling workFlow. *BMC Bioinformatics*, *22*(1), 402. https://doi.org/10.1186/s12859-021-04317-y

Bellone, R. R., Holl, H., Setaluri, V., Devi, S., Maddodi, N., Archer, S., Sandmeyer, L., Ludwig, A., Foerster, D., Pruvost, M., Reissmann, M., Bortfeldt, R., Adelson, D. L., Lim, S. L., Nelson, J., Haase, B., Engensteiner, M., Leeb, T., Forsyth, G., … Brooks, S. A. (2013). Evidence for a Retroviral Insertion in TRPM1 as the Cause of Congenital Stationary Night Blindness and Leopard Complex Spotting in the Horse. *PLoS ONE*, *8*(10), e78280. https://doi.org/10.1371/journal.pone.0078280

Brow, D. A., & Guthrie, C. (1988). Spliceosomal RNA U6 is remarkably conserved from yeast to mammals. *Nature*, *334*(6179), 213–218. https://doi.org/10.1038/334213a0

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w ^{1118} ; iso-2; iso-3. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Cone, R. D., Lu, D., Koppula, S., Vage, D. I., Klungland, H., Boston, B., Chen, W., Orth, D. N., Pouton, C., & Kesterson, R. A. (1996). The melanocortin receptors: Agonists, antagonists, and the hormonal control of pigmentation. *Recent Progress in Hormone Research*, *51*, 287–317; discussion 318.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Du, J., Miller, A. J., Widlund, H. R., Horstmann, M. A., Ramaswamy, S., & Fisher, D. E. (2003). MLANA/MART1 and SILV/PMEL17/GP100 Are Transcriptionally Regulated by MITF in Melanocytes and Melanoma. *The American Journal of Pathology*, *163*(1), 333–343. https://doi.org/10.1016/S0002-9440(10)63657-7

Durward-Akhurst, S. A., Schaefer, R. J., Grantham, B., Carey, W. K., Mickelson, J. R., & McCue, M. E. (2021). Genetic Variation and the Distribution of Variant Types in the Horse. *Frontiers in Genetics*, *12*, 758366. https://doi.org/10.3389/fgene.2021.758366

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Federici, M., Gerber, V., Doherr, M. G., Klopfenstein, S., & Burger, D. (2015). Association of skin problems with coat colour and white markings in three-year-old horses of the Franches-Montagnes breed. *Schweiz Arch Tierheilkd*, *157*(7), 391–398. https://doi.org/10.17236/sat00026

Field, M. A., Cho, V., Andrews, T. D., & Goodnow, C. C. (2015). Reliably Detecting Clinically Important Variants Requires Both Combined Variant Calls and Optimized Filtering Strategies. *PLOS ONE*, *10*(11), e0143199. https://doi.org/10.1371/journal.pone.0143199

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *ArXiv:1207.3907 [q-Bio]*. http://arxiv.org/abs/1207.3907

GATK team. (2022a, January). *Germline short variant discovery (SNPs + Indels)*. GATK. https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels-

GATK team. (2022b, April). *HaplotypeCaller in a nutshell*. GATK. https://gatk.broadinstitute.org/hc/en-us/articles/360035531412-HaplotypeCaller-in-a-nutshell

Haase, B., Brooks, S. A., Tozaki, T., Burger, D., Poncet, P.-A., Rieder, S., Hasegawa, T., Penedo, C., & Leeb, T. (2009). Seven novel *KIT* mutations in horses with white coat colour phenotypes. *Animal Genetics*, *40*(5), 623–629. https://doi.org/10.1111/j.1365-2052.2009.01893.x

Haase, B., Signer-Hasler, H., Binns, M. M., Obexer-Ruff, G., Hauswirth, R., Bellone, R. R., Burger, D., Rieder, S., Wade, C. M., & Leeb, T. (2013). Accumulating Mutations in Series of Haplotypes at the KIT and MITF Loci Are Major Determinants of White Markings in Franches-Montagnes Horses. *PLOS ONE*, *8*(9), 1–10. https://doi.org/10.1371/journal.pone.0075071

Hauswirth, R., Haase, B., Blatter, M., Brooks, S. A., Burger, D., Drögemüller, C., Gerber, V., Henke, D., Janda, J., Jude, R., Magdesian, K. G., Matthews, J. M., Poncet, P.-A., Svansson, V., Tozaki, T., Wilkinson-White, L., Penedo, M. C. T., Rieder, S., & Leeb, T. (2012). Mutations in MITF and PAX3 Cause "Splashed White" and Other White Spotting Phenotypes in Horses. *PLoS Genetics*, *8*(4), e1002653. https://doi.org/10.1371/journal.pgen.1002653

Hauswirth, R., Jude, R., Haase, B., Bellone, R. R., Archer, S., Holl, H., Brooks, S. A., Tozaki, T., Penedo, M. C. T., Rieder, S., & Leeb, T. (2013). Novel variants in the KIT and PAX3 genes in horses with white-spotted coat colour phenotypes. *Animal Genetics*, *44*(6), 763–765. https://doi.org/10.1111/age.12057

Høiseth, M. (2017). *Genetic variation and colour genetics of the Norwegian Fjord horse* [Master]. NMBU.

Holsinger, K. E., & Weir, B. S. (2009). Genetics in geographically structured populations: Defining, estimating and interpreting FST. *Nature Reviews Genetics*, *10*(9), 639–650. https://doi.org/10.1038/nrg2611

Imsland, F., McGowan, K., Rubin, C.-J., Henegar, C., Sundström, E., Berglund, J., Schwochow, D., Gustafson, U., Imsland, P., Lindblad-Toh, K., Lindgren, G., Mikko, S., Millon, L., Wade, C., Schubert, M., Orlando, L., Penedo, M. C. T., Barsh, G. S., & Andersson, L. (2016). Regulatory mutations in TBX3 disrupt asymmetric hair pigmentation that underlies Dun camouflage color in horses. *Nature Genetics*, *48*(2), 152–158. https://doi.org/10.1038/ng.3475

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., … MacLeod, J. N. (2018). Improved reference genome for the domestic horse increases assembly contiguity and composition. *Communications Biology*, *1*(1), 197. https://doi.org/10.1038/s42003-018-0199-z

Khalfan, M. (2020, March 25). *Variant Calling Pipeline using GATK4 – Genomics Core at NYU CGSB*. https://gencore.bio.nyu.edu/variant-calling-pipeline-gatk4/

Klungland, H., & Våge, D. I. (2000). Molecular Genetics of Pigmentation in Domestic Animals. *Current Genomics*, *1*(3), 223–242.

Kobayashi, T., Imokawa, G., Bennett, D. C., & Hearing, V. J. (1998). Tyrosinase Stabilization by Tyrp1 (the brown Locus Protein). *Journal of Biological Chemistry*, *273*(48), 31801–31805. https://doi.org/10.1074/jbc.273.48.31801

Kvist, L., & Niskanen, M. (2021). Modern Northern Domestic Horses Carry Mitochondrial DNA Similar to Przewalski's Horse. *Journal of Mammalian Evolution*, *28*(2), 371–376. https://doi.org/10.1007/s10914-020-09517-6

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for

analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Nestaas, T. (2002). *Colours and markings*. FHI. https://www.fjordhorseinternational.org/FjHI/images/handbook/09_colours%20and%20markings.pdf

Nestaas, T. (2014). *Colours of the Fjord horse*. NFHR.

NFHR. (n.d.). *NFHR Breed Standard*. Retrieved October 24, 2021, from https://www.nfhr.com/catalog/index.php?breedstd=1

Patterson Rosa, L., Martin, K., Vierra, M., Foster, G., Lundquist, E., Brooks, S. A., & Lafayette, C. (2021). Two Variants of KIT Causing White Patterning in Stock-Type Horses. *Journal of Heredity*, *112*(5), 447–451. https://doi.org/10.1093/jhered/esab033

Picard Tools—By Broad Institute (2.26.10). (2022). [Computer software]. The Broad Institute. http://broadinstitute.github.io/picard/

R Core Team. (105 C.E.). *R: A language and environment for statistical computing.* (4.1.1) [Windows x64]. R Foundation for Statistical Computing. https://www.R-project.org/

Rieder, S. (2009). Molecular tests for coat colours in horses. *Journal of Animal Breeding and Genetics*, *126*(6), 415–424. https://doi.org/10.1111/j.1439-0388.2009.00832.x

Rieder, S., Hagger, C., Obexer-Ruff, G., Leeb, T., & Poncet, P.-A. (2008). Genetic Analysis of White Facial and Leg Markings in the Swiss Franches-Montagnes Horse Breed. *Journal of Heredity*, *99*(2), 130–136. https://doi.org/10.1093/jhered/esm115

Rieder, S., Stricker, C., Joerg, H., Dummer, R., & Stranzinger, G. (2000). A comparative genetic approach for the investigation of ageing grey horse melanoma. *Journal of Animal Breeding and Genetics*, *117*(2), 73–82. https://doi.org/10.1111/j.1439-0388.2000x.00245.x

Rieder, S., Taourit, S., Mariat, D., Langlois, B., & Guérin, G. (2001). Mutations in the agouti (ASIP), the extension (MC1R), and the brown (TYRP1) loci and their association to coat color phenotypes in horses (Equus caballus). *Mammalian Genome*, *12*(6), 450–455. https://doi.org/10.1007/s003350020017

Rosengren Pielberg, G., Golovko, A., Sundström, E., Curik, I., Lennartsson, J., Seltenhammer, M. H., Druml, T., Binns, M., Fitzsimmons, C., Lindgren, G., Sandberg, K., Baumung, R., Vetterlein, M., Strömberg, S., Grabherr, M., Wade, C., Lindblad-Toh, K., Pontén, F., Heldin, C.-H., … Andersson, L. (2008). A cis-acting regulatory mutation causes premature hair graying and susceptibility to melanoma in the horse. *Nature Genetics*, *40*(8), 1004–1009. https://doi.org/10.1038/ng.185

Sandmeyer, L. S., Bellone, R. R., Archer, S., Bauer, B. S., Nelson, J., Forsyth, G., & Grahn, B. H. (2012). Congenital stationary night blindness is associated with the leopard complex in the miniature horse. *Veterinary Ophthalmology*, *15*(1), 18–22. https://doi.org/10.1111/j.1463-5224.2011.00903.x

Santschi, E. M., Purdy, A. K., Valberg, S. J., Vrotsos, P. D., Kaese, H., & Mickelson, J. R. (1998). Endothelin receptor B polymorphism associated with lethal white foal syndrome in horses. *Mammalian Genome: Official Journal of the International Mammalian Genome Society*, *9*(4), 306–309. https://doi.org/10.1007/s003359900754

Searle, A. G. (1968). *Comparative Genetics of Coat Colour in Mammals*. Logos Press.

Tian, S., Yan, H., Kalmbach, M., & Slager, S. L. (2016). Impact of post-alignment processing in variant discovery from whole exome data. *BMC Bioinformatics*, *17*, 403. https://doi.org/10.1186/s12859-016-1279-z

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013).

From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*, *43*(1), 11.10.1-11.10.33. https://doi.org/10.1002/0471250953.bi1110s43

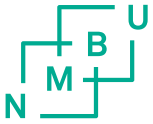
Wade, C. M., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., Lear, T. L., Adelson, D. L., Bailey, E., Bellone, R. R., Blöcker, H., Distl, O., Edgar, R. C., Garber, M., Leeb, T., Mauceli, E., MacLeod, J. N., Penedo, M. C. T., Raison, J. M., … Lindblad-Toh, K. (2009). Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science (New York, N.Y.)*, *326*(5954), 865–867. https://doi.org/10.1126/science.1178158

Weir, B., & Cockerham, C. (1984). Weir BS, Cockerham CC.. Estimating F-Statistics for the Analysis of Population-Structure. Evolution 38: 1358-1370. *Evolution*, *38*, 1358–1370. https://doi.org/10.2307/2408641

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A Tool for Genome-wide Complex Trait Analysis. *The American Journal of Human Genetics*, *88*(1), 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011

Yu, X., Guda, K., Willis, J., Veigl, M., Wang, Z., Markowitz, S., Adams, M. D., & Sun, S. (2012). How do alignment programs perform on sequencing data with varying qualities and from repetitive regions? *BioData Mining*, *5*(1), 1–12. https://doi.org/10.1186/1756-0381-5-6

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, *44*(7), 821–824. https://doi.org/10.1038/ng.2310



Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway