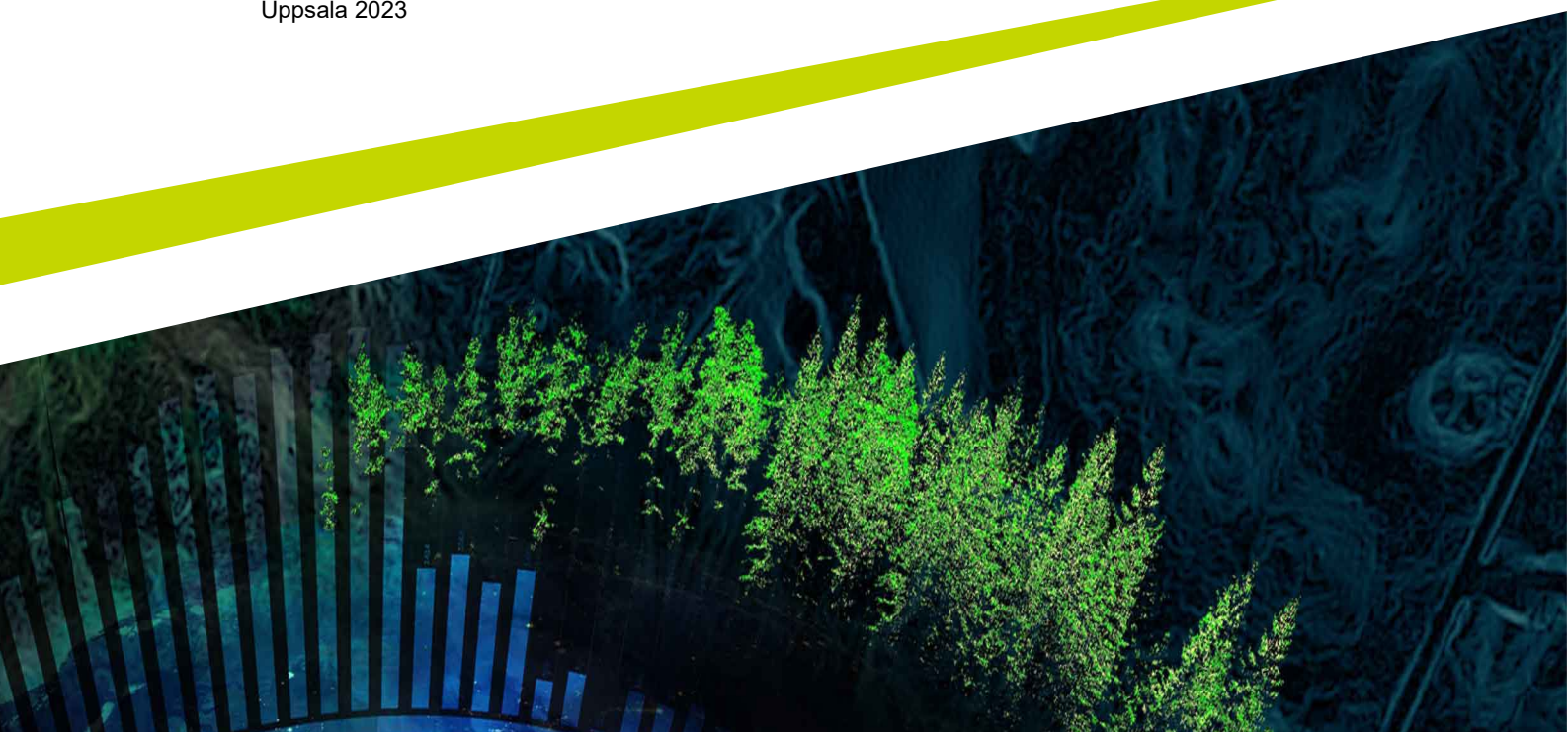
Towards mapping the genetic cause of polydactyly in a miniature Shetland pony family

Delyan Georgiev

Independent project • 30 credits
Swedish University of Agricultural Sciences, SLU
Department of Animal Breeding and Genetics
Uppsala 2023



Towards mapping the genetic cause of polydactyly in a miniature Shetland pony family

Kartläggning av den genetiska orsaken till polydaktyli genom analys av helgenomssekvensdata från rasen miniatyr shetlandsponny

Delyan Georgiev

Supervisor:	Gabriella Lindgren, SLU, Department of Animal Breeding and Genetics
Examiner:	Tomas Bergström, SLU, Department of Animal Breeding and Genetics

Credits:	30
Level:	Master's level (A2E)
Course title:	Independent project in Bioinformatics, A2E
Course code:	EX1002
Programme/education:	Bioinformatics
Course coordinating dept:	Department of Animal Breeding and Genetics
Year of publication:	2023
Copyright:	All featured images are used with permission from the copyright owner.
Keywords:	polydactyly, miniature Shetland pony, WGS, bioinformatics

Swedish University of Agricultural Sciences
Department of Animal Breeding and Genetics

Abstract

Polydactyly is a vertebrate developmental condition characterized by the presence of supernumerary digits. Its manifestation and severity can vary, ranging from cosmetic to debilitating, and it is commonly inherited in a dominant manner. The present study investigates a family of miniature Shetland ponies exhibiting a rare case of polydactyly, which appears to be inherited recessively and is accompanied by limb deformities. To shed light on the genetic basis of this observed phenotype, bioinformatic analyses were conducted on the whole genome sequences of five family members (consisting of one stallion, two mares, and two foals), along with several samples from healthy controls. The aim was to compare and contrast their genomes, in order to identify the mutation responsible for this change in development. Although these analyses succeeded in narrowing down the potential variations from millions to thousands, the limited number of samples proved inadequate to definitively pinpoint the precise genetic change underlying this distinct form of polydactyly.

Table of contents

List of tables

List of figures

Abbreviations

WGS	whole genome sequencing
SNP	single nucleotide polymorphism
InDel	insertion/deletion
GATK	Genomic Analysis Tool Kit
SRA	Sequence Read Archive
Fst	fixation index
DNA	deoxyribonucleic acid
bp	base pair
AER	apical ectodermal ridge
PZ	progress zone
ZPA	zone of polarizing activity

1. Introduction

The vertebrate limb is a versatile adaptation, enabling animals to complete all manner of tasks, from locomotion and survival, to tool usage and social interaction. It is also an object of interest of developmental biology, due in part to the observable homology between the limbs of different vertebrates (Coates 1994). Understanding the evolutionary history that connects the fins of a fish to the opposable thumbs of a primate can shed light on the existence of the entire terrestrial tetrapod group.

The word chiral is used to describe molecules that have no axis of symmetry, meaning their mirror image cannot be superimposed on themselves no matter the combination of rotations and translations. It derives its meaning from the Greek word for hand, which is one of the most commonly encountered chiral objects. To properly describe the hand, or any vertebrate limb in general, all the axes have specific names, which describe them in relation to the body. The direction that goes from the trunk to the fingers is called the proximal to distal axis; the direction from the thumb to the ring finger is called anterior to posterior axis; and finally the direction from the back of the hand to the palm is called the dorsal to ventral axis.

The bones of vertebrate limb can be used as a guide to distinguish the three morphologically distinct regions, following the proximal to distal axis. The humerus represents the stylopod, the ulna and radius represent the zeugopod, and the autopod contains all bones normally associated with the hand. This is useful when talking about different species, as the wing of a chicken, for example, has only three fingers, a human hand has five, while the horse only has a single digit. However, all three species have an stylopod, zeugopod and an autopod from a developmental point of view

The determination of tissues along an axis is strongly associated with the homeobox (hox) genes (Gehring 1993). This is true for the whole embryo, as well as the limb in particular, only this time the rule of collinearity (the correlation of hox genes position along the chromosome to tissue order along the axis) operates on the proximal to distal axis (Duboule 1998). Any loss of function of limb-associated hox genes results in severe limb deformities (Small & Potter 1993). The buds that would grow to form the limbs bulge off the main body, forming due to cell migration from the lateral plate mesoderm and increased proliferation. This is followed by formation of the bones, cartilage, muscles, as well as nerve and blood vessel infiltration.

Even at the limb bud stages, all the axes are already established. This is determined by three distinct regions of significance, the first being a ridge of epithelium at the most distal part of the bud. This structure, called the apical ectodermal ridge (AER), delineates the dorsal and ventral sides of the bud. The clump of cells growing beneath AER is called the progress zone (PZ), and the third and final region is known as the zone of polarizing activity (ZPA), located on the posterior-proximal side of the bud. While it doesn't have a distinct physical appearance, experiments with chicken embryos, in which the ZPA is grafted to the anterior-proximal side of the bud, resulted in a mirror-image limb morphology (Summerbell 1979). Conversely, removing the AER would stop the bud development at whatever stage the manipulation was performed. And replacing the tissue of the PZ of a forelimb with a hind limb would result in a leg growing where an arm would be (Muneoka & Bryant 1984). These are just some examples of what the different zones of the limb bud do and their interactions.

Limb bud formation is determined by the interplay of several known morphogens. Initiation of bud formation is influenced by the expression of the Tbx5 (forelimb) and Tbx4 (hind limb) transcription factors, with some species variation (Takeuchi et al. 1999). No matter precisely which gene is the initiator, it all leads to the expression of fibroblast growth factor 10 (FGF10) in PZ, which is the main cell fate contributing factor. AER affects the limb bud development in the proximal-distal axis through FGF8 signalling (Mahmood et al. 1995). Removal of the AER stops limb development as mentioned previously, but supplementing FGF8 is enough to restore normal function. The anterior-posterior axis development is determined through a sonic hedgehog (Shh) gradient originating from the ZPA, and Wnt and BMP determine the dorsal-ventral polarity.

In the limb bud, Shh is solely expressed in the ZPA, and its gradient determines the anterior-posterior axis development (McGlinn & Tabin 2006). Mutations in Shh regulatory elements have been found in several species, and they result in mirror-image duplication of the digits, similar to the grafting experiments done with chicken embryos (Lettice et al. 2003). An increase in the expected digit number for a species is called polydactyly, and in the case of the horse that can present a major challenge, since the entire upper limb makes contact with the ground through that single digit, so structural changes can affect the gait or balance of the animal, or even make it impossible to walk in severe occurrences.

This work examines the occurrence of a unique type of polydactyly in a small family of miniature Shetland ponies. The breed is related to the Shetland pony and is distinguished by the height measured at the withers, which shouldn't exceed 87 cm. The group that is the object of this thesis consists of 5 individuals. The family tree of these horses is represented in Figure 1.

- One healthy stallion, named SU1, that has previously had 26 healthy offspring before producing a foal with polydactyly.

- One healthy mare, named MS2, that has had 2 healthy offspring before giving birth to a foal with polydactyly.
- One healthy mare, named MK1, daughter of SU1 and a mare for which no genetic information or pedigree was provided.
- One foal with polydactyly, named FO1, son of SU1 and MS2
- One foal with polydactyly, named FO2, result of inbreeding between MK1 and SU1

Figure 1. Family tree of studied horses

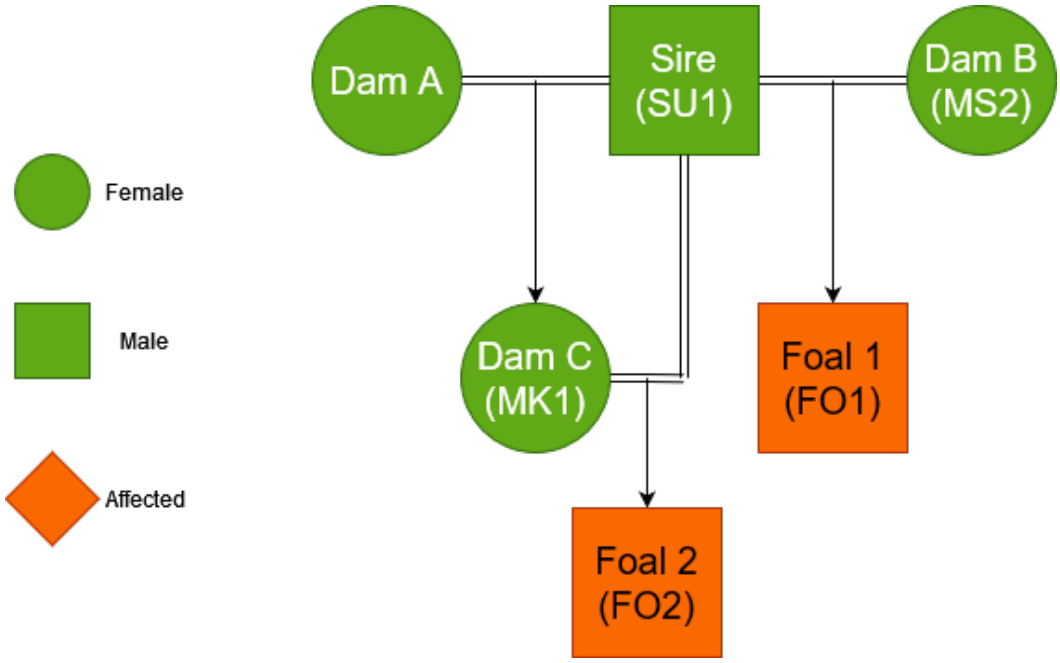


Diagram representing the familial relationships between the 5 studied horses. Dam A has no provided genetic data and is thus not part of the study. The sire is entered as sample SU1, Dam B is MS2, Dam C is MK1, Foal 1 is FO1 and Foal 2 is FO2 when referred further into the text.

Polydactyly in horses is still not fully understood on the genetic level. In humans, the body of research is much larger, and even there the genetic pathway is not completely elucidated. More than 10 loci and 6 genes are involved in non-syndromic polydactyly, and even more genes can be considered when including developmental defects that have polydactyly as just one of their symptoms (Umair et al. 2018).

To better understand the genetic causes behind this particular form of polydactyly in the studied family, sequencing data from publicly available repositories was chosen to compare and contrast with. Where possible, only sequencing data from miniature Shetland ponies was selected to minimise the risk for creating false positives due to breed-specific loci. However, sometimes regular Shetland ponies were included in the selection due to the low volume of available sequences.

The first dataset was selected from a paper studying skeletal atavism in Shetland ponies (Rafati et al. 2016), because it used a large and diverse pool of manually chosen individuals, which provides an excellent genetic background for all future comparisons. Another dataset was obtained from a paper exploring the quantitative trait loci behind height in Shetland ponies (Frischknecht et al. 2015), the major factor that distinguishes "regular" from miniature. This dataset was selected for its excellent sequence quality and coverage, and because it uses miniature Shetland ponies as sequencing targets, despite the low count of samples studied. The last dataset was from a study that identifies a potential causative mutation resulting in dwarfism in miniature Shetland ponies (Metzger et al. 2017). While that individual was afflicted, its phenotype did not include any polydactyly-like symptoms, so for the purposes of this work it was considered healthy.

2. Methods and Materials

2.1 Sequence data acquisition

The thesis project was exclusively in the area of computational biology, so no handling of animals or any biological samples was done. All sequencing data used was either obtained from public databases or provided in digital form.

2.2 Data collection table

Some of the sequences in this table are not yet publicly available.

Table 1. List of sequencing data accessions used for this project.

Sample	Original Name	Accession
CG1	CG_1	SAMN04538183
CG2	CG_2	SAMN04538182
CG3	CG_3	SAMN04538181
CG4	CG_4	SAMN04538180
CG5	CG_5	SAMN04538179
CG6	CG_6	SAMN04538178
GP1	CG_7-tag12_CTTGTA	SAMN04538177
GP2	CG_7	SAMN04538176
NG1	NGSHORSE028	SAMN05440129
SP1	SPH041	SAMEA3367609
SP2	SPH020	SAMEA3367610
SU1	Sire	Own sample
MK1	Dam C	Own sample
MS2	Dam B	Own sample
FO1	Foal 1	Own sample
FO2	Foal 2	Own sample

2.3 Whole-Genome Sequencing

The data for this work was obtained from several sources and was generated using short read whole genome sequencing (WGS) methods. This was taken into consideration when choosing an appropriate software for data analysis. Due to the diversity of sources, there was no single platform that was used across all samples, with examples including Illumina Next-Seq 500, Illumina HiSeq 2000, and others.

2.4 Computational devices used

The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

Some of the lighter data analysis steps were performed on my personal computer.

A work computer was provided for the purposes of writing the thesis and accessing the supercomputing cluster (UPPMAX) while on university grounds.

2.5 Software and languages used

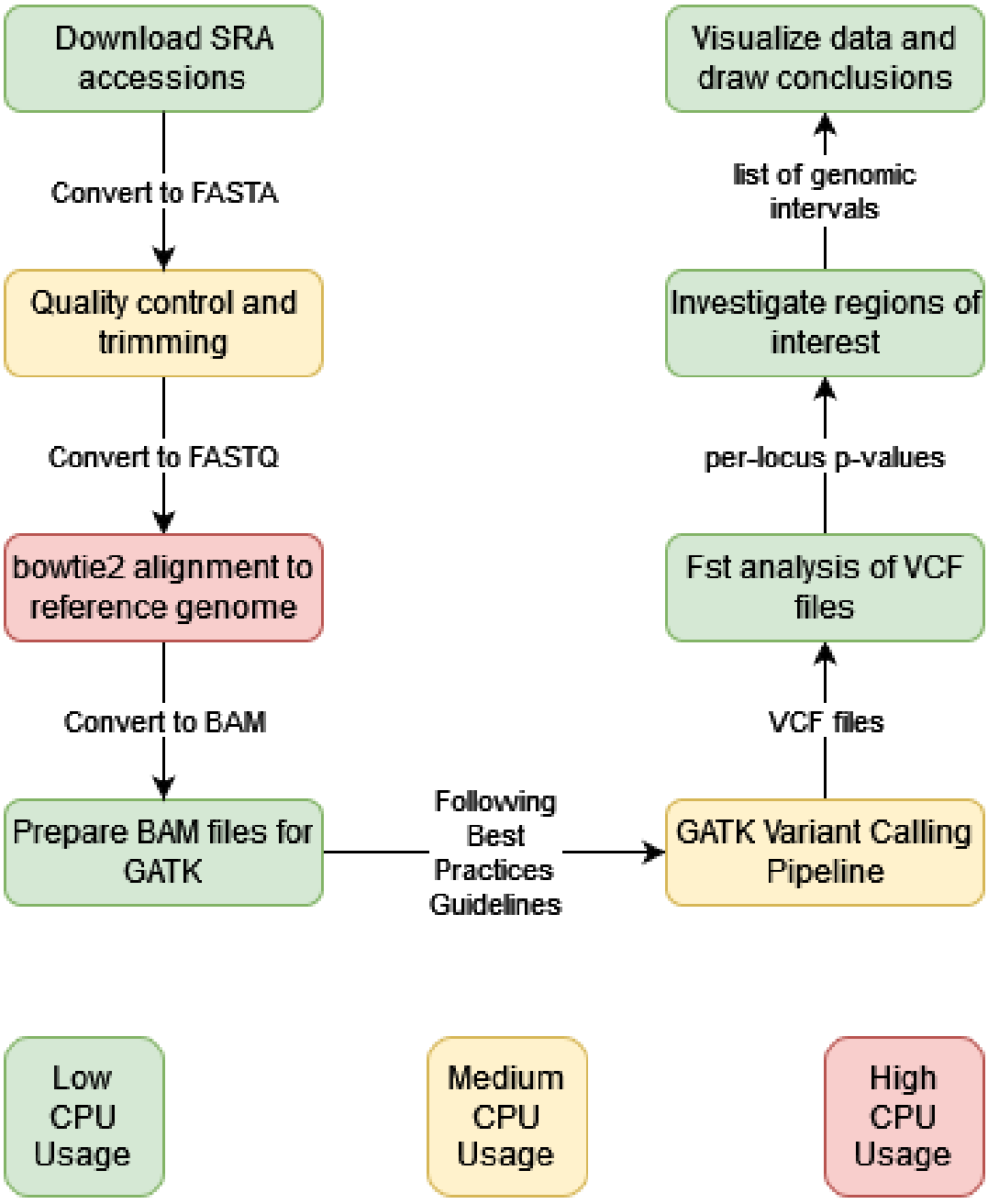
Languages used to perform the computational analysis of the data include bash, Perl, Java, Python, C#, R.

Software used in addition to the ones outlined in the next section include the Microsoft Office suite, RStudio 2023.03.0, PyCharm 2022.2.1 (Community Edition) and Visual Studio Code 1.78.2.

2.6 Data analysis

A detailed breakdown of the scripts and commands used to obtain the results presented in this work is provided in Appendix 1. The brief summary of the workflow is illustrated on Figure 2. The intent is to process the raw sequences by removing low-quality reads, then aligning the output against the horse reference genome, and converting the output to a format that can be processed by an established data analysis pipeline. The results from that pipeline show which samples contain mutations in comparison to the reference, and by using statistical methods, information can be obtained that can lead to pinpointing the molecular origin of the phenotype being investigated.

Figure 2. Workflow chart



Workflow chart depicting the key moments of the variant calling workflow. Data is being downloaded from public repositories or provided on a local machine in the FASTA format. After quality control and aligning to the reference genome, the data is output as a binary file (BAM) to speed up downstream analysis. Steps are taken to prepare this binary file in a way that is consistent with the Genomic Analysis Toolkit pipeline of the Broad Institute. The result of that pipeline is a text file (VCF) that contains all genomic loci with differences, compared to the reference genome provided (EquCab 3.0 at the time of writing). These files are then used in various ways, such as obtaining fixation statistic information or cross-referencing regions of interest within the genome.

3. Results

3.1 Data pool

In total, the genomes of 16 different individuals were analysed. Where applicable, several low-coverage sequencing runs from the same biological sample were combined into one to increase the depth of coverage for further downstream analysis. For example, the miniature Shetland pony family data was supplied in the form of 3 separate runs, so it required a merging step as described in the Methods section. Other samples (like NG1, Table 2) were done in a single run and were directly processed further.

Table 2. Sample sequencing run counts

Sample name	Run counts
CG1	4
CG2	4
CG3	4
CG4	4
CG5	4
CG6	4
GP1	4
GP2	4
NG1	1
SP1	1
SP2	1
SU1	3
MK1	3
MS2	3
FO1	3
FO2	3

3.2 Statistics and quality control of sequencing

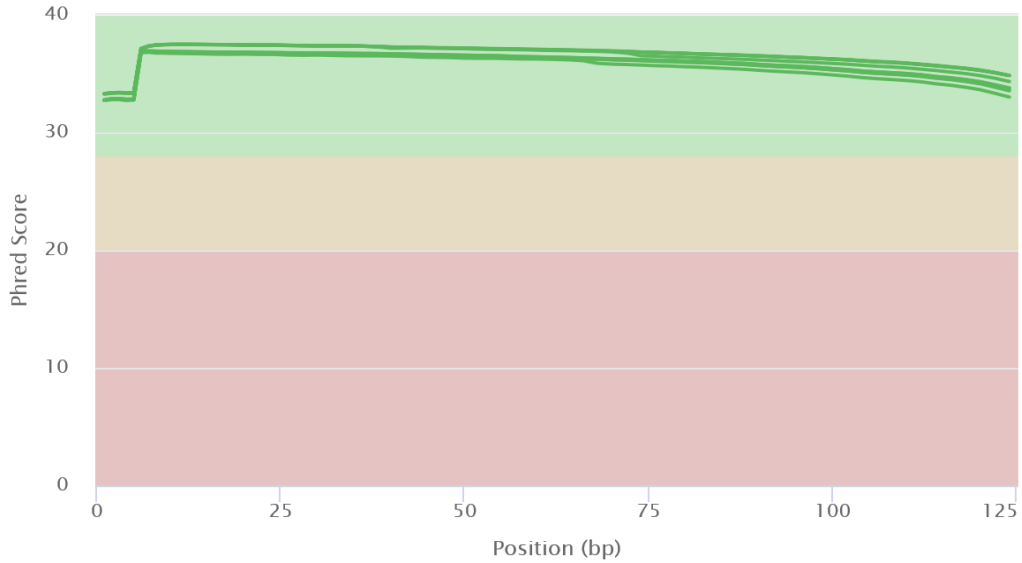
Similar steps were performed for all sequencing runs. with this section mostly focusing on the miniature Shetland pony family data. While there are some slight deviations between samples and runs, there were no significant differences between runs from the viewpoint of sequencing quality. Therefore, these results can be considered representative for the overall quality of all samples, unless stated otherwise.

Table 3. MultiQC statistics of studied samples

Sample	%Dups	%GC	Read Length	M. pairs
SU1	7.6	42	124	37.4
MK1	8.1	42	124	36.7
MS2	15.7	45	124	48.7
FO1	7.7	44	124	38.3
FO2	8.1	42	124	40.0

Sequencing quality across the whole read was very high for all samples, with average of 10% of reads failing the quality control checks, and ranging from 0% for some samples to 20% for sample NG1, but that one uses a different sequencing technology (NextSeq 500), has longer reads (150bp), and more than 3 times as much in quantity (145M).

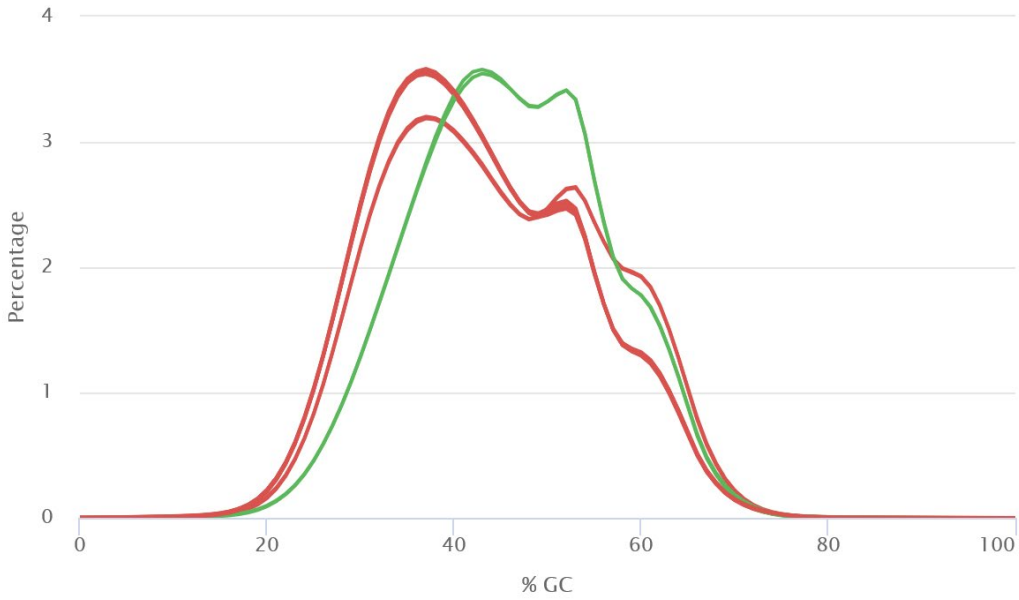
Figure 3.Per base sequence quality score.



The X axis represents the base position of the read, and the Y axis represents the PHRED33 score assigned to that base. PHRED score is a measure of the confidence of the base call and is in base 10 logarithmic scale. 10-point increase in the score corresponds to a 10-fold increase in confidence.

MultiQC produced a warning when assessing the per sequence GC content of the samples. However, despite the unusual distribution, this trait is consistent across all horse genomes analysed, from several different laboratories and sequencing technologies. Online forums and verbal communication with experts in the field also confirm this oddity of the horse genome, so the GC content distribution will not be taken into consideration further.

Figure 4. Per sequence GC content.



The X axis represents relative GC content as percentage, the Y axis represents how many of the reads fall into each category as percentage. The graph includes all 5 of the individuals being analysed. The green line is sample MS2, marked green because it seems to fall just barely into FastQC's quality tresholds.

Based on the quality thresholds set in the tool trimmomatic, some reads were disqualified. This resulted in sequences being sorted into four distinct categories – forward paired, reverse paired, forward unpaired and reverse unpaired. All reads initially are paired, but if one of the sequences fails the quality control checks, it is assigned to an unpaired bin. The overall drop rate varied by sample and sequencing run, but the average of 100 bins (50 sequencing runs times forward and reverse reads) is 8.85% with a standard deviation of 5.63%.

3.3 Statistics and quality control of mapping

Mapping was performed only on the paired reads from each run, which while reducing the overall number of sequences available, increases the confidence in their alignment later on. The mapper tool of choice was bowtie2, due to my familiarity with it and its good performance when aligning Illumina DNA

sequencing outputs (Canzar & Salzberg 2017). When working in paired-end mode, bowtie2 categorises reads into several groups based on their alignment. Results were similar for all sequencing runs analysed, and Table 4 shows some of the statistics for my samples.

Table 4. Alignment statistics.

Sample	Reads, 10e6	Concordant %	Unique Hit %	Overall Alignment %
SU1	32.0	95.66	70.70	99.70
MK1	31.7	96.05	70.65	99.70
MS2	46.9	97.78	70.85	99.79
FO1	32.6	85.58	62.98	88.30
FO2	34.6	96.66	71.07	99.60

Reads are represented in millions, rounded to the first decimal. Concordant reads are reads that mapped as expected by bowtie2's parameters, for example if the two mates of a pair map 1000 bases apart when the insert is supposed to be 125, then the reads mapped discordantly. Unique hits are the percentage of reads that mapped concordantly and only once. Overall alignment is the percentage of reads that mapped in any capacity, regardless of mates or concordance.

After the clean-up of mapping results (like duplicate marking, sorting, quality recalibration), a useful metric is to check the overall coverage of the genome. This can give insight into the mapping distribution, any potential sequencing errors or large-scale chromosomal changes, such as deletions, duplications, changes in ploidy. Since the data is a lot and hard to represent visually in its entirety, only sample plots will be shown here, with notable exceptions pointed out when necessary.

Table 5. Average coverage per sample.

Sample	Chr01	Chr16	ChrX	Sex
SU1	11.44	11.54	6.15	M
MK1	10.95	11.08	11.16	F
MS2	13.18	13.84	11.64	F
FO1	10.32	10.45	5.56	M
FO2	12.29	12.38	6.62	M
GP1	14.15	14.31	7.81	M
GP2	23.47	23.74	12.88	M
NG1	9.1	9.29	8.8	F
SP1	19.57	19.66	20.64	F
SP2	17.7	17.89	18.81	F

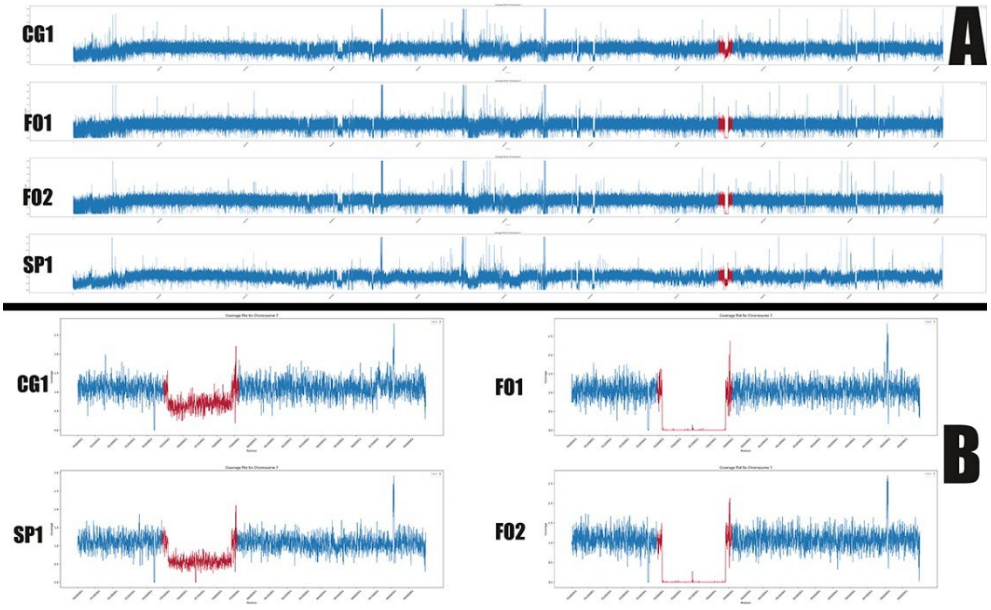
Average coverage in reads per locus shown for 3 chromosomes and 10 samples. The chromosomes are chosen to represent a size distribution. The X chromosome was selected to help determine the sex of samples for which that information was missing.

As seen in Table 5, coverage is consistent between chromosomes of a sample, with a notable exception of chromosome X, which is showing around half the average coverage for males, as expected. Chromosome 1 was chosen as a representative since it's the biggest, and chromosome 16 was chosen as the midpoint.

3.4 Low-coverage region in chromosome 5 hints at potential large deletion

While examining the coverage data, a standout feature was discovered in affected individuals a 0.5MB region of almost zero coverage spanning the same location in one of the chromosomes. This observation is depicted in more detail on Figure 5.

Figure 5. Average coverage for chromosome N



Average coverage for chromosome 5 for two controls – CG1 (pool of male Shetland ponies), SP1 (female Shetland pony), and the two affected horses FO1 (individual with polydactyly), FO2 (individual with polydactyly). Plot A shows average normalized coverage across the whole chromosome N. Image is used as illustration to show that coverage profile looks similar across individuals. An interesting region is marked in red and zoomed in further. Plot B shows only the region of interest that shows almost 0 coverage in the affected individuals and normal coverage in the control samples. The region is located in coordinates N:75500001-75900001.

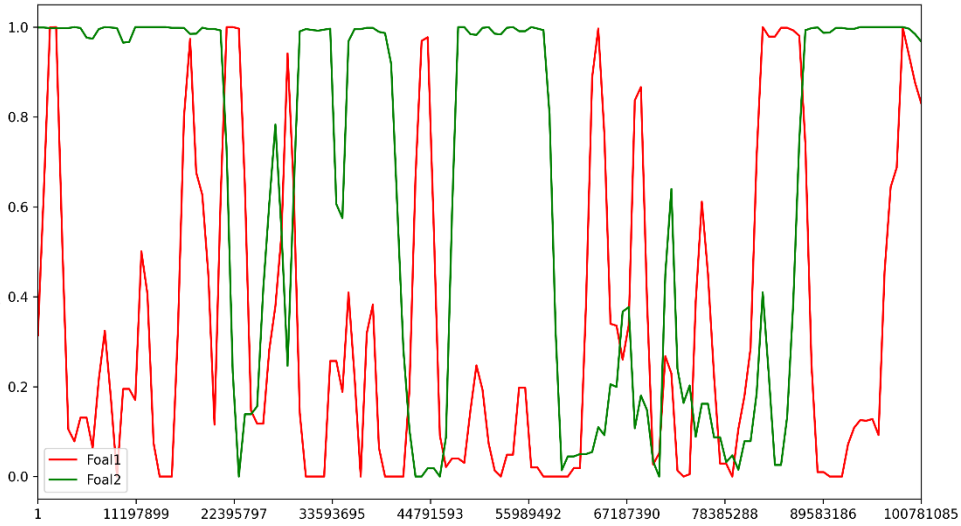
This region contains 28 genes that all belong to the olfactory receptor family, as well as a few olfactory pseudogenes. Interestingly the region is downstream of a gene that is involved with a developmental syndromic disease that features polydactyly as one of its symptoms in humans. The gene is around 1MB away from

the low-coverage region discussed here. No other large regions that show unusual coverage patterns were found. Upon further investigation of more control sequences, this deletion is present in all samples of the breed miniature Shetland pony (regardless of phenotype status), and none of the regular Shetland pony individuals, which diminishes the possibility of the mutation being the causative change for the investigated phenotype.

3.5 Run of homozygosity analysis

Because of the unique biological circumstances of sample FO2 (product of inbreeding of 2 suspected carriers), a run of homozygosity plot could potentially be used to identify the region of interest, since it comes once from the P generation and once from the F1 generation, so the F2 generation must carry the same "physical" piece of DNA, along with its markers. The results of that analysis are shown on Figure 4.

Figure 6. ROH plot for chromosome 7.

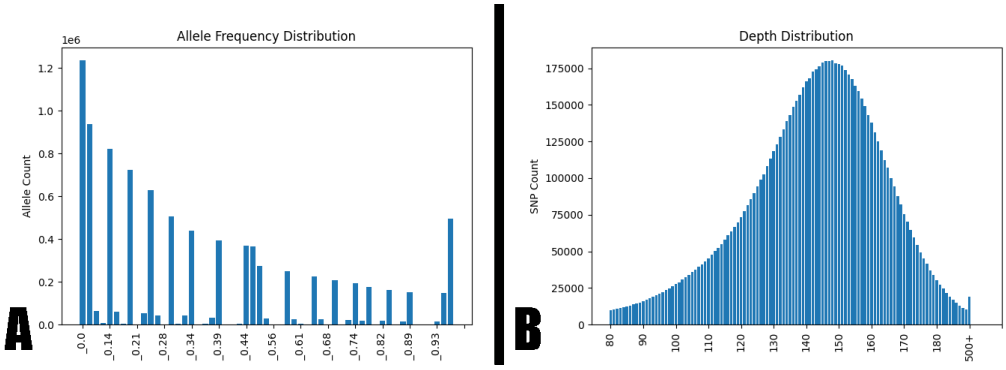


The graph shows relative homozygosity for a floating region of around 5000 SNP window with a half-size step across chromosome 7 of FO1 (red, not a product of inbreeding) and FO2 (green, inbred offspring). Big regions of the genome are completely homozygous for all called SNPs across all chromosomes (data not shown). Chromosome 7 was chosen at random.

3.6 Variant filtration of initial data

After alignment, variant calling and initial filtration, a total of 9 213 679 SNPs were called across 10 samples. Figure 5 shows some stats about the distribution of depth and frequency of alleles across the samples. 501 461 SNPs were removed when filtering for the standard minor allele frequency of 0.05.

Figure 7.SNP distribution statistics



The distribution of allele frequencies and depth of coverage per SNP across samples are shown. On Plot A the allele frequency distribution has some "gaps" at regular intervals, but those are most likely caused by the binning process. The gaps have fewer alleles so they don't display properly at this scale (millions versus tens of thousands for some bins). On Plot B the SNPs per depth follow a normal distribution, as expected. Values that contributed less than 0.1% of the total SNPs were cut off for clarity. The peak at 500+ depth is due to the inclusion of the mitochondrial genome in the initial SNP analysis.

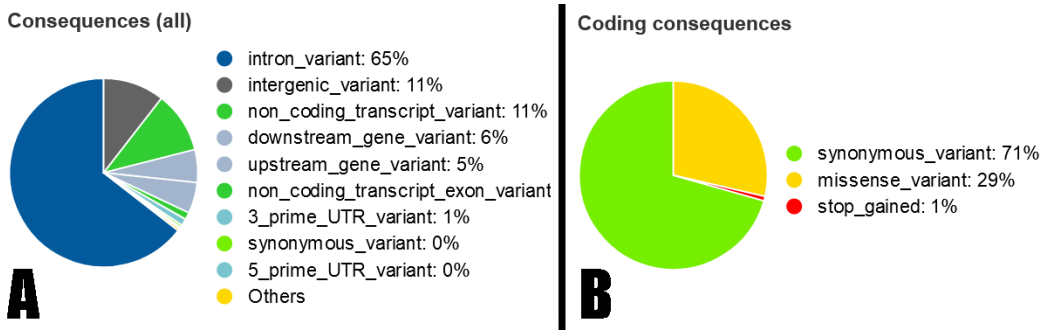
Since the analysis focuses on a small group of related individuals, the rules of inheritance can be used to further filter SNPs to only those that are biologically relevant. Since this type of polydactyly is severe, and pedigree analysis shows that all parents previously produced multiple healthy offspring, the causative mutation must be inherited recessively. Together with the family relationships of the 5 samples, the following SNP constraints can be placed:

- The allele must not be homozygous reference
- The allele must be homozygous in both foals
- Parents must be heterozygous for the same allele at that locus.
- The allele must not be present in a homozygous state in any of the control samples.

Additionally, a constraint that the coverage at that locus must be at least 8x for each included sample was placed to ensure that low coverage and random sequencing errors do not interfere with the selection process. The results of these constraints reduced the overall SNPs from more than 9 million to 20 268. Of those 20 255 were SNPs already added to the Ensembl database and 13 that were new to this sample set. SNPs that follow these criteria span 1794 genes which produce 5916 transcripts.

The categories these variants fall into are shown on Figure 8.

Figure 8. Distribution of SNP effects



This figure generated by Ensembl's Variant Effect Predictor tool shows all consequences that SNPs in the filtered list could have. This does not represent a guarantee that all these effects are occurring in the genome, since this needs to be manually confirmed by looking at the sequence of the regions of interest, but is a narrowed down list that can serve as a starting point for deeper investigation. Plot A shows the distribution of SNP consequences, with most variants landing in intergenic or intronic regions, as expected. Plot B shows the distribution of SNPs within known protein coding regions and their predicted effects.

Out of all these variants, the 13 novel SNPs land in 9 genes, 4 of which are protein coding. However, all of the SNPs are in intronic regions, and none of the genes are directly involved in development of the skeleton or the limbs. Using this list of 1794 potential candidate genes, further refinements can be made. For example, genes can be filtered by their Gene Ontology annotation, such as limb development or DNA binding activity. Results of this initial filtering are shown in table 7 on the next page. Regions may also be selected by correlating them with signals from other statistical methods, such as calculating the fixation index (Fst) between the two experimental "populations", namely individuals in the family (that share the mutant allele) and control groups (presumed to not be allele carriers). The results of that analysis will be shown in the next section.

Table 6. Genes per Gene Ontology label

Gene Ontology Label	**Count**
Bone Development (GO:0060348)	17
Nuclear Receptor Activity (GO:0004879)	7
Nuclear Receptor Binding (GO:0016922)	6
DNA Binding (GO:0003677)	107
Transcription Factor Binding (GO:0008134)	31
Limb Development (GO:0060173)	11
Limb Morphogenesis (GO:0035108)	9
DNA Binding TF Activity (GO:0003700)	64

To confirm that these results are not caused by an unaccounted methodical bias, 5 samples of randomly selected SNP sets were analysed through Ensembl's Variant Effect Predictor tool, then genes were filtered for the same Gene Ontology labels. The results are shown on Figure 9.

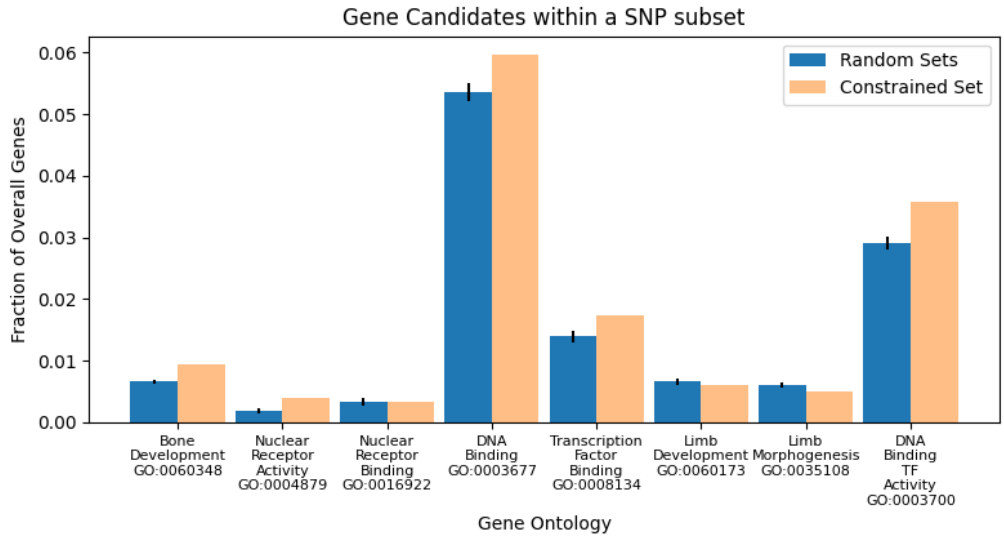


Figure 9. Distribution of SNP effects. Comparison of the set of genes left after imposing biological constraints on the SNP data versus 5 random sets of SNPs. Random SNP sets include a lot more genes, but the fraction of genes related to skeletal development is lower.

A random set of similar number of SNPs (around 20 000 in this case) are distributed among much greater number of genes, which results in a lot more genes with the desired Gene Ontology labels being included. However, when comparing the fraction of genes for each set, the one with imposed biological constraints about the inheritance pattern of the SNPs gives a significantly enriched fraction of genes related to skeletal development and transcription regulation (p value < 0.0001).

3.7 Fst analysis of genome-wide SNP data

Calculating the fixation index for the whole genome was done using vcftools and the results were plotted in R using qqman in linear mode (Turner 2018). The recommended window size was 100 000 bp, with a step of half that size, 50 000. The 10 samples were split into two populations of 5, one which includes all horses of the affected family, and another group that only contains phenotypically healthy individuals (assumed to not be carriers). Figure 10 shows the results of the analysis.

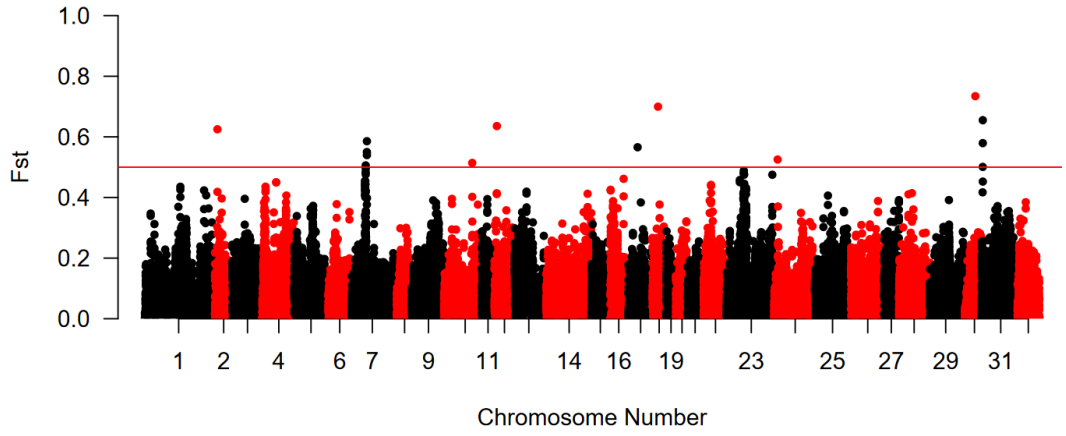


Figure 9. Fst analysis of SNP data. The Manhattan plot of Fst values for each window across all 32 chromosomes of the horse genome. The MT and Y scaffolds are not included in this plot. Genome-wide threshold of 0.5 (red horizontal line) was considered as marking a region significant. The Y axis represents the Fst index, where 0 is complete Hardy-Weinberg equilibrium, while 1 is complete fixation for different alleles between the two populations.

The calculated Fst value for the whole genome was 0.068922. This is very low and implies there is no difference between the populations (Charlesworth 1998), but this result is to be expected, since the two "populations" being compared are comprised of very few individuals from closely related breeds. Still, it is a check on the overall validity of the analysis, as high value of the F statistic would imply there might be something wrong with the dataset or the data processing. For example, the difference between the three lineages of ancient domestic horses (Alaskan, Iberian and Eurasian) is as low as 0.1 in some cases (Cieslak et al. 2010). High signals are observed on several chromosomes, with the only statistically significant peak seen on chromosome 7. All bins that overlap significant signals were then converted to a list of chromosome coordinates and those were used as an additional filter to cross-reference genes marked by the Variant Effect Predictor tool. There were only 8 regions that passed all the required thresholds stated previously, and they contain 19 genes. Highest Fst value for these regions was 0.65, and lowest was 0.50 (the cutoff).

4. Discussion

Short read sequencing and its analysis produces enormous amounts of data, as evidenced by the almost 2000 GB of intermediary data produced during the analysis. Some of it could be optimised, and some of it can be offloaded to temporary storage, but the reality is that it is a very computationally demanding process. The 50 analysed sequencing runs were all of very good quality, which made the downstream data processing easier, since a majority of the reads could be preserved and thus obtain higher confidence when drawing conclusions from the data. Anecdotally, the newer the deposition date was, the better the overall quality of the sequences, but this could be due to many reasons, including newer library preparation chemistry, improved machines or even higher proficiency of the researchers who prepare the DNA for sequencing.

One curious deviation that was observed was the skewed shape of the GC content distribution. Although the GC distribution varies across species, this deviation from the expected normal distribution shape was consistent across all samples tested, regardless of library preparation method, sequencing device or research institution that prepared the sample. While conferring with a colleague that works in a different breed of horses, they observed the same distribution, so it is not even breed-specific, but rather an oddity of the horse genome itself.

Mapping with bowtie2 produced very good results, with nearly 70% of the read pairs mapped uniquely, and overall alignment rate of 99% or higher. Several regions of the genome showed pairs mapping thousands of times, even after duplicate removal, but after manual checking several of those areas, they contained ribosomal genes. This explains the unique behaviour as ribosomal genes are present in high copy numbers (Dawid et al. 1978), and this interferes with the alignment process.

Another overall challenge was representing the data in a way that shows as much as possible while maintaining readability. Fortunately, this study tries to find differences between samples, rather than represent all of them, so a choice was made to show data that is indicative of the findings, and point out exceptions to the norm when present.

As the analysis progressed, several sanity checks of the data were performed along the way, for questions for which the answer was known. If the analysis would show a different outcome than base reality, then steps could be taken to rectify the

error, since this would have further downstream applications. One such example was already demonstrated previously when analysing the super-high coverage hotspots in the genome. This could have been an interesting signal, or a mistake in duplicate removal process. Finding out a plausible biological reason eliminates those concerns. Another analysis was estimating the sex of the samples based on average coverage. If everything went as expected, the average coverage across the X chromosome should be half that of the coverage for the rest of the chromosomes. Such was the case for all male samples, and none of the female samples, with slight deviations due to imperfect sequencing and alignment.

This sanity checks on coverage and alignment allowed me to be more confident when noticing the low-coverage area of chromosome 5. This could have been a sequencing or alignment error, but since these were made very unlikely based on the state of the rest of the samples, this appears to be a large-scale chromosomal alteration. However, this region only contains olfactory genes and related pseudogenes, so it is unlikely to be contributing to skeletal development.

Due to the unique nature of the sampled individuals – closely related family with a case of inbreeding – an opportunity presented itself to use homozygosity as a genomic region filter. Since the information provided by the owner state that all parents (P) had multiple healthy offspring before these two cases, the inheritance pattern of this mutation must be recessive. Additionally, due to the case of inbreeding there is information about both first generation (F1) and second generation (F2) offspring. The sample marked as FO2 is coming from the individual who's the product of inbreeding, and both parents (a P generation and a F1 generation horse) are phenotypically healthy. This means that the chromosomal region responsible for this mutation should be in a highly homozygous state, since it essentially comes twice from the same individual (SU1), once directly in a gamete, and once from a gamete of MK1. However, this did not produce the desired outcome, since the inbreeding event was very recent, and large regions of the genome were in a homozygous state. If this mutation was desirable and not deleterious, several outcrossing events and a few generations further would have normalized the heterozygosity across the rest of the genomic regions, making this a viable strategy to identify the associated trait locus.

The initial analysis produced more than 9 million SNPs across more than a million sites. Even eliminating sites with minor allele frequency of 0.05 or lower, that is still a considerable number. Fortunately, due to the known pedigree of this population, the rules of genetics and inheritance can be used to narrow down the number of SNPs to be investigated considerably. A single nucleotide polymorphism may not be the cause for a particular phenotype, but since the mechanisms of meiotic crossing over rearrange the genome in chunks, the closer a SNP is to the actual genetic cause, the more likely it is to segregate together. Additionally, since FO2 is born from inbreeding, the same chunk of DNA, together with its

polymorphism pattern, must be present on both copies of the respective chromosome. One of those copies must come from the parent, MK1, and the other from the father, SU1. And since both of them do not exhibit the studied phenotype, they must both be heterozygous at this particular locus. Lastly, if the locus in FO2 is homozygous reference, that site is also excluded. Coverage is also taken into consideration, since low coverage increases the chance that this variant could be the result of a technical error and not a true mutation.

Taking all of this into consideration reduces the number of sites to be considered dramatically, down to a bit over 20 000. Most of them were already existing in the Ensembl database, but 13 were novel to this sample set, making them prime candidates for investigation. The genes where these SNPs are found are 9 in total, of which 5 produce long non-coding RNAs with unknown functions.

The rest of the SNPs encompass almost 1800 genes, which can be further selected based on associated function using the Gene Ontology labels, as outlined in Table 6. This has some interesting candidates, which can be further checked, but so far these filters have been logic-based and while certainly helpful, are not backed up by statistics. The genome-wide Fst analysis of SNP data is the most powerful filtering tool that can be applied to this dataset. Establishing the fixation index for the whole "population" with vcftools gives a very low score of around 0.07, which is to be expected given that these are a small group of horses of the same breed. Much more informative is the plot of per-site Fst, shown on Figure 10.

Cross-referencing the list of genes that obey the inheritance rules (homozygous in offspring, heterozygous in parents, not homozygous reference or similar to the controls) with the regions that show a strong signal in the Fst analysis results in 0 genes being left. This could be due to several reasons. The SNPs may be less strongly associated with the mutation than assumed, so further analysis that looks at haplotypes could reveal a better association between phenotype and genotype. Fst analysis can also be inconclusive, since the sample size is really small. Most studies that performed Fst or GWAS had at minimum 10 times the number of samples. The fact that the sample size might be insufficient can also be found in the raw Fst data. The values should range from 0 (no difference) to 1 (completely fixed for different alleles), and negative values indicate a statistical fluke. More than 15% of all entries in the input data were negative, while the consensus is that this should rarely happen if ever.

Another possibility is that the mutation is due to a large structural change, such as deletion, duplication, inversion or translocation. Such differences are hard to detect with traditional SNP analysis. Even though the deletion in chromosome 5 showed up on the Fst plot, that was more of a coincidence, since choosing different parameters erases that signal in some of the analyses I did. It is possible to detect chromosomal structural changes with GATK, but these are a very broad category, and the pipeline that is made to process them is outside of the scope of the time

allocated for this work. Despite that, structural variation is an interesting possibility to explore, since some of the control samples used in this thesis were taken from a work that discovered exactly such an underlying mutation cause.

Searching for the causative mutation for any phenotype will always be like looking for a needle in a haystack. If we expand the metaphor, statistics is then like using a magnet to aid the search, a very powerful, albeit not omnipotent tool. Unfortunately, it is entirely possible that due to the low number of samples that magnet is not particularly strong, making the search for the causative mutation for this rare type of polydactyly a yet unresolved task.

5. References

Canzar, S. & Salzberg, S.L. (2017). Short Read Mapping: An Algorithmic Tour. *Proceedings of the IEEE*, 105 (3), 436–458. https://doi.org/10.1109/JPROC.2015.2455551

Charlesworth, B. (1998). Measures of divergence between populations and the effect of forces that reduce variability. *Molecular Biology and Evolution*, 15 (5), 538–543. https://doi.org/10.1093/oxfordjournals.molbev.a025953

Cieslak, M., Pruvost, M., Benecke, N., Hofreiter, M., Morales, A., Reissmann, M. & Ludwig, A. (2010). Origin and History of Mitochondrial DNA Lineages in Domestic Horses. *PLOS ONE*, 5 (12), e15311. https://doi.org/10.1371/journal.pone.0015311

Coates, M.I. (1994). The origin of vertebrate limbs. *Development (Cambridge, England). Supplement*, 169–180

Dawid, I.B., Wellauer, P.K. & Long, E.O. (1978). Ribosomal DNA in Drosophila melanogaster: I. Isolation and characterization of cloned fragments. *Journal of Molecular Biology*, 126 (4), 749–768. https://doi.org/10.1016/0022-2836(78)90018-9

Duboule, D. (1998). Vertebrate Hox gene regulation: clustering and/or colinearity? *Current Opinion in Genetics & Development*, 8 (5), 514–518. https://doi.org/10.1016/S0959-437X(98)80004-X

Frischknecht, M., Jagannathan, V., Plattet, P., Neuditschko, M., Signer-Hasler, H., Bachmann, I., Pacholewska, A., Drögemüller, C., Dietschi, E., Flury, C., Rieder, S. & Leeb, T. (2015). A Non-Synonymous HMGA2 Variant Decreases Height in Shetland Ponies and Other Small Horses. *PLOS ONE*, 10 (10), e0140749. https://doi.org/10.1371/journal.pone.0140749

Gehring, W.J. (1993). Exploring the homeobox. *Gene*, 135 (1), 215–221. https://doi.org/10.1016/0378-1119(93)90068-E

Lettice, L.A., Heaney, S.J.H., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E. & de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics*, 12 (14), 1725–1735. https://doi.org/10.1093/hmg/ddg180

Mahmood, R., Bresnick, J., Hornbruch, A., Mahony, C., Morton, N., Colquhoun, K., Martin, P., Lumsden, A., Dickson, C. & Mason, I. (1995). A role for FGF-8 in the initiation and maintenance of vertebrate limb bud outgrowth. *Current Biology*, 5 (7), 797–806. https://doi.org/10.1016/S0960-9822(95)00157-6

McGlinn, E. & Tabin, C.J. (2006). Mechanistic insight into how Shh patterns the vertebrate limb. *Current Opinion in Genetics & Development*, 16 (4), 426–432. https://doi.org/10.1016/j.gde.2006.06.013

Metzger, J., Gast, A.C., Schrimpf, R., Rau, J., Eikelberg, D., Beineke, A., Hellige, M. & Distl, O. (2017). Whole-genome sequencing reveals a potential causal mutation for dwarfism in the Miniature Shetland pony. *Mammalian Genome*, 28 (3), 143–151. https://doi.org/10.1007/s00335-016-9673-4

Muneoka, K. & Bryant, S.V. (1984). Cellular contribution to supernumerary limbs resulting from the interaction between developing and regenerating tissues in the

axolotl. *Developmental Biology*, 105 (1), 179–187. https://doi.org/10.1016/0012-1606(84)90273-2

Rafati, N., Andersson, L.S., Mikko, S., Feng, C., Raudsepp, T., Pettersson, J., Janecka, J., Wattle, O., Ameur, A., Thyreen, G., Eberth, J., Huddleston, J., Malig, M., Bailey, E., Eichler, E.E., Dalin, G., Chowdary, B., Andersson, L., Lindgren, G. & Rubin, C.-J. (2016). Large Deletions at the SHOX Locus in the Pseudoautosomal Region Are Associated with Skeletal Atavism in Shetland Ponies. *G3 Genes\textbarGenomes\textbarGenetics*, 6 (7), 2213–2223. https://doi.org/10.1534/g3.116.029645

Small, K.M. & Potter, S.S. (1993). Homeotic transformations and limb defects in Hox A11 mutant mice. *Genes & Development*, 7 (12a), 2318–2328. https://doi.org/10.1101/gad.7.12a.2318

Summerbell, D. (1979). The zone of polarizing activity: evidence for a role in normal chick limb morphogenesis. *Development*, 50 (1), 217–233. https://doi.org/10.1242/dev.50.1.217

Takeuchi, J.K., Koshiba-Takeuchi, K., Matsumoto, K., Vogel-Höpker, A., Naitoh-Matsuo, M., Ogura, K., Takahashi, N., Yasuda, K. & Ogura, T. (1999). Tbx5 and Tbx4 genes determine the wing/leg identity of limb buds. *Nature*, 398 (6730), 810–814. https://doi.org/10.1038/19762

Turner, S.D. (2018). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *Journal of Open Source Software*, 3 (25), 731. https://doi.org/10.21105/joss.00731

Umair, M., Ahmad, F., Bilal, M., Ahmad, W. & Alfadhel, M. (2018). Clinical Genetics of Polydactyly: An Updated Review. *Frontiers in Genetics*, 9. https://www.frontiersin.org/articles/10.3389/fgene.2018.00447

6. Popular science summary

Polydactyly (from Greek, meaning many fingers) is a condition in which an organism has more fingers than typical. For humans that would be six or more, while horses normally have only one finger. For most cases of this condition, the inheritance is dominant, meaning that a parent would pass this trait to all of its offspring. However, in the unique case present in this study, the mutation is inherited recessively, meaning both parents had to have been carriers, in order for the offspring to be affected.

Organisms of the same species share an overwhelming similarity in their DNA. For example, people can look very different from each other, but on the sequence level that's caused by less than 0.1% difference. In the case of domesticated animals, like horses, those genetic differences are even smaller, due to selective breeding. Even then, when a small percentage is multiplied by the billions of DNA letters in a genome, the number becomes unmanageable for anyone to analyse by hand.

By employing supercomputers and advanced algorithms, many genomes from many individuals can be compared DNA base by base. The data can then be analysed statistically, so that random mutations can be averaged out. This leaves only spots in the genome that are significantly different between groups of individuals – for example horses with polydactyly compared to those without. That way the potential list of causes is reduced to a comprehensible level, hopefully leading to discovering what DNA mutation causes this unique change in the limb of the miniature Shetland pony horses being studied.

Acknowledgements

I would like to thank my supervisor Gabriella Lindgren for providing me the opportunity to work on such an interesting project, as well as helping me on this complicated, but interesting journey. I'm also grateful to Rakan Naboulsi for the many bioinformatics questions answered throughout the thesis.

Appendix 1

This section outlines in detail the workflow and the commands used to obtain the data presented in this work.

6.1 Data analysis steps and workflow: Part One

The following paragraph and all sub-paragraphs will outline the workflow I followed to arrive at my current results. For all intents and purposes this is not a ready to be executed pipeline script, but with some minor adjustments of variables and parameters it could be used as such.

6.1.1 Load software packages used in the workflow

Below is a list of CLI packages used during the workflow. The version of each package is also specified, when possible, although in theory newer versions of the tools should produce the same outcome. UPPMAX uses lmod to dynamically load different packages and adjust the $PATH, so the first step is to specify what is going to be used.

```
### load modules

module load bioinfo-tools

module load sratools/3.0.0

module load trimmomatic/0.39

module load FastQC/0.11.9

module load MultiQC/1.12

module load samtools/1.17

vcftools/0.1.16

bcftools/1.17
```

```
module load bowtie2/2.4.5

module load picard/2.27.5

module load GATK/4.3.0.0
```

6.1.2 Download sequences to local storage

Sequence files that were provided by the thesis supervisor for the purposes of the thesis work were already present in the local storage of the project. Addidional sequences need to be downloaded from a dedicated server, such as NCBI's Sequence Read Archive.

Sample names may contain multiple SRA entries, and all of those need to be downloaded (and later on combined) to create a complete WGS.

```
### download SRA entry to current directory.

srr=ERR868004

fasterq-dump -x -e 10 --split-files --skip-technical -t . -O . $srr
```

This code will use 10 threads to download the sequencing run onto the local storage in the form of $srr.1.fastq and $srr.2.fastq FASTQ files, in the case of a paired-end sequencing run.

6.1.3 Trimming of low-quality reads

Illumina sequencing creates millions or even billions of short reads, and no matter how good the library preparation or the sequencing process is, some reads will be faulty, which can compromise the downstream analysis. This step aims to remove reads that don't follow generally accepted quality standards.

```
### trim low quality sequences

r1=$srr_1.fastq

r2=$srr_2.fastq

trimmomatic PE -phred33 -threads 10 $r1 $r2
-trimlog ./$srr.log -baseout $srr.fq.gz
ILLUMINACLIP:$TRIMMOMATIC_ROOT/adapters/TruSeq3-PE.fa:2:30:10
SLIDINGWINDOW:5:20 MINLEN:80

rm $r1 $r2
```

This code will use 10 threads to trim PHRED33-based FASTQ file, using supplied known sequences of Illumina adapters, stopping when the quality in a sliding

window of 5 nt drops below 20, and discarding all reads that are shorter than 80bp, regardless of quality. It will output 4 files containing reads from 4 different categories – forward paired, reverse paired, forward unpaired and reverse unpaired. Despite reads always coming in pairs, quality control may delete one member of the pair, in which the read is sorted into the unpaired file. Those will be discarded for consequtive analyses.

6.1.4 Manual quality control

Reads sorted into the 4 files from the previous step can be checked for various parameters before proceeding with the downstream analysis.

FastQC is a useful tool that calculates various statistics for the sequencing run and outputs them in a handy visual reqpesentation in the form of a webpage.

MultiQC aggregates several quality-control files from tools like FastQC and gives a broader overview of the whole batch of samples.

```
### Quality control of reads

for file in *.fq.gz; do

  fastqc $file

done

multiqc .
```

This code block will fun FastQC for each of the files generated in the previous step and output a report for each. MultiQC will aggregate all these reports into one final file, ready for downloading and viewing.

6.1.5 Downloading the reference genome

My thesis work is seeking to find the location of a marker that strongly associates with the observed mutant phenotype in a family of horses, so the required reference genome is that of the horse, *Equus cabalus*. The version as of the time of writing of this work is EquCab3.0 (GCA_002863925.1).

First step is to download the genome on the local storage, which can be done with the bash utility of wget:

```
### download the genome

wget ftp.linktogenome.fa.gz
```

Note that this is not the actual real link to the genome, as links are very long and are subject to change, so in the interest of formatting it was omitted.

6.1.6 Preparing the reference genome for sequence alignment

After the genome has been downloaded to local storage, the file needs to be converted to be more easily accessible during the sequence alignment step. Bowtie2, the tool that will be used for alignment can also do the reference indexing. This step needs to be performed only once per species analyzed.

```
### reference building

bowtie2-build
--threads 10 -o 3
EquCab3.0.dna.toplevel.fa.gz
reference/ref
```

6.1.7 Sequence alignment to reference genome

Now that the reference index has been created, the millions of quality-controlled reads can be mapped to the horse genome. Bowtie2 is among the many similar tools that can do that, but this is the one I chose based on the characteristics of the sequencing run and the performance of the alignment implementation.

```
### bowtie2 sequence alignment

t1=$srr_1P.fq.gz

t2=$srr_2P.fq.gz

ref=./reference/ref

bowtie2 -p 10 --very-sensitive-local -x $ref -1 $t1 -2 $t2 |
samtools view -@ 10 -bS -o $srr.bam

rm $t1 $t2
```

This code will use 10 threads to attempt local alignment of the reads using the provided reference genome, the two reads files (forwards and reverse paired). Since the tool outputs a text file in the SAM format, which is very large and inefficient, the output is directly converted to BAM format. All further downstream analyses will use BAM files.

6.1.8 Preparation of alignment for downstream analysis

The pipeline of choice GATK has several requirements for the format of files that go into it. This step will sort the individual reads according to their alignment coordinates, will add a read group tag to distinguish samples that come from the

same individual, but were ran on different lanes or chips, and will remove reads that are deemed technical replicates, so that they don't skew further analyses.

```
samtools sort $srr.bam -o $srr.sorted.bam

java -jar $PICARD_ROOT/picard.jar AddOrReplaceReadGroups -I $srr.sorted.bam -O $srr.sorted.rg.bam $rg

java -jar $PICARD_ROOT/picard.jar MarkDuplicates -REMOVE_DUPLICATES true -I $srr.sorted.rg.bam -O $srr.dedup.bam -M $srr.metrics.txt

samtools index $srr.dedup.bam
```

This code stores the read group text into a variable $rg, since it has several fields. In addition, this enables a different read group to be automatically assigned to each sequencing run by modifying the variable's contents accordingly.

6.1.9 Optional preparation of multi-part samples

This step is optional and needs to be performed only for some reads. If a sequencing of a single sample has been performed on multiple lanes of an Illumina flow cell, those samples need to be merged into a single file, since they represent information from the same individual, and treating them as independent data points would affect analyses and the conclusions that are drawn from them.

```
find -type f -name $srr.dedup.bam > $srr.txt

samtools merge -@10 -c -p -o $srr.merged.bam -b $srr.txt

rm $srr.txt

samtools sort -@10 $srr.merged.bam -o $srr.sorted.bam

samtools index -@10 $srr.sorted.bam -o $srr.sorted.bam.bai
```

This code first uses a bash command to create a list of all BAM files that have the same accession number, then passes that list to samtools to be used as input when merging. After that, sorting and merging is performed and the output files are ready to be fed into the GATK pipeline, which will be discussed in the next paragraph.

6.2 Data analysis steps and workflow: Part Two

The second part of the data analysis involves mainly tools from the GATK pipeline. This part will describe the file preparation, file analysis, variant calling procedure and final variant filtration criteria before a file containing DNA polymorphisms is obtained.

6.2.1 Sequencing coverage metrics

Before proceeding, average coverage depth needs to be calculated as a general quality control. Variant calling relies on the fact that if many reads show a different DNA sequence than what is in the reference genome, there is a high confidence that this is not a sequencing error but a real mutation. So having too low coverage (usually under 10x across the chromosome/genome) places doubt on the obtained SNP detection.

```
gatk DepthOfCoverage -R $gen -O ./coverage -I $srr.sorted.bam --intervals 1
```

This code uses the horse genome ($gen) as reference and calculates per-locus coverage for the given interval (in this case chromosome 1) and the given sample ($srr.sorted.bam). It can also calculate coverage for many or all chromosomes, and multiple samples at the same time. Since the process is single-threaded and takes a long time, a possible acceleration would be to calculate the depth of coverage per chromosome simultaneously, which would require 32 separate scripts to be ran in parallel.

6.2.2 Indexing of reference files

GATK recommends performing base quality score recalibration of the BAM files in order to trim low quality base scores (resulting from technical biases during sequencing). This is necessary so that the confidence that a base is correctly identified is paramount to deciding if a SNP should be called further downstream.

Since working with text files is slow, GATK first needs to index both the reference genome of the horse, as well as the VCF file containing all known variants in the horse genome. Both of these files can be obtained from Ensembl or NCBI, and I chose to obtain them from Ensembl (this matters because chromosomes have different naming conventions across the two databases).

```
gatk IndexFeatureFile -I $vcf

samtools faidx $dna

java -jar $PICARD_ROOT/picard.jar CreateSequenceDictionary -R $dna
```

This code first indexes the file containing the known horse variants ($vcf variable). It then uses samtools to index the reference genome ($dna variable), and finally creates a sequence dictionary of the genome using picard.

6.2.3 Base Quality Score Recalibration

The next step is the base quality score recalibration of the pre-processed BAM file, which will yield an analysis-ready alignment. Variant calling or other analyses can be performed on the final output.

```
gatk BaseRecalibrator -I $srr.sorted.bam -R $gen -O $srr.recalibration.table --known-sites $vcf

gatk ApplyBQSR -I $srr.sorted.bam -R $gen --bqsr-recal-file $srr.recalibration.table -O $srr.bqsr.bam
```

This code first uses the BaseRecalibrator tool to create the recalibration table, and then applies it on the next step. The final output ($srr.bqsr.bam) is the file that can be kept in local storage for future analyses or various quality controls, or if the storage is limited, can also be deleted after gVCF files are obtained.

6.2.4 Genomic Variant Calling

The next step will produce a gVCF file, which contains records for every base in the genome, even if no polymorphisms are present. The file is large and contains a lot of extra information, but is a necessary step before obtaining a final variant call file which can be datamined.

```
gatk HaplotypeCaller -ERC GVCF -R $gen -I $srr.bqsr.bam -O $srr.raw.variants.vcf
```

This code takes the base quality score recalibrated BAM file obtained in the previous steps and outputs a genomic VCF file to be used further downstream. The process is long and slow, but needs to be performed only once. The output file can be kept in local storage so that when further samples are added and processed, it can be used for joint genotyping again.

6.2.5 Merge gVCF files for joint genotyping

At this point each sequencing run has been reduced to a file containing all detected variants. Each file contains one sample. Merging them will allow inferences to be drawn for the whole sample set, as well as simplifies the process if more samples need to be added at a later point in time. The new gVCF file can simply be merged into the existing sample group, and then variant calling needs to be performed again on the new output file.

```
find . -type f -name \*.raw.g.vcf | sort -n > $name.gvcf.list
```

```
gatk CombineGVCFs --java-options "-Xmx10g" -R \$gen --variant $name.gvcf.list -O
combined.$name.g.vcf.gz
```

6.2.6 Cohort genotyping and variant calling

The final step is to perform joint genotyping on the set of samples, resulting in a VCF file that contains unfiltered variants (SNP and indel), as well as a lot of additional information such as quality, depth, allele frequencies, genotype likelihood and so on. These are crucial when filtering variants later on and datamining.

```
gatk GenotypeGVCFs -R $gen -V combined.$name.g.vcf.gz
-O cohort.$name.vcf.gz
```

The code takes as input the horse reference genome ($gen) and the combined gVCF file and outputs a VCF file that can be filtered based on certain criteria or used as is for data analysis.

6.2.7 Separate SNPs and InDels

While indels are just as important as SNPs when it comes to biological consequences, some tools struggle when analysing them, so for the purposes of downstream statistics, SNPs and InDels will be split and filtered separately, then only SNPs will be used to calculate genome association. However InDels will be considered when elucidating biological consequences in potential regions of interest.

```
java -jar $PICARD_ROOT/picard.jar SplitVcfs -I cohort.$name.vcf.gz
-SNP_OUTPUT $name.snp.vcf.gz
-INDEL_OUTPUT $name.indel.vcf.gz
--STRICT false
```

6.2.8 SNP variant filtration

Part of the SNP filtration code has been adapted from (https://raw.githubusercontent.com/kpatel427/YouTubeTutorials/main/variant_filtering_annotation.sh).

```
gatk VariantFiltration \

        -R $gen \

        -V $name.snp.vcf.gz \

        -O $name.filtered.snps.vcf \
```

```
        -filter-name "QD_filter" -filter "QD < 2.0" \
        -filter-name "FS_filter" -filter "FS > 60.0" \
        -filter-name "MQ_filter" -filter "MQ < 40.0" \
        -filter-name "SOR_filter" -filter "SOR > 4.0" \
        -filter-name "MQRankSum_filter" -filter "MQRankSum < -12.5" \
        -filter-name "ReadPosRankSum_filter" -filter "ReadPosRankSum < -8.0" \
        -genotype-filter-expression "DP < 10" \
        -genotype-filter-name "DP_filter" \
        -genotype-filter-expression "GQ < 10" \
        -genotype-filter-name "GQ_filter"
gatk SelectVariants \
        --exclude-filtered \
        -V $name.filtered.snps.vcf \
        -O $name.qc.snps.vcf
```

6.2.9 InDel variant filtration

Part of the InDel filtration code has been adapted from (https://raw.githubusercontent.com/kpatel427/YouTubeTutorials/main/variant_filtering_annotation.sh).

```
gatk VariantFiltration \
        -R $gen \
        -V $name.indel.vcf.gz \
        -O $name.filtered.indel.vcf \
        -filter-name "QD_filter" -filter "QD < 2.0" \
        -filter-name "FS_filter" -filter "FS > 200.0" \
        -filter-name "SOR_filter" -filter "SOR > 10.0" \
        -genotype-filter-expression "DP < 10" \
```

```
        -genotype-filter-name "DP_filter" \
        -genotype-filter-expression "GQ < 10" \
        -genotype-filter-name "GQ_filter"
gatk SelectVariants \
        --exclude-filtered \
        -V $name.filtered.indel.vcf \
        -O $name.qc.indel.vcf
```

6.3 Data mining and data visualization

6.3.1 Variant selection

Only certain variants make biological sense to be chosen for further analysis. This initial filtration was done using a simple Python script that compares if the variants follow certain criteria and deletes the line in the VCF file if they don't.

6.3.2 Homozygosity plots

The homozygosity of genomic regions was calculated using the tool bcftools, and then converted into a plot using matplotlib in Python.

```
bcftools roh -I -s sample_name -o roh_data.txt sample_name.vcf
```

6.3.3 Variant Effect Prediction

Even with several rounds of filtration there are tens of thousands of SNPs left to consider. Potential SNPs with high impact can further be selected using Ensembl's Variant Effect Predictor tool. This can then be converted into a list of candidate genes and loci to examine further downstream.

6.3.4 Fst Analysis

The fixation index statistic is an important metric when considering association of traits with a genomic location. The tool vcftools can be used to calculate Fst index over a specified region and output it to a text file that can be visualised. Higher values imply a difference in that particular region between compared populations.

```
vcftools --vcf input.vcf --weir-fst-pop samples.txt --weir-fst-pop controls.txt
--fst-window-size 100000 --fst-window-step 50000 --out results.fst
```

6.3.5 Depth of coverage plots

Depth of coverage is a useful metric to get an overview of the average outcome of the sequencing run alignment to the reference genome. Using the output of the GATK tool DepthOfCoverage (outlined in 2.7.1), which outputs depth of coverage for each position in a given interval, a simple Python script was made that calculates the floating average in a 1000bp interval, normalizes it against the calculated by GATK average coverage, then plots it on a per-chromosome basis.

6.3.6 Selecting genes of interest

Genes of interest can be selected with another Ensembl tool, BioMart. This allows genes to be filtered based on a variety of criteria, so that they can be narrowed down to a subset that is most likely of interest to the current research.

6.3.7 Diagram drawing

Some graphs and diagrams have been generated using the online tool draw.io found at https://app.diagrams.net/.

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, the checked box will be applied to all authors. You will find a link to SLU's publishing agreement here:

- https://libanswers.slu.se/en/faq/228318.

☒ 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.