Coursework Declaration and Feedback Form

The Student should complete and sign this part

Student Number: 2440230	Student Name: Qikai Yang
Programme of Study (e.g. MSc in Electronics and Electrical Engineering): MSc in Computer System Engineering	
Course Code: ENG5059P	Course Name: MSc Project
Name of **First** Supervisor: Dr Umer Zeeshan Ijaz	Name of **Second** Supervisor: Dr Anubhab Khan
Title of Project: Comparative genomics of known horse genomes	

Declaration of Originality and Submission Information

I affirm that this submission is all my own work in accordance with the University of Glasgow Regulations and the School of Engineering requirements Signed (Student) : Qikai Yang	E N G 5 0 5 9 P
Date of Submission :	

Feedback from Lecturer to Student – to be completed by Lecturer or Demonstrator	
Grade Awarded: Feedback (as appropriate to the coursework which was assessed):	
Lecturer/Demonstrator:	Date returned to the Teaching Office:

Comparative genomics of known horse genomes

MSc Computer System Engineering

Student Name: Qikai Yang

Student ID: 2440230y

Supervisors: Dr Umer Zeeshan Ijaz

Dr Anubhab Khan

Dr Barbara Mable

Dr Ciara Keating

August 27,2021

A report submitted in partial fulfillment of the requirements

for Computer System Engineering Degree

at the University of Glasgow

Acknowledgements

First of all, I would like to thank the two supervisors Dr. Umer Zeeshan Ijaz and Dr. Anubhab Khan for their great help and guidance in this project. In addition, there are also the workflow guide provided by Dr. Barbara Mable and the management of project promotion by Dr. Ciara Keating. It was their pioneering work and exploration that made it possible for my project to be completed without much unnecessary trouble.

Next, the author needs to thank the University of Glasgow for providing the platform and the opportunity to further research knowledge.

In addition, the author also needs to thank teachers, classmates and parents for their care and encouragement during the project. It is because of this strength that the research was successfully completed.

Abstract

This paper mainly studies to achieve 87 breeds of biological sequence matching to EquCab3.0 horse reference genome assembly method for bioinformatics database after study the history population size drawing of the 25 kinds of breeds and analysis is made on the three variants in the genome, like single nucleotide variants, small insert/deletions variation and large structural variation (Jagannathan et al., 2019). Trim-Galore (Lindgreen, 2012) will be used for trimming the reads and the trimmed reads will be mapped to the EquCab3 genome assembly using BWA (H. Li & Durbin, 2009). Duplicates will be removed using samtools (H. Li et al., 2009) and store the read alignment of duplicate markers in BAM format. Pairwise Sequentially Markovian Coalescent (PSMC) provides a historical map of population size changes from approximately 2 million to 10,000 years ago. And found that climate change can affect population size. Similar trend of population size change for Exmoor Pony is Noriker, German Riding Pony and Icelandic. Sequence variants will be called using Strelka variant caller (Saunders et al., 2012). A total of 19,578,627 SNPs sites were found by counting the variation of each species. However, 1012,116 SNPs sites were obtained by integrating 26 species and then filtering them.

Key Words: Horse, History population size, Trim-Galore, EquCab3, bwa-mem, Samtools, PSMC, SNP.

Table of Conetent

1. Introduction

1.1 Background

Horses not only occupy a special place among farm animals (Raudsepp, Finno, Bellone, & Petersen, 2019), but also participate in the Mammalian Genome Project as a representative of the perissodactyla purpose (Jagannathan et al., 2019). Furthermore, horses have been domesticated by humans since around 3500 BC and played an important role in transportation, communication and warfare in prehistoric times (Outram et al., 2009). In addition, although horses have contributed to the food and production industry in modern times, their strong economic impact has come from the development of sports and recreation (Chowdhary & Bailey, 2003). This means that the development of horse health and breeding is one of the core issues in the industry, because the results of this research will help to better understand the causes of disease, and to develop scientific methods for diagnosis and treatment, which will ultimately produce immediate and long-term economic benefits. All kinds of organisms have their own unique genetic information. To understand the formation process of organisms and all kinds of life activities deeply is inseparable from the research results of genome. Similarly, a strong interest and value in equine biomedicine, evolutionary mapping and basic science has led to an organized study of the equine genome (Raudsepp et al., 2019).

According to the Food and Agriculture Organization of the United Nations, as of 2015, there were 905 stallions in the world (Scherf & Pilling, 2015). Horses have influenced the political and economic trajectory of human society at the same time that human activity has had a huge impact on horses, through the thought that selection has created hundreds of domestic horse breeds while driving all wild populations close to extinction (Librado et al., 2016).

Although closely related to humans, several aspects of domestication remain controversial. First, whether domestication of domesticated horses originated from a single source, that is, horses were domesticated at the Botai site in modern-day Kazakhstan 5,500 years ago (Outram et al., 2009). Or is there more time and address of origin than that? Secondly, of all the domesticated species in existence, which are more closely related to their common ancestor? In addition, which breeds have contributed greatly to the formation of so many horse breeds in the world today? These are things to consider in future research.

From a genetic point of view, the update of EquCab3.0 reference library greatly improves the quality of the genomic reference assembly. The contiguity of the new assembly increased by about 40 times and the number of gaps decreased by about 10 times (Raudsepp et al., 2019). This means that advances in sequencing technology provide a solid basis for in-depth understanding of single-nucleotide variations in the equine genome and for exploring changes in horse population size over history.

Advances in genome sequencing technology and significant reductions in the cost of sequencing, as well as combinatorial sequencing strategies that allow simultaneous sequencing of multiple samples in a reaction pool (Mardis, 2008), have made it possible to perform large scale, high throughput and high depth whole genome sequencing. As a result, more and more organisms are getting their whole genomes sequenced, and many that have already been mapped are being resequenced in large numbers to actually study the genomes of populations of that species. These studies reveal the characteristics and laws of heredity, evolution, domestication and selection of organisms, and provide a deeper understanding of the evolutionary history of population origin, diffusion, population dynamics and adaptive differentiation, as well as the genetic and molecular mechanisms of species' traits.

Single nucleotide polymorphism (SNPS) refers to the change of a single base on different chromosomes of different individuals of the same species, which is mainly manifested as the variation at the nucleotide level of the genome and eventually leads to DNA sequence polymorphism, including the change of base conversion, single base insertion or single base deletion (Wang et al., 1998). SNP is the nucleotide difference at a specific position in different individuals at the genome level. SNPS usually have biallelic polymorphisms, which can detect variations in a single base and determine that the gene polymorphism is a stable mutation, more useful than previous molecular markers. Currently, SNP marker is an important method to study functional gene polymorphism and is also the preferred marker to establish genetic linkage map for quantitative trait genome mapping.

Most of the first population genome studies focused on the evolutionary history of human populations (Shriner, Tekola-Ayele, Adeyemo, & Rotimi, 2018). With the development of animal population genomics research, it has become an important frontier field of animal genetics and animal breeding to reveal the evolutionary relationship between species and explore the excellent genetic resources of various species. Many studies of animals that are relevant to human history and present life have also been reported, the analysis of the genomes of these animal populations, combined with relevant archaeological data, can not only help us understand their evolutionary history, but also provide great help to understand the migration in the process of human evolution, the domestication of wild animals, and the origin and development of agriculture and animal husbandry.

1.2 Related research Progress

Assembled the first genome of a domesticated horse in 2009 (Wade et al.,

2009), marking the first milestone in horse genomics research. The ABI 3730 sequencer was used to shotgun sequence the entire genome of a thoroughbred horse with a genome size of approximately 2.689Gb and a coverage rate of 6.8x. But because the samples were female horses, the horse Y genome was not assembled. The assembly and publication of the genome provided a model for subsequent studies of related aspects of the equine genome and became the first relatively complete equine reference genome. Based on the assembly of this genome, many equine genomes have been developed.

Equine reference genomes represented by digital mitochondrial DNA with low bias were examined for 82 mitochondrial pseudogenes in 2010. (Nergadze et al., 2010). Mitochondrial pseudogenes are derived from nucleotide sequences derived from the combination of mitochondrial DNA with nuclear genomes. They were randomly inserted into random parts of the horse's nuclear genome. In contrast to humans, insertional polymorphisms in equine mitochondrial pseudogenes are very frequent in equine populations, proving the hypothesis that species were quickly evolving.

The continuous development of sequencing technology has not only greatly accelerated the research progress of modern horse genome, but also provided new solutions for the research of ancient DNA. The first single molecule of ancient DNA was sequenced in 2011 (Orlando et al., 2011). However, in the process of exploring the molecular richness of ancient DNA, the sequencing technology introduced the concept of bias repair, which slowed down the efficiency of the sequencing process.

In 2012, DNA hybridization microarray chips for horses were developed by a consortium of research institutions in the United States, Australia, Germany, France, Japan, Ireland, Switzerland, Sweden, Norway and the United Kingdom (McCue et al., 2012). There are more than 54,000 polymorphic SNPS on the

chip. The chip was evaluated with sample sets representing 14 domestic horse breeds and 18 evolutionarily related species. It was also proved that other odactyla species can be genotyped with this microarray. In 2013, horse chips were able to genotype 670K OF SNPs scattered throughout the genome (Petersen et al., 2013). This is another milestone in the history of equine genome research. This high-quality SNP genotyping resource will have wider use in the analysis of various equine genomes and related species.

In 2012, a female quarter horse genome was the first to be assembled after a 2009 pedigree genome (Doan, Cohen, Sawyer, et al., 2012). This time, the researchers used second-generation sequencing technology to obtain 59.6Gb of quarter horse DNA sequence with an average coverage of 24.7× and discovered 19.1 Mb of new genome sequence. Functional clustering of the equine genome revealed that the genetic variation tended to develop sensory, immune and communication functions.

Also in 2012, a study of equine CNV showed similar results in terms of genetic variation (Doan, Cohen, Harrington, et al., 2012). A copy number variant virus that is common in the horse genome is used to regulate the biological processes that different horse breeds exhibit.

In 2013, Researchers at the University of Copenhagen used the 700,000-year-old remains of a horse's leg to sequence its DNA molecules, producing a complete genome sequence, the oldest genome ever sequenced (Orlando et al., 2013). The sequencing was performed using a second-generation Illumina instrument and a third-generation tool, the Helicos platform. The Helicos platform, on the other hand, allows single-molecule sequencing, further improved access to endogenous DNA fragments with less risk. In addition, researchers have sequenced an ancient horse and several contemporary horses and found that zebras, horses and donkeys may share a common

ancestor 4 million years ago. Some positive selection genes associated with olfactory and immune systems in domesticated horses may serve as genetic markers for domestication. The significance of the study is that it pushes back the time frame of palaeogenomics by nearly 10 times, a period when DNA samples are not easy to obtain because DNA degrades rapidly into shorter fragments after an organism dies.

The genome of Przewalski's horse also made new progress in 2014. The researchers found that males had 5,879,868 SNPs, while females had 160,910 more SNPs than males. (Do et al., 2014). The X chromosome of Przewaldhorse has more rearrangements than other chromosomes, indicating that the rearrangement of its genome is not evenly distributed across all chromosomes. and have far fewer reversals than insertions and relocations in the horse genome. These results shed more light on chromosome rearrangement and karyotype evolution in Equus.

In 2015, the genetic basis for the rapid adaptation of the Yakut horse to extreme Arctic temperatures was found in the study of the yakut horse population genome (Librado et al., 2015). The study found that the Yakut came along with the Yakuma to populate the region, so they needed to adapt metabolically, anatomically and physiologically to arctic extremes in a very short period of time. Cis-regulated mutations in Yakut horses contribute more to their adaptation than non-synonymous changes, and genes involved external phenotype observable signs are important components of the genetic component of rapid adaptation in Yakut horses.

In 2017, researchers analyzed the evolution and diversification of ancient horses from the Neogene to the Quaternary (Cantalapiedra, Prado, Fernández, & Alberdi, 2017). A phylogenetic approach was used to assess the rate of lineage-specific morphology in horses in this study. under Neogene and

Quaternary radiation in relation to evolutionary patterns of body size and tooth morphology. The pulsation of diversity was shown to be independent of the rate of phenotypic evolution, as it was a constant feature of horse evolution. Thus, rapid branching is associated with external factors, such as increased productivity or dispersal into the old world.

In 2018, researchers remapped the ancestors of Przewalskii's horse and domestic horses in conjunction with studies of the ancient genome. All horses seen today are domesticated, and all wild horses are extinct, according to the study (Gaunitz et al., 2018). This means that large-scale genome transformation set the stage for the expansion of equine gene flow that formed to modern horses.

Advances in sequencing and analysis techniques have led to a deeper understanding of the equine genome, providing a more comprehensive understanding of the evolutionary and domestication history of the equine species, the structural characteristics of the genome, and the genetic basis of population diversity.

1.3 Aim and objectives

The first stage of the project will be to develop the equine genome database, The establishment of the database needs to identify SNV, INDEL and SV genetic variation information. In addition, PSMC method will be use to infer the effective population size of the individual's population at various periods in history using the resequencing data of the individual, and then wait for the sequencing of Exmoor ponies to be matched. Provides a comprehensive set of annotated horse variants that map to the Equcab3 reference assembly.

2. Theory

2.1 Establish the horse genome database

2.1.1 Data collection

Detailed annotation of the horse genome is a task in the Animal Genome Functional Annotation FAANG project, in which a large number of high quality and continuous reference genomes are preserved (Andersson et al., 2015). In total, the project used whole genome sequencing results from 87 horses. Its SRA data number was taken from 88 horses mentioned in the results of a whole genome sequencing study (Jagannathan et al., 2019). Except for one that could not be found, a total of 122 records were downloaded from FAANG according to the SRA index, for a total of 25 horse species. Fifty-six of the horses were males and 31 were females. The main download data from https://data.faang.org/home. Each sample fastq file is divided into fq1 and fq2. be sure to download both files at the same time. In addition, it is important to note that the vast majority of sample files are over 10G. Therefore, a good file management organization can help manage the process of these samples in different stages. One solution in this project is to classify and download samples according to the Study Accession index, because the genetic data under the same study accession index will have some ascending order. See Appendix A for data collection procedures.

2.1.2 Use the Trim Galore

Trim Galore is actually a wrapper for FastQC and Cutadapt. It is suitable for all high-throughput sequencing platforms, including RRBS(Reduced Representation Bisulfeet-SEq), Illumina, Nextera, and smallRNA sequencing platforms with double-ended and single-ended data (Krueger, 2021). The main function of Trim Galore consists of two steps. The first step is to remove the low

quality base before removing the 3 'end adapter. The program will automatically detect the first 1 million of the sequence without specifying the corresponding adapter, and then try the first 12 to 13bp of the sequence to see if there are any adapters that meet the following requirements, like: Illumina: AGATCGGAAGAGC, Small RNA: TGGAATTCTCGG and Nextera: CTGTCTCTTATA. More specifically, the current algorithm basically recovers the true portion of DNA by making the correct alignment between the sequence of joints and the 3 'end of the reading. They divided the double-ended reads into two groups of single-ended reads and modified adapter of each group independently (Y.-L. Li et al., 2015). This means that the overall quality of the sequence is guaranteed after the low-quality bases are filtered out. In other words, after filtering out the low-quality bases, the overall quality of the sequence will be ensured.

When using Trim Galore for quality control, that specify the parameter paired, which indicates that the file is double-ended, and that if one of the reads is removed, the other reads will be discarded regardless of whether they meet the criteria. This option discards read pairs that are too short without interfering with the sequential Fastq files required for many aligners. See Appendix B for data collection procedures.

2.1.3 Mapping

The history of next-generation sequencing technology is growing faster than Moore's Law in computer architecture (Zhang, Liu, & Dong, 2019). In other words, the process by which the cost of next-generation sequencing continues to fall makes IT more common, and the evolution of the field shifts from sequencing itself to IT. In addition, it is important to note that NGS sequencing does not read complete chromosomes, so the sequencer takes short, long

pieces of DNA (Houtgast, Sima, Bertels, & Al-Ars, 2018). Furthermore, if sequencing is required to replicate a complete genome it is done through a process called the genomics pipeline. The process is to compare each short-read fragment from the mapping stage to the reference genome to find the most suitable match position.

Bwa-mem is one of the short read mapping jobs that is superior to most other software and can find results quickly and accurately (H. Li, 2013). The principle is to locate the appropriate location on the reference genome for each short read in the input data set based on seed-and-extend (Houtgast et al., 2018). In this algorithm, maximal exact matches (MEM) were used to extend seeds alignments, and affine-Gap Smith-waterman (SW) algorithm was used to extend seeds. The BWA--MEM algorithm performs local alignment and splicing. There may be multiple optimal matches in different parts of the Query sequence, resulting in multiple optimal matches in reads.

What needs to be done in this project is to use the BWA-MEM tool to map the two parts of the quality-controlled gene dataset to the equine gene reference EquCab3.0.

2.2 Remove PCR duplicate reads

Next generation sequencing platforms Analyzing next generation sequencing data with large data sets is difficult because multiple sequencing is done for each location of the target gene, like the transcriptome and exome (Ebbert et al., 2016). Next generation sequencing platforms Analyzing next generation sequencing data with large data sets is difficult. So, many bioinformatics algorithms have been developed to analyze these data sets under the challenge of an urgent need for a high quality data set. PCR is usually used to

amplify the library if the amount of starting material is small and/or to increase the number of cDNA molecules sufficient for sequencing. Run as few amplification cycles as possible to avoid PCR artifacts. The goal of these algorithms is to find the right balance between error rate, computation time, data loss, and memory consumption. An important step in many process methods is the removal of duplications by PCR in order to prevent detection of variations due to false positives. This process involves the removal of repeated portions of multiple PCR products from the same template molecule bound to the flowing pool (Ebbert et al., 2016). In detail, a PCR repeat is a sequence reading obtained when two or more identical copies of a DNA fragment are sequenced. In the worst case, PCR has a higher frequency of repeated sequencing during amplification due to the introduction of incorrect mutations in the presence of alleles compared to other non-haploid organisms. Ideally, the circulation pool hybridizes with each PCR copy from the same DNA molecule fragment resulting in each being sequenced and, as a result, the reading from this step becomes a PCR repeat. First, the process of hybridization between sequences in the PCR product library and the flow cell is uncontrollable. Second, the original DNA molecule is always more or less biased in its amplification. These are the things that lead to repetition (Ebbert et al., 2016).

Two software programs that are primarily used to remove PCR duplicates are Picard (MarkDuplicates) and SAMTools (rmdup) (Ebbert et al., 2016). Both software generally use similar methods to double mark or delete. samtools identifies PCR duplications by reading pairs at exact locations in the genome and reverse readings from the same locations on the 3 'end map (Ebbert et al., 2016). Moreover, doing so will only screen out the read pairs saved with the highest map quality score, removing some possible drawbacks. Both software generally use similar methods to double mark or delete. Samtools identifies PCR duplications by reading pairs at exact locations in the genome and reverse readings from the same locations on the 3 'end map. Further, doing so will only

screen out the read pairs saved with the highest map quality score, removing some possible drawbacks. In addition, this approach does not work for read pairs that are not paired in end-mode, such as RMDUP, which assumes that all reads in BAM files are under the same library, can cause accidents when there are multiple libraries in BAM (H. Li, 2011).

The BAM and Sam format file formats generated by samtools provide a clear unified interface between downstream analysis such as mutation detection, genotyping, and assembly (H. Li et al., 2009). Additionally, the tool also supports read alignment, implementing various utilities such as indexes, variant callers, and aligned viewers for the SAM format. The BAM format, however, is a binary representation of Sam, with exactly the same information as Sam to improve subsequent performance (H. Li et al., 2009).

As shown above, samtools is a powerful piece of software for parsing and manipulating SAM/BAM alignment. It can not only remove PCR duplicates, generate each location information in a stacked format, but also perform format conversion, sorting and merging operations. Refer to Appendix C for the procedure

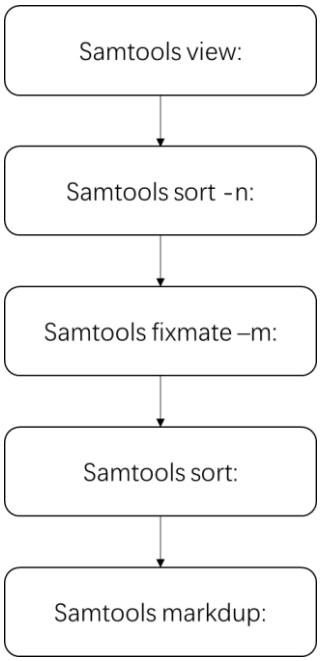

Figure 1: PCR repeat sequence removal process

2.3 PSMC

Information about the historical development of population size over time is provided by the temporal distribution of the two alleles from the most recent common ancestor (TMRCA) in the species (H. Li & Durbin, 2011). Because the statistical resolution of inferences from any one locus is poor over time, and few independent lineages explore depth time depth. Thus, there are hundreds of thousands of separate alleles each carrying their own version of TMRCA (H. Li & Durbin, 2011). In order to reconstruct TMRCA distribution on autosomal and X chromosomes, changes in the local density of heterozygous sites in the genome can be analyzed (H. Li & Durbin, 2011).

Pairwise sequential Markov combined PSMC can estimate its change trajectory over a period of time based on the effective population size simulated by genetic data (Nadachowska-Brzyska, Burri, Smeds, & Ellegren, 2016). The local heterozygous patterns used to estimate the characteristics of TMRCA make the mean genome-wide coverage indicator important for filtered PSMC analysis because the true genotype status in the genome can be influenced by

SNP calls (Han, Sinsheimer, & Novembre, 2014).

When studying the distribution of natural populations, it is found that a continuous habitat is subdivided by density or local dispersion (N. H. Barton, 2008). In other words, a species on a larger scale can form distinct geographical races because of the isolation of interbreeding areas. At the same time, the effects of different environments and hybrid adaptations on species impede gene exchange between these races. They may be eliminated regardless of whether the allele itself is beneficial to the population, because it's associated with alleles at selected loci (N. Barton & Bengtsson, 1986). There is a positive correlation between genetic diversity and whether a population has a stable range. However, the rate of population range contraction will affect population diversity. For example, rapid range contraction will indicate a higher level of biodiversity and trigger lower genetic differentiation between refuge areas (Arenas, Ray, Currat, & Excoffier, 2012). It is also worth noting that the ability of species to spread and the rate of change determine the level of diversity between protected areas after climate change (Arenas et al., 2012).

The average sequencing depth obtained using qualimap was approximately 2 times that used when calling the variation sites (Armstrong et al., 2021). For details, see Appendix E.

2.4 Qualimap

The system's detection of unnecessary deviations caused by sequencing technology and mapping algorithms facilitates downstream analysis (Garcia-Alcalde et al., 2012). Therefore, the presentation process of quality control results will intuitively discover the overall quality of the genetic data set. To provide a theoretical support for the credibility of the resulting results. qualimap provides an additional assessment of the performance of post-overlapping

mapping of genome characteristics with research reading technologies. These read counts are loaded into the program as text by specialized tools in the qualimap to be computed directly (Garcia-Alcalde et al., 2012). Get the results of qualimap using the methods in Appendix D.

2.5 Strelka

Strelka is a scheme for detecting somatic SNV and small insertion deletion from sequencing data of normal samples (Saunders et al., 2012). This approach is a unique Bayesian approach, using the normally expected genotype structure to represent the normal sample by mixing the germ line variation with noise, and then the normal sample with the somatic variation to represent the tumor sample.

Normal sequencing experiments have more stringent requirements on the accuracy and efficiency of the method of identification of somatic variation, especially the number of somatic variation in SNV and small deletion of insertion is easy to exceed the manual review. Therefore, the ideal scenario for detecting somatic variation in matched tumor and normal samples is one in which no external purity estimation is required, i.e. robust resolution of impurities and copy number variation in tumor samples (Saunders et al., 2012).

Strelka's workflow, in addition to its core modeling scheme, improves accuracy by performing read realigning and indel searching for connections in context between two samples. This structure does not need to be assessed for purity because its structure can account for changes in allele frequency at any level occurring in a tumor sample (Saunders et al., 2012). See Appendix F for usage procedures.

2.5 VCFTools

Variation Call Format (VCF) is a common format for storing DNA polymorphism data, including annotated SNPS, inserts, deletions, and structural variations. This is a format that stores and can be indexed in a compressed way to quickly retrieve variant data from a series of locations on the reference genome. Vcftools is a utility software program for processing VCF files that can validate, merge, and compare genetic data. In addition, it provides a general-purpose Perl API (Danecek et al., 2011).

The convenience and versatility of the VCF makes it possible to develop a standardized format to store the most common type of sequence variation. The main representative development of this format is human genetic variation, but its uses can also be applied to different environments, such as the diploid genome of horses. It is essential that the representation of various genomic variations in a single reference sequence be developed based on flexibility and user scalability (Danecek et al., 2011). See Appendix F for usage procedures.

2.7 BCFTools

With the rapid increase in the number of samples obtained from exon and whole genome sequencing, it becomes more important to be able to screen the required variation data accurately and rapidly. The core task is to sequence the variants and annotate their functional effects, because predicting functional outcomes is important for clinical, evolutionary, and genotyping downstream interpretation (Danecek & McCarthy, 2017).

Bcftools are used to identify homozygotes by hidden markov models. The HMM

is based on the VCF format of the sample genetic variation data, corresponding the chain positions to the isolated sites in the population, and then analyzing the possibility of genotype invocation. Two of these hidden states represent extended pure sum and impure sum in the sample. However, the genotype consisted of RR representing the pure sum site matched to the reference, AA representing the pure sum site and RA representing the heterozygous site. Therefore, the H region can contain only RR and AA sites, while the N region can contain sites of any genotype (Narasimhan et al., 2016). See Appendix F for usage procedures.

3. Result

3.1 Trim galore results

An example of the effect of Trim Galore is shown in ERR1527964, a male horse with Franches Montagnes. Use FastQC to generate a report on the second part of both the original data downloaded from FAANG and the Trim Galore quality control data.

The following table shows the statistical results of the basic data. The input format, encoding format and overall Sequence information of the sample can be obtained from this. It can be seen from the table that the number of sequencing this time is 174789544.

Measure	Value
Filename	ERR1527965_2.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	174789544
Filtered Sequences	0
Sequence length	101
%GC	40

Measure	Value
Filename	ERR1527964_2_val_2.fq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	173287896
Filtered Sequences	0
Sequence length	20-101

%GC	39

From these two tables, 1,501,648 has been reduced after quality control of Trim Galore across total sequences.

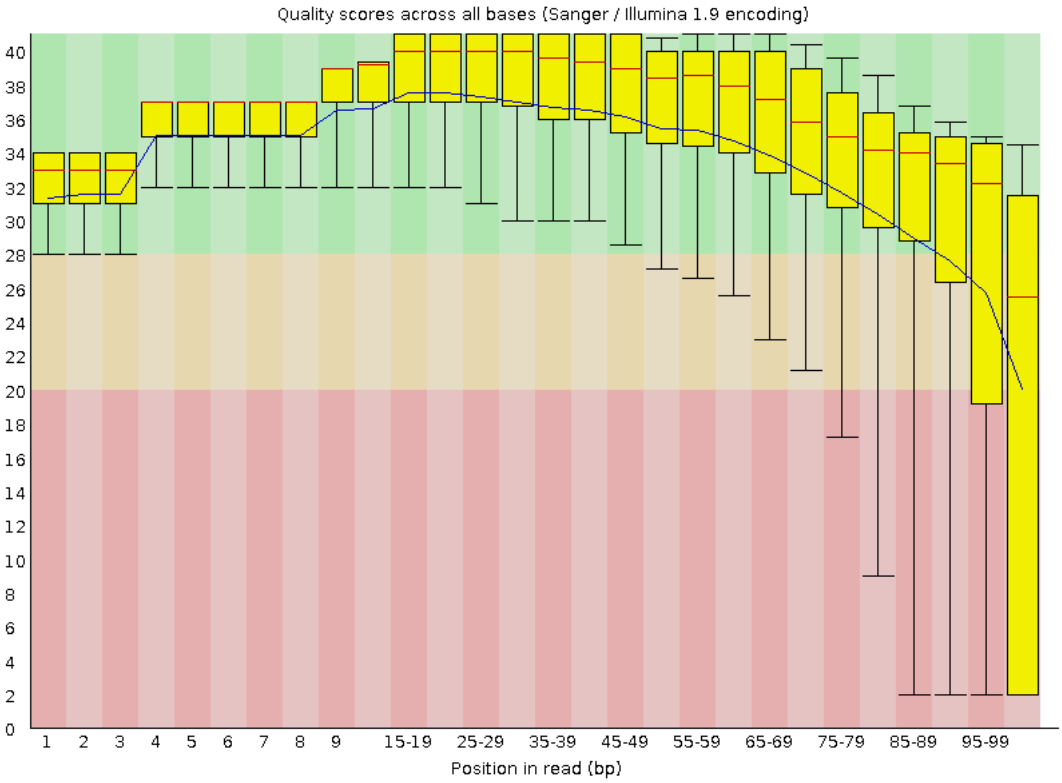

Figure 2: The second part of the original data of ERR1527964 downloaded from FAANG

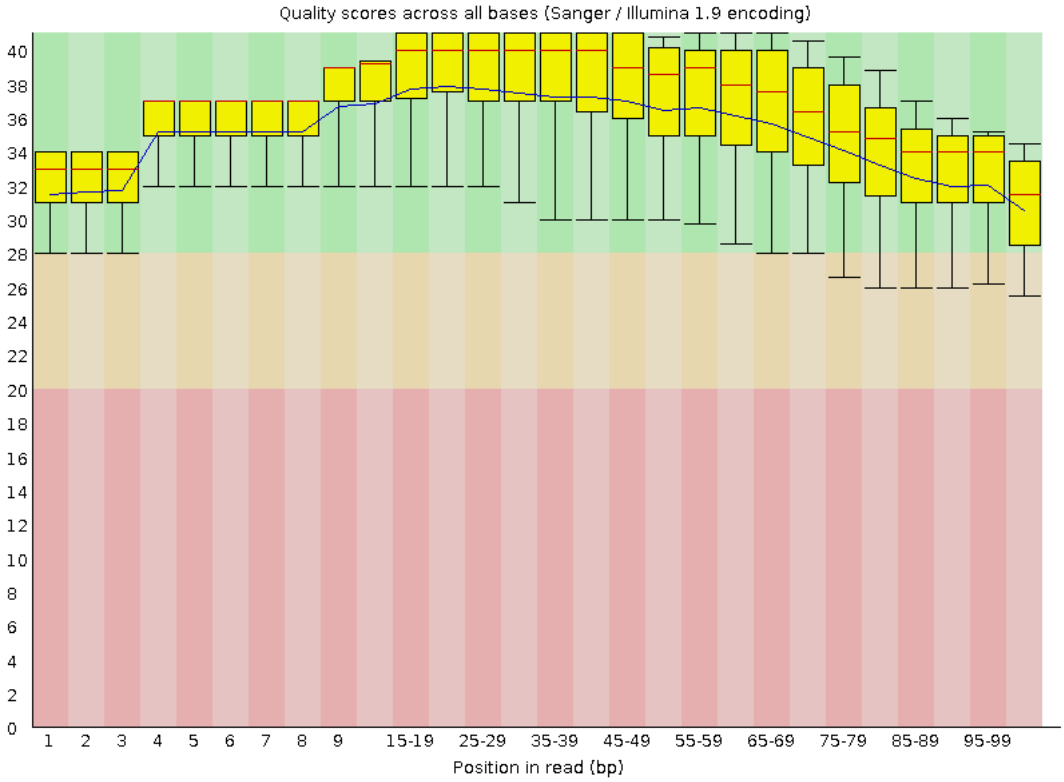

Figure 3: The ERR1527964 data were carried out in the second part after trim-Galore quality control

Read length is 100, the position of the read on the horizontal axis, and the quality on the vertical axis. The quality results are given according to quality. The normal range is 28 to 40, the warning range is 20 to 28, and the error range is less than 20.

Total reads processed	174,789,544
Reads with adapters	50,693,412 (29.0%)
Reads written (passing filters)	174,789,544 (100.0%)
Total basepairs processed	17,653,743,944 bp
Quality-trimmed	1,003,045,792 bp (5.7%)
Total written (filtered)	16,570,782,678 bp (93.9%)

The table shows the total written rate at 93.9%, which illustrates the need for

quality control and a good performance result.

3.2 Qualimap result

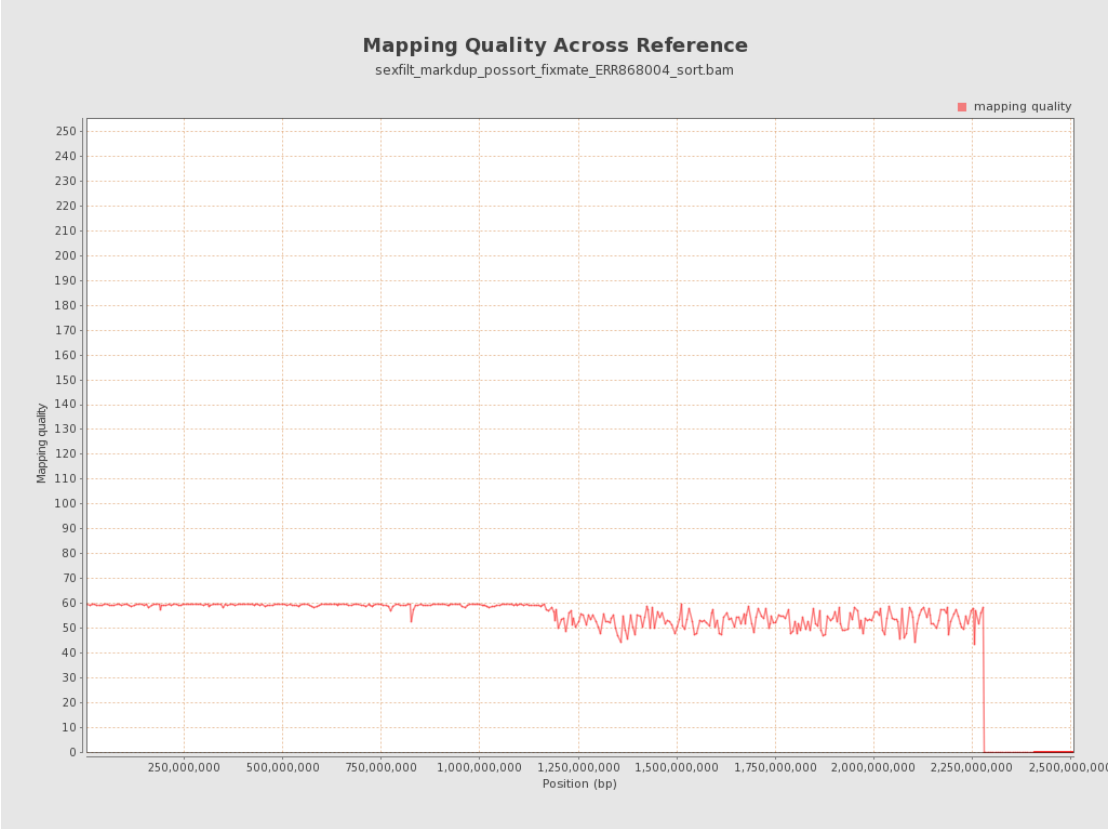

Figure 4: Qualimap for Mapping Quality Across Reference

The line chart in figure 4 shows that the quality of the comparison result between mapping quality and reference data set of horses fluctuates around 60, and its value fluctuates downward less than 15.

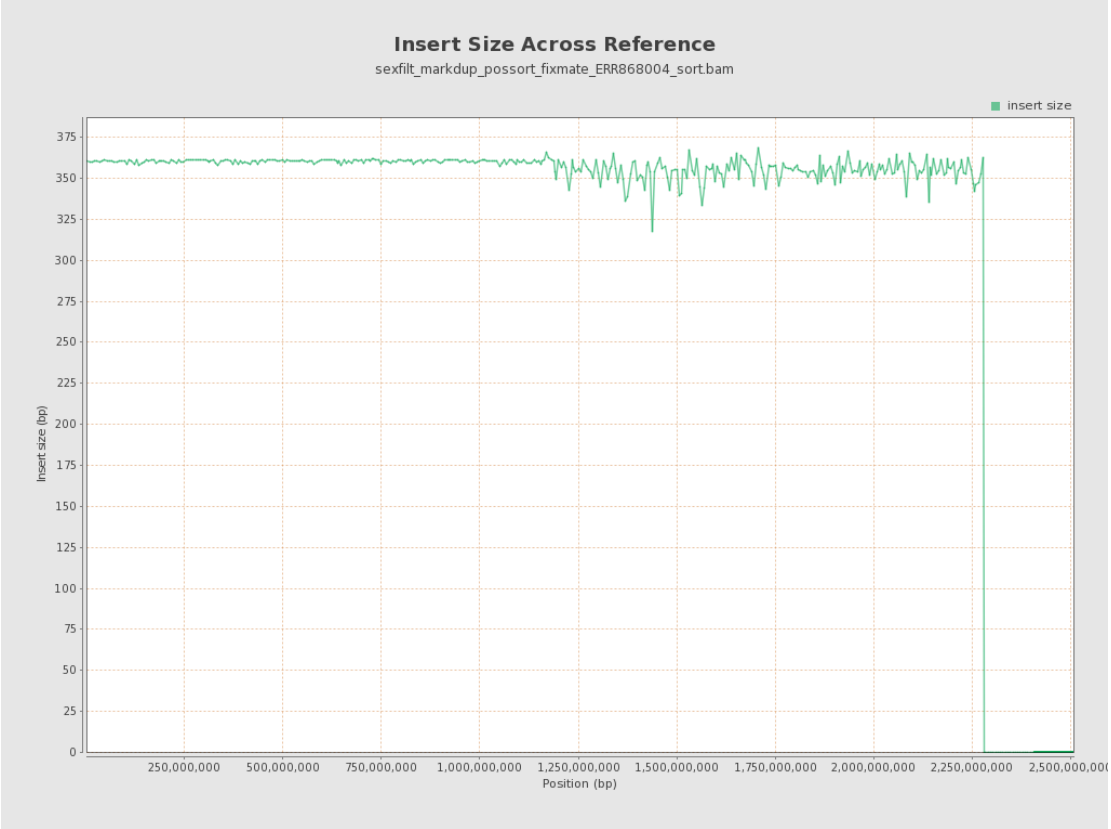

Figure 5: Qualimap for Insert Size Across Reference

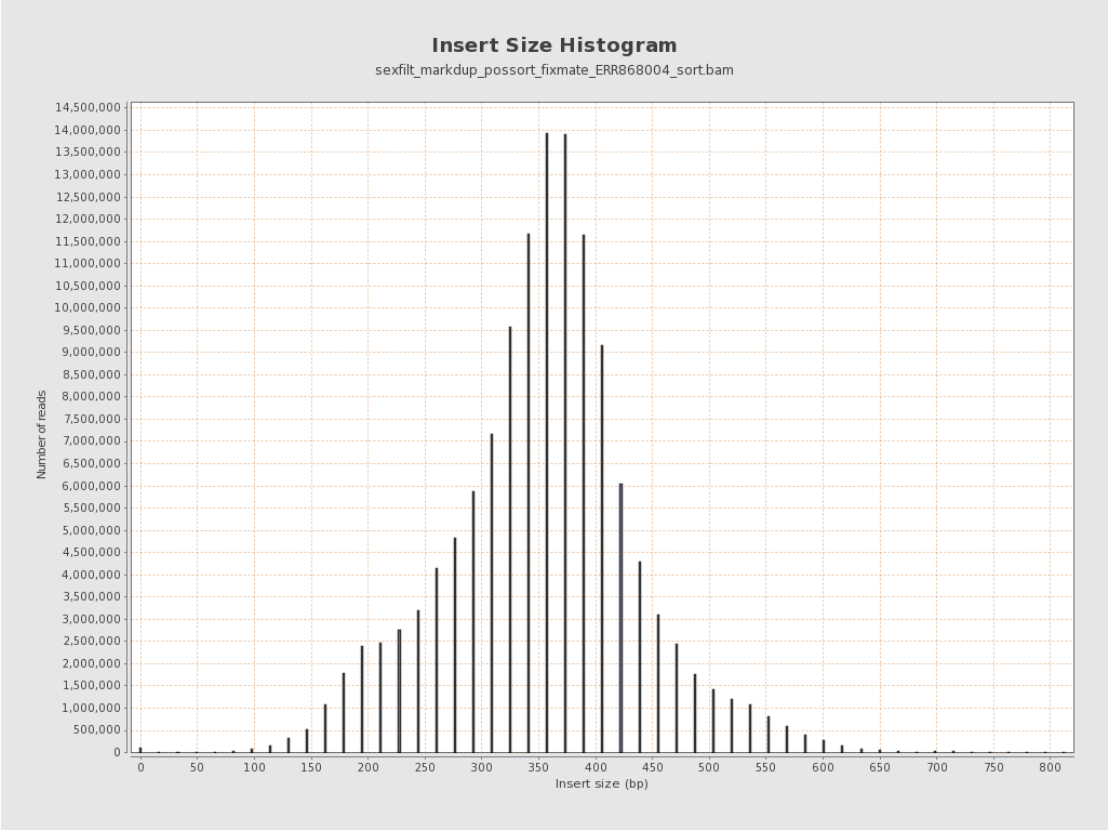

Figure 6: Qualimap for Insert Size Histogram

Figure 5 and Figure 6 show that about 360bp of insert size can be obtained in the library from horizontal and vertical perspectives.

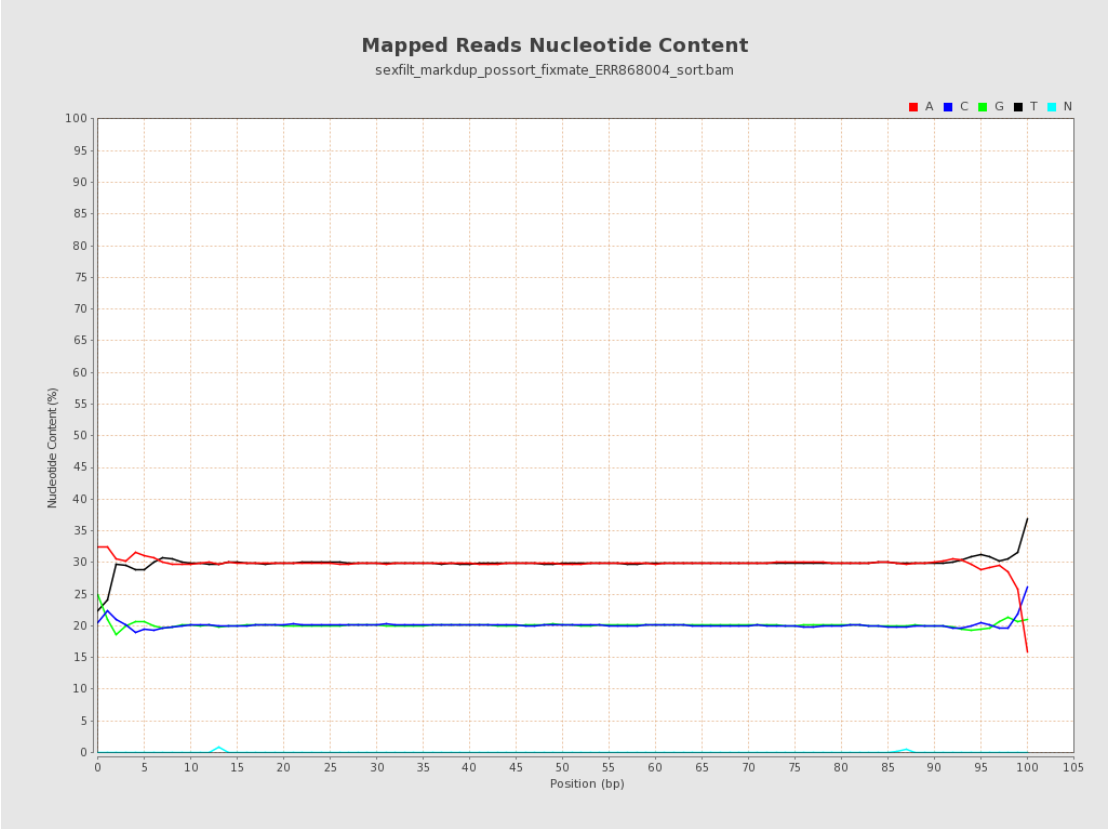

Figure 7: Qualimap for Mapped Reads Nucleotide Content

The four nucleotide mapped reads remain parallel to each other about 90% of the time. Dimensionally, that's a good sample set.

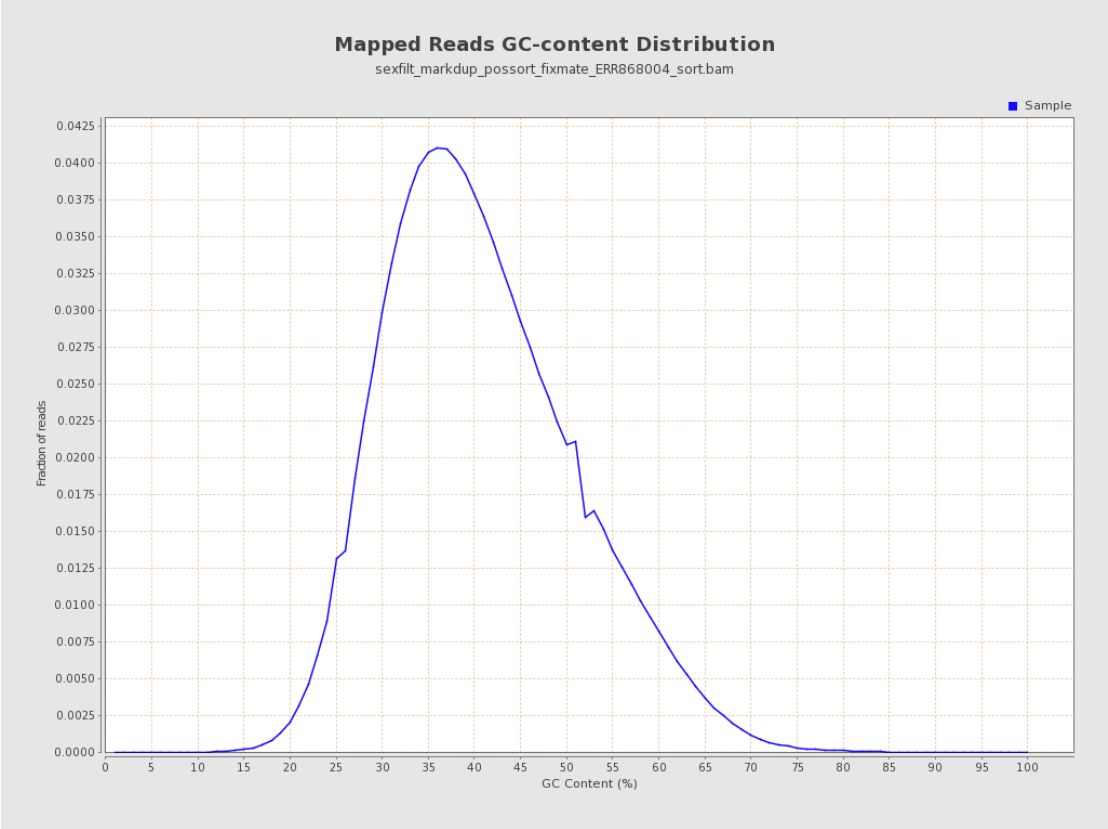

Figure 8: Qualimap for Mapped Reads GC-content Distribution

The straight-line graph in Figure 8 shows that GC content starts to rise at 20%, then peaks at around 35%, and then drops significantly to near zero at around 80%, with a slight fluctuation at 25% and 50%.

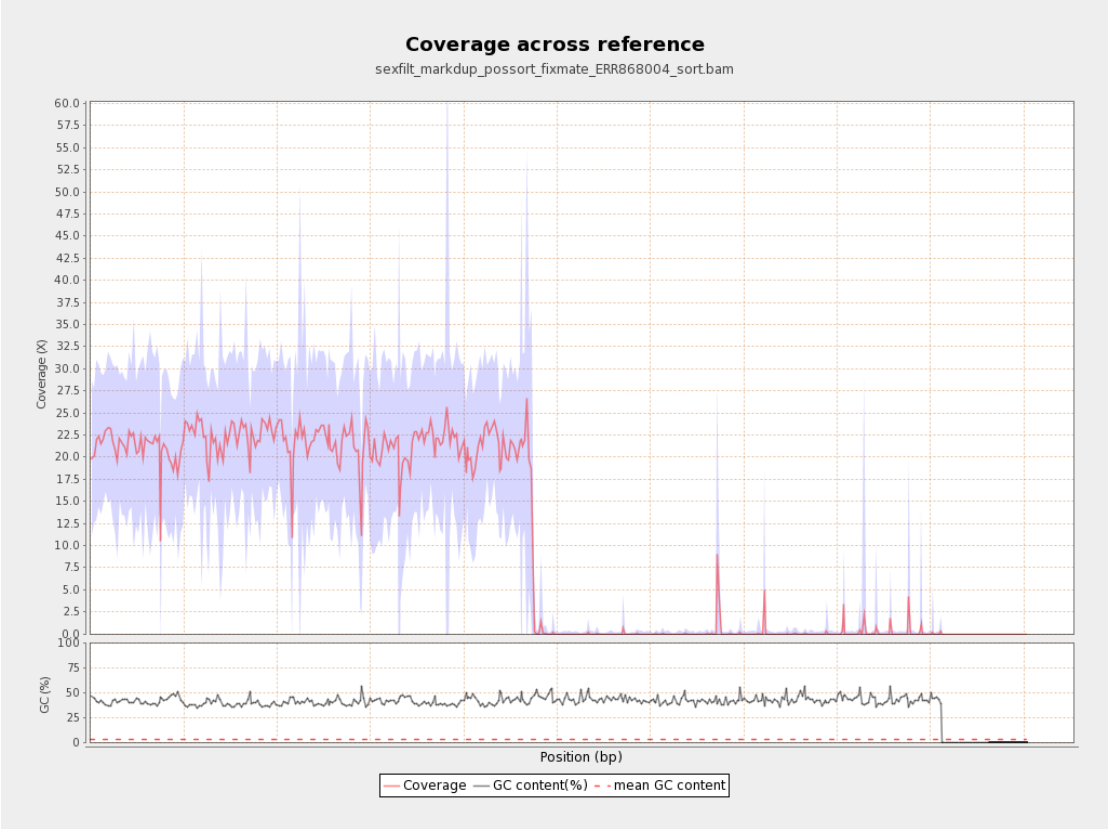

Figure 9: Qualimap for Coverage across reference

Figure 9 shows that mean coverage was about 22.5x when the sample was matched to the reference. GC content fluctuates within a margin of error of approximately 50% or less.

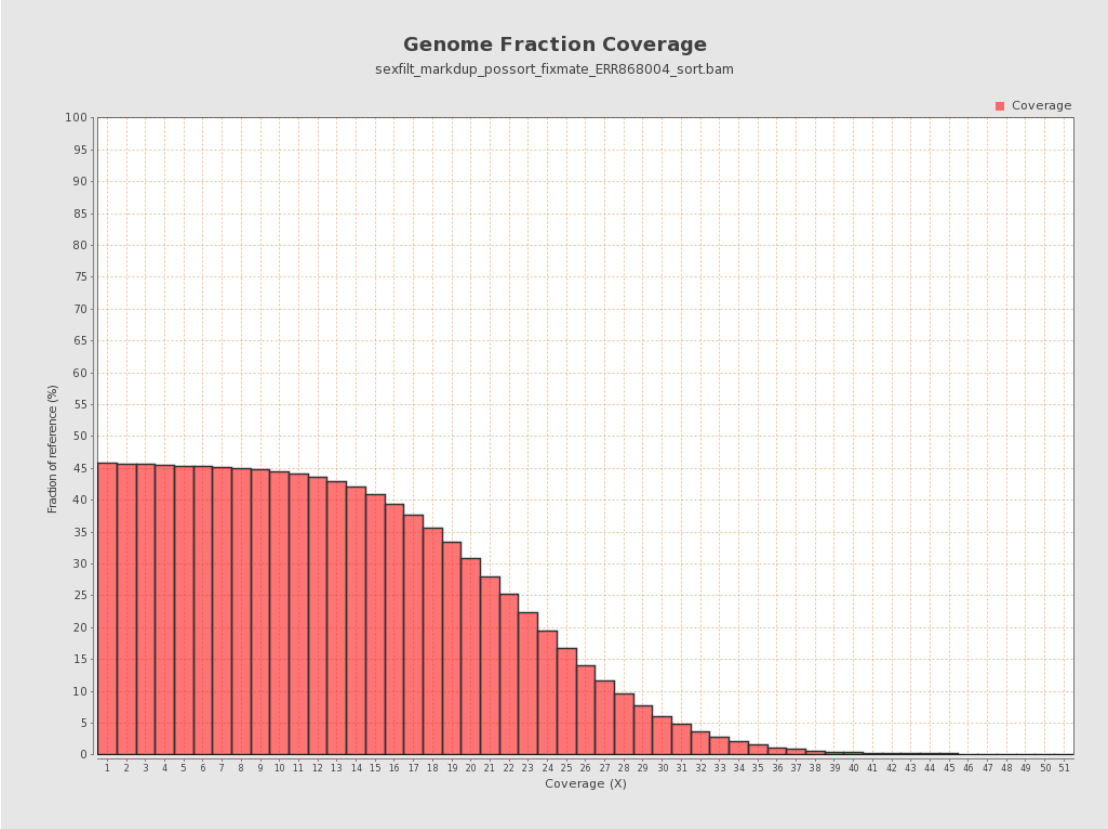

Figure 10: Qualimap for Genome Fraction Coverage

In Figure 10, it can be observed from the way that most of the genome coverage is within the depth of 25x, and its distribution is very sparse after the depth of 30.

3.3 PSMC result

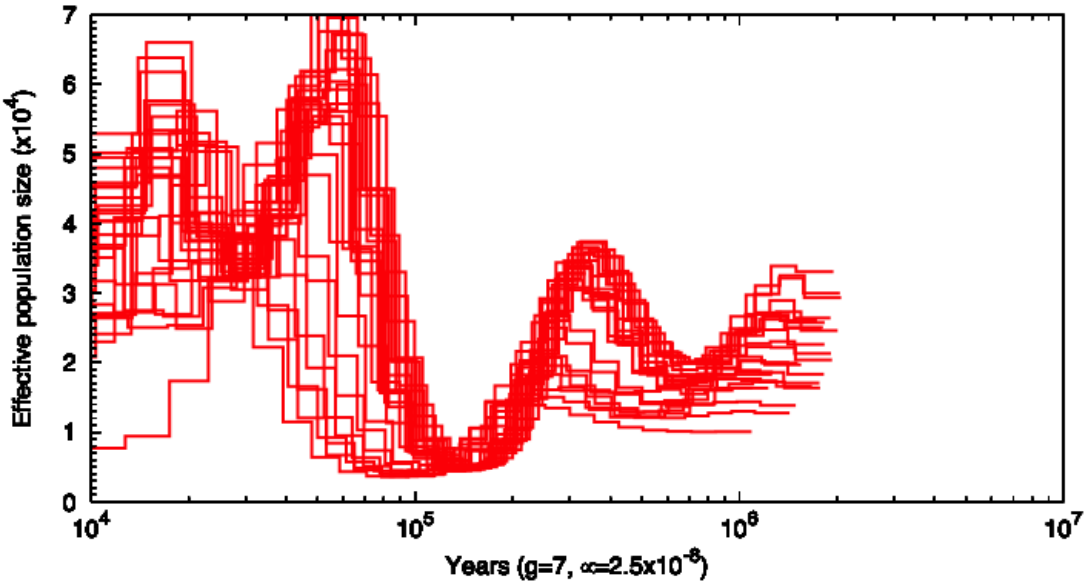

Figure 11: PSMC has no labels for 26 species

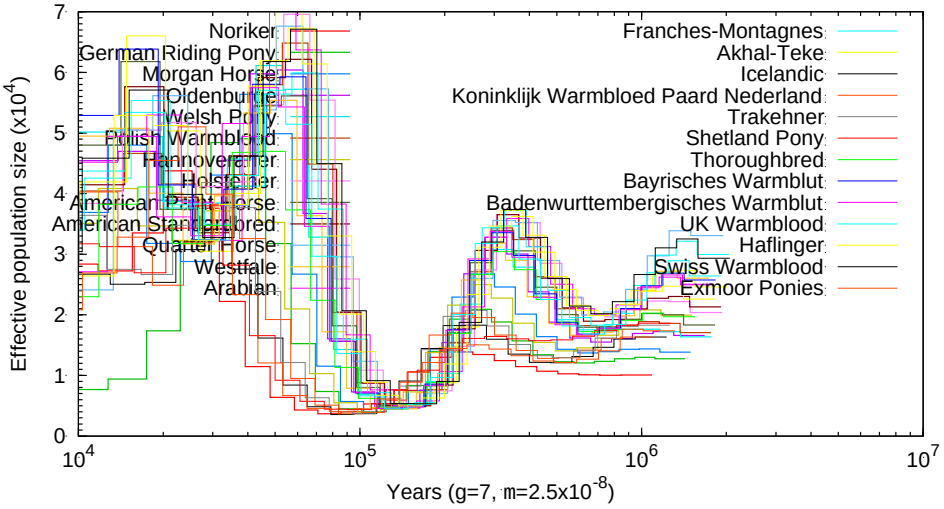

Figure 12: PSMC has labels for 26 species

Figure 11 and 12 shows the historical population size changes of 26 equine populations calculated based on the mutation rate of $2.91 * 10^{-8}$ times per equine generation of 7 years. Overall, the historical change in total group size has gone through three periods of underestimation, approximately $8 * 10^5$, $1.1 * 10^5$ and $5 * 10^4$ respectively. However, four horse breeds remained unaffected during the final period of population collapse Noriker, German Riding Pony, Icelandic and Exmoor Ponies. In addition, the PSMC can only extrapolate to around 10,000 years ago. The effective population size was

basically the same from the sequencing sample population 10,000 years ago, indicating that these samples may not have been differentiated before 10,000 years ago.

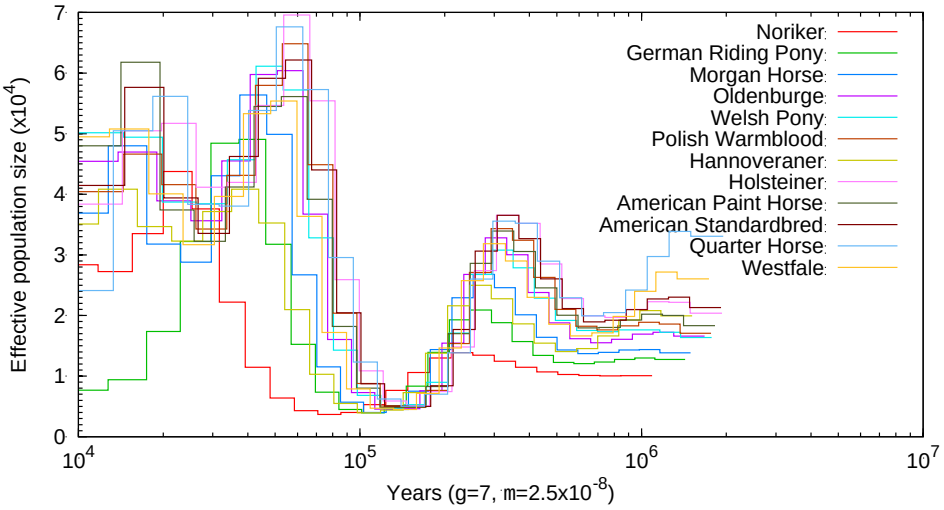

Figure 13: PSMC for 12 species

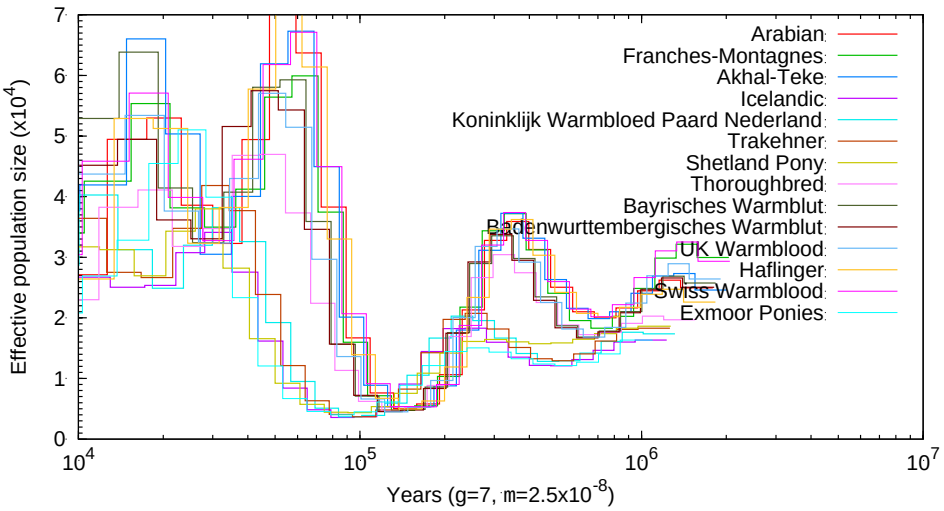

Figure 14: PSMC for 14 species

Due to the large amount of data, horse species are divided into two parts, Figure 13 and Figure 14, respectively.

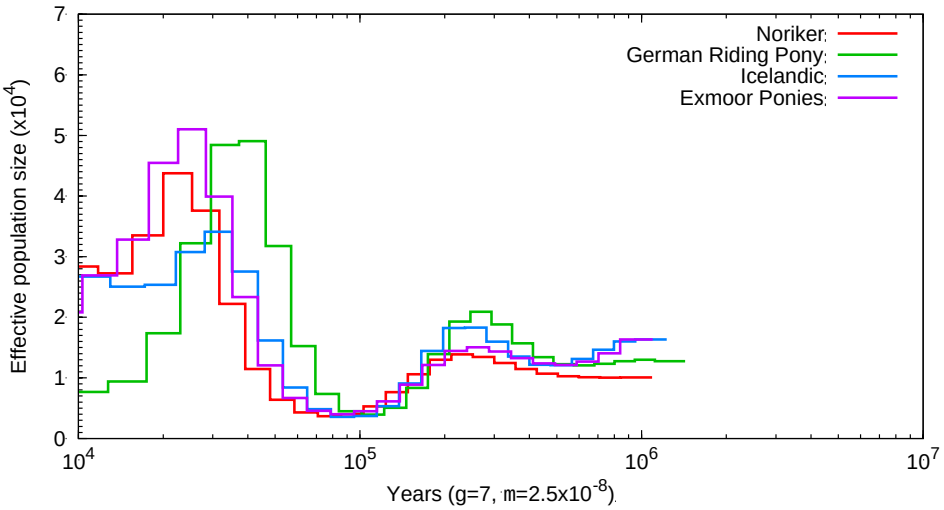

Figure 15: PSMC results had the same trend as Exmoor Ponies

The study looked at the overall trend of four horse species that were different from the other 22. They had similar characteristics, with only one population decline in the last 100,000 years, and the change in population size was relatively slow.

3.4 Strelka Result

Breeds	Region	Records	SNPs
Noriker	Austrian	1	5531341
German Riding Pony	Germany	2	8411819
Morgan Horse	United States	1	5784236
Oldenburger	Oldenburger	3	9138847
Welsh Pony	Wales	2	8648027
Polish Warmblood	Poland	1	5621141
Hannoveraner	Germany	2	8139031
Holsteiner	Germany	8	11393815
American Paint Horse	North America	6	9702329
American Standardbred	American	1	5895142
Quarter Horse	American	3	7081804
Westfale	Westfalen	2	7889537
Arabian	Arabian Peninsula	2	8019544
Franches-Montagnes	Switzerland	64	14493857
Akhal-Teke	Turkmen	4	8950652
Icelandic	Iceland	2	8582857
Koninklijk Warmbloed Paard Nederland	Nederland	1	4981716
Trakehner	Prussia	2	7620588
Shetland Pony	Shetland Islands,	3	6502423

	Scotland		
Thoroughbred	England	1	4853868
Bayrisches Warmblut	Germany	1	5453355
Badenwürttembergisches Warmblut	Germany	1	5439708
UK Warmblood	England	2	7636110
Haflinger	Austria and northern Italy	4	9152631
Swiss Warmblood	Switzerland	4	10864249

The table shows the strelka calculation results of all samples of each horse breed and the SNPs obtained through vcftools filtering. A total of 19,578,627 SNPs sites were found. By integrating the results of strelka analysis of 26 horse species, 1,012,116 SNPs sites were obtained after filtering.

4. Discussion

Through PSMC calculation, the evolution of the horse was reconstructed. Combined with relevant meteorological and geological data, it was found that about 2 million years ago, the earth entered the Quaternary great ice age, and the global temperature gradually cooled (Walker et al., 2012). It turns out that mammals were better able to adapt to ice ages than reptiles. Says it affects horse populations during cold weather. Over the ensuing millions of years, the earth's climate fluctuated frequently and dramatically, changing the effective population of horses. Around 50,000 years ago, the effective population of horses reached a new high in the last two million years, but with the arrival of the last ice age, most mammals died out and the effective population of horses declined significantly. Although it is impossible to rule out the impact of human hunting on the horse population, it is undeniable that changes in the earth's environment have had a profound impact on the evolution of the horse.

Further research on Exmoor Pony can be considered with 3 other breeds with the same historical population change including Noriker, German Riding Pony and Icelandic. As they adapt to changes in the earth's climate and species evolve, they have more genetic resonance, such as a tolerance to cold. In addition, these horse breeds can be classified and compared with other horse breeds, so that it is easier to extract excellent genes conducive to breeding, and at the same time, to take care of the health of the horse breeds through gene research.

Because the SNP in the distribution of the chromosome with relative homogeneity and density is far higher than that of microsatellite DNA loci, and normality easier to achieve rapid and high-throughput automation testing, therefore is considered to be the most potential applications of a new generation of genetic markers, their disease for species complex traits in post

genome era and pharmacogenomics studies play an increasingly important role. In future studies, the establishment and accumulation of data on the relationship between susceptible genes and disease phenotypes and SNP or SNP type in equine diseases will help us to discover and determine the susceptible genes of equine complex genetic diseases.

In the follow-up study, the differences between other breeds can be further analyzed, in order to find more valuable positive selection genes and mutations and provide scientific basis for future research on various physiological and pathological phenomena of equine.

5. Conclusion

The study collected and established genetic databases for 87 horse breeds from 25 species. This study collected and sorted out a large number of data sets, which laid a solid foundation for subsequent research and expansion. The data obtained by sequencing were quality-controlled to ensure the reliability of gene reads, and then the quality-controlled data were mapped to the equine gene reference set. After the data underwent the process of removing duplicates from PCR, subsequent PSMC analysis and SNPs statistics could be performed.

Design workflow in a scientific and rigorous way, and strictly control the data processing results of each step, so that the analysis conclusions will be more reliable. The historical population size of the horse was then analyzed in combination with climate and geological changes on earth. In this way, we can find similar correlations among different horse breeds, and verify the conjecture through further observation and comparison of these correlations in subsequent studies.

Reference List:

Andersson, L., Archibald, A. L., Bottema, C. D., Brauning, R., Burgess, S. C., Burt, D. W., . . . Sveriges, I. (2015). Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome biology, 16*(1), 57-57. doi:10.1186/s13059-015-0622-4

Arenas, M., Ray, N., Currat, M., & Excoffier, L. (2012). Consequences of Range Contractions and Range Shifts on Molecular Diversity. *Molecular Biology and Evolution, 29*(1), 207-218. doi:10.1093/molbev/msr187

Armstrong, E. E., Khan, A., Taylor, R. W., Gouy, A., Greenbaum, G., Thiéry, A., . . . Barsh, G. (2021). Recent evolutionary history of tigers highlights contrasting roles of genetic drift and selection. *Molecular Biology and Evolution, 38*(6), 2366-2379.

Barton, N., & Bengtsson, B. O. (1986). The barrier to genetic exchange between hybridising populations. *Heredity, 57*(3), 357-376.

Barton, N. H. (2008). The effect of a barrier to gene flow on patterns of geographic variation. *Genetics research, 90*(1), 139-149.

Cantalapiedra, J. L., Prado, J. L., Fernández, M. H., & Alberdi, M. T. (2017). Decoupled ecomorphological evolution and diversification in Neogene-Quaternary horses. *Science, 355*(6325), 627-630.

Chowdhary, B. P., & Bailey, E. (2003). Equine genomics: galloping to new frontiers. *Cytogenetic and Genome Research, 102*(1-4), 184-188. doi:10.1159/000075746

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . Genomes Project Analysis, G. (2011). The variant call format and VCFtools. *Bioinformatics, 27*(15), 2156-2158. doi:10.1093/bioinformatics/btr330

Danecek, P., & McCarthy, S. A. (2017). BCFtools/csq: haplotype-aware

variant consequences. *BIOINFORMATICS, 33*(13), 2037-2039. doi:10.1093/bioinformatics/btx100

Do, K.-T., Kong, H.-S., Lee, J.-H., Lee, H.-K., Cho, B.-W., Kim, H.-S., . . . Park, K.-D. (2014). Genomic characterization of the Przewalski′ s horse inhabiting Mongolian steppe by whole genome re-sequencing. *Livestock Science, 167*, 86-91.

Doan, R., Cohen, N., Harrington, J., Veazy, K., Juras, R., Cothran, G., . . . Dindot, S. V. (2012). Identification of copy number variants in horses. *Genome research, 22*(5), 899-907.

Doan, R., Cohen, N. D., Sawyer, J., Ghaffari, N., Johnson, C. D., & Dindot, S. V. (2012). Whole-genome sequencing and genetic variant analysis of a Quarter Horse mare. *BMC genomics, 13*(1), 1-12.

Ebbert, M. T. W., Wadsworth, M. E., Staley, L. A., Hoyt, K. L., Pickett, B., Miller, J., . . . for the Alzheimer's Disease Neuroimaging, I. (2016). Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC bioinformatics, 17 Suppl 7*(S7), 239-239. doi:10.1186/s12859-016-1097-3

Garcia-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., Goetz, S., Tarazona, S., . . . Conesa, A. (2012). Qualimap: evaluating next-generation sequencing alignment data. *BIOINFORMATICS, 28*(20), 2678-2679. doi:10.1093/bioinformatics/bts503

Gaunitz, C., Fages, A., Hanghøj, K., Albrechtsen, A., Khan, N., Schubert, M., . . . Bignon-Lau, O. (2018). Ancient genomes revisit the ancestry of domestic and Przewalski's horses. *Science, 360*(6384), 111-114.

Han, E., Sinsheimer, J. S., & Novembre, J. (2014). Characterizing bias in population genetic inferences from low-coverage sequencing data. *Molecular biology and evolution, 31*(3), 723-735.

Houtgast, E. J., Sima, V.-M., Bertels, K., & Al-Ars, Z. (2018). Hardware acceleration of BWA-MEM genomic short read mapping for longer read

lengths. *Computational biology and chemistry, 75*, 54-64. doi:10.1016/j.compbiolchem.2018.03.024

Jagannathan, V., Gerber, V., Rieder, S., Tetens, J., Thaller, G., Drögemüller, C., & Leeb, T. (2019). Comprehensive characterization of horse genome variation by whole‐genome sequencing of 88 horses. *Animal genetics, 50*(1), 74-77. doi:10.1111/age.12753

Krueger, F. (2021). Trim Galore. doi:10.5281/zenodo.5127899

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics, 27*(21), 2987-2993.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

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

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *nature, 475*(7357), 493-496.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Li, Y.-L., Weng, J.-C., Hsiao, C.-C., Chou, M.-T., Tseng, C.-W., & Hung, J.-H. (2015). PEAT: an intelligent and efficient paired-end sequencing adapter trimming algorithm. *BMC bioinformatics, 16 Suppl 1*(Suppl 1), S2-S2. doi:10.1186/1471-2105-16-S1-S2

Librado, P., Der Sarkissian, C., Ermini, L., Schubert, M., Jónsson, H., Albrechtsen, A., . . . Seguin-Orlando, A. (2015). Tracking the origins of Yakutian horses and the genetic basis for their fast adaptation to subarctic environments. *Proceedings of the National Academy of Sciences, 112*(50), E6889-E6897.

Librado, P., Fages, A., Gaunitz, C., Leonardi, M., Wagner, S., Khan, N., . . . Al-Rasheid, K. A. (2016). The evolutionary origin and genetic makeup of domestic horses. *Genetics, 204*(2), 423-434.

Lindgreen, S. (2012). AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC research notes, 5*(1), 337-337. doi:10.1186/1756-0500-5-337

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in genetics, 24*(3), 133-141.

McCue, M. E., Bannasch, D. L., Petersen, J. L., Gurr, J., Bailey, E., Binns, M. M., . . . Hill, E. W. (2012). A high density SNP array for the domestic horse and extant Perissodactyla: utility for association mapping, genetic diversity, and phylogeny studies. *PLoS genetics, 8*(1), e1002451.

Nadachowska-Brzyska, K., Burri, R., Smeds, L., & Ellegren, H. (2016). PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white Ficedula flycatchers. *Molecular ecology, 25*(5), 1058-1072. doi:10.1111/mec.13540

Narasimhan, V., Danecek, P., Scally, A., Xue, Y., Tyler-Smith, C., & Durbin, R. (2016). BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *BIOINFORMATICS, 32*(11), 1749-1751. doi:10.1093/bioinformatics/btw044

Nergadze, S. G., Lupotto, M., Pellanda, P., Santagostino, M., Vitelli, V., & Giulotto, E. (2010). Mitochondrial DNA insertions in the nuclear horse genome. *Animal genetics, 41*, 176-185.

Orlando, L., Ginolhac, A., Raghavan, M., Vilstrup, J., Rasmussen, M., Magnussen, K., . . . Zazula, G. (2011). True single-molecule DNA sequencing of a pleistocene horse bone. *Genome research, 21*(10), 1705-1719.

Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., Stiller,

M., . . . Moltke, I. (2013). Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature, 499*(7456), 74-78.

Outram, A. K., Stear, N. A., Bendrey, R., Olsen, S., Kasparov, A., Zaibert, V., . . . Evershed, R. P. (2009). The Earliest Horse Harnessing and Milking. *Science (American Association for the Advancement of Science), 323*(5919), 1332-1335. doi:10.1126/science.1168594

Petersen, J. L., Mickelson, J. R., Rendahl, A. K., Valberg, S. J., Andersson, L. S., Axelsson, J., . . . Borges, A. S. (2013). Genome-wide analysis reveals selection for important traits in domestic horse breeds. *PLoS genetics, 9*(1), e1003211.

Raudsepp, T., Finno, C. J., Bellone, R. R., & Petersen, J. L. (2019). Ten years of the horse reference genome: insights into equine biology, domestication and population dynamics in the post‐genome era. *Animal genetics, 50*(6), 569-597. doi:10.1111/age.12857

Saunders, C. T., Wong, W. S. W., Swamy, S., Becq, J., Murray, L. J., & Cheetham, R. K. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *BIOINFORMATICS, 28*(14), 1811-1817. doi:10.1093/bioinformatics/bts271

Scherf, B. D., & Pilling, D. (2015). The second report on the state of the world's animal genetic resources for food and agriculture.

Shriner, D., Tekola-Ayele, F., Adeyemo, A., & Rotimi, C. N. (2018). Genetic ancestry of Hadza and Sandawe peoples reveals ancient population structure in Africa. *Genome biology and evolution, 10*(3), 875-882.

Wade, C. M., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., . . . Bellone, R. R. (2009). Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science, 326*(5954), 865-867.

Walker, M. J. C., Berkelhammer, M., Björck, S., Cwynar, L. C., Fisher, D. A., Long, A. J., . . . Weiss, H. (2012). Formal subdivision of the Holocene

Series/Epoch: a Discussion Paper by a Working Group of INTIMATE (Integration of ice-core, marine and terrestrial records) and the Subcommission on Quaternary Stratigraphy (International Commission on Stratigraphy). *Journal of quaternary science, 27*(7), 649-659. doi:10.1002/jqs.2565

Wang, D. G., Fan, J. B., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., . . . Spencer, J. (1998). Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome. *Science (American Association for the Advancement of Science), 280*(5366), 1077-1082. doi:10.1126/science.280.5366.1077

Zhang, L., Liu, C., & Dong, S. (2019). PipeMEM: A Framework to Speed Up BWA-MEM in Spark with Low Overhead. *Genes, 10*(11), 886. doi:10.3390/genes10110886

Appendices

Appendix excerpts from only one sample as an example to show the workflow and all the scripts can access git homepage:
https://github.com/yechiyu/Equine-genome-analysis.git

Appendix A

For example, the SRA reference number ERR1527951 to ERR1527972 is a continuous reference that can be done automatically and quickly with a followed shell script under Linux screen. One of the links has to be found on the FAANG website.

Linux script for downloading SRA data from FAANG

```
for ((i=51; i<=72;i++));
do let "g=$i%10";
wget
"ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR152/00"$g"/ERR15279"$i"/ERR15279"
$i"_1.fastq.gz";
wget
"ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR152/00"$g"/ERR15279"$i"/ERR15279"
$i"_2.fastq.gz";
done
```

In addition, different folders and naming data should be created for each stage, which means that the file needs to be labeled with the operation name for each operation result.

Appendix B

Use Trim Galore for quality control.

Step1: Activating the Toolkit environment

Linux script for activated galore trim environment
export PATH=/shared3/Anubhab/miniconda3/bin/ *source activate trim-galore*

Step2: Using Galore Trim

Linux script for using galore trim
*trim_galore --output_dir . --paired ::: *_1.fastq.gz ::: *_2.fastq.gz*

This step does quality control for fastq parts 1 and 2 for all samples in the current directory and generates a report.

Appendix C

Map the dataset to the reference directory.

Step1: Download the equine reference gene locally.

Step2: Activating the Environment Toolkit

Linux script for activate environment toolkit
export PATH=/home/opt/miniconda2/bin:$PATH *source activate popgen*

Step3: Use the BWA-MEM tool for mapping

Linux script for use the BWA-MEM tool for mapping
bwa mem */shared5/studentprojects/Qikai/Horse_database/EquCab3_ref/GCF_002863* *925.1_EquCab3.0_genomic.fna ERR978603_1_val_1.fq.gz* *ERR978603_2_val_2.fq.gz > ERR978603.sam &*

Using a sample ERR978603 from Franches Montagnes breeds as an example, specify the input as a quality control result and reference genome set, and then format the output in SAM format to facilitate the use of downstream software.

Appendix D

Remove PCR duplicate reads.

A sample of Shetland Pony number ERR868004 is used as an example to illustrate the workflow.

Step1: samtools view.

Linux script for samtools view
samtools view -q 30 -Sb ERR868004.sam > ERR868004.bam

The main function of the view command is to convert the SAM file into a BAM file and then execute the BAM file using various operations similar to data sorting. Given the parameter '-q 30', the minimum mass of the comparison is 30.

Step2: samtools sort -n

Linux script for samtools sort
samtools sort -n ERR868004.bam ERR868004_sort &

Sort BAM files using the ‘-n’ parameter to sort them by the ID of the short read.

Step3: samtools fixmate –m

Linux script for samtools fixmate
samtools fixmate -m ERR868004_sort.bam fixmate_ERR868004_sort.bam

Fix the mate information

Step4: samtools sort

Linux script for samtools sort
samtools sort fixmate_ERR868004_sort.bam possort_fixmate_ERR868004_sort

Sort again by position

Step5: samtools markdup

Linux script for samtools markdup
samtools markdup possort_fixmate_ERR868004_sort.bam markdup_possort_fixmate_ERR868004_sort.bam &

Remove PCR redundancy but need to be supported by the steps above.

Step6: samtools index

Linux script for samtools index
samtools index markdup_possort_fixmate_ERR868004_sort.bam &

Index bam format files for quick subsequent access.

Step7: samtools view

Linux script for sexfilt

```
samtools view -h markdup_possort_fixmate_ERR868004_sort.bam | awk '{if($3 == "NC_009144.3" && $3 == "NC_009145.3" && $3 == "NC_009146.3" && $3 == "NC_009147.3" && $3 == "NC_009148.3" && $3 == "NC_009149.3" && $3 == "NC_009150.3" && $3 == "NC_009151.3" && $3 == "NC_009152.3" && $3 == "NC_009153.3" && $3 == "NC_009154.3" && $3 == "NC_009155.3" && $3 == "NC_009156.3" && $3 == "NC_009157.3" && $3 == "NC_009158.3" && $3 == "NC_009159.3" && $3 == "NC_009160.3" && $3 == "NC_009161.3" && $3 == "NC_009162.3" && $3 == "NC_009163.3" && $3 == "NC_009164.3" && $3 == "NC_009165.3" && $3 == "NC_009166.3" && $3 == "NC_009167.3" && $3 == "NC_009168.3" && $3 == "NC_009169.3" && $3 == "NC_009170.3" && $3 == "NC_009171.3" && $3 == "NC_009172.3" && $3 == "NC_009173.3" && $3 == "NC_009174.3"){print $0}}' | samtools view -Sb - > sexfilt_markdup_possort_fixmate_ERR868004_sort.bam &
```

Appendix E

The historical population size changes of 25 horse breeds were analyzed by PSMC.

Step1: Generate qualimap result report.

Linux script for generating qualimap result report.

```
qualimap bamqc -bam sexfilt_markdup_possort_fixmate_ERR868004_sort.bam --java-mem-size=100G
```

Mean coverage was 10.3X.

Step2: To generate diploid sequences are common throughout the genome

Linux script for generating diploid sequences is common throughout the genome
*samtools mpileup -C50 -uf ../EquCab3_ref/GCF_002863925.1_EquCab3.0_genomic.fna sexfilt_markdup_possort_fixmate_ERR868004_sort.bam | bcftools call -c | vcfutils.pl vcf2fq -d 10 -D **21** | gzip > ERR868004_diploid.fq.gz*

The value behind the parameter '-D' is about twice that of mean coverage.

Its main function is to generate BCF, VCF files, or pileup one or more BAM files. The comparison record uses the sample name in @RG as the distinguishing identifier. If the sample identifier is missing, each input file is considered a sample.

Step3: Convert the common sequence to a fasta-like format

Linux script for convert the common sequence to a fasta-like format
../psmc/utils/fq2psmcfa -q20 ERR868004_diploid.fq.gz > ERR868004_diploid.psmcfa

Step4: Infer the history of population size

Linux script for Infer the history of population size
*../psmc/psmc -N25 -t15 -r5 -p "4+25*2+4+6" -o ERR868004_diploid.psmc ERR868004_diploid.psmcfa &*

The first parameter spans the first 4 atomic intervals, each of the next 25 spans 2 intervals, the 27th spans 4 intervals, and the last parameter spans the last 6 intervals. Thus, after 20 iterations, at least 10 reorganizations are inferred to have occurred in the interval that each parameter spans.

Step5: Plot the population's historical size

Linux script for plot PSMC
../psmc/utils/psmc_plot.pl -Y 7 -m 2.91e-8 -g 7 PSMC_ERR868004 ERR868004_diploid.psmc
../psmc/utils/psmc_plot.pl -Y 7 -m 2.91e-8 -g 7 -M "Noriker,German Riding Pony,Morgan Horse,Oldenburge,Welsh Pony,Polish Warmblood,Hannoveraner,Holsteiner,American Paint Horse,American Standardbred,Quarter Horse,Westfale,Arabian,Franches-Montagnes,Akhal-Teke,Icelandic,Koninklijk Warmbloed Paard Nederland,Trakehner,Shetland Pony,Thoroughbred,Bayrisches Warmblut,Badenwurttembergisches Warmblut,UK Warmblood,Haflinger,Swiss Warmblood,Exmoor Ponies" 26Horse_Breeds_Population ERR2179552_diploid.psmc ERR2179545_diploid.psmc ERR2179546_diploid.psmc ERR2179548_diploid.psmc ERR2179543_diploid.psmc ERR2179544_diploid.psmc ERR1545180_diploid.psmc ERR2731058_diploid.psmc ERR1527949_diploid.psmc ERR1512897_diploid.psmc ERR1527969_diploid.psmc ERR1545187_diploid.psmc ERR1527951_diploid.psmc ERR1527952_diploid.psmc ERR1527947_diploid.psmc ERR863167_diploid.psmc ERR1527967_diploid.psmc ERR1545185_diploid.psmc ERR868004_diploid.psmc ERR1735862_diploid.psmc ERR1545179_diploid.psmc ERR1545178_diploid.psmc ERR1306526_diploid.psmc ERR2179553_diploid.psmc ERR1545188_diploid.psmc Exmoor_ponies/s2010_diploid.psmc

The time of equine generation was 7 years, and the mutation rate was $2.91 * 10^{-8}$

Appendix F

Strelka was used to detect variations in sample data.

Step1: Download Strelka locally.

Strelka must be downloaded at least 2.9.5, otherwise an error will be reported.

Step2: configuration

Linux script for configure strelka at American Paint Horse
/shared5/studentprojects/Qikai/Horse_database/strelka-2.9.10.centos6_x86_64/bin/configureStrelkaGermlineWorkflow.py \ *--bam /shared5/studentprojects/Qikai/Horse_database/fixmate_File/markdup_possort_fixmate_ERR1527949_sort.bam* \ *--bam /shared5/studentprojects/Qikai/Horse_database/fixmate_File/markdup_possort_fixmate_SRR6507149_sort.bam* \ *--bam /shared5/studentprojects/Qikai/Horse_database/fixmate_File/markdup_possort_fixmate_ERR1305962_sort.bam* \ *--bam /shared5/studentprojects/Qikai/Horse_database/fixmate_File/markdup_possort_fixmate_ERR1305963_sort.bam* \ *--bam /shared5/studentprojects/Qikai/Horse_database/fixmate_File/markdup_possort_fixmate_ERR1305964_sort.bam* \

```
    --referenceFasta
/shared5/studentprojects/Qikai/Horse_database/EquCab3_ref/GCF_002863
925.1_EquCab3.0_genomic.fna \
    --runDir Result_American_Paint_Horse1
```

Step3: Execute workflow

Linux script for run Strelka
./runWorkflow.py -m local -j 48

Running on a node with 64 threads, 48 is specified.

Step4: Output head information

Linux script for get head
less -S variants.vcf.gz | grep -e "#" > head

Step5: Update head as sample number

Edit the head using vi or Nano, find the last line #CHROM, and replace the default sample tag.

Linux script for tabix rehead
tabix -r head variants.vcf.gz > variants_rehead.vcf.gz

Step6: Merge 26 horse breeds using bcftools

Linux script for bcftools merge
bcftools merge Result_Akhal_Teke/results/variants/variants_rehead.vcf.gz Result_American_Paint_Horse/results/variants/variants_rehead.vcf.gz Result_American_Standardbred/results/variants/variants_rehead.vcf.gz

Result_Arabian/results/variants/variants_rehead.vcf.gz

Result_Polish_Warmblood/results/variants/variants_rehead.vcf.gz

Result_Badenwurttembergisches_Warmblut/results/variants/variants_rehea

d.vcf.gz

Result_Bayrisches_Warmblut/results/variants/variants_rehead.vcf.gz

Result_German_Riding_Pony/results/variants/variants_rehead.vcf.gz

Result_Holsteiner/results/variants/variants_rehead.vcf.gz

Result_Morgan_Horse/results/variants/variants_rehead.vcf.gz

Result_Franches_Montagnes/results/variants/variants_rehead.vcf.gz

Result_Hannoveraner/results/variants/variants_rehead.vcf.gz

Result_Haflinger/results/variants/variants_rehead.vcf.gz

Result_Icelandic/results/variants/variants_rehead.vcf.gz

Result_Koninklijk_Warmbloed_Paard_Nederland/results/variants/variants_r

ehead.vcf.gz Result_Noriker/results/variants/variants_rehead.vcf.gz

Result_Oldenburger/results/variants/variants_rehead.vcf.gz

Result_Quarter_Horse/results/variants/variants_rehead.vcf.gz

Result_Swiss_Warmblood/results/variants/variants_rehead.vcf.gz

Result_Shetland_Pony/results/variants/variants_rehead.vcf.gz

Result_Thoroughbred/results/variants/variants_rehead.vcf.gz

Result_Trakehner/results/variants/variants_rehead.vcf.gz

Result_UK_Warmblood/results/variants/variants_rehead.vcf.gz

Result_Westfale/results/variants/variants_rehead.vcf.gz

Result_Welsh_Pony/results/variants/variants_rehead.vcf.gz

Exmoor_ponies/variants_rehead.vcf.gz -Oz -o merged_26breed.vcf.gz

Step7: Vcftools Filters and removes Indels

Linux script for vcftools Filters and removes Indels
vcftools --vcf merged_26breed.vcf.gz --minQ 30 --minGQ 30 --remove-

indels --out merged_26breed_minQ30_minGQ30_rmvIndels --recode

'--minQ 30' Set the quality value higher than 30 sites, '--minGQ 30' All genotypes with mass less than 30 were excluded, '--remove-indels' Exclude sites that contain indel.

Step8: vcftools filter2

Linux script for vcftools filter2
vcftools --vcf merged_26breed_minQ30_minGQ30_rmvIndels.recode.vcf --mac 3 --remove-filtered-all --out merged_26breed_minQ30_minGQ30_rmvIndels_mac3_passonly --recode

'--remove-filtered-all' Delete all sites with the FILTER flag instead of the PASS flag.

Step9: vcftools filter3

Linux script for vcftools filter3
vcftools --vcf merged_26breed_minQ30_minGQ30_rmvIndels_mac3_passonly.recode.vcf --max-missing 0.5 --hwe 0.01 --out merged_26breed_minQ30_minGQ30_rmvIndels_mac3_passonly_mm0.5_hwe0.01 --recode

'-- hwe' Sites with p values below the threshold defined by this option are considered not in hwe and are therefore excluded.

'--max-missing' Sites are excluded based on the percentage of missing data