

# The genetic tale of hog species (*Suidae*), inbreeding, genetic load and their demographic history.

## Bachelor Thesis

Nino Menger

### Bachelor Thesis:

Version 1

Submitted on 16/01/2022

Thesis period: 01/04/2022 – 02/01/2023

### Commissioned by:

Wageningen University and research

Animal Breeding and Genomics

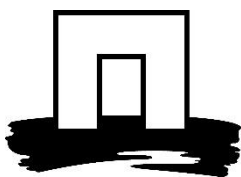
Supervision by: dr. Mirte Bosse

&

University of Applied Science Leiden

Bioinformatics

Supervision by: dr. André Klein



**WAGENINGEN**  
UNIVERSITY & RESEARCH



**hogeschool**  
**Leiden**

## ABSTRACT

Human influences have reduced wildlife population sizes drastically. Small population sizes lead to inbreeding, which can lead to reduced fitness, called inbreeding depression. Additionally, inbreeding can lead to the expression of recessive deleterious alleles. These recessive deleterious alleles accumulate when the population size is bigger and can also be referred to as genetic load. Thus, the current genetic state of a species depends on its demographic history. Hogs, also known as Suids, inhabit a wide range of habitats across Eurasia, all with different demographic histories, which makes Suids an excellent model organism to investigate the effects of demographic history on the amount of inbreeding and genetic load that currently can be measured within the genome.

This project aims to develop a pipeline capable of analyzing the demographic history, inbreeding, and genetic load within different populations. Additionally, a second pipeline will be developed to visualize data structure and quality, to aid in selecting populations and individuals for the earlier-mentioned population genomics pipeline.

These pipelines were built to analyze a multi-sample VCF file containing 424 Suid samples, 405 of which contain appropriate amounts of annotation. Using the sample selection pipeline, a PCA using plink, and a sequencing depth analysis using vcftools were performed. Both analyses were visualized using the R programming language. The PCA resulted in distinct clustering which corresponds with the known history and habitats of Suid populations. Based on the results a total of 62 individuals were selected across 13 distinct groups.

The population genomics pipeline performed an effective population size analysis using SMC++ to get further insight into the demographic history of the groups. These results correspond with the known history of the different groups. Furthermore, inbreeding patterns were analyzed by deriving ROHs. Two methods were tested, bcftools and plink, because there does not exist a golden standard for non-model organisms. Overall, both methods performed relatively similarly, and the choice between them properly depends on the data. European Wild Boars contained relatively high amounts of inbreeding compared to Asian Wild boars, this is likely due to the high population size fragmentation and decrease in Europe. Lastly, the pipeline performed a deleterious mutations analysis using the pCADD method. The analysis found a surprising amount of highly deleterious mutations in European populations compared to Asian populations. This is not as expected since the demographic history shows a larger bottleneck for the European populations during the LGM, which should have led to the purging of these highly deleterious alleles. However, the European population sizes have increased in the last few decades, which could explain an increase in deleterious alleles.

To conclude, using the sample selection pipeline to visualize data structure and quality proved to be a useful asset for the selection of individuals for further analysis. Furthermore, the population genomics pipeline successfully gave more insight into the demographic history, inbreeding, and genetic load of different groups of hogs. Both pipelines can be used on other comparable datasets, like cattle, chicken, and elephant.

## ABBREVIATIONS

<b>IBD</b>	Identity By Decent
<b>ROH</b>	Region of Homozygosity
<b>VCF</b>	Variant Call Format
<b>HMM</b>	Hidden Markov Model
<b>CADD</b>	Combined Annotation-Dependent Depletion
<b>SNV</b>	Single Nucleotide Variant
<b>SNP</b>	Single Nucleotide Polymorphism
<b>SMC++</b>	Sequential Markov Coalescent + Plenty of Unlabeled Samples
<b>TMRCA</b>	Time to Most Recent Common Ancestor
<b>MSEA</b>	Mainland of South East Asia
<b>ISEA</b>	Island of South East Asia
<b>PCA</b>	Principle Component Analysis
<b>LGM</b>	Last Glacial Maximum

# CONTENTS

1	Introduction .....	4
1.1	Inbreeding .....	4
1.2	Deleterious mutations .....	4
1.3	Demographic history .....	5
1.4	Hogs as a model system .....	5
1.5	Aim .....	8
2	Materials and Methods .....	9
2.1	Data .....	9
2.2	Sample selection pipeline .....	10
2.3	Population genomics pipeline .....	12
2.3.1	Effective population size .....	13
2.3.2	Inbreeding analysis .....	13
2.3.3	Deleterious mutation analysis .....	14
3	Results .....	15
3.1	Sample selection pipeline .....	15
3.1.1	PCA analysis .....	15
3.1.2	Sequencing depth .....	16
3.2	Population genomics pipeline .....	18
3.2.1	Effective population size analysis .....	18
3.2.2	Inbreeding analysis .....	19
3.2.3	Deleterious mutations analysis .....	22
4	discussion .....	25
4.1	Population structure and data quality .....	25
4.1.1	Population structure .....	25
4.1.2	Sequencing depth .....	25
4.2	Population genomics analysis .....	26
4.2.1	Effective population size analysis .....	26
4.2.2	Inbreeding .....	26
4.2.3	Deleterious mutation analysis .....	27
5	Conclusion .....	27
6	Data availability .....	28
7	Acknowledgments .....	29
8	References .....	30
9	Supplementary material .....	33

# 1 INTRODUCTION

## 1.1 INBREEDING

The rising ecological footprint of the human race<sup>1</sup> has led to an average decrease of 69% in wildlife population sizes since 1970.<sup>2</sup> Because of this it is more important than ever to be aware of the effects of small population sizes. One of these effects is the increase of inbreeding within these populations. Inbreeding can lead to inbreeding depression, which causes a reduction of fitness due to the loss of heterozygosity.<sup>3</sup> For example, inbreeding within ancient royal families caused an overrepresentation of the recessive 'royal disease' within these families. The disease leads to deadly blood clotting and thus a reduction of fitness.<sup>4</sup>

Traditionally inbreeding was estimated by inspecting pedigree data. However, modern genetics allowed us to measure inbreeding within the genome.<sup>5</sup> Related individuals are likely to have alleles that are inherited from a common ancestor, these alleles are Identical By Decent (IBD). In the case, an individual inherits two of these IBD alleles, a Region of Homozygosity (ROH) arises. When these regions arise they tend to be large. However, due to DNA recombination, the regions will get smaller and more fragmented over generations. Hence, ROHs indicate more recent inbreeding.<sup>6</sup> A ROH can be detected within a Variant Call Format (VCF) file by using a sliding window technique.<sup>7</sup> The homozygosity of the window is evaluated before it slides to the following variants and evaluates again. Another way to detect ROHs is by using a Hidden Markov Model (HMM). This model assigns every variant within a VCF file with one out of two states, ROH and non-ROH. This is achieved by calculating which sequence of states has the highest probability of explaining the observed sequence of variants. The likeliness of a certain state is calculated based on the associated zygosity of the observed variant and on the state that came before.<sup>8,9</sup> The intensity of inbreeding within an individual or population can be estimated by measuring the ROH coverage and the time frame of inbreeding can be derived by measuring the ROH sizes.

## 1.2 DELETERIOUS MUTATIONS

A healthy non-inbred population accumulates various deleterious recessive mutations over time. Because of the recessive nature of these mutations, they do not directly influence the fitness of an individual, and thus can be accumulated by the population over generations until they reach a mutation-selection balance. When the amount of inbreeding in such a population would increase, so would the chance of these mutations becoming homozygous, and thus expressing their deleterious nature. This accumulation of deleterious mutation is called genetic load.<sup>10</sup>

Genetic load or deleteriousness can be estimated by the Combined Annotation Dependent Depletion (CADD) method. This method generates PHRED-like scores for all three potential Single Nucleotide Variants (SNVs) for each genomic position.<sup>11,12</sup> This score can range from 0 to ~95. All scores above 30 represent the top 0.1% of most deleterious SNVs.<sup>13</sup> These scores are generated by using a logistic regression model, which is a machine learning algorithm that can be trained to estimate a binary outcome based on input variables.<sup>14</sup> For the CADD method, this means it can predict whether an SNV is deleterious or not based on the type and location of the SNV. The CADD model is trained by a set of SNVs that are classified as either non-deleterious or deleterious. The non-deleterious SNVs are collected by finding nearly fixed alleles within the target species that differ from the ancestral genome. In theory, these nearly fixed alleles have undergone enough selection to be considered non-deleterious. The deleterious SNVs are collected by generating SNVs de novo. These SNVs have not undergone selection, and thus could contain different levels of deleteriousness.<sup>11-13</sup>

### 1.3 DEMOGRAPHIC HISTORY

As touched on above, the current genetic state of a species is highly dependent on the demographic history of that species. Bottlenecks lead to inbreeding and inbreeding leads to the loss of heterozygosity and can lead to the purging of deleterious alleles.

The amount of inbreeding within a population is directly affected by the effective population size. Effective population size describes the size of a population that contributes to producing the next generation.<sup>15</sup> The historic effective population size can be estimated by using the Sequential Markov Coalescent + Plenty of Unlabeled Samples (SMC++) method. This method uses an HMM to estimate the Time to Most Recent Common Ancestor (TMRCA) for each variant based on a combination of allele frequency distribution and linkage.<sup>16–18</sup> The TMRCA can be translated into historic effective population size, because of the correlation that exists between TMRCA and the effective population size of the ancestral population.<sup>19,20</sup>

The interplay between demographic history, deleterious mutations, and recent inbreeding provides a unique combination of genetic diversity within populations. To better understand how these factors jointly shape patterns of genetic diversity, a model system is required that contains various closely related populations with different evolutionary and demographic histories.

The demographic history of hogs is incredibly diverse. While some species/populations have undergone major bottlenecks, others have not.<sup>21,22</sup> Different levels of inbreeding and deleterious mutations are to be expected within these species and populations. Hence, hogs form an excellent model system to investigate the effects of demographic history on the current genetic state of the species and populations.

### 1.4 HOGS AS A MODEL SYSTEM

Hogs belong to the family of Suidae, which can also be referred to as Suids. Suids are part of the even-toed ungulates superfamily: Suoidae. A closely related pig-like family to the Suids is Tayassuidae, also called peccaries. Suids and Tayassuidae diverged during the late Eocene or the early Oligocene, as seen in Figure 1.<sup>22</sup>

## Phylogenetic timeline in MYA

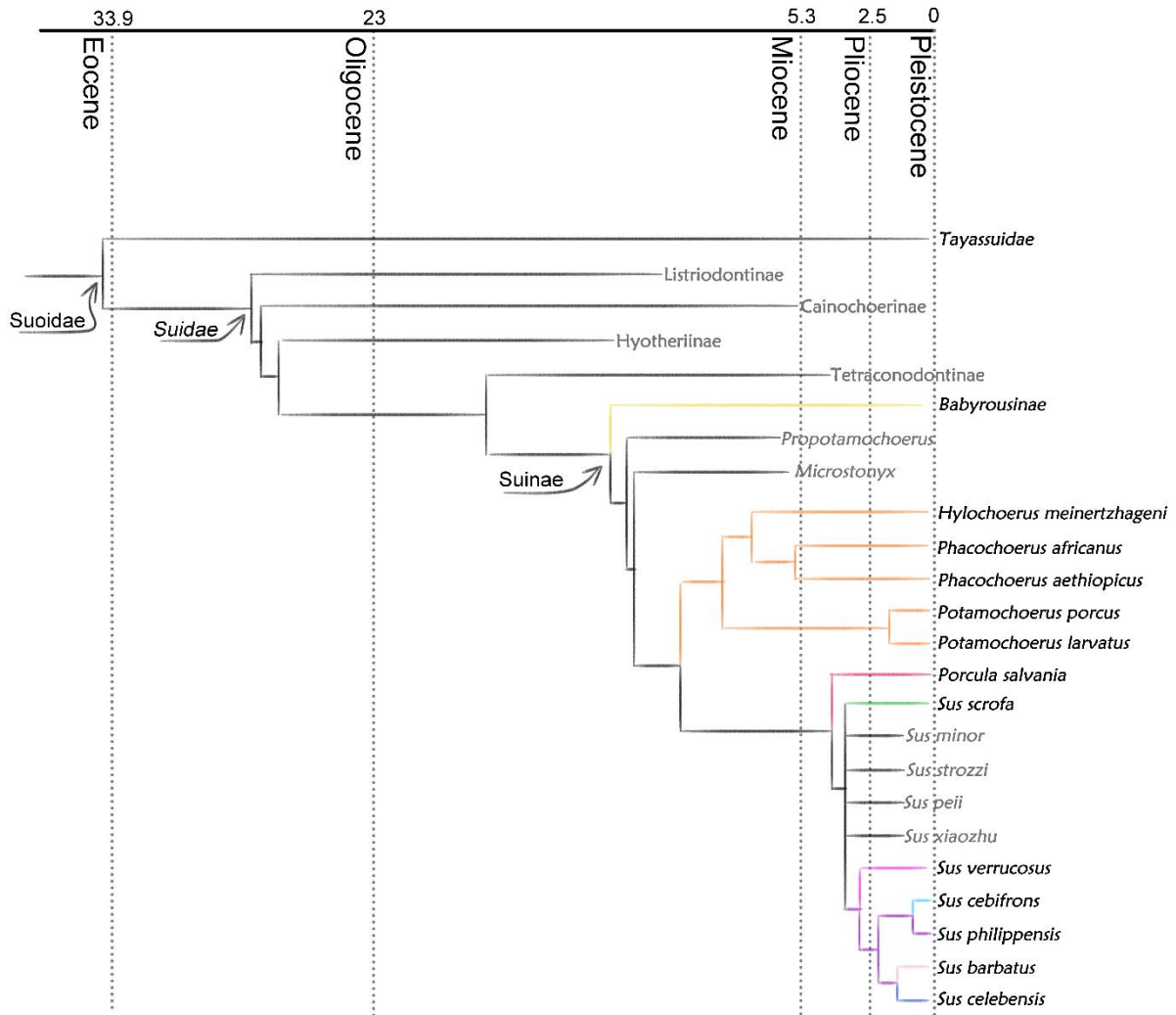
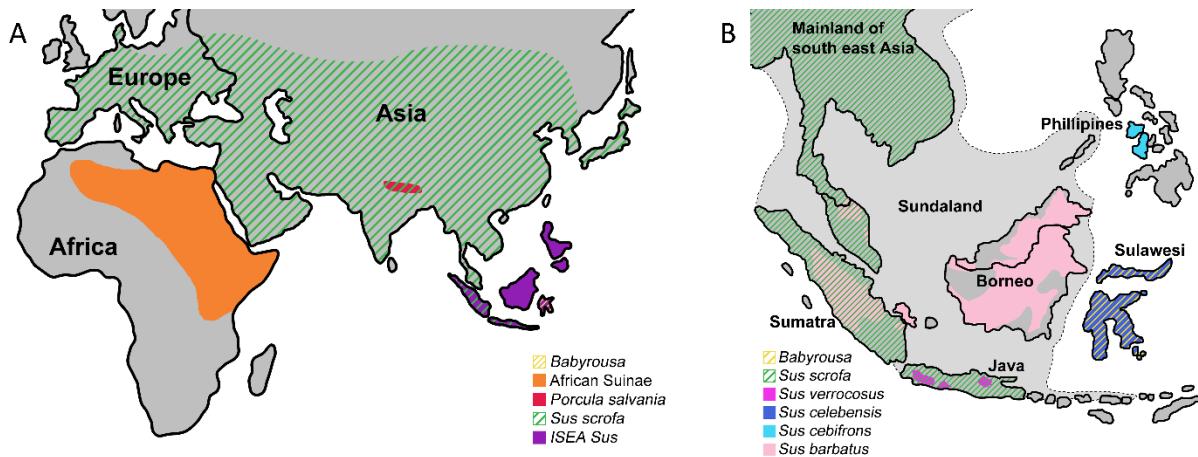


FIGURE 1: SCHEMATIC PHYLOGENETIC TIMELINE ADOPTED FROM FRANTZ L.<sup>22</sup> SPECIES WITH GREYED-OUT LABELS ARE EXTINCT AND COLORED LINES CORRESPOND WITH THE COLOR USAGE IN FIGURE 2.

During the late Oligocene and early Miocene, Suids diverged into four now-extinct subfamilies: Listriodontinae, Cainochoerinae, Hyotheriinae, and Tetraconodontinae. These subfamilies lived in a great variety of habitats throughout Eurasia and Africa, resulting in a vast morphological diversity.<sup>22,23</sup>

Another Suids subfamily, Suinae, emerged during the late Miocene. The Suinae subfamily was highly successful, resulting in the extinction of almost all other Suids subfamilies by the end of the Miocene. The only extant *Suid* species is *Babyrousa*, which is endemic to the island of Sulawesi, see Figure 2. Suinae diverged multiple times throughout Miocene and Pliocene. In Africa, Suinae evolved into the ancestor of the Sub-Saharan Suid species during the Miocene/Pliocene boundary. Extant sub-Saharan Suid species include *Hylochoerus* (forest hogs), *Potamochoerus* (river hogs and bush pigs), and *Phacochoerus* (warthogs). Meanwhile, in Eurasia, the *Sus* genus emerged.<sup>22,23</sup>



**FIGURE 2: A SCHEMATIC OVERVIEW OF SUID DISTRIBUTION ACROSS EURASIA ADOPTED FROM LIU L.<sup>23</sup> MOST OF MAINLAND EURASIA IS SOLELY INHABITED BY *SUS SCROFA*, WITH THE EXERTION OF A SMALL STROKE OF EASTERN ASIA GRASSLAND, WHICH IS INHABITED BY *PORCULA SALVANIA*. AFRICA IS INHABITED BY AFRICAN SUIDS (*HYLOCHOERUS*, *POTAMOCHOERUS*, AND *PHACOCHOERUS*). IN FIGURE B AND MORE DETAILED MAP OF THE ISLANDS OF SOUTH EAST ASIA CAN BE SEEN. THE ISLES OF SUMATRA AND JAVA ARE BOTH INHABITED BY *SUS SCROFA*. HOWEVER, JAVA IS ALSO INHABITED BY *SUS VERRUCOSUS* AND SUMATRA IS ALSO INHABITED BY *SUS BARBATUS*. *SUS BARBATUS* ALSO INHABITS BORNEO. FURTHERMORE, *SUS CEBIFRONS* INHABITS THE ISLES OF NEGROS AND PANAY WITHIN THE PHILIPPINES. LASTLY, SULAWESI IS INHABITED BY *SUS CELEBENSIS* AND *BABYROUSSA*.**

During the Pliocene and Pleistocene, the *Sus* genus diverged into multiple species. At the start of the Pliocene, the divergence of *Sus scrofa* took place on either Mainland South East Asia (MSEA) or Sundaland, a landmass made out of the Islands of South East Asia (ISEA) when the sea level was low. Because of glacial and interglacial periods, the sea level descended and ascended, resulting in the presence and absence of Sundaland for continuous periods.<sup>22,24</sup> Following the emergence of *S. scrofa*, was the dispersal of the ancestor of *Sus verrucosus* to Java. Subsequently, during the early Pleistocene, the ancestor of *Sus cebifrons* migrated to the Philippines, just before the Philippines was permanently excluded from Sundaland due to tectonic activity. During this period the divergence of Sumatran (ISEA) and MSEA *Sus scrofa* took place. During the late Pleistocene, *Sus celebensis* migrated from Borneo to Sulawesi, leaving behind *Sus barbatus* on Borneo.<sup>24</sup>

The periods of low sea level allowed the ISEA *Sus* species to not only migrate to the other ISEA islands but to hybridize with species from other islands as well. Analyzing hybridization can answer interesting questions regarding the past interactions between the different species, but is as well interesting for studying the ongoing speciation of these species. Hybridization can be observed in the genome of the ISEA *Sus* species by performing an admixture analysis. The admixture analysis compares the different *Sus* genomes and scans for pieces of a genome that originated from another *Sus* genome. Thus, revealing the inter-species gene flow caused by hybridization. All ISEA *Sus* species show admixture, especially the Sundaland *Sus* species, *S. scrofa* (Sumatra), *S. verrucosus* (Java), and *S. barbatus* (Borneo). However, admixture can as well be observed between the Sundaland *Sus* species and the non-Sundaland *Sus* species, which are: *S. cebifrons* (The Philippines) and *S. celebensis* (Sulawesi). Furthermore, gene flow between Sumatran *Sus scrofa* and MSEA *Sus scrofa* has as well been revealed.<sup>24</sup>

At the start of the Pliocene *S. scrofa*, also called wild boar, emerged in South East Asia. *S. scrofa* was highly successful in spreading itself throughout Eurasia during the Pleistocene. During this spread of *S. scrofa*, most other *Sus* species living in Eurasia disappeared.<sup>22,23</sup> Admixture analysis shows that hybridization between *Porcula salvania* (pygmy hog) and *S. scrofa* took place during the expansion of



*S. scrofa*. This indicates that *S. scrofa* did not simply replace the species it encountered, but hybridized with them. This may suggest that hybridization drives the evolution of a species, and in the case of *S. scrofa*, giving it the possibility to inhabit a wide range of different kinds of habitats.<sup>23</sup>

The only non-*Sus* Suinae species shown to have survived the *S. scrofa* expansion during the Pleistocene is *P. salvania*. *P. salvania* diverged from the *Sus* genus during the Pliocene and inhabited a wide area of grassland in MSEA during the late Pleistocene. However, in the modern age, its habitat is restricted to a small stroke of grassland and thus is considered critically endangered.<sup>22,23</sup>

Approximately 10.000 years ago domestication of *S. scrofa* happened in multiple places throughout Eurasia independently.<sup>4</sup> The European and Asian pigs were separated until the 18<sup>th</sup>/19<sup>th</sup> century when the Asian pigs were imported into Europe for hybridization to improve traits such as fertility, growth, and fatness in the domestic, but not European wild lineages.<sup>25</sup>

## 1.5 AIM

As described above, all these Suid species have varying demographic histories, and therefore can be expected to have different levels of inbreeding and genetic load. This project aims to develop a pipeline capable of illuminating how demographic history shapes the current genetic state of different species and populations. To achieve this aim, the pipeline will perform a historic effective population size analysis, an ROH analysis, and a deleterious mutations analysis.

Furthermore, to select populations and individuals for above mentioned analyzes, a pipeline capable of visualizing the data structure and quality will be built.

The aim is to build both pipelines in a manner that they can easily be run on other datasets of similar structure.

## 2 MATERIALS AND METHODS

### 2.1 DATA

In this project sequence data from a total of 424 different Suid individuals was used, these individuals can be categorized in an arrangement of different species, geographical origins, and domestication statuses. The samples were sequenced, trimmed, assembled, and called for variants during various projects using the same pipeline. The resulting files are saved within the in-house sequence archive of WUR-ABG. Sequencing was performed using short-read paired-end Illumina technology. The reads have been trimmed using the sickle paired-end trimmer<sup>26</sup> version 1.33, using a length threshold of 50. The trimmed reads were assembled utilizing the MEM algorithm of the Burrow-Wheeler Aligner<sup>27</sup> version 0.7.5a, using the *S. scrofa*/Duroc reference genome 11.1<sup>28</sup> as reference. Variant calling was achieved by using the freebayes caller<sup>29</sup> version 1.3.1 with the reference genome mentioned above. Alleles with a quality score below 10 and alignments with a quality score below 20 were excluded from the variant calling process. Alternate alleles were kept if 0.2 of the total observations represented the alternative, with a minimum of 2 observations. The vcfilter command of the vcflib tool packages<sup>30</sup> version 0.00.2019.07.10 was used to filter out alleles with a quality score below 20. The vcfkeepgeno command of the vcflib tool packages was used to reduce the file size by removing all format fields except for combined- and allelic depth.

This resulted in a multi-sample VCF containing 424 individuals, of which 405 had sufficient annotation information present for the purposes of this project. These 405 individuals originate from different species, populations, and domestication levels. An overview of the individuals and their annotations are shown in Table 1. The majority of the individuals are part of the commercial domesticated category. These 253 pigs are bred with optimal commercial profit in mind, using modern (genetic) techniques. Original Asian and/or European domestic pig breeds were used to create these modern commercial pigs. The second majority of individuals consist of 125 wild and domesticated *S. scrofa* originating from Asia and Europe. The other individuals, a small minority, can be classified as ISEA *Sus*, African *Suinae*, *Babyrussa*, and *Pecari tajacu*. All available breed and population annotations can be found in supplements 1.

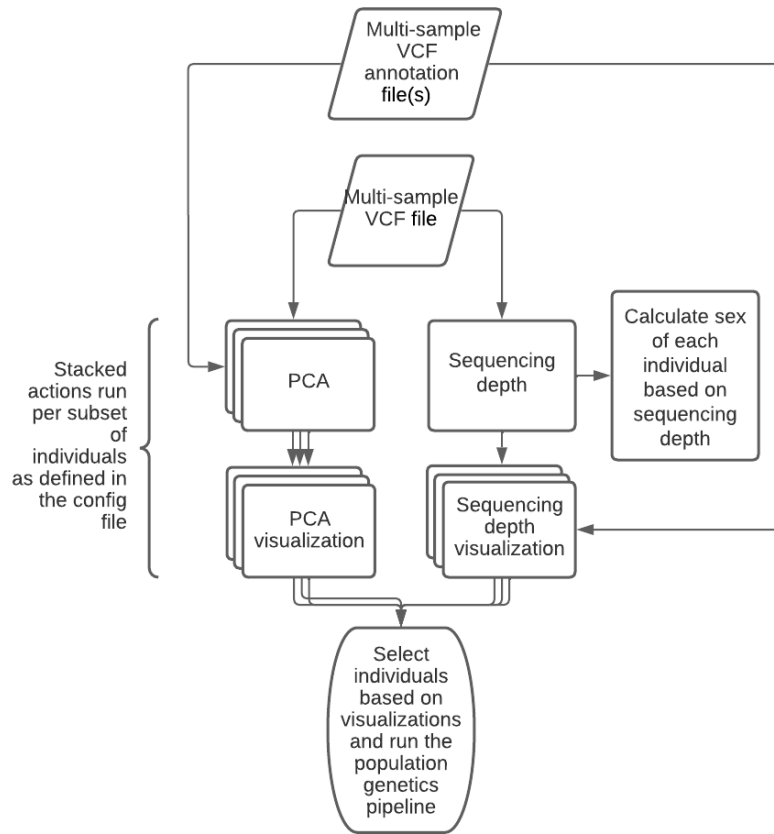
**TABLE 1: IN THIS TABLE, ALL INDIVIDUALS PRESENT WITHIN THE MULTI-SAMPLE VCF USED IN THIS PROJECT ARE SHOWN. THE INDIVIDUALS ARE GROUPED AND COUNTED ACCORDING TO THE AVAILABLE ANNOTATION DATA. THE INDIVIDUALS ARE ANNOTATED USING THREE CATEGORIES, SPECIES, CONTINENTAL ORIGIN, AND DOMESTICATION LEVEL. THE MAJORITY OF THE INDIVIDUALS ARE PART OF THE 'COMMERCIAL' CONTINENTAL ORIGIN. THESE 253 DOMESTIC INDIVIDUALS ARE BRED USING A COMBINATION OF ASIAN AND EUROPEAN DOMESTIC BREEDS FOR COMMERCIAL PURPOSES.**

Species	Continental origin	Domestication status	Number of individuals
<i>Sus scrofa</i>	Asia	Wild	21
		Domesticated	24
	Europe	Wild	35
		Domesticated	45
	Commercial (Mixture of Asian and European breeds)	Domesticated	253
	Islands of South East Asia	Wild	1
<i>Sus barbatus</i>	Asia	Wild	4
<i>Sus cebifrons</i>			7
<i>Sus celebensis</i>			2
<i>Sus verrucosus</i>			2
<i>Porcula salvania</i>			6
African Suidae	Africa		3
<i>Babyrussa</i>	Islands of South East Asia		1
<i>Pecari</i>	America		1
<b>Total</b>			<b>405</b>

## 2.2 SAMPLE SELECTION PIPELINE

As can be seen in Table 1, the groups within the data do not contain an equal amount of individuals. A selection of individuals for each group was performed to prevent a sample size bias and because analyzing all 405 individuals would be computationally unnecessary for the purposes of this project. A Principle Component Analysis (PCA) was used to select clusters of between three and six individuals, with a preference for five individuals. If a cluster consisted of more than six individuals, five of them were selected based on which had the highest sequencing depth.

To make the individual selection process reproducible, a pipeline was developed using Snakemake version 7.8.3. This pipeline can be used to visualize the structure and quality of a multi-sample VCF. An overview of the pipeline can be seen in Figure 3. The pipeline uses a YAML formatted config file to enable the user to specify the following settings: the user can select one chromosome for the PCA analyses and can select a range of chromosomes for the sequencing depth analysis. Furthermore, Including all individuals in one PCA would decrease the visual distance between relatively closely related groups. Therefore, the user can define subsets that need to be visualized separately. The PCA and sequencing depth visualization rules run per predefined subset to give the user the possibility to dynamically add or remove subsets as deemed necessary. The whole pipeline can be downloaded at: <https://git.wur.nl/NinoMenger/vcf-quality-and-population-structure-pipeline>



**FIGURE 3: WORKFLOW VISUALIZING THE STEPS OF THE SAMPLE SELECTION PIPELINE. THE PIPELINE STARTS WITH A MULTI-SAMPLE VCF FILE AND ENDS WITH PCA AND SEQUENCING DEPTH VISUALIZATIONS. AN ANNOTATION FILE WAS USED TO ADD ANNOTATIONS TO THE VISUALIZATIONS. A CONFIG FILE WAS USED TO ALLOW THE USER TO CONFIGURE THE PIPELINE. STACKED STEPS RUN PER BY THE USER-PREDEFINED SUBSET.**

The PCA was performed using the plink program<sup>31</sup> version 1.90b3.38. The allele frequency of closely related loci can be correlated due to physical linkage. To eliminate this bias the linked variants have been pruned from the dataset.<sup>32</sup> For linkage pruning the ‘Indep-pairwise’ command was used with a window size of 50, a window step size of 5 bp, and an  $r^2$  threshold of 0.1. After linkage pruning, the actual PCA was performed by using the ‘pca’ command. Only the first chromosome was included in this analysis. The PCA analysis was visualized using the R programming language version 4.2.0 supported by the ggplot2 and tidyverse packages.

The sequencing depth was derived using the ‘--depth’ command from the vcftools<sup>33</sup> version program version 0.1.16 and visualized as a heatmap using the R programming language version 4.2.0 supported by the tidyr, stringr, ggplot2, and pheatmap packages.

After inspecting the results of the pipeline, 14 different groups of individuals with distinct evolutionary and domestication histories were selected, which can be seen in Table 2. For each of these groups, three to six individuals were selected based on sequencing depth.

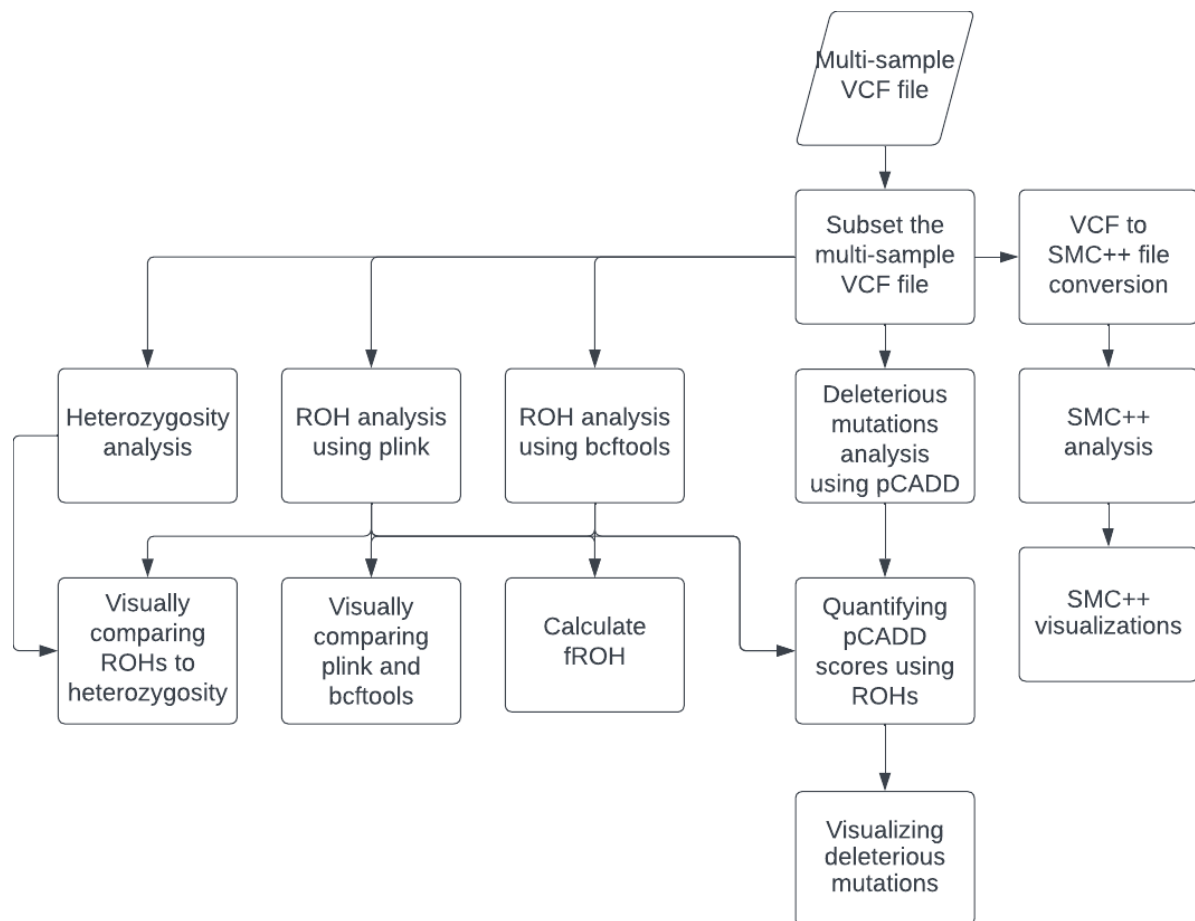
Additionally, it was tested if the sequencing depth could be used to estimate the sex of all individuals. All individuals were assigned male if their X chromosome depth is at least a quarter lower than the mean autosome depth, and the others were assigned female.

**TABLE 2: IN THIS TABLE, ALL GROUPS SELECTED FOR FURTHER ANALYSIS ARE SHOWN. FOR EVERY GROUP, THE AMOUNT OF SELECTED INDIVIDUALS IS SHOWN. A FULL LIST OF SELECTED INDIVIDUALS CAN BE FOUND IN SUPPLEMENTS 1 .**

Species	Continent	Domestication status	Group (Geographical origin or breed)	Number of individuals selected
<i>Sus scrofa</i>	Asia	Wild	North China	3
			South China	5
			Japan	5
		Domesticated	Meishan	5
	Europe	Wild	Netherlands, Mijnweg	4
			Switzerland	5
		Domesticated	Mangalica	5
	Commercial (Asia and Europe mixed)	Domesticated	Duroc	5
			Landrace	5
			Large White	5
<i>Sus barbatus</i>	Islands of South	Wild	-	4
<i>Sus cebifrons</i>	East Asia		-	5
<i>Porcula salvania</i>	Asia		-	6
Total				62

## 2.3 POPULATION GENOMICS PIPELINE

A population genomics pipeline was developed to analyze inbreeding, genetic load, and the demographic history of different groups/populations of Suids. A workflow of the pipeline can be seen in Figure 4. The first step in the pipeline was to subset the multi-sample VCF according to the sample selection made by the user in the YAML formatted config file. Using the sub-stetted data the pipeline performed and visualized an effective population size analysis using SMC++. Thereafter the pipeline performed an ROH analysis using two methods, bcftools and plink. The results of both methods were compared to each other and a bin-wise heterozygosity analysis by vcftools. Lastly, the pipeline performed a deleterious mutation analysis and compared the results to the ROH analysis. The whole pipeline can be downloaded at: <https://git.wur.nl/NinoMenger/vcf-population-genomics-pipeline>



**FIGURE 4: WORKFLOW VISUALIZING THE POPULATION GENETICS PIPELINE.** THE PIPELINE STARTS WITH A MULTI-SAMPLE VCF AND SUBSETS IT ACCORDING TO THE CONFIG FILE CREATED BY THE USER. THE PIPELINE WILL THEN PERFORM AN SMC++ ANALYSIS AND VISUALIZATION. THE PIPELINE WILL USE TWO DIFFERENT METHODS FOR AN ROH ANALYSIS AND WILL VISUALLY COMPARE THESE METHODS TO EACH OTHER. LASTLY, THE PIPELINE WILL PERFORM A DELETERIOUS MUTATION ANALYSIS AND VISUALLY COMPARE THE DELETERIOUS MUTATIONS TO THE ROH ANALYSIS.

### 2.3.1 EFFECTIVE POPULATION SIZE

To get further insight into the demographic past of the different species and populations, a historic effective population size analysis was performed using the SMC++ program<sup>16</sup> version 1.15.5. The estimate command with a mutation rate of 2.5-8. The ftol threshold setting was set on 1e-2, to stop the algorithm when the relative improvement is below the threshold. The number of data points to be derived was set to 20 with the knots parameter. The analysis was visualized using the 'smc++ plot' command and the R programming language version 4.2.0 supported by the ggplot2 and scales packages. A generation time of 5 was used to transform the x-axis from generations to time in years.

### 2.3.2 INBREEDING ANALYSIS

Inbreeding patterns within the different species and populations were revealed by analyzing ROHs. There is no golden standard for ROH analysis in non-model organisms. In this project two different methods were tested, the sliding window method provided by plink<sup>31</sup> version 1.90b3.38 and the HMM method provided by bcftools<sup>34</sup> version 1.9.

The bcftools 'roh' command was used to initiate the analysis. This command was set to use only genotype data (GT) with a safe value of 30 to account for genotype errors. Additionally only ROHs with a quality score above 95 were kept for all future analyses.

The plink method gives the user many possibilities to alter settings. For this project, the window size was set to be 10kb long and a window would be annotated as an ROH if it contains a maximum of 500 missing calls and 18 heterozygous calls. The maximum of heterozygous calls was derived by taking the average heterozygosity of male x chromosomes. Furthermore, only ROHs containing at least 10 kb Single Nucleotide Polymorphisms (SNPs) were kept.

The results from both techniques were visually compared to each other by annotating chromosome 1 of all individuals with the derived ROHs, using the R programming language version 4.2.0 supported by the ggplot2 and cowplot packages. Furthermore, in bins of 1000kb, the number of bases covered by an ROH across all individuals was derived and visualized in a distribution plot.

Both plink and bcftools results were visually compared to a bin-wise nucleotide diversity analysis. This analysis was run using the window-pi command of vcftools<sup>33</sup> version 0.1.16 with bins of 100 kb. The combined results were visualized using the R programming language version 4.2.0 supported by ggplot2 and cowplot.

The proportion of the ROH-covered genome to the entire genome (fROH) was calculated by taking the total length of all ROHs and dividing that by the total genome length. The fROH was visualized using R version 4.2.0 supported by ggplot2.

### 2.3.3 DELETERIOUS MUTATION ANALYSIS

The deleteriousness of the mutations present in all analyzed individuals is derived using the pCADD program which uses the CADD method specialized for pigs.<sup>13</sup>

For all groups/populations, all mutations have been categorized as either heterozygous or homozygous and as either within an ROH or outside an ROH. Thereafter, two proportions were calculated. The first one calculated the proportion for each category, between the high deleterious mutations against all mutations within a group. The second calculated the proportion for each category, between high deleterious mutations against all mutations within that category. These proportions were calculated and visualized as a grouped bar plot using the R programming language supported by the ggplot2 and data.table packages.

### 3 RESULTS

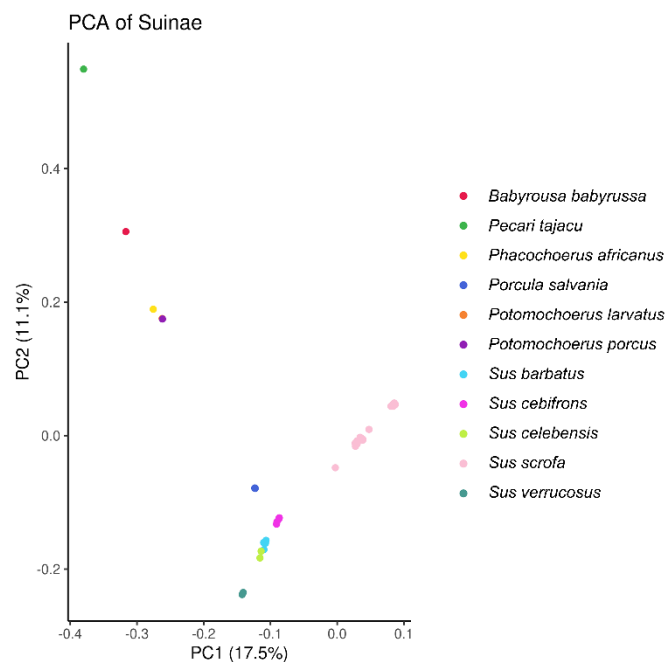
For this project, two pipelines were developed. The first pipeline visualized population structure using a PCA, which could be used to select distinct clusters for further analysis. Additionally, the pipeline performed a sequencing depth analysis, which could be used to select up to five individuals out of a larger cluster for further analysis.

The second pipeline was responsible for a population genomics analysis. This included a historic effective population size analysis, an ROH analysis, and a deleterious mutation analysis.

#### 3.1 SAMPLE SELECTION PIPELINE

##### 3.1.1 PCA ANALYSIS

Using the sample selection pipeline, a PCA analysis was performed to get an insight into the clustering of the different hog species and populations. Below in Figure 5, we can see the most general overview PCA including all 405 samples. The phylogenetic history described earlier can be observed by following the clusters from the high end of PC2 and the low end of PC1 to the low end of PC2 and the low end of PC1.

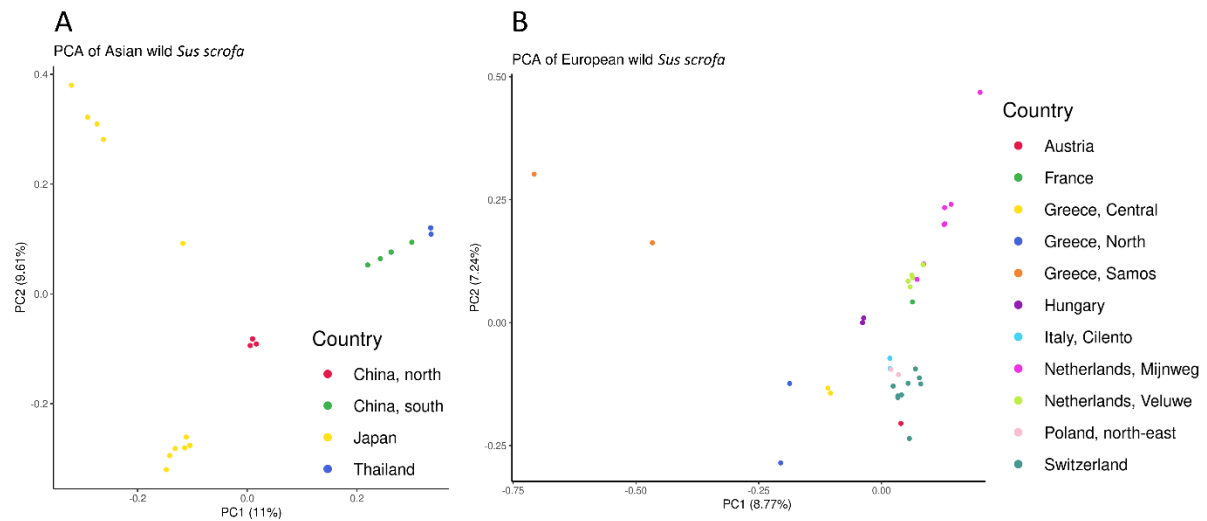


**FIGURE 5: PCA PLOT OF ALL 405 INDIVIDUALS COLORED BASED ON THE SPECIES NAME. THE CLUSTERING RESEMBLES THE EARLIER DISCUSSED PHYLOGENETIC HISTORY OF SUIDS.**

Below in Figure 6A a PCA containing the Asian wild *S. scrofa* is shown. The position of the clusters illustrates the geographical origin of the clusters quite well. The clusters that lie on the high end of PC1 represent South East Asia, by containing Thai and South Chinese individuals in close proximity. The North Chinese individuals are located in the middle of PC1 and the Japanese individuals are located in 2 or 3 clusters on the low end of PC1. Based on this PCA it was deemed interesting to further investigate both Chinese and the Japanese clusters because of the clear distinction between them.



The Japanese cluster on the lower end of PC2 was chosen because of its compact clustering, thus representing a homogeneous group. See Table 2: for an overview of the selected groups.



**FIGURE 6: PCA PLOTS OF EURASIAN *S. SCROFA*, COLORED BASED ON COUNTRY OF ORIGIN. PLOT A CONTAINS THE ASIAN *S. SCROFA*. THE CLUSTERING RESEMBLES THE GEOGRAPHY OF ASIA IN SOME WAY. PLOT B CONTAINS THE EUROPEAN *S. SCROFA*. MOST CLUSTERS CAN BE FOUND ON THE HIGH END OF PC1, EXCEPT FOR THE INDIVIDUALS FROM SAMOS. FURTHERMORE, THE DUTCH INDIVIDUALS SEEM TO BE SHIFTED TO THE HIGH END OF PC2.**

Above in Figure 6B, a PCA including all European wild *S. scrofa* is shown. The majority of the individuals are located on the high end of PC1. However, the individuals from the island of Samos (Greece) are located on the low end of PC1. This can be explained by the geographical location of Samos, which lies closer to Turkey than to Greece. Furthermore, the Dutch samples are shifted toward the upper end of PC2. Based on this PCA it was deemed interesting to further investigate the Swiss cluster since it seems to lie at the center of most European clusters. Additionally, the Netherlands Mijweg cluster was included because of the seemingly shifted position on PC2 compared to the Swiss cluster.

The pipeline successfully generated PCA plots for subsets of the ISEA, Asian domesticated, European domesticated, and commercially domesticated individuals. These plots can be found in supplements 2.1. Briefly, all PCAs show a relatively clear and distinct clustering which is sufficient for selecting populations for further analysis. See Table 2 for an overview of selected groups.

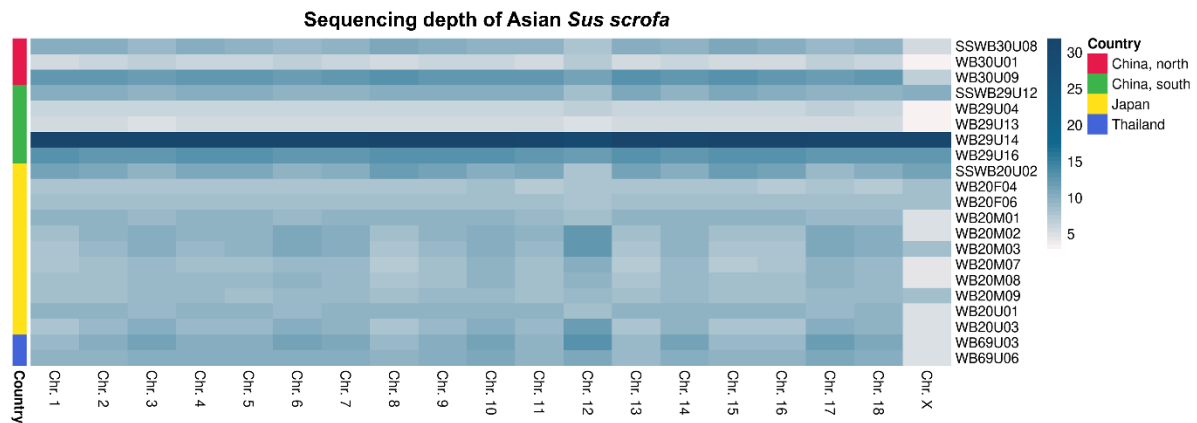
### 3.1.2 SEQUENCING DEPTH

After the group selection step in the pipeline, the specific individuals representing each group could be selected. For this, a sequencing depth analysis was performed to select the individuals with the highest overall sequencing depth. Below in Figure 7 and Figure 8 sequencing depth heatmaps of Asian and European *S. scrofa* are shown respectively.

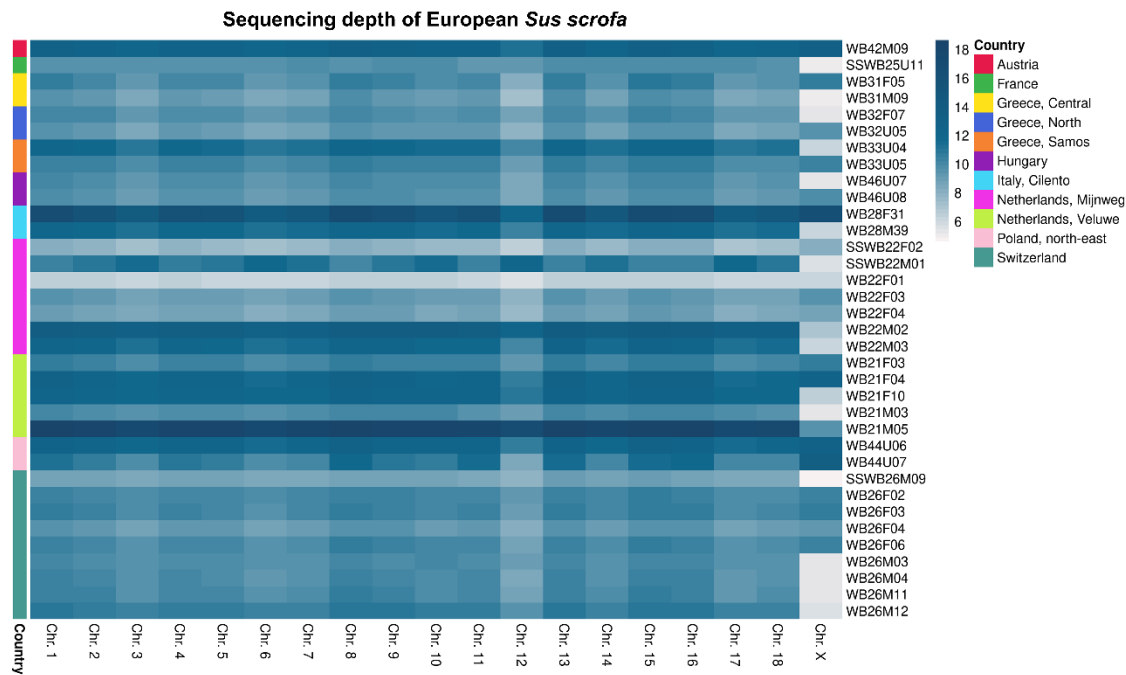
Most individuals seem to have a comparable depth across all chromosomes. However, chromosome 12 seems to deviate a bit compared to the other chromosomes. The sequencing depth seems to be either always lower or higher than the other chromosomes of an individual. Furthermore, the X chromosome highlights an interesting feature. The sequencing depth of the X chromosome is in some cases about half the depth compared to the other chromosomes of the same individual. This is likely caused by the male individuals only having one X chromosome, and thus half the depth. It seems like this information can be used to estimate the sex of all individuals. This has been tested by assigning male to all individuals whose X chromosome depth is at least a quarter lower than the means autosome depth, and female to all others. This algorithm was capable of successfully assigning 155

out of 166 individuals with correct sex annotation. Furthermore, the algorithm assigned sex annotation to 258 previously unknown individuals.

The in 3.1.1 mentioned Swiss *S. scrofa* will be used as an example to illustrate how the sequencing depth heatmap aids in the sample selection process. As can be seen in Figure 6B and Figure 8, there are more than 5 individuals within this cluster, so five of them were selected based on sequencing depth. Out of all nine individuals, SSWB26M09 stands out as lower than the others. The other eight individuals seem to have a comparable sequencing depth. Thereby, it was decided to randomly select 5 individuals from the other eight, which can be seen in supplements 1.



**FIGURE 7: SEQUENCING DEPTH HEATMAP OF ASIAN *S. SCROFA* INDIVIDUALS.** THE HEATMAP INCLUDES THE AVERAGE SEQUENCING DEPTH OF ALL CHROMOSOMES. THE ANNOTATION COLORS ARE ASSIGNED BASED ON THE COUNTRY OF ORIGIN. EACH INDIVIDUAL HAS A COMPARABLE DEPTH ACROSS ITS CHROMOSOMES, WITH THE X CHROMOSOME AS AN EXCEPTION. FURTHERMORE, CHROMOSOME 12 STANDS OUT FOR EITHER HAVING A LOWER OR HIGHER DEPTH COMPARED TO THE OTHER CHROMOSOMES OF THE SAME INDIVIDUAL.



**FIGURE 8: SEQUENCING DEPTH HEATMAP OF EUROPEAN *S. SCROFA* INDIVIDUALS.** THE HEATMAP INCLUDES THE AVERAGE SEQUENCING DEPTH OF ALL CHROMOSOMES. THE ANNOTATION COLORS ARE ASSIGNED BASED ON THE COUNTRY OF ORIGIN. EACH INDIVIDUAL HAS A COMPARABLE DEPTH ACROSS ITS CHROMOSOMES, WITH THE X CHROMOSOME AS AN EXCEPTION. FURTHERMORE, CHROMOSOME 12 STANDS OUT FOR EITHER HAVING A LOWER OR HIGHER DEPTH COMPARED TO THE OTHER CHROMOSOMES OF THE SAME INDIVIDUAL.

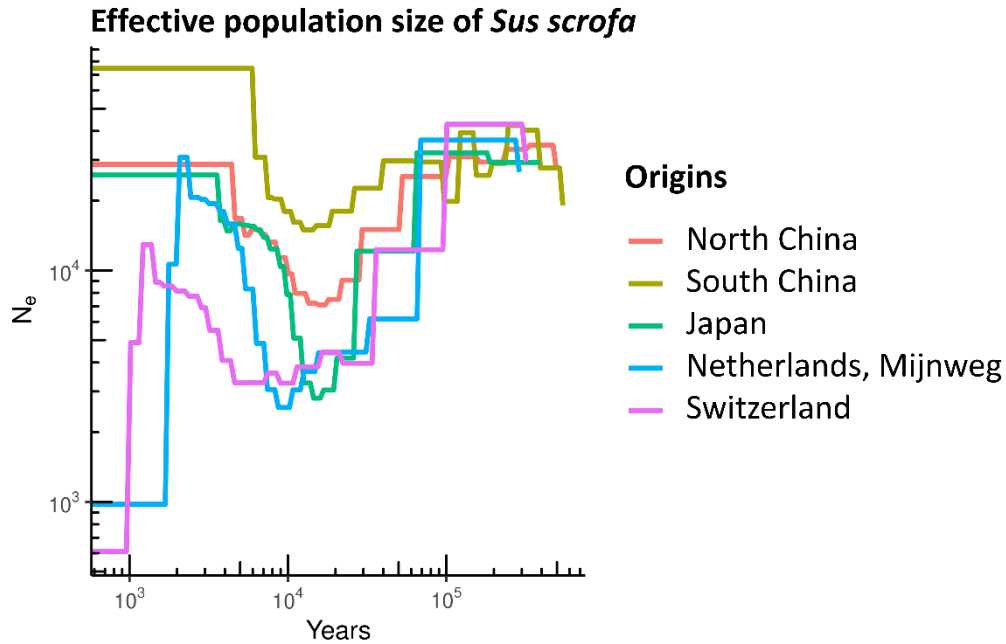
Furthermore, the pipeline successfully generated sequencing depth heatmaps for subsets of the ISEA, Asian domesticated, European domesticated, and commercially domesticated individuals. Briefly, the plots were successful in visualizing the sequencing depth across all chromosomes. These plots formed a great foundation for selecting individuals for further analysis. See supplements 1 for all individuals that were selected for further analysis.

## 3.2 POPULATION GENOMICS PIPELINE

A population genetics pipeline was developed to analyze inbreeding, genetic load, and the demographic history of the selected groups of individuals.

### 3.2.1 EFFECTIVE POPULATION SIZE ANALYSIS

Using the earlier mentioned population genetics pipeline, the historic effective population size of Eurasian *S. scrofa* was derived, which can be seen in Figure 9. The results seem to indicate a shared ancestry between the groups 100.000 years ago. The groups have a common bottleneck around 10.000 years ago. However, the intensity of the bottleneck varies. The South Chinese individuals are the least affected and seem to have the highest recovery. Following the South Chinese are the North Chinese individuals, and lastly the Japanese and European individuals, which seem to be affected an equal amount. The European individuals seem to indicate a recent drop in effective population size around 1000 years ago.

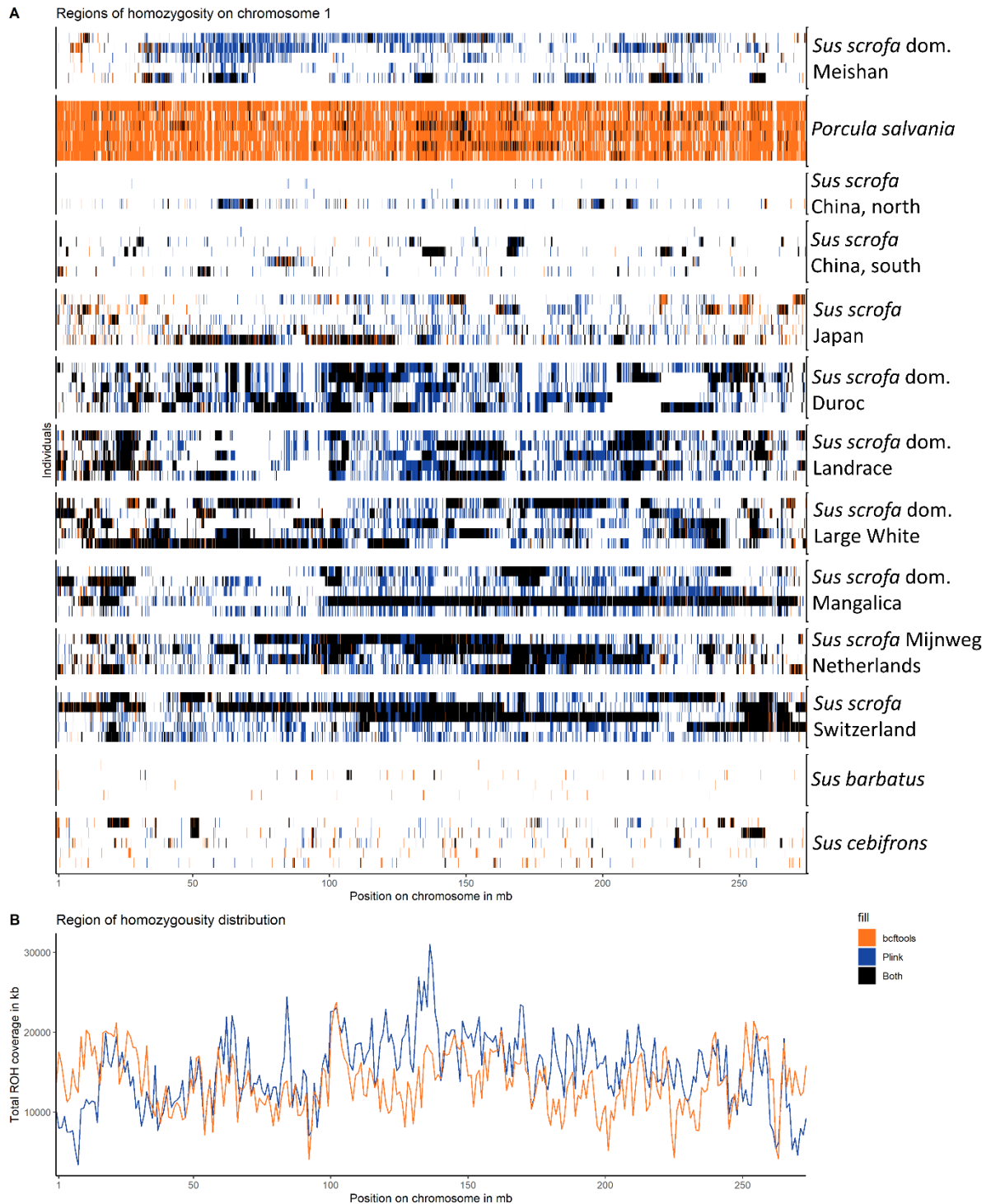


**FIGURE 9: HISTORIC EFFECTIVE POPULATION SIZE PLOT OF EURASIAN *S. SCROFA*.** THE YEARS ARE LOGARITHMICALLY DISPLAYED ON THE X-AXIS AND THE EFFECTIVE POPULATION SIZE IS LOGARITHMICALLY DISPLAYED ON THE Y-AXIS. FIVE EURASIAN *S. SCROFA* ARE SHOWN IN THIS FIGURE. ALL FIVE SEEM TO SHARE AN ANCESTRY ABOUT 100,000 YEARS AGO AND A BOTTLENECK ABOUT 10,000 YEARS AGO. HOWEVER, THE INTENSITY OF THE BOTTLENECK VARIES IN SIZE. THE EFFECTIVE POPULATION SIZE OF THE EUROPEAN INDIVIDUALS HAS BEEN DERIVED UP TO 1000 YEARS AGO WHEN A CLEAR DROP CAN BE WITNESSED.

The pipeline successfully generated effective population size plots for subsets of the ISEA, Asian domesticated, European domesticated, and commercially domesticated individuals. These plots can be found in supplements 2.2. Briefly, the domestic breeds seem to share a clear ancestry with their wild origins. The same is true for the historic effective population size of the commercial domestic breeds, which seem to be close to the European *S. scrofa*. Furthermore, the historic effective population size of the ISEA species are most notably different. These species are not affected by the same bottleneck, their effective population size seems to be quite stable throughout the years. Overall, all effective population size analyses visualize the demographic history of a species/population.

### 3.2.2 INBREEDING ANALYSIS

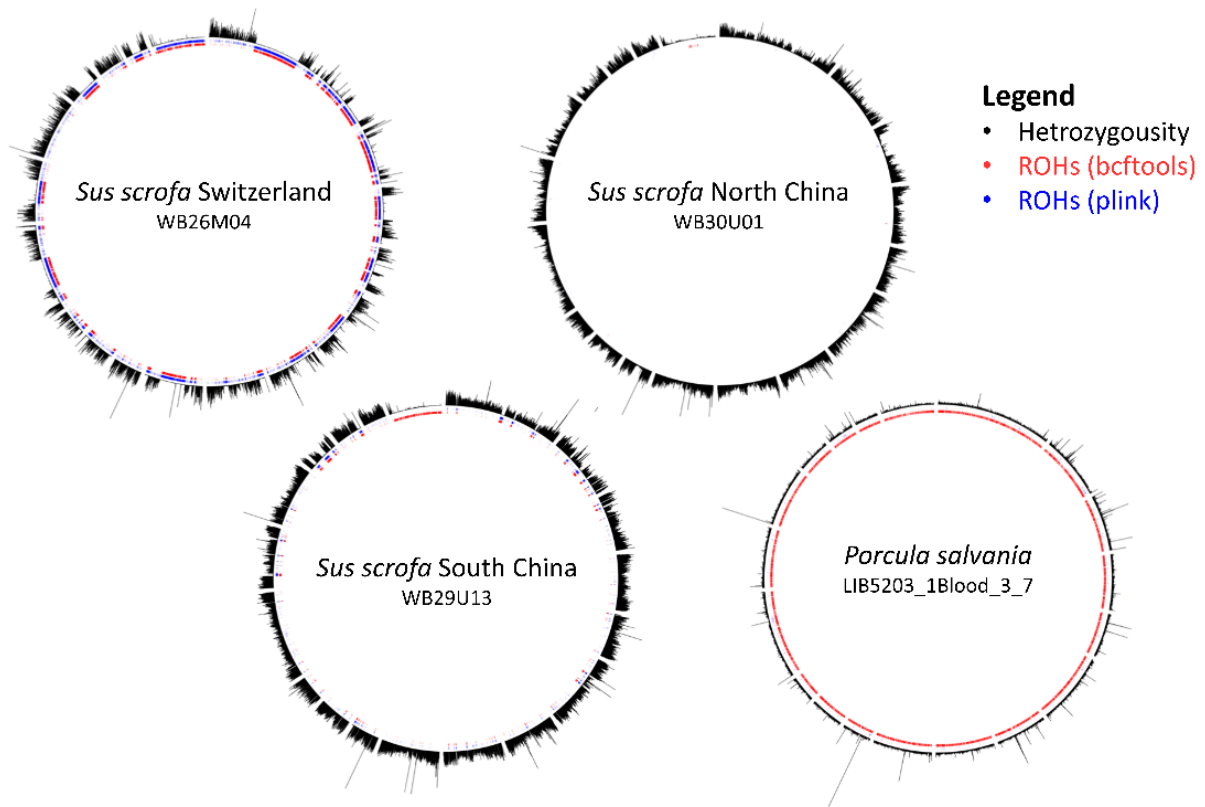
The pipeline performed an ROH analysis to get insight into inbreeding patterns. Two different methods were tested, one provided by plink and the other provided by bcftools. In Figure 10A, a comparison of these techniques of chromosome 1 of all individuals is shown. Even though there are some areas in which both techniques find the same ROHs, there are also a lot of differences. The biggest difference seems to be within the *P. salvania* individuals. Bcftools basically considers everything as an ROH, while plink barely finds any ROHs. In the other samples ROHs found by plink seem to be more present in the middle while ROHs found by bcftools are more present at the edges. In Figure 10B, an ROH distribution plot is shown which visualizes in bins of 1000kb how many bases of all individuals are covered by an ROH. In this plot can be seen that the ROH coverage of bcftools seems to be pretty constant compared to plink which drops at the edges.



**FIGURE 10: VISUAL COMPARISON OF ROH ANALYSES PERFORMED BY BCFTOOLS AND PLINK. FIGURE A DISPLAYS THE ROH COVERAGE ACROSS CHROMOSOME ONE OF ALL INDIVIDUALS. THE ORANGE VISUALIZES THE ROHS FOUND BY BCFTOOLS WHILE THE BLUE VISUALIZES THE ROHS FOUND BY PLINK. BLACK AREAS VISUALIZE ROHS FOUND BY BOTH TECHNIQUES. FIGURE B VISUALIZES THE ROH DISTRIBUTION. ALL BASES COVERED BY AN ROH IN BINS OF 1000KB FOR ALL INDIVIDUALS WERE COUNTED AND VISUALIZED IN THE LINE CHART. THE BLUE LINE REPRESENTS ROH FOUND BY BCFTOOLS AND THE ORANGE LINE REPRESENTS ROHS FOUND BY PLINK.**

To further investigate the ROHs found by the two methods, they were compared to a heterozygosity distribution analysis. Heterozygosity patterns and ROHs of all chromosomes of four individuals can be found in Figure 11. The Swiss individuals show a really clear pattern of regions with a low amount of

heterozygosity being annotated as ROHs by both techniques. The South Chinese individual has more heterozygosity compared to the Swiss individual. However, the small regions of reduced heterozygosity seem to be annotated as ROHs by both techniques. The North Chinese individuals seem to show a comparable amount of heterozygosity to the South Chinese individuals, however not many ROHs have been found within this individual. The most peculiar case is the *P. salvania* individual, which has really low levels of heterozygosity across its genome. As a result, the entire genome is annotated as an ROH by bcftools. However, plink seems to come to the opposite conclusion. The overall low level of heterozygosity can be explained by the distance toward the reference genome used for variant calling. The reference genome originates from domestic *S. scrofa*. This likely caused false positives to arise within the *P. salvania* genome, which drastically reduce the overall heterozygosity.

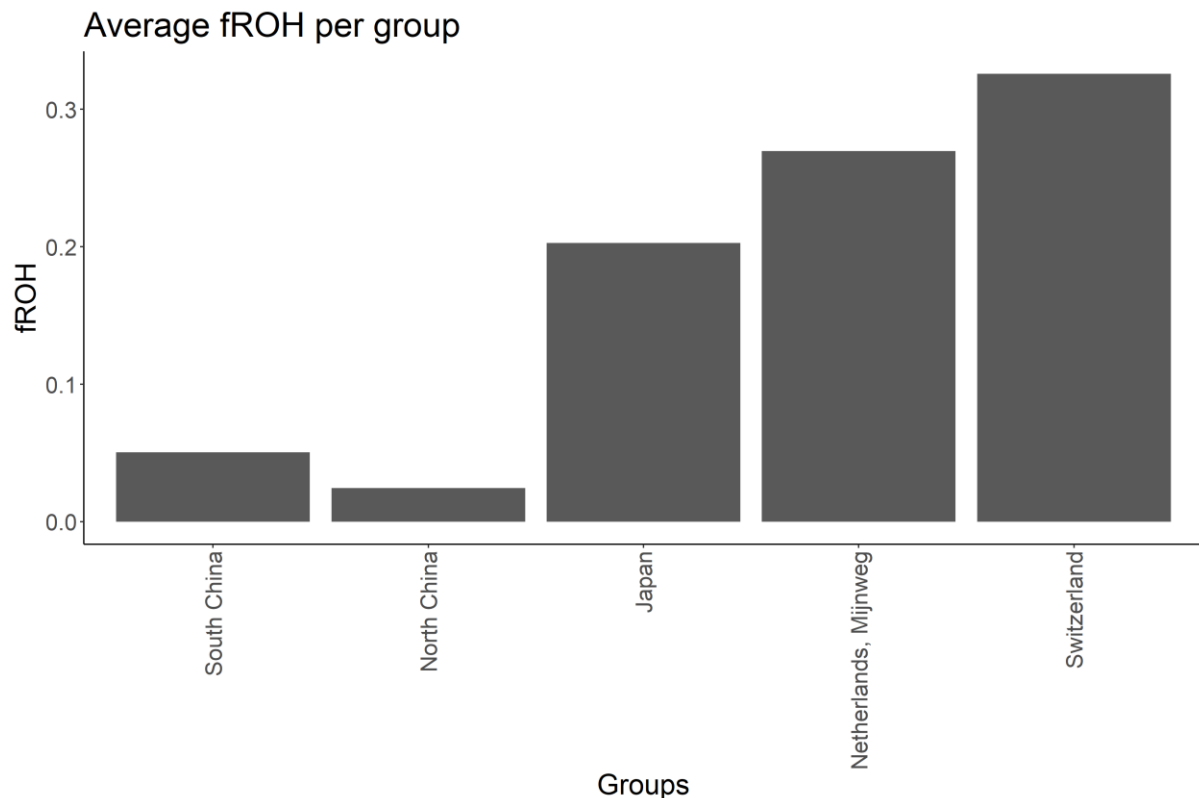


**FIGURE 11: VISUAL COMPARISON OF HETEROZYGOSITY ANALYSIS TO ROH ANALYSES PERFORMED BY BCFTOOLS AND PLINK. EACH CIRCLE REPRESENTS THE GENOME OF ONE INDIVIDUAL. AT THE TOP OF THE CIRCLE LIES CHROMOSOME ONE, AND BY FOLLOWING THE CIRCLE CLOCKWISE THE OTHER CHROMOSOMES CAN BE ENCOUNTERED IN CHRONOLOGICAL ORDER.**

Figure 11 seems to indicate that both methods of determining ROHs perform relatively similarly, especially when the heterozygosity patterns are very clear. However, Figure 10 shows some really clear differences between the techniques. Based on the consistent distribution of ROHs that bcftools seems to provide, it was decided to base further analyses on the bcftools results.

To quantify the amount of inbreeding within all groups, the proportion of the genome covered by ROHs (fROH) was calculated. The average fROH of the Eurasian *S. scrofa* groups has been visualized in Figure 12. Chinese individuals are covered less with ROHs than European and Japanese individuals. The fROH of all other groups can be found in supplements 2.3. Shortly, the average fROH of the domestic breeds seems very comparable to the average fROH of their wild origins. The ISEA seem to have a relatively low fROH. *S. cebifrons* has an fROH of around 0.1 and *S. barbatus* has an fROH of

around 0.02. Lastly, and to no surprise, *P. salvania* has the highest fROH, which can already be expected by looking at Figure 10 and Figure 11.

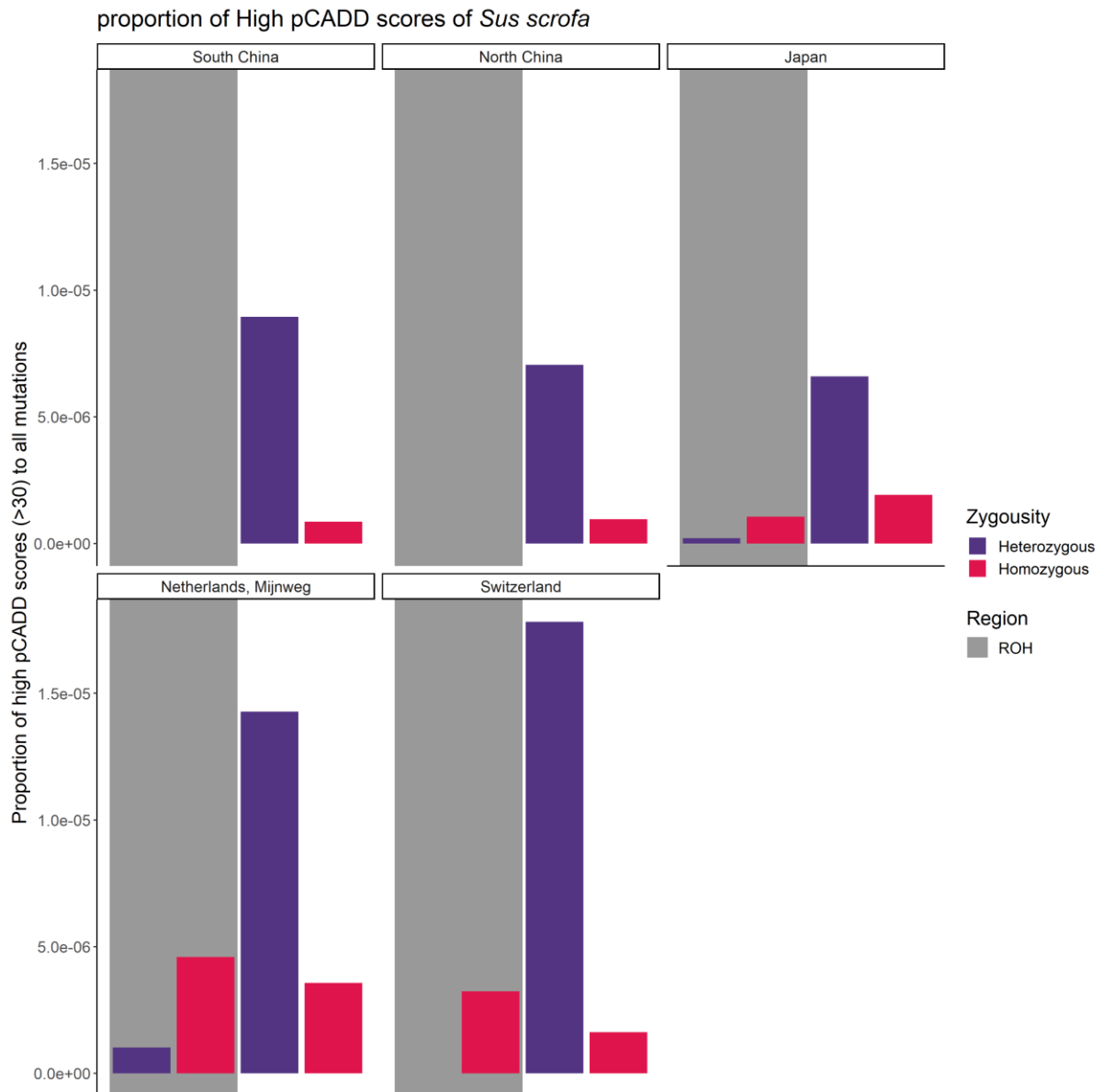


**FIGURE 12: AVERAGE fROH OF EURASIAN *S. SCROFA*. WITH AN fROH OF AROUND 0.05 AND LOWER, THE CHINESE GROUPS HAVE THE LOWEST AVERAGE fROH. THE JAPANESE AND EUROPEAN GROUPS HAVE AN fROH BETWEEN AROUND 0.2 TO AROUND 0.3.**

### 3.2.3 DELETERIOUS MUTATIONS ANALYSIS

To get an insight into the genetic load of the different groups and populations, a deleterious mutation analysis was performed using the pCADD method. This method assigns all mutations with a pCADD score, indicating their deleteriousness. Highly deleterious mutations were quantified by taking the proportion between the highly deleterious mutations to all mutations within a group across multiple categories. The results have been visualized and can be found in Figure 13. most highly deleterious mutations lie outside of ROHs. This is likely due to the underrepresentation of ROH coverage increasing the chances of any mutation lying outside of ROHs instead of in them. Furthermore, most highly deleterious mutations outside of ROHs are heterozygous. The opposite is true for highly deleterious mutations inside ROHs, which are mostly homozygous. Furthermore, proportionally the European groups have more highly deleterious mutations compared to the Asian groups.

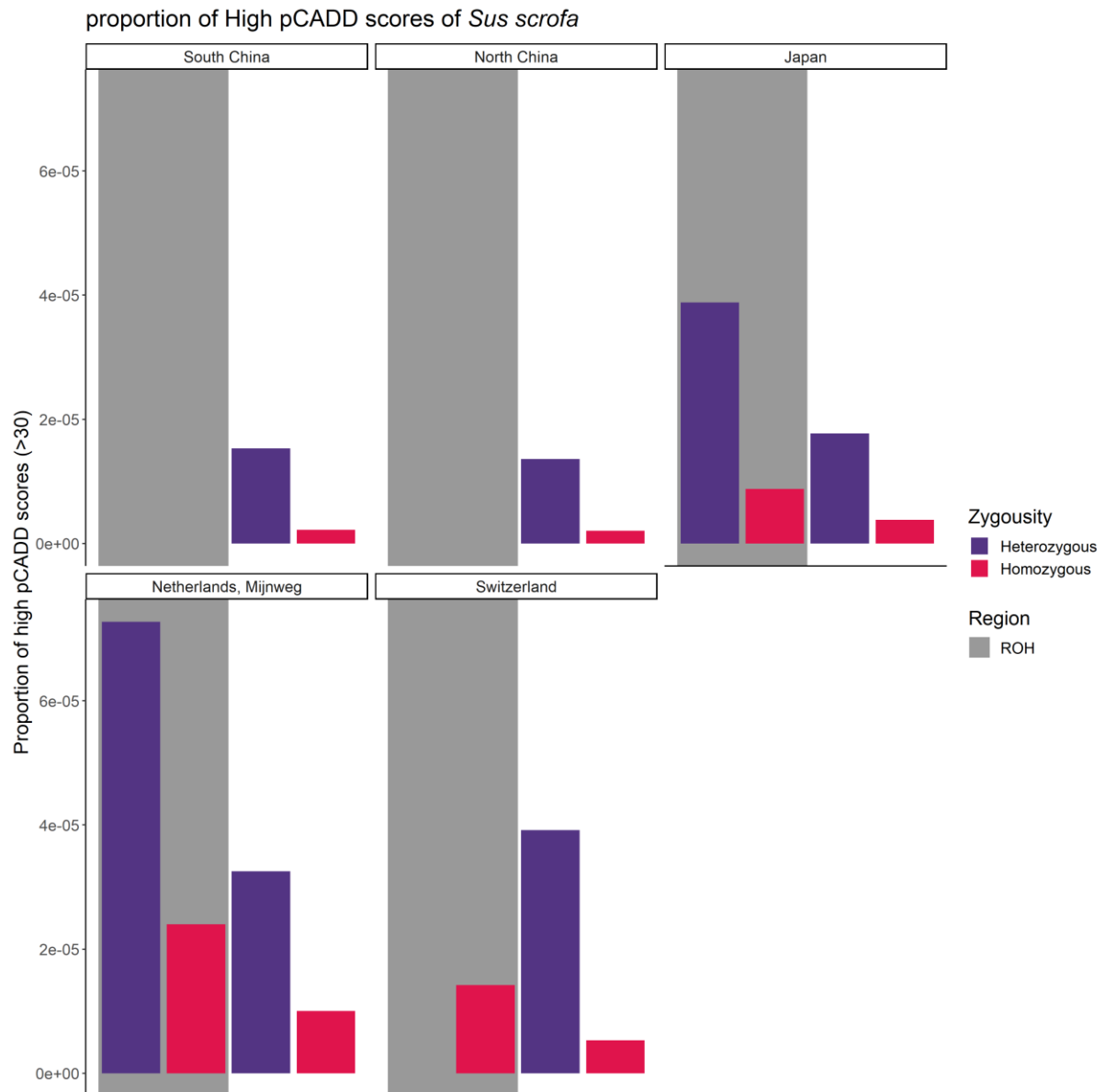
The pipeline successfully generated the same visualization for the other groups as well, which can be found in supplements 1.4.1. The other groups display similar patterns as described above. However, the most notable differences are the *S. barbatus* and *S. cebifrons*, which contain an almost equal amount of heterozygous and homozygous highly deleterious mutations outside of ROHs.



**FIGURE 13: GROUPED BAR PLOT VISUALIZATION OF HIGHLY DELETERIOUS MUTATIONS IN EURASIAN *S. SCROFA*.** FOR THIS VISUALIZATION, ALL MUTATIONS HAVE BEEN CATEGORIZED AS EITHER HETEROZYGOUS OR HOMOZYGOUS AND AS WITHIN AN ROH AND OUTSIDE AN ROH. WITHIN EACH OF THESE CATEGORIES, THE PROPORTION OF HIGHLY DELETERIOUS MUTATIONS TO ALL MUTATIONS WITHIN THE OVERARCHING GROUP WAS CALCULATED AND VISUALIZED IN THIS FIGURE. MUTATIONS WERE CONSIDERED HIGHLY DELETERIOUS IF THEY HAD A pCADD SCORE HIGHER THAN 30.

To get more insight into deleterious mutations, they were quantified in another way. The deleterious mutations were categorized, and for all categories, the proportion of highly deleterious mutations to all mutations within that category was calculated. The visualization can be found in Figure 14. Calculating the proportion this way results in proportional more highly deleterious mutations within ROHs in European and Japanese *Sus scrofa*. The Chinese groups do not follow this same pattern, this is because they contain no high mutations within their ROHs. The same applies to the heterozygous high mutations within the Swiss group.





**FIGURE 14: GROUPED BAR PLOT VISUALIZATION OF HIGHLY DELETERIOUS MUTATIONS IN EURASIAN *S. SCROFA*.** FOR THIS VISUALIZATION, ALL MUTATIONS HAVE BEEN CATEGORIZED AS EITHER HETEROZYGOUS OR HOMOZYGOUS AND AS WITHIN AN ROH OR OUTSIDE AN ROH. WITHIN EACH OF THESE CATEGORIES, THE PROPORTION OF HIGH MUTATION TO ALL MUTATIONS WITHIN THE CATEGORY WAS CALCULATED AND VISUALIZED IN THIS FIGURE. MUTATIONS WERE CONSIDERED HIGH IF THEY HAD A pCADD SCORE HIGHER THAN 30.

The pipeline successfully generated the same visualization for the other groups, which can be found in supplements 2.4.2. The patterns described above can also be found within the other groups as well.

## 4 DISCUSSION

### 4.1 POPULATION STRUCTURE AND DATA QUALITY

#### 4.1.1 POPULATION STRUCTURE

The sample selection pipeline was built to visualize a multi-sample VCF and select individuals for further analysis. First, a PCA analysis was used to visualize population structure and select clusters for further analysis. The PCA analysis seems to represent the known demography and phylogeny of *Suids* quite well. For example, a clear separation between the Northern Chinese and Southern Chinese *S. scrofa* clusters can be seen. This deep split is likely caused by the geographical separation of these populations during The Last Glacial Maximum (LGM), commonly known as the ice age.<sup>21</sup> Furthermore, the Thai individuals cluster close to the Southern Chinese individuals, which geographically makes sense. The Japanese individuals are separated into two or three clusters. This separation could be explained by the archipelagic nature of Japan. However, at the time of writing the exact origin of the Japanese individuals is unknown. It would be very interesting for a future project to uncover this origin and see how these origins cluster in a PCA. This could form the basis for a project aiming to get further insight into the speciation within Japan.

The PCA of European *S. scrofa* seems to explain the geographical origins quite well, just like the Asian variant. For a future project, it would be interesting to compare the distance between PCA clusters to the geographical distance between these clusters. An analysis like this could give some insight into the correlation between geographical distance and genomic distance, and what influences this correlation.

Overall, the PCA was proven to be an effective tool for visualizing population structure and using those visualizations as a base to select groups for the population genomics analysis. However, the population genomics analysis revealed a problem that occurs because of the distance certain groups have to the universally used domestic *S. scrofa* reference genome, namely *P. salvania*. It is a good practice to be aware of these technical aspects, and therefore it would be sensible to include an analysis that visualizes the distance to the universally used reference genome, like a phylogenetic analysis that highlights which group the reference genome was based on.

#### 4.1.2 SEQUENCING DEPTH

The sequencing depth was derived to select good-quality individuals for further analysis. The first thing that can be noticed is the consequent differing depth of chromosome 12 compared to the other chromosomes, which can be explained by the relatively high GC% of this chromosome. Chromosome 12 of the reference genome has a GC% of 47.7%, which is 5.4% higher than the average of 42.3%.<sup>35</sup>

Another pattern that can be observed is the sequencing depth of the X chromosome of some individuals which is about half compared to the other chromosomes of that individual. This can be explained by the male individuals only having one copy of the X chromosome. Using sequencing depth and a simple algorithm has proven to correctly guess the sex of 155 out of 166 individuals with sex annotation. However, it is known that the available sex annotation has its flaws, future research can validate this method of estimating the sexes by testing it against a validation set with trustworthy sex annotations. Furthermore, future projects could improve on the method by not using all the chromosomes as a baseline but just the first three for example. This would reduce computation costs.

Overall selecting individuals for further analysis based on sequencing depth turned out to be effective. However, even though it is good to be aware of the above mentioned patterns, it does not seem to

be necessary to derive the sequencing depth from every chromosome. Future projects could consider only using the first three chromosomes for example.

## 4.2 POPULATION GENOMICS ANALYSIS

### 4.2.1 EFFECTIVE POPULATION SIZE ANALYSIS

To obtain insight into the demographic history of the different groups, a historic effective population size analysis was performed using SMC++. The Eurasian *S. scrofa* seems to share a clear shared ancestry from more than 100.000 years ago. *S. scrofa* is thought to have expanded from Southern Eastern Asia to Europe during the Pleistocene. The timeframe of the found shared ancestry corresponds with the known expansion of *S. scrofa*. The demography of the species seems to split around 20.000 years ago. The LGM occurred during this same timeframe, which is known for splitting populations by making the terrain between them untraversable. The effects of the LGM were more significant in Europe than in Asia, the results seem to confirm this.<sup>21</sup> The bottleneck within the European populations is more significant than the Asian populations. Furthermore, the bottleneck is least present in the Southern Chinese populations, which also corresponds with earlier findings.<sup>21</sup> The past effective population size of Japanese individuals lies between European and Asian. To confirm these findings, a future project could focus on deriving the heterozygosity between ROHs. These bottlenecks likely caused ROHs to emerge into the genome. However, due to DNA recombination, these hypothetical ROHs have been fragmented by now, and thus influence the levels of heterozygosity between the current ROHs. Hence, lower levels of heterozygosity between ROHs are to be expected in European individuals. However, this theory will have to be tested.

### 4.2.2 INBREEDING

ROHs have been derived to get an insight into inbreeding patterns. Two methods were tested to get further knowledge of which technique to use for non-model organisms. Both techniques found believable ROHs if the individual has clear heterozygosity patterns. However, on the opposite spectrum, when an individual has vague heterozygosity patterns, then both tools handle the problem oppositely. *P. salvania* has really low and monotone heterozygosity patterns. While bcftools annotates practically the entire genome as an ROH, plink does not seem to find any ROHs. As mentioned before, this specific problem may be caused by the distance between *P. salvania* and the reference genome, which can be solved by using a more comparable reference genome.

The choice between one method or the other depends heavily on the usage of the user. Bcftools is more straightforward to use with fewer options to optimize. However, plink offers many options the user can use to define what an ROH looks like. Having many options is great for when the user wants to control how stringent the algorithm works. However, having more options also raises the question: how do you define an ROH? The answer to this question is important since all results depend on how this question is answered. Future projects should be devoted to answering this question by investigating the phenomenon of ROHs further.

For this project bcftools was chosen to further investigate ROH patterns within the groups because of the more consistent distribution. The proportion of ROH coverage found by bcftools against the length of the entire genome (fROH) was calculated. The fROH of Chinese individuals is lower than that of European and Japanese individuals. Which suggests more inbreeding within the latter three groups. The earlier mentioned effective population size results showed a drop in effective population size about 1000 to 2000 years ago in European individuals. Additionally, it is known that wildlife populations endured more fragmentation and decrease of populations than Asian wildlife, which could explain the increase in inbreeding.<sup>36,37</sup>

### 4.2.3 DELETERIOUS MUTATION ANALYSIS

To get insight into the genetic load of different populations a deleterious mutations analysis was performed using the pCADD method. The pCADD scores were quantified using the ROH results. The groups contain more highly deleterious mutations outside of ROHs than inside ROHs. Since the bigger part of the genome is not an ROH, the chances are higher of any mutation lying outside of ROHs. Furthermore, there are more homozygous mutations inside ROHs than heterozygous mutations, which can be explained by the homozygous nature of an ROH. Proportionally, European individuals contain more highly deleterious mutations than Asian individuals. Because of the earlier discussed bottleneck during the LGM, which was more significant within European populations, the opposite pattern could be expected. Bottlenecks lead to inbreeding, which should lead to the purging of deleterious alleles. The unexpected outcome could be explained by the proportional nature of the analysis. If Asian individuals have many overall mutations then that would reduce the proportion of high mutations. A future project could clarify this by deriving the total amount of mutations in each group and comparing these.

Previously was established that the chance of finding a mutation inside an ROH is smaller than finding it outside the ROH. However, the results seem to hint that even though not many mutations lie within ROHs, the mutations that lie within an ROH, have a proportionally high deleterious score. This phenomenon can be observed within the European and Japanese groups. Since ROHs are an indication of recent inbreeding, this could hint at the recent acquisition of these deleterious alleles. What makes this more convincing is the fact that this effect seems the strongest on heterozygous alleles. Because of the homozygous nature of an ROH, the heterozygous alleles within an ROH are expected to be gained recently. These recently gained deleterious alleles could be explained by the recent increase in European *S. scrofa* population sizes, that occurred in the last few decades.<sup>36,38</sup> However, future research will have to confirm this theory by including more European populations with well-known recent demographic histories and comparing these to each other.

## 5 CONCLUSION

To conclude, the sample selection pipeline with the PCA and sequencing depth analysis served its purpose well. It is capable of providing insight into the population structure and data quality. This insight was proven successfully to select samples for further analysis.

The population genomics pipeline was proven successful to give an insight into the demographic history, inbreeding patterns, and deleterious mutations. This gained insight can form the base to further investigate the patterns encountered in the populations, and the interplay of these population characteristics.

Even though improvements will always be possible, both pipelines were successfully made to be reproducible by the department on other datasets. These datasets can include but are not limited to cattle, chicken, and elephant. The possibilities are endless.

## 6 DATA AVAILABILITY

The data used in this project is in-house and available. It can be requested at either WUR-ABG or TopigsNorsvin (<https://topignorsvin.nl/contact-ons/>). At the time of writing, these organizations can be reached at:

### **TopigsNorsvin:**

Phone number: +31 411 64 88 70

E-mail address: [info@topignorsvin.com](mailto:info@topignorsvin.com)

Contact page: <https://topignorsvin.nl/contact-ons/>

### **WUR-ABG:**

WUR-ABG can be contacted by filling in the contact form that can be found at the following website:

<https://www.wur.nl/en/research-results/chair-groups/animal-sciences/cluster-population-dynamics-and-genomics/animal-breeding-and-genomics-group/about-us.htm>

## 7 ACKNOWLEDGMENTS

For giving me the opportunity and honor to work on an interesting and ecologically important project, I want to thank Mirte Bosse. Mirte guided me through the complex subject matter and gave me my first insight into population genomics. By allowing me to work on this project she gave me new insight into what I want for my future career. Lastly, Mirte gave me the freedom to challenge myself and take the project in directions I found interesting. For all of these things, I am incredibly thankful for Mirte.

Secondly, I want to thank my mentor assigned by The University of Applied Science, André Klein. André was not just my mentor for my first bioinformatics project, but also this project, my last. I want to thank André for guiding me through the important parts of the project and the technicalities that came with it.

Lastly, I want to thank everyone from the Monday genomics meeting. Listening to their presentations allowed me to learn a lot of new things. Furthermore, they all showed a sincere interest in my project and gave me priceless feedback.

## 8 REFERENCES

1. Cafaro P, Hansson P, Götmark F. Overpopulation is a major cause of biodiversity loss and smaller human populations are necessary to preserve what is left. *Biol Conserv.* 2022;272:109646. doi:10.1016/j.biocon.2022.109646
2. Almond R.E.A, Grooten M, Juffe Bignoli D, Petersen T. *Living Planet Report 2022 – Building a Naturepositive Society.*; 2022.
3. Nonaka E, Sirén J, Somervuo P, Ruokolainen L, Ovaskainen O, Hanski I. Scaling up the effects of inbreeding depression from individuals to metapopulations. *Journal of Animal Ecology.* 2019;88(8):1202-1214. doi:10.1111/1365-2656.13011
4. Rogaev EI, Grigorenko AP, Faskhutdinova G, Kittler ELW, Moliaka YK. Genotype Analysis Identifies the Cause of the “Royal Disease.” *Science (1979).* 2009;326(5954):817-817. doi:10.1126/science.1180660
5. Doekes HP, Veerkamp RF, Bijma P, de Jong G, Hiemstra SJ, Windig JJ. Inbreeding depression due to recent and ancient inbreeding in Dutch Holstein–Friesian dairy cattle. *Genetics Selection Evolution.* 2019;51(1):54. doi:10.1186/s12711-019-0497-z
6. Bosse M, Megens H, Derks MFL, Cara ÁMR, Groenen MAM. Deleterious alleles in the context of domestication, inbreeding, and selection. *Evol Appl.* 2019;12(1):6-17. doi:10.1111/eva.12691
7. Meyermans R, Gorssen W, Buys N, Janssens S. How to study runs of homozygosity using PLINK? A guide for analyzing medium density SNP data in livestock and pet species. *BMC Genomics.* 2020;21(1):94. doi:10.1186/s12864-020-6463-x
8. Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-Smith C, Durbin R. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics.* 2016;32(11):1749-1751. doi:10.1093/bioinformatics/btw044
9. Eddy SR. What is a hidden Markov model? *Nat Biotechnol.* 2004;22(10):1315-1316. doi:10.1038/nbt1004-1315
10. Robinson J, Kyriazis CC, Yuan SC, Lohmueller KE. Deleterious Variation in Natural Populations and Implications for Conservation Genetics. *Annu Rev Anim Biosci.* 2023;11(1). doi:10.1146/annurev-animal-080522-093311
11. Groß C, de Ridder D, Reinders M. Predicting variant deleteriousness in non-human species: applying the CADD approach in mouse. *BMC Bioinformatics.* 2018;19(1):373. doi:10.1186/s12859-018-2337-5
12. Kircher M, Witten DM, Jain P, O’Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet.* 2014;46(3):310-315. doi:10.1038/ng.2892
13. Groß C, Derks M, Megens HJ, et al. PCADD: SNV prioritisation in *Sus scrofa*. *Genetics Selection Evolution.* 2020;52(1). doi:10.1186/s12711-020-0528-9
14. Ranganathan P, Pramesh C, Aggarwal R. Common pitfalls in statistical analysis: Logistic regression. *Perspect Clin Res.* Published online 2017:148-151.

15. Wang J, Santiago E, Caballero A. Prediction and estimation of effective population size. *Heredity (Edinb)*. 2016;117(4):193-206. doi:10.1038/hdy.2016.43
16. Terhorst J, Kamm JA, Song YS. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat Genet*. 2017;49(2). doi:10.1038/ng.3748
17. Harris K, Sheehan S, Kamm JA, Song YS. Decoding Coalescent Hidden Markov Models in Linear Time. In: ; 2014:100-114. doi:10.1007/978-3-319-05269-4\_8
18. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. *Nature*. 2011;475(7357):493-496. doi:10.1038/nature10231
19. Mailund T, Dutheil JY, Hobolth A, Lunter G, Schierup MH. Estimating Divergence Time and Ancestral Effective Population Size of Bornean and Sumatran Orangutan Subspecies Using a Coalescent Hidden Markov Model. *PLoS Genet*. 2011;7(3):e1001319. doi:10.1371/journal.pgen.1001319
20. Tavaré S, Balding DJ, Griffiths RC, Donnelly P. Inferring Coalescence Times From DNA Sequence Data. *Genetics*. 1997;145(2):505-518. doi:10.1093/genetics/145.2.505
21. Groenen MAM, Archibald AL, Uenishi H, et al. Analyses of pig genomes provide insight into porcine demography and evolution. *Nature*. 2012;491(7424):393-398. doi:10.1038/nature11622
22. Frantz L, Meijaard E, Gongora J, Haile J, Groenen MAM, Larson G. The Evolution of Suidae. *Annu Rev Anim Biosci*. 2016;4(1):61-85. doi:10.1146/annurev-animal-021815-111155
23. Liu L, Bosse M, Megens HJ, et al. Genomic analysis on pygmy hog reveals extensive interbreeding during wild boar expansion. *Nat Commun*. 2019;10(1):1992. doi:10.1038/s41467-019-10017-2
24. Frantz LA, Schraiber JG, Madsen O, et al. Genome sequencing reveals fine scale diversification and reticulation history during speciation in *Sus*. *Genome Biol*. 2013;14(9):R107. doi:10.1186/gb-2013-14-9-r107
25. Bosse M, Megens HJ, Madsen O, et al. Untangling the hybrid nature of modern pig genomes: a mosaic derived from biogeographically distinct and highly divergent *Sus scrofa* populations. *Mol Ecol*. 2014;23(16):4089-4102. doi:10.1111/mec.12807
26. Joshi N, Fass J. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33). Published online 2011.
27. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv*. Published online 2013.
28. Warr A, Affara N, Aken B, et al. An improved pig reference genome sequence to enable pig genetics and genomics research. *Gigascience*. 2020;9(6). doi:10.1093/gigascience/giaa051
29. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. *ArXiv*. Published online 2012.
30. Garrison E, Kronenberg ZN, Dawson ET, Pedersen BS, Prins P. A spectrum of free software tools for processing the VCF variant call format: vcflib, bio-vcf, cyvcf2, hts-nim and slivar. *PLoS Comput Biol*. 2022;18(5):e1009123. doi:10.1371/journal.pcbi.1009123



31. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 2015;4(1):7. doi:10.1186/s13742-015-0047-8
32. Meier J, Ravinet M. Speciation & Population Genomics: a how-to-guide. Physalia-courses. Published December 2021. Accessed August 29, 2022. <https://speciationgenomics.github.io/pca/>
33. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27(15):2156-2158. doi:10.1093/bioinformatics/btr330
34. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. *Gigascience*. 2021;10(2). doi:10.1093/gigascience/giab008
35. Genome: *Sus scrofa* (pig). National Center for Biotechnology Information. Accessed January 14, 2023. <https://www.ncbi.nlm.nih.gov/genome?term=sus%20scrofa%20%5BOrganism%5D&cmd=DetailsSearch>
36. Tack J. *Wild Boar (Sus Scrofa) Populations in Europe: A Scientific Review of Population Trends and Implications for Management.*; 2018.
37. Ellis E. Ecology in an anthropogenic biosphere. *Ecol Monogr*. Published online 2015:287-331.
38. Veličković N, Ferreira E, Djan M, et al. Demographic history, current expansion and future management challenges of wild boar populations in the Balkans and Europe. *Heredity (Edinb)*. 2016;117(5):348-357. doi:10.1038/hdy.2016.53

## 9 SUPPLEMENTARY MATERIAL

The supplementary materials contain the items below. All supplementary materials can be found in:  
NinoMenger\_s1098386\_Bafstu\_Thesis\_Supplements\_V1.0.0.pdf

1. Data overview
2. Results:
  - 2.1. PCAs:
    - 2.1.1. South East Asian Island Suids
    - 2.1.2. Eurasian domesticated
    - 2.1.3. Commercial breeds
  - 2.2. Effective population size analyses:
    - 2.2.1. South East Asian Islands Sus
    - 2.2.2. Asian wild and domesticated *Sus scrofa*
    - 2.2.3. European wild and domesticated *Sus scrofa*
    - 2.2.4. Commercial breeds
  - 2.3. ROH analysis
  - 2.4. Deleterious mutations analysis:
    - 2.4.1. pCADD scores proportional to all mutations
    - 2.4.2. pCADD scores proportional to all categorized mutations

All supplementary materials can be found in:  
NinoMenger\_s1098386\_Bafstu\_Thesis\_SupplementaryMaterials\_V1.0.0.pdf