Norwegian University of Life Sciences

Master's Thesis 2023 60 ECTS
Faculty of Biosciences

PRNP guided selection against CWD in reindeer – Impact on local and total genetic variation

Aurora Hofsvang
Faculty of Chemistry, Biotechnology and Food Science

Acknowledgements

I would like to thank my supervisors Dag Inge Våge, Michael A. Tranulis, Thu-Hien To and Matthew Peter Kent, for their help with this thesis. They have participated in long meetings to discuss results and methods and shared their knowledge and enthusiasm for the project.

I would also like to thank my mom for all the support and encouragement I have received while writing this thesis. And both her and Axel for taking the time to help me proofread.

The author acknowledge the Orion High Performance Computing Center (OHPCC) at the Norwegian University of Life Sciences (NMBU) for providing computational resources that have contributed to the research results reported within this paper. URL (internal): https://orion.nmbu.no

Abstract

The emergence of Chronic Wasting Disease (CWD) in Norway poses a significant threat to the populations of free-ranging and semi-domesticated reindeer. Controlling the spread of the disease is of utmost importance, and studies have indicated that different PRNP genotypes give varying levels of sensitivity to CWD in reindeer. Some genotypes offer more protection against the disease, making it desirable for reindeer herders to breed selectively to increase the ratio of CWD resistant genotypes in the populations. In all artificial selection it is important to avoid reducing the genetic variation in the population. To this end it is important to know to what extent genetic variation in the population is associated with the PRNP-genotypes.

DNA samples from reindeer were sequenced with Illumina sequencing, followed by alignment and variant calling to generate a set of SNP genetic markers for population structure and variation analysis. Analysis of population structure revealed that individuals with identical PRNP genotypes did not exhibit increased relatedness to each other compared to the rest of the individuals. However, analysis of positions surrounding the PRNP gene indicated that the PRNP genotypes influenced the variation found in the surrounding regions, suggesting the presence of linkage disequilibrium (LD) tied to PRNP alleles. Notably, the A allele, which causes sensitivity to CWD in reindeer, showed less signs of LD compared to the other alleles. This implies that the A allele has less association with specific variants than the other genotypes and could indicate a smaller chance of removing alleles from the population when selecting away from the A allele. However, importantly, as the A allele exhibited the most variation in the regions surrounding the PRNP gene, selectively removing A alleles would reduce the genetic variation in this area. As a high portion of the variants in the positions surrounding the PRNP gene is found together with the A allele this could potentially lead to the loss of additional alleles. As LD was investigated by visually inspecting clustering in MDS plots in this study more specific analysis is needed to conclude the impact of selective breeding against CWD on genetic variation in the area surrounding the PRNP gene.

In addition, this pilot study uncovered an important discrepancy. The genotypes identified through PCA amplification and Sanger sequencing differed from those identified through

whole-genome Illumina sequencing for nine of the animals. Further investigation is necessary to determine the causes of these inconsistencies.

Table of Contents

1.Introduction

After the first cases of Chronic Wasting Disease (CWD) was discovered in Norway in 2016, it has been considered a serious threat to both semi-domesticated and wild populations of reindeer. Selective breeding to make the semi-domesticated reindeer more resistant to this disease is discussed as a possible part of the solution. This is the context for this study where the goal is to assess the possible effect of selective breeding based on PRNP alleles on genetic variation.

Wild (free-ranging) and semi-domesticated (herded) reindeer in Norway and the threat from CWD

Norway is one of the countries in Europe with the highest number of semi-domestic reindeer, with a population of around 225.000 animals (Maraud & Roturier, 2021; Veterinærinstituttet). Reindeer herding is practiced over an area of 140.000 square kilometers or 40% of the land area in Norway (Forbes & Kumpula, 2009) In addition, Norway is the only country in West Europe with populations of wild reindeer (*Rangifer tarandus*) and wild tundra reindeer (*Rangifer tarandus tarandus*) (Sylvie L. Benestad et al., 2016). The population of wild reindeer in Norway has been stable at around 25.000 (+/- 3.000) individuals over the last decade (Eldegard K, 2021). The management and conservation efforts of wild reindeer have largely been concentrated on area management and the impact of human disturbance, but the discovery of the first cases of CWD in wild reindeer in 2016 drastically changed this (Mysterud et al., 2020). CWD is considered a serious threat to the Norwegian, and thus the Eurasian, population of wild reindeer. If the disease spread to populations of semi-domesticated reindeer, it may threaten the culture, traditions and income of a large part of the indigenous Sami people (Mysterud & Rolandsen, 2018).

Reindeer (Rangifer tarandus)

Figure 1.1: Foto of reindeer, taken by Per Jordhøy (Fremstad, 2020)

Distribution:
Reindeers inhabit a broad range of territories spanning from 50 to 81 degrees north around the Arctic region. They can be found in various locations across the globe, including the northwestern region of the United States (Alaska), Canada, Greenland, Norway, Finland, Russia, and Mongolia (Gunn, 2016).
In Europe, approximately half of the population of wild reindeer, and nearly the entire population of mountain reindeer R. t. tarandus can be found on the mainland of Norway (Eldegard K, 2021)

As the numbers above show, most of the reindeers in Norway are parts of semi-domesticated herds. For a large part of the Sami indigenous population, reindeer herding is considered an important part of their culture and tradition, and it is also a significant source of income. Troms and Finnmark account for around 75% of the population of semi-domesticated reindeer in Norway (Veterinærinstituttet), but reindeer herding is also present in the area north of Røros, and in the Femundsmarka area south of Røros. There are four non-Saami herding groups (Filefjell tamreinlag, Vågå tamreinlag, Lom tamreinlag, and Fram tamreinlag) that practice reindeer husbandry in central parts of southern Norway (Kaltenborn et al., 2014). Wild reindeer can be found in the Mountain ranges in southern Norway and is a common species here (see Figure 1. map with the various areas of wild reindeer). Some of the wild habitats and the habitats of domestic reindeer are overlapping, for example in the Jotunheimen–Breheimen area (Kaltenborn et al., 2014). Filefjell tamreinlag share borders with the CWD-affected wild reindeer population in Nordfjella.

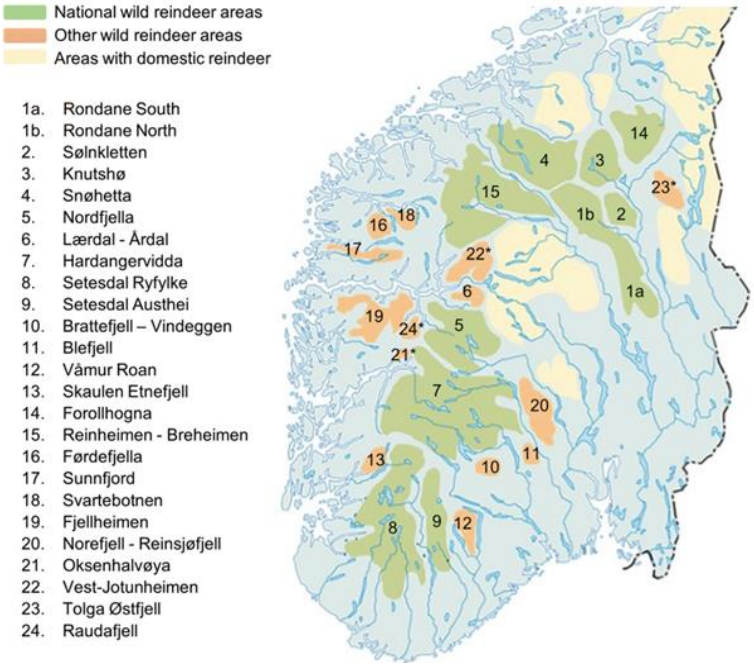


Figure 1.2: Map of the reindeer management areas in Norway. Retrieved, as a modified version of a map from villrein.no, from (Kvie et al., 2019)

Current and historic importance of reindeer

Reindeer have been a key component of Eurasian high-latitude ecosystems for at least two million years. There is evidence that reindeer were a source of food for Neolithic humans as they were extensively hunted (Forbes & Kumpula, 2009) and within Fennoscandia, writings from as early as the ninth century describe people, most likely the indigenous Saami, engaging in regular reindeer husbandry (Forbes & Kumpula, 2009). This illustrates that herding of reindeer has been important for centuries. Today reindeer is used for nutrition, clothing, shelter and plays a role in spiritual beliefs. Commercial products from reindeer is

mostly meat, but skins, handicrafts and even antler velvet are also sold (Forbes & Kumpula, 2009). Many Indigenous people depend on Rangifer species for their livelihoods as the reindeer meat remains an important food source for the herders' communities (Maraud & Roturier, 2021).

The wild reindeer in Norway is of special significance as it is considered the last remnants of wild tundra reindeer (*Rangifer tarandus tarandus*) in Europe. Norway therefore has a responsibility to protect the species (Sylvie L. Benestad et al., 2016).

Ecological importance of reindeer and threats effecting the herds

Both the semi-domesticated reindeer and the wild reindeer populations play an important role in the ecosystems. Reindeer is sometimes referred to as a keystone species or an "ecosystem engineer" and it is widely agreed that reindeer play an important role in the boreal forest and tundra ecosystems (Forbes & Kumpula, 2009). In a recent study in the north of Fennoscandia it was shown that reindeer contribute to maintaining the regional biodiversity, as on relatively rich dolomitic substrates reindeer tends to promote rare and threatened plants (Forbes & Kumpula, 2009).

The availability of winter ranges is widely recognized as one of the primary factors that limit the carrying capacity for reindeer populations (Nellemann et al., 2001). Various factors influence this, including the fragmentation of original habitats and climate change. Habitat fragmentation has long been identified as a major threat to population size, and it has been advised that habitat connectivity should be a focus for conservation efforts (Kvie et al., 2019). When habitats become fragmented, reindeer have less territory available for migration, resulting in more sedentary populations. This, in turn, can lead to overgrazing issues and a decrease in the carrying capacity of the pastures.

Furthermore, climate change, attributed to global warming, is predicted to exacerbate this problem. Fluctuating climate patterns can result in harsher winters, with more ice instead of snow making it more difficult to find food, and this have been shown to increase reindeer mortality rates. Additionally, global warming may cause increased winter precipitation, leading to larger areas of reindeer summer ranges being permanently covered in snow. Consequently, this limits the availability of nutrients for reindeer (Heggberget et al., 2002).

Another concern associated with climate change is the increased likelihood of emerging infectious diseases in wildlife, including reindeer populations (Jones et al., 2008). Chronic Wasting Disease (CWD) has posed a new challenge to the management and protection of reindeer populations in Norway since 2016. This further emphasizes the need for effective management strategies.

Chronic wasting disease (CWD)

CWD is a deadly, transmissible disease which affects the brain and nervous system of reindeer and other members of the *Cervidae* family (Haley & Hoover, 2015). It is a prion disease, i.e., the disease-causing agent consists of aggregates of a misfolded conformer of the prion protein (PrP). The disease was first discovered in Colorado, USA, in 1967, and by 2022, it has been recognized in wild or farmed deer in 30 US states and four Canadian provinces (Richards, 2021). The disease was introduced in South Korea with imported live cervids (Kim et al., 2005; Sohn et al., 2002). In 2016 the first CWD case was discovered in Eurasian reindeer (*Rangifer tarandus*) (S. L. Benestad et al., 2016) and moose (Pirisinu et al., 2018) in Norway. Spread of the disease is considered a threat to both wild populations of cervids and animal husbandry, and measures to fight the disease is high on the nature management agenda in Norway.

The early history of CWD

The first case of CWD was registered in Colorado in captive mule deer (Odocoileus hemionus) at a research facility in 1967, and it was subsequently found in similar research facilities in Wyoming (Leiss, 2017; Otero et al., 2021). Although this was the first recognized case of the disease there have likely been cases in the states before this (Miller & Fischer, 2016). In 1980 the disease was classified as a transmissible spongiform encephalopathy (TSE) (Williams & Young, 1980). Other forms of TSE include scrapie in sheep and goats, bovine spongiform encephalopathy (BSE), transmissible mink encephalopathy (TME), and Creutzfeldt-Jakob disease (CJD) in humans (Otero et al., 2021).

Due to both lack of and imperfect surveillance, the history of how CWD spread through North America is somewhat unclear. In the first years after the discovery in 1967, new cases were mostly found in Colorado and Wyoming. From 1996 to 2016 the disease range expanded rapidly, and by 2016 it was registered in 21 states and two Canadian provinces (Leiss, 2017; Miller & Fischer, 2016).

This history of how CWD has spread is important when we are looking for ways to manage the disease. Much of the early spread of CWD in captive herds can be attributed to the transfer of individuals between facilities as this was a common practice as the knowledge of CWD was limited. The disease has spread to both Canada (1996) and South Korea (2021) through the import of infected captive cervids (Kim et al., 2005; Otero et al., 2021). In wild populations, Cervid migration is an important factor in the geographic spread of CWD (Otero et al., 2021).

CWD stands out from other TSEs in that it is highly contagious, and the infectious particles can survive in the environment for prolonged periods (Sylvie L. Benestad et al., 2016; Otero et al., 2021). It is also the only TSE found to affect both wild and farmed animals, with the first reported case in a wild cervid (an elk) in 1981 in Colorado (Leiss, 2017; Miller & Williams, 2004). As a consequence of the high contagiousness of CWD, the disease is on the rise, unlike other Prion diseases which are under control or declining (Rivera et al., 2019).

The contagiousness of CWD is one reason why the disease is challenging to eradicate when it is present in an area, the stability of infectious particles in the environment is another. Cervids can be infected with CWD via ingestion of PrP-CWD from sources in the environment and the persistence of these particles can make an area unsuitable even though the infected animals have been removed (Kahn et al., 2004). Infected cervids may shed PrP-CWD in excretions such as feces and saliva, or the contamination may come from the decomposition of diseased carcasses (Kahn et al., 2004; Otero et al., 2021).
Dense populations as a consequence of keeping captive herds can contribute to increased likelihood of transmission both from the environment and between individuals (Kahn et al., 2004).

To sum up, eradication of CWD from areas of endemicity is likely to be impossible due to the long-term stability of infectious prions in the environment, the ease of transmission from animal to animal, and the lack of an effective vaccine or treatment (Race et al., 2018).

The discovery of CWD in Norway

When the first case of CWD was discovered in Norway in 2016, this also represented the first CWD case in Europe, and the first case found in reindeer globally. Only in 2018, two years after the first reindeer case in Europe, was the first case of CWD in reindeer

discovered in North America (*Statusrapport CWD for 2018*, 2019).

Because of the difficulty of stopping the spread of CWD in North America the Norwegian government introduced drastic measures when a case was discovered in Norway. It was decided to eradicate the entire sub-population of reindeer in Nordfjella, and this entire sub-population where the positive case was found was culled between 2016-2018 (Güere et al., 2020). Nordfjella sone 1 is adjacent to other zones of wild and herded reindeer and the find raised serious concern that the disease would spread to nearby populations (S. L. Benestad et al., 2016).

A total of 19 animals in the culled Nordfjella population tested positive for CWD. In 2020 and 2022 two further cases were diagnosed in the Hardangervidda area, which holds the largest population of wild reindeer in Western Europe. In addition to the 21 cases of CWD in reindeer, distinct forms of CWD with sporadic occurrence have been recognized in moose in Norway (11 cases), Sweden (4 cases) and Finland (3 cases), and tree red deer in Norway (Tranulis et al., 2021). From 2016 to 2023 35 cases of CWD have been reported in Norway, 21 of them in reindeer (Veterinærinstituttet, 2023).

CWD and other prion diseases

CWD is a prion disease. While the term prion was first used in 1982, the first description of what we now know as prion diseases date back to 1732 (Liberski, 2012). This was a disease in sheep that caused altered behavior including excessive licking, scratching and altered gait, this disease was named Scrapie. More recently, similar diseases, which like Scrapie have neurological characteristics, have been found in humans. Creutzfeldt-Jakob disease was described in 1920, and Kuru was described in 1957 (Rivera et al., 2019). By 1959 the three diseases were linked and in 1967 a theory that the diseases were caused by a proteinase agent emerged (Rivera et al., 2019).

Prusiner (Prusiner, 1998) and collaborators, in 1982, were the first to prove that the causative agent for scrapie is a protein, for which they won the Nobel Prize for in 1997. Also in 1982, Prusiner coined the term “prion” to describe the transmissible proteinaceous agent that was the cause of TSEs (Rivera et al., 2019).

Prions proteins are present in almost, if not all, mammalian species, making it highly evolutionary conserved. Prion proteins (PrPC) are cell-surface glycoproteins consisting

mostly of alfa-helical conformations. Prions have around 42% alpha-helix content and are essentially devoid of B-sheets which is only present in 3% (Pan et al., 1993). The protein is encoded by the single-copy PRNP gene with the entire open reading frame contained in one exon (Huang et al., 1994)

Prions can be found in several different tissues and cell types, like epithelial, endothelial and immune cells, but has particularly high expression levels in neurons and neurological cells of the central and peripheral nervous system (Rivera et al., 2019).

Prion diseases is a collective term for diseases where the transmissible agent that causes the disease is a misfolded prion protein (PrPSc). The diseases manifest as sporadic, inherited, and infectious disorders, and include, among others, scrapie, chronic wasting disease, and bovine spongiform encephalopathy in animals (Huang et al., 1994)

The disease causing mechanism is that when PrPSc comes into contact with the cellular prion protein (PrPC) it acts as a catalyst and causes a configurational change of the cellular protein, where a portion of its α-helical and coil structure is refolded into β-sheet (Pan et al., 1993). This change in configuration triggers aggregation of the prions as this is energetically favorable in the new configuration. This then leads to accumulation of abnormal prions which is harmful especially in the central and peripheral nervous system.

CWD susceptibility based on PRNP genotype

Variation in disease susceptibility has been linked to PRNP variation in elk (O'Rourke et al., 2007; Perucchini et al., 2008), mule deer (Fox et al., 2006) and white-tailed deer (Johnson et al., 2006; Johnson et al., 2011). It was theorized that this would be the same for reindeer.

As most of the CWD outbreaks have happened in North America most of the research into CWD is also from here. The first outbreak in Europe made it evident that more European research was needed, and especially because the case found in Norway seemed to differ from the disease in North America (Maraud & Roturier, 2021). As the first case of CWD in Norway also represented the first case found in reindeer globally, it was necessary to map the occurrence of different PRNP-genotypes in reindeer (Güere et al., 2020; *Statusrapport CWD for 2018*, 2019).

The culling of the reindeer sub-population in Nordfjella made valuable genetic material available for research. The genetic material from the eradication was used to research the variation in the prion protein gene (PRNP) in Norwegian wild reindeer, and if prion genotypes showed signs of impacting the susceptibility of the reindeer to CWD. For this purpose, 19 reindeer affected with CWD and 101 healthy animals, all from the Norefjella population, were genotyped. In the study, five different prion alleles were found from the genetic material collected from the culled population (Güere et al., 2020). A study in 2022 which included individuals from several populations found one additional allele (Güere et al., 2022). These alleles were named A-F. Allele A was the allele matching the reference sequence used. An overview of the sequence differences between the different alleles found in 2020 can be seen in Table 1.1 (Güere et al., 2020). Allele F found in the 2022 study was a substitution of Lysine (207) to Methionine (Güere et al., 2022)

These six alleles were found to combine into 15 different genotypes. The most abundant genotypes were A/A, A/B and B/B while other genotypes were detected at proportions <.10. No animals homozygous for the alleles C (deletion) or F (207M) were detected (Güere et al., 2022).

Table 1.1: The different alleles found by Güere in 2020 (Güere et al., 2020).

	PRNP open reading frame variant positions						
	4G>A Val2Met	249_272del Trp84_Gly91del	385G>A Gly129Ser	505G>A Val169Met	526A>G Asn176Asp	674C>A Ser2a25Tyr	Study population =240
Allele							
A	Val	Trp84_Gly91	Gly	Val	Asn	Ser	46.3%
B	-	-	-	-	-	Tyr	30,4 %
C	-	Trp84_Gly91del	-	-	-	-	9,6 %
D	-	-	-	-	Asp	-	7,9 %
E	Met	-	Ser	Met	-	-	5,8 %

The paper published by Güere (Güere et al., 2020) showed that there were signifcant differences between the allele frequencies of healthy reindeers and the ones affected with CWD and subsequently the genotype distribution was also different. Allele A and allele C was significantly overrepresented in the cases compared to the controls. All the CWD affected animals had a PRNP-genotype with at least one A or C allele. On the basis of this, it

was concluded that some genotypes found among Norwegian reindeer provide more resistance to CWD.

This indicates that populations with higher frequencies of certain genotypes would be beneficial to reduce the risk of CWD. A study on farmed whitetail deer describes how a breeding program for increasing the frequencies of disease-resistant PRNP genotypes is already implemented in a farm in an area of Nort America where CWD is endemic. This study reports success in altering the genotype frequencies to higher occurrence of less susceptible genotypes, but it does not mention if there has been any considerations of potential loss in genetic variation when implementing this breeding program (Haley et al., 2021).

The importance of genetic variation

Genetic variation is vital for the survival and adaptability of populations, and preserving genetic diversity is crucial for population viability and long-term survival (Lacy, 1997). Genetic variation is often defined as "the difference in DNA sequences between individuals within a population"[1]

Genetic variation is essential for a population's ability to adapt and gives resilience to environmental changes (Lacy, 1997). In experimental populations, it has been demonstrated that reduced levels of genetic variation can limit the adaptive potential of a population (Wright, 1968).

Low genetic variation can be linked with inbreeding. Inbreeding can result in, or amplify loss of genetic variation. In one study of a population of lions that suffered a bottleneck where the population was reduced to 10 animals, the resulting loss of genetic variation seen in the population consisting of the descendants led to reduced fitness in the population. A higher rate of sperm abnormalities, and lower sperm motility than the nearby population in the Serengeti is observed (Packer et al., 1991). In a study on desert topminnows it was demonstrated that genetic depletion could cause slower growth and increased vulnerability to stress and parasites(Lacy, 1997).

[1] https://www.ebi.ac.uk/training/online/courses/human-genetic-variation-introduction/what-is-genetic-variation/

Livestock populations are often small, isolated and subjected to strong artificial selection to maximize production of desired traits. These factors affect genetic variation and play a major role in the fact that there is often less genetic variation in livestock populations than in wild populations (Notter, 1999).

In a study where the genetic variability in semi-domesticated reindeer in Norway was examined it was found that the populations included in the study contain a moderate amount of genetic variation (RØED, 1985). The genetic variation found was mostly contributed by variation within the populations and the different populations were found to differ only very slightly. Compared to the general genetic variability in mammals inferred by Nevo (Nevo, 1978) the genetic variation found amongst semi-domesticated reindeer was lower, but not remarkably so. No indication of inbreeding was found in the populations (RØED, 1985).

Genetic differences between wild and semi-domesticated reindeer have been reported in several studies (Kharzinova et al., 2018; RØED, 1986). Røed (RØED, 1986) investigated the genetic variation in wild reindeer with electrophoresis at 9 loci. By comparing the results to that found for semi-domesticated reindeer in a similar study (RØED, 1985), he found that there is a slightly higher amount of genetic variation in the wild populations. In the study he underscores that both wild and semi-domesticated reindeer have a considerable amount of genetic variation.

Research of the genetic composition of semi domesticated reindeer show that there have been a bottleneck sometime in the past (Røed et al., 2021).

As adequate genetic variation is an important part of maintaining population fitness it is generally part of the considerations in breeding programs (RØED, 1985).

Measuring genetic variation

Before molecular methods became the standard practice for investigating genetic variation it was usual to investigate genetic variation in a few loci by looking at changes in protein structures (Lewontin & Hubby, 1966; RØED, 1986). A common way to do this was to use electrophoresis to separate the proteins based on different characteristics like net charge. The problem with this method is that not all genetic differences give measurable changes in proteins and this causes an underestimation of actual variation. Another difficulty is that it

that with this method one chose some enzymes to represent the variation found in the whole genome, but it is not always certain that these are representative of the genetic variation found in all the loci in the population (Lewontin & Hubby, 1966).

In an article published in 1993, it was stated that “microsatellites may become the makers of choice for molecular population genetics”(Bruford & Wayne, 1993). Microsatellites are highly polymorphic short tandem repeats of sequence units. The polymorphism is attributed to variation in the number of repeated units. The discovery of microsatellites represent a leap in investigation of population structure as they can be used to investigate genetic variation from small amounts of material and they can be highly precise (Bruford & Wayne, 1993).

Single nucleotide polymorphisms (SNPs) emerged as a new genetic marker after the use of microsatellites became common (Brumfield et al., 2003; Liu et al., 2005). In studies comparing the two molecular markers, it is shown that microsatellites are more informative than SNPs and fewer microsatellites than SNPs are needed to infer population structure and determine whether sub-populations are present (Haasl & Payseur, 2011; Liu et al., 2005). Since SNPs can only vary to four different characters, they may therefore be less suited for parentage analysis (Brumfield et al., 2003). Still, SNPs have become the marker of choice for many researchers. The lower and uniform mutation rate makes it easier to infer historic population demographics, and makes it less likely that information is lost to mutation in the same positions (Brumfield et al., 2003).
The fact that SNPs are less informative than microsatellites can be offset by including a larger set of markers in the data. Brumfield (Brumfield et al., 2003) highlights the fact that the use of SNPs as markers make more tests of deviations from neutrality, population size and recombination available. The use of SNPs also improves the ease of modeling, they are simple to genotype and as many are available it can make up for the lack of statistical power compared to microsatellites (Brumfield et al., 2003).

There are different ways to investigate genetic variation once a set of genetic markers has been compiled. Whether heterozygosity differs from that expected under Hardy Weinberg equilibrium is one common test (Lacy, 1997). Studies investigating genetic variations also investigate population structure (Muli et al., 2022).

That statistical power increases with number of individuals sampled is a widely accepted premise. Low sample sizes are in some cases taken as indicating low power (Ryman et al., 2006). It is difficult to determine the required sample size for an accurate measure of genetic variation, but as higher sample numbers result in higher statistical confidence of the analysis, more samples are usually preferred. The limiting considerations are often computational resources, expenses and samples available.

Thesis aim

It is of great importance to find solutions to the CWD problem, but without creating other problems or weaknesses in the reindeer population. Because susceptibility to CWD vary with PRNP genotype, breeders are interested in selecting reindeer with genotypes making them less susceptible to CWD. Changes in the genetic makeup of a population can have unintended consequences, so before a breeding program with selection of (PRNP genotype) can be implemented, potential consequences needs to be investigated. An important question is if removing the alleles making reindeers more susceptible to CWD also will reduce the genome wide genetic variation in the population.

In this thesis I will process sequence data from reindeer to produce SNP data of all the variable positions in the genome, and use this to study the population structure of a herd of semi-domesticated reindeer with the goal of detecting if additional genetic variation will be lost if susceptible genotypes are lost.

If it is possible to remove alleles causing susceptibility to CWD from semi-domesticated reindeer populations this can help prevent the spread of CWD. This is important for reindeer husbandry but also for conservation of wild reindeer in Norway.

2. Method

Genotyping and Sequencing

The DNA material used in this analysis is genomic DNA samples obtained from ears of reindeer. The samples were collected from the four non-Sami reindeer herding groups, Filefjell Reinlag, Fram Reinlag, Vågå tamrein and Lom Tamrein. The DNA from these samples was extracted using QIAGEN DNeasy Blood & Tissue kit (Qiagen, German) and the DNA concentration of the samples was assessed with Nanodrop (Themo Fisher Scientific, USA). The samples were genotyped with SANGER sequencing after amplification of the PRNP reading frame with PCR. The genotyping was performed before the master project started.

Out of the genotyped individuals 27 reindeer from the Filefjell population were chosen for sequencing. Only samples from one herd were chosen as the number of samples that could be sequenced was limited (due to budget) and it would be more informative for the goal of the study to investigate the genetic variation within one population as opposed to between different populations. Seven of the individuals had the genotype A/A which makes them susceptible to CWD, 10 were classified as intermediate susceptible because they had the genotypes A/B, A/D or A/E. Lastly 10 individuals with the genotype E/E, B/D, B/B, B/D, B/E or D/D were chosen to represent the less sensitive group. The samples were sent to Novogene UK, a commercial provider, for sequencing. This company prepared the DNA library with the NEB Next Ultra II DNA Library prep kit (NEB, USA) and the DNA was sequenced by a Novoseq6000 machine using S4 Flow-cells (Illumina USA). The requested amount of sequence was 60Gb of each sample. As the reindeer genome is 2,92Gb (Weldenegodguad et al., 2020) this amounts to a coverage of around 20X.

Pre-processing

There are several steps in the process of performing a variant calling from raw-reads sequenced with Illumina (Illumina USA). These steps were performed in a linux environment at the high performance computing (HPC) server Orion. The raw sequence data files were unzipped and then combined into two files for each of the samples, one for the forward reads and one for the reverse reads produced by Illumina “paired end sequencing”. The quality of the reads was assessed with FastQC version 0.11.9-Java-11(Babraham

Bioinformatics). FastQC measures different parameters about the reads for each sample, among other things the base-quality and the GC content. The FastQC results for all the samples were summarized with MultiQC version 1.9-foss-2019b-Python-3.7.4 (Ewels et al., 2016) to make them easier to asses.

Samples that were flagged as failed in the “per tile quality” check by FastQC underwent quality trimming. If a sample failed the “per tile quality” check it means that specific positions on the flow cell produced reads with lower quality scores than the average read quality score (Alnasir & Shanahan, 2020). To remove some of the reads with low quality the failed sequences were filtered with FilterByTile from BBtools (Bushnell, 2020). FilterByTile is a tool that filters reads based on position specific quality over the flowcell. The default values for the program were used.

Alignment

To align the reads BWA version 0.7.17-GCC-10.2.0 (Li & Durbin, 2009) was used. The reads were aligned to a partially assembled genome of a reindeer form Hardangervidda in Norway (Accession:PRJEB35834) (Kiel, 2021). The reference genome was indexed with BWA index and the alignment was performed using BWA mem (Li, 2013). The Readgroup information from the Illumina sequence files were included as a parameter in the alignment so that the information still was accessible in the resulting SAM files, as later steps in the pipeline require this information (GATK). To reduce file size SAMtools (Danecek et al., 2021) was used to convert the SAM files to BAM files in the same operation as the alignment. The completeness of the alignments was assessed with the “flagstat” command from SamTools.

To flag reads that were likely artifacts from the PCR amplification MarkDuplicates from Picard tools version 2.9.2 was used (*Picard tools - by Broad Institute*). This step is important to ensure that PCR and optical/sequencing errors do not get included in downstream analyses.

The reference genome was originally soft masked, before the variant calling it was changed to hard masked with text editing in Rstudio version 4.1.0 (RStudio Team, 2020). This was done to decrease run time and memory use. As the goal in this analysis is to discover the distribution of genetic variation within the population, specificity and speed is favored over

sensitivity. Aligning short reads to repeated sequence is not specific, as they can align at several positions in the genome. This is the case when the repeated region is longer than the individual read.

Variant calling

For the variant calling HaplotypeCaller, a tool form GATK (McKenna et al., 2010; O'Connor, 2020), was used with the GVCF mode. HaplotypeCaller call SNPs and indels via local re-assembly of haplotypes. The GVCF mode is used when the input data consists of sequences from multiple individuals, when it is used the output file format is GVCF. The samples were run through HaplotypeCaller individually and one GVCF file was produced for each sample (Caetano-Anolles, 2023a).

HapplotypeCaller works by using a local de-novo assembly of haplotypes to call SNPs and indels simultaneously. When it discovers variation, it disregards current mapping information for that region and does a local re-alignment. This is what is making the tool able to detect different types of variants even if they are close together (Van der Auwera et al., 2013).

To combine the 27 GVCF files to one file that can be used as input for GenotypeGVCFs the program CombineGVCFs was used (McKenna et al., 2010; O'Connor, 2020). To limit the amount of memory and time consumed by the program the GVCF files were merged 5 at a time in several runs until they all were merged into one file. This file was used as input for the genotyping which produced a VCF file for all the samples.

Filtering

To make sure the variants produced in the final VCF file are as correct as possible the VCF file was filtered. For this the VCF file was sorted into two new files, one for SNPs and one for indels. This was done with SelectVaiant by GATK (McKenna et al., 2010; O'Connor, 2020). The filtering was done with GATK VariantFiltration (McKenna et al., 2010; O'Connor, 2020), a tool that performs hard filtering, which means that it filters out variants based on a threshold limit for specific parameters. The filtering focused on removing variants with low quality scores and strand bias. An overview of the parameters used can be seen in Table 2.1. This table also provide a description of the parameters, the recommended threshold values from GATK (Caetano-Anolles, 2023b) and the values used for the filtering.

The threshold value for each parameter was determined by looking at density plots that displayed the distribution of variants for the different quality parameters. The plots were produced from the quality data in the VCF files which were converted into table format by using VariantsToTable from GATK (McKenna et al., 2010; O'Connor, 2020) to make it possible to plot in Rstudio (Team, 2020). A density plot for each parameter was created in Rstudio with ggplot2 (Wickham, 2016). Based on the distribution of values and the recommendations from the company that provides the software (GATK) the threshold values were chosen.

Table 2.1: The filtering parameters used for filtering of the variants called with GenotypeGVCFs. The table includes a description of the filtering parameters, the threshold value recommended by GATK (Caetano-Anolles, 2023b) for which variants to filter out and the threshold values used in this study.

Filter	Description	Recommended threshold for variants to filter out	Used threshold for variants to filter out
Quality By Depth	The variant confidence normalized to account for the higher quality caused by increased depth when sequencing	<2	<2
FisherStrand	The probability that one strand of the DNA is preferred over the other measured with Fishers exact test. The output is given as a Phred-scale p-value. High values indicate strand bias.	>60	>60
RMSMapping-Quality	The root mean square mapping quality over all the reads at the site. This is a parameter used to include the standard deviation of the mapping quality. A good MQ value is around 60	>40	>55
StrandOddsratio	A measure of strand bias that accounts for the fact that the reads at the ends of exons tend to only be covered by reads in one direction which gives them an inaccurate bad score with the FS test.	<3	<3
MappingQuality-RankSumTest	Compares the mapping quality of the reads supporting the reference allele and the mapping quality of the reads supporting the alternate allele. A score around 0 indicates little difference between the strands.	<-12,5	<-5 and >5
ReadPos-RankSumTest	Compares the reference and the alternate strand. It looks at positional differences in the two strands by comparing if one of the alleles is more commonly found in the ends of reads. A score close to zero indicates little positional differences.	<-8	<-5 and >5

VariantFiltration flags variants that fail one or more of the quality checks. These variants

were removed by using SelectVariants from GATK which gather the variants that passed the filtration into one file. After this new density plots were produced to look at the outcome of the filtering.

Base quality score recalibration (BQSR)

To further improve the result of the variant calling Base Quality Score Recalibration was performed. This step recalibrates the quality scores from the sequencing by using a set of known variants. As there was no available set of known variants for reindeer the BQSR was performed after the variant calling. The set of variants produced in the earlier variant calling was used as the required input for the BQSR. Base quality score recalibration is a common step in many next-generation sequencing workflows since inaccurate quality score values affect all subsequent analyses. The step has shown to lead to better variant calls with less false positive variants called (Jade & Swaine, 2017).

The first step in BQSR was to use BaseRecalibration by GATK (McKenna et al., 2010; O'Connor, 2020) to produce a recalibration table. This step required the set of known variants and the bam files that contain the alignment with flagged duplicates, the output form MarkDuplicates, as input. The recalibration table was then used to apply the BQSR to the bam files with ApplyBQSR. Lastly BaseRecalibration is run again with the corrected bam files as input. This produces a table that can be used to evaluate the BQSR.

Re-running the variant calling

The bam files produced by ApplyBQSR (McKenna et al., 2010)were used in the second round of variant calling which performed in the same way as previously described for the first variant calling. The individual sample files were run through HaplotypeCeller, combined with CombineGVCFs and joint genotyped with GenotypeGVCFs which produced the VCF file used for the analysis.

Analysis

Quality control

PLINK 1.9 (Purcell; Purcell et al., 2007) and ggplot2(Wickham, 2016) in Rstudio 4.1.0 (Rstudio Team2020) was used for analysis of the finished VCF file. The file was converted into bed file format and underwent a standard quality control. As the file contains a lot of scaffolds the "--allow-extra-chr" flag was used in all the PLINK runs. Variants were filtered out based on

different parameters that was set; --maf 0.01, --mind 0.1, --geno 0.1 and --hwe 1e-5. The maf option filters out variants (genetic markers) that have a minor allele frequency of less than 0.01. This removes rare variants that may have unreliable genotype calls. "–mind 0.1" filters out individuals based on their genotype missingness. The threshold is sett to 0.1 which means that individuals with more than 10% missing genotype data will be excluded from the analysis. With "—geno 0.1" variants with low call rates are filtered out. Variants that have a call rate of less than 10% is filtered out. The last option "—hwe 1e-5" tests if the SNPs are in Hardy-Weinberg equilibrium (HWE) and filters out variants that deviate significantly from HWE. The threshold is set to 1e-5 which means that variants where the distribution of SNPs found is less than 0.001% likely to observe under HWE is filtered out (Rentería et al., 2013, pp. 193–213).

Variations in the PRNP gene

To look at segregation close to the location of the PRNP gene, its position in the genome was located. This was done by aligning the sequence of the PRNP gene (accession: DQ154293) (Happ, 2005) to the reference genome with Bowtie2 (Langmead & Salzberg, 2012). SAMtools (Danecek et al., 2021) was then used to locate where in the genome the sequence had aligned.

To investigate if any new variation could be found within the PRNP reading frame the VCF file produced with GenotypeGVCFs was phased and the variants found within the reading frame were extracted. The phasing was performed with Beagle (Browning et al., 2021) and to reduce the amount of memory required by the process, the position of the PRNP gene was provided with the flag "chrom" specifying that the program should only phased variants found in this area. The command "query" by BCFtools (Danecek et al., 2021) was used to extract the variants from the selected positions to text format. The BCFtools query command requires the VCF file to be BGZF compressed and indexed, this was done with the command bgzip and index by BCFtools. The Resulting text file contained information about the variation found in the PRNP gene and which combination of variants each allele of each sample had. As the variants found in this table had not undergone quality control with PLINK the variants found in the PRNP reading frame was extracted from the quality filtered bed file to compare if the same variants were found. This was done with "--recode" in PLINK.
The mutations effect on protein primary structure was investigated by using the program

EMBROSS Transeq (Rice et al., 2000) to translate the nucleotide sequence of the reference allele and the alternate sequence to amino acid sequences. By comparing the alleles found in this study to the alleles found by Güere (Güere et al., 2020) the alleles were named.

The genotypes found for the samples after Illumina sequencing was compared with the genotypes found with Sanger sequencing and the legitimacy of the allele calls was investigated by looking at the reads aligned to the reference genome for the positions that showed variation within the PRNP gene with Integrative genomics viewer (IGV) (Robinson et al., 2011).

Determining the population structure and investigating linkage disequilibrium (LD)

To investigate the population structure a multidimensional scaling (MDS) plot was made with Plink (Purcell; Purcell et al., 2007) and ggplot2(Wickham, 2016). Multidimensional scaling is used to reduce the dimensionality of the variation found between data points in a data set to be able to visualize the relationships between the data points. The MDS plot displays the positions of data points in two or three dimensions, with the distance between points indicating their relative similarity or dissimilarity (Dzemyda & Sabaliauskas, 2022; Kruskal, 1964).

PLINK 1.9 (Purcell; Purcell et al., 2007) was used to make a matrix of genetic distances used for the MDS plot. The genetic distances between individuals were calculated with identity by state (IBS) and Identity by Decent (IBD) as the "genome" option was provided (Purcell et al., 2007).

The options "—cluster" and "--mds-plot" were used to create a multidimensional scaling report from the distance matrix calculated with IBS, which can be used to plot the MDS results. The "mds-plot" option was set to 2 and a two-dimensional plot was made. The plot was created in Rstudio (Rstudio Team2020) with ggplot2 (Wickham, 2016). The samples were colored based on if their genotype makes them less sensitive, sensitive or very sensitive to CWD.

To investigate if there was LD tied to the different PRNP-alleles, variants form different stretches around the PRNP gene was extracted and used as input for a MDS analysis. Four MDS plots were made. One containing variants found between position 3.427.296-3.430.067 (3.000bp up and downstream from the PRNP gene), one containing variants

found between position 3.420.296-3.437.067 (8.000bp up and downstream for the PRNP gene) and the two last containing respectively variants from position 3.408.296 to 3.449.066 (20.000bp up and downstream from the PRNP gene) and variants found from position 3.328.296 to 3.529.067 (100.000bp up and downstream from the PRNP gene). All the indicated positions were found on scaffold JAHWTM010000007.1. The four files created were used in MDS analysis with PLINK 1.9 (Purcell; Purcell et al., 2007) and four MDS plots were made using ggplot2 (Wickham, 2016) the same way as previously described. An additional MDS plot was made from the variants found 20.000bp up and downstream from the PRNP gene, where the individuals were colored based on the genotype found with Illumina sequencing.

Reducing the number of SNPs

To investigate how many SNPs that are needed to retain the information of the population structure the ped file was pruned based on Linkage Disequilibrium (LD). Files containing 50.000, 500.000, 3.000.000 and 5.000.000 SNPs were created and used to create four different MDS plots. The option "--indep-pairwise" was used to remove SNPs in high LD. To be able to perform this step the option "--set-all-var-ids @:#[b37]$r,$a" in Plink2 (Chang et al., 2015; Shaun Purcell) was used to make all the variant IDs unique. "indep-pairwise" requires three values as input; the window size, how many bases to shift the window at the end of each step and the R^2 value. The Window size was set to 50 and it moved 5 steps further at the end of every step. The R^2 value is the multiple correlation coefficient between a SNP and all other SNPs in the window based on allele counts, and is the threshold that determines which variants pass (Purcell et al., 2007). For the file that contained 3.000.000 SNPs it was set to 0.85, for the file containing 1.000.000 SNPs it was set to 0.4 and for the files containing 500.000 and 5.000 SNPs it was set to 0.2. "indep-pairwise" works by determining the r^2 value for any given pairs of SNPs within the window and then moving the window the set amount of variant counts. A r^2 value of 0 means no correlation is observed (Harvard.edu). The output from "indep-pairwise" is two files on for the variants that passed and one for the variants that failed. The file containing the variants that passed was used with the option "extract" to exclude the variants in high LD from the quality controlled ped file. The option "--thin-count" was then used to retain a specific number of SNPs in the file.

The SNPs are removed based on position, which means that the flag tries to keep the distance between the remaining SNPs as evenly as possible.

3.Results

Quality

FastQC and MultiQC were used to assess the quality of the raw reads produced by Ilumina sequencing and selected results can be seen in figure A1 and A2 in the appendix. In figure 3.1 the mean quality of the bases in each position is summarized for all the reads per sample.

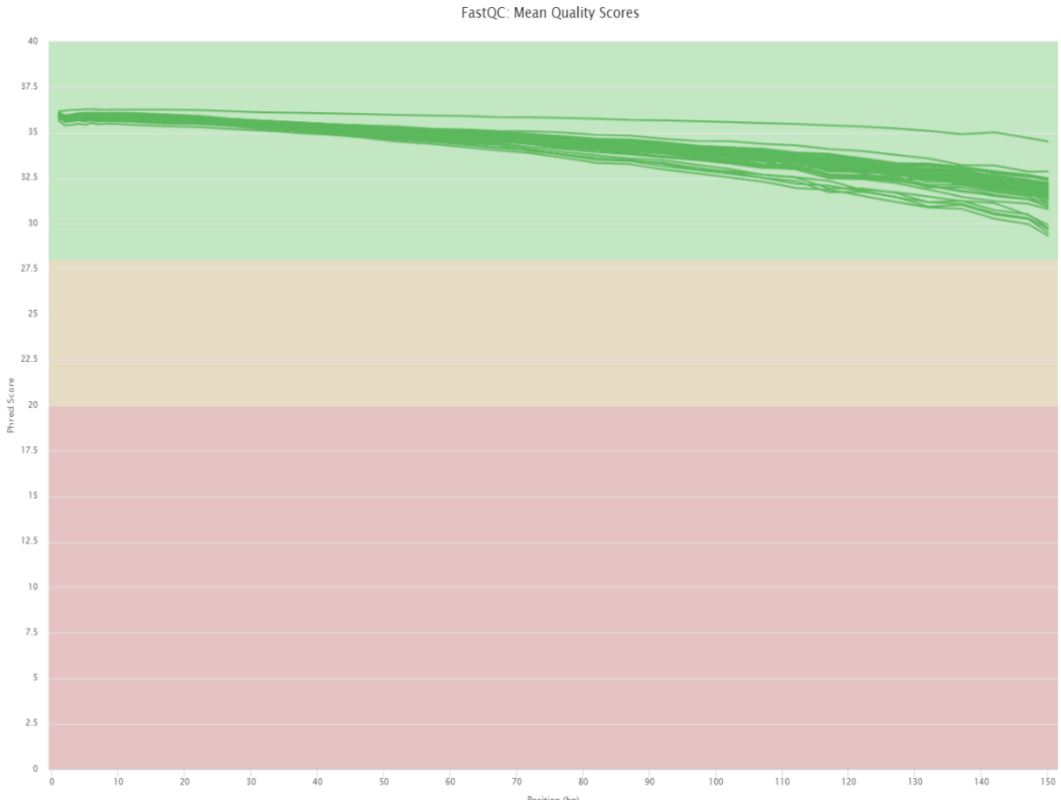


Figure 3.1: The mean quality scores for all the samples for each position. Each green line represents either the forward(R1) or reverse(R2) reads for a sample. The quality score from a position within each read is averaged for all the sample`s R1 and R2 reads separately and is plotted with the average score on the x-axis and the position in the read on the y-axis. The green field indicates an interval of quality scores that FastQC defines as acceptable quality scores, the yellow field indicates an interval of quality scores which FastQC issues a warning for, and the red field indicates an interval of quality scores for which FastQC considers problematic.

Figure 3.1 shows that all the samples are well within the green section with the lowest scores being over 27,5 which is the threshold for indicating that the mean quality passed fastQC filters. The average quality scores drop from over 35 in the first positions to between 28-35 towards the end. This shows that the quality scores drop at the end of reads. The lines

representing average quality for each sample`s either forward or reverse reads are close together indicating a similar distribution of quality score within all files.

Another quality measure taken was the level of duplication controlled with FastQC. The results can be seen in figure 3.2 The average duplication was 19,86% with no sample containing more than 24% duplication. FastQC reports mark reads that are identical as duplicates. Only the first 75bp in the reads are scanned to reduce running time (Babraham Bioinformatics).

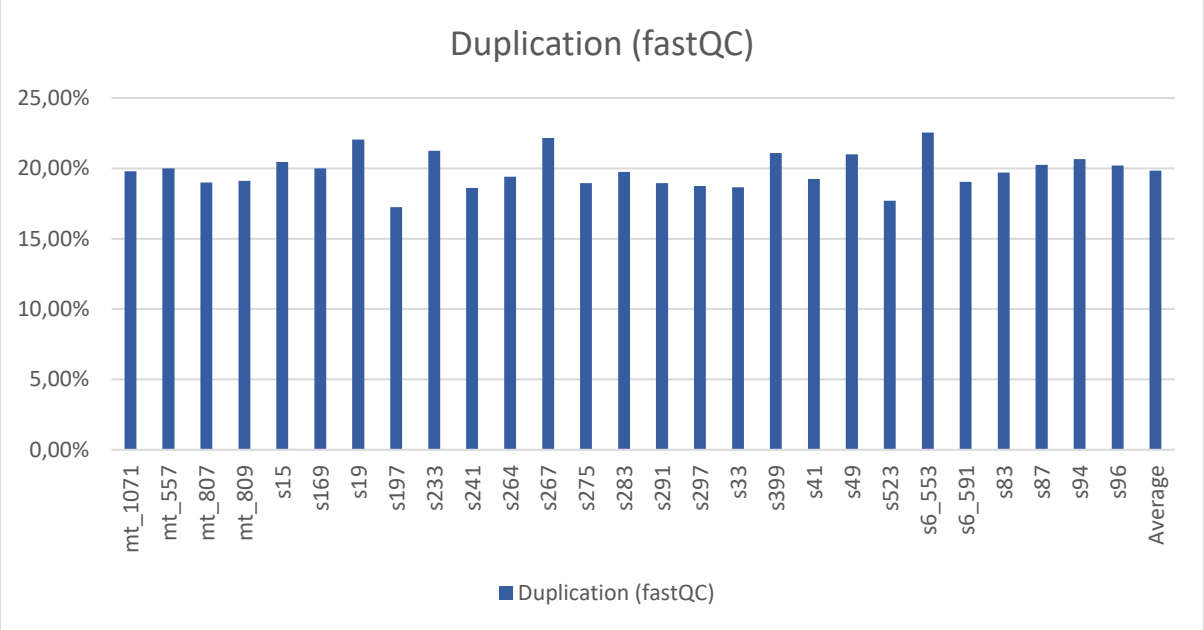


Figure 3.2 Duplication levels amongst the reads for each sample measured with by FastQC. The figure shows the duplication percentage for each sample calculated by averaging the duplication of the forward and reverse read for that sample. The last post shows the average duplication for all the samples.

Quality measures were taken from the alignments to investigate how well the alignment performed. The results can be seen in figure 3.3.

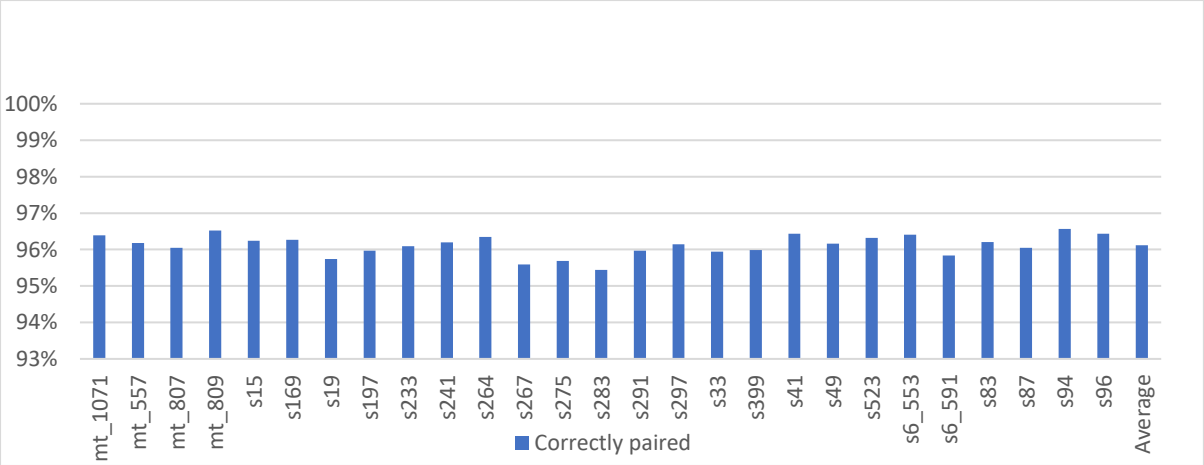


Figure 3.3 The percentage of correctly paired reads for each sample. The calculated average for all the samples is shown in the last post.

The average correctly paired read percentage was 96.12% and none of the samples had a value less than 95%. This means that over 95% of the reads for each sample were mapped to the reference genome with its pair in the correct orientation and expected position.

Base score recalibration was performed to adjust the quality scores to make the variant calling as accurate as possible. Density plots for the distribution of the quality scores and the quality by depth are shown in respectively figure 3.4A and 3.4B.

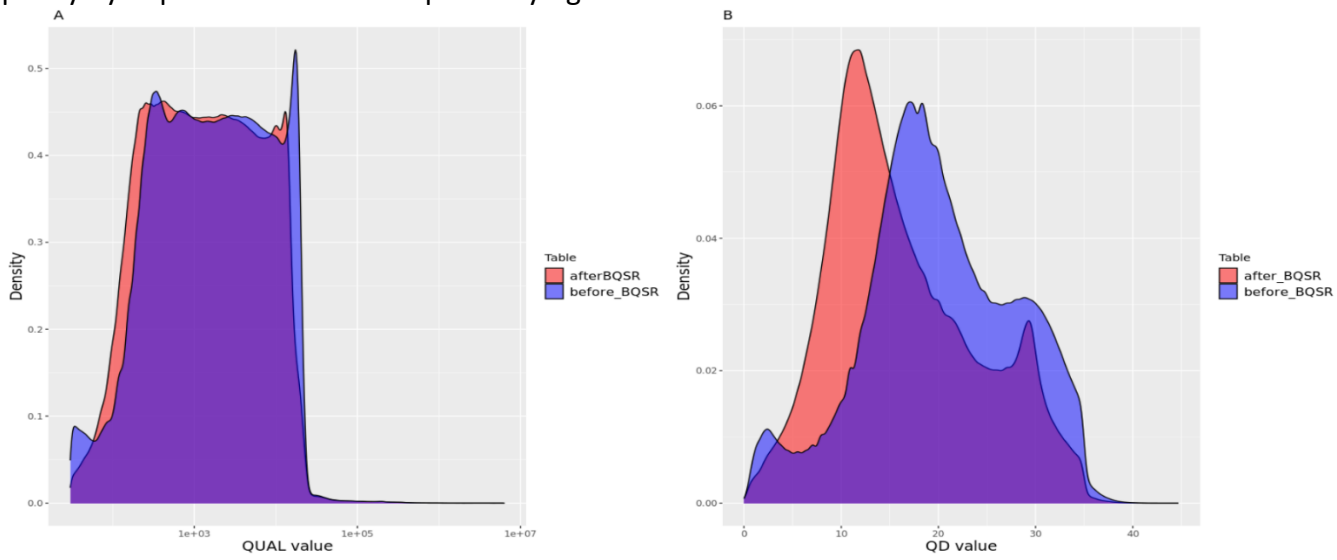


Figure 3.4: Comparison of quality scores(A) and quality by depth scores (B) before and after base quality score recalibration. The figures show the distribution of their quality parameter before the BQSR in blue and the distribution after BQSR in red.

Density plots show the distribution of quality scores within the data set. After the recalibration the quality scores are lower than before. This in turn affects the quality by depth score and the values here are also lower after recalibration than before.

The finished VCF file was filtered in PLINK, an overview of the number of variants filtered out can be seen in table 3.1. Variants were filtered out if they failed one or more filtering parameters.

	Before quality control	Removed due to missing genotype data (geno)	Failed Hardy-Weinberg exact test	Removed due to minor allele threshold(s)	Remaining variants after quality control
Variants	12 049 675	1 484 019	108 459	464 393	9 992 804

Table 3.1: An overview of the number of variants removed in the quality filtration performed with PLINK. The table displays how many variants each of the quality filters removed.

Out of the over 12 million SNPs present in the pre quality controlled VCF file over 2 million were filtered out and just under 10 million variants remained. Over half of the variants that were removed in the quality filtering failed the missing genotype test. This filters out variants with a call rate lower than 0.9. Samples with a higher missingness rate than 0.1 across all positions were also filtered out, but none of the samples failed this test.

PRNP variation

The PRNP variation was extracted from the VCF file and formatted into table 3.2 which shows all the alleles found in the data and how their primary protein structure varies. Of the 8 different alleles found the A allele, which matches the reference genome in all the positions showing variation, were the most abundant, followed by the B allele. The G, H and I alleles only appear once in the data set. The table includes all variants which caused a change in the amino acid sequence and the synonymous mutation at position six in the PRNP reading frame where a guanine (G) was changed to adenine (A) was not included.

Table 3.2: The different alleles of the PRNP gene found after variant calling. The first row of the table contains information about the location within the PRNP gene and effect of each of the seven positions where non-synonymous variation was discovered in the PRNP reading frame. Both the DNA position and change and the amino acid position and change are included. The combination of non-synonymous mutations that make up each allele are noted. The frequency denotes how many of the total 54 alleles that were found to have that specific variation.

Allele	4G>A Val2Met	237_260del Pro79_Pro88del	385G>A Gly129Ser	505G>A Val169Met	526A>G Asn176Asp	674C>A Ser225Tyr	Frequency (total 54)
A	Val	Pro79_Pro88	Gly	Val	Asn	Ser	21
B						Tyr	14
C		Pro79_Pro88del					4
D					Asp		7
E	Met		Ser	Met			5
G			Ser	Met			1
H			Ser				1
I					Asp	Tyr	1

The PRNP genotype of each individual was determined with PCR amplification and Sanger genotyping to determine which samples to send to sequencing. After the Illumina whole

genome sequencing and variant calling the PRNP genotype was also determined by the observed combination of alleles for each sample. Table 3.3 shows the genotype for all the samples as it was determined by Sanger genotyping and after variant calling. The rows marked in green are samples where the two genotypes matches and the yellow rows show where only one allele differs. The table shows that out of the 27 samples there are 9 that have genotypes that do not match for the two methods of determining genotype and that for 3 of the samples with genotypes that do not match none of the alleles are the same.

Table 3.3: The genotype of each sample as determined by Sanger sequencing and variant calling from Illumina sequence data. Green-colored rows have genotypes that are the same for the two columns, and the rows marked with yellow have one allele that match.

Individuals	Genotypes from Illumina sequence	Genotypes from Sanger sequencing
Mt_1071	A/C	A/E
Mt_557	B/D	B/D
Mt_807	B/E	B/E
Mt_809	B/D	B/D
S15	B/G	A/E
S169	A/D	A/A
S19	A/B	A/B
S197	A/A	A/D
S233	A/E	A/E
S241	A/D	A/D
S264	A/A	A/A
S267	A/H	A/E
S275	B/D	B/D
S283	A/A	A/A
S291	B/B	B/B
S297	A/B	A/B
S33	A/I	B/D
S399	A/A	A/A
S41	A/B	A/B
S49	A/A	A/A
S523	I/C	A/D
S6_553	B/C	B/B
S6_591	D/D	D/D
S83	B/C	B/B
S87	B/E	B/E
S94	E/E	E/E
S96	A/A	A/A

The genome viewing program IGV (Robinson et al., 2011) was used to inspect the reads that had aligned to the positions within the PRNP reading frame. The coverage and base frequencies at positions where the genotypes found with Illumina sequencing and Sanger sequencing differ for the same sample were extracted to table 3.4. The proportion of reads that support the genotype found with Sanger sequencing is marked in green and the proportion of reads that supported the genotype found with Illumina sequencing is marked in yellow. Only the proportions of reads that either supported the genotype found with Illumina sequencing or Sanger sequencing is included in the table. Individual s33 who also had a genotype that was observed differently with Illumina sequencing and Sanger sequencing is not included in the table as both methods agreed that the B and D mutation was present and the inconsistency is based on which haplotype the mutations are found on.

Table 2.4: The sequencing depth and base frequencies at the positions where the genotypes found with Illumina sequencing and Sanger sequencing differs. The table shows the number of reads covering a position and the percentage of reads found with each base. The positions are given in nucleotides from the start of the reeding frame and only the proportion of bases that support either the genotype found with Illumina sequencing or Sanger sequencing is included.
The bases and proportions marked in green are those that support the genotype found with Sanger sequencing and the bases marked in yellow are those that support the genotype found with Illumina sequencing.

		Position in the PRNP reding frame given in nucleotide position from the start of the reding frame					
Individual		4G>A	237_260del	385G>A	505G>A	526A>G	674C>A
mt_1071	coverage	568	625	784	676		
	frequency	A:2%		A: 1%	A:2%		
		G:97%	del:19%	G:99%	G:98%		
S15	coverage	24					23
	frequency	A:25%					C:48%
		G:75%					A:48%
S169	coverage					18	
	frequency					A:44%	
						G:56%	
s197	coverage					35	
	frequency					G:0%	
						A:100%	
S267	coverage	15			13		
	frequency	A:7%			A:15%		
		G:93%			G:85%		
S523	coverage		70				51

	frequency						C:80%
			del:1%				A:18%
S6_553	coverage		105				116
	frequency						C:64%
			Del:1%				A:36%
S83	coverage		48				57
	frequency						C:28%
			Del:1%				A:72%

MDS analysis

To visualize the variation in the dataset a MDS plot was created with PLINK. A MDS plot reduces the variation down the specified number of dimensions, in this case 2, so the difference between data points can be quantified and visualized.

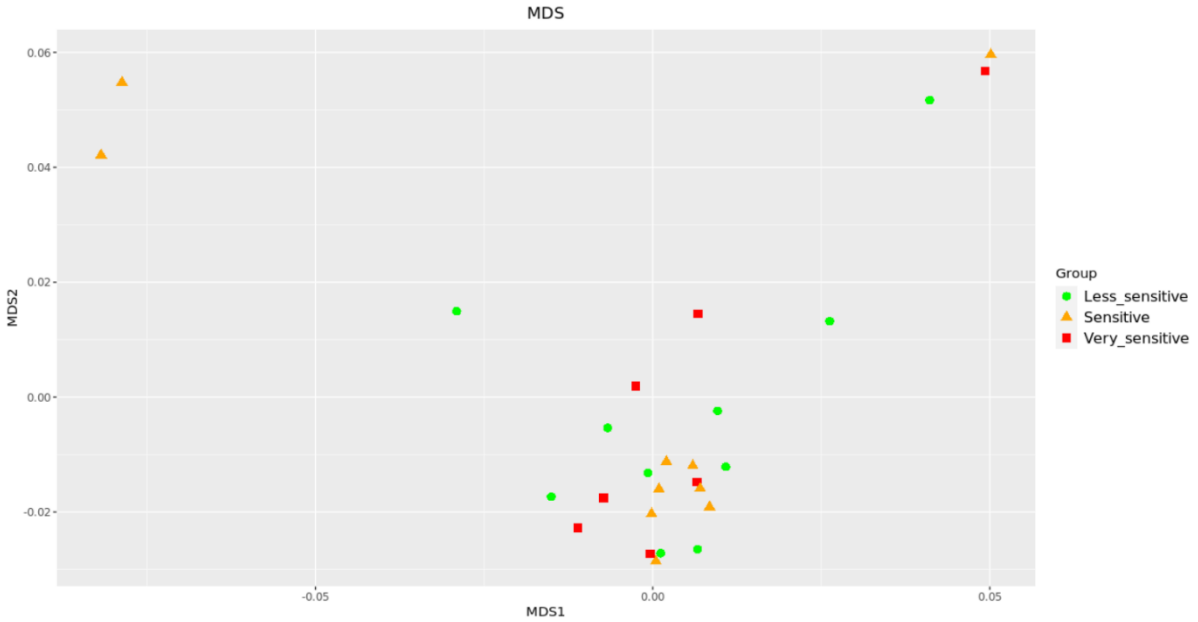


Figure 3.5: MDS plot created from the variant data. The figure shows the genetic distance between each sample, calculated with IBS, as physical distance in the plot. The samples are grouped based on their susceptibility to CWD. Very sensitive individuals are shown in red, sensitive individuals in yellow and less sensitive individuals in green.

In the plot (figure 3.5) similar objects are clustered together. The distance between individuals are calculated by Plink based on Identity-by-state. In other words it shows the samples that have similar genetic content close together. Figure 3.5 shows that when all the variants from the whole genome are included in the analysis there is no clear pattern in the distribution based on the samples sensitivity to CWD.

The stress value calculated for the model is 1.06759e-31 which is close to zero.

LD closer to the PRNP gene

By aligning the sequence of the PRNP gene to the reference genome its position was located. The PRNP gene is located on scaffold JAHWTM010000007.1 in position 3428296-3429077. After quality filtering 227 861 variants were found by Plink to belong on this scaffold. To investigate if LD could be found close to the PRNP gene four MDS plots were made (Figure 3.6). In the plots the genotypes are marked with different symbols and the CWD susceptibility with different colors.

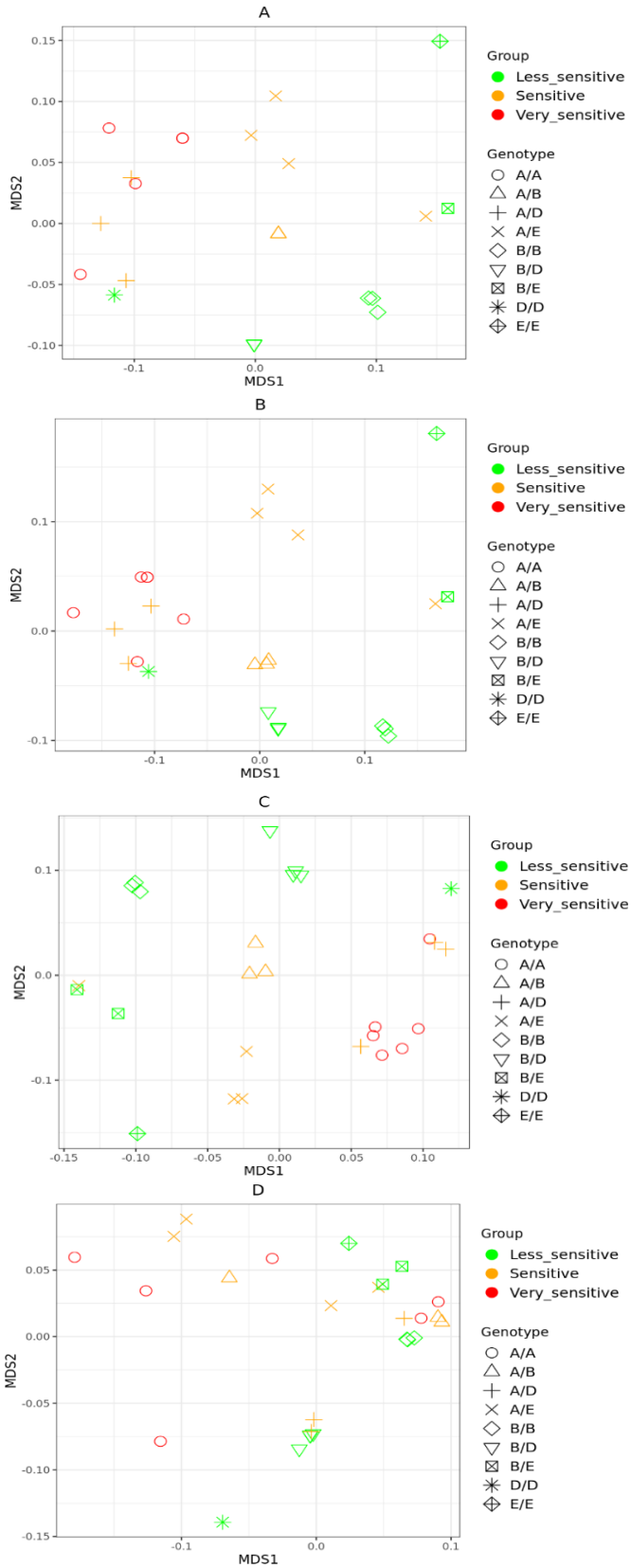


Figure3.5: Four MDS plots created with variants extracted from around the PRNP gene on scaffold JAHWTM010000007. A Includes variants from positions 3000bp up and downstream from the gene, B includes variants 8000bp up and downstream for the gene, C includes variants 20 000bp up and downstream from the gene and D includes variants up to 100 000bp away from the gene. The four plots shows the genetic distance between each sample as physical distance in the plot. The samples are grouped based on their susceptibility to CWD and their genotype is marked with different shapes. The genotypes found with Sanger sequencing is used.

Figure 3.6 shows four MDS plots made with variants found in different stretches around the PRNP gene. Figure 3.6A shows the genetic distances between the individuals when only the 19 variants found in the PRNP gene and 3000bp up and downstream for it was included in the MDS analysis. Plot 3.6B shows the same, but the 49 variants found from up to 8000bp away from the PRNP gene is included. Figure 3.6C shows the same for the 156 variants form positions 20 000bp up and downstream from the PRNP gene and 3.6D includes the 817 variants from positions 100 000bp up and downstream for the PRNP gene. The distribution of the individuals in the MDS plots are changing based on how many positions that are included. A general trend one can see is that the individuals with identical genotypes are less clustered when more positions are included.

As different genotypes were found with Illumina sequencing compared to the one found with Sanger sequencing. A MDS plot with the variants found within 20 000bp away from the PRNP gene was made where the individuals were marked with the genotypes found with Illumina sequencing (Figure 3.7).

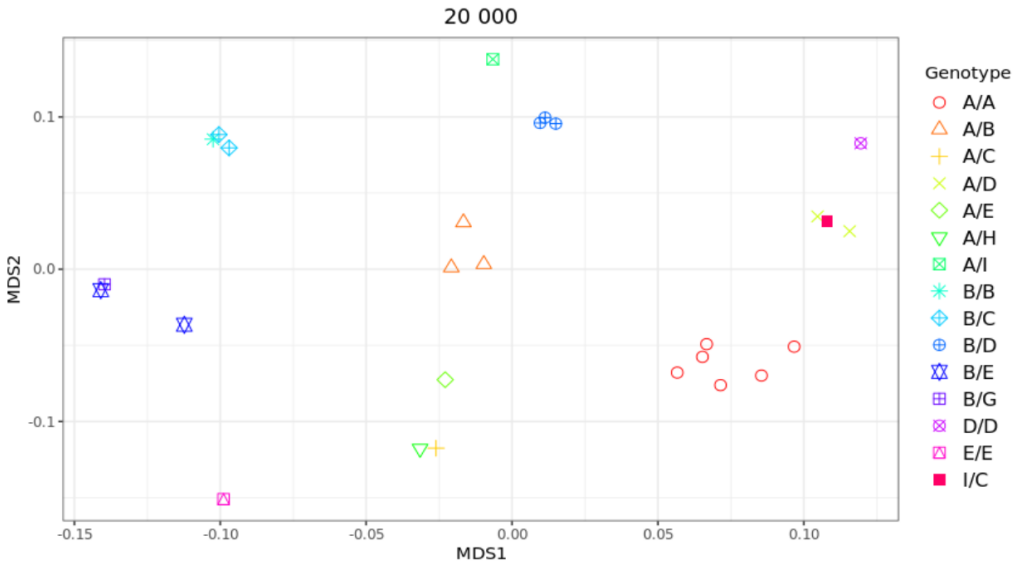


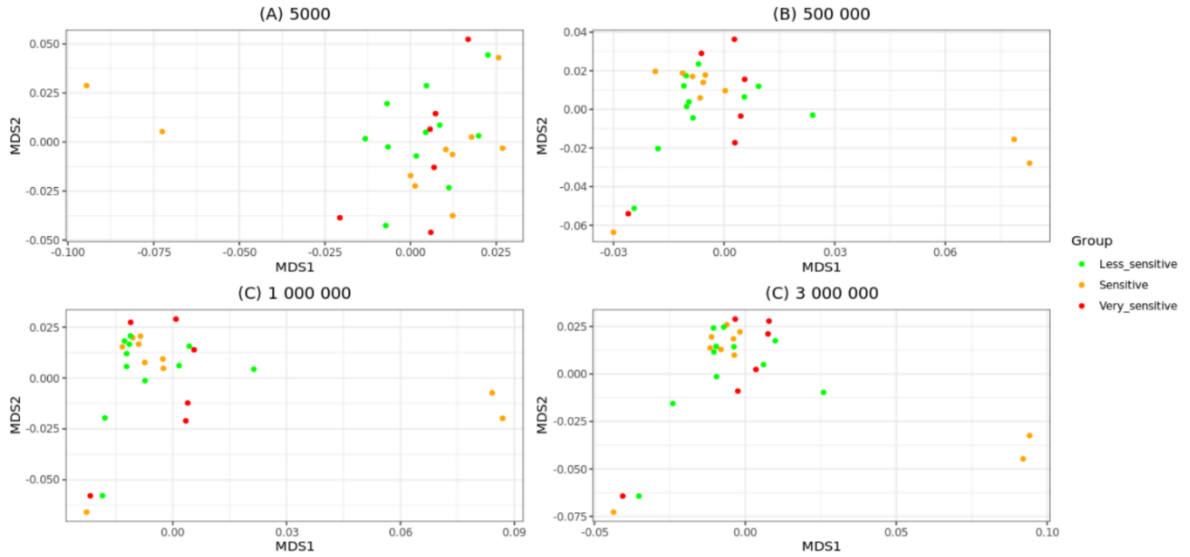
Figure 3.7: MDS plot created from the variants extracted from position 3.408.296-3.449.066 on scaffold JAHWTM010000007.1. The figure shows the genetic distance between each sample as physical distance in the plot. The individuals are marked with the genotype found with Illumina sequencing. The ledger shows the color and shape that corresponds to each genotype.

The figure shows that mostly the individuals with identical PRNP-genotypes clusters together, but notably the individual with genotype B/G clusters together with the individuals with B/E genotypes. The same goes for the sample with I/C genotype which

clusters together with samples with A/D genotype and the sample with B/B genotype which clusters with the samples with B/C genotypes.

How many SNPs are informative?

The finished VCF file contained almost 10 million (9 992 804) variants after quality filtration. By filtering out variants based on Linkage disequilibrium and position the amount of variants were reduced and a MDS plot was created for each of the different amounts of SNPs retained in the file before the analysis. The different plots can be seen in figure 3.8 where 3.8A-D show the MDS plot created respectively with 5000, 500 000, 1 000 000 and 5 000 000 variants.



Figur 3.6: Four different MDS plots created from the same variant dataset, but with different number of variants included in the analysis. The plots shows the genetic distance between samples as physical distance in the plot. A show the distribution of samples when 5000 variants were included in the samples, B shows the same for 500 000 variants, C for 1 000 000 variants and D for 5 000 000 variants. The samples are divided into groups based on their sensitivity to CWD and the groups are marked with different colors in the plot. Very sensitive individuals are marked in red, sensitive individuals with green and less sensitive individuals with green.

The pattern of the distribution of samples can be recognized in all the plots regardless of the number of variants retained, but it is clearer the more variants that were retained in the file. Plot 3.8C and 3.8D, which are created with one million and five million variants respectively, closely resembles that of the MDS plot created with all the obtained variants seen in figure 3.5. For plot 3.8A and 3.8B the distance between the samples has been reduced and though the pattern of distribution of samples remains the same it is less distinct.

4. Discussion

In this study, the population structure of the 27 individuals has been investigated with respect to possible genetic similarities linked to PRNP genotype and sensitivity to CWD. This type of information will be important for understanding if selecting for a population that is more resistant to CWD would significantly reduce the overall genetic variation in the semi-domesticated reindeer populations. If genotypes that make reindeer more susceptible to CWD can be removes from the population without a detrimental loss of genetic variation, this strategy could be an attractive way of reducing the probability of CWD-outbreaks in the semidomesticated herds and consequently a reduced risk of contaminating the neighboring wild populations.

A first step to look into this is to examine the population structure of a captive reindeer herd with known PRNP genotypes.

MDS analysis

A MDS plot is a way to visualize the population structure. The plot will show whether there are any sub-divisions within the population by comparing the genetic distance between all the samples examined. The "goodness of fit" was estimated by calculating a stress value for the MDS plot. Stress values are a measure of discrepancy between the distances in the MDS plot and the original pairwise dissimilarities. Essentially, stress is a measure of how well the MDS plot represents the original data. They are a value between 1 and 0, where values close to zero indicate that the model is a good fit. As the stress value calculated for the MDS plot shown in figure 3.5 is close to zero (1.06759e-31) it indicates that the model is a good fit for displaying the variation in the dataset (Joseph B. Kruskal, 1978; Sturrock & Rocha, 2000).

Figure 3.5 shows the population structure of the study population. There is no clear clustering of the individuals marked with the same colors indicating that there is no correlation between PRNP genotype, and subsequently CWD susceptibility, and genetic similarity when looking at all variable positions in the genome.
Overall figure 3.5 indicates that the population is homogeneous, with the exception of 5 individuals that are genetically different from the rest. These individuals are divided into two clusters one with two individuals and one with 3 individuals as seen in figure 3.5. The clustering indicates that the samples are genetically similar, and different from the

individuals in the other cluster and the rest of the individuals. The reason for these differences could be traced back to a common origin for example a recent ancestor from a wild reindeer population, but as the two clusters show high dissimilarities it is unlikely that they both consists of samples that have an ancestor from a wild population. It would be interesting to perform the same analysis and include samples from wild reindeer to see if the individuals from one of the clusters group together with the individuals from the wild populations. This is however outside of the scope of this study.

The main focus of this study is to examine if removing the A-allele from the population will cause loss of genetic variation in other loci as well. If the samples with identical genotypes had clustered closely together, removing those from the population would obviously cause a reduction in genetic variation. If figure 3.5 showed clustering it would mean that all the samples with the same genotype inherited that genotype from a common ancestor and therefore are closer related to each other than to the rest of the individuals. This is not the case here, plot 3.5 shows no pattern of genetic similarities between individuals with the same genotype. This is a favorable result for the possibility of selective breeding based on PRNP genotypes.

The closer one gets to the PRNP gene the more influence the genotype of the PRNP should have on the clustering of samples. This is because the variants within the PRNP gene and the variants in the flanking regions may be in LD. As plot 3.5 includes variants from the whole genome it is not expected to see an effect of LD here. By looking at MDS plots where gradually more SNPs further away from the PRNP gene is added one can estimate how far in the genome the influence of the PRNP genotype reaches.

The position of the PRNP gene was found to investigate whether clearer clustering could be seen closer to the gene. The PRNP gene was located in position 2428296-3428967 on scaffold JAHWTM010000007. Four MDS plots were created with variants from different stretches around the PRNP gene (figure 3.6). The figure indicates that there is some LD around the PRNP gene as there is clustering of individuals with some genotypes even when the 816 variants from up to 100 000bp away from the PRNP gene is included. As the sample set contains few samples with each genotype the pattern observed is not statistical proof of LD. Trends that can be seen could appear partly coincidentally. A sample set with homozygous animals would have been more suited to investigate LD as the pattern

becomes more distinct when the animals have two of the same allele in the position of interest.

Interestingly Figure 3.6 shows less clustering of the very CWD susceptible animals with A/A genotype. The animals that have genotypes that make them less susceptible to CWD show a more distinct clustering based on genotypes than the other animals. This could indicate that the A allele is an older allele than the others and therefore much recombination of the surrounding area has occurred through time.

The same MDS analysis was performed, but the individuals were marked with the genotype found with Illumina sequencing and only the plot made from the analysis with variants from the positions up to 20 000bp away from The PRNP gene is shown (figure 3.7). The plot shows the same as figure 3.6 with regards to LD.

The MDS analysis indicates that Segregation of samples with different genotypes is small. This is a good sign, but to use these results to implement a breading program one would have to assess the credibility of the result. Important factors in evaluating the predictive value of the analysis is that the quality of the data is good and that the sample pool is representative and contains enough samples that the results are statistically sound.

PRNP variation

A prerequisite for the purpose of this study is the differences in CWD sensitivity based on PRNP genotypes reported by Güere (Güere et al., 2020; Güere et al., 2022). As the research into PRNP variation is an evolving field of study it is interesting to analyze more samples for new variation in the PRNP protein coding region. The variation found in the PRNP reading frame for the 27 animals in this study was extracted to table 3.2. The table shows the variable positions found and how they combine into alleles. Table 1.1 displays the same information from the study performed by Güere at al., (Güere et al., 2020; Güere et al., 2022).

Three additional alleles were found in this study that have not previously been reported by Güere or in other studies. As each of these new alleles are found in a smal number of individuals, two at the most, it is important to evaluete wether these are actual new alleles or stem form sequencing errors or errors in the variant calling. These alleles are part of the

genotypes of four of the nine individuals that have genotypes that are called differently with Sanger sequencing and Illumina sequensing and variant calling.

Table 3.3 shows a worrying statistic for this study. For 9 of the 27 samples, the genotype observed with PCR apmlification and SANGER sequencing (SANGER genotyping) and the genotypes found with whole genome sequencing and variant calling (variant calling) do not match. This means that either the Sanger genotyping or the variant calling determined the genotype of these 9 animals wrong. To investigate possible causes of these errors the genome viewing programe IGV (Robinson et al., 2011) was used to inspect the bam files containing the reads aligned to the reference genome. The program made it posible to see all the reads that had aligned to a spesific position and the bases called in that position for each read. The sequence depth and proportion of reads with each base has been used to evaluate each genotype that differs. Table 3.4 displays this information.

Out of the nine samples that show a mismach between the genotype found with the different methods, four were found to have a C allele with variant calling and not with Sanger genotyping. The 24bp deletion found in this study was not found in the exact position as the C allele repported by Güere (Güere et al., 2020), but as the deletion is found within a repeated region of the PRNP genome, it is almost impossible to determine the exact position of the deletion. Since the deletion is found within the range of positions reported by Güere (nucleotide positions 238 and 272) and is the same length we conclude that the deletion is the same as the one reported by Güere and refer to the deletion as the C allele.

All of the individuals found with C alleles in variant calling have additional differences between the two observed genotypes. Mt_1071 should have an E allele according to the SANGER genotyping, but this is not found in variant calling. Table 3.4 shows that the three mutations making up the E allele are found, but in very smal proportions (less then 3%). HaplotypeCaller therfore discards this as a valid allele.

For sample s523 a D mutation is found with variant calling, but not with SANGER genotyping. For sample s6_553 and s83 the Sanger genotyped geontype was BB and the genotype found with variant calling was B/C, which means that a B mutation was observed with Sanger-genotyping but not with variant calling as the C allele is identical to the A allele exept for the 24bp deletion.

These four samples with C alleles also has a unusuall high coverage in the PRNP region, compaired to the expected coverage (around 20x) and the coverage in the surounding regions. Sample mt_1071 a coverage is over 600 reads for the entire PRNP reading frame which was several magnitudes higher than that found in the adjacent regions. Based on this one possible explanation for the discrepancies in the genotypes observed with the two different methods could be that there are copy number variation of the PRNP gene for these samples. This would explain why there are more reads aligning to this region than to the rest of the genome. It would also explain additional variation found with variant calling as all the reads of the different copies would align to the same position.

For two other samples with mismatches in genotype the descrepencies stem from the observation of the novel G and H allele with variant calling. The two alleles respectively contain two (385G>AGly129Ser and 505G>AVal169Met) and one (385G>AGly129Ser) of the three (4G>AVal2Met, 385G>AGly129Ser and 505G>AVal169Met) mutations that make up the E allele. The individuals s15 and s267 where respectively the G and H allele were observed were both found to have A/E genotype with Sanger sequencing. For sample s267 the observed genotypes would be identical if the observed H allele was called as an E allele with variant calling. Fore sample s15 there is one additional missmach as the genotype observed with variant calling is B/G. The mutation causing the B allele was not found with Sanger sequencing.

Including individual mt_1071 (which is among the four individuals for wich a C allele was found), that makes three of the seven samples that had at least one E allele according to the SANGER-genotyping not called as E alleles in the variant calling. The three mutations making up the E allele are all pressent for the samples with G and H allele, but as for sample mt_1071 the proportion of reads with these mutations is lowe in some of the positions and not all three mutations are called. Because the three mutations making up the E allele always has appear together in previous reaserch performed on larger sample sizes (n=120 and n=364) (Güere et al., 2020; Güere et al., 2022) it is likely thet the variant calling is wrong and needs to be more sensitive for these three mutations. As the proportions of reads with each base is extrapolated from the raw reads aligned to the reference genome the potential problem must be caused in the sequencing process.

The remaining samples with missmaches in genotypes could be explained by two different errors. It could have been an error during phasing that caused the mismach for sample s33 and sample s523. The genotype observed for sample s33 with Sanger sequencing was B/D and the genotype observed with Illumina sequencing was A/I. I is one of the novel alleles found in this study and consists on the B and D mutation (674C>A and 526C>A) found on the same haplotype. Erroneous phasing could have caused this missmach. The I allele is observed for s523 as well and erroneous phasing could have introducet the I allele for this individual as well.
A possible explanation for the missmach in genotypes observed for individual s169 and s197 is that the samples could have been swiched. Sample s196 has the genotype A/A found with variant calling and A/D with genotyping and sample s169 has the exact opposite. This can be controlled by re-sequencing the samples.

The MDS plot in Figure 3.7 show how the individuals cluster when variants close to the PRNP gene is included. As the individuals are marked according to the genotype found with Illumina sequencing one can inspect how individuals with different genotypes clusters and use this to infere if the genotypes are correct. The plot shows that all the individuals that were observed to have a genotype with a C allele with Illumina sequencing clusters together with the individuals they share a genotype with according to Sanger sequencing. The individual observed with B/G genotype with Illumina, clusters together with the individuals with B/E genotype. As this genotype found for this sample with Sanger sequencing was A/E the MDS plot could indicate that both genotype methods called the genotype wrong.

The mismatches in genotypes is very important to investigate furter, and is of great concern for future reaserch. Accurate genotyping of the reindeers is a requirement for assesing whether removing the animals with A alleles and C alleles from the population would lead to loss of genetic variation. It could also slightly change the interpretation of already published studies based on the Sanger genotyping.

Is the quality good?

All the steps in the process must be performed adequately and give outputs that are reliable to ensure the accuracy and trustworthiness of the final analysis. This is why different estimates of the quality, and several quality filtering steps are performed during the process

of compiling a VCF file from the Illumina reads. Errors from sequencing and inaccurate variant calls can introduce false positives in downstream analyses.

Several quality filtration steps were carried out during the pre-processing and variant calling processes. FilterByTile was used to remove reads from tiles that consistently produced reads with lower average quality than the rest of the flowcell, hard filtering was performed on the variants after the first round of variant calling and the VCF file produced in the second variant calling was filtered with Plink. Plink filters out variants based on the composition of SNPs in the VCF file. The filtering by plink removed over 2 million variants from the VCF file. Most of the SNPs were filtered out due to low call rates at specific positions. The filtering thresholds used needs to be determined based on the composition of the dataset and the goal of the study. In this study strict filtering thresholds were used to avoid false positives.

The quality of the reads produced in the sequencing was assessed with FastQC. A summary of all the checks FastQC performed (Figure A, appendix) indicates that the quality of the reads were acceptable. Figure 3.1 shows that the mean quality score for each position in the reads for each sample was good, as the graph for all the samples never leaves the green sone which indicates good quality scores. A high quality score for a position indicates that the probability that the sequencing has called a correct base for that position is high. The reads were therefore used in further analysis without further quality filtration except for the three samples filtered with “FilterByTile”.

Quality measures from the alignment showed that over 95% of the reads for all the samples were “correctly aligned”(figure 3.3), indicating that the read were aligned to a position close together.

Base quality score recalibration was performed to improve the accuracy of the variant calling. On average the base quality score recalibration shifted the quality scores towards lower values. Several studies reports that when the GATK pipeline is used for variant calling the results are more accurate when base score recalibration is performed (DePristo et al., 2011; Pirooznia et al., 2014).

Sample size

An other crucial factor for the applicability of this study is that the sample pool is representative of the entire population. This was considered when choosing samples, as

samples with different genotypes were chosen.
In an article that investigates how many samples it is necessary to include in analyses to gain accurate estimates of population structure and genetic variation, it was suggested that by including as few as 6-8 individuals one gained accurate results (ALISON G. NAZARENO, 2017). This is in contrasts with Røed who suggested that inaccurate estimates of genetic variation of reindeer by Baccus (Baccus et al., 1983) could be caused by an inadequate number of samples, as only 20 individuals were used in the analysis (RØED, 1986). These studies (Baccus et al., 1983; RØED, 1986) were performed with electrophoresis and just a few loci were included in the analysis. Nazareno (ALISON G. NAZARENO, 2017) suggests that small sample sizes can be made up for by including an adequate number of loci when looking at total genetic variation.

To study genetic variation and linkage disequilibrium (LD) in specific areas of the genome it is important to have enough samples, but the genotypes of the included individuals play an important role too. To investigate LD tied to a specific allele the patterns in plots can become clearer by using homozygote animals. For the analysis it is important to have several animals with the same genotype. This is because one needs to compare how similar animals with the same genotypes is to be able to infer the extent of LD tied to an allele. More individuals give more statistical power and make it less likely that the observed differences are coincidental.

In a study performed by Ryman (Ryman et al., 2006) it was noted that usual numbers of individuals for studies on population structure are between 50 and 100. And that a set of samples from this many individuals give results with high statistical power for several different measures of genetic variation.

Further studies

In a study performed on species in the Vitis genus the genetic diversity between the major groups was assessed with PCA. They then further went on to assess the LD pattern by performing an analysis on the LD found in the population (Liang et al., 2019).
A recent study investigating the genetic diversity and population structure of wild and cultivated Crotalaria species used genotyping by sequencing to discover variation. They then

used several methods to investigate the population structure and genetic variation, among others they used correspondence analysis, PCoA, PCA and calculation of LD estimates. They also looked into the genome wide He (Muli et al., 2022).

As these articles show there are more analyses that could be performed to determine the genetic variation in this study, but due to time constraint and limited data it was decided to not proceed with further analyses. The trustworthiness of results of further analysis would also have to be considered thoroughly as there is some uncertainty about the genotypes observed for some of the individuals as discussed previously.

In further studies it could be informative to compare the population structure observed with MDS analysis to other analyses for inferring population structure like Principal component analysis. If matching patterns of population structure was seen with PCA, this would strengthen the credibility of the results.

To create a biological context for the eventual loss of genetic variation caused by removing the A allele from the population one could annotate the variants found close to the PRNP-gene and investigate which traits would be affected if one were to perform selection on reindeer based on PRNP genotype. This could be important information when considering the risk of implementing such a breeding program.

Reducing the number of SNPs

The final VCF file produced in this study contained approximately 10 million variants. The reindeer genome is estimated to be 2,92Gb (Weldenegodguad et al., 2020), that means that approximately 0,3% of the positions in the genome varies.

To estimate whether all the variants in the dataset were essential to retaining the same level of information about the population structure the amount of variants was filtered down to specified numbers based on LD and their position in the genome. MDS plots were then created as a way to compare the level of information kept in the file. Figure 3.8 shows the MDS plots created with 4 different numbers of variants. The figure shows that even when only 5000 were included in the analysis trends of the population structure can still be seen, though it is a lot less distinct and much of the differences between the samples has been lost. Figure 3.8D shows that with 1 million variants included in the analysis the plot remains very similar to the MDS plot created with all the variants.

Based on these results one can argue that filtering the down the number of variants in the

dataset based on LD can make the data more manageable while keeping much of the variation. A study comparing the informativeness of microsatellites and SNPs found that when the least informative SNP markers were included in the analysis the results were less accurate than when only the more informative markers were included (Liu et al., 2005). As the differences between the samples were less obvious when smaller numbers of samples were retained one would still need to keep an adequate number of variants.

5. Conclusion

The results from this study highlight the importance of performing a pilot project. As the genotypes found with Sanger sequencing and with Illumina sequencing differ one needs to further investigate possible causes of these discrepancies before a project involving large scale PRNP-genotyping can be undertaken. In this study we have proposed several explanations for the differences, but more research is needed to conclude exactly what is occurring.

The goal of this study was to investigate the genetic variation in the population and to what extent genetic variation was tied to PRNP genotypes. The study shows no general relatedness between individuals with the same genotype, indicating that there is no common recent ancestor that animals with identical genotype has inherited the genotype from.

Investigation of clustering close to the PRNP gene revealed that there is some LD tied to the different alleles. The result also indicates that there is less LD tied to the A allele than to the other PRNP-alleles. This fits with the fact that the A allele is the most abundant PRNP allele in wild populations of reindeer and is therefore likely an old allele. When it comes to the impact on genetic variation this indicates that few variants are tied closely to the A-allele and therefore the risk of removing additional alleles because of LD is less. It is however worth noting that if one were to remove the A allele from the populations one would lose the allele with most variation around the PRNP-gene and in that way lessen the variation found in the surrounding positions. As a higher proportion of different alleles are found with the A allele there is a chance that some variants may be lost.

The extent of LD around the PRNP gene needs to be examined in more detail with more suitable statistical tests. MDS plots are useful to indicate general trends, but as they show

relative distances precise measures of LD can be better picked up with other models. The sample set used in this study consists of few individuals with the same genotype which makes the information deduced about LD less certain. With few individuals it is a chance that individuals with identical PRNP- genotype are similar by coincidence.

It could also be informative to perform statistical measures of the genetic variation in the population to figure out if the population is diverse. This could be done by measuring heterozygosity.

References

Uncategorized References

ALISON G. NAZARENO, J. B. B., † CHRISTOPHER W. DICK† and LUC IA G. LOHMANN. (2017). Minimum sample sizes for population genomics: an empirical study from an Amazonian plant species. *Molecular Ecology Resources*. https://doi.org/doi: 10.1111/1755-0998.12654

Alnasir, J. J., & Shanahan, H. P. (2020). Intra-exon motif correlations as a proxy measure for mean per-tile sequence quality data in RNA-Seq. *bioRxiv*, 2020.2008.2023.262055. https://doi.org/10.1101/2020.08.23.262055

Baccus, R., Ryman, N., Smith, M. H., Reuterwall, C., & Cameron, D. (1983). Genetic Variability and Differentiation of Large Grazing Mammals. *Journal of Mammalogy*, *64*(1), 109-120. https://doi.org/10.2307/1380756

Benestad, S. L., Mitchell, G., Simmons, M., Ytrehus, B., & Vikøren, T. (2016). First case of chronic wasting disease in Europe in a Norwegian free-ranging reindeer. *Veterinary Research*, *47*(1), 88. https://doi.org/10.1186/s13567-016-0375-4

Benestad, S. L., Mitchell, G., Simmons, M., Ytrehus, B., & Vikøren, T. (2016). First case of chronic wasting disease in Europe in a Norwegian free-ranging reindeer. *Vet Res*, *47*(1), 88. https://doi.org/10.1186/s13567-016-0375-4

Bioinformatics, B. *FastQC: A quality control tool for high throughput sequence data*. Retrieved 12.05 from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Browning, B. L., Tian, X., Zhou, Y., & Browning, S. R. (2021). Fast two-stage phasing of large-scale sequence data. *The American Journal of Human Genetics*, *108*(10), 1880-1890. https://doi.org/https://doi.org/10.1016/j.ajhg.2021.08.005

Bruford, M. W., & Wayne, R. K. (1993). Microsatellites and their application to population genetic studies. *Current Opinion in Genetics & Development*, *3*(6), 939-943. https://doi.org/https://doi.org/10.1016/0959-437X(93)90017-J

Brumfield, R. T., Beerli, P., Nickerson, D. A., & Edwards, S. V. (2003). The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology & Evolution*, *18*(5), 249-256. https://doi.org/https://doi.org/10.1016/S0169-5347(03)00018-1

Bushnell, B. (2020). *BBtools*. In https://sourceforge.net/projects/bbmap/

Caetano-Anolles, D. (2023a, 08.05.2023). *Germline short variant discovery (SNPs + Indels)*. Retrieved 13.05 from https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels-GATK

Caetano-Anolles, D. (2023b, 17.04). *Hard-filtering germline short variants*. Retrieved 13.05 from https://gatk.broadinstitute.org/hc/en-us/articles/360035890471-Hard-filtering-germline-short-variants

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1). https://doi.org/10.1186/s13742-015-0047-8

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

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491-498. https://doi.org/10.1038/ng.806

Dzemyda, G., & Sabaliauskas, M. (2022). Geometric multidimensional scaling: efficient approach for data dimensionality reduction. *Journal of Global Optimization*. https://doi.org/10.1007/s10898-022-01190-8

Eldegard K, S. P., Bjørge A, Kovacs K, Støen O-G og van der Kooij J (2021, 24.11.2021). *Pattedyr: Vurdering av rein Rangifer tarandus for Norge*. Retrieved 23.04.2023 from https://www.artsdatabanken.no/lister/rodlisteforarter/2021/19057

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

Forbes, B. C., & Kumpula, T. (2009). The Ecological Role and Geography of Reindeer (Rangifer tarandus) in Northern Eurasia. *Geography Compass*, *3*(4), 1356-1380. https://doi.org/https://doi.org/10.1111/j.1749-8198.2009.00250.x

Fox, K. A., Jewell, J. E., Williams, E. S., & Miller, M. W. (2006). Patterns of PrPCWD accumulation during the course of chronic wasting disease infection in orally inoculated mule deer (Odocoileus hemionus). *J Gen Virol*, *87*(Pt 11), 3451-3461. https://doi.org/10.1099/vir.0.81999-0

Fremstad, J. J. (2020, 16.07.2020). *Lokalbefolkningen har eierskap til Hardangervidda*. Retrieved 12.05 from https://www.hjortevilt.no/lokalbefolkningen-har-eierskap-til-hardangervidda/

Gunn, A. (2016). *Rangifer tarandus* The IUCN Red List of Threatened Species 2016. https://dx.doi.org/10.2305/IUCN.UK.2016-1.RLTS.T29742A22167140.en

Güere, M. E., Våge, J., Tharaldsen, H., Benestad, S. L., Vikøren, T., Madslien, K., Hopp, P., Rolandsen, C. M., Røed, K. H., & Tranulis, M. A. (2020). Chronic wasting disease associated with prion protein gene (PRNP) variation in Norwegian wild reindeer (Rangifer tarandus). *Prion*, *14*(1), 1-10. https://doi.org/10.1080/19336896.2019.1702446

Güere, M. E., Våge, J., Tharaldsen, H., Kvie, K. S., Bårdsen, B.-J., Benestad, S. L., Vikøren, T., Madslien, K., Rolandsen, C. M., Tranulis, M. A., & Røed, K. H. (2022). Chronic wasting disease in Norway—A survey of prion protein gene variation among cervids [https://doi.org/10.1111/tbed.14258]. *Transboundary and Emerging Diseases*, *69*(4), e20-e31. https://doi.org/https://doi.org/10.1111/tbed.14258

Haley, N., Donner, R., Merrett, K., Miller, M., & Senior, K. (2021). Selective Breeding for Disease-Resistant PRNP Variants to Manage Chronic Wasting Disease in Farmed Whitetail Deer. *Genes (Basel)*, *12*(9). https://doi.org/10.3390/genes12091396

Haley, N. J., & Hoover, E. A. (2015). Chronic wasting disease of cervids: current knowledge and future perspectives. *Annu Rev Anim Biosci*, *3*, 305-325. https://doi.org/10.1146/annurev-animal-022114-111001

Happ, G. M., Huson,H.J. and Beckmen,K.J. (2005). *Prion genotypes in feral herds of Alaska caribou*. https://www.ncbi.nlm.nih.gov/nuccore/DQ154293.1?report=genbank#sequence_DQ154293.1

Harvard.edu. https://www.cog-genomics.org/plink/2.0/ld

Heggberget, T. M., Gaare, E., & Ball, J. P. (2002). Reindeer (Rangifer tarandus) and climate change: Importance of winter forage. *Rangifer*, *22*(1), 13-31. https://doi.org/10.7557/2.22.1.388

Huang, Z., Gabriel, J. M., Baldwin, M. A., Fletterick, R. J., Prusiner, S. B., & Cohen, F. E. (1994). Proposed three-dimensional structure for the cellular prion protein. *Proceedings of the National Academy of Sciences*, *91*(15), 7139-7143. https://doi.org/10.1073/pnas.91.15.7139

Haasl, R. J., & Payseur, B. A. (2011). Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites. *Heredity*, *106*(1), 158-171. https://doi.org/10.1038/hdy.2010.21

Jade, C. S. C., & Swaine, L. C. (2017). Lacer: accurate base quality score recalibration for improving variant calling from next-generation sequencing data in any organism. *bioRxiv*, 130732. https://doi.org/10.1101/130732

Johnson, C., Johnson, J., Vanderloo, J. P., Keane, D., Aiken, J. M., & McKenzie, D. (2006). Prion protein polymorphisms in white-tailed deer influence susceptibility to chronic wasting disease. *J Gen Virol*, *87*(Pt 7), 2109-2114. https://doi.org/10.1099/vir.0.81615-0

Johnson, C. J., Herbst, A., Duque-Velasquez, C., Vanderloo, J. P., Bochsler, P., Chappell, R., & McKenzie, D. (2011). Prion protein polymorphisms affect chronic wasting disease progression. *PLoS One*, *6*(3), e17450. https://doi.org/10.1371/journal.pone.0017450

Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L., & Daszak, P. (2008). Global trends in emerging infectious diseases. *Nature*, *451*(7181), 990-993. https://doi.org/10.1038/nature06536

Joseph B. Kruskal, M. W. (1978). *Multidimensional Scaling* (11 ed.).

Kahn, S., Dubé, C., Bates, L., & Balachandran, A. (2004). Chronic wasting disease in Canada: Part 1. *Can Vet J*, *45*(5), 397-404.

Kaltenborn, B. P., Andersen, O., & Gundersen, V. (2014). The role of wild reindeer as a flagship species in new management models in Norway. *Norsk Geografisk Tidsskrift - Norwegian Journal of Geography*, *68*(3), 168-177. https://doi.org/10.1080/00291951.2014.904400

Kharzinova, V. R., Dotsev, A. V., Deniskova, T. E., Solovieva, A. D., Fedorov, V. I., Layshev, K. A., Romanenko, T. M., Okhlopkov, I. M., Wimmers, K., Reyer, H., Brem, G., & Zinovieva, N. A. (2018). Genetic diversity and population structure of domestic and wild reindeer (Rangifer tarandus L. 1758): A novel approach using BovineHD BeadChip. *PLoS One*, *13*(11), e0207944. https://doi.org/10.1371/journal.pone.0207944

Kiel, i. o. c. m. b. (2021). *Genome assembly of Norwegain reindeer (R. tarandus)*. https://www.ncbi.nlm.nih.gov/genome/7845?genome_assembly_id=1699206

Kim, T. Y., Shon, H. J., Joo, Y. S., Mun, U. K., Kang, K. S., & Lee, Y. S. (2005). Additional cases of Chronic Wasting Disease in imported deer in Korea. *J Vet Med Sci*, *67*(8), 753-759. https://doi.org/10.1292/jvms.67.753

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1-27. https://doi.org/10.1007/BF02289565

Kvie, K. S., Heggenes, J., Bårdsen, B.-J., & Røed, K. H. (2019). Recent large-scale landscape changes, genetic drift and reintroductions characterize the genetic structure of Norwegian wild reindeer. *Conservation Genetics*, *20*(6), 1405-1419. https://doi.org/10.1007/s10592-019-01225-w

Lacy, R. C. (1997). Importance of Genetic Variation to the Viability of Mammalian Populations. *Journal of Mammalogy*, *78*(2), 320-335. https://doi.org/10.2307/1382885

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357-359. https://doi.org/10.1038/nmeth.1923

Leiss, W., Westphal, M., Tyshenko, M.G., Croteau, M.C., Oraby, T., Adamowicz, W., Goddard, E., Cashman, N.R., Darshan, S. and Krewski, D. (2017). Challenges in managing the risks of chronic wasting disease. *International Journal of Global Environmental Issues*, *16*(4), 277-302. https://doi.org/10.1504/ijgenvi.2017.086716

Lewontin, R. C., & Hubby, J. L. (1966). A MOLECULAR APPROACH TO THE STUDY OF GENIC HETEROZYGOSITY IN NATURAL POPULATIONS. II. AMOUNT OF VARIATION AND DEGREE OF HETEROZYGOSITY IN NATURAL POPULATIONS OF DROSOPHILA PSEUDOOBSCURA. *Genetics*, *54*(2), 595-609. https://doi.org/10.1093/genetics/54.2.595

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

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

Liang, Z., Duan, S., Sheng, J., Zhu, S., Ni, X., Shao, J., Liu, C., Nick, P., Du, F., Fan, P., Mao, R., Zhu, Y., Deng, W., Yang, M., Huang, H., Liu, Y., Ding, Y., Liu, X., Jiang, J., . . . Dong, Y. (2019). Whole-genome resequencing of 472 Vitis accessions for grapevine diversity and demographic history analyses. *Nature Communications*, *10*(1), 1190. https://doi.org/10.1038/s41467-019-09135-8

Liberski, P. P. (2012). Historical overview of prion diseases: a view from afar. *Folia Neuropathol*, *50*(1), 1-12.

Liu, N., Chen, L., Wang, S., Oh, C., & Zhao, H. (2005). Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure. *BMC Genetics*, *6*(1), S26. https://doi.org/10.1186/1471-2156-6-S1-S26

Maraud, S., & Roturier, S. (2021). Chronic Wasting Disease (CWD) in Sami Reindeer Herding: The Socio-Political Dimension of an Epizootic in an Indigenous Context. *Animals (Basel)*, *11*(2). https://doi.org/10.3390/ani11020297

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297-1303. https://doi.org/10.1101/gr.107524.110

Miller, M. W., & Fischer, J. R. (2016). The first five (or more) decades of chronic wasting disease: lessons for the five decades to come. Transactions of the North American Wildlife and Natural Resources Conference,

Miller, M. W., & Williams, E. S. (2004). Chronic wasting disease of cervids. *Curr Top Microbiol Immunol*, *284*, 193-214. https://doi.org/10.1007/978-3-662-08441-0_8

Muli, J. K., Neondo, J. O., Kamau, P. K., Michuki, G. N., Odari, E., & Budambula, N. L. M. (2022). Genetic diversity and population structure of wild and cultivated Crotalaria species based on genotyping-by-sequencing. *PLoS One*, *17*(9), e0272955. https://doi.org/10.1371/journal.pone.0272955

Mysterud, A., & Rolandsen, C. M. (2018). A reindeer cull to prevent chronic wasting disease in Europe. *Nature Ecology & Evolution*, *2*(9), 1343-1345. https://doi.org/10.1038/s41559-018-0616-1

Mysterud, A., Strand, O., & Rolandsen, C. M. (2020). Embracing fragmentation to save reindeer from disease. *Conservation Science and Practice*, *2*(8), e244. https://doi.org/https://doi.org/10.1111/csp2.244

Nellemann, C., Vistnes, I., Jordhøy, P., & Strand, O. (2001). Winter distribution of wild reindeer in relation to power lines, roads and resorts. *Biological Conservation*, *101*(3), 351-360. https://doi.org/https://doi.org/10.1016/S0006-3207(01)00082-9

Nevo, E. (1978). Genetic variation in natural populations: Patterns and theory. *Theoretical Population Biology*, *13*(1), 121-177. https://doi.org/https://doi.org/10.1016/0040-5809(78)90039-4

Notter, D. R. (1999). The importance of genetic diversity in livestock populations of the future1. *Journal of Animal Science*, *77*(1), 61-69. https://doi.org/10.2527/1999.77161x

O'Connor, V. d. A. G. (2020). *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra* (1st Edition ed.). O'Reilly Media.

O'Rourke, K. I., Spraker, T. R., Zhuang, D., Greenlee, J. J., Gidlewski, T. E., & Hamir, A. N. (2007). Elk with a long incubation prion disease phenotype have a unique PrPd profile. *Neuroreport*, *18*(18), 1935-1938. https://doi.org/10.1097/WNR.0b013e3282f1ca2f

Otero, A., Velásquez, C. D., Aiken, J., & McKenzie, D. (2021). Chronic wasting disease: a cervid prion infection looming to spillover. *Veterinary Research*, *52*(1), 115. https://doi.org/10.1186/s13567-021-00986-y

Packer, C., Pusey, A. E., Rowley, H., Gilbert, D. A., Martenson, J., & O'Brien, S. J. (1991). Case Study of a Population Bottleneck: Lions of the Ngorongoro Crater. *Conservation Biology*, *5*(2), 219-230. http://www.jstor.org/stable/2386196

Pan, K. M., Baldwin, M., Nguyen, J., Gasset, M., Serban, A., Groth, D., Mehlhorn, I., Huang, Z., Fletterick, R. J., Cohen, F. E., & et al. (1993). Conversion of alpha-helices into beta-sheets features in the formation of the scrapie prion proteins. *Proc Natl Acad Sci U S A*, *90*(23), 10962-10966. https://doi.org/10.1073/pnas.90.23.10962

Perucchini, M., Griffin, K., Miller, M. W., & Goldmann, W. (2008). PrP genotypes of free-ranging wapiti (Cervus elaphus nelsoni) with chronic wasting disease. *J Gen Virol*, *89*(Pt 5), 1324-1328. https://doi.org/10.1099/vir.0.83424-0

Picard tools - by Broad Institute. In. http://broadinstitute.github.io/picard/
Pirisinu, L., Tran, L., Chiappini, B., Vanni, I., Di Bari, M. A., Vaccari, G., Vikøren, T., Madslien, K. I., Våge, J., Spraker, T., Mitchell, G., Balachandran, A., Baron, T., Casalone, C., Rolandsen, C. M., Røed, K. H., Agrimi, U., Nonno, R., & Benestad, S. L. (2018). Novel Type of Chronic Wasting Disease Detected in Moose (Alces alces), Norway. *Emerg Infect Dis*, *24*(12), 2210-2218. https://doi.org/10.3201/eid2412.180702
Pirooznia, M., Kramer, M., Parla, J., Goes, F. S., Potash, J. B., McCombie, W. R., & Zandi, P. P. (2014). Validation and assessment of variant calling pipelines for next-generation sequencing. *Human Genomics*, *8*(1), 14. https://doi.org/10.1186/1479-7364-8-14
Prusiner, S. B. (1998). Prions. *Proceedings of the National Academy of Sciences*, *95*(23), 13363-13383. https://doi.org/10.1073/pnas.95.23.13363
Purcell, S. *PLINK 1.9*. In http://pngu.mgh.harvard.edu/purcell/plink/
Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, *81*(3), 559-575. https://doi.org/10.1086/519795
Race, B., Williams, K., Orrú, C. D., Hughson, A. G., Lubke, L., & Chesebro, B. (2018). Lack of Transmission of Chronic Wasting Disease to Cynomolgus Macaques. *Journal of Virology*, *92*(14), e00550-00518. https://doi.org/doi:10.1128/JVI.00550-18
Rentería, M. E., Cortes, A., & Medland, S. E. (2013). Using PLINK for Genome-Wide Association Studies (GWAS) and Data Analysis. In C. Gondro, J. van der Werf, & B. Hayes (Eds.), *Genome-Wide Association Studies and Genomic Prediction* (pp. 193-213). Humana Press. https://doi.org/10.1007/978-1-62703-447-0_8
Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, *16*(6), 276-277. https://doi.org/https://doi.org/10.1016/S0168-9525(00)02024-2
Richards, B. J. (2021). Chronic Wasting Disease distribution in the United States by state and county: U.S. Geological Survey data release. In.
Rivera, N. A., Brandt, A. L., Novakofski, J. E., & Mateus-Pinilla, N. E. (2019). Chronic Wasting Disease In Cervids: Prevalence, Impact And Management Strategies. *Vet Med (Auckl)*, *10*, 123-139. https://doi.org/10.2147/vmrr.S197404
Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nat Biotechnol*, *29*(1), 24-26. https://doi.org/10.1038/nbt.1754
Ryman, N., Palm, S., André, C., Carvalho, G. R., Dahlgren, T. G., Jorde, P. E., Laikre, L., Larsson, L. C., Palmé, A., & Ruzzante, D. E. (2006). Power for detecting genetic divergence: differences between statistical methods and marker loci. *Molecular Ecology*, *15*(8), 2031-2045. https://doi.org/https://doi.org/10.1111/j.1365-294X.2006.02839.x
RØED, K. H. (1985). Genetic variability in Norwegian semi-domestic reindeer (Rangifer tarandus. *Hereditas*, *102*(2), 177-184. https://doi.org/https://doi.org/10.1111/j.1601-5223.1985.tb00612.x
RØED, K. H. (1986). Genetic variability in Norwegian wild reindeer (Rangifer tarandus L.). *Hereditas*, *104*(2), 293-298. https://doi.org/https://doi.org/10.1111/j.1601-5223.1986.tb00542.x
Røed, K. H., Kvie, K. S., Bårdsen, B.-J., Laaksonen, S., Lohi, H., Kumpula, J., Aronsson, K.-Å., Åhman, B., Våge, J., & Holand, Ø. (2021). Historical and social–cultural processes as drivers for genetic structure in Nordic domestic reindeer. *Ecology and Evolution*, *11*(13), 8910-8922. https://doi.org/https://doi.org/10.1002/ece3.7728
Shaun Purcell, C. C. *PLINK 2.0*. In www.cog-genomics.org/plink/2.0/
Sohn, H. J., Kim, J. H., Choi, K. S., Nah, J. J., Joo, Y. S., Jean, Y. H., Ahn, S. W., Kim, O. K., Kim, D. Y., & Balachandran, A. (2002). A case of chronic wasting disease in an elk imported to Korea from Canada. *J Vet Med Sci*, *64*(9), 855-858. https://doi.org/10.1292/jvms.64.855

Statusrapport CWD for 2018. (2019). https://www.hjortevilt.no/wp-content/uploads/2019/10/statusrapportcwdfor2018.pdf
Sturrock, K., & Rocha, J. (2000). A multidimensional scaling stress evaluation table. *Field methods*, *12*(1), 49-60.
Team, R. (2020). *RStudio: Integrated Development Environment for R*. In (Version 4.1.0) http://www.rstudio.com/
Tranulis, M. A., Gavier-Widén, D., Våge, J., Nöremark, M., Korpenfelt, S.-L., Hautaniemi, M., Pirisinu, L., Nonno, R., & Benestad, S. L. (2021). Chronic wasting disease in Europe: new strains on the horizon. *Acta Veterinaria Scandinavica*, *63*(1), 48. https://doi.org/10.1186/s13028-021-00606-x
Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*, *43*(1), 11.10.11-11.10.33. https://doi.org/https://doi.org/10.1002/0471250953.bi1110s43
Veterinærinstituttet. *Tamrein*. Retrieved 23.4 from https://www.vetinst.no/dyr/tamrein
Veterinærinstituttet. (2023, 26.02.2023). *Skrantesjukestatistikk*. Retrieved 26.02.2023 from http://apps.vetinst.no/skrantesykestatistikk/NO/#kasus
Weldenegodguad, M., Pokharel, K., Ming, Y., Honkatukia, M., Peippo, J., Reilas, T., Røed, K. H., & Kantanen, J. (2020). Genome sequence and comparative analysis of reindeer (Rangifer tarandus) in northern Eurasia. *Sci Rep*, *10*(1), 8980. https://doi.org/10.1038/s41598-020-65487-y
Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. In Springer-Verlag New York. https://ggplot2.tidyverse.org
Williams, E. S., & Young, S. (1980). Chronic wasting disease of captive mule deer: a spongiform encephalopathy1. *Journal of Wildlife Diseases*, *16*(1), 89-98. https://doi.org/10.7589/0090-3558-16.1.89
Wright, S. (1968). *Evolution and the genetics of populations a treatise*. University of Chicago Press.

Appendix

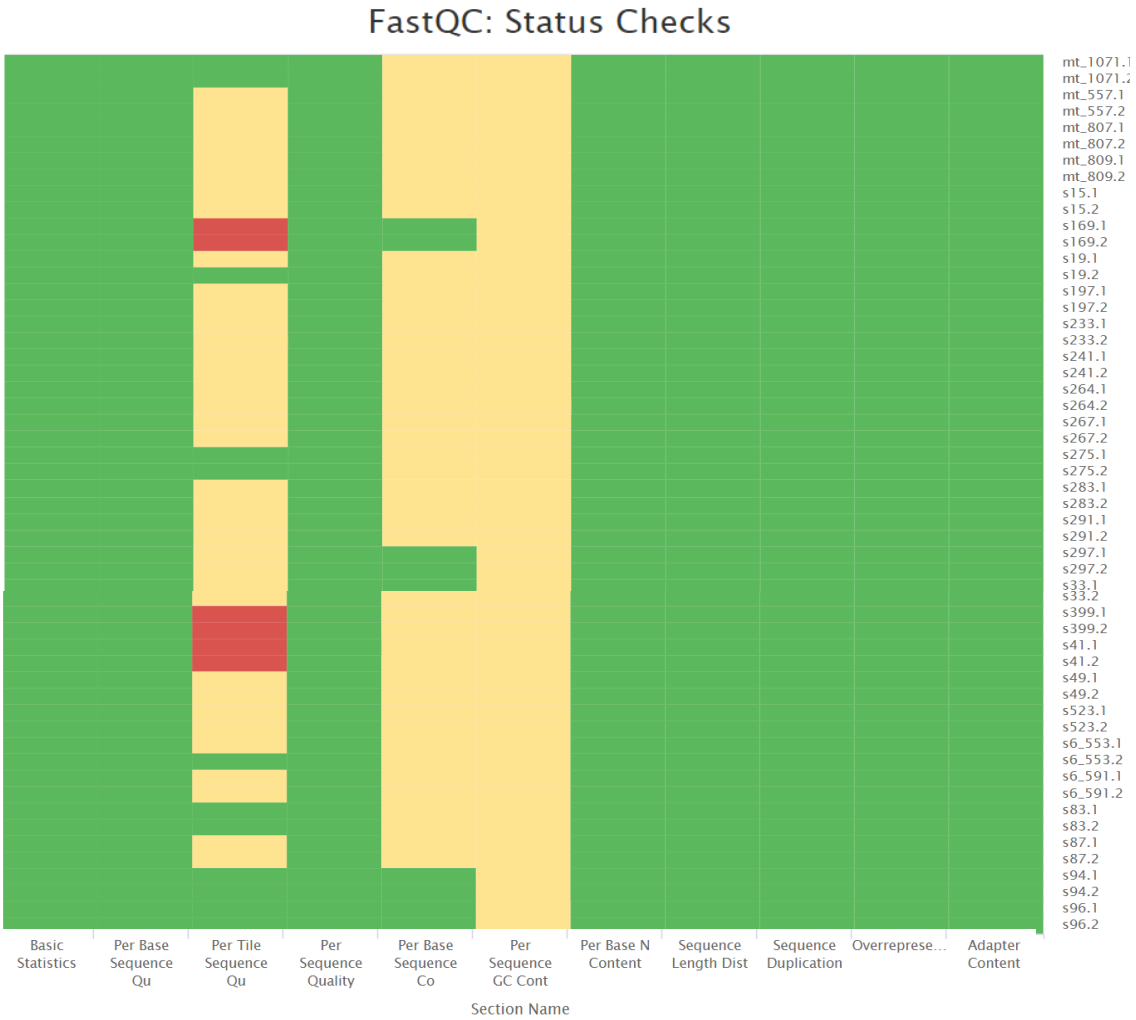


Figure A1 shows whether the samples passed certain quality controls by fastQC. Most of the fields are green which indicates that they passed the quality check. The fields which are yellow is to indicate that fast QC produced a warning for that quality measure. Red fields means that the values were outside the threshold that FastQC consider acceptable. The forward and reverse read files of three samples were marked as failed in the quality control check "per tile sequence quality". This indicates that part of the flowcell consistently produced reads with poor quality.

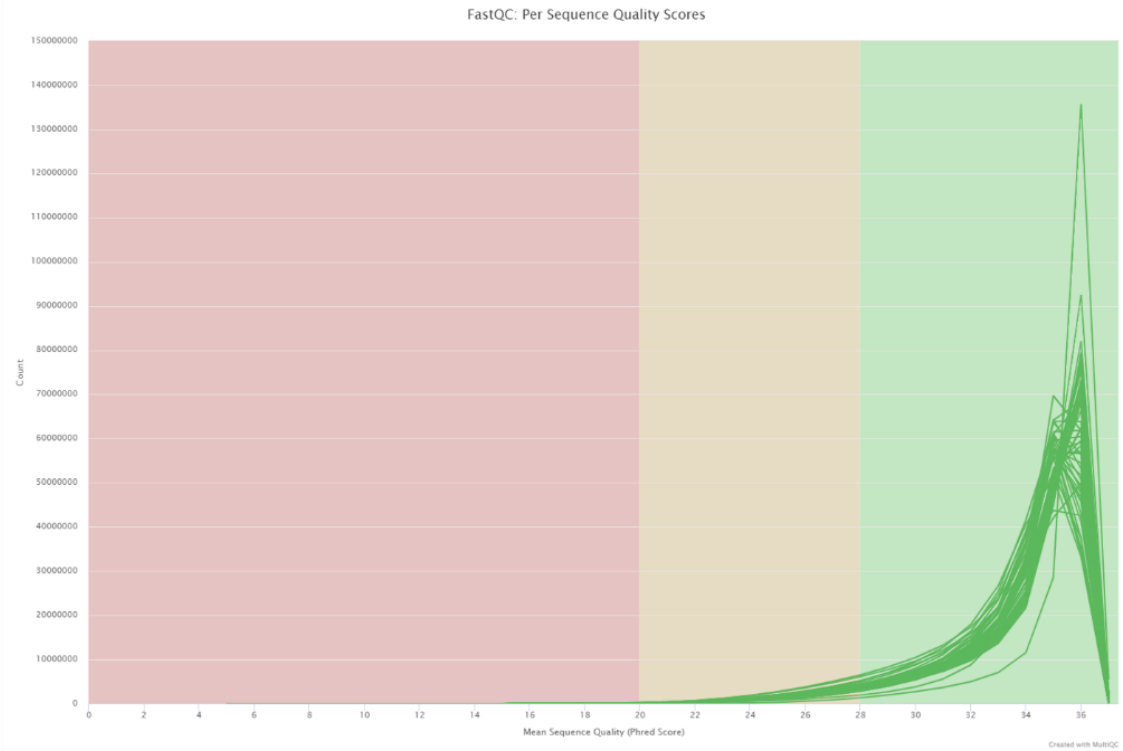


Figure A2: Count of reads with a certain quality score. Each line represents one sample. The x-axis shows possible quality scores while the y-axis shows amount of reads.

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

Postboks 5003
NO-1432 Ås
Norway