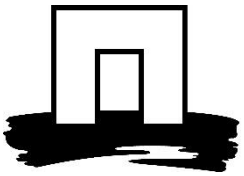
DNA test for Dutch native cattle breeds and its influence on genetic diversity

Ineke Koning
REG.NR.: 1022805

Major thesis Animal Breeding and Genomics (ABG-80436)
August 2022

Supervisors/examiners:
Jack Windig
Mira Schoon

Thesis: Animal Breeding and Genomics



WAGENINGEN
UNIVERSITY & RESEARCH

Preface

First of all I would like to thank my supervisors Jack Windig and Mira Schoon for the help and feedback they gave me! Next to that, I would like to thank Ina Hulsegge for providing additional support with R studio and Beagle. Finally, I would like to thank everyone involved in the Master Thesis Rings of ABG for the tips and tricks they have given.

Abstract

In 2018 a DNA test was designed to distinguish six native Dutch cattle breeds from each other and determine the resemblance to the purebred individuals of these breeds. For the original test 133 SNP's were selected from a reference population based on the genetic diversity between these breeds. In the three years that the test has been used new individuals have been added to the studbooks and to the reference populations. In this paper the effects of the DNA test on genetic diversity of the reference population was analysed, and an evaluation of the DNA test was done to check whether improvements can and have to be made. The effects of the test on genetic diversity were analysed by calculating F statistics for the original versus the new populations and comparing the PCA results on these populations. The evaluation of the DNA test was done by making a new SNP selection, based on the same method as the original selection, and comparing the quality of the new selection with the original selection. Analysis of the effects of the test showed that most animals that were tested were determined as Dutch Friesians, and an overall increase in the total expected heterozygosity was found while the Fst value was lower. Evaluation of the test showed that the updated and original SNP selection had similar results overall with some differences. The updated selection is slightly better at separating Dutch Friesian from Dutch Friesian Red and White, and the original is slightly better at separating Deep Red from Meuse-Rhine-Yssel. Overall, the DNA test has a positive effect on genetic diversity and provides an accurate way to assign animals with missing pedigree information to one of the purebred populations of the six native Dutch cattle breeds.

Table of Contents

1. Introduction

Genetic diversity is the genetic difference between different species, breeds or individuals (Oldenbroek, 2017). It's important for wildlife populations to adapt to their environment through the process of natural selection. Individuals with a genotype that is more suited to its environment will generally live longer and contribute more to the next generation, both in numbers of offspring and genetics. The more genetically diverse a population is, the more likely it is to contain individuals with suitable genetics to the current situation and adapt to the future environment. Likewise, genetic diversity has benefits for livestock as well, as it creates the possibility to respond to artificial selection and can be used to optimize breeding programmes.
Genetic diversity is influenced by the appearing and disappearing of alleles (Ellegren & Galtier, 2016; Oldenbroek, 2017). New alleles can be introduced in a species or population by mutations caused by, for example, replication errors, and alleles can disappear again by a process called genetic drift. Genetic drift is the disappearing of alleles by chance due to random sampling which happens more often in smaller populations. A population can also gain new alleles by the process of immigration. When two populations are separated, they will start to differ genetically due to genetic drift. Which alleles disappear is determined by chance and they will therefore likely be different in both populations and lead to genetic differences between populations. Letting individuals from one population immigrate into the other, will increase genetic diversity within populations again as lost alleles may be reintroduced into the population. In small populations, the impact of genetic drift is faster than immigration of new alleles, causing the genetic diversity to decrease which might results in breeds disappearing completely.

Genetic diversity can be measured in various ways. In the past it used to be done by looking at the pedigree of individuals and applying statistical models (Oldenbroek, 2017). Nowadays DNA analysis has become more prominent for determining genetic diversity, as the costs are becoming less expensive while techniques and knowledge improve. When a reference genome is available, an analysis of single nucleotide polymorphisms (SNP's) is enough to determine genetic diversity (Oldenbroek, 2017).
This can be done by, for example, calculating the proportion of polymorphic markers within a breed or population, which means the proportion of SNP's that are varying between the individuals and are thus not fixed in the breed or population (Berg & Hamrick, 1997; Oldenbroek, 2017). Another method is the expected heterozygosity, which looks at the frequency of all different alleles at a marker and calculates the chance that a random individual will be heterozygous at that marker. does take this into account (Nei, 1973).

In the Netherlands there are six dual-purpose native cattle breeds (Table 1): Deep Red (DR), Dutch Friesian (DF), Groningen White Headed (GWH), Dutch Belted (DB), Meuse-Rhine-Yssel (MRY) and Dutch Friesian Red and White (DFR) (SZH, 2021). Five of these breeds have less than 3,000 female breeding animals (CGN, 2022) and therefore receive the risk status "threatened" according to the FAO risk classification (Food and Agriculture Organization of the United Nations., 2013). Begin 2022, MRY is the only breed that has the risk status "normal" with more than 6,000 female breeding animals. The possibility of infectious disease outbreaks and preventive culling of animals in the same area could decrease the number of individuals until there are too little individuals left to sustain the populations.
This is one of the reasons why the Dutch government asked the 'Dutch rare breed survival trust' (*Stichting Zeldzame Huisdierrassen*) to set up a system to assign a rare breed status to individual animals, called 'Umbrella data base' (Paraplubestand) (SZH, 2021). In the case of an infectious disease outbreak, the government will see whether other solutions are possible instead of preventative culling, such as vaccination or quarantine. To qualify for these exemptions, individual animals have to be registered as purebred animal at one of the Dutch native rare breed studbooks.

Table 1 Overview of the dual-purpose native cattle breeds with their English name and abbreviation, Dutch name and number of female individuals.

English abbreviation	English Name	Dutch Name	Female individuals in 2020/2021 (CGN, 2022)
DR	Deep Red	Brandrood rund	1,121
DF	Dutch Friesian	Fries-Hollands vee	2,789
GWH	Groningen White Headed	Groninger Blaarkop	1,408
DB	Dutch Belted	Lakenvelder	1,608
MRY	Meuse-Rhine-Yssel	Maas-Rijn-Ijsselvee	9,968
DFR	Dutch Friesian Red and White	Roodbont Friesvee	594

Unfortunately, not all individuals from native cattle breeds are registered in studbooks. To register in a studbook, individuals officially need sufficient pedigree information to prove their purebred status (*Verordening (EU) 2016/1012 van Het Europees Parlement En de Raad*, 2016). However, when the conservation of rare breeds is involved, exceptions can be made. This is why a DNA test was designed to distinguish purebred individuals, with (partial) missing pedigree information and assign them to one of the native breeds (Hulsegge et al., 2019). Adding these animals to the studbook population could increase the population size, decrease inbreeding, and decrease the impact of circumstantial measures such as for example disease outbreaks.

Hulsegge et al. 2019 used the differences in genetic diversity between the breeds to design the DNA test. Based on a reference population, they selected a total of 133 SNP's that were needed to correctly distinguish between DR, DF, GWH, DB, MRY and DFR and assign individuals to the purebred breeding populations of these breeds. The use of this DNA test is expected to have increased the population sizes of the native Dutch cattle breeds by assigning new individuals to the studbook population. The populations outside the studbook are expected to be genetically similar to the corresponding studbook individuals, but with some genetic differences, making it likely that the newly added animals increase the genetic diversity within and between the native Dutch cattle breeds as well. The changes in genetic diversity might influence the power of the DNA test, as the test is based on the genetic diversity of the old studbook population.

In this research I looked into the effect of the DNA test on the genetic diversity of the native dual-purpose Dutch cattle breeds and how the DNA test can be improved with these changes in genetic diversity. This will be done by answering the following research questions:

- What is the effect of the individuals added by the DNA test on the genetic diversity within the breeding populations of the Dutch native cattle breeds, and how does this affect the power of the DNA test?
    - Which animals have been added by the use of the DNA test?
    - How do the test animals and newly added individuals change the genetic diversity within the reference populations?
    - Does the SNP selection of the DNA test have to be improved based on the new reference population?

2. Material & Methods

2.0. Preparation genotypic data

Genotypic data in ACGT format is available of the current reference population and of all individuals that have been tested with the DNA test (tested individuals). Both datasets consist of 53,219 SNP's and include the 133 SNP's that were previously selected for the DNA test (Hulsegge et al., 2019). These datasets were combined for the quality check and imputation.

The quality of the data was checked in Rstudio. SNP's and individuals with more than 5% missing data were removed from the dataset leaving 706 individuals and 37,123 SNP's. Missing SNP's were then imputed with Beagle (reference) using 20 iterations. Each breed was imputed separately, except for the Holstein individuals as those were already imputed.
The dataset was then split into current reference population (630 individuals) and the tested individuals (76 individuals). The current reference population includes the original reference population and the individuals that have been newly added to the reference population since the DNA test has been put into use. Therefore, the current reference population was also split into the original reference population (621 individuals) and the newly added individuals (9 individuals).

Table 2 Overview of the number of individuals per breed of each of the populations.

	Cref	Test	Oref	New
DR	45	0	40	5
DF	150	65	150	0
GWH	129	9	129	0
DB	45	2	43	2
MRY	146	0	146	0
DFR	65	0	63	2
HOL	50	0	50	0

2.1. Animals added by DNA test

An overview was available of the individuals that have been tested with the DNA test, including the assignment scores for DF, DFR, GWH, DB, DR, MRY and Holstein, and if applicable their assigned breed. The scores of DF and DFR are added together for the breed assignment as they are registered in the same Dutch Friesian studbook with a distinction based on coat colour. The scores of DR and MRY are also added together for the breed assignment of DR, as these breeds were only officially separated in 2004. The genetic similarity between DF and DFR, and between DR and MRY can also be seen in the PCA plots of Hulsegge et al. (2019) of various breeds within the reference population.

Animals are assigned to a breed if their score for the corresponding breed is above the threshold value of 77.5%, set by Hulsegge et al. (2019).. Individuals are said to fit within the purebred population for a specific breed if they score above 77.5% for a specific breed. In this paper borderline cases were also identified with a score between 70.0% and 77.5% for a specific breed, and potential crossbreds with a score above 30% for two different breeds. Categorization of DF and DFR, and MRY and DR was done both for the separate scores and for the combined scores.

2.2. Change in genetic diversity

The complete genotypic datasets of the current reference population and the tested individuals were used after data preparation had been performed (37,123 SNPs). A second dataset was created by selection of the 133 SNP's from the DNA test.

Genetic diversity within and between the breeds of the current reference population, the tested individuals and of the combined population was investigated with the use of F-statistics calculated by the hierfstat (version 0.5-11) package in R (Goudet et al., 2022), both for the complete datasets and for the 133 SNP datasets. The observed heterozygosity (Ho) and within population expected heterozygosity (Hs) were calculated for each breed separately and averaged for the total population. The total expected heterozygosity (Ht), i.e based on the allele frequencies in the whole dataset, and the proportion of Hs relative to Ht (Fst = (Ht-Hs)/Ht) were only calculated for the total population. F-statistics of the datasets will then be compared with each other to gain insight in the level of inbreeding. It is expected that the Fst of the 133 SNP datasets will be higher than the complete datasets as the SNP's are selected based on a high genetic diversity between breeds and a low genetic diversity within the breeds.

2.3. Improvements DNA test

The complete genotypic dataset of all populations after data preparation was used to analyse possible improvements of the DNA test. PCA analysis was performed on the combination of the tested individuals with the current reference population, and on the current reference population on its own to analyse the effect of the tested individuals and the newly added individuals on the reference populations. An ANOVA was used to determine the significance of the differences in PCA scores between the populations and breeds. The ANOVA of the tested individuals and current reference population was only run with the breeds included in the test individuals, and the ANOVA of the original reference population and the newly added individuals was only run with the breeds included in the newly added individuals.

The complete genotypic dataset of the current reference population after data preparation was then used to make an updated SNP selection. This was done according to the same methods used to design the original DNA test (Hulsegge et al., 2019).
First SNP's with high linkage to each other were filtered with LD pruning. Another PCA was performed and compared to the previous PCA to ensure that the SNP's that were left were still representative for the total.
A barplot was made to find the most relevant PC's from which the top 500 SNP's with the highest relevance were selected. From these 1,000 SNP's a top 100 was selected based on the mean decrease gini index (MDG). Next to that, an additional 20 SNP's were selected based on the difference in allele frequency to separate DR and MRY, and DF and DFR.
The selected SNP's were then combined to form the updated SNP selection and were used to form a new dataset of the current reference population of only these SNP's. Next to that, another dataset of the current reference population was made with only the SNP's that were selected for the original DNA test.
On both datasets a Random Forest (Liaw & Wiener, 2002) was performed to assess the quality of classification for both SNP selections. This was done with the R package *randomForest* (version 4.7-1) which gives a classification matrix showing to which breed each individual belongs and to which it is assigned.
After that STRUCTURE (version 2.3.4) (Pritchard et al., 2000) was used for additional analysis of the classification quality. Within STRUCTURE the Ancestry model was used with default settings, a 100,000 burn in period and 100,000 iterations. This was done with 6, 7 and 8 clusters (K) for both de original and updated SNP selection.
Results from Random Forest and STRUCTURE (K = 7) of both SNP selections were then compared to evaluate the original DNA test and determine whether improvement is needed. Next to that, the SNP's selected for the original DNA test and SNP's of the updated selection were compared with each other to determine the overlap between both selections.

3. Results

3.1. Animals added by DNA test

A total of 129 individuals were tested on their purebred status since the DNA test was put into use in 2018 ((Ministerie van Landbouw Natuur en Voedselkwaliteit, 2018)**.** Of the tested individuals, 100 had a breed assignment score above the threshold value of 77.5% for one of the breeds, and were thus given purebred status.
A total of 85 individuals were determined as DFR/DF with 77 individuals scoring above 90%. These individuals mainly scored high for DF with 84 individuals scoring above the threshold value and 74 individuals scoring above 90% based on their score for DF, and none scoring above the threshold value based on their score for DFR (Table 33).
Another 10 individuals were determined as GWH with 3 individuals scoring above 90% (Table 33). DB had 6 individuals assigned with 2 individuals scoring above 90%. Finally, no individuals scored above the threshold value for DR/MRY.

Table 3 Breed assignment scores of all individuals tested by the DNA test since it has been put into use. The black line represents the threshold value of 77.5%. Above this value all breeds are considered a purebred for the corresponding breed.

Score	DFR	DF	GWH	DB	DR	MRY
>0.90	0	74	3	2	0	0
0.80-0.90	0	7	7	2	0	0
0.775-0.80	0	3	0	2	0	0
0.70-0.775	0	3	1	0	0	0
0.50-0.7	1	1	5	1	1	3

A total of 29 individuals scored below the threshold value for all breeds and were thus not given purebred status. Of these individuals, 8 were identified as borderline cases with a score between 0.7 and 0.775, if scores of DFR and DF, and of DR and MRY are combined. Half of these borderline cases can be found under DFR/DF, three under DR/MRY and one under GWH. No borderline cases can be found under DB.

Another six individuals had a score above 30% for two breeds (Table 44). Three of these individuals scored above 30% for DR/MRY and GWH. The other three scored above 30% for DFR/DF and another breed, including one high score for Holstein Friesian.

Table 4 Overview of individuals with a breed assignment score above 30% for two breeds, and for which breeds this was.

Breeds	DFR/DF	GWH	DB	DR/MRY	HOL
DFR/DF					
GWH	0				
DB	1	0			
DR/MRY	1	3	0		
HOL	1	0	0	0	

3.2. Change in genetic diversity

Observed heterozygosity (Ho) and within population heterozygosity (Hs) of the reference population calculated from all SNP's differ slightly from each other for all subpopulations, leading to a small difference in overall Ho and Hs as well with values of 0.3323 and 0.3333 respectively (Table 5). Both these values are lower than the total expected heterozygosity (Ht) which has a value of 0.3641.

An Fst value of 0.0847 can be found, meaning that 8% of the genetic diversity can be found between the different breeds.

Table 5 Based on all SNP's of only the reference population.

	Overall	DB	DF	DFR	DR	GWH	HOL	MRY
Ho	0.3323	0.3355	0.3179	0.3326	0.3413	0.3063	0.3574	0.3351
Hs	0.3333	0.3340	0.3269	0.3351	0.3431	0.3055	0.3552	0.3336
Ht	0.3641							
Fst	0.0847							

The overall Ho and Hs of the reference population calculated from the 133 SNP dataset hardly differ with values of 0.3545 and 0.3548 while a clearly higher Ht can be found with a value of 0.4392 (Table 6). The larger difference between Hs and Ht leads to a higher Fst value of 0.1923 compared to the Fst value based on all SNP's. Both Ho, Hs and Ht are lower when calculated from all SNP's than when calculated from the 133 SNP dataset.
Within the subpopulations only small differences can be found between the Ho and Hs. However, both Ho and Hs are clearly higher for the 133 SNP dataset than for the complete SNP set.

Table 6 Based on 133 SNP's of only the reference population.

	Overall	DB	DF	DFR	DR	GWH	HOL	MRY
Ho	0.3545	0.3871	0.3184	0.3563	0.3744	0.3278	0.3732	0.3446
Hs	0.3548	0.3861	0.3245	0.3641	0.3773	0.3252	0.3684	0.3384
Ht	0.4392							
Fst	0.1923							

The Ho and Hs based on the complete dataset differs slightly overall and per breed for the tested animals (Table 7). The Ht is higher than both Ho and Hs with a value of 0.3513, leading to a Fst of 0.0873.
The difference between Hs and Ht based on the complete dataset for the tested animals is slightly larger than the difference for the reference population, leading to the slightly higher Fst for the tested animals.

Table 7 Based on all SNPs of tested animals that have been approved for one of the breeds

	Overall	DB	DF	GWH
Ho	0.3244	0.3406	0.3148	0.3177
Hs	0.3206	0.3395	0.3133	0.3141
Ht	0.3513			
Fst	0.0873			

The difference between Ho, Hs and Ht is higher when the 133 SNP dataset is used, with values of respectively 0.3644; 0.3508 and 0.4236 (Table 8). The larger difference between Hs and Ht leads to a higher Fst compared to the Fst based on the complete dataset of the tested animals. This shows that for the 133 SNP dataset relatively more genetic diversity can be found between the breeds than within, compared to the complete dataset.

Table 8 Based on 133 SNP's of the tested animals that have been approved for one of the breeds

	Overall	DB	DF	GWH

Ho	0.3644	0.4474	0.3075	0.3383
Hs	0.3508	0.4135	0.3107	0.3388
Ht	0.4236			
Fst	0.1719			

Small differences can be found between the Ho and Hs for the different breeds and for the Ho and Hs overall for the combined population of tested animals and the reference population (Table 9; Table 10). Ht is higher than Hs leading to a high Fst. The Fst of the combined population is slightly higher than for both populations separately when looking at the 133 SNP dataset (Table 6; Table 8; Table 10). When looking at the complete dataset the Fst of the test individuals is shown to be higher than both the combined population and the reference population (Table 5; Table 7; Table 9).
We also see that the Hs of the Dutch Friesians, the breed of which most test individuals have been added, shows a slight increase for the combined population compared to the reference population for the complete dataset (Table 5; table 9).

Table 9 Based on all SNPS of combination of tested and reference population

	Overall	DB	DF	DFR	DR	GWH	HOL	MRY
Ho	0.3323	0.3357	0.3170	0.3326	0.3413	0.3070	0.3574	0.3351
Hs	0.3332	0.3341	0.3250	0.3351	0.3431	0.3063	0.3552	0.3336
Ht	0.3641							
Fst	0.0850							

Table 10 Based on 133 SNP's of combination of tested and reference population

	Overall	DB	DF	DFR	DR	GWH	HOL	MRY
Ho	0.3545	0.3897	0.3151	0.3563	0.3744	0.3285	0.3732	0.3446
Hs	0.3548	0.3875	0.3221	0.3641	0.3773	0.3262	0.3684	0.3384
Ht	0.4393							
Fst	0.1925							

3.3. Improvement DNA test

3.3.1. Evaluation original DNA test

The PCA's of the original reference population and the newly added individuals show a clear distribution of the different breeds and similarities between the two populations (Figure 1). Considering the original reference population, PC 1 shows a separation of DF and DFR on the far left side (range = -19.44;-5.35), GWH on the far ride sight (range = 19.42;36.19) and the DR, HOL, DB and MRY in between (range = -7.80;0.03). PC 2 shows MRY and DR clustered together on the right side (range = 8.3;28.33), with DF, DFR and GWH clustered on the left side (range = -20.23;-4.46), and DB and HOL in the middle (range = 0.59;5.47). Finally, PC 3 shows HOL (range = -33.04;-21.91) and DB (range = -17.88;-5.89) in the lower part of the graph, and MRY, DFR, DR, DF and GWH (range = --0.95;11.02) in the higher part of the graph. All three PC's show overlap for DFR and DF, and for DR and MRY.

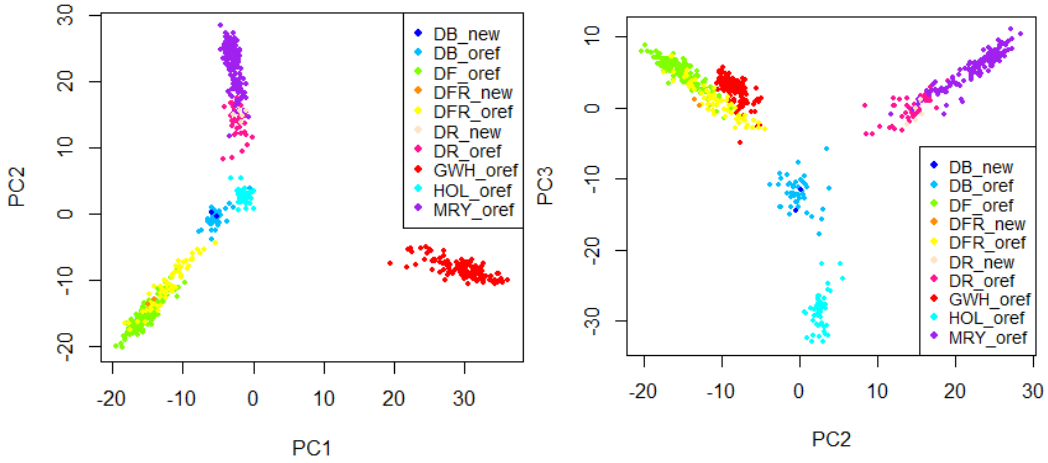


Figure 1 PCA of original reference population (oref) and the newly added individuals (new) forming the current reference population. Based on all SNP's after quality check.

The newly added individuals all fit within the range of the breeds of the original reference population (figure 2), but a significant difference is shown for population for all three PC's when testing with ANOVA. The average of the newly added individuals of DB and DR shows a slight deviation from the average of the original reference population with a maximum of 1.02, and for DFR the population averages deviate more with a maximum of 2.57.
The ANOVA also showed a significant difference for breed for PC 1 and PC 2. Overlap can be found in the range of the three breeds for PC 1, but the averages are clearly different (figure 2). However, no significant difference for breed can be found for PC 3. DR and DFR show both a similar range and average, but both are clearly different from DB (figure 2).
Finally, the ANOVA showed no significant differences for the interaction of breed and population for PC 1, PC 2 and PC 3.

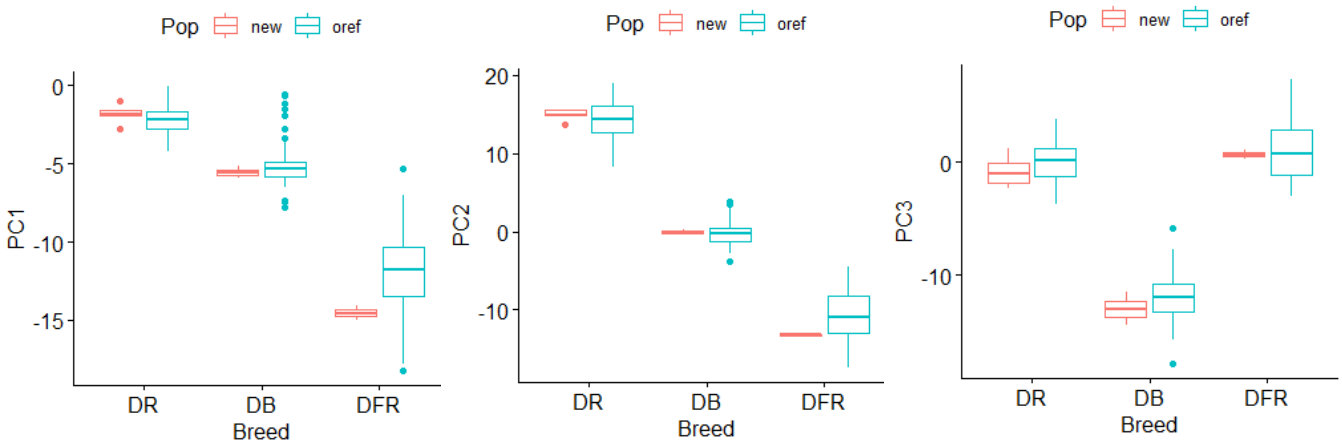


Figure 2 Boxplots of ANOVA comparing DR, DB and DFR for the original reference population (oref) and the newly added individuals.

The PCA's of the current reference population and the tested individuals show a clear distribution of the different breeds and similarities between the two populations (figure 3).

Considering the current reference population, PC 1 shows a separation of DF and DFR on the far right side (range = 3.28;22.79), GWH on the far left sight (range = -31.09;-17.19) and the DR, HOL, DB and MRY in between (range = -8.82;4.96). PC 2 shows MRY and DR clustered together on the left side (range = 10.32;28.53), with DF and DFR (range = -10.66;0.14) clustered on the right side close to GWH (range = -22.01;12.06), and DB and HOL in the middle (range = 1.05;7.92). Finally, PC 3 shows HOL (range = 20.42;30.43) and DB (range = 7.52;19.51) in the higher part of the graph, and MRY, DFR, DR, DF and GWH (range = -12.46;7.04) in the lower part of the graph. All three PC's show overlap for DFR and DF, and for DR and MRY.

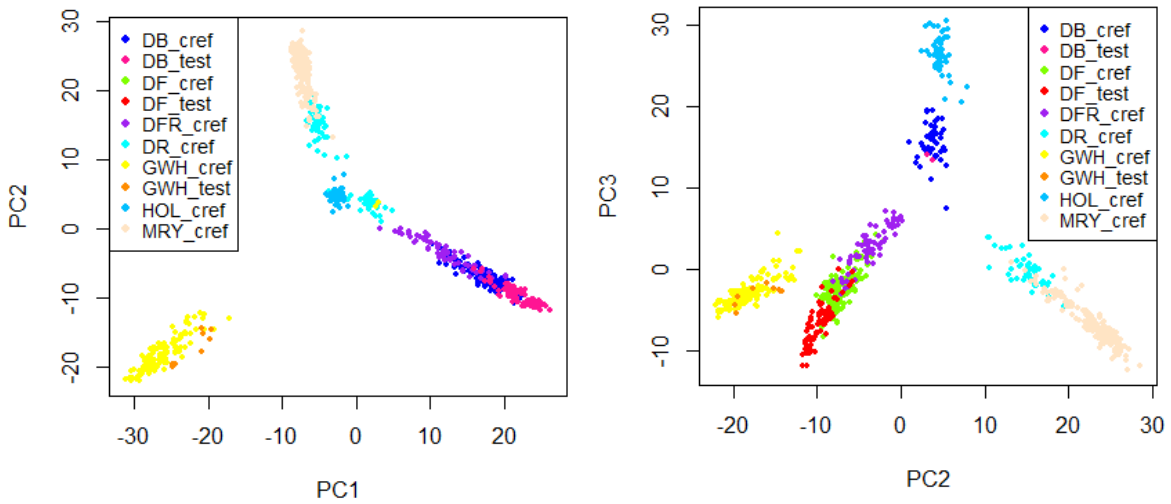


Figure 3 PCA of current reference population (cref) and the tested inidviduals (test). Based on all SNP's after quality check

The tested individuals of GWH and DB all fit within the range of the corresponding breeds of the current reference population (figure 4). The tested individuals of DF mostly fit within the range of the DF of the current reference population, but show an expansion to the higher values for PC 1 and 3, and to the lower values for PC 2. An ANOVA test showed that difference between the two populations is significant for both PC 1 and PC 2, but not for PC 3.
The ANOVA also showed a significant difference for breed for all three PC's. For PC 1 and PC 2 no overlap can be found for the ranges of the different breeds (figure 4). PC 3 does show overlap for GWH and DF, but the averages are different. Interaction is significant for PC 2 and PC 3 but not PC 1.

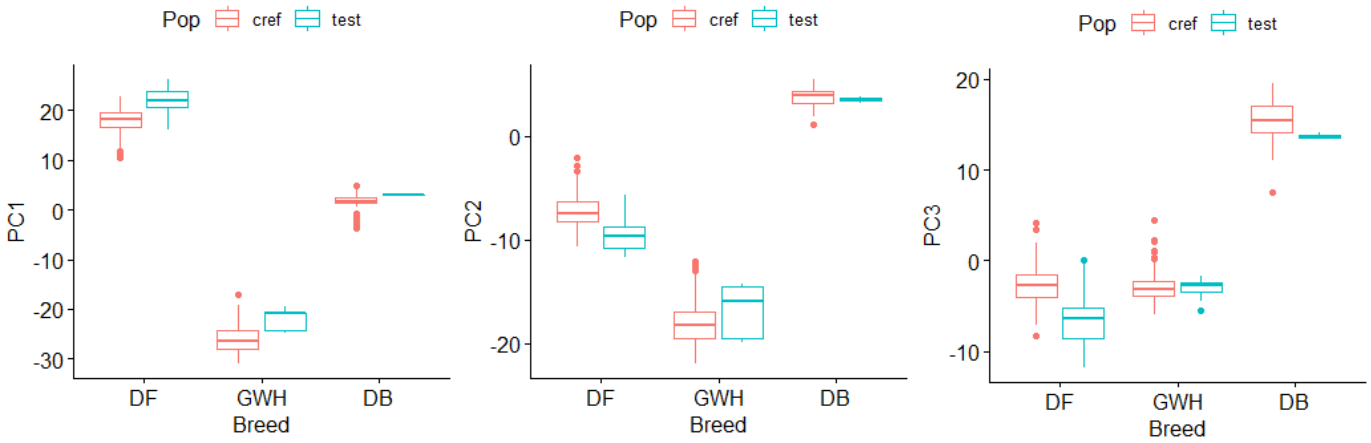


Figure 4 Boxplot of ANOVA comparing DFR, GWH and DB of current reference population (cref) and tested individuals (test).

3.3.2. Updated SNP selection

The new SNP selection was performed on the current reference population (N = 630). The proportion of variance explained by PC 1 and PC 2 was respectively 0.06 and 0.05 (Figure 5). After that, all PC's scored 0.02 or lower.

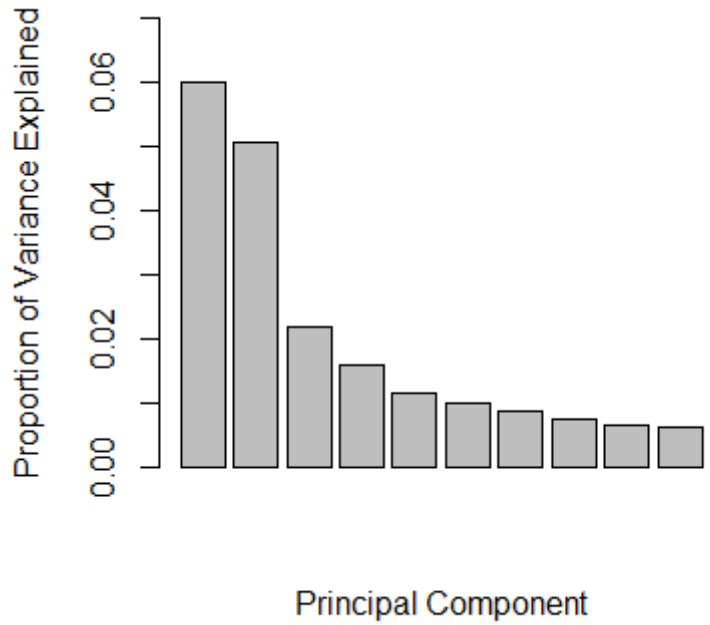


Figure 5 Barplot with proportion of variance explained of the 10 PC's with highest score.

After the top 500 SNP's of PC 1 and PC 2 were combined, Random Forest breed assignment was performed (table 11). The Random Forest shows an error rate below 0.05 for GWH, HOL and MRY. DR has a high error rate as a significant amount is classified as MRY. DFR and DF also have a relatively high error rate due to some DF animals being classified as DFR and vice versa. The DB individuals are a relatively small group and have some misclassifications for several breeds. Overall error rate is 0.09.

Table 11 Random Forest results of 1,000 selected SNP's from the top 500 of PC 1 and PC 2. Rows show the breed an individual belongs to, and columns show the classification of the individual according to the Random Forest. Error rate is calculated per breed. Overall error rate equals 8.57%.

Breed	DB	DF	DFR	DR	GWH	HOL	MRY	Class error
DB	40	1	0	0	2	0	2	0.111
DF	0	141	9	0	0	0	0	0.060
DFR	1	18	46	0	0	0	0	0.292
DR	0	0	0	29	1	0	14	0.356
GWH	0	0	0	0	128	1	0	0.008
HOL	0	1	0	0	0	49	0	0.020
MRY	0	0	1	2	0	0	143	0.021

A top 100 was selected from PC 1 and PC 2, and an additional selection of SNP's to separate DFR and DF, and DR and MRY was added. Overall error rate of 133 SNP's is 0.03, which is lower than the error rate of the 1,000 SNP's. DFR and DR are the only breeds with an error rate above 0.05, but do have a lower error rate than the Random Forest with 1,000 SNP's.

DR and DFR still have an error rate above 0.05 as there are still some misclassifications. However, the amount of misclassification and thus the error rate did decrease for both breeds.

Table 12 Random Forest results of final 132 selected SNP's. Rows show the breed an individual belongs to, and columns show the classification of the individual according to the Random Forest. Error rate is calculated per breed. Overall error rate equals 2.22%.

Breed	DB	DF	DFR	DR	GWH	HOL	MRY	Class error
DB	42	0	0	1	0	1	1	0.067
DF	0	150	0	0	0	0	0	0.000
DFR	0	4	61	0	0	0	0	0.062
DR	0	0	0	39	0	0	6	0.133
GWH	0	0	0	0	129	0	0	0.000
HOL	0	0	0	0	0	50	0	0.000
MRY	0	0	0	1	0	0	145	0.007

STRUCTURE was used to identify the correct number of clusters. DR and MRY were clustered together in the analysis of 6 clusters, but DR is split between cluster 4 and 7 for the analysis with 8 clusters (table 13; figure 6). The analysis with 7 clusters showed one score above 0.85 for each breed, all in separate clusters.
GWH has the highest score with 0.952 and DFR has the lowest score with 0.852 (table 13). The next highest score of DFR is 0.099 in cluster 4, which is the same cluster of DF's highest score. The same goes for DR which scores 0.866 for its highest cluster, and for which the next highest score (0.044 in cluster 1) is in the same cluster as MRY's highest score.

Table 13 Overview of STRUCTURE results of newly selected SNP's for K = 6, K = 7 and K = 8.

6 clusters (0.867)						
Breed	1	2	3	4	5	6
HOL	0.014	**0.951**	0.010	0.006	0.014	0.005
DR	**0.559**	0.156	0.033	0.074	0.149	0.290
DF	0.009	0.012	0.008	0.034	0.020	**0.917**
GWH	0.007	0.012	**0.953**	0.011	0.011	0.006
DB	0.017	0.051	0.008	0.018	**0.896**	0.010
DFR	0.007	0.008	0.009	**0.850**	0.022	0.014
MRY	**0.946**	0.013	0.009	0.010	0.013	0.009

7 clusters (0.906)							
Breed	1	2	3	4	5	6	7
HOL	0.012	0.014	**0.946**	0.005	0.009	0.006	0.008
DR	0.044	0.027	0.019	0.014	0.011	0.019	**0.866**
DF	0.008	0.017	0.011	**0.918**	0.007	0.030	0.008
GWH	0.006	0.010	0.010	0.006	**0.952**	0.010	0.006
DB	0.013	**0.886**	0.046	0.010	0.007	0.019	0.019
DFR	0.006	0.020	0.008	0.099	0.008	**0.852**	0.006
MRY	**0.919**	0.013	0.010	0.008	0.008	0.010	0.032

8 clusters (0.870)

Breed	1	2	3	4	5	6	7	8
HOL	0.015	0.005	0.011	0.011	0.006	0.009	0.008	**0.936**
DR	0.014	0.006	0.036	0.250	0.006	0.006	**0.674**	0.008
DF	0.017	**0.912**	0.006	0.012	0.028	0.007	0.007	0.010
GWH	0.010	0.005	0.005	0.008	0.009	**0.948**	0.006	0.010
DB	**0.871**	0.008	0.009	0.044	0.015	0.006	0.010	0.037
DFR	0.020	0.101	0.005	0.010	**0.843**	0.007	0.006	0.007
MRY	0.012	0.007	**0.903**	0.028	0.009	0.008	0.026	0.009

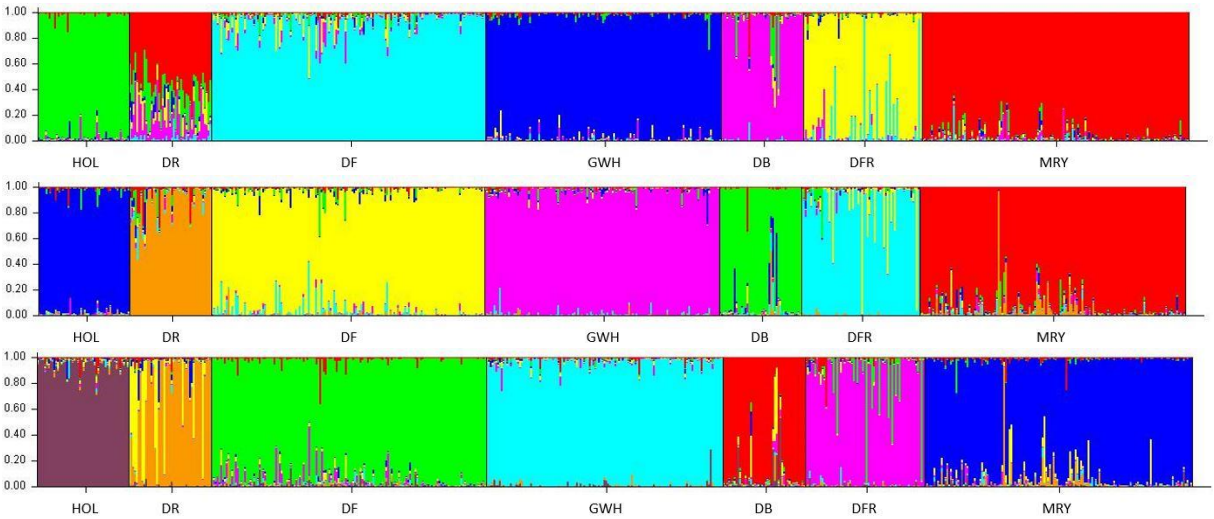


Figure 6 STRUCTURE output of 132 newly selected SNP's for 6, 7 and 8 clusters (from top to bottom).

3.3.3. Comparing original and updated DNA test

Average of highest score of each breeds is slightly higher for the new selection than for the original selection (0.007) (table 14). The biggest differences can be found for the distinction between MRY and DR, and between DFR and DF. The new selection shows a higher score of DFR with a difference of 0.089 and a similar score for DF (difference = 0.008), while the original selection scores higher for both MRY and DR with differences of 0.028 and 0.010.

Table 14 Overview of STRUCTURE results of K = 7 for new and original SNP selection

New selection (0.906)							
Breed	1	2	3	4	5	6	7
HOL	0.012	0.014	**0.946**	0.005	0.009	0.006	0.008
DR	0.044	0.027	0.019	0.014	0.011	0.019	**0.866**
DF	0.008	0.017	0.011	**0.918**	0.007	0.030	0.008
GWH	0.006	0.010	0.010	0.006	**0.952**	0.010	0.006
DB	0.013	**0.886**	0.046	0.010	0.007	0.019	0.019
DFR	0.006	0.020	0.008	0.099	0.008	**0.852**	0.006
MRY	**0.919**	0.013	0.010	0.008	0.008	0.010	0.032

Original selection (0.899)							
Breed	1	2	3	4	5	6	7
HOL	0.014	0.007	0.007	0.006	**0.948**	0.006	0.012
DR	**0.894**	0.016	0.01	0.005	0.022	0.015	0.038
DF	0.01	0.017	0.026	0.005	0.01	**0.926**	0.005
GWH	0.008	0.007	0.011	**0.950**	0.012	0.008	0.005
DB	0.024	**0.882**	0.015	0.01	0.048	0.009	0.011
DFR	0.008	0.009	**0.763**	0.005	0.009	0.197	0.009
MRY	0.038	0.006	0.009	0.005	0.009	0.005	**0.929**

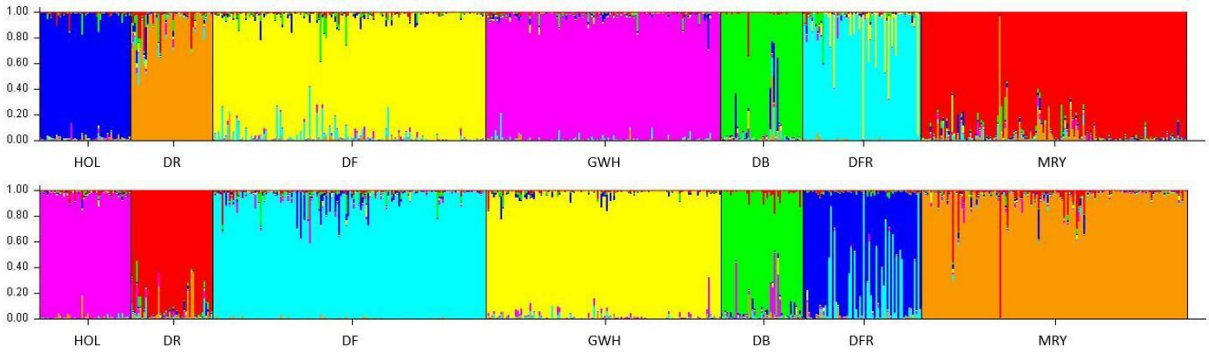


Figure 1 STRUCTURE output for 7 clusters of 132 newly selected SNP's (top) and 133 original selected SNP's (bottom)

From the 132 newly selected SNP's there are 24 that were also selected for the original selection. Most overlap of SNP's between the original selection and the new selection can be found for the SNP selection based on the separation of DFR and DF (35%), then for separating MRY and DR (20%) and least overlap was found for the 100 SNP's selected based on MDG (14%).

Most overlap can be found on chromosome 3 (3 out of 9 SNP's of the original), chromosome 18 (4 out of 5 SNP's of the original), chromosome 25 (3 out of 3 SNP's of the original) (table 15). Chromosome 25 was the only one where are SNP's from the original SNP selection were also included in the new SNP selection.

Table 15 Overview of amount of selected SNP's per BTA for the updated and original SNP selection.

BTA	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Updated	6	4	7	5	5	15	5	1	6	4	6	2	3	1	2
Original	2	7	9	4	12	9	6	1	3	6	7	5	2	1	7

BTA	16	17	18	19	20	21	22	23	24	25	26	27	28	29
Updated	5	2	8	3	4	6	4	7	5	8	1	1	1	5
Original	4	2	5	4	4	6	5	7	4	3	1	4	2	1

4. Discussion

4.1. Effect of the DNA test

Most individuals that have been tested by the original DNA test have been assigned to DF and no animals have been assigned to DR native or MRY, of which MRY is not at risk and excluded from registering animals based on the DNA test (*Verordening (EU) 2016/1012 van Het Europees Parlement En de Raad*, 2016). Though, MRY is included in the DNA-test as a reference population to better distinguish the DR individuals.
Scores of DR and MRY were added together for the breed assignment. Their recent separation made them difficult to separate which could also be seen in the PCA.

The addition of the animals to the studbook, most of all the DF, is likely to be positive. In the last couple of years an average of 1149 breeding females and 155 breeding males have been used within the DF breed (personal communication Jack Windig). Assuming that all individuals have an equal chance of contributing to the next generation this gives an inbreeding rate of 0.0915% per generation ($\Delta F = 1/8N_m + 1/8N_f$) which is far below the maximum of 1.00% allowed by the FAO (Food and Agriculture Organization of the United Nations., 2013). When all 85 DF individuals will be added to the breeding population and that they include an equal number of males and females, and that all individuals have an equal chance of contributing to the next generation, the inbreeding rate will decrease with 0.0177 to 0.0738% ($\Delta F = 1/8N_m + 1/8N_f$). Calculated inbreeding rates are likely different due to individuals not contributing equally in practice and because most added individuals will likely be female.
Effects on other populations will likely be even larger if the same amount of individuals would be added, as DF is one of the larger native breeds. Especially for DR adding 85 individuals could make a huge difference for both population size and, if used for breeding, a decrease in inbreeding and increase in genetic diversity.
When looking at the average inbreeding as calculated by CRV (CRV, 2020) a small decrease can already be seen for DF, but for DR, for example, an overall increase can be seen over the last couple of years. Adding individuals with the DNA test is therefore likely to further decrease the inbreeding rate of DF, but is likely to be of higher importance for DR once individuals will be added by the test.

The difficulty of the DNA test to separate DF from DFR and DR from MRY could also make you question whether these breeds should even be considered separate breeds. In the case of DF and DFR the breeds are even registered within the same breeding programmes and only separated based on coat colour. Especially for DFR, which is the smaller breed considering population size, combining these populations into one might be beneficial. However, combining these population does also increase the chance that unique traits of these breeds disappear.
The same goes for MRY and DR, of which DR has the smaller population size. As mentioned before, the inbreeding rate of DR is currently increasing and no individuals have been added yet by the DNA test. Calculations of CRV, (CRV, 2020) also show an increase in inbreeding for MRY over the last couple of years. Combining DR with MRY could decrease the inbreeding rate for both populations, but is also likely to lead to an overall decrease in genetic diversity. This decrease is likely to be minimal due to the already high genetic similarity but is still something to consider when combining the populations.

The original SNP selection was based on high variation between the breeds, making the breeds easier to separate. In combination with a smaller increase within the breeds this leads to the higher Ht from the F statistics for the 133 SNP dataset than for the complete dataset with all SNP's. The smaller increase in variation within the breeds than between the breeds also leads to the higher Fst for the 133 SNP dataset.
The PCA plots of the test individuals and the current reference population shows proper functioning of the test as the tested individuals mostly fit within the range of the current reference population,

but also add genetic diversity due to some individuals slightly expanding the range and other individuals scoring at the edge of the range of the current reference population.

The addition of the tested individuals increases the total number of individuals in the studbook population. The larger population size may lead to a decrease in inbreeding within the breeds if managed properly in breeding programs. An increase in population size will make the breeds more resilient against, for example, circumstantial measures surrounding infectious disease outbreaks (SZH, 2021). The increased genetic diversity that these tested individuals bring could also have a positive outcome on the resilience of these breeds. A higher genetic diversity makes it easier for the breeds to adapt to possible changes in the environment (Oldenbroek, 2017) which increases in importance due to global warming.

4.2. Updated DNA test

The newly selected 132 SNP's show a good separation of most breeds, but similar to the original selection it has some difficulty in separating DFR and DF, and MRY and DR. The additional selection specifically to separate these breeds showed an overall decrease of the error rate from 8.6 to 2.7 for the new SNP selection, and a specific decrease in error rate of 0.15 and 0.19 respectively for DR and DFR. STRUCTURE showed a slight overall improvement of the new SNP selection due to easier separation of DFR and DF, but showed more difficulty with separating MRY and DR.
The original DNA test adds the scores for DF and DFR together and separates them based on coat colour. This makes the genetic separation of these two breeds less important than the genetic separation of MRY and DR.

The difference between the selected SNP's is likely caused by the LD pruning. LD pruning is a selection on SNP's that are in high or complete linkage disequilibrium and are thus (almost) always appearing together. For each pair of SNP's that is in high or complete linkage disequilibrium the LD pruning removes one of those SNP's. It is possible that during the LD pruning for the original SNP selection other SNP's were removed from each pair that is in linkage disequilibrium than for the new SNP selection, leading to a different selection of SNP's.
These SNP's can usually be found close to each other on the autosome. Some examples of this can be found on BTA 6 which is approximately 1.2 Mb long (Zimin et al., 2009). On this autosome ARS.BFGL.NGS.2525 from the new selection and ARS.BFGL.NGS.97136 from the original selection only differ 95 kb, and ARS.BFGL.BAC.2481 from the new selection and HAPMAP27851.BTC.071689 from the original selection only differ 84 Kb. It is likely that these SNP's are in disequilibrium to a certain extent and can thus likely be found together more often.
Next to that, the addition of new DFR, DR and DB individuals to the reference population is also likely to influence the difference in SNP selection, although to a lesser extent. The PCA showed a slight expansion of the genetic diversity of these breeds, causing possible changes in allele frequencies of SNP's. These changes in allele frequency could then, in turn, decrease or increase the importance of certain SNP's in the distinction of the different breeds leading to a different SNP selection.

4.3. Other selection methods

In this research a combination of a PCA and Random Forest, and difference in allele frequency was used to select the SNP's, and a combination of Random Forest and Structure was used to evaluate the SNP selection. Wilmot et al., (2022) compared a similar combination of PCA and Random Forest to other selection and classification methods for the breeds dual-purpose Belgian Blue, East Belgian Red and White and Red-Pied of Ösling. They showed that accuracy was approximately 84% with 205 SNP's selected. The best classification model showed an accuracy of 99% but had 7153 SNP's selected, and the second best showed an accuracy of 98.5% with 2005 SNP's selected.

The model used in this research already shows an accuracy of 90% for the breeds used in the DNA test, which has shown to be sufficient. Next to that, possible improvement will also lead to a higher amount of SNP's selected which means it will take longer to get results.

4.4. Application on other (native) breeds or species

The DNA test is shown to be able to distinguish the different native Dutch cattle breeds successfully. It increased the population size and genetic diversity of the Dutch Friesians which had the highest number of test individuals score above the threshold. Earlier research also showed that similar methods can be successful in separating native cattle breeds in other countries (Wilmot et al., 2022). Applying a test similar to the DNA test discussed in this report could therefore be useful for cattle breeds in other countries as well. However, it should be taken into account that individuals can only be added to the studbook with the DNA test if the breed is at risk; that enough genotypic information should be available of the breeds included in the DNA test; and that the similarity of the breeds also influences the amount of SNP's needed to distinguish them.
For other native Dutch species that include breeds at risk, a DNA test could have a positive effect on population size and genetic diversity. A similar DNA test could be designed, but efficiency will depend on the available genotypes and SNP's, and current genetic diversity of the breeds within that species.

5. Conclusion

Overall a positive effect can be seen on the population size and genetic diversity of mainly the Dutch Friesians, potentially decreasing their average level of inbreeding and therefor possibly increasing the resilience of the population in the long term. As the DNA test has only been in use for three years, more time might be needed to see similar effects on the other breeds included in the DNA test.

Next to that, the DNA test does not have to be improved at this moment as it is currently still sufficient to distinguish the different breeds and determine the resemblance to the purebred individuals. However, the test should be monitored regularly as improvements may be needed in the future.

6. References

Berg, E. E., & Hamrick, J. L. (1997). *Quantification of genetic diversity at allozyme loci*.

CGN. (2022). *Rassenlijst Nederlandse landbouwhuisdierrassen en hun risicostatus op basis van aantal volwassen vrouwelijke dieren in Nederland*.

CRV. (2020). *Animal Evaluation Unit*.

Ellegren, H., & Galtier, N. (2016). Determinants of genetic diversity. *Nature Reviews Genetics*, *17*(7), 422–433. https://doi.org/10.1038/nrg.2016.58

Food and Agriculture Organization of the United Nations. (2013). *In vivo conservation of animal genetic resources.* 242.

Goudet, J., Jombart, T., Kamvar, Z. N., Archer, E., & Hardy, O. (2022). *Estimation and Tests of Hierarchical F-Statistics*. https://github.com/jgx65/hierfstat

Hulsegge, I., Schoon, M., Windig, J., Neuteboom, M., Hiemstra, S. J., & Schurink, A. (2019). Development of a genetic tool for determining breed purity of cattle. *Livestock Science*, *223*, 60–67. https://doi.org/10.1016/J.LIVSCI.2019.03.002

Liaw, A., & Wiener, M. (2002). *Classification and Regression by randomForest*. *2*(3). http://www.stat.berkeley.edu/

Ministerie van Landbouw Natuur en Voedselkwaliteit. (2018). Regeling van de Minister van Landbouw, Natuur en Voedselkwaliteit van 17 oktober 2018, nr. WJZ/18034451, tot wijziging van de Regeling houders van dieren en enige andere regelingen in verband met de uitvoering van een Europese verordening over fokkerij. *Staatscourant*, *59100*. https://zoek.officielebekendmakingen.nl/stcrt-2018-59100.html

Nei, M. (1973). Analysis of Gene Diversity in Subdivided Populations. *Proceedings of the National Academy of Sciences of the United States of America*, *70*(12 Pt 1-2), 3321. https://doi.org/10.1073/PNAS.70.12.3321

Oldenbroek, K. (2017). Genomic management of animal genetic diversity. In *Wageningen Academic Publishers*. Wageningen Academic Publishers. https://doi.org/10.3920/978-90-8686-850-6

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). *Inference of Population Structure Using Multilocus Genotype Data*. http://www.stats.ox.ac.uk/pritch/home.html.

SZH. (2021). *Stichting Zeldzame Huisdierrassen*. https://szh.nl

Verordening (EU) 2016/1012 van het Europees Parlement en de Raad. (2016). https://www.rvo.nl/sites/default/files/2017/06/Europese-Fokkerijverordening-Verordening-EU-2016-1012.pdf

Wilmot, H., Bormann, J., Soyeurt, H., Hubin, X., Glorieux, G., Mayeres, P., Bertozzi, C., & Gengler, N. (2022). Development of a genomic tool for breed assignment by comparison of different classification models: Application to three local cattle breeds. *Journal of Animal Breeding and Genetics*, *139*(1), 40–61. https://doi.org/10.1111/JBG.12643

Zimin, A. V., Delcher, A. L., Florea, L., Kelley, D. R., Schatz, M. C., Puiu, D., Hanrahan, F., Pertea, G., Van Tassell, C. P., Sonstegard, T. S., Marçais, G., Roberts, M., Subramanian, P., Yorke, J. A., & Salzberg, S. L. (2009). A whole-genome assembly of the domestic cow, Bos taurus. *Genome Biology*, *10*(4), R42. https://doi.org/10.1186/GB-2009-10-4-R42

7. Appendix

7.1. Selected SNP's

7.1.1. Original SNP selection

BTA Bos Taurus Autosome
Position Physical chromosome position
SNP SNP name

SNP's overlapping with the updated selection are indicated in bold with a light green background colour.

BTA	Position	SNP
1	87829634	BTB.01858021
1	89431170	HAPMAP27729.BTA.108726
2	**34015788**	**ARS.BFGL.NGS.65913**
2	46268964	HAPMAP50824.BTA.92238
2	52176845	ARS.BFGL.NGS.118744
2	64300614	HAPMAP47221.BTA.22428
2	70559853	HAPMAP49404.BTA.100549
2	107747873	ARS.BFGL.NGS.19219
2	136305920	ARS.BFGL.NGS.118957
3	65684294	BTB.01580562
3	66402051	BTA.86788.NO.RS
3	69123629	BTB.00133827
3	77975986	BTB.00137287
3	84195302	ARS.BFGL.NGS.44176
3	**85429665**	**BTB.01463885**
3	**87451543**	**HAPMAP57348.RS29011302**
3	**102810330**	**ARS.BFGL.NGS.102829**
3	116801749	HAPMAP48857.BTA.69699
4	5262797	ARS.BFGL.NGS.26036
4	8943433	ARS.BFGL.NGS.70391
4	20915847	BTB.01054295
4	48945677	BTB.00181232
5	9157214	HAPMAP48875.BTA.74500
5	32983741	ARS.BFGL.NGS.12553
5	48080259	BTA.11044.RS29016809
5	**59665562**	**HAPMAP40158.BTA.104729**
5	77780682	ARS.BFGL.NGS.106744
5	86502735	ARS.BFGL.NGS.111850
5	87389613	ARS.BFGL.NGS.85748
5	91262694	BTA.111865.NO.RS
5	91947526	ARS.BFGL.NGS.111260
5	98338510	ARS.BFGL.NGS.110933
5	103821233	HAPMAP49735.BTA.74776
5	106538208	ARS.BFGL.NGS.116906
6	8104004	ARS.BFGL.NGS.105221

6	38746212	HAPMAP43470.BTA.114677
6	43713187	ARS.BFGL.NGS.58762
6	45909053	ARS.BFGL.NGS.97136
6	**46970861**	**ARS.BFGL.NGS.53753**
6	53159092	HAPMAP27851.BTC.071689
6	63071799	ARS.BFGL.NGS.118302
6	73057109	HAPMAP53308.RS29024839
6	119057557	ARS.BFGL.NGS.90043
7	56766754	ARS.BFGL.NGS.116933
7	68436319	BTA.88775.NO.RS
7	81492059	BTB.00322735
7	**84145466**	**BTB.01865608**
7	92201609	BTA.20841.NO.RS
7	109139835	ARS.BFGL.NGS.109842
8	828888	ARS.BFGL.NGS.38053
9	12707144	ARS.BFGL.NGS.59162
9	30638895	BTB.01708014
9	82781711	HAPMAP40603.BTA.84409
10	29028329	BTB.00415635
10	63166747	ARS.BFGL.NGS.34398
10	**72658520**	**BTA.109792.NO.RS**
10	**91258020**	**ARS.BFGL.NGS.25471**
10	95923129	BTB.01147175
10	97372769	ARS.BFGL.NGS.116407
11	33886580	BTB.01683514
11	45851216	ARS.BFGL.NGS.21435
11	56617997	ARS.BFGL.NGS.108438
11	70201515	HAPMAP41131.BTA.103178
11	71090191	BTA.07322.NO.RS
11	75352649	ARS.BFGL.NGS.119401
11	101602103	ARS.BFGL.NGS.2573
12	**18584913**	**ARS.BFGL.NGS.5229**
12	**67249579**	**BTB.00502546**
12	83225374	ARS.BFGL.NGS.105587
12	86077889	HAPMAP41824.BTA.31393
12	90015695	ARS.BFGL.NGS.106663
13	17452477	HAPMAP32400.BTA.31937
13	80239131	ARS.BFGL.NGS.36793
14	9606819	HAPMAP31564.BTC.007633
15	13416704	HAPMAP59263.RS29021295
15	32637662	ARS.BFGL.NGS.57210
15	35990313	ARS.BFGL.NGS.21827
15	37338267	BTB.00483457
15	53801075	HAPMAP43128.BTA.105550
15	59020999	BTB.00607669
15	74269138	UA.IFASA.8983

16	**17304547**	**HAPMAP51075.BTA.86900**
16	19340691	ARS.BFGL.NGS.41932
16	59271825	ARS.BFGL.NGS.24236
16	63048628	BTA.39328.NO.RS
17	29612223	ARS.BFGL.NGS.44538
17	32377757	ARS.BFGL.NGS.74608
18	10117349	HAPMAP39226.BTA.42295
18	**11245392**	**ARS.BFGL.NGS.56801**
18	**11836226**	**ARS.BFGL.NGS.14524**
18	**15225943**	**UA.IFASA.6050**
18	**16654560**	**BTB.00703990**
19	16649459	HAPMAP53971.RS29017666
19	35735450	ARS.BFGL.NGS.25445
19	55174260	ARS.USMARC.PARENT.EF164803.RS29011141
19	62390766	UA.IFASA.9390
20	770272	HAPMAP48367.BTA.50266
20	15478658	UA.IFASA.9433
20	**16044811**	**BTA.86837.NO.RS**
20	19187276	BTB.00775883
21	15404665	ARS.BFGL.NGS.100506
21	35184852	ARS.BFGL.NGS.81131
21	44907303	ARS.BFGL.NGS.118121
21	55712777	ARS.BFGL.NGS.60795
21	59081896	ARS.BFGL.NGS.34766
21	63236456	HAPMAP46949.BTA.119904
22	10998835	BTB.01184669
22	28161068	ARS.BFGL.NGS.35996
22	**31841994**	**HAPMAP40441.BTA.54131**
22	35442308	ARS.BFGL.NGS.103404
22	59612919	ARS.BFGL.NGS.24884
23	20215611	ARS.BFGL.NGS.113880
23	**22080127**	**HAPMAP39064.BTA.55787**
23	22848975	ARS.BFGL.BAC.30741
23	**30039692**	**HAPMAP36280.SCAFFOLD155216_10397**
23	34516184	ARS.BFGL.NGS.113689
23	45586528	BTA.107518.NO.RS
23	49260004	HAPMAP49546.BTA.25249
24	2186221	BTB.01448403
24	11888700	ARS.BFGL.NGS.31355
24	27381101	HAPMAP60145.RS29013637
24	55570429	ARS.BFGL.NGS.26447
25	**6239475**	**HAPMAP30117.BTC.019731**
25	**11885927**	**HAPMAP25405.BTC.026733**
25	**15459513**	**ARS.BFGL.NGS.67659**
26	6692455	ARS.BFGL.NGS.98526
27	7561409	UA.IFASA.1517

27	16926528	ARS.BFGL.NGS.113025
27	22100455	ARS.BFGL.NGS.20982
27	31000749	ARS.BFGL.NGS.107550
28	**16783056**	**ARS.BFGL.NGS.115739**
28	24175313	ARS.BFGL.NGS.22483
29	20706709	ARS.BFGL.NGS.104458

7.1.2. Updated SNP selection

BTA Bos Taurus Autosome
Position Physicial chromosome position
SNP SNP name

SNP's overlapping with the original selection are indicated in bold with a light green background colour.

BTA	Position	SNP
1	8837296	HAPMAP59566.RS29024165
1	96791940	ARS.BFGL.NGS.14357
1	109521534	ARS.BFGL.NGS.66914
1	131227520	ARS.BFGL.NGS.6392
1	143831554	ARS.BFGL.NGS.110098
1	145971743	ARS.BFGL.NGS.100302
2	12138830	HAPMAP44546.BTA.48673
2	**34015788**	**ARS.BFGL.NGS.65913**
2	87133202	ARS.BFGL.NGS.100636
2	131041728	ARS.BFGL.NGS.54356
3	20501858	ARS.BFGL.NGS.36889
3	72519744	ARS.BFGL.NGS.67327
3	83242603	ARS.BFGL.NGS.89351
3	**85429665**	**BTB.01463885**
3	**87451543**	**HAPMAP57348.RS29011302**
3	**102810330**	**ARS.BFGL.NGS.102829**
3	118167062	HAPMAP46717.BTA.69803
4	2239242	ARS.BFGL.NGS.108486
4	7338370	BTB.00685072
4	19132537	BTB.01238546
4	106831111	ARS.BFGL.NGS.69855
4	112317788	BTB.01755781
5	10862100	ARS.BFGL.NGS.117126
5	19458651	HAPMAP42303.BTA.72846
5	42605932	BTA.119068.NO.RS
5	**59665562**	**HAPMAP40158.BTA.104729**
5	72390588	ARS.BFGL.NGS.18105
6	8810726	BTA.76544.NO.RS
6	12169113	ARS.BFGL.NGS.27865
6	13502824	BTB.00244579

6	14449728	BTB.00244629
6	23617252	ARS.BFGL.NGS.114096
6	42294971	BTA.75961.NO.RS
6	43811863	BTB.00253074
6	45813845	ARS.BFGL.NGS.2525
6	46556246	ARS.BFGL.NGS.31525
6	**46970861**	**ARS.BFGL.NGS.53753**
6	53242994	ARS.BFGL.BAC.2481
6	71073768	ARS.BFGL.NGS.30652
6	84968944	BTA.121763.NO.RS
6	89001383	BTA.68275.NO.RS
6	116306266	HAPMAP60400.RS29014035
7	16882597	ARS.BFGL.NGS.22940
7	72264142	ARS.BFGL.NGS.95195
7	**84145466**	**BTB.01865608**
7	107014408	BTB.00328969
7	109210429	ARS.BFGL.NGS.43318
8	57619690	ARS.BFGL.NGS.108096
9	11526739	ARS.BFGL.NGS.105758
9	31669195	BTA.93494.NO.RS
9	32049723	ARS.BFGL.NGS.106238
9	76776811	ARS.BFGL.NGS.20794
9	98784166	ARS.BFGL.NGS.83787
9	103755777	ARS.BFGL.NGS.2859
10	45069536	HAPMAP42289.BTA.67247
10	**72658520**	**BTA.109792.NO.RS**
10	86242499	ARS.BFGL.NGS.32234
10	**91258020**	**ARS.BFGL.NGS.25471**
11	4946563	ARS.BFGL.NGS.12127
11	55809281	HAPMAP42798.BTA.109203
11	64269224	ARS.BFGL.NGS.119320
11	71251616	ARS.BFGL.NGS.20385
11	78545016	HAPMAP41140.BTA.107332
11	100901534	ARS.BFGL.NGS.116589
12	**18584913**	**ARS.BFGL.NGS.5229**
12	**67249579**	**BTB.00502546**
13	3143331	HAPMAP57657.RS29025501
13	16466935	BTB.00515019
13	61426148	ARS.BFGL.NGS.83014
14	10812434	HAPMAP54453.RS29015999
15	30076410	ARS.BFGL.NGS.86583
15	38852264	ARS.BFGL.NGS.111723
16	13542437	ARS.BFGL.NGS.103869
16	**17304547**	**HAPMAP51075.BTA.86900**
16	58543206	ARS.BFGL.NGS.1152
16	68981703	BTA.39797.NO.RS

16	73711818	BTB.00660988
17	5987292	ARS.BFGL.NGS.39467
17	22713714	ARS.BFGL.BAC.35968
18	9403168	BTA.119338.NO.RS
18	**11245392**	**ARS.BFGL.NGS.56801**
18	11774889	ARS.BFGL.NGS.102077
18	**11836226**	**ARS.BFGL.NGS.14524**
18	**15225943**	**UA.IFASA.6050**
18	**16654560**	**BTB.00703990**
18	42856641	HAPMAP59019.RS29021918
18	43165294	ARS.BFGL.NGS.113068
19	22157176	ARS.BFGL.NGS.105622
19	55646408	BTB.00761187
19	58226592	ARS.BFGL.NGS.15995
20	5539268	BTB.01104181
20	**16044811**	**BTA.86837.NO.RS**
20	32878037	ARS.BFGL.NGS.89478
20	39538676	HAPMAP42401.BTA.102906
21	5335492	UA.IFASA.4137
21	14304199	HAPMAP30803.BTA.135661
21	15833866	BTB.00804375
21	54186710	BTB.01046586
21	67904403	HAPMAP59281.RS29027629
21	69490730	ARS.BFGL.NGS.111074
22	2069272	HAPMAP32015.BTA.110997
22	26206868	BTB.01464126
22	**31841994**	**HAPMAP40441.BTA.54131**
22	46211533	ARS.USMARC.PARENT.EF093510.RS29010035
23	3544538	ARS.BFGL.NGS.85750
23	11692615	ARS.BFGL.NGS.68466
23	**22080127**	**HAPMAP39064.BTA.55787**
23	**30039692**	**HAPMAP36280.SCAFFOLD155216_10397**
23	30987898	ARS.BFGL.NGS.37027
23	46373705	BTB.00867928
23	49476179	ARS.BFGL.NGS.39104
24	8071750	BTA.58996.NO.RS
24	19698377	HAPMAP46805.BTA.38898
24	45300946	ARS.BFGL.NGS.87679
24	58125298	HAPMAP60813.RS29010178
24	59703255	ARS.BFGL.NGS.94293
25	1086505	ARS.BFGL.NGS.61709
25	3148958	HAPMAP24934.BTC.017367
25	**6239475**	**HAPMAP30117.BTC.019731**
25	9258986	BTA.111433.NO.RS
25	**11885927**	**HAPMAP25405.BTC.026733**
25	**15459513**	**ARS.BFGL.NGS.67659**

25	40862647	ARS.BFGL.NGS.112894
25	41274779	ARS.BFGL.NGS.22044
26	39390959	BTA.93527.NO.RS
27	26142340	ARS.BFGL.NGS.110867
28	**16783056**	**ARS.BFGL.NGS.115739**
29	724198	HAPMAP58534.RS29018685
29	12661834	ARS.BFGL.NGS.113634
29	34531698	ARS.BFGL.NGS.16773
29	40695267	ARS.BFGL.NGS.86495
29	43269744	ARS.BFGL.NGS.5027

7.2. Overview of PC values

7.2.1. Current reference population

PC 1	DR	DF	GWH	HOL	DB	MRY	DFR
Min	-6.6	10.41	-31.09	-3.96	-3.61	-8.82	3.28
Max	-1.16	22.79	-17.19	-1	4.96	-3.19	20.39
Mean	-5.11	17.95	-26.23	-2.51	1.46	-7.14	12.13
Variance	1.02	6.26	6.72	0.37	3.55	0.67	14.57

PC 2	DR	DF	GWH	HOL	DB	MRY	DFR
Min	10.32	-10.66	-22.01	2.53	1.05	13.18	-8.14
Max	19.45	-2.06	-12.06	7.92	5.45	28.53	0.14
Mean	15.05	-7.25	-18.05	4.73	3.8	22.8	-3.54
Variance	4.42	2.2	5.38	0.85	0.97	9.16	4.58

PC 3	DR	DF	GWH	HOL	DB	MRY	DFR
Min	-4.54	-8.31	-5.97	20.42	7.52	-12.46	-2.98
Max	3.91	4.21	4.39	30.43	19.51	0.8	7.04
Mean	-0.3	-2.69	-2.89	26.5	15.47	-6.72	2.75
Variance	3.58	4.68	2.76	5.05	5.34	7.26	6.55

7.2.2. Comparison of current reference population and test individuals

DF	PC 1 cref	PC 1 test	PC 2 cref	PC 2 test	PC 3 cref	PC 3 test
Min	10.41	16.04	-10.66	-11.66	-8.31	-11.88
Max	22.79	26.24	-2.06	-5.73	4.21	-0.01
Mean	17.95	21.93	-7.25	-9.61	-2.69	-6.59
Variance	6.26	5.55	2.2	2.09	4.68	7.17
GWH	**PC 1 cref**	**PC 1 test**	**PC 2 cref**	**PC 2 test**	**PC 3 cref**	**PC 3 test**
Min	-31.09	-24.84	-22.01	-19.86	-5.97	-5.49
Max	-17.19	-19.6	-12.06	-14.25	4.39	-1.69
Mean	-26.23	-21.85	-18.05	-16.75	-2.89	-3.12
Variance	6.72	4.65	5.38	5.81	2.76	1.4
DB	**PC 1 cref**	**PC 1 test**	**PC 2 cref**	**PC 2 test**	**PC 3 cref**	**PC 3 test**

Min	-3.61	2.78	1.05	3.22	7.52	13.37
Max	4.96	3.12	5.45	3.75	19.51	14.05
Mean	1.46	2.95	3.8	3.49	15.47	13.71
Variance	3.55	0.06	0.97	0.14	5.34	0.23

7.2.3. Original reference population

PC 1	DR	DF	GWH	HOL	DB	MRY	DFR
Min	-4.23	-19.44	19.42	-3.14	-7.80	-4.71	-18.26
Max	-0.08	-9.64	36.19	0.03	-0.54	-0.66	-5.35
Mean	-2.21	-15.18	29.96	-1.3	-4.93	-2.87	-11.98
Variance	0.75	3.3	10.91	0.46	2.82	0.62	7.07
PC 2	**DR**	**DF**	**GWH**	**HOL**	**DB**	**MRY**	**DFR**
Min	8.3	-20.23	-10.58	0.59	-3.85	11.69	-17.46
Max	18.93	-8.67	-4.95	5.47	3.76	28.33	-4.46
Mean	14.13	-15.19	-8.41	2.64	-0.24	22.61	-10.68
Variance	5.88	4.65	1.7	0.82	2.82	10.26	10.18
PC 3	**DR**	**DF**	**GWH**	**HOL**	**DB**	**MRY**	**DFR**
Min	-3.71	-1.44	-4.85	-33.04	-17.88	-0.95	-2.99
Max	3.78	8.8	5.68	-21.91	-5.89	11.02	7.29
Mean	-0.17	5	2.69	-28.62	-11.97	5.66	0.92
Variance	3.23	3.34	2.74	6.22	4.86	6.13	6.81

7.2.4. Comparison of newly added individuals and original reference population

DR	PC 1 cref	PC 1 test	PC 2 cref	PC 2 test	PC 3 cref	PC 3 test
Min	-4.23	-2.8	8.3	13.75	-3.71	-2.29
Max	-0.08	-1.00	18.93	15.59	3.78	1.14
Mean	-2.21	-1.84	14.13	14.92	-0.17	-0.82
Variance	0.75	0.42	5.88	0.54	3.23	1.88
DB	**PC 1 cref**	**PC 1 test**	**PC 2 cref**	**PC 2 test**	**PC 3 cref**	**PC 3 test**
Min	-7.8	-5.95	-3.85	-0.45	-17.88	-14.44
Max	-0.54	-5.19	3.76	0.21	-5.89	-11.53
Mean	-4.93	-5.57	-0.24	-0.12	-11.97	-12.99
Variance	2.82	0.29	2.82	0.22	4.86	4.25
DFR	**PC 1 cref**	**PC 1 test**	**PC 2 cref**	**PC 2 test**	**PC 3 cref**	**PC 3 test**
Min	-18.26	-15	-17.46	-13.56	-2.99	0.24
Max	-5.35	-14.09	-4.46	-12.9	7.29	1.08
Mean	-11.98	-14.55	-10.68	-13.23	0.92	0.66
Variance	7.07	0.41	10.18	0.22	6.81	0.35

7.3. ANOVA

```
> summary(aov1.cref.test)
             Df Sum Sq Mean Sq   F value Pr(>F)
Pop           1  20260   20260  3286.747 <2e-16 ***
Breed         2 149783   74891 12149.436 <2e-16 ***
Pop:Breed     2     13       7     1.058  0.348
Residuals   394   2429       6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> aov2.cref.test <- aov(PC2 ~ Pop + Breed + Pop*Breed, data = mydata.cref.test)
> summary(aov2.cref.test)
             Df Sum Sq Mean Sq  F value   Pr(>F)
Pop           1      3       3    0.868    0.352
Breed         2  18851    9426 2957.264  < 2e-16 ***
Pop:Breed     2     87      43   13.590 1.96e-06 ***
Residuals   394   1256       3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> aov3.cref.test <- aov(PC3 ~ Pop + Breed + Pop*Breed, data = mydata.cref.test)
> summary(aov3.cref.test)
             Df Sum Sq Mean Sq  F value   Pr(>F)
Pop           1   1763    1763  381.659  < 2e-16 ***
Breed         2  15097    7548 1634.066  < 2e-16 ***
Pop:Breed     2     80      40    8.707 0.000199 ***
Residuals   394   1820       5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> summary(aov1.oref.new)
             Df Sum Sq Mean Sq F value Pr(>F)
Pop           1     23    22.7   5.176 0.0243 *
Breed         2   3383  1691.6 386.198 <2e-16 ***
Pop:Breed     2     13     6.4   1.466 0.2342
Residuals   149    653     4.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> aov2.oref.new <- aov(PC2 ~ Pop + Breed + Pop*Breed, data = mydata.oref.new)
> summary(aov2.oref.new)
             Df Sum Sq Mean Sq  F value   Pr(>F)
Pop           1    311     311   49.038 8.02e-11 ***
Breed         2  15994    7997 1259.858  < 2e-16 ***
Pop:Breed     2     10       5    0.825     0.44
Residuals   149    946       6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> aov3.oref.new <- aov(PC3 ~ Pop + Breed + Pop*Breed, data = mydata.oref.new)
> summary(aov3.oref.new)
             Df Sum Sq Mean Sq F value Pr(>F)
Pop           1      1     0.8   0.136  0.713
Breed         2   5976  2987.8 529.221 <2e-16 ***
Pop:Breed     2      3     1.5   0.271  0.763
Residuals   149    841     5.6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

7.4. Data management plan

Data management plan belonging to the MSc thesis performed at the Animal Breeding and Genomics Group by Ineke Koning completed in August 2022.

Agreements

1. The data used in this thesis project have been described in this document and have been stored in a systematic manner (at least in separate folders for all sections as described below). Data includes all data as mentioned in the results section of your report.
2. The data management plan has been discussed with the MSc thesis supervisor and he/she has agreed on the location for data storage.
3. In case of confidentiality, contact details of the responsible person from the company/institution that has ownership of the data are mentioned in this document.
4. **The data can be found through Jack Windig (jack.windig@wur.nl)**

Section A - Raw data

File names	Received from	On date
20210622 – overzicht uitslagen.xlsx	Mira Schoon	20-12-2021
Bijlage – geselecteerde 133 SNPs.xlsx	Mira Schoon	20-12-2021
genoACGT_fer_new2.txt	Mira Schoon	28-12-2021
genoACGT_tested_animals.txt	Mira Schoon	28-12-2021
ID_CODE_ALL_20210622_V5.xslx	Mira Schoon	20-12-2021
REF_635samples_133snp_V5_20210622.txt	Mira Schoon	20-12-2021
Snp55map.txt	Mira Schoon	23-2-2021

Comments: These were the raw data files that I received for my analyses. The files include results of the DNA test, selected SNP's of the original DNA test, genotypes of all individuals, ID code of the individuals and an overview of all SNP's including BTA and position.

Section B – Data analysis (e.g. script files)

Mention here the (script) files you used for the analysis of your data

File names	Created in (month, year)	Remarks
R Analysis – DNA test Dutch Cattle.R	August 2022	R script for all analyses and making plots and tables. Includes comments explaining the different steps.
Beagle.bat	January 2022	Imputes missing SNP of individuals. Script received from Ina Hulsegge.

Section C – Final data

All data files that were used for the Results section of your report.

File names	Created in (month, year)	Remarks
Figure 1.1 (PCA 1,2 cref).png	May 2022	First part of Figure 1: PCA of PC 1 and PC 2 of current reference population.
Figure 1.2 (PCA 2,3 cref).png	May 2022	Second part of Figure 1: PCA of PC 2 and PC 3 of current reference population.

Figure 2.1 (ANOVA 1 cref).png	May 2022	First part of Figure 2: ANOVA of PC 1 of current reference population.
Figure 2.2 (ANOVA 2 cref).png	May 2022	Second part of Figure 2: ANOVA of PC 2 of current reference population.
Figure 2.3 (ANOVA 3 cref).png	May 2022	Third part of Figure 2: ANOVA of PC 3 of current reference population.
Figure 3.1 (PCA 1,2 cref+test).png	May 2022	First part of Figure 3: PCA of PC 1 and PC 2 of current reference population and test individuals combined.
Figure 3.2 (PCA 2,3 cref+test).png	May 2022	Second part of Figure 3: PCA of PC 2 and PC 3 of current reference population and test individuals combined.
Figure 4.1 (ANOVA 1 cref+test).png	May 2022	First part of Figure 4: ANOVA of PC 1 of current reference population.
Figure 4.2 (ANOVA 2 cref+test).png	May 2022	Second part of Figure 4: ANOVA of PC 2 of current reference population.
Figure 4.3 (ANOVA 3 cref+test).png	May 2022	Third part of Figure 4: ANOVA of PC 3 of current reference population.
Figure 5 (barplot PCA).png	May 2022	Figure 5: Barplot of first 10 PC's.
Figure 6.1 (132 SNP K = 6).jpg	May 2022	First part of Figure 6: STRUCTURE results of new SNP selection, K = 6
Figure 6.2 & 7.1 (132 SNP K = 7).jpg	May 2022	Second part of Figure 6 & first part of Figure 7: STRUCTURE results of new SNP selection, K = 7
Figure 6.3 (132 SNP K = 8).jpg	May 2022	Third part of Figure 6: STRUCTURE results of new SNP selection, K = 8
Figure 7.2 (133 SNP K = 7).jpg	May 2022	Second part of Figure 7: STRUCTURE results of original SNP selection, K = 7
Tables used in 1. And 2. (Introduction and M&M).xlsx	August 2022	Tables used in the Introduction (1.) and Material & Methods (2.): tables 1-2
Tables used in 3.1. (Results).xlsx	August 2022	Tables used in the first part of the Results (3.1.): tables 3-4
Tables used in 3.2. (Results).xlsx	August 2022	Tables used in second part of the Results (3.2.): tables 5-10
Tables used in 3.3. (Results).xlsx	August 2022	Tables used in third part of the Results (3.3.): tables 11-15
Tables used in 7.1. (Appendix).xlsx	August 2022	Tables used in the first part of the Appendix (7.1.)
Tables used in 7.2 (Appendix).xlsx	August 2022	Tables used in the second part of the Appendix (7.2.)
Figure used in 7.3. (ANOVA of cref+test) (Appendix).png	May 2022	Figure used in the third part of the Appendix: ANOVA results of the PC's of the current reference population and tested individuals.
Figure used in 7.3. (ANOVA of oref+new) (Appendix).png	May 2022	Figure used in the third part of the Appendix: ANOVA results of the PC's of the original reference population and the newly added individuals.