Genomics applied to conservation of genetic diversity in Dutch livestock

Ina Hulsegge

Propositions

1. Without genomic information gene banks are of limited value.
   (this thesis)

2. Genomics makes non herd book animals valuable for conservation.
   (this thesis)

3. Machine learning stops learning.

4. Colour is an indicator of meat quality, but it does not guarantee it.

5. Slow learning, it´s time well spent.

6. Dialect diversity is worth preserving.

Propositions belonging to the thesis, entitled

Genomics applied to conservation of genetic diversity in Dutch livestock

Ina Hulsegge
Wageningen, 27 September 2023

Genomics applied to conservation of genetic diversity in Dutch livestock

Ina Hulsegge

Thesis committee

Promotor
Prof. Dr R.F. Veerkamp
Special professor of Numerical Genetics
Wageningen University & Research

Co-promotors
Dr J.J. Windig
Researcher, Animal Breeding and Genomics
Wageningen University & Research

Dr A.C. Bouwman
Researcher, Animal Breeding and Genomics
Wageningen University & Research

Other members
Prof. Dr B.J. Zwaan, Wageningen University & Research
Prof. T.H.E. Meuwissen, Norwegian University of Life Sciences, Ås, Norway
Prof. Dr D. Hinrichs, Universität Kassel, Germany
Dr J. Peippo, Nordic Genetic Resource Center, Alnarp, Sweden

This research was conducted under the auspices of the Graduate School of Wageningen Institute of Animal Sciences (WIAS).

Genomics applied to conservation of genetic diversity in Dutch livestock

Ina Hulsegge

Thesis

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Wednesday 27 September 2023
at 4. p.m. in the Omnia Auditorium.

Ina Hulsegge
Genomics applied to conservation of genetic diversity in Dutch livestock
161 pages

PhD thesis, Wageningen University, Wageningen, the Netherlands (2023)
With references, with summary in English

ISBN: 978-94-6447-765-8
DOI: https://doi.org/10.18174/633578

Abstract

Hulsegge, I. (2023). Genomics applied to conservation of genetic diversity in Dutch livestock. PhD thesis, Wageningen University, the Netherlands

Conserving genetic diversity is essential for the sustainability of populations. In livestock, the amount of genetic diversity should be large enough to enable the adaptation of populations to changing environments and market requirements, and for selection to genetically improve economically important traits. Unfortunately, the current trend in populations is often for reduced genetic diversity due to intense selection or random drift. Consequently, breeding methods and gene banks were developed to avoid the risk of losing genetic diversity. As genomic information becomes more accessible, we now have the option to better manage genetic diversity. In this thesis, I applied genomics to conservation practises. More specifically, I applied genomic tools and methods to prove their relevance for the conservation of Dutch livestock breeds. I demonstrated that the use of genomics led to a more detailed understanding of the genetic diversity conserved in gene banks or in living populations of numerically small breeds in The Netherlands. Moreover, I reported the implications for genetic diversity of (1) lines or supposed lines within a numerically small breed, (2) merging and terminating lines of the Dutch Landrace pig breed, and (3) the replacement over time of traditional local cattle breed (Dutch Friesian Cattle) with just productive breed (Holstein Friesian). Subsequently, I illustrated that only a small set of informative SNPs is needed to differentiate among Dutch local cattle breeds. Using such a small set of informative SNPs a genetic tool (DNA test) was developed for the determination of breed purity of cattle. Lastly, I addressed the recent developments in genomics and how they can be used effectively for genetic conservation, and in particular how gene banks can benefit from these developments, and I outline possible future directions for (a more effective) conservation of breeds using genomic methods. More specially, I propose a strategy for conservation and stated that gene banks should transform from “traditional gene banks” into “digital gene banks”.

Contents

1

General introduction

1.1 Introduction

Conserving genetic diversity is essential for the sustainability of populations. In livestock, the amount of genetic diversity should be large enough to enable the adaptation of populations to changing environments (such as production systems and climate) and market requirements, and for selection to genetically improve the economically important traits (FAO 2015; Woolliams and Oldenbroek 2017; Doekes 2020). Genetic diversity is important in stable environments as well. When breeds are reduced to a small number of breeding individuals (i.e., a few hundred), negative consequences of inbreeding can occur, reducing amongst other traits the fitness of animals, which threatens breeds' survival (Hoban et al. 2021). Unfortunately, the current trend is often a reduced genetic diversity due to intense selection or random drift. Consequently, breeding methods (e.g., optimal contributions) and gene banks have been developed to avoid the risk of losing genetic diversity. Genomic techniques are developed to improve the efficacy of breeding programs and create an opportunity to describe and conserve the genetic diversity more accurately.

1.2 Genetic diversity under threat

Genetic diversity is the set of differences between species, populations within species, and individuals within populations present in their DNA or observed in individuals as a result (Woolliams and Oldenbroek 2017). Genetic diversity within livestock species is generally defined by the number of breeds in the species of interest and their level of similarity and uniqueness from a genetic point of view. Within breeds genetic diversity is defined as the level of genetic similarity and uniqueness of individuals, i.e., differences in their DNA content. This variation is influenced by selection and random drift. Nowadays, genetic diversity is rapidly being lost. According to the Food and Agriculture Organization of the United Nations (FAO), numerous livestock breeds have gone extinct or are threatened (FAO 2015), because local breeds have been upgraded or replaced by a few very strong selected mainstream breeds (FAO 2006; Oldenbroek 2019). Genetic diversity is threatened by strong selection on a few traits, as it increases genetic relationships between animals and reduces effective population size. This will result in a higher rate of inbreeding and associated negative aspects, such as a reduction in the viability of populations (Frankham 1995; Oldenbroek and Windig 2022). To be able to handle future challenges in agriculture, it is important to conserve within and between breed genetic diversity and efforts will have to be made to achieve this.

1.3 Conservation of genetic diversity

Awareness has increased that diversity within livestock species should be conserved. Therefore, local breeds should be protected from extinction, although commercial aspects still favour the advancement of a few mainstream breeds (Felius et al. 2011). The importance of conserving genetic diversity has been recognised by many countries by signing the Convention of Biodiversity in 1992 and adopting the Global Plan of Action for Animal Genetic Resources of the FAO in 2007 (FAO 2007a). These initiatives highlight the responsibility and commitment of each country to conserve their native livestock breeds, and to take action to prevent loss of genetic diversity (Meuwissen 2009; Hiemstra et al. 2010; Engelsma 2012). The conservation of genetic diversity requires actions. The first action should be a detailed description and monitoring of the breeds and their potential risks for losing genetic diversity. Next, in order to prevent the loss of genetic diversity, conservation actions can be initiated for the current population (in situ). These actions focus on the selection of breeding individuals, the management of mating designs as well as the control over the individuals' contributions to the next generation (Meuwissen 1997; Caballero and Toro 2000). As a result, it primarily limits the rate of inbreeding and supports a viable population. In addition to in situ conservation, genetic material can be conserved ex situ in gene banks. Gene banks allow to conserve the overall population genetic diversity in the form of reproductive material, such as semen or embryos, for an indefinite period of time. The genetic diversity conserved in germplasm is not subject to evolution or drift (Eynard 2018). For the conservation of animal genetic resources, both in situ and ex situ approaches are used and they are generally considered complementary to each other (FAO 2019). The conservation of genetic diversity is a costly process and budgets are often limited. Therefore, conservation of genetic diversity should be carried out in the most effective and efficient manner possible. Determination and evaluation of genetic diversity within livestock breeds is of crucial importance for making the right conservation decisions and for an efficient use of resources available for conservation.

1.4 Genetic diversity in the genomics era

The development and use of genomics provide tools to obtain a more complete picture of the parameters of genetic diversity that can be used for breed prioritisation, conservation or management decisions. Traditionally, genetic diversity has been estimated and managed with the help of inbreeding and kinship coefficients based on pedigree data (Meuwissen and Luo 1992; Engelsma 2012). However, pedigree data present drawbacks that limit their use in genetic diversity

analyses. First, pedigrees may be incomplete, incorrect or not available for a reasonable number of generations. Secondly, it is assumed that founders are unrelated, while they are likely to be at least slightly related in practice. Thirdly, the assumption when measuring genetic relationships based on pedigree information is that full siblings share exactly 50% of their alleles, while in reality they are subject to mendelian sampling that creates variation in the amount of allele sharing between full siblings; the same holds for other relationships. Fourthly, pedigrees are generally recorded per breed, making analysis of between-breed diversity by pedigree analysis impossible (Eusebi et al. 2019; Doekes 2020; Galla et al. 2022). With the wide availability of genomic data, it has become possible to study and quantify genetic diversity directly rather than using pedigree information to make statistical inferences. This is true even in populations that do not have genealogical records (Frankham et al. 2012) and in populations with these records it corrects pedigree errors. Previous studies (de Cara et al. 2011; Engelsma et al. 2011; Eynard 2018) investigated the impact of using genomic information from SNP array data instead of pedigree information for evaluation of genetic diversity within breeds. They showed that genomic information assesses more accurately the genetic diversity, both for the whole genome and for specific regions of the genome, which improves the management of genetic diversity. Engelsma et al. (2012) showed that SNPs preserved the genetic diversity within breeds better than pedigrees. Previous studies have described methods and opportunities of genomics for conservation of genetic diversity (e.g. Eding et al. 2002; Engelsma 2012; Eynard 2018). Because of the availability of a wide range of genomics tools and methods, the expectation for the practical application for conservation of genetic diversity is high. However, customisation is required and depends on the questions related to conservation. In the meantime, genomics is still evolving and provides an increasing amount of information and detail. In this thesis, I used genomics for conservation practices. I applied genomic techniques and methods to prove their potency for the improvement of the present conservation activities for Dutch livestock breeds.

1.5 Aim and outline of the thesis

The overall aim of this thesis is to analyse the genomic data of Dutch livestock breeds. The aim of these analyses is to support the conservation of genetic diversity in these breeds. The results can make conservation decisions more accurate. Breeds involved varied from the numerically small and at one point almost extinct Dutch Red and White Friesian to the worldwide number one dairy cattle breed Holstein Friesian.

The first objective was to investigate how to deal with lines or supposed lines within a numerically small breed and the consequence for conservation. To this end, genomics was used to quantify genetic diversity within and between lines of the Dutch Red and White Friesian cattle, as well as their relationship with other Dutch cattle breeds (**Chapter 2**).

Over the past decades, various commercial pig breeding lines were merged or discontinued due to consolidation in the pig breeding industry. Fortunately, their semen was conserved in the Dutch gene bank. The second objective was therefore to assess the implications for genetic diversity of merging lines of the Dutch Landrace pig and the discontinuation of lines in this breed using genomics. (**Chapter 3**).

Over the last century, genetic diversity has been affected by the replacement of traditional local breeds with just a few highly productive breeds. The third objective was to evaluate the consequences for the whole genome and especially for its rare allelic variants of the replacement of the Dutch Friesian cattle (DF) by the Holstein Friesian breed (HF). To do so, we evaluated genome-wide genetic diversity between three groups of bulls, chosen from the historic (1961–1989) and recent (2003–2015) DF population and the recent HF (1998–2014) population using whole genome sequencing (WGS) (**Chapter 4**).

Maintaining genetic diversity can be achieved not only by characterising genetic diversity, but also by actively increasing the population size of a breed. This can be done by identifying whether unregistered animals belong to a certain breed. The use of small sets of informative SNPs can be a cost-effective option for the estimation of breed composition. The fourth objective was to evaluate methods of SNP selection and determine the minimum number of SNPs needed to differentiate among Dutch cattle breeds (**Chapter 5**). The last objective was to develop a low cost genomic test, using a small set of informative SNPs, to identify breed of origin and breed purity of unregistered individuals to assign them to one of the Dutch local cattle breeds (**Chapter 6**). Including them actively may increase the (effective) population size and therefore enlarge the genetic diversity.

All studies in this thesis benefitted from material stored in the Dutch gene bank, and in the general discussion (**Chapter 7**) I reflect on the role of the gene bank and how it may develop and benefit from developments in genomics and bioinformatics and outline possible directions for (better) conservation of breeds using genomic methods in the future.

2

Conservation priorities for the different lines of Dutch Red and White Friesian cattle change when relationships with other breeds are taken into account

B. Hulsegge[1] , M.P.L. Calus[1] , J.K. Oldenbroek[2] & J.J. Windig[1]

[1] Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, Wageningen, The Netherlands; [2] Centre for Genetic Resources The Netherlands, Wageningen, The Netherlands.

Journal of Animal Breeding and Genetics (2017) 134: 69–77

Abstract

From a genetic point of view, the selection of breeds and animals within breeds for conservation in a national gene pool can be based on a maximum diversity strategy. This implies that priority is given to conservation of breeds and animals that diverge most and overlap of conserved diversity is minimized. This study investigated the genetic diversity in the Dutch Red and White Friesian (DFR) cattle breed and its contribution to the total genetic diversity in the pool of the Dutch dairy breeds. All Dutch cattle breeds are clearly distinct, except for Dutch Friesian breed (DF) and DFR and have their own specific genetic identity. DFR has a small but unique contribution to the total genetic diversity of Dutch cattle breeds and is closely related to the Dutch Friesian breed. Seven different lines are distinguished within the DFR breed and all contribute to the diversity of the DFR breed. Two lines show the largest contributions to the genetic diversity in DFR. One of these lines comprises unique diversity both within the breed and across all cattle breeds. The other line comprises unique diversity for the DFR but overlaps with the Holstein Friesian breed. There seems to be no necessity to conserve the other five lines separately, because their level of differentiation is very low. This study illustrates that, when taking conservation decisions for a breed, it is worthwhile to take into account the population structure of the breed itself and the relationships with other breeds.

Key words: Conservation, genetic diversity, population structure, relationships with other breeds.

2.1 Introduction

Farm animal breeds are recognized for different values, with economic, social, historical and cultural aspects (Gandini and Oldenbroek 2007). Genetic diversity is the basis for the development and survival of animal breeds. However, many traditional, local, farm animal breeds have small (effective) population sizes, leading to a loss of their genetic diversity. It is, therefore, especially important to maintain genetic diversity in these small populations of farm animals (Fernandez et al. 2011). Small populations of local breeds often comprise genetic variation with cultural, historical, sociological and environmental values (Hiemstra et al. 2010) generally not present in the global highly productive breeds that dominate modern intensive livestock production systems. Genetic management of local breeds, is crucial for their own survival and for maintaining diversity in the entire species, because the genetic diversity between breeds is a substantial part of the genetic diversity within the species (Woolliams and Toro 2007).

Maintaining high levels of within-breed genetic diversity is the second important aim in conservation genetic diversity within the species. Traditionally, animal breeders quantify genetic diversity by analysing pedigrees and estimating average kinships and inbreeding levels (Gutierrez et al. 2003; Woolliams and Toro 2007). Pedigree analysis may not be adequate, as pedigrees are often not available in depth, so that a reliable quantification of within-breed variation may not be possible. Moreover, pedigrees are generally only known as breed formation, making analysis of between-breed diversity by pedigree analysis impossible. Methods based on pedigree analysis can now be complemented with molecular genetic information facilitating analysis of diversity both within and across breeds (Boettcher et al. 2010).

Besides small effective population size, local breeds may be threatened by indiscriminate crossing with other breeds. Crossing may lead to increased genetic diversity in a population, however, at the expense of losing part or eventually all of the original genetic diversity in the population (FAO 2007b). Thus, both within- and across-breed variations need to be considered to preserve genetic diversity within species (Bennewitz et al. 2007; Woolliams and Toro 2007; Boettcher et al. 2010; Roberts and Lamberson 2015).

Eding et al. (2002) provided a framework to quantify relative amounts of both within- and across-population genetic diversity using marker-estimated kinships. In this method, kinships are estimated with the help of markers and the genetic diversity

within a breed is estimated as one minus the average kinship in that breed. The average kinship is also estimated across breeds, so that the genetic diversity of a set of breeds can be determined. Moreover, for each breed, its contribution to the diversity of the total set can be quantified, thereby quantifying both its unique diversity and the overlap with other breeds.

After the study of Eding et al. (2002), progress in genotyping techniques has increased the number of available markers. The availability of dense molecular marker maps can provide a more precise picture of the genetic background of breeds (e.g., distances, uniqueness), which increase the capabilities for making decisions aimed at maintaining genetic diversity.

In this study, the maximum diversity strategy was used to quantify the genetic diversity (Bennewitz et al. 2007). This strategy selects breeds that contribute in a significant way to the overall genetic diversity considering both within- and across-breeds diversity.

For local breeds, next to setting conservation priorities at breed level, a more detailed division into lines can be helpful to determine conservation priorities within the breed.

The objective of this study was to quantify the genetic diversity in a numerically small breed and its contribution to the total genetic diversity in other breeds of the same species in the same country. For these objectives, we used the Dutch Red and White Friesian cattle (DFR) and quantified the relationship with other Dutch dairy breeds. We assessed the following:

i. The relationship of DFR with other Dutch dairy breeds and the contribution of the DFR to the total genetic diversity in Dutch dairy cattle breeds.
ii. The genetic differences between lines within the DFR.
iii. The contribution of the within-line genetic diversity to the total genetic diversity in the DFR and to the gene pool of the Dutch dairy cattle breeds.

2.2 Materials and methods

2.2.1 Animals and genotypes

A total of 68 Dutch Red and White Friesian cattle (DFR) animals (26 bulls and 42 cows) were sampled. The DFR is a local breed in the North of the Netherlands. Anecdotally and according to herdbook information, it is closely related to the Dutch Friesian (DF) breed, which is one of the founding breeds of the Holstein Friesian, which is now the dominant dairy cattle breed in the world (Felius et al. 2011). Of the 68 sampled DFR animals, 48 animals were assigned to different lines, based on their ancestry from (founding) sires, within the breed by the Dutch herdbook 'Stichting Roodbont Fries Vee' (Table 2.1). Two other groups consist of animals not (yet) registered in the herdbook: one group from two farms with some Holstein Friesian (HF) blood and another group of isolated animals originating from the Dutch island Terschelling, from here on referred to as line 6 and 7, respectively.

Table 2.1. Number of samples per line of Dutch Red and White Friesian animals.

Line	Name	#Bulls	#Cows	Total
1	Jet	5	4	9
2	Marco-Kei	3	5	8
3	Koos	5	5	10
4	Reitsma	4	7	11
5	DF-line	8	2	10
6	Elsinga line		11	11
7	Terschelling	1	8	9
Total		26	42	68

To obtain DNA, we collected hair samples from the cows. From the bulls, semen straws were provided by the Centre for Genetic Resources, the Netherlands (CGN). Samples were chosen, based on pedigree information of the herdbook, so that they represent a wide variation in origin within a line. Samples were genotyped using the BovineSNP50 BeadChip (Illumina Inc., San Diego, CA, USA). All samples had a genotype call rate >85%. During the quality check, SNPs with a GenCall score ≤0.20 and call rate ≤85% were deleted from the analyses (n = 2635). Missing genotypes were imputed using Beagle with 20 iterations (Browning and Browning 2009). The imputation was carried out for each chromosome independently. The mean r^2 value for the accuracy of imputation provided by Beagle was 0.98. After these editing steps, 51,974 of the initial 54,609 SNPs remained.

Data from the DFR cattle were supplemented with data originating from studies with four other Dutch breeds (Maurice - Van Eijndhoven 2014; Pryce et al. 2014; Maurice-Van Eijndhoven et al. 2015). These data included 1,287 purebred cows; 989 were Holstein Friesian (HF), 97 Groningen White headed (GWH), 137 Meuse-Rhine-Yssel (MRY) and 64 Dutch Friesian (DF). Previously performed editing steps to remove uninformative SNP are described by Hulsegge et al. (2013). In short, Holstein Friesian animals were genotyped with a BovineSNP50 BeadChip and imputed to the BovineHD BeadChip using Beagle (Browning and Browning 2009). The mean Beagle r^2 was 0.96 across the imputed loci. Animals from the three other breeds (GWH, MRY and DF) were genotyped with the BovineHD BeadChip. The editing steps comprised deleting SNP with call rate <95%, GenCall score ≤0.20 and GenTrain score ≤0.55. No MAF (minor allele frequency) thresholds were applied in the editing procedure. To investigate whether differences in results could arise with edits based on MAF, as is commonly done in other studies or applications, the impact of MAF threshold 0.02 was evaluated. The preliminary analyses indicated that our results and conclusions were hardly affected when not applying such editing step (results not shown). After the editing steps, 750,457 of the 777,962 SNPs remained. These 750,457 SNPs contained 36,625 SNP that were also included in the DFR data after editing. For all animals, genotypes on those 36,625 SNPs were used in further analyses.

2.2.2 Breed identity of DFR

To investigate whether DFR is a breed with its own genetic identity and to visualize the relationship between DFR and the four other Dutch cattle breeds, principal component analysis (PCA) was performed on the SNP genotypes (Patterson et al. 2006; Price et al. 2006) using the R-package Hierfstat (Goudet 2005). Genetic divergence between each breed pair was quantified by calculating pairwise F_{ST} (Weir and Cockerham 1984) using the R-package Hierfstat (Goudet 2005).

2.2.3 Contribution of DFR to total genetic diversity in Dutch dairy cattle

To quantify the importance of DFR relative to the other breeds, the marker-estimated kinships and the core set method of Eding et al. (2002) were used. In this method, kinships are estimated with the help of markers and the genetic diversity within a breed is estimated as one minus the average kinship in that breed. The average kinship is also estimated across breeds, so that the genetic diversity of the whole set can be determined. The total genetic diversity of a set depends on the contribution of each breed to the total set. If all breeds contribute equally, the total

genetic diversity is equal to one minus the average within- and across-breed kinships. Otherwise, breed kinships have to be weighted by their contribution, for example:

$$g_{div} = \mathbf{1} - \boldsymbol{c}'\boldsymbol{M}\boldsymbol{c}$$

with $\boldsymbol{c}$ being the vector with n (number of breeds) contributions of each breed (summing up to 1) and $\boldsymbol{M}$ being the *n* × *n* matrix with within- and across-breed kinships. So, if a relatively uniform breed contributes more to the total set, the genetic diversity of the total set will be lower compared to when a relatively diverse breed contributes more.

In the core set method of Eding et al. (2002), the contribution of each of the breeds that maximize the genetic diversity is estimated as follows:

$$\boldsymbol{C}_{max} = \frac{\boldsymbol{M}^{-1}\mathbf{1}_n}{\mathbf{1}_{n'}\boldsymbol{M}^{-1}\mathbf{1}_n}$$

where $\boldsymbol{C}_{max}$ is a vector with the contributions that maximizes the diversity in the total set, and $\mathbf{1}_n$ is a vector of *n* ones. The total diversity in the set is then estimated as follows:

$$\boldsymbol{Div}_{set} = \mathbf{1} - \boldsymbol{c}_{max'}\boldsymbol{M}\boldsymbol{c}_{max} = \frac{\mathbf{1}}{\mathbf{1}_{n'}\boldsymbol{M}^{-1}\mathbf{1}_n}$$

The contribution of each breed to this core set thus depends both on the between- and within-breed components of genetic diversity. However, not only the contribution determines the relative importance of a breed for the total genetic diversity. A breed may contribute a small amount to the core set (e.g., when their within-breed kinship is high) but nevertheless increase the total genetic diversity considerably (e.g., when it's across-breed kinships are low). Therefore, the average kinship of the core set when the breed is included is compared to the average kinship of the core set when the breed is excluded (Eding et al. 2002).

The required kinships were obtained by first computing a genomic relationship matrix (G) according to Yang et al. (2010) using the software Calc_grm (Calus 2013). Using those genomic relationships, average within- and between-breed kinships were computed across all pairwise relationships within and between breeds, including self-kinships.

2.2.4 Contribution of lines to genetic diversity within DFR

To visualize the separation of the different lines based on molecular genetic data, PCA was used. The core set method was used to determine the relative contribution of each line to the total genetic diversity in the DFR. The core set method was also performed with both the DFR lines and the other breeds simultaneously, to determine the overlap of the contribution of the individual DFR lines to the total genetic diversity with the contribution of other breeds.

2.3 Results

2.3.1 Relationship of DFR cattle breed with other Dutch dairy breeds

The combination of the first and second principal components (PC1 and PC2) separated individual animals according to their breed (Figure 2.1). PC1 distinguished the four local breeds from the commercial breed HF. PC2 separated the local breeds MRY on the one hand and GWH on the other hand from the Friesian breeds (DF, DFR and HF). Based on the first two principal components, overlap existed between the DF and DFR.

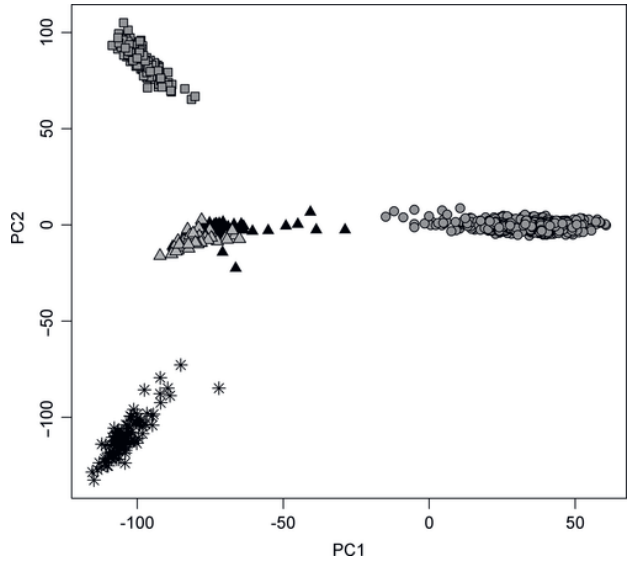

Figure 2.1. Principal component analysis (PCA) of five Dutch dairy cattle breeds based on 36,625 single-nucleotide polymorphisms (SNP's) [circle grey = Holstein Friesian (HF); star = Groningen White headed (GWH); triangle grey = Dutch Friesian (DF); square grey = Meuse-Rhine-Yssel (MRY); triangle black = Dutch Red and White Friesian (DFR)].

Genetic differentiation (pairwise F_{ST}) among breeds, confirmed that DFR is genetically closest to DF (F_{ST} = 0.056) (Table 2.2). Pairwise F_{ST} values ranged from 0.056 (between DFR and DF) to 0.156 (between GWH and DF). The kinship values also indicated that DFR and DF were more related to each other than to the other breeds. DFR and DF had the highest average between-breed kinship (0.033) (Table 2.2). Average between-breed kinship ranged from −0.078 to 0.033.

Table 2.2. Estimated pairwise F_{ST} as a measure of genetic differentiation (below diagonal) and average genomic kinship (above diagonal) between five Dutch dairy cattle breeds.

	GWH	DF	MRY	HF	DFR
GWH	–	−0.078	−0.057	−0.053	−0.068
DF	0.156	–	−0.067	−0.056	0.033
MRY	0.155	0.135	–	−0.031	−0.050
HF	0.132	0.111	0.110	–	−0.036
DFR	0.136	0.056	0.111	0.088	–

Abbreviations: GWH = Groningen White headed; DF = Dutch Friesian; MRY = Meuse-Rhine-Yssel; HF = Holstein Friesian; DFR = Dutch Red and White Friesian.

Dutch Red and White Friesian showed the lowest average within-breed kinship (0.106) and GWH the highest (0.248) (Table 2.3). The total diversity of the Dutch cattle breeds was 0.926. All five breeds contributed almost equal to the overall genetic diversity (varying from 19.55% to 20.64%). The highest unique genetic diversity was observed for GWH (0.015) and the lowest for DFR (0.006). Nevertheless, the DFR contains some unique genetic diversity not present in the other Dutch breeds, although it is less than the unique diversity of the other breeds (Table 2.3).

Table 2.3. Average genomic kinship (*f*) within breeds and contribution of breeds to a core set in which the diversity is maximized (= average *f* minimized). Unique diversity is measured as the increase in *f* when the core set is formed without a contribution of that breed.

	f	Contribution	Unique diversity
DFR (all lines)	0.106	19.84%	0.006
GWH	0.248	19.93%	0.015
DF	0.155	19.55%	0.007
MRY	0.199	20.04%	0.012
HF	0.174	20.64%	0.010
Core set	0.074		–

Abbreviations: DFR = Dutch Red and White Friesian; GWH = Groningen White headed; DF = Dutch Friesian; MRY = Meuse-Rhine-Yssel; HF = Holstein Friesian.

2.3.2 Genetic differences between DFR lines

Principal component analysis distinguished DFR line 7 from the other lines by the first principal component (Figure 2.2). There was some differentiation among the other lines along the second principal component, but with a large overlap between the different lines. Genetic differentiation between the different DFR lines was also confirmed by the pairwise F_{ST}, which varied between 0.012 and 0.190 (Table 2.4). Consistent with the PCA results, the F_{ST} values indicated that line 7 clearly diverged from the other lines. Pairwise F_{ST} between DFR line 7 and the other six lines ranged from 0.149 to 0.190, while the maximum pairwise F_{ST} between the lines 1 to 6 was 0.078 (between DFR lines 3 and 4). The F_{ST} values between DFR lines 1 to 6 were lower than the F_{ST} values between breeds (Table 2.2), meaning that the DFR lines 1 to 6 were more related to each other than the breeds were. The F_{ST} values between DFR line 7 and the other lines were somewhat higher than the values found between the breeds as presented in Table 2.2.

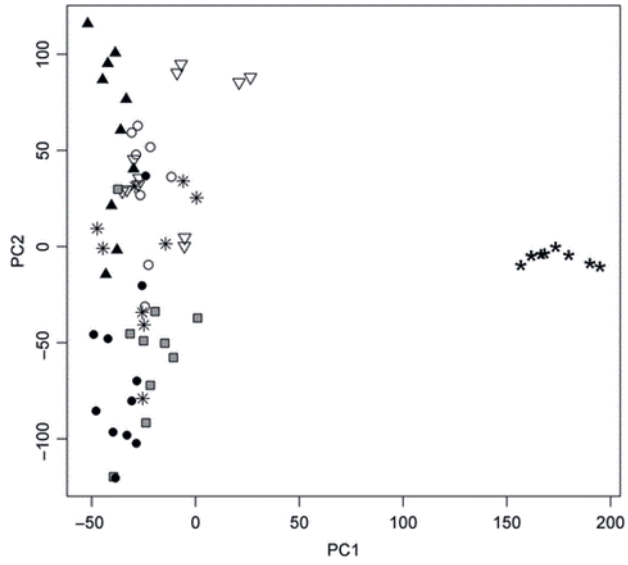

Figure 2.2. Principal component analysis (PCA) of seven Dutch Red and White Friesian (DFR) lines on 36,625 single-nucleotide polymorphisms (SNPs) [star = DFR line 1; circle white = DFR line 2; triangle point up black = DFR line 3; circle black = DFR line 4; square grey = DFR line 5; triangle point down = DFR line 6; asterisk = DFR line 7].

The average kinships between-line and within-line of the DFR breed are presented in Table 2.4 and 2.5. Within-line kinships were higher (Table 2.5; varying between 0.131 and 0.478) compared with the between-line kinships (Table 2.4; varying

between 0.041 and 0.157). The lines 1 to 5 were more related to each other than to the lines 6 and 7. DFR line 7 showed the highest within-line kinship (0.478) and the lowest between-line kinship (ranging from 0.041 to 0.053). DFR line 6 had the lowest level of within-line and the second lowest level of between-line kinship.

Table 2.4. Estimated pairwise F_{ST} as a measure of genetic differentiation (below diagonal) and average genomic kinship (above diagonal) between seven Dutch Red and White Friesian (DFR) lines.

	DFR line 1	DFR line 2	DFR line 3	DFR line 4	DFR line 5	DFR line 6	DFR line 7
DFR line 1	–	0.119	0.123	0.135	0.095	0.078	0.052
DFR line 2	0.042	–	0.140	0.114	0.091	0.080	0.053
DFR line 3	0.061	0.058	–	0.112	0.110	0.108	0.045
DFR line 4	0.036	0.056	0.078	–	0.157	0.069	0.043
DFR line 5	0.040	0.046	0.059	0.012	–	0.063	0.042
DFR line 6	0.048	0.049	0.057	0.063	0.046	–	0.041
DFR line 7	0.158	0.166	0.190	0.171	0.149	0.149	–

Table 2.5. Average genomic kinship (*f*) within lines and contribution of lines to a core set in which the diversity is maximized (= average *f* minimized). Unique diversity is measured as the increase in *f* when the core set is formed without a contribution of that breed/ line.

	f	DFR lines		All breeds/lines	
		Contribution	Unique diversity	Contribution	Unique diversity
DFR line 1	0.176	12.63%	0.005	13.26%	0.0002
DFR line 2	0.192	11.70%	0.004	11.90%	0.0002
DFR line 3	0.265	6.81%	0.001	15.37%	0.0003
DFR line 4	0.205	10.01%	0.002	16.35%	0.0002
DFR line 5	0.140	19.02%	0.008	14.97%	0.0002
DFR line 6	0.131	26.02%	0.020	14.82%	0.0002
DFR line 7	0.478	13.81%	0.014	13.21%	0.0004
Core set	0.126		–		

Abbreviations: DFR = Dutch Red and White Friesian.

The contribution of each line (in %) to the DFR breed is shown in Table 2.5. All lines contributed to the diversity of the DFR breed. The highest contribution to the total diversity of the DFR breed was observed for line 6 (26.02%), while line 3 showed the smallest contribution (6.81%). The total diversity of the DFR was 0.874. The largest part of diversity of most lines is represented in the other lines as well. The highest impact on the diversity was observed when line 6 or line 7 was removed, leading to

a decrease in overall diversity of the DFR breed by approximately 2.3% and 1.6%, respectively. Removing one of the lines 1 to 5 had only a small impact on the diversity. Apparently, the diversity contained in these lines is almost completely present in the other lines as well.

2.3.3 Contribution of the DFR lines to the total genetic diversity

The average kinship between DFR lines and the Dutch cattle breeds are presented in Table 2.6. This kinship varied from −0.079 to 0.085. The highest values were estimated between DFR lines and DF, while the lowest values were observed between DFR lines and GWH. Line 6 was the line most closely related to HF, and line 7 was the line least related to DF.

Table 2.6. Average genomic kinship between Dutch cattle breeds and Dutch Red and White Friesian (DFR) lines.

	GWH	DF	MRY	HF
DFR line 1	−0.065	0.027	−0.046	−0.040
DFR line 2	−0.065	0.030	−0.048	−0.040
DFR line 3	−0.073	0.031	−0.054	−0.045
DFR line 4	−0.074	0.050	−0.058	−0.051
DFR line 5	−0.079	0.085	−0.065	−0.056
DFR line 6	−0.058	0.010	−0.046	−0.003
DFR line 7	−0.060	−0.007	−0.030	−0.017

Abbreviations: DFR = Dutch Red and White Friesian; GWH = Groningen White headed; DF = Dutch Friesian; MRY = Meuse-Rhine-Yssel; HF = Holstein Friesian.

Results of assessing the impact of removing one line from the DFR breed and calculating the contribution of each line (in %) to the pool of Dutch dairy cattle breeds with maximal genetic diversity are shown in Table 2.5. When considering all Dutch dairy cattle breeds, removing one of the DFR lines has a small impact on the diversity (loss of 0.0002 to 0.0004; Table 2.5). When considering all breeds, the contribution of DFR line 6 was considerably smaller (14.82%) compared to DFR lines analysed in separation (26.02%). This was due to the inclusion of the HF breed, removing the HF breed increased the contribution of line 6 with 4.7% (results not shown). The contribution of DFR line 5 to the diversity across all breeds is also smaller (14.97%) compared to DFR lines only (19.02%). For DFR line 3, the contribution to the diversity across all breeds is larger (15.37%) compared to DFR lines only (6.81%).

Removing DF increased the contribution of DFR, especially by the contribution of line 5. Thus, analysing DFR in isolation of the other breeds suggests, for some lines, a

larger proportion of unique diversity, while part of this diversity apparently is due to influences of the other breeds, in particular DF and HF, as revealed by the analysis including other breeds.

2.4 Discussion

2.4.1 Relationship of DFR cattle breed with other Dutch dairy breeds

Genetically, Dutch cattle breeds are clearly distinct from each other as shown by the PCA results, except for DF and DFR. As expected from breed history, the DFR breed is closely related to the DF breed (FAO 2007b). These breeds were recorded as separate breeds for slightly more than 100 years. Red offspring of the DF breed, born out of the combination of two red factor carriers, could be incorporated in the DFR breed. From 1970, DF and DFR became rare (Porter 2002). Genetic differentiation between the breeds (pairwise F_{ST}) and the between-breed kinship also indicated that DFR and DF were more related to each other than to the other Dutch breeds. In European cattle breeds, pairwise F_{ST} values have been reported, that is ranging from 0.035 to 0.132 (Gautier et al. 2007) and from 0.059 to 0.142 (Neuditschko 2011). The F_{ST} between DFR and DF of 0.056 is at the lower end of these ranges. DFR showed a reasonable contribution (19.84%) to the total genetic diversity of Dutch cattle breeds and contains a small amount of genetic diversity not present in the other Dutch breeds. This contribution is comparable to the contribution of each of the other breeds. Thus, although DFR and DF are closely related, the results of this study showed that DFR has its own genetic identity, containing some genetic diversity not present in other breeds.

2.4.2 Genetic management of lines within breeds

Management of breeds subdivided in lines implies a compromise of different factors: first, the maintenance of the highest possible levels of genetic diversity for the whole breed; second, the preservation of the genetic differentiation between lines; and third, the restriction of within-line diversity to acceptable levels, so inbreeding would not increase beyond these acceptable levels (Fernandez et al. 2008). The results of our study revealed a high level of admixture between lines 1 and 5. This reflects the similar origin of these lines. Consequently, there seems to be no necessity to conserve these 5 lines separately, because their level of differentiation is very low. The line with the highest overall contribution to diversity in DFR is line 6. However, part of this diversity is due to some HF blood and therefore of lower conservation value.

The pairwise F_{ST} values indicated that DFR line 7 had a high level of genetic differentiation from other lines. This line has been bred for a considerable time in isolation from the other lines and apparently conserved genetic diversity not present anymore in the rest of the population. However, this line showed high levels of inbreeding, and a low level of diversity.

2.4.3 Contribution of lines within breeds to the total genetic diversity across breeds

A way to measure the influence of one line over the others in the DFR breed is to ascertain its genetic contribution to diversity by removing this line from the whole DFR breed and determining the remaining genetic diversity (Caballero and Toro 2002; Eding et al. 2002). However, the results are different when relationships of other Dutch cattle breeds are taken into account. Some DFR lines contains a portion of genetic diversity which is also represented in the other Dutch cattle breeds. Maximizing genetic diversity within a breed is therefore not always the best strategy. Thus, our results demonstrate that when establishing conservation programmes, it is necessary to take relationships with other breeds into account as well. Lenstra (2006) also indicated that for decisions on conservation priorities, the diversity of all local breeds related to the endangered population should be taken into account to assess their unique contribution to diversity.

2.4.4 Assessing contributions of lines without pedigree relationship to herdbook animals

Previously, pedigree information was the most important information used for registration of animals in a herdbook. Use of genome-wide SNP information now provides a way to assess the relationship of animals without pedigree to animals registered in a herdbook. The Dutch DFR herdbook 'Stichting Roodbont Fries Vee' had assigned 48 sampled animals in this study to five different lines. Two additional DFR lines were defined consisting of animals that were not registered (DFR line 6 and 7). The lines might be considered as subpopulations, but there are no formal restrictions on pairing animals from different lines with each other, whereas crosses between animals of different breeds are considered cross-breds and not registered as belonging to either breed. Consequently, in the context of diversity, relationships between lines are generally much higher than relationships between breeds.

For the lines without an official pedigree, the results of this study showed similarities and differences to the five lines (DFR lines 1 to 5) with an official pedigree. This study indicated that line 6, a group with some HF blood, indeed represents part of the HF

genetic diversity. Currently, there seems to be no necessity to conserve DFR line 6. However, conserving line 6 in situ may be useful in practice for several reasons: first, this line consists of approximately 100 animals, while the total population size of DFR is 500; second, to increase the milk production of the DFR breed; and third, to increase the genetic diversity of DFR and consequently to decrease the chance of inbreeding. However, conserving line 6 should not be at the expense of other lines.
This study distinguished DFR line 7 from the other DFR lines. However, considering all Dutch cattle breeds, line 7 is closely related to DFR and DF. This isolated group of animals will maximize the level of genetic diversity for the whole DFR breed and will increase genetic differentiation between lines, despite its high levels of inbreeding. Therefore, line 7 makes a unique contribution to the DFR cattle, and it is worthwhile to include this line without an official pedigree in the herdbook. The DFR herdbook and breeders are now considering the inclusion of line 6 and line 7 in the herdbook. It is often not possible and may also not be desirable, to conserve all breeds/lines, mostly due to financial limitations (Bennewitz et al. 2007). As shown in this study, taking relationships with other breeds into account can change conservation priorities within a breed and thus may affect conservation decisions made for this breed. This is applicable not only to the lines within a breed in this study, but also for breeds within a species or in a gene pool of national breed as in this study.

Conservation decisions also should take into account the degree of endangerment and costs of conservations and economic, cultural and historical values of different characteristics of a breed (Simianer et al. 2003; Bennewitz et al. 2007). Endangerment of most DFR lines is similar; however, line 7 is clearly more endangered since the owner has stopped active farming. DFR line 6 had the highest overall contribution to diversity in DFR; however, when considering HF, the contribution of DFR line 6 was considerably smaller, indicating that the endangerment of line 6 is not really a threat for the DFR breed as a whole. Consequently, conservation priorities based on genetic diversity coincides with priority based on degree of endangerment.

2.5 Acknowledgements

This work was supported by Centre for Genetic Resources, the Netherlands, funded by the Ministry of Economic Affairs, program ‘Kennisbasis Dier’, code KB-12-005.03.001 and program ‘WOT’, code WOT-03-003-056. The RobustMilk project and the National Institute of Food and Agriculture (NIFA) are acknowledged for providing the 50k genotypes of the HF cows, and the gDMI consortium is acknowledged for

imputing those to 777k genotypes. The Dutch Milk Genomics Initiative and the project 'Melk op Maat', funded by Wageningen University (the Netherlands), the Dutch Dairy Association (NZO, Zoetermeer, the Netherlands), the cooperative cattle improvement organization CRV BV (Arnhem, the Netherlands), the Dutch Technology Foundation (STW, Utrecht, the Netherlands), the Dutch Ministry of Economic Affairs (The Hague, the Netherlands) and the Provinces of Gelderland and Overijssel (Arnhem, the Netherlands) are thanked for providing the 777k genotypes of the GWH, FH and MRY cows. The authors acknowledge Myrthe Maurice – van Eijndhoven for collecting the data of the GWH, FH and MRY cows and the herd owners for their help in collecting the data. The text represents the author's views and does not necessary represent a position of the Ministry who will not be liable for the use made of such information.

3

Impact of merging commercial breeding lines on the genetic diversity of Landrace pigs

Ina Hulsegge[1,2], Mario Calus[1], Rita Hoving-Bolink[1,2], Marcos Lopes[3,4], Hendrik-Jan Megens[1] and Kor Oldenbroek[2]

[1] Wageningen University & Research, Animal Breeding and Genomics, P.O. Box 338, 6700 AH, Wageningen, the Netherlands; [2] Centre for Genetic Resources, the Netherlands, Wageningen University & Research, P.O. Box 338, 6700 AH, Wageningen, the Netherlands; [3] Topigs Norsvin Research Center, P.O. Box 43, 6640 AA Beuningen, the Netherlands; [4] Topigs Norsvin, 80420-210, Curitiba PR, Brazil.

Genetics Selection Evolution (2019) 51:60

Abstract

Background

The pig breeding industry has undergone a large number of mergers in the past decades. Various commercial lines were merged or discontinued, which is expected to reduce the genetic diversity of the pig species. The objective of the current study was to investigate the genetic diversity of different former Dutch Landrace breeding lines and quantify their relationship with the current Dutch Landrace breed that originated from these lines.

Results

Principal component analysis clearly divided the former Landrace lines into two main clusters, which are represented by Norwegian/Finnish Landrace lines and Dutch Landrace lines. Structure analysis revealed that each of the lines that are present in the Dutch Gene bank has a unique genetic identity. The current Dutch Landrace breed shows a high level of admixture and is closely related to the six former lines. The Dumeco N-line, which is conserved in the Dutch Gene bank, is poorly represented in the current Dutch Landrace. All seven lines (the six former and the current line) contribute almost equally to the genetic diversity of the Dutch Landrace breed. As expected, the current Dutch Landrace breed comprises only a small proportion of unique genetic diversity that was not present in the other lines. The genetic diversity level, as measured by Eding's core set method, was equal to 0.89 for the current Dutch Landrace breed, whereas total genetic diversity across the seven lines, measured by the same method, was equal to 0.99.

Conclusions

The current Dutch Landrace breed shows a high level of admixture and is closely related to the six former Dutch Landrace lines. Merging of commercial Landrace lines has reduced the genetic diversity of the Landrace population in the Netherlands, although a large proportion of the original variation is maintained. Thus, our recommendation is to conserve breeding lines in a gene bank before they are merged.

Key words: genetic diversity, SNP, pig breeds, consolidation breeding lines

3.1 Introduction

The pig is a major livestock species, which in 2016 accounted for 37% of the meat production worldwide (FAO 2016). The global pork production primarily relies on the use of a limited number of international commercial breeds, specifically Duroc, Large White, and Landrace. In the mid-twentieth century, a large number of breeding associations that operated regionally were responsible for pig breeding. Each of these breeding associations and breeding companies had their own breeding stock, which was usually based on the same limited number of commercial breeds, but often originated from national or regional, and therefore unique, populations.

Over the past decades, the commercial breeding industry has seen considerable business consolidation through mergers and take-overs, which have resulted in a limited number of remaining internationally operating breeding companies (De Man 2008). Consequently, the breeding lines owned by these companies have experienced a high degree of consolidation as well. Breeding lines that lost the competition in terms of performance and genetic gain were often discontinued but perhaps more often, breeding lines were merged 'asymmetrically', keeping the old breeding line's name, but with extraneous influences.

The process of consolidation of breeding lines in domestic farm animals is most advanced in poultry, where both for broiler and laying chickens, the global market relies on just a handful of breeding lines/populations. Currently, the global poultry breeding market is primarily covered by just a few breeding companies, which has led to a loss of genetic diversity in these breeds (Muir et al. 2008). Pig breeding shares similarities with poultry breeding in that it relies on a limited number of international breeds. Nevertheless, consolidation of pig breeding lines (and breeding companies, for that matter) has not yet progressed to the same extent. However, worldwide, genetic variation in pigs is threatened by the progressive marginalization of local breeds for the benefit of commercial breeds (Herrero-Medrano et al. 2014; FAO 2015). The continued merging of the many distinct local populations of these commercial pig breeds and lines is expected to further increase the loss of genetic potential for pig production.

Traditional pig breeds and pure breeding lines are valued resources, not only for meat production, but also for cultural, historical, sociological, and environmental aspects. The underlying genetic variation may disappear, or may already have disappeared, from the global highly productive breeds that dominate modern

intensive livestock production systems. Thus, the continued merging of breeding companies increases the concern of losing essential genetic variation (Hillel et al. 2003).

The consolidation of breeding lines is often poorly documented, with public records usually limited or absent. Even for breeding lines listed by the FAO, which include data on their current status and vulnerability, information is often limited or outdated. A post hoc evaluation of loss of diversity in the aftermath of company mergers by genotyping is further hampered by the absence of reference samples from the pre-merger breeding lines. Here, we present a relatively well-documented case of the merging of a number of breeding associations that operated at the national level (the Netherlands) into an internationally operating breeding company (Topigs Norsvin). Although consolidation affected all breeding lines owned by the breeding companies, we will focus on one particular breed in this paper, i.e. the Dutch Landrace.

Our objective was to investigate the consequences of merging and discontinuing breeding populations on the genetic diversity of the Dutch Landrace breed over the past decades. To achieve this objective, we used genotype data of boars from the former Dutch Landrace breeding lines that have been conserved in the Dutch Gene bank to quantify their relationship with the current Dutch Landrace breed, and to estimate the loss (if any) of genetic diversity as a result of the merging of lines.

3.2 Methods

3.2.1 Description of the Landrace breed

The Dutch Landrace breed originated from the original native Landrace pig, with infusions of the German Landrace and the Danish Landrace around 1900 (Haring 1961). By 1933, the Dutch Landrace was officially recognized as a Dutch native breed. By 1960, different breeding associations started selecting their own Dutch Landrace populations for their specific breeding goals. In the 1970s, Finnish and Norwegian Landrace pigs were imported into the Netherlands for use in crossbreeding programs (Hoving et al. 2017). During the 1990s, the Cofok, Dumeco, Fomeva, and Stamboek breeding associations, which together represent the majority of pig sales in the Netherlands, merged into a new internationally operating breeding organization called Topigs (Slaghuis 2009). During this period, semen from breeding lines that were owned by the parent breeding organizations was deposited into the Dutch Gene bank (http://www.genebankdata.cgn.wur.nl/). This practice has been continued by Topigs (now called Topigs Norsvin) during the last two decades,

resulting in a unique collection of material from breeding lines that were either discontinued or altered by merging lines. The timeline of consolidation of the different Landrace breeding lines in the Netherlands since 1960s is illustrated in Figure 3.1. (Hoving et al. 2017).

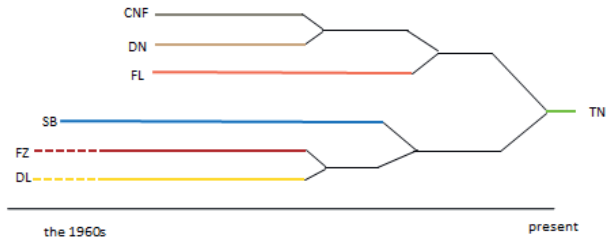

Figure 3.1. Timeline showing the consolidation of the Landrace breeds in the Netherlands since the 1960s (after Hoving et al. 2017). Abbreviations: CNF Cofok Norwegian and Finnish Landrace, DL Dumeco L-line, DN Dumeco N-line, FL Stamboek Finnish Landrace, FZ Fomeva Z1-line, SB Stamboek Dutch Landrace, TN Topigs Norsvin N-line.

3.2.2 Animals and genotypes

The Centre for Genetic Resources, the Netherlands (CGN) of Wageningen UR, i.e., the Dutch Gene bank, stores cryopreserved genetic material, primarily semen, from the former pig breeding associations in the Netherlands. From 1998 to 2003, CGN collected genetic material from six Landrace breeding lines of breeding associations that existed at that time. Merging of Dutch Landrace lines was in full progress and consequently the number of animals was already reduced. To select the group of boars, from the available animals, with minimal kinship and maximum diversity, optimal contributions were estimated using Gencont (Meuwissen 2002). From 2011 to 2016, CGN has preserved genetic material from the current Dutch Landrace line (Topigs Norsvin N-line; hereafter referred to as “TN line”) in the Dutch Gene bank. Genotype data, provided by CGN and Topigs Norsvin, were available for 187 animals from six former Dutch Landrace lines (Dutch lines from Fomeva, Dumeco and Stamboek, and Dutch Norwegian/Finnish lines from Cofok, Dumeco and Stamboek) and the current TN line (Table 3.1).

The 187 animals were genotyped using the PorcineSNP80 BeadChip (Illumina Inc., San Diego, CA, USA). All samples had a genotype call rate higher than 90%. For quality control, SNPs with a GenCall score lower than 0.20, a minor allele frequency lower than 0.02 and a per SNP genotype call rate less than 100% were removed from further analyses, the latter because some of the subsequent analyses cannot deal

with missing genotypes. Imputing missing genotypes was not appropriate for this dataset, since it requires more animals for each of the lines involved to be genotyped. In addition, applying a call rate threshold of 100% left a sufficient number of SNPs in the dataset for subsequent analyses. The final dataset included 42,655 SNPs with calls for all 187 animals.

Table 3.1. Number of genotyped animals in six former and the current Dutch Landrace line (TN line).

Line	Abbreviation	Origin of the lines[a]	Semen collection year	Number of animals
Cofok Norwegian and Finnish Landrace	CNF	FN	2000–2002	46
Dumeco L-line	DL	NL	1998–2002	49
Dumeco N-line	DN	FN	1998–2002	24
Stamboek Finnish Landrace	FL	FN	2002	11
Fomeva Z1-line	FZ	NL	2000	11
Stamboek Dutch Landrace	SB	NL	2002–2003	12
Topigs Norsvin N-line	TN	TN	2011–2016	34

[a]Origin of the lines: FN: Finnish/Norwegian; NL: Dutch; TN: current line.

3.2.3 Population structure

To examine relatedness between the Landrace lines, a principal component analysis (PCA) was performed using the prcomp function in R (R Core Team 2013). To identify subpopulations (clusters), genotypes of all individual animals were analysed by the model-based clustering algorithm implemented in the software Structure (version 2.3.4) (Pritchard et al. 2000; Falush et al. 2003). Subpopulation numbers (K) ranging from 2 to 7 were evaluated by repeating each analysis 10 times. A burn-in of 10,000 iterations and subsequent 50,000 iterations of the Markov chain Monte Carlo were applied, with all other program parameters set to their default values. The most likely number of subpopulations was inferred with the ΔK method of Evanno (Evanno et al. 2005), implemented in the R package pophelper (version 2.2.3) (Francis 2017). The program CLUMPP (Jakobsson and Rosenberg 2007) implemented in pophelper was used to align the 10 independent runs for each K. Pophelper was also used to plot results for K = 2 to 7. The Structure analysis was performed a second time by applying the “Use Population Information” setting, such that individuals of the TN line (POPFLAG = 0) were assigned to clusters that were defined by the allele frequencies of the other lines (POPFLAG = 1). A neighbour-joining tree (Saitou and Nei 1987) was computed based on the resulting distance matrix using the R package

APE (version 4.1) (Paradis et al. 2004). Genetic divergence between each pair of Landrace lines was quantified by calculating pairwise F_{ST}, as defined by Weir and Cockerham (Weir and Cockerham 1984), using the R-package 'hierfstat' (version 0.04–22) (Goudet 2005).

3.2.4 Genetic diversity

The contribution of breeds to genetic diversity was analysed using the marker-estimated kinships and the core set method of Eding et al. (2002). In this method, kinship coefficients are estimated based on SNP genotypes, and the genetic diversity within a breed is estimated as one minus the average kinship coefficient in that breed. The average kinship coefficient was also estimated across breeds to determine the genetic diversity of the whole set. The total genetic diversity of a set depends on the contribution of each breed to the total set. If all breeds contribute equally, the total genetic diversity is equal to one minus the average within- and across-breed kinship coefficients. Otherwise, the kinship coefficients of each breed have to be weighted by their contribution, as:

$$g_{div} = \mathbf{1} - \boldsymbol{c}'\boldsymbol{M}\boldsymbol{c}$$

where $\boldsymbol{c}$ is a vector of the n (number of breeds) contributions of each breed (summing to 1) and $\boldsymbol{M}$ is a $n \times n$ matrix with within- and across-breed kinship coefficients. Thus, if a relatively uniform breed contributes more to the total set, the genetic diversity of the total set will be lower than when a relatively diverse breed contributes.

In the core set method of Eding et al. (2002) the contribution of each breed that maximizes the genetic diversity is estimated as:

$$\boldsymbol{C}_{max} = \frac{\boldsymbol{M}^{-1}\mathbf{1}_n}{\mathbf{1}_{n'}\boldsymbol{M}^{-1}\mathbf{1}_n}$$

where $\boldsymbol{c}_{max}$ is the vector of contributions that maximizes the diversity in the total set, $\mathbf{1}_n$ is a vector of n ones, and $\boldsymbol{M}$ is the $n \times n$ matrix with the average within- and between-breed kinships. Then, the total diversity in the set is estimated as:

$$\boldsymbol{Div}_{set} = \mathbf{1} - \boldsymbol{c}_{max'}\boldsymbol{M}\boldsymbol{c}_{max} = \frac{\mathbf{1}}{\mathbf{1}_{n'}\boldsymbol{M}^{-1}\mathbf{1}_n}$$

Thus, the contribution of each breed to this core set depends on both the between- and within-breed components of genetic diversity. However, this contribution is not the only one that determines the relative importance of a breed to total genetic diversity. A breed that only contributes a small amount to the core set (e.g. when their within-breed kinship is high) can, nevertheless, increase the total genetic diversity considerably, e.g., when its across-breed kinships are low. Therefore, the average kinship coefficient of the core set when the breed is included is compared to the average kinship coefficient of the core set when the breed is excluded (Eding et al. 2002).

The required kinship coefficients were obtained by first computing the genomic relationship matrix (**G**) according to Yang et al. (2010), using the software Calc_grm (Calus 2013). Using **G**, average within- and between-breed kinship coefficients were computed across all pairwise relationships within and between breeds, including self-kinship coefficients.

3.2.5 Identification of selection signatures by using F_{ST}

Selection signatures were detected for each pairwise comparison between the current TN and the six former lines, by using the F_{ST}-outlier approach implemented in the BayeScan software (version 2.1), using default settings (Foll and Gaggiotti 2008). SNPs with a q-value lower than 0.05 were considered as outliers, which indicate regions potentially under selection. Genes that are located within 10 kb (5 kb downstream/upstream) of the SNP outliers were identified as candidate genes, based on the Ensembl annotation of Sscrofa10.2 (https://may2017.archive.ensembl.org/Sus_scrofa/Info/Index). The candidate genes were characterized using the PANTHER Classification System version 14.1 (http://geneontology.org/) (Mi et al. 2019), in particular, with the GO-Slim Biological Process annotation dataset. Overrepresentation analysis of GO-Slim Biological Process terms was also done using PANTHER; GO terms with a $p \leq 0.05$ after Bonferroni correction were deemed significant. Compared to using the entire GO term database, GO-Slim uses a limited set of GO terms to provide a more general list of functions that map to genes.

3.3 Results

3.3.1 Population structure

The current Dutch Landrace (TN: Topigs Norsvin N-line) is the result of the consolidation of six former Landrace lines that existed from the 1960s until early 2000 (CNF: Cofok Norwegian and Finnish landrace, DL: Dumeco L-line, DN: Dumeco

N-line, FL: Stamboek Finnish Landrace, FZ: Fomeva Z1-line and SB: Stamboek Dutch Landrace). The PCA clearly indicates a division of the seven Landrace lines into two main clusters; on the one hand, the former Norwegian/Finnish Landrace lines (CNF, DN and FL lines), which were introduced in the Netherlands between 1970 and 1980, and, on the other hand, the former Dutch Landrace lines (DL, FZ and SB) (Figure 3.2a). Clearly, the current commercial Dutch Landrace line (TN) is a mixture of the former breeding lines, since the old breeding lines included the extremes of the first principal component (PC1). The widespread distribution of the animals along PC1 for the current TN line shows that the contribution of the Dutch and Norwegian/Finnish lines to the current line differs between pigs. The second principal component (PC2) distinguished the DN line from the other six lines.

A unique genetic identity was identified for each of the six former Landrace lines based on the cluster analysis using the Structure software (Figure 3.2c). At K = 2, the two ancestries clearly reflected Dutch Norwegian/Finnish versus Dutch Landrace origins. At K = 3, DN was separated from CNF and FL (representative of Dutch Norwegian/Finnish Landrace). Based on ΔK, the most likely number of genetic groups (clusters) was equal to 5. While all parent lines appeared to be well separated at K = 5 (with the exception of FZ and SB), TN is clearly an admixed population with substantial contributions from the former breeding lines CNF, FL, and FZ/SB. At K = 5, the average proportion of membership of the founder breeds to TN was 0.205, 0.043, 0.350, 0.290, and 0.112 for CNF (cluster 1), DN (cluster 2), FL (cluster 3), FZ/SB (cluster 4), and DL (cluster 5), respectively. Results of the Structure analysis with no prior population information (POPFLAG = 0 for TN line) is shown in Figure S1 (Additional file 1: Figure S1), and confirmed the results of the PCA, i.e., that the contributions from the former lines differed between individuals. A neighbour-joining tree separated the breeding lines from each other in separate clades, except for the current TN line (Figure 3.2b).

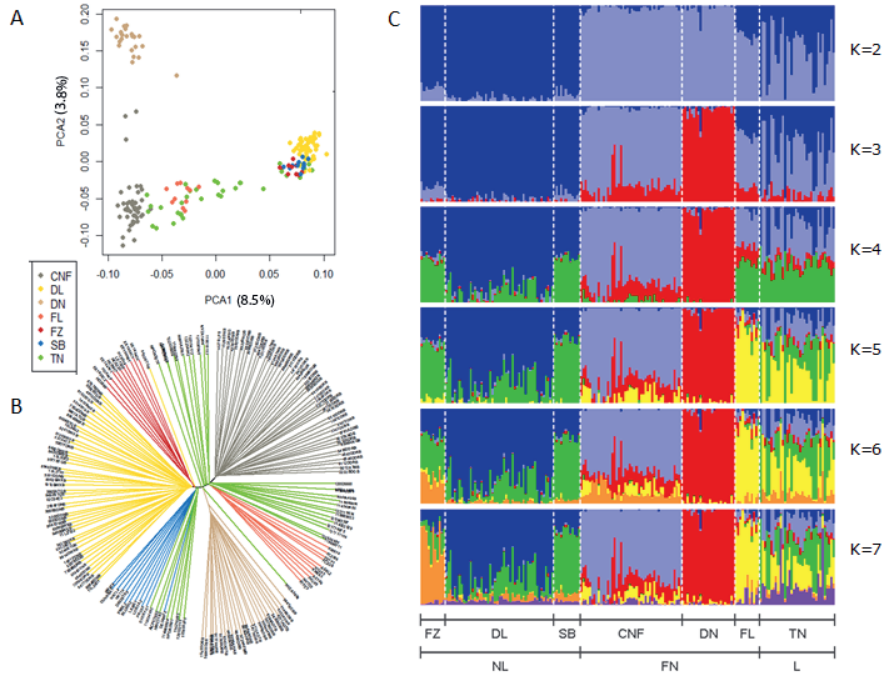

Figure 3.2. Population structure and relationships of Landrace breeding lines in the Netherlands. A) Principal component (PC) analysis, PC 1 against PC 2. B) Neighbour-joining tree of the relationships between the seven lines. C) Proportion of ancestry for each individual assuming different numbers of ancestral populations (K = 2 to 7). Colours of each vertical line represent the estimated proportion of an animal's genome that is assigned to a source population. Abbreviations: CNF Cofok Norwegian and Finnish Landrace, DL Dumeco L-line, DN Dumeco N-line, FL Stamboek Finnish Landrace, FZ Fomeva Z1-line, SB Stamboek Dutch Landrace, TN Topigs Norsvin N-line; FN: Finnish/Norwegian; NL: Dutch; L: current line.

Genetic differentiation among the Landrace lines was low to moderate, as indicated by the pairwise F_{ST} values that ranged from 0.02 to 0.10 (Table 3.2). The genetic differentiation of the current TN breeding line from the six former lines was low, which indicates that the current breeding line is closely related to the former breeding lines.

Table 3.2. Estimated pairwise F_{ST} as a measure of genetic differentiation (below the diagonal) and average genomic kinship (above the diagonal) between the Landrace breeding lines.

	CNF	DL	DN	FL	FZ	SB	TN
CNF	–	− 0.072	0.044	0.008	− 0.092	− 0.091	0.013
DL	0.066	–	− 0.070	− 0.072	0.025	0.048	− 0.019
DN	0.051	0.077	–	− 0.018	− 0.109	− 0.105	− 0.033
FL	0.036	0.044	0.074	–	− 0.074	− 0.087	0.010
FZ	0.055	0.030	0.098	0.088	–	0.025	− 0.034
SB	0.054	0.024	0.094	0.085	0.0588	–	0.039
TN	0.032	0.035	0.061	0.029	0.0412	0.0310	–

Abbreviations: CNF Cofok Norwegian and Finnish Landrace, DL Dumeco L-line, DN Dumeco N-line, FL Stamboek Finnish Landrace, FZ Fomeva Z1-line, SB Stamboek Dutch Landrace, TN Topigs Norsvin N-line.

3.3.2 Genetic diversity

The average kinship coefficients between and within the Landrace lines are in Tables 3.2 and 3.3. As expected, within-line kinship coefficients were higher (Table 3.3; ranging from 0.051 to 0.249) than the between-line kinship coefficients (Table 3.2; ranging from − 0.092 to 0.074). The higher negative between-line kinship coefficients between the former Dutch Norwegian/Finnish and the Dutch breeding lines indicates that the distance between these lines was greater than between individuals within the lines. The within-line kinship coefficient was lowest (0.051) for the current TN.

Table 3.3. Average genomic kinship coefficient ($\bar{f}$) within lines and the contribution of lines to a core set in which the diversity is maximized (= $\bar{f}$ minimised).

Line	$\bar{f}$	Contribution (%)	Unique diversity
CNF	0.170	15.74	0.005
DN	0.249	17.79	0.008
FL	0.158	14.70	0.007
DL	0.143	12.45	0.007
FN	0.186	13.28	0.004
SB	0.121	10.84	0.004
TN	0.051	15.18	0.003
Core set	0.007		–

Abbreviations: CNF Cofok Norwegian and Finnish Landrace, DL Dumeco L-line, DN Dumeco N-line, FL Stamboek Finnish Landrace, FZ Fomeva Z1-line, SB Stamboek Dutch Landrace, TN Topigs Norsvin N-line.

The contribution of each line (in %) to the genetic diversity in the overall Landrace population is shown in Table 3.3. All lines contributed to the diversity of the core set. The largest contribution to the total genetic diversity of the Landrace breed was observed for DN (17.79%), whereas it was smallest for SB (10.84%). Each line had a certain proportion of unique genetic diversity. The total genetic diversity of the Landrace breeding lines, estimated by Eding's core set method, was 0.993, and that of the six former breeding lines was 0.990, while the genetic diversity of TN was 0.894.

3.3.3 Identification of selection signatures using F_{ST}

As breeding lines are merged, selection continues, although in some cases the breeding goal may be different in the consolidated line compared to the parent lines. SNP genotypes were used to estimate allele frequency differentiation (measured as F_{ST}) in pairwise comparisons between the current TN and the six former lines. Outlier (high allele frequency differentiation) SNPs are an indication of regions that are potentially under selection. The log10 Bayes factor values for each SNP are shown in Figure 3.3. The number of loci with statistically significant patterns of divergent genetic differentiation (q-value ≤ 0.05), which were identified by pairwise comparisons, revealed that CNF and TN had the largest number (93) of outlier SNPs (Table 3.4). The outlier SNPs were located close to or within 20 candidate genes. Among these outlier SNPs, 29% (n = 27) were located almost at the end of chromosome 13 (SSC13: 191,713,636–196,766,412). Almost all of these 27 outliers are intergenic variants, which lie in-between genes (Additional file 2: Table S1). Additional file 2: Table S1 lists the outlier SNPs, candidate genes, and their respective assigned GO-slim terms (Biological Processes). Fifty-three SNPs were identified as loci that were under diversifying selection between the DL and TN lines, and these corresponded to 20 candidate genes. Seven outlier SNPs were located within small nucleolar RNAs (snoRNAs). Pairwise comparison between DN and TN revealed 46 SNP outliers (q-value ≤ 0.05) with 13 candidate genes. Pairwise comparisons of TN with each of the other four lines revealed 46 significant SNPs (q-value ≤ 0.05) between DN and TN, 18 between FL and TN, 21 between FZ and TN, and 7 between SB and TN. No candidate genes were found for the comparison between SB and TN. GO annotation of the candidate genes showed that most genes were linked to biological processes associated with cellular processes, metabolic processes, and intracellular transport (Table 3.4 and Additional file 2: Table S1). However, no significant over-representation was observed for any biological process.

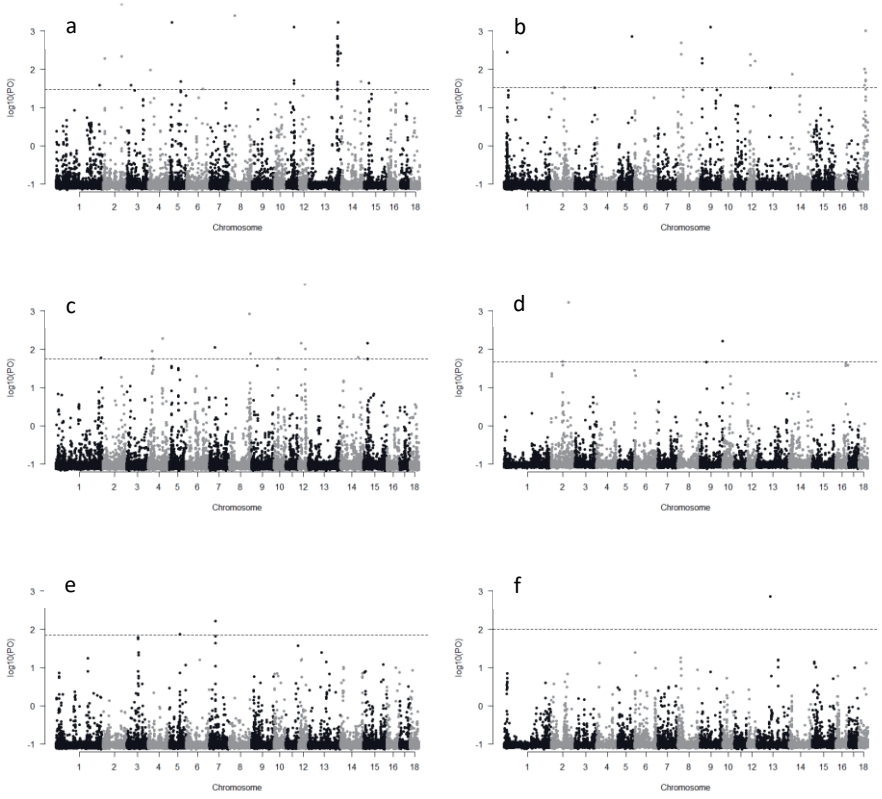

Figure 3.3. Genome-wide distribution of log10 Bayes factor values in the pairwise comparison between the current TN and the six former lines. a CNF versus TN, b DL versus TN, c DN versus TN, d FL versus TN, e FZ versus TN and f SB versus TN. Abbreviations: CNF Cofok Norwegian and Finnish Landrace, DL Dumeco L-line, DN Dumeco N-line, FL Stamboek Finnish Landrace, FZ Fomeva Z1-line, SB Stamboek Dutch Landrace, TN Topigs Norsvin N-line. The threshold for significance of signatures of selection is denoted with a line (q-value ≤ 0.05)

Table 3.4. Number of outlier SNPs detected (q-value ≤ 0.05) by BayeScan and their respective candidate genes within 5 kb up- or downstream.

Pairwise comparison of lines	Number of outlier SNPs	Candidate genes	General Term GO BP
CNF – TN	93	*CDC6*, *CIB4*, *CLEC1A*, *CLEC7A*, *ENSSSCG00000000959*, *ENSSSCG00000007221*, *ENSSSCG00000008799*, *ENSSSCG00000012012*, *ENSSSCG00000020566*, *ENSSSCG00000025389*, *ENSSSCG00000027643*, *ENSSSCG00000027841*, *ENSSSCG00000028250*, *GALM*, *GDE1*, *LAPTM5*, *LRRK2*, *RAB39A*, *SLC35F2*, *TCN1*	cell cell signalling / immune, cell communication, cell cycle, intracellular transport, metabolic process, skeletal muscle function and regeneration, system process, transmembrane transport
DL –TN	53	*ABRACL*, *BECN1*, *CCDC6*, *ENSSSCG00000010218*, *GARS*, *KIAA0513*, *MINDY4*, *NOL10*, *REPS1*, *SPNS2*, *SPNS3*, *UST*, *WNK4*	(cell) development/differentiation, cell communication, cellular processes, gene expression, intracellular transport, metabolic process
DN – TN	46	*CACNG3*, *CD48, CFAP45*, *DDX42*, *ENSSSCG00000024706*, *ENSSSCG00000026756*, *ENSSSCG00000027460*, *GLO1*, *GOT1, KSR1*, *PKHD1L1*, *ssc-mir-4331*, *TSPAN11*	cell communication, cellular processes, immune, intracellular transport, metabolic process
FL – TN	18	*EHBP1L1*, *KCNK7*, *MPP7*, *VAT1L*, *ZNF354C*	intracellular transport, metabolic process
FZ – TN	21	*FZD2*, *IL17REL*, *PLEKHM1*, *WDR92*	cellular process, developmental processes, intracellular transport
SB - TN	7	-	

Abbreviations: CNF Cofok Norwegian and Finnish Landrace, DL Dumeco L-line, DN Dumeco N-line, FL Stamboek Finnish Landrace, FZ Fomeva Z1-line, SB Stamboek Dutch Landrace, TN Topigs Norsvin N-line.

3.4 Discussion

In this study, we investigated the consequences for genetic diversity of the merging of lines within a breed, with the individual lines being discontinued thereafter. The data were derived from samples of boars that are included in the Dutch Gene bank collection, which led us to assume that these would encompass the genetic variation present in the modern breeding population (Berg and Windig 2017). Sample size of some Dutch Landrace lines used in this study were relatively small, due to limited availability of samples, and differed between lines, ranging from 11 samples for the lines FL and FZ lines to 49 for the DL line. Small sample size can lead to incorrect estimates of allele frequencies (Abi-Rached et al. 2018) and a proportion of genetic diversity present in the lines may remain undetected. Nevertheless, the results showed that the sampled animals formed genetic clusters that corresponded to their line designations (Figure 3.1). The results also showed that, genetically, the current commercial Dutch Landrace line (TN) is a mixture of the six former Landrace lines in the Netherlands. In general, the results reported here are in good agreement with the known history of the different Landrace lines examined (Slaghuis 2009; Hoving et al. 2017).

3.4.1 Genetic diversity

Genetic differentiation between the lines (pairwise F_{ST}) was moderate to low. Wilkinson et al. (2011a) reported a mean F_{ST} value of 0.156 between three British Landrace lines. For wild pigs sampled across different locations of the state of Florida (USA), pairwise F_{ST} values ranged from 0.020 to 0.256 (Hernández et al. 2018). The F_{ST} values (0.02 to 0.10) found in our study are at the lower end of this range. According to Willing et al. (2012), F_{ST} can be accurately calculated based on small sample sizes (as small as n = 4 to 6) if the number of markers examined is large, i.e. larger than 1000.

The results reported here showed that the merging of commercial Landrace lines has reduced the genetic diversity of the Landrace population in the Netherlands. For poultry, Besbes et al. (2008) also reported that the merging of lines leads to a decrease in genetic diversity of the available gene pool. However, our results also showed that, after merging, a large proportion of the genetic variability was maintained, and that all former lines showed a lower genetic diversity than the current TN. This indicates that merging lines is a better strategy for maintaining genetic diversity than just continuing with one line and discontinuing the other lines.

In this study, the total genetic diversity of the Landrace lines was estimated using the optimal contribution strategy. The optimal contributions of breeding lines were derived such that the average kinship coefficient in the core set was minimal, and thus the genetic diversity was maximal. Because breeding programs compete for market share, they select their lines intensively. Due to the breeding strategies that were followed over time, the actual genetic contributions of the different parent lines to the current Landrace line differed from the optimal contributions, indicating that part of the genetic diversity was lost. In addition, the DN line was poorly represented in the current Dutch Landrace. These observations support the recommendation that all breeding lines should be conserved before merging and discontinuing them.

3.4.2 Identification of selection signatures using F_{ST}

Commercial pig breeds have been subject to intense artificial selection for production traits. Functional analysis of regions under positive selection in pig breeds has identified genes that are involved in the development of the nervous system and of muscle, and in growth, pigmentation, metabolism, visual/odour perception, immune and inflammatory responses, and reproduction (Gouveia et al. 2014). Functional annotation analyses of the candidate genes in our study are shown in Table 3.4. For the interpretation of our results, it should be noted that we used the Ensembl annotation of Sscrofa10.2 and not the latest version Sscrofa11.1 (Warr et al. 2020). Furthermore, a small sample size can lead to poor population structure estimates, which affects the ability to differentiate between loci that were under selection and neutral population structure (Ahrens et al. 2018). However, in our study at least 11 animals per line were used, in line with a previous study that suggested that detecting regions under selection with F_{ST} methods requires at least 10 samples (Willing et al. 2012).

We detected no over-representation of any GO biological process among the candidate genes in our study. It should be noted that most traits that are under selection in pigs are complex traits that are regulated by many genes (Te Pas et al. 2017). We identified a number of candidate genes that were located within 10 kb (5 kb downstream/upstream) of the SNP outliers, most of them being associated with cellular processes, metabolic processes and intracellular transport (Table 3.4 and Additional file 2: Table S1). The candidate genes that were found in the comparison between the CNF and FN lines are involved in fertility (LAPTM5 (Abd El Naby et al. 2013); CIB4 (sheep) (Yu et al. 2010)), the immune system (RAB39A (Zhi et al. 2018)), and intramuscular fat content (ENSSSCG00000012012 (Wang et al. 2019)). In the

comparison between the DL and TN lines, we identified BECN1, which is a muscle-related gene (Liu et al. 2015; Lloyd et al. 2017), GARS and NOL10, which are associated with meat quality (Fontanesi et al. 2017; Xu et al. 2018), and KIAA0513, which is associated with the male reproduction trait "Seminiferous tubule diameter" (Zhao et al. 2016). In the comparison between the DN and TN lines, we detected several candidate genes: GLO1, which is assumed to be involved in fatness (Fowler et al. 2013), is important for nutrition energy intake and obesity (Kumar et al. 2007), and is connected with pig birth weight variability (Wang et al. 2016); GOT1 and PKHD1L1, which have been reported as candidate genes for intramuscular fat content (Ros-Freixedes et al. 2016) and variation in pH of meat (Chung et al. 2015), respectively; and TSPAN11, which was associated with metabolic body weight in a study on Holstein dairy cows (Hardie et al. 2017). In the comparison between the FZ and TN lines, we found the candidate gene WDR92, which is associated with total fat in Duroc and Yorkshire F2 intercrosses (Pant et al. 2014). It should be kept in mind that, although these associated SNPs and respective genes may be involved in certain biological processes related to selection events, further experimentation needs to be performed to verify these associations.

As shown by our results, differences can be pronounced even between populations that have common origins, which stresses the value of gene banks to record and preserve variation that is lost in the process of merging, even over short periods of time.

3.4.3 Consolidation

The breeding industry has undergone a strong consolidation process in the past decades and this will likely continue (Gura 2007; FABRE Technology Platform 2008; Franz and Rolfsmeier 2016). Economic reality forces breeding companies to discard breeding lines that are not of immediate value for product formulation or do not have potential to be used in the near future. Inevitably, maintaining genetic diversity in breeds and breeding lines has a cost, while the benefits are not immediately translated into profit. However, the consequences of losing genetic diversity are generally acknowledged; maintaining it is essential to provide future opportunities of selection for changing markets, consumer preferences, products etc., to allow sustained genetic improvement, to develop alternatives to intensive management, to decrease disease incidence and increase health, and to anticipate future changes in climate (Notter 1999; FAO 2007b; FABRE Technology Platform 2008; Boettcher et al. 2014).

3.5 Conclusions

The current Dutch Landrace (TN line) shows a high level of admixture and is closely related to the six former Dutch Landrace lines. However, the merging of commercial Landrace lines has reduced the genetic diversity of the Landrace population in the Netherlands, and the DN line is poorly represented in the current Dutch Landrace. Thus, it is recommended to conserve selection lines in a gene bank before merging. Our findings also showed that the merging of lines results in a large proportion of the original variability being maintained.

Additional files

The online version of this article (10.1186/s12711-019-0502-6) contains supplementary material, which is available to authorized users.

Availability of data and materials

The data that support the findings of this study are available from the Centre for Genetic Resources, the Netherlands and Topigs Norsvin but restrictions apply to the availability of these data, which were used under license for the current study, and thus are not publicly available. However, data are available from the authors upon reasonable request and with permission from the Centre for Genetic Resources, the Netherlands and Topigs Norsvin.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by Centre for Genetic Resources, the Netherlands, funded by the Ministry of Agriculture, Nature and Food Quality, program 'Kennisbasis Dier', code KB-21-004-001 and program 'WOT', code WOT-03-003-056 and the IMAGE project which received funding from the European Union's Horizon 2020 Research and Innovation Programme under the grant agreement n° 677353.

Authors' contributions

IH, MC, RH, ML, HM and KO conceived and designed the study. IH performed the data analysis. IH wrote the paper, with input from MC, RH, ML, HM, and KO. All authors read and approved the final manuscript.

Acknowledgements

The Centre for Genetic Resources, the Netherlands and Topigs Norsvin (the Netherlands) are acknowledged for providing the data.

4

Selection and Drift: A Comparison between Historic and Recent Dutch Friesian Cattle and Recent Holstein Friesian Using WGS Data

Ina Hulsegge[1,2] , Kor Oldenbroek[1,2], Aniek Bouwman[1], Roel Veerkamp[1] and Jack Windig[1]

[1] Animal Breeding and Genomics, Wageningen University & Research, P.O. Box 338, 6700 AH Wageningen, The Netherlands; [2]Centre for Genetic Resources, The Netherlands, Wageningen University & Research, P.O. Box 338, 6700 AH Wageningen, The Netherlands

Animals (2022), 12(3), 329

Simple Summary

Over the last century, genetic diversity in the cattle species has been affected by the replacement of many local, dual-purpose breeds with a few specialized, high-output dairy breeds. This replacement caused a sharp decline in the population size of local breeds. In the Netherlands, the local Dutch Friesian breed has gradually been replaced by the Holstein Friesian. This resulted in a rapid decrease in numbers of the Dutch Friesian breed with an associated risk of loss of genetic diversity due to drift. The objective of this study was to investigate genome-wide genetic diversity between a group of historic and recent Dutch Friesian bulls and a group of recently used Holstein Friesian bulls. Our findings showed that a large amount of diversity is shared between the three groups, but each of them has some unique genetic identity (12% of the single nucleotide polymorphism were group-specific). The genetic diversity of the Dutch Friesians reduced over time, but this did not lead to higher inbreeding levels—especially, inbreeding due to recent ancestors has not increased. Genetically, the recent Dutch Friesians were slightly more different from Holstein Friesians than the historic Dutch Friesians. Our results also highlighted the presence of several genomic regions that differentiated between the groups.

Abstract

Over the last century, genetic diversity in many cattle breeds has been affected by the replacement of traditional local breeds with just a few milk-producing breeds. In the Netherlands, the local Dutch Friesian breed (DF) has gradually been replaced by the Holstein Friesian breed (HF). The objective of this study was to investigate genome-wide genetic diversity between a group of historically and recently used DF bulls and a group of recently used HF bulls. Genetic material of 12 historic (hDF), 12 recent DF bulls (rDF), and 12 recent HF bulls (rHF) in the Netherlands was sequenced. Based on the genomic information, different parameters—e.g., allele frequencies, inbreeding coefficient, and runs of homozygosity (ROH)—were calculated. Our findings showed that a large amount of diversity is shared between the three groups, but each of them has a unique genetic identity (12% of the single nucleotide polymorphisms were group-specific). The rDF is slightly more diverged from rHF than hDF. The inbreeding coefficient based on runs of homozygosity (Froh) was higher for rDF (0.24) than for hDF (0.17) or rHF (0.13). Our results also displayed the presence of several genomic regions that differentiated between the groups. In addition, thirteen, forty-five, and six ROH islands were identified in hDF, rDF, and rHF, respectively. The genetic diversity of the DF breed reduced over time, but this did not lead to higher inbreeding levels—especially, inbreeding due to recent ancestors was not increased.

Key words: genetic diversity; Dutch Friesian; Holstein Friesian; cattle breeds; WGS

4.1 Introduction

Over the last century, genetic diversity in many European national cattle populations has been affected by replacement of traditional local breeds with just a few specialized milk producing breeds, e.g., the Holstein Friesian, Brown Swiss, and Jersey. The predominant use of these breeds caused a sharp decline in the population size of local dual-purpose breeds (Marchitelli and Consortium 2006; Medugorac et al. 2009). Although less productive under intense production conditions, these local breeds may carry alleles that enabled them to adapt to local conditions. Moscarelli et al. (2021) reported the presence of several genomic regions that vary between original and modern Brown cattle populations, in line with their different breeding histories. Selection and genetic drift will both have contributed to the genetic differentiation between original and modern breeds. Therefore, local breeds might represent an important genetic resource to facilitate animal breeding when changes occur in production systems and market requirements.

The change in use from local breeds to specialized breeds has been observed in many industrialized European countries including the Netherlands (Oldenbroek 2007; Hiemstra et al. 2010). Until 1975, the Dutch Friesian cattle (DF) dominated the Dutch national cattle population (76%) (van Breukelen et al. 2019; Doekes 2020). By the late nineteenth century, the Dutch Friesians were internationally known as exceptionally productive dairy cattle. American dairy farmers imported them in the 1870s and 1880s for this reason (Theunissen 2012). In the United States, where their progeny became known as Holstein Friesians, the farmers continued to breed them as high-yielding dairy cows (Theunissen 2012). In the Netherlands, there was emphasis on conformation and beef production in addition to milk production, because the Dutch Friesians were kept as a dual-purpose breed. Since the 1960s and 1970s, Holstein Friesians (bulls, semen, and embryos)—descendants of the original Dutch Friesian cattle—were imported from the United States into the Netherlands and used to improve the genetic ability for milk production. Consequently, DF in the Netherlands has gradually been replaced by Holstein Friesian (HF) during the past decades. Currently, more than 90% of the Dutch dairy cattle population consists of HF (Maurice - Van Eijndhoven 2014; van Breukelen et al. 2019). This upgrading process resulted in a rapid decrease in numbers and, therefore, a potential loss in genetic diversity due to drift in the DF breed (Fimland and Oldenbroek 2007). Since the beginning of the 1960s, genetic material from Dutch local cattle breeds and later from the HF as well has been collected and stored in the Dutch gene bank. Stored material contains genetic diversity of the breed at the time of sampling, which may

include diversity that, since then, has been lost in situ due to selection and genetic drift.

Currently, single nucleotide polymorphism (SNP) chips are available for the majority of livestock species, targeting genetic variants widely spread along their entire genomes. Importantly, SNPs detected in commercial breeds were selected and used to design the chips leading to some ascertainment bias when using these chips in studies with local breeds (Perez-Enciso et al. 2015). The latest advances and increasing economic accessibility of whole-genome sequencing (WGS) brings new perspectives exploring the genetic information of local breeds. Unlike SNP chips, WGS is the complete genome sequence containing all polymorphisms present on the genome. Thus, WGS does not have the problem of ascertainment biases. Another advantage of WGS is that it contains information on rare variants (Eynard et al. 2016) and, additionally, maps genomic regions highly affected by selection pressure (Eusebi et al. 2019). WGS enables the estimation of relationships between individuals more accurately because it is based on both common and rare variants. Furthermore, WGS has information on common variants in local breeds, which might be rare or absent in specialized dairy breeds, such as Holstein Friesian. However, to date, much of the effort has been devoted to dominant commercial breeds, with local breeds rarely studied. Furthermore, changes to genetic diversity in breeds over time is also rarely studied.

The objective of this study was to investigate genome-wide genetic diversity and loss of alleles between three groups of bulls, chosen from the historic (1961–1989) and recent (2003–2015) DF population and the recent HF (1998–2014) population. Differences in allele frequencies and in homozygosity will provide insights into the mechanisms underlying their genomic differences caused by selection or a sharp decrease in the number of breeding animals.

4.2 Materials and Methods

4.2.1 Animals

Genetic material from purebred Dutch Friesian animals born from 1961 onwards has been preserved by the Centre for Genetic Resources, the Netherlands (CGN) of Wageningen University and Research, i.e., the Dutch Gene bank (https://www.genebankdata.cgn.wur.nl, accessed on 27 July 2021). From this genetic material, a group of 12 historic (1961–1989; hDF) and a group of 12 recent (2003–2015; rDF) Dutch Friesian bulls were sequenced. The animals were selected based on their year of birth by taking the oldest and youngest DF sires, while avoiding

closely related animals in the selection. Furthermore, sequence data from a group of 12 recently used Holstein Friesian bulls in the Netherlands (1998–2014; rHF) were available for this study. These bulls were a selection of unrelated animals born in different years and sequenced for a project in 2017 on efficiency and health indices of the breeding company in dairy cattle CRV.

4.2.2 Short Read Sequencing Mapping and Variant Calling

DNA was isolated from sperm using the Echolution Sperm DNA kit (BioEcho Life Science GmBH, Köln, Germany). Library preparation and sequencing of the DF animals were performed at the Institute national de la Recherche Agronomique (INRA), France, following their established protocols. Library preparation and sequencing of the rHF animals were performed at BGI, China, following their established protocols. Paired-end sequencing was performed on the Illumina HiSeq platform. All animals were sequenced with short reads at 10× coverage. We followed the 1000 Bull Genomes Project Run 7 guideline (1000 bulls GATK fastq to GVCF guidelines; version: 18 June 2018) to process the raw sequence data into both binary alignment map (BAM) and genomic variant call format (GVCF) files (Hayes and Daetwyler 2019). A per-base sequence quality, for the raw sequence reads, was examined using the fastQC software (version: 0.11.7) (Andrews 2010). The reads were trimmed and filtered using Trimommatic (version: 0.38) (Bolger et al. 2014) and then mapped against the bovine reference genome ARS-UCD1.2_Btau5.0.1Y (version: 8 May 2018) using the Burrows–Wheeler Aligner (BWA; version: 0.7.17) (Li and Durbin 2009). Samtools (version: 1.8) (Li et al. 2009) was used to sort the BAM files and create index files. Polymerase chain reaction (PCR) duplicates were identified using the 'MarkDuplicates' function of Picard (version: 2.18.2) software (http://broadinstitute.github.io/picard, accessed on 27 July 2021). Base quality recalibration (BQSR) was performed with 'BaseRecalibrator' and 'PrintReads' of the Genome Analysis Toolkit (GATK; version: 3.8-1-0-gf15c1c3ef). The known variants file (ARS1.2PlusY_BQSR.vcf.gz; version: 15 June 2018) generated by the 1000 Bull Genomes Project was used to mask out positions with known variation to avoid confusing real variation with errors. The before/after BQSR reports were checked using 'AnalyzeCovariates' to ensure that base quality scores were corrected as expected. SNPs were called using the GATK 'HaplotypeCaller' with '-ERC GVCF' option. The rate of genome alignment and average sequencing depth were determined with Qualimap (version 2.2.1) software (Okonechnikov et al. 2016). The 'GenotypeGVCFs' arguments of GATK were used to identify variants simultaneously in all samples. 'VariantRecalibration' and 'ApplyRecalibration' were used to produce filtering information for SNPs. The process has been shown to outperform the 'hard'

filtering of variants (Pirooznia et al. 2014). For the recalibration steps, the truth and training datasets described by Jagt et al. (2018) were used, replacing the Run6 datasets of the 1000 bulls by datasets of Run 7.

The variants were called across all three groups combined. Only biallelic SNPs were kept, and filtration for Minor Allele Frequency was not applied at this stage. Per group (hDF, rDF, and rHF), a maximum of 2 out of 12 animals were allowed to have a missing value per SNP. These criteria resulted in a total of 10,780,681 SNPs, genotyped in all three groups, for further analysis.

4.2.3 Group Structure and Identification of Group-Specific SNPs

To identify group-specific SNPs, each SNP that passed the applied filtering criteria was analysed according to the information about the three groups. An allele was labelled as group-specific if it was only present in one of the three groups and not detected in any of the other two, also called private allele (Ramos et al. 2011).

To explore the genetic distance between animals of the three groups, a Principal Component Analysis (PCA) was performed using the --pca function in PLINK (version: 1.90) (Purcell et al. 2007). The graphical representation was depicted using the statistical R software (http://www.R-project.org/, accessed on 4 November 2021).

4.2.4 Genetic Diversity Parameters

Various parameters were used to estimate genetic diversity within the groups: Observed (Ho) and Expected Heterozygosity (He) and Minor Allele Frequency (MAF). Observed and Expected Heterozygosity were calculated using VCFtools (version 0.1.13) (--het) (Danecek et al. 2011). The MAFs were calculated using the—freq option in PLINK.

4.2.5 Selection Signature Analysis

The fixation index (F_{ST}) was used to characterize the differentiation between the groups, i.e., to identify selection signatures. The pairwise estimates of F_{ST} among the three groups (hDF-rDF, hDF-rHF, and rDF-rHF) were calculated using VCFtools with the Weir and Cockerham approach (--weir-fst-pop) (Weir and Cockerham 1984). Windows of SNPs were used to minimize the stochastic effect of a single SNP. The F_{ST} values were averaged across 40-kb windows, with a sliding frame of 20 kb at a time. The parameters for the VCFtools program were "—fst-window-size 40,000--fst-window-step 20,000", following Rafiepour et al. (2021). To normalize the mean

F_{ST} values, Z-transformation was performed $ZF_{ST}=\frac{F_{ST}-\mu F_{ST}}{aF_{ST}}$, where F_{ST} is mean F_{ST} in a window, μF_{ST} is an average F_{ST} over all windows, and αF_{ST} is a standard deviation of F_{ST} values of all windows tested for a given comparison (Rubin et al. 2010; Turner 2014). The ZF_{ST} was visualized in the form of a Manhattan plot by the R package 'qqman' (version 0.1.8) (Turner 2014). Candidate genomic regions under selection were defined as regions where the ZF_{ST} value > 8. To reduce the number of false positives, windows with less than 5 SNPs were removed.

4.2.6 Measure of Runs of Homozygosity

Runs of homozygosity (ROHs) were calculated to identify contiguous regions of the genome where an animal is homozygous across sites. ROHs were calculated individually using PLINK with adjusted parameters: --homozyg --homozyg-window-snp 50 --homozyg-snp 50 --homozyg-kb 300 --homozyg-density 50 --homozyg-gap 1000 --homozyg-window-missing 5 --homozyg-window-threshold 0.05 --homozyg-window-het 3, following Cheng et al. (2020). No linkage disequilibrium (LD)-based pruning was performed before calculating ROHs. Individual degree of inbreeding based on ROH analysis (Froh), genomewide as well as chromosome-wide, was calculated using the function Froh_inbreeding of the R package 'detectRUNS' (version 0.9.6) (Biscarini et al. 2018). In addition, Froh (Froh > 2 Mb, Froh > 4 Mb, Froh > 8 Mb, and Froh > 16 Mb) derived from ROHs of different length (>2, >4, >8, and >16) were calculated.

4.2.7 Runs of Homozygosity Islands

To identify genomic regions most commonly associated with ROH, i.e., ROH Islands, the percentage of the occurrences of a SNP in ROH was calculated by counting the number of times the SNP was detected in those ROHs across animals (Gorssen et al. 2020) within each group using the R package 'detectRUNS'. The most common runs were retrieved using the function 'tableRuns' of the package 'detectRUNS' with a threshold value of 0.8. This means that a ROH has to be present in at least 80% (10 of the 12 animals) of each group hDF, rDF, and rHF to be included in a ROH island.

4.3. Results

4.3.1. SNP Distribution

Among the set of 10,780,681 SNPs, 11.98% were identified as putatively group-specific (Figure 4.1), indicating that the genotype of one of the alleles was present in only one of the three groups (hDF, rDF, and rHF). SNPs specific for rHF were the most abundant (6.69%), while rDF displayed the lowest number of group-specific SNPs

(2.14%). It is notable that the percentage of group-specific SNPs for the DF (hDF and rDF) groups was 7.82%, which is slightly higher as for rHF (6.69%). Over 77% of the SNPs occurred in all three groups.

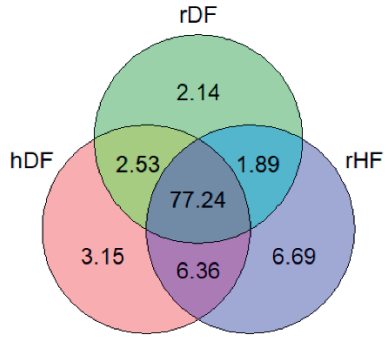

Figure 4.1. Venn diagram showing percentage of shared and group-specific variants in each group. Abbreviations: hDF = historic Dutch Friesian; rDF = recent Dutch Friesian; rHF = recent Holstein Friesian.

The genetic structure of the three groups assessed with the first three principal components of PCA accounted for 13.5% (PC1), 6.4% (PC2), and 5.6% (PC3) of the total variation (Figure 4.2). The first principal component (PC1) distinguished hDF and rDF from rHF. The hDF group differentiated across the second principal component (PC2), while rDF grouped more together, although overlapping with hDF. One rDF animal was positioned away from the other rDF animals along PC2. The third PC distinguished variation within the rHF, indicating one sire more distantly from the others.

4.3.2. Genetic Diversity Parameters

Minor Allele Frequency together with Observed and Expected Heterozygosity (Ho and He), were used to determine the levels of genetic variability in the three groups (Table 4.1). The average MAF and He were almost similar for the three groups (MAF: 0.16; He: 0.25). For all three groups, Ho was lower than He. There were small differences in Ho between three groups: rHF had the lowest Ho value (0.19) and rDF had the highest Ho value (0.20).

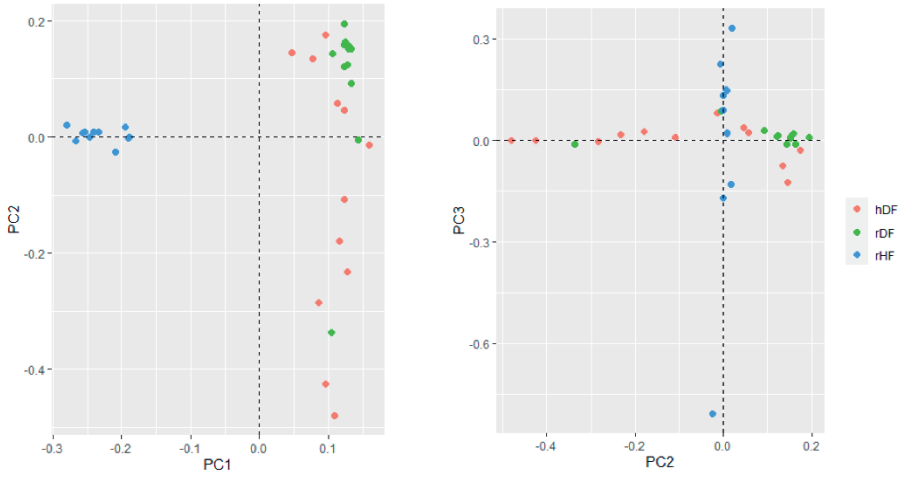

Figure 4.2. Genetic relationships based on PCA between the three groups. Abbreviations: hDF = historic Dutch Friesian; rDF = recent Dutch Friesian; rHF = recent Holstein Friesian.

Table 4.1. Genetic diversity parameters of within-group diversity of hDF, rDF, and rHF (mean ± standard deviation).

Group	Abbreviation	MAF	Ho	He
Historic Dutch Friesian	hDF	0.165 ± 0.152[a]	0.195 ± 0.025	0.250 ± 0.0005[a]
Recent Dutch Friesian	rDF	0.164 ± 0.155[b]	0.201 ± 0.015	0.250 ± 0.0003[a]
Recent Holstein Friesian	rHF	0.161 ± 0.153[c]	0.188 ± 0.035	0.249 ± 0.0016[b]

[a, b, c] Different letters within a column indicates significant differences at $p < 0.05$.

Genetic differentiation among the three groups ranged from low to moderate, as indicated by the weighted pairwise F_{ST} values that ranged from 0.01 to 0.11 (Table 4.2). Recent and historic DF were genetically very similar, and somewhat different from rHF.

Table 4.2. Estimated pairwise F_{ST} (fixation index) as a measure of genetic differentiation between the three groups (Weir and Cockerham mean F_{ST}, above diagonal; Weir and Cockerham weighted F_{ST}, below diagonal).

	hDF	rDF	rHF
hDF	-	0.0005	0.0624
rDF	0.0100	-	0.0719
rHF	0.0978	0.1105	-

Abbreviations: hDF = historic Dutch Friesian; rDF= recent Dutch Friesian; rHF= recent Holstein Friesian.

4.3.3. Genomic Inbreeding Coefficients

The inbreeding coefficient derived from ROH (Froh) in different length categories differentiated past and recent inbreeding (Table 4.3). Recent DF tended to have a larger fraction of the genome covered by ROH compared with hDF and rHF. The general average inbreeding coefficient was significantly higher for rDF (0.24) than rHF (0.13) ($p < 0.05$). The level of ancient inbreeding reached 0.05–0.13 (Froh > 2 Mb) with rDF having the highest level, whereas the recent inbreeding load was 0.01 (Froh > 16 Mb) for all three groups. So, Froh decreased as the minimum length of the ROH increased.

Table 4.3. Mean and standard deviation of inbreeding coefficients (Froh) calculated from runs of homozygosity (ROH) with minimum length of 2 (ROH > 2), 4 (ROH > 4), 8 (ROH > 8), and 16 (ROH > 16) Mb for the three groups. Between brackets is the number of animals in the classes.

Group	General Mean	Froh			
		ROH > 2 Mb	ROH > 4 Mb	ROH > 8 Mb	ROH > 16 Mb
Historic Dutch Friesian	0.169 ± 0.095[ab] (12)	0.081 ± 0.061[ab] (11)	0.058 ± 0.034 (9)	0.019 ± 0.012 (9)	0.010 ± 0.005 (3)
Recent Dutch Friesian	0.243 ± 0.062[a] (12)	0.132 ± 0.066[a] (12)	0.078 ± 0.055 (12)	0.035 ± 0.028 (11)	0.011 ± 0.007 (5)
Recent Holstein Friesian	0.130 ± 0.067[b] (12)	0.047 ± 0.036[b] (10)	0.031 ± 0.022 (7)	0.019 ± 0.013 (5)	0.012 ± 0.005 (2)

[a, b] Different letters within a column indicates significant differences at $p < 0.05$.

The mean Froh values for each chromosome followed the same pattern as those computed for the whole genome, but there was variation across the chromosomes (Additional file 3, Figure S1). For most chromosomes, the mean Froh was highest for rDF and lowest for rHF. Only chromosome 25 showed lower values for rDF compared with hDF and rHF. The chromosomal Froh variability within groups was high.

4.3.4. Measure of Runs of Homozygosity

The number and length of ROHs differed among animals and across groups (Table 4.4). The rDF group had the highest number of ROHs (513), whilst rHF had the lowest number (424). Additionally, rDF had the highest average length of ROHs (1.19 Mb) and rHF the lowest (0.73 Mb). Variation existed in the distribution of the various ROH length classes, but a common pattern was observed across the groups (Figure 4.3, Figure S1). The majority of ROH segments (~85 to 95%) is found in the length class 0 to 2 Mb for all three groups. The number of ROHs was the highest for rDF and the lowest for rHF in all classes, except class >16 Mb. The range of number of ROHs was more variable in the hDF and rHF groups in comparison with a more even number of ROHs for rDF group.

Table 4.4. Summary of specific regions of homozygosity (ROHs) in the three groups.

Group	# Animals	Number of ROH	Total ROH Length (Mb)	Average ROH Length (Mb)
		Mean ± sd	Range	Mean ± sd
hDF	12	449.75 ± 180.96	132–745	421.04 ± 235.43[ab]
rDF	12	513.50 ± 44.14	421–570	603.54 ± 154.92[a]
rHF	12	424.00 ± 161.48	75–653	323.92 ± 166.72[b]

[a, b] Different letters within a column indicates significant differences at $p < 0.05$.

Abbreviations: hDF = historic Dutch Friesian; rDF = recent Dutch Friesian; rHF = recent Holstein Friesian.

4.3.5. Genome-wide Selection Signature Analysis

To identify the differentiated genomic regions among the groups, the Z-transformed F_{ST} (ZF_{ST}) values based on SNPs in 40-kb sliding windows with 20-kb steps were calculated. The ZF_{ST} varied markedly across the genome in all three comparisons (hDF-rDF, hDF-rHF, and rDF-rHF) (Figure 4.4). We identified highly differentiated genomic regions ($ZF_{ST} > 8$) across autosomal chromosomes, i.e., thirty-eight for hDF versus rDF, nine for hDF versus rHF, and seven for rDF versus rHF (Figure 4.4, Additional file 1: Table S1). For the comparisons between hDF and rDF, the strongest differentiated genomic regions were detected on BTA1 (96.50–96.58 Mb and 98.80–98.86 Mb) and BTA2 (72.54–72.62 and 75.80–75.86 Mb). For hDF–rHF comparison, the strongest differentiated region was detected on BTA4 (44.66–44.72 Mb). In the case of the rDF versus rHF group, the strongest differentiated regions were located

on BTA1 (101.26–101.30 Mb), BTA16 (9.88–9.92 Mb), BTA20 (28.84–28.88 MB), BTA22 (three regions between 52.80–52.88 Mb), and BTA24 (44.16–44.20 Mb).

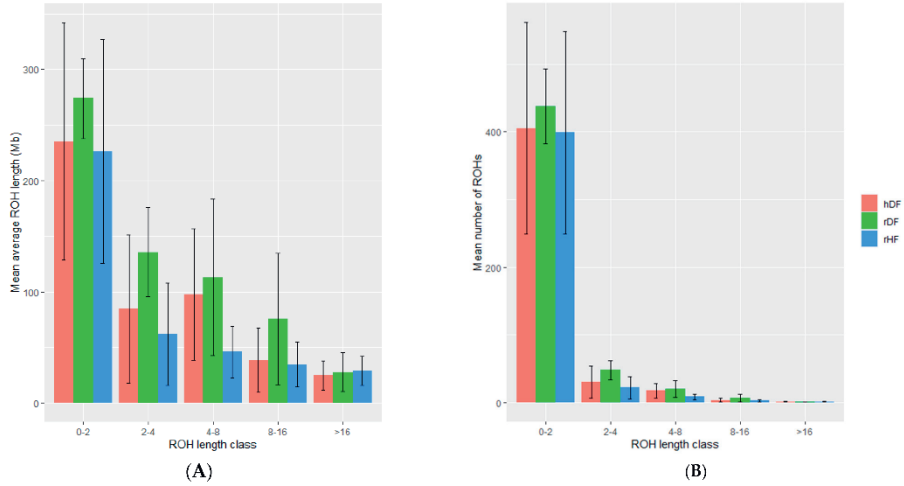

Figure 4.3. The mean and standard deviation of the average length of runs of homozygosity (ROH) (A) and mean number of ROH within each ROH length class (B). Abbreviations: hDF = historic Dutch Friesian; rDF = recent Dutch Friesian; rHF = recent Holstein Friesian.

4.3.6. Runs of Homozygosity Islands

The genomic distribution of ROH islands was nonuniform across chromosomes, regardless of the group (Additional file 2: Table S2). Differences in the segments of ROH islands on the chromosomes were identified between the three groups. In total, we identified thirteen, forty-five, and six ROH islands for hDF, rDF, and rHF, respectively. Almost all ROH islands found in hDF overlapped with ROH islands found in rDF. No genomic regions were common to all the three groups. Only one of the ROH islands identified ((BTA1: 101.26–101.30 Mb in rDF) overlapped with the genomic regions identified using pairwise F_{ST} (hDF vs. rDF and rDF vs. rHF).

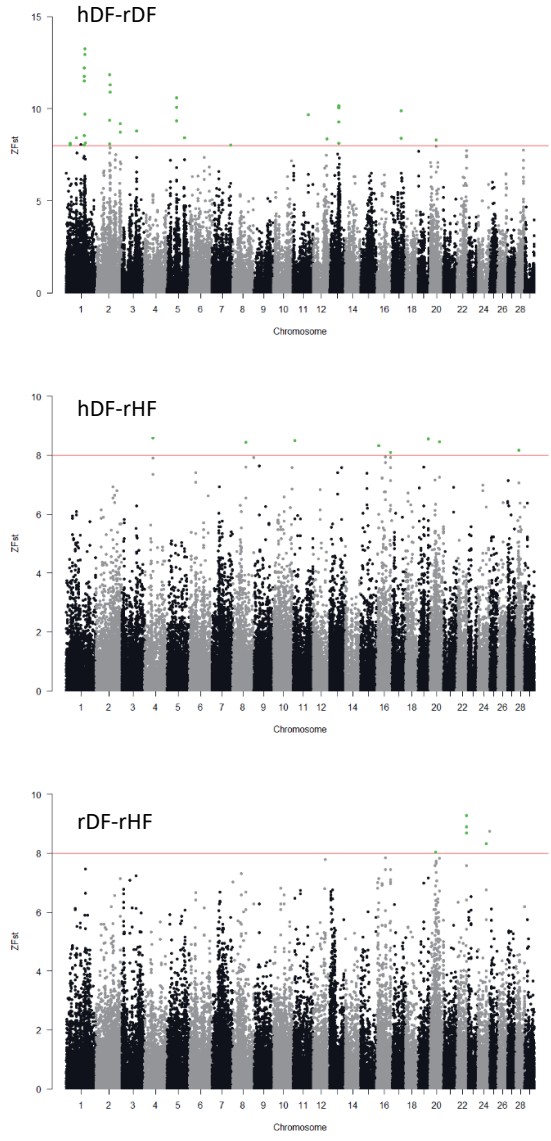

Figure 4.4. Manhattan plots of Z-transformed fixation index (ZF_{ST}) across all autosomes. The ZF_{ST} values were calculated for each sliding 40-kb window with steps of 20 kb across all autosomes. The solid red line indicates ZF_{ST} values > 8; differentiated genomic regions (ZF_{ST} values > 8 and the number of SNPs in the region > 5) are highlight green. Abbreviations: hDF = historic Dutch Friesian; rDF = recent Dutch Friesian; rHF = recent Holstein Friesian.

4.4. Discussion

4.4.1. General

In this study, we investigated genome-wide genetic diversity within and between groups of historic and recent DF bulls, and a group of recent HF bulls. In the Netherlands, local, dual-purpose cattle breeds, including the DF breed, have gradually been replaced by the specialized dairy breed HF during the past decades. This has caused a decline in the population size of local breeds and potentially a loss of genetic diversity. The historic Dutch Friesian animals used in this study were born between 1961–1989, when the population size of DF was still large. The recent DF animals were born between 2003 and 2015. In the approximately 5–10 generations between hDF and rDF, the population size (number of adult cows) of DF has declined significantly, from 629,410 in 1970 to 3153 in 2017 (van Breukelen et al. 2019). DF is now classified as being at risk (https://www.fao.org/dad-is; accessed on 15 October 2021).

4.4.2. Divergence between Groups

Our findings indicate that a large amount of diversity is common to the three groups. A high percentage of shared SNPs was found for the 3 groups, which is expected since all groups descend from the same ancestors. The founders of the Holstein Friesian breed originated from the Dutch Friesian breed (Felius et al. 2011). Furthermore, all three groups have a small number of group-specific SNPs (2–7%), indicating that each has some unique genetic identity. The PCA analysis displayed that the rDF group has diverged slightly from hDF group over the last approximately 5–10 generations, presumably as a result of genetic drift. Genetically, DF is distinct from HF, probably resulting from the selection of HF as a specialized dairy breed, whereas farmers aimed to maintain DF as a dual-purpose breed. The genetic distinction between DF and HF is in agreement with results reported by van Breukelen et al. (2019) and Hulsegge et al. (2019b). Likewise, a PCA analysis separated the Swedish Holstein Friesian breed from native Swedish cattle breeds (Upadhyay et al. 2019). In our study, the results of the PCA are confirmed by the pairwise F_{ST}. Although the DF and HF groups are selected for different purposes, we expected some similarities between them, and these are indicated in this study by the moderate average F_{ST} values (0.1). A similar pairwise F_{ST} value between DF and HF, based on SNP array data and a larger number of animals, was reported by Hulsegge et al. (2017).

4.4.3. Genetic Diversity within Groups

A decimation in numbers of a population is expected to reduce its genetic diversity and increase inbreeding levels. Indeed, diversity in the DF has reduced, e.g., rDF contains fewer specific alleles than hDF and, in the Principal Component Analysis, members of hDF are spread out across PCA2 while rDF animals cluster. However, based on manually checking the individual pedigrees of the hDF and rDF group, inbreeding levels have not increased. On the contrary, Ho, an indication of a lack inbreeding has increased in rDF. This is confirmed by the inbreeding level determined by pedigree for the whole population (CRV 2021b), which increased initially from around 3% in 1990 to above 5% in 2005 and decreased since then to under 4% in 2020; in 1970, the average inbreeding level was around 0% (Mill and Nauta 2010). One explanation for lower Observed Heterozygosity than expected is local inbreeding.

Manual checking of individual pedigrees indicated local breeding in historic DF. Breeders generally used their own bulls and certainly no bulls from other regions. In particular, there was a separation between Friesian bulls and bulls from the North Holland region. The animals from North Holland were, for example, slightly larger and produced more milk, but had less conformation than the animals from Friesland (Theunissen 2008). Currently, this separation has largely disappeared, and most animals have similar ancestry. However, one breeder went against the tide and eliminated from his stock all influence of an ancient Friesian bull who is ancestor to most other animals in the breed (pers. comm. Henk Sulkers). The deviating bull in rDF in PCA2 was bred by this breeder.

In the 1990s, when the DF rapidly declined, the DF herdbook initiated a strategy called fundament breeding to counter the loss of diversity. In this strategy, the breed is divided into several fundaments, each consisting of one or a few herds. Within each fundament, 4–5 own bulls are used and rotated over groups of cows so that inbreeding is postponed for at least three generations. Bulls should not be exchanged across fundaments to safeguard their genetic distinctness. This latter point was not strictly adhered to (Mill and Nauta 2010) and our data show no clear separation of fundaments in rDF; however, the strategy to postpone inbreeding seems to have worked. Although ROH levels are higher in rDF, this is due to ROH segments of shorter length only. These shorter segments indicate inbreeding due to ancestors further back in the pedigree.

In conclusion, although the diversity has reduced, this has not led to higher inbreeding levels—particularly, inbreeding due to recent ancestors has not increased. The policy of the breeding organization has influenced inbreeding levels but has not prevented the loss of some diversity, and diversity conserved in the gene bank has been lost from the live population. Therefore, to maintain and improve the genetic diversity in the current DF population, material from historic individuals present in the gene bank, should be used in the life population. Furthermore, the current strategy of rotating bulls within the fundaments should be maintained to limit the increase in inbreeding.

4.4.4. Differentiated Genomic Regions

The pairwise F_{ST} highlighted the presence of several genomic regions that differentiated between the groups. The F_{ST}-based approach does not directly indicate in which group selection is operating. In this study, the region with the highest ZF_{ST} values for the comparison hDF–rDF are observed from BTA1 (96.50–96.58 and 98.80–98.86 Mb) and BTA2 (72.54–72.62 and 75.80–75.86 Mb). In two of the four regions, no genes are located, while three genes are located in the other two regions: EIF5A2, RPL22L1, and ENSBTAG00000051422. EIF5A2 is associated with fertility traits. EIF5A2 has been reported as a candidate gene for age at sexual maturity in Indian Buffalo (Vohra et al. 2021) and for infertility in human (Christensen et al. 2005). RPL22L1 is also described as a candidate gene for age at sexual maturity in Indian Buffalo (Vohra et al. 2021). Furthermore, RPL22L1 is reported as associated gene in low-fertility buffalo bull spermatozoa (Paul et al. 2020). This gene is also mentioned as a candidate gene for birth weight in Holstein Friesian (Cole et al. 2014; Zaborski et al. 2014). This is in agreement with Estimated Breeding Values (EBV) for DF reported between 1980 and 2020, which indicate a decrease in fertility and birth weight (CRV 2021a). For the hDF–rHF comparison, we detected the strongest signal on BTA4 (44.66–44.72 Mb). In this region, the gene RELN is located. As stated by Cerri et al. (2012), RELN is involved in the regulation of pregnancy and lactation in Holstein cows. The latter is also reported by Lonergan et al. (2016). Furthermore, RELN affected aggressive behaviour in pigs (Terenina et al. 2012).

In the case of the rDF versus rHF group, we identified seven highly differentiated regions. Genes are only found in two regions on BTA22: ALS2CL, LRRC2, and TDGF1.

4.4.5. Runs of Homozygosity Detection and Distributions

Almost all ROH islands found in hDF partially overlapped with ROH islands found in rDF. These partially overlapped regions probably preserve segments in high

homozygosis, characteristic of the ancient selection of the population. In these regions, several known candidate genes, such as HCHD7, FBXO2, MAD2L2, MOS, and PLAG1, are mapped (Additional file 2: Table S2). These candidate genes are predominantly related to biological regulation (32.8% of the candidate genes) and metabolic processes (26.2% of the candidate genes (http://www.pantherdb.org/; accessed on 24 October 2021). Some traits are associated with these candidate genes as well. For example: the PLAG1-CHCHD7 region (BTA14: 23.33–3.38 Mb) is associated with stature; body size, including height; and weight in many cattle breeds (Nishimura et al. 2012; Bouwman et al. 2018; An et al. 2019; Smith et al. 2019; Zinovieva et al. 2020).

The largest ROH island in the rDF group was found on BTA7 between 50.03–50.86 Mb. This region seems to coincide with an ROH island reported for taurine and indicine cattle breeds by Sölkner et al. (2014) and for eight Chinese local cattle breeds reported by Xu et al. (2019). There are 15 candidate genes located within this ROH island: CTNNA1, DNAJC18, ECSCR, LRRTM2, MATR3, MZB1, PAIP2, PROB1, SIL1, SLC23A1, SMIM33, SNORA74, SPATA24, STING1, UBE2D2. Among them, we highlight the CTNNA1 gene, which has been associated with muscle development, skeletal muscle growth, and meat tenderness (Bongiorni et al. 2016; Jang et al. 2021).

The largest ROH island in the rHF group was found on BTA8 between 105.89–106.21 Mb. This region contains one gene: ASTN2. The ASTN2 gene has been related to carcass weight of cattle (Junior et al. 2016) and meat traits in pigs (Hlongwane et al. 2020).

4.4.6. Gene Bank

Our results revealed that the Dutch national gene bank has stored material containing genetic diversity that has been lost in-vivo by selection and drift. Gene bank collections have been shown to capture more diversity than some in situ populations thanks to periodic resampling (Blackburn 2012; Paiva et al. 2016; Boitard et al. 2021). It is also important that the gene bank pool stores genetic variation existing in the whole population. Van Breukelen et al. (2019) reported that within the DF populations there are fundamental breeding groups, which have a unique genetic diversity. For pigs, Hulsegge et al. (2019a) reported that merging of commercial Landrace lines has reduced the genetic diversity of the Landrace population in the Netherlands, although a large proportion of the original variation is maintained. This stresses the value of gene banks to record and preserve variation that is lost in the process of merging lines, even over short periods of time.

4.4.7. Limitation of the Study

The accuracy with which allele frequencies and, therefore, inbreeding is estimated will depend on the sample size and number of SNPs (Schmidt et al. 2021). In this study, we used 12 animals per group, which may have influenced the results. Although the sample size is small, it has previously been shown that a small sample size does accurately estimate population parameters when a large number of SNPs are used (Nazareno et al. 2017). Our study contains a large number of SNPs (n = 10,780,681). According to Willing et al. (2012) and Nazareno et al. (2017), F_{ST} can be accurately calculated based on small sample sizes (as small as n = 4 to 6) if the number of markers examined is large, i.e., larger than 1000. A small sample size can lead to poor population structure estimates, which affects the ability to differentiate between loci that were under selection and neutral population structure (Ahrens et al. 2018). However, in our study, 12 animals per group were used, in line with a previous study that suggested that detecting regions under selection with F_{ST} methods requires at least 10 samples (Willing et al. 2012).

4.5. Conclusions

Through the present study with WSG data, we have described the genetic differences between historic and recent Dutch Friesian groups, and a recent Holstein Friesian group. Our findings revealed that a large amount of diversity is shared in the three groups and each of the groups has a small number of group-specific SNPs. The two DF groups are genetically distinct from the HF group. rDF is slightly more diverged from rHF than hDF. We identified changes in the genetic composition of the DF population in the approximately 5–10 generations between the historic and the recent DF group. The genetic diversity has reduced and a more homogeneous group has emerged. Although diversity was reduced, this did not lead to higher inbreeding levels—especially, inbreeding due to recent ancestors has not increased.

Supplementary Materials

The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ani12030329/s1, Table S1: Differentiated genomic regions (ZF_{ST} > 8) across autosomal chromosomes for the hDF versus rDF population, the hDF versus rHF population and the rDF versus rHF population; Table S2: Genomic regions with the highest frequency of runs of homozygosity (ROH islands) occurrence across all animals per group; Figure S1. Distribution of inbreeding coefficients (Froh) based on runs of homozygosity (ROH) for each chromosome across groups.

Author Contributions

Conceptualization, I.H., K.O. and J.W.; methodology, I.H. and J.W.; formal analysis, I.H.; writing—original draft preparation, I.H.; writing—review and editing, I.H., K.O., A.B., R.V. and J.W. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by The Centre for Genetic Resources, the Netherlands (CGN) of Wageningen University and Research, funded by the Ministry of Agriculture, Nature and Food Quality, program 'Kennisbasis Dier', code KB-34-013-002 and program 'WOT', code WOT-03-003-056 and the IMAGE project which received funding from the European Union's Horizon 2020 Research and Innovation Programme under the grant agreement n° 677353.

Informed Consent Statement

Not applicable.

Data Availability Statement

The data presented in this study from The Centre for Genetic Resources, the Netherlands (CGN) of Wageningen University and Research will be online available before December 2022. Availability of the data of the "Melkveefonds" (HF sequences) are restricted to be used only for the current study, and thus, are not publicly available.

Acknowledgments

The Centre for Genetic Resources, the Netherlands (CGN) of Wageningen University and Research and the Dutch “Melkveefonds” are acknowledged for providing the data.

Conflicts of Interest

The authors declare no conflict of interest.

5

Selection of SNP from 50K and 777K arrays to predict breed of origin in cattle

B. Hulsegge[1], M. P. L. Calus[1], J. J. Windig[1], A. H. Hoving-Bolink[1], M. H. T. Maurice-van Eijndhoven[1], S. J. Hiemstra[2]

[1] Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, P.O. Box 65, 8200 AB, Lelystad, The Netherlands; [2]Centre for Genetic Resources, The Netherlands, Wageningen University and Research Centre, P.O. Box 65, 8200 AB, Lelystad, The Netherlands

Journal of Animal Science (2013) 91;5128–5134

Abstract

Reliable breed assignment can be performed with SNP. Currently, high density SNP chips are available with large numbers of SNP from which the most informative SNP can be selected for breed assignment. Several methods have been published to select the most informative SNP to distinguish among breeds. In this study, we evaluated Delta, Wright's F_{ST}, and Weir and Cockerham's F_{ST}, and extended these methods by adding a rule to avoid selection of sets of SNPs in high linkage disequilibrium (LD) providing the same information. The SNP that had a r^2 value >0.3 with any of the SNP already selected were discarded. The different selection methods were evaluated for both the 50K SNP and 777K Bovine BeadChip. Animals from four cattle breeds (989 Holstein Friesian, 97 Groningen White headed, 137 Meuse-Rhine-Yssel, and 64 Dutch Friesian) were genotyped. After editing 30,447 and 452,525 SNP were available for the 50K and 777K SNP chip, respectively. All selection methods showed that only a small set of SNPs is needed to differentiate among the four Dutch cattle breeds, whereas comparison of the selection methods showed only small differences. In general, the 777K performed marginally better than the 50K BeadChip, especially at higher confidence thresholds. The rule to avoid selection of SNP in high LD reduced the required number of SNP to achieve correct breed assignment. The Global Weir and Cockerham's F_{ST} performed marginally better than other selection methods. There was little overlap in the SNP selected from the two BeadChips, whereas the number of SNP selected was about the same.

Key words: assignment test, cattle breeds, high density SNP chips, SNP selection methods

5.1 Introduction

Known origin of individual animals is important in multiple aspects of animal production, such as breeding and tracing of animal products. Genetic markers can be used to infer the relationship among individuals. In several livestock species, including cattle, panels comprising tens of thousands of SNP are now available at affordable cost. While many SNP are needed to obtain commonalities among cattle breeds (de Roos et al. 2008), only a small set of SNP, if accurately chosen, is needed to differentiate among breeds (Wilkinson et al. 2011b). Currently, most studies on cattle breed identification are based on the BovineSNP50 BeadChip. The SNP with high minor allele frequency across cattle breeds were preferentially selected in the design of the BovineSNP50 BeadChip. Therefore, SNP data may suffer from an ascertainment bias. This ascertainment bias may be overcome by using whole genome sequence data or may perhaps be alleviated using a higher density SNP chip, such as the BovineHD BeadChip that contains 777K SNP. In our study, it is hypothesized that using a selected subset of high density SNP will lead to more accurate assignment of cattle to their known breed of origin because of less ascertainment bias. One important consequence of increased SNP density is that the linkage disequilibrium (LD) among SNP increases. Proposed SNP selection methods select most informative SNP (Ding et al. 2011; Wilkinson et al. 2011b) but do not account for the fact that some selected SNP may be highly correlated, due to LD, and therefore explain largely the same variance. As a result, the selected subset may contain redundant SNP.

The objective of this study was to compare different SNP selection methods, using the BovineSNP50 BeadChip and BovineHD 777K BeadChip, in terms of the minimum required number of informative SNP to differentiate among cattle breeds. Published SNP selection methods were extended with simple rules to avoid selection of redundant SNP.

5.2 Material and Methods

Animal Care and Use Committee approval was not obtained for this study because data were generated from multiple studies.

5.2.1 Genotypes and Allele Frequencies

Genotyped animals included 1,287 cows from The Netherlands, of four different breeds [989 Holstein Friesian (HFR), 97 Groningen White headed (G), 137 Meuse-Rhine-Yssel (MRY), and 64 Dutch Friesian (FH)]. All animals were 100% purebred.

Holstein Friesian animals were genotyped with a 50K chip and, together with 2,349 additional HFR cows from outside The Netherlands that were also genotyped with a 50k chip, imputed to high density (777K), based on a reference population consisting of 3,150 Holstein Friesian animals (1,366 males and 1,784 females). This imputation step was performed using Beagle (Browning and Browning 2009). The mean Beagle r^2 value, which reflects the accuracy of imputation, was 0.96 across the imputed loci, indicating that the imputation was highly accurate and that imputation will have influenced breed assignment results marginally, at most. During an initial quality check, 50K SNP with a call rate <95% were deleted. Animals from the other three breeds were genotyped with a high density chip (777K). The quality check for this dataset involved deleting SNP with a GenCall score ≤0.2, GenTrain score ≤0.55, and call rate ≤95% in one of the breeds. The GenCall and GenTrain score are quality measures on genotype calls from the genotyping assay (Illumina 2005). Within both datasets, SNP without known map position, as well as SNP located on the sex chromosomes, were deleted. After combining genotypes from all cows of the four breeds in this entire dataset, SNP with a minor allele frequency ≤0.5% and SNP in complete LD with a neighbouring SNP were deleted. After all these editing steps, 452,525 of the 777,962 SNP remained. Because one of the objectives of our study was to investigate whether the initial number of SNP affects the number of SNP required to predict breed of origin with high accuracy, a subset of those 452,525 SNP was selected. This subset contained SNP that are also included on the second version of the Illumina 50K SNP chip (Illumina Inc., San Diego, CA), being 30,447 in total, hereafter is termed “50K.” The term “777K” will be used for the 452,525 SNP.

5.2.2 Selection Methods to Find Most Informative Markers

To find the most informative markers to distinguish among breeds, the statistical selection methods described by Wilkinson et al. (2011b) were used, namely: Delta, Wright's F_{ST}, and Weir and Cockerham's F_{ST}. For Wright's F_{ST} and Weir and Cockerham's F_{ST}, the global F_{ST} among all breeds were calculated, as well as pairwise F_{ST} between each pair of breeds. All these parameters were calculated for each SNP, using data in the reference populations (i.e., data of animals with known breed of origin).

Delta, the absolute allele frequency difference observed between two populations, is the most commonly used measure of marker informativeness. For a biallelic marker, it is calculated as $|P_{Bi} - P_{Bj}|$, where P_{Bi} and P_{Bj} are the frequencies of allele B in the i^{th} and j^{th} population, respectively. Pairwise comparisons for each SNP were

averaged to obtain an overall estimate of the level of genetic information contained in each SNP.

Wright's F_{ST} statistic, the standardized variance in allele frequencies among populations, was calculated as $\boldsymbol{var(P_B)/\overline{P_B}(1-\overline{P_B})}$, where is $\boldsymbol{var(P_B)}$ the variance of the frequency of allele B across breeds and $\boldsymbol{\overline{P_B}}$ is the mean allele frequency of allele B across breeds.

The Weir and Cockerham's unbiased estimator (W&C) was calculated using an R script written by Chan (2012). This script estimates the variance components and fixation indices as described by Weir and Cockerham. The W&C's F_{ST} (Weir and Cockerham 1984) was calculated as

$$\widehat{\boldsymbol{\theta}} = \frac{\boldsymbol{a}}{\boldsymbol{a}+\boldsymbol{b}+\boldsymbol{c}}$$

where $\boldsymbol{a}$ = variance of allele frequencies among populations, $\boldsymbol{b}$ = variance of allele frequencies among individuals within populations, and $\boldsymbol{c}$ = variance of allele frequencies among gametes within individuals.

When applying SNP selection methods as described above, the possibility exists that pairs of SNPs that are selected are in strong LD, implying that one of those SNP will contribute very little to breed assignment, because it largely explains the same variation as the other selected SNP. To avoid selection of such redundant SNP for each measure, we included a rule that avoided selection of SNP with a r^2 value >0.3 (calculated according to Hill and Robertson (1968)), with any SNP that were already selected.

5.2.3 Individual Assignment

The individual assignment method of Paetkau et al. (Paetkau et al. 1995) is used in this study, because it is most frequently used in empirical studies (Wilkinson et al. 2011b).

The individual assignment analysis encompasses 3 steps. First, the probability that genotypes 0, 1, and 2 occur were calculated per locus (***j***) and breed (***k***), whereby 1 indicates heterozygote genotype and 0 and 2 indicate the 2 homozygotes. If $\mathbf{p}_{jk}$ is the frequency of the allele that homozygote genotype is coded as 2, then the probability of genotype 0 is calculated as $\mathbf{P}_{jk}\mathbf{(0) = (1 - p}_{jk}\mathbf{)^2}$; $\mathbf{P}_{jk}\mathbf{(1) = 2p}_{jk}\mathbf{(1 - p}_{jk}\mathbf{)}$, and $\mathbf{P}_{jk}\mathbf{(2) = p}_{jk}^{\mathbf{2}}$.

Because the next step involves calculation of the log of those probabilities and log(0) is not defined, allele frequencies with values of zero or one were, respectively, replaced by a value of 1 × 10–5 or 0.99999 (Wilkinson et al. 2011b). Next, for each genotype, the log-likelihoods for all breeds were calculated, using the probabilities from the first step as the sum of the log(10) probabilities for all selected loci. Finally, the log-likelihood ratios (LLR) were calculated by subtracting the log-likelihood of an individual being assigned to any of the breeds from the likelihood of it being assigned to another breed, resulting in three LLR values for each of the four breeds.

Four different confidence thresholds were applied as confidence levels of assignment precision: LLR > 0, LLR > 1, LLR > 2, and LLR > 3. The LLR > 0 implies that a genotype is more likely in one breed than another. The LLR > 1, LLR > 2, and LLR > 3 means that an observed multilocus genotype has to be 10, 100, or 1,000 times more likely in one breed than any other to be assigned. For details, see Wilkinson et al. (2011b). Breed assignment, following the different methods of SNP selection, followed when all of the three calculated LLR for a specific breed were greater than the selected confidence threshold. If the LLR was lower than the selected confidence threshold, the individual animal could not be definitely assigned to a specific breed. To enable this evaluation, the complete genotype dataset was split into two subsets: 80% of each breed were randomly selected to represent a sample of animals that have genotypes and breed known, hereafter referred to as reference data (n = 1,032), and the remaining 20% represented a sample of individual animals for which we wanted to predict their breed of origin (n = 255).

All analyses were performed using a combination of in-house R (www.r-project.org) and Perl (www.perl.org) scripts.

5.3 Results

5.3.1 Comparison of SNP Selection Methods

Distributions of calculated Wright's and W&C's F_{ST} statistics for individual SNPs were predominantly right skewed (Figure 5.1). The distributions of Delta estimates were more symmetric but were also slightly right skewed. Distributions of measures for 777K were almost similar in shape to the 50K distributions, indicating that ascertainment bias was similar for 50K and 777K. Summary statistics of the different selection methods of SNP informativeness are shown in Supplementary Table S5.1. The average level of Delta, Global Wright's F_{ST}, Pairwise Wright's F_{ST}, Global W&C's F_{ST}, and Pairwise W&C's F_{ST} were 0.17, 0.10, 0.06, 0.11, and 0.11, respectively. Average estimates were similar for 50K and 777K. The estimates of individual SNP

were highly correlated among the different methods, with Spearman rank correlation coefficients ranging from 0.78 between Delta and Global W&C's F_{ST} to 0.97 between Pairwise Wright's F_{ST} and Pairwise W&C's F_{ST}. The correlation coefficients among methods were comparable for 50K and 777K.

5.3.2 Removing Redundant SNP

Filtering the most informative SNP, by discarding SNP that had a $r^2 > 0.3$ with any of the SNP that were already selected, reduced the number of SNP necessary for correct assignment of individual animals (Table 5.1). This filtering had the most effect for the selection method Global W&C's F_{ST}, where the reduction ranged between 47 and 57 SNP, and 58 to 69 SNP for 50K and 777K, respectively. The reduction in the required number of SNP is found for the 50K, as well as for the 777K. However, the degree of reduction was larger for the 777K than for the 50K.

5.3.3 Breed Assignment of Individual Animals

Overall, 95% of the sample of individual animals (n = 255), whose breed of origin was predicted, were assigned correctly, using ≤37 SNP (Table 5.2). Differences in the percentage of individuals assigned correctly to their breed of origin using the different SNP selection methods were small. Global W&C's F_{ST} and Delta performed marginally better than other selection methods. Figure 5.2 shows overlap of the top 50 SNP selected by the methods Delta and Global W&C's F_{ST}, using both types of chips. There was almost no overlap between the 50K and 777K, as well as between Global W&C's F_{ST} and Delta. There were only 3 and 6 SNP selected by both chips (50K and 777K) for Delta and Global W&C's F_{ST}, respectively.

The number of SNP required for accurate breed assignment depends on the confidence level used for breed assignment. We used the confidence LLR > 0, LLR > 1, LLR > 2, and LLR > 3 (Figure S5.1). Correctly assigning 95% of the individual genotypes at LLR > 0 required ≤13 SNP, whereas LLR > 3 required ≤37 SNP (Table 5.2). The Global W&C's F_{ST} required the smallest number of SNP to reach 90%, 95%, and 100% correct assignment at the confidence values of LLR followed by Delta, whereas Pairwise Wright's F_{ST} performed the worst. The number of SNP necessary for correct assignment of individual animals differed very little between the 50K and 777K. The difference in the required number of SNP to reach 100% correct assignment at confidence threshold LLR > 0 between the 50K and 777K ranged from 1 to 8. In general, the 777K performed marginally better than the 50K, especially at higher confidence thresholds.

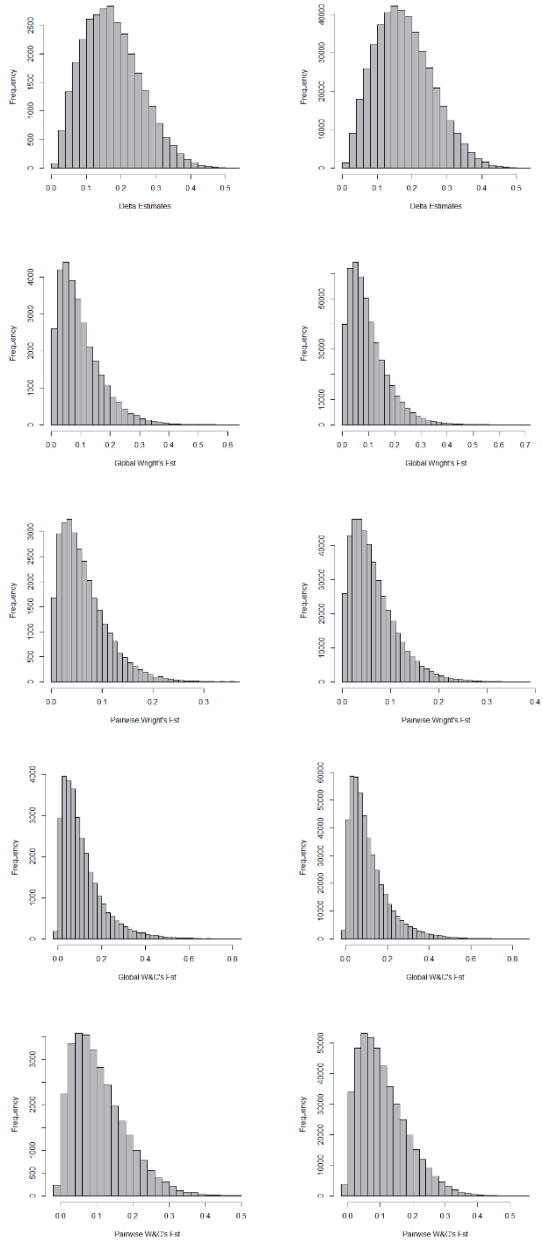

Figure 5.1. Distribution of estimates for each selection method. (Left estimates for the BovineSNP50 BeadChip (50K) and right for the BovineHD 777K BeadChip (777K)).

Table 5.1. Number of SNPs required to achieve 100% correct assignment for the 50K and 777K at the four confidence thresholds for each SNP selection method, with and without LD restriction ($r^2 > 0.3$).

	No LD restriction				LD restriction ($r^2 > 0.3$)			
	LLR0	**LLR1**	**LLR2**	**LLR3**	**LLR0**	**LLR1**	**LLR2**	**LLR3**
50K								
Delta	33	35	36	38	24	33	39	39
Global Wright's F_{ST}	38	53	57	60	22	30	33	34
Pairwise Wright's F_{ST}	30	35	37	41	29	32	38	42
Global W&C's F_{ST}	66	79	80	81	19	22	25	29
Pairwise W&C's F_{ST}	21	31	36	39	24	30	30	37
777K								
Delta	33	45	46	62	17	23	35	36
Global Wright's F_{ST}	63	76	86	87	24	38	46	52
Pairwise Wright's F_{ST}	57	57	59	62	28	29	37	42
Global W&C's F_{ST}	76	77	78	90	11	15	20	21
Pairwise W&C's F_{ST}	64	68	70	77	25	33	38	42

Abbreviations: LD = linkage disequilibrium; LLR0 = log-likelihood ratio > 0; LLR1 = log-likelihood ratio > 1; LLR2 = log-likelihood ratio > 2; LLR3 = log-likelihood ratio > 3.

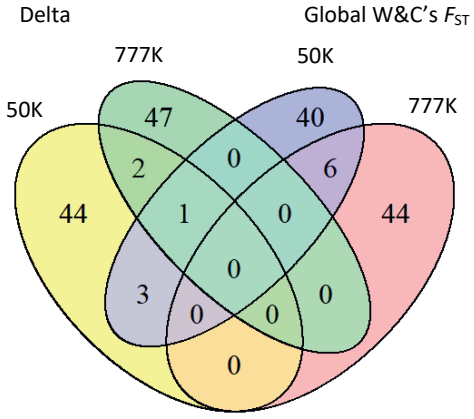

Figure 5.2. Overlap for the top 50 selected SNP between the Global Weir and Cockerham's (W&C's) F_{ST} and Delta selection methods. (50K = BovineSNP50 BeadChip; 777K = BovineHD 777K BeadChip).

Table 5.2. Number of SNP required to achieve 90%, 95%, and 100% correct assignment for the 50K and 777K at the 4 confidence thresholds for each SNP selection method.

		50K			777K		
		90%	**95%**	**100%**	**90%**	**95%**	**100%**
Delta							
	LLR0	7	11	24	5	8	17
	LLR1	13	19	33	10	13	26
	LLR2	20	22	39	13	17	35
	LLR3	24	27	39	17	20	36
Global Wright's F_{ST}							
	LLR0	8	10	22	7	8	24
	LLR1	12	14	30	17	20	38
	LLR2	16	19	33	21	30	46
	LLR3	20	22	35	28	37	52
Pairwise Wright's F_{ST}							
	LLR0	9	13	29	6	12	28
	LLR1	16	22	32	14	16	29
	LLR2	20	24	38	16	21	37
	LLR3	23	26	42	21	21	42
Global W&C's F_{ST}							
	LLR0	4	5	19	3	3	11
	LLR1	7	9	22	7	7	15
	LLR2	9	12	25	8	11	20
	LLR3	16	17	29	13	14	21
Pairwise W&C's F_{ST}							
	LLR0	7	10	24	2	8	25
	LLR1	13	15	30	11	12	33
	LLR2	15	15	30	12	15	38
	LLR3	18	19	37	19	21	42

Abbreviations: LLR0 = log-likelihood ratio > 0; LLR1 = log-likelihood ratio > 1; LLR2 = log-likelihood ratio > 2; LLR3 = log-likelihood ratio > 3; 50K = BovineSNP50 BeadChip; 777K = BovineHD 777K BeadChip.

The percentage of individuals from a known breed of origin allocated to each breed category was calculated using the Delta selection method, which is an efficient and simple method, and at the lowest confidence threshold (LLR > 0; Table 5.3).

Table 5.3. Percentage of individuals from a known breed of origin allocated to each breed category, using the Delta selection method at the lowest confidence threshold level (LLR > 0). Breed of origin is indicated in columns, whereas assigned breed is indicated in rows.

	50K				777K			
	G	**FH**	**MRY**	**HFR**	**G**	**FH**	**MRY**	**HFR**
5 SNPs								
G	100	8.3	0	0.5	89.4	8.3	0	2.0
FH	0	91.7	0	8.6	5.3	83.4	0	1.5
MRY	0	0	92.6	4.6	0	8.3	77.7	3.0
HFR	0	0	7.4	86.3	5.3	0	22.4	93.4
15 SNPs								
G	100	0	0	0	100	0	0	0.5
FH	0	100	0	1.5	0	100	0	1.0
MRY	0	0	96.3	1.0	0	0	100	0.5
HFR	0	0	3.7	97.5	0	0	0	98.0
25 SNPs								
G	100	0	0	0	100	0	0	0
FH	0	100	0	0	0	100	0	0
MRY	0	0	96.3	0	0	0	100	0
HFR	0	0	3.7	100	0	0	0	100
35 SNPs								
G	100	0	0	0	100	0	0	0
FH	0	100	0	0	0	100	0	0
MRY	0	0	100	0	0	0	100	0
HFR	0	0	0	100	0	0	0	100

Abbreviations: G = Groningen White headed; FH = Dutch Friesian; MRY = Meuse-Rhine-Yssel; HFR = Holstein Friesian; 50K = BovineSNP50 BeadChip; 777K = BovineHD 777K BeadChip.

The evaluation was performed considering 5, 15, 25, or 35 SNP. In the range of 15 to 30 selected SNP, the 777K performed marginally better than the 50K. Using five selected SNP, the 50K performed slightly better than the 777K for three out of four breeds. To achieve 100% correct assignment for all breeds, more SNP were needed using the 50K in comparison to using the 777K. Incorrectly assigned animals with the same breed of origin tended to be spread across the other breeds. One exception is MRY, where incorrectly assigned animals were always assigned to HFR.

5.4 Discussion

Numerous studies have shown that SNP can be used to identify the breed of origin of an individual (for cattle: e.g., Lewis et al. 2011; Pant et al. 2012). Several selection methods are available to determine which SNP panel contains most information to predict breed of origin (Rosenberg et al. 2003; Ding et al. 2011; Wilkinson et al. 2011b). In this study, we compared five different SNP selection methods, using the BovineSNP50 BeadChip (50K) or BovineHD BeadChip (777K), to investigate whether the minimum required number of informative SNP to differentiate among cattle breeds depends on the chip used and whether the SNP selection methods perform similar for different chips. Our results clearly show that using the 777K instead of the 50K chip only marginally increased the percentage of correct breed assignments, and this limited increase does not justify the additional costs of using the 777K instead of the 50K chip for the purpose of breed assignment.

In addition, the SNP selection methods were extended with a rule that the maximum allowed squared correlation among selected SNP was 0.3 to avoid selection of redundant SNP as a result of high LD between, generally, neighbouring SNP. This rule directly accounts for the overlap of the variance explained by two SNP and is therefore expected to be more efficient than avoiding selection of SNP with a small physical distance between them (e.g., Ding et al. 2011). The current results show that adding the rule of a maximum allowed squared correlation among selected SNP of 0.3 to the SNP selection methods reduced the required number of SNP for correct breed assignment, especially when using the BovineHD BeadChip, because the higher LD among SNP leads to more redundancy in the selected SNP. Most studies on cattle breed identification using SNP are based on the 50K SNP chip (e.g., Wilkinson et al. 2011b; Dimauro et al. 2013). During the development of this chip, SNP with high minor allele frequency across cattle breeds were preferentially selected in the design (Matukumalli et al. 2009). Therefore, SNP data may suffer from an ascertainment bias (Nielsen 2004; Albrechtsen et al. 2010; Wang and Nielsen 2012), which may also have an impact on the selection of SNP to differentiate among populations. The BovineHD BeadChip tended to give only marginally better results than the Bovine50SNP BeadChip and this indicates that the assumed larger ascertainment bias of the Bovine50SNP BeadChip hardly affected the results. Most likely, the selection methods are efficient enough to select a small set of SNPs from the Bovine50SNP BeadChip that differentiate the breeds with high accuracy, despite the ascertainment bias. In other words, the selection methods used in this study are

robust to the effect of ascertainment biases and adding SNP from the 777K only marginally improves results.

All selection methods showed that only a small set of SNPs is needed to differentiate among populations and there were little differences among the different selection methods (Table 5.2). The number of SNP necessary for correct assignment of individual genotypes differed very little between the Bovine SNP50 and BovineHD. At a confidence threshold of LLR > 0 and considering the level of 95% correct assignment, the required number of SNP ranged from 5 to 13 and from 3 to 12 for BovineSNP50 and BovineHD, respectively (Table 5.2). Developing a reduced SNP panel that is tailored to breed assignment lowers the animal genotyping cost and therefore would be more cost effective if animals are genotyped only for breed verification purposes.

Our findings indicated that all SNP selection methods performed well. The Global W&C's F_{ST} performed marginally better than other selection methods in the individual assignment of Dutch cattle breeds. However, different studies (e.g., Kersbergen et al. 2009; Ding et al. 2011; Wilkinson et al. 2011b) reported that using Global F_{ST}'s, including Global W&C's F_{ST}, is not optimal for selection of informative SNP when analysing >2 different breeds at the same time. Global F_{ST} will most likely select SNP that are specific for only the most distinct population.

Since the Delta selection method was a very useful estimator and it is easy to compute, as reported by Yang et al. (2005), we used this method to assign individuals to breed of origin. Using this selection method, the minimum number of SNP needed for accurate breed assignment (≥95%) was < 27 SNP, regardless of the chosen degree of confidence (LLR > 0, LLR > 1, LLR > 2, and LLR > 3). The number of SNP required to reach 95% correct assignment increased using higher confidence threshold levels, whereas the percentage of correct assignment decreased. This is in agreement with results reported by Wilkinson et al. (2011b). The number of SNP needed, according to our analyses, is less than what was reported by Wilkinson et al. (2011b). This is probably due to two differences between both studies. First, our analyses include only four breeds, whereas the analysis by Wilkinson et al. (2011b) included 17 breeds. Assignment precision is expected to decrease with an increase in number of populations included in the reference population, simply because there are more populations in which an animal erroneously can be assigned. Decreasing assignment success with an increase in number of populations is also observed using microsatellites (e.g., Talle et al. 2005). Second, in our analysis, the SNP selection

methods avoided selection of redundant SNP, which makes them more efficient as shown in our results.

The percentage of individual genotypes correctly assigned to their known breed of origin (Figure S5.1) shows a typical pattern across different numbers of SNP used for breed assignment. It consists of a rapid increase in correct assignment percentage at a low number of SNP, then plateaus and only a marginally higher percentage of individual genotypes could be assigned after a certain number of SNP were already selected. This pattern was observed for all selection methods and was also found by Ding et al. (2011) and Wilkinson et al. (2011b). Bjørnstad and Røed (2002) reported for crossbred animals that the percentage of individuals correctly assigned increased at a much slower rate compared with purebred animals. For all selection methods and confidence thresholds, 100% correct assignment was reached in our study using a maximum of 52 SNP. This means that the individual animals for which we predicted their breed of origin were genetically typical for their breed. As suggested by others (Bjornstad and Roed 2002; Talle et al. 2005; Dalvit et al. 2008), there is the possibility that some individual animals will never be assigned correctly, even when using a high number of SNP, because the considered breeds are too closely related or because some individuals are genetically atypical of their breeds. To avoid the latter from happening, it is important that the reference population reflects the complete range of genotypes in a breed. This implies that when setting up a reference population for breed verification purposes, especially when its size is limited, it is important to sample widely from the population and avoid sampling of closely related animals.

5.5 Conclusions

Although tens of thousands of SNP markers are now available, only a small set of SNPs, if accurately chosen, is needed to differentiate among the cattle breeds G, FH, MRY, and HFR, with high accuracy. Using the 777K instead of the 50K chip only marginally increased the percentage of correct breed assignments and this limited increase does not justify the additional costs of using the 777K instead of the 50K chip for the purpose of breed assignment. From both chips, the number of SNP selected was about the same. Avoiding selection of SNP in high LD reduced the required number of SNP for correct breed assignment. The selection methods investigated in this study showed very few differences. The Global W&C's F_{ST} performed marginally better than other selection methods in the individual assignment of individuals to four cattle breeds. We have shown that ≤37 SNP are

needed to successfully assign 95% of unknown cattle individuals to one of four cattle breeds.

Funding

This work was supported by the Centre for Genetic Resources, The Netherlands (CGN) funded by the Ministry of Economic Affairs, project “Kennisbasis Dier,” code KB-04-002-021. The text represents the author's views and does not necessarily represent a position of the Ministry, which will not be liable for use made of such information.

Supplementary Materials

Table S5.1. Summary statistics of different measures used for SNP selection.

Measure	Mean	Std Dev	Min	Max
50K				
Delta	0.1737	0.0820	0.0057	0.5209
Global Wright's F_{ST}	0.0977	0.0729	0.0001	0.6359
Pairwise Wright's F_{ST}	0.0639	0.0468	0.0001	0.3633
Global W&C F_{ST}	0.1103	0.0951	-0.0037	0.8370
Pairwise W&C F_{ST}	0.1067	0.0743	-0.0054	0.4839
777K				
Delta				
Global Wright's F_{ST}	0.1768	0.0823	0.0013	0.5361
Pairwise Wright's F_{ST}	0.0980	0.0733	0.0000	0.7070
Global W&C F_{ST}	0.0644	0.0473	0.0000	0.3857
Pairwise W&C F_{ST}	0.1104	0.0946	-0.0038	0.8646

Abbreviations: 50K = BovineSNP50 BeadChip; 777K = BovineHD 777K BeadChip.

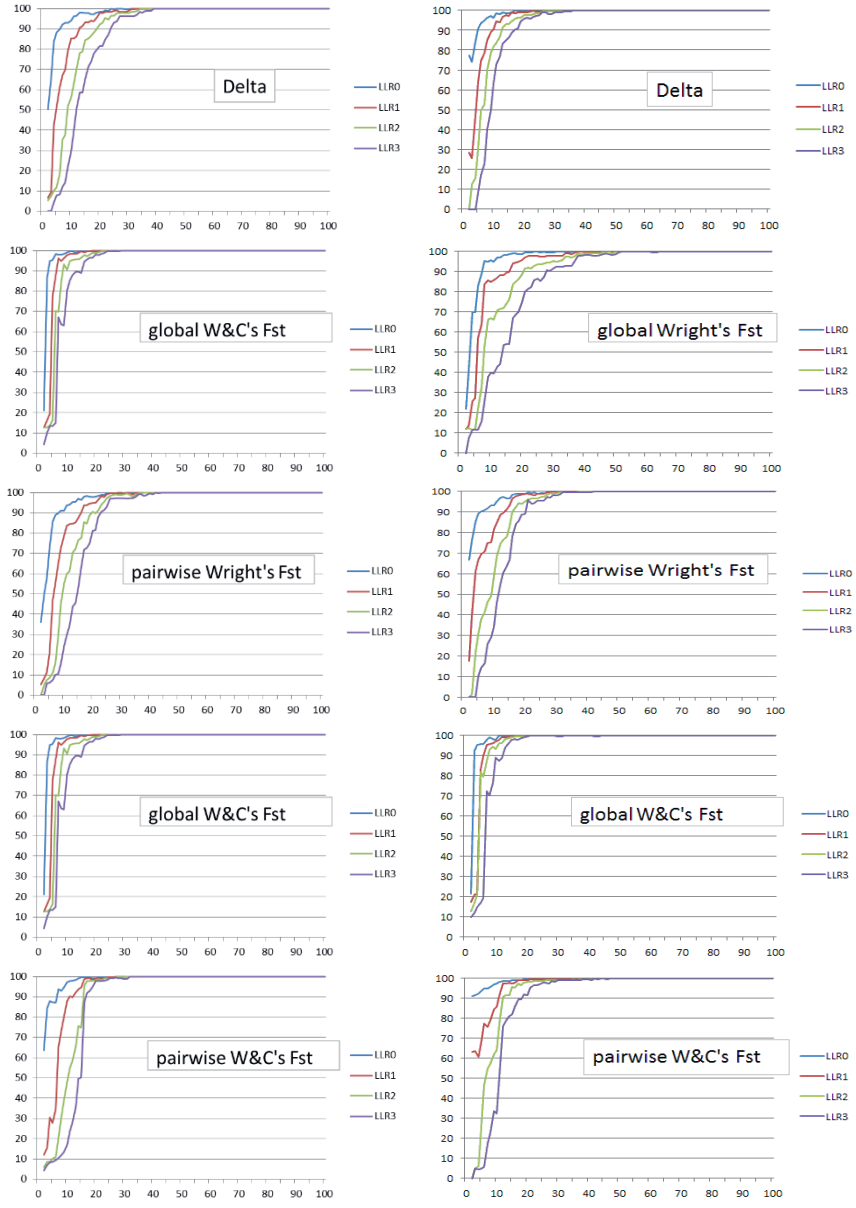

Figure S5.1. The percentage of correctly assigned individuals across different numbers of selected SNP. X axis values are numbers of SNP and Y axis values are correct assignment percentage of the individual genotypes . (Left for the 50k and right for the 777K).

Abbreviations: LLR0 = log-likelihood ratio > 0; LLR1 = log-likelihood ratio > 1; LLR2 = log-likelihood ratio > 2; LLR3 = log-likelihood ratio > 3.

6

Development of a genetic tool for determining breed purity of cattle

Ina Hulsegge[1,2], Mira Schoon[1], Jack Windig[1,2], Marjolein Neuteboom[3], Sipke Joost Hiemstra[1,2], Anouk Schurink[1,2]

[1] Animal Breeding and Genomics, Wageningen Livestock Research, P.O. Box 338, 6700 AH, Wageningen, the Netherlands; [2] Centre for Genetic Resources The Netherlands, PO Box 16, 6700 AA Wageningen, the Netherlands; [3] Stichting Zeldzame Huisdierrassen, Dreijenlaan 2, 6703 HA Wageningen, the Netherlands

Livestock Science (2019) 223:60-67

Highlights

- Only a small set of SNPs, when accurately chosen, was needed to differentiate among the Dutch local cattle breeds.
- The reference population of purebred animals showed genetic clusters that corresponded to their breed designations and its usefulness for assignment of future unknowns.
- A genetic test was developed to unequivocally determine the breed origin of animals without pedigree data.

Abstract

Breed registries have been established for livestock species to maintain the purity of breeds and to document the ancestry of animals. However, a significant number of animals are unregistered with no or incomplete pedigree data and uncertain ancestral breed origin. Although many local livestock breeds are "at risk" on the basis of the number of purebred breeding females in a breed registry, there is often also a reservoir of unregistered animals that may belong to the same breed. However, due to the missing pedigree it is not possible for breed societies or herdbooks to include those animals in their breeding program for purebred animals. A genetic test was developed to unequivocally determine the breed origin of cattle without pedigree data. Such a test will open up the possibility to incorporate animals without pedigree data in the breed registry that turn out to be purebred based on the test results. In this study we developed and validated such a test. Genotype data (50k SNP array) were used to compose reference populations for six local Dutch cattle breeds. The combination Principal Component Analysis and Random Forest was used to perform SNP selection. A total of 133 informative SNPs were selected to determine breed composition of individual animals. Overall, 82.0% of the animals in the test population are correctly assigned to the breed in question. For Dutch Red and White Friesian and Deep Red Cattle we suggest that if an animal has a percentage for its own breed <0.775 to use the combined percentage of two breeds (Deep Red Cattle with Meuse-Rhine-Yssel and Dutch Red and White Friesian with Dutch Friesian). Using this criterion 88.9% (104 out of 117) of the animals in the test population is correctly assigned.

The developed test was successful and will be implemented in practice to identify (partly) unregistered individuals as being purebred (or not) for one of the Dutch local cattle breeds.

Key words: Genetic test, Breed purity, Assignment, SNP, Cattle breeds

6.1 Introduction

Modern livestock production is dominated by global use of highly productive breeds, while many local breeds have become endangered. Nowadays, most of these local farm animal breeds are at risk of extinction on the basis of their small (effective) population sizes (www.fao.org/dad-is). Moreover, in numerically small populations inbreeding can increase rapidly and consequently genetic variation will be eroded. Breed registries have been established to maintain the purity of breeds and to document the ancestry of breeding animals, and to enable breed specific breeding programs. However, there is also a significant number of unregistered animals that have no or incomplete pedigree or ancestral breed composition data.

According to Regulation (EU) 2016/1012 on Animal Breeding (EU 2016b) this potential "reservoir" of animals without pedigree data cannot enter the main section of the herdbook. However, with reference to article 19 of the Regulation, Member States can decide to implement a specific derogation for the conservation or reconstruction of endangered breeds. Furthermore, in the event of disease outbreaks that could threaten the survival of local breeds, derogations are also allowed on the basis of the EU animal health legislation (EU 2016a). It allows competent authorities to take specific measures to protect purebred animals of local breeds.

Traditionally, the determination of purebred animals is derived from pedigree information. When pedigree information is lacking, alternatively, molecular markers can be used to estimate breed purity. In several livestock species, including cattle, tens of thousands of Single Nucleotide Polymorphisms (SNP) markers located across the whole genome are available (Matukumalli et al. 2009). The availability of genotypes of these SNPs allows estimation of breed composition of individual animals using genomic data (Manel et al. 2005; Kuehn et al. 2011; Frkonja et al. 2012; Hulsegge et al. 2013).

On the basis of established methods it is possible to estimate breed composition and purity and to allow incorporating purebred animals in the breed registry for purebred animals.

A purity test requires genotypes of reference individuals whose breed of origin is known, a so called reference population. The individuals in a reference population should match the full range of genetic diversity within a particular breed. Based on

these reference individuals, SNP markers can be selected, which contain sufficient genetic information to be able to discriminate amongst the breeds. Preferably the number of SNP markers should be limited, in order to simplify the test, to reduce the costs and to speed up computations. The information of the selected SNPs from the reference populations subsequently could be used to infer the ancestry of individuals with unknown origin. For a purity test it is necessary to draw a threshold value for which an allocation of an unknown individual to a breed is accepted.

For implementing the methodology in practice, a rapid and reliable method for genetic purity testing of animals is needed, distinguishing crossbred animals from purebred animals and to determine the breed composition. Furthermore, there is genetic variation within breeds and consequently a breed purity test will depend on how well this genetic variation will be reflected in the reference populations dataset. Finally, some introgression of genes of other breeds is generally accepted, e.g. animals registered with 87.5% pedigree purity are generally considered purebred, so the challenge is to determine a threshold value for purity that is generally accepted.

The general aim of this study was to set up an easy applicable, highly accurate and affordable breed composition and purity test for the purpose of breed purity determination where pedigree is unknown or unable to verify with traditional methods. The specific objectives of this study were to: (1) build reference populations with individuals whose breed of origin is known; (2) select SNP markers that contain sufficient genetic information to be able to discriminate amongst the cattle breeds, (3) demonstrate the effectiveness of the test and (4) validate the test.

6.2. Materials and methods

6.2.1. Animals and genotypes

Six local cattle breeds in the Netherlands were incorporated in the purity test: Deep Red Cattle, Dutch Belted, Dutch Friesian, Dutch Red and White Friesian, Groningen White Headed and Meuse-Rhine-Yssel. Genotype data for these local breeds were available from former studies (Maurice-Van Eijndhoven et al. 2015; Francois et al. 2017; Hulsegge et al. 2017; Manzanilla-Pech et al. 2017) and the recently available genotype data from bulls in the Dutch gene bank, born between 1960 and 2015. Data on the six local breeds were provided by the Centre for Genetic Resources, The Netherlands (CGN). Individuals were genotyped with the Illumina BovineSNP50 or BovineHD Beadchip. The dataset includes data from bulls in the Dutch gene bank collection, suggesting they would include the genetic variation present in the population (Berg and Windig 2017). The cows were selected from several farms for

each breed. As the local breeds are sometimes crossed with Holstein Friesians, we included genotype data of a small group of Holstein Friesian animals as an outgroup to the dataset with the local cattle breeds. Data of Holstein Friesians were from cows of the Dairy Campus Research dairy herd (Wageningen University & Research, Wageningen Livestock Research, Lelystad, The Netherlands). Previously performed editing and imputation steps of these data are described by Manzanilla-Pech et al. (2017). After combining the different genotype datasets, a total of 36,148 SNPs remained for a total of 1850 animals with pedigree breed percentage > 87.5% (8/8 breed fraction)

6.2.2. Quality control

Prior to the analysis, several quality control measures were applied to the genotype data. The dataset was pruned by excluding SNPs and animals with a call rate < 90%. Missing genotypes were imputed using Beagle with 20 iterations (Browning and Browning 2009). Imputation was carried out for each breed and chromosome independently, except for the Holstein Friesian samples which were already imputed. Rare alleles were not excluded, because these are important for the differentiation between breeds (Bertolini et al. 2015). SNPs were pruned for Linkage Disequilibrium (LD, threshold: > 0.2) with the SNP Relate (version 1.12.2) package in R (Zheng et al. 2012). After quality control, a total of 10,449 SNPs and 1774 purebred animals remained for the analysis.

6.2.3. Reference and test population

Each cattle breed was divided into a reference population and a test population. The test population was generated by randomly sampling 10% of the animals within each breed with a maximum of *n* = 20. The test population included 4 Deep Red Cattle, 4 Dutch Belted, 5 Dutch Red and White Friesian, 20 Dutch Friesian, 12 Groningen White Headed and 20 Meuse-Rhine-Yssel (Table 6.1). The remaining animals formed the reference population (Table 6.1). The test population was supplemented with 59 crossbred animals with known breed composition, 29 purebred- and 9 crossbred animals of other breeds (20 Improved Red Cattle, 8 Lineback Cattle and 1 Belgian Red Cattle) (Table 6.1).

The difference in number of samples per breed could bias the analysis. Therefore, we performed the analysis using a maximum of 150 randomly selected animals per breed. We included genotype data of a small group of Holstein Friesian animals as an outgroup to the dataset with the local cattle breeds (test population *n* = 19; reference population *n* = 50). The final reference population included a total of 572

purebred animals (36 Deep Red Cattle, 32 Dutch Belted, 43 Dutch Red and White Friesian, 150 Dutch Friesian, 111 Groningen White Headed, 150 Meuse-Rhine-Yssel and 50 Holstein Friesian (Table 6.1).

Table 6.1. Number of animals per breed in the reference population (REF) and test population (TEST). Reference is the population used to develop the breed composition and purity test; test population are animals with known breed composition used to validate the developed test.

Breed Name	REF	TEST population by breed percentage (12.5%)				
		>87.5%	>75%	>62.5%	50%	>37.5%
Deep Red Cattle	36	4	0	0	1	1
Dutch Belted	32	4	1	0	0	0
Dutch Friesian	150	20	4	1	0	1
Dutch Red and White Friesian	43	5	0	1	0	2
Groningen White headed	111	12	13	6	1	4
Meuse-Rhine-Yssel	150	20	14	1	0	2
Holstein Friesian	50	19	1	4	0	1
Other breed		29	5	1	0	3

6.2.4. Selection of informative SNPs

A combined approach of Principal Component Analysis (PCA) and Random Forest (RF) (Bertolini et al. 2015) was used to determine which SNPs contained the most information to discriminate among breeds. PCA was performed using the prcomp function in R (Core Team 2016). The first two principal components (PC1 and PC2) were used to reduce the number of SNPs needed to discriminate between breeds. The contribution of each SNP to PC1 and PC2 was estimated using the function get_pca_var incorporated in the factoextra package (version 1.0.5) in R (Kassambara 2017). The contribution of each SNP to each of the PCs was ranked and the 500 SNPs with highest contribution were selected, leading to 1000 selected SNPs. After removing duplicates, 976 SNPs remained. Random Forests based on the selected 976 SNPs were built using the Random Forests (RF, version 4.6 12) R package (Liaw and Wiener 2002), where the number of trees was set to *ntree* = 10,000 and the number of candidate predictors considered at each split to *mtry* = 500. The classification confusion matrix, an error matrix, as well as the out-of-bag error (OOB), the estimated prediction error, were used to evaluate the quality of classification. It has been shown that the Mean Decrease in Gini Index (MDGI), a relevance measures, is most likely to promote SNPs with high minor allele frequencies (Boulesteix et al.

2012), which was found to be beneficial in a similar study investigating the selection of informative SNPs to differentiate four cattle breeds (Bertolini et al. 2015). Based on the ranked MDGI score of the SNPs the 100 most informative SNPs were selected.

6.2.5. Clustering animals

The model-based clustering method implemented in the program STRUCTURE (version 2.3.4) (Pritchard et al. 2000) was used to infer the most probable number of genetically distinct clusters present in the reference population and to estimate admixture proportions within each of those clusters. The software clustered the data according to allele frequencies into K populations (clusters). The admixture model, correlated allele frequencies (Falush et al. 2003) and the number of populations *K* = 6 to 8 were used for the STRUCTURE analyses, a total of 200,000 Markov chain Monte Carlo (MCMC) iterations were run, with a burn-in period of 100,000 iterations. The seed was set at 1234. Results of clustering based on higher and lower numbers of clusters (K) confirmed that seven clusters were the best fit to the data at hand.

6.2.6. Validation

Predicting individual breed composition and purity of the test population based on the 133 informative SNPs was calculated using the program STRUCTURE (version 2.3.4) (Pritchard et al. 2000; Porras-Hurtado et al. 2013). The data of the test population was treated as having unknown affinity and the program assigned the test individuals to the seven genetic clusters from the reference population. The USEPOPINFO model was used, whereby the reference populations were used to estimate the ancestry of the test population with unknown origin. Clustering and allele frequencies were updated using only individuals from the reference populations (*POPFLAG*=1) so that individuals from the test population were forced to cluster with one or more of the reference population clusters. Based on preliminary analysis (data not shown), the GENSBACK ("generations back" infers only whether an individual itself is a migrant) was set to 1 and the prior on migration rate (MIGRPRIOR) to 0.01. Again, a total of 200,000 MCMC iterations were run, with a burn-in period of 100.000 iterations. STRUCTURE assigned each individual to the inferred clusters based on the individual proportion of membership (Q-value) and its confidence interval (90% CI). In order to distinguish purebreds from crossbreds a threshold value needed to be set. The threshold value was set based on achieving an optimal balance between false positives (a crossbred animal assigned as purebred) and false negatives (a purebred animal assigned as crossbred). Therefore, the proportion of membership of the purebred animals (≥ 87.5%; 7/8 and 8/8) of the test

population was determined and subsequently the proportion of membership of the crossbred animals (that must be excluded). The set threshold, as best as possible, assigned the purebred animals but excluded the crossbred animals.

6.3. Results

6.3.1. Selection of informative SNPs

The first three PCs separated the 572 individual animals from the reference population according to their breed (Figure 6.1). PC1 accounted for 6.3% of the total variation and separated the Dutch Friesian breeds (Dutch Red and White Friesian and Dutch Friesian) on the one hand and Groningen White Headed on the other hand from Holstein Friesian, Meuse-Rhine-Yssel and Deep Red Cattle. PC2 (5.5%) separated Meuse-Rhine-Yssel and Deep Red Cattle on the one hand and the Dutch Friesian breeds and Groningen White Headed on the other hand from Dutch Belted and Holstein Friesian, while PC3 distinguished all local breeds from Holstein Friesian. A partial overlap between Dutch Friesian and Dutch Red and White Friesian as well as between Meuse-Rhine-Yssel and Deep Red Cattle was observed as expected based on their history.

Assigning the reference population animals to breeds rendered too many misclassifications when based on RF and 976 SNPs (Table 6.2). Therefore, a second selection step was performed to render a more (and reduced) informative set of SNPs. Based on the ranked MDGI score of the SNPs the 100 most informative SNPs were selected.

To improve the assignments of the closely related breeds (Dutch Friesian and Dutch Red and White Friesian, as well as Meuse-Rhine-Yssel and Deep Red Cattle), additional SNPs were selected. For both comparisons, the 20 SNPs with the highest differences in allele frequency between the two breeds were selected. These 40 SNPs and the 100 most informative SNPs selected with RF were combined. After removal of duplicates 133 SNPs remained. This set of 133 SNPs resulted in less misclassification as the error rate reduced from 6.3% when using 976 SNPs to 4.4%. However, the error rate within some breeds was still unacceptably high (Table 6.2) when keeping in mind an application in practice. We therefore considered breed assignment using the STRUCTURE program in which for each animal proportions of membership to each of the seven clusters (that is, breeds) was provided.

Figure 6.2 shows the distribution of the 133 SNPs over the different chromosomes. SNP name, chromosome and location of the SNPs is available in Suppl. Table 6.1. The

selected 133 informative SNPs were located across all chromosomes, where the number of SNPs per chromosome ranged from one to 12.

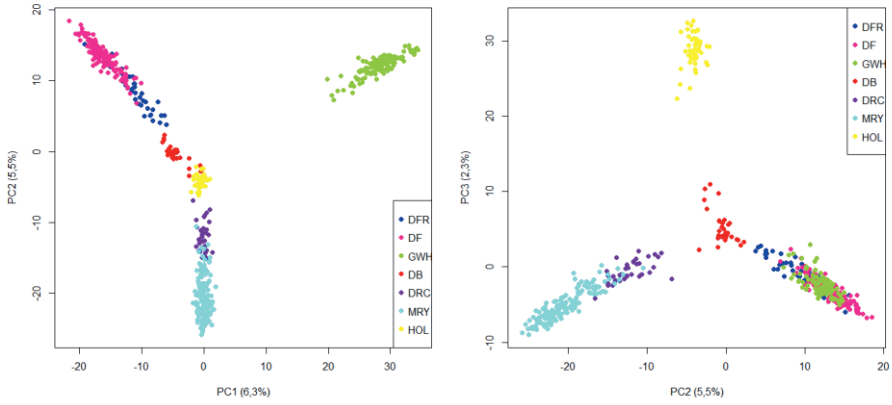

Figure 6.1. PCA results visualizing individuals of various breeds within the reference population using 10,449 SNPS, with the percentage of variance explained in Brackets. Abbreviations: DRC = Deep Red Cattle, DB = Dutch Belted, DF = Dutch Friesian, DRF = Dutch Red and White Friesian, GWH = Groningen White Headed, MRY = Meuse-Rhine-Yssel and HOL = Holstein Friesian.

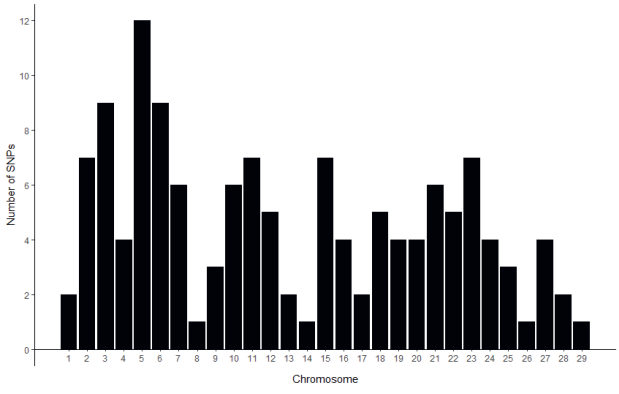

Figure 6.2. Distribution of the 133 SNPs over the chromosomes.

Table 6.2. Assignment of reference population animals to breeds based on Random Forest (RF) classification using 976 and 133 SNPs.

Breed	DRF	DF	GWD	DB	DRC	MRY	Hol	Error rate
RF classification 976 SNPs								
DRF	25	16	1			1		0.419
DF	1	149						0.007
GWH			111					0.000
DB			1	29	1		1	0.094
DRC	1				20	15		0.444
MRY	1					149		0.007
HOL							50	0.000
RF classification 133 SNPs								
DRF	29	13				1		0.325
DF		150						0.000
GWH			111					0.000
DB	1			30			1	0.063
DRC					29	7		0.194
MRY		1			1	148		0.013
HOL							50	0.000

Abbreviations: DRC = Deep Red Cattle, DB = Dutch Belted, DF = Dutch Friesian, DRF = Dutch Red and White Friesian, GWH = Groningen White Headed, MRY = Meuse-Rhine-Yssel and HOL = Holstein Friesian.

6.3.2. Breed assignment

The STRUCTURE analysis (*K* = 6 to 8) using the 572 animals in the reference population showed the lowest cross/validation error at *K* = 7 and confirmed the presence of seven breeds. The purebred animals of the Dutch Friesian, Groningen White Headed, Dutch Belted, Meuse-Rhine-Yssel and Holstein-Friesian breeds within the reference population showed large proportion of membership in one of the inferred clusters (mean proportion of membership was > 0.9; Table 6.3). These animals were therefore correctly assigned to their breed of origin. However, this did not hold for the purebred animals of the Dutch Red and White Friesian and Deep Red Cattle breeds within the reference population. Mean proportion of membership of purebred Dutch Red and White Friesian animals to inferred cluster 5, the cluster representing this breed, was 0.731 (Table 6.3). A considerable average proportion of membership (0.227; Table 6.3) was also assigned to inferred cluster 2, the Dutch

Friesian breed. Similarly, the average proportion of membership of purebred Deep Red Cattle animals to inferred cluster 6, the cluster representing this breed, was 0.894 (Table 6.3). The second largest average proportion of membership for the purebred Deep Red Cattle animals was 0.044 to inferred cluster 3, the Meuse-Rhine-Yssel breed.

Table 6.3. Average proportion of membership of the animals in the reference population to the seven clusters. The highest contributions per breed are in boldface.

Breed	Inferred clusters							Number of animals
	1	2	3	4	5	6	7	
DRF	0.012	0.227	0.008	0.004	**0.731**	0.007	0.011	43
DF	0.009	**0.935**	0.005	0.004	0.023	0.009	0.015	150
GWH	0.009	0.005	0.004	**0.959**	0.008	0.008	0.007	111
DB	0.042	0.009	0.011	0.007	0.013	0.013	**0.907**	32
DRC	0.023	0.011	0.044	0.004	0.013	**0.894**	0.012	36
MRY	0.007	0.004	**0.937**	0.003	0.005	0.038	0.006	150
HOL	**0.949**	0.006	0.016	0.006	0.006	0.011	0.006	50

Abbreviations: DRC = Deep Red Cattle, DB = Dutch Belted, DF = Dutch Friesian, DRF = Dutch Red and White Friesian, GWH = Groningen White Headed, MRY = Meuse-Rhine-Yssel and HOL = Holstein Friesian.

6.3.3. Assignment testing

In general, animals from the test population showed a high proportion of membership to the same cluster as the reference population representatives of the same breed (Supp. Table 2). Average proportion of membership of the animals in the test population ranged from 0.687 for the Dutch Red and White Friesian to 0.929 for the Groningen White Headed (Figure 6.3). The 90% probability interval of the purebred test population of Groningen White Headed was smaller than that of the other breeds, suggesting that the genetic diversity within Groningen White Headed (or at least within this data set) is lower than within the other breeds and/or Groningen White Headed has more unique alleles compared to the other breeds.

A low proportion of membership to their breed of origin was observed for several test animals (Figure 6.3). For example, proportion of membership of one Dutch Red and White Friesian animal was 0.251. For this particular animal a higher proportion of membership was observed for the Dutch Friesian breed (0.627), which could be explained by its ancestors (mostly from Dutch Friesian).

The threshold value for which an allocation of an unknown individual to a breed is accepted was set to 0.775 (proportion of membership). The threshold value was set based on achieving an optimal balance between false positives (a crossbred animal assigned as purebred) and false negatives (a purebred animal assigned as crossbred). Accuracy in breed assignment of the test population as determined by the number of animals correctly assigned to their breed of origin using the threshold value for proportion of membership of 0.775 is shown in Table 6.4. Overall, 82.0% (96 out of 117) of the animals in the test population is correctly assigned to the breed in question. No animals were assigned to another breed and no animals from the other breeds (Improved Red Cattle, Lineback Cattle and Belgian Red Cattle) were assigned to the Dutch local breeds in question. As previously indicated, the Dutch Red and White Friesian cattle is closely related to the Dutch Friesian breed, as well as Deep Red Cattle is closely related to Meuse-Rhine-Yssel. For these breeds, if an animal is not correctly assigned, but the combined (Meuse-Rhine-Yssel and Deep Red Cattle or Dutch Red and White Friesian and Dutch Friesian) proportion of membership is ≥0.775, the animal can be considered as purebred, provided that the phenotype, colour and/or pattern and meets the requirements for the breed, as determined by the herdbook. Using this criterion 88.9% (104 out of 117) of the animals in the test population is correctly assigned. Of the 34 purebred animals (≥87.5%) that are composed of breeds not in the reference populations (Improved Red Cattle, Lineback Cattle and Belgian Red Cattle), 33 animals were not assigned as belonging to one of the seven breeds in the reference population. One Improved Red Cattle was incorrectly assigned as purebred Deep Red Cattle.

The proportion of membership of crossbred animals should be below the threshold value of 0.775. In total 73.3% of the crossbred animals were indeed assigned as admixture (Table 6.5). Noticeably, almost half of the crossbred animals of the Groningen White Headed were assigned as purebred Groningen White Headed. One Dutch Red and White Friesian crossbred animal (75% Dutch Red and White Friesian and 25% unknown) was assigned as purebred (Table 6.5). It is very plausible that the unknown breed Dutch Red and White Friesian breed or Dutch Friesian was.

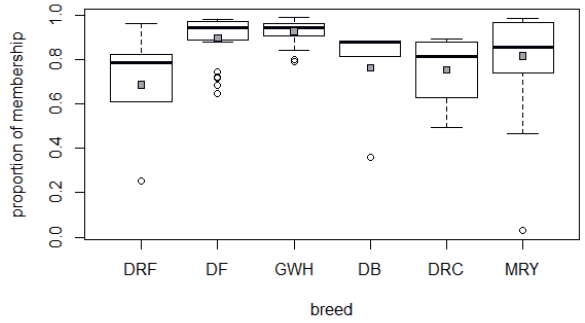

Figure 6.3. Boxplot of the breed proportion of membership for the six local cattle breeds. (greysquare= average proportion of membership; Dutch Red and White Friesian (DRF) n = 5; Dutch Friesian (DF) n = 23; Groningen White Headed (GWH) n = 25; Dutch Belted (DB) n = 5; Deep Red Cattle (DRC) n = 4 and Meuse-Rhine-Yssel (MRY) n = 34).

Table 6.4. Assignment accuracy of the test population.

Breed**	# purebred (≥87.5%)	# assigned			# assigned from other breeds**
		Purebred (Q-value ≥ 0.775	Crossbred (Q-value < 0.775)	Other breed**	
DRF	5	3	2	0	0
DF	24	19	5	0	0
GWH	25	25	0	0	0
DB	5	4	1	0	0
DRC	4	3	1	0	0
MRY	34	24	10	0	0
HOL	20	18	2	0	0
Total	117	96	21	0	0
DRF + DF*	29	25	4	0	0
MRY + DRC*	38	32	6	0	0
Total	117	104	13	0	0

*Combined membership proportion.

**Other breed(s) = breeds within the reference population: Deep Red Cattle, Dutch Belted, Dutch Friesian, Dutch Red and White Friesian, Groningen White Headed, Meuse-Rhine-Yssel and Holstein Friesian.

Abbreviations: DRC=Deep Red Cattle, DB = Dutch Belted, DF = Dutch Friesian, DRF = Dutch Red and White Friesian, GWH = Groningen White Headed, MRY = Meuse-Rhine-Yssel and HOL = Holstein Friesian.

Table 6.5. Assignment accuracy of the crossbred animals and animals from other breeds.

Breed	# crossbred	# correctly assigned crossbred (Q-values < 0.775)	# assigned purebred
DRF	3	2	1
DF	2	2	0
GWH	11	6	5
DB	–	–	–
DRC	2	2	0
MRY	3	2	1
HOL	5	5	0
OTH	4	3	1
Total	30	22	8

Abbreviations: DRC=Deep Red Cattle, DB = Dutch Belted, DF = Dutch Friesian, DRF = Dutch Red and White Friesian, GWH = Groningen White Headed, MRY = Meuse-Rhine-Yssel, HOL = Holstein Friesian and OTH] Other Breed: Improved Red Cattle, Lineback Cattle and Belgian Red Cattle.

6.4. Discussion

In this study we set up a test to determine breed composition and purity and quality control where pedigree is unknown or unable to verify with traditional methods

6.4.1. Breeds

Six local Dutch cattle breeds were incorporated in the purity test: Deep Red Cattle, Dutch Belted, Dutch Friesian, Dutch Red and White Friesian, Groningen White Headed and Meuse-Rhine-Yssel.

Anecdotally and according to breed registry information, the Dutch Red and White Friesian cattle is closely related to the Dutch Friesian breed, as well as the Deep Red Cattle is closely related to the Meuse-Rhine-Yssel. The Dutch Red and White Friesian Cattle originated from Dutch Friesian. With the increasing demand for black and white pied animals for export, the red pied Dutch Friesians were no longer allowed to be registered as Dutch Friesian. However, some farmers kept breeding with red pied animals and in 1975 the Dutch Red and White Friesian became an official cattle breed. Both are now registered as one breed, with an additional notification for colour. Similarly, the Deep Red Cattle and Meuse-Rhine-Yssel are closely related. These two breeds have a common history. With the increasing interest in highly productive dairy cattle the number of purebred Meuse-Rhine-Yssel decreased

rapidly. Farmers attempted to improve production in local cattle breeds through crossing with more productive breeds. In Meuse-Rhine-Yssel white colouring was preferred because a link of this colouring to milk production was suspected. Farmers opposing these changes, moved back to the old type of dual-purpose cattle with its typical deep red coat colour, creating a new line within the breed: Deep Red Cattle (de Haas et al. 2009). The separation of Deep Red Cattle as an official studbook was in 2004. This clarifies why the PCA, RF and STRUCTURE had difficulties to distinguish between these breeds.

6.4.2. Reference population

The genotype data available for this study was not specifically gathered to build a reference population for the purpose to setup a breed composition and purity test. The genotype data of the different breeds used to compose the reference populations originated from different studies (Maurice-Van Eijndhoven et al. 2015; Francois et al. 2017; Hulsegge et al. 2017; Manzanilla-Pech et al. 2017) and the recently available genotype data from bulls of which semen is stored in the Dutch national gene bank of CGN, born between 1960 and 2015, suggesting they would include the genetic variation present in the population (Berg and Windig 2017). Cows were selected from several farms for each breed, suggesting that they represent different families and thereby relevant variation in the population. The variation present in a population should be represented by a reference population, to avoid exclusion of atypical animals or even whole breeding lines or families (Hulsegge et al. 2013). Dalvit et al. (2008) and Rosenberg et al. (2001) suggested for real and practical use of breed assignment methods to verify the suitability of collected samples to be used as a reference population. For pigs, Funckhouser et al. (2017) indicated that subpopulations within a breed may differ in allele en haplotype frequencies, highlighting the importance of having a representative reference population that capture the genetic variation existing among animals to be tested. Our results showed that the animals of the reference population form genetic clusters that correspond to their breed designations and that these animals can be used in a reference population for assignment of future unknowns. We have no indications that the genetic diversity range of the reference population is too small. Another important aspect for a reference population is the minimum number of animals that would be required to accurately assign an animal to a breed using genotype data (Connolly et al. 2014). The data used in this study included an unequal number of animals in the breeds of the reference population. Connolly et al. (2014) indicated that at least 50 animals are required in a reference population when attempting to discriminate between distantly related breeds, and many more (400

to 500) if the breeds are closely related. This latter number is probably difficult to realize in regard to the small population sizes of most of the Dutch cattle breeds. Frkonja et al. (2012) reported that a very small number of samples of purebred (ancestral) individuals (10) is sufficient to provide accurate estimates of admixture. Although the results showed that the breed assignment of the test population using the current reference population was successful, we propose, based on the arguments mention above, to add additional animals to the reference population. When adding additional animals to the reference population one should sample widely from the breed and avoid adding closely related animals. So, the reference populations could still be improved on numbers and potentially representation of the total genetic diversity.

6.4.3. Selection of informative SNPs

Genotyping and analysing a large number of SNPs is costly and time-consuming. Therefore, selecting a subset of SNPs that is sufficiently informative is an important step toward a breed composition and purity test. Several methods can be used to determine which SNPs contain the most information to discriminate between populations (Ding et al. 2011; Wilkinson et al. 2011b; Bertolini et al. 2015). In this study we used the combination of PCA and RF to perform SNP selection (Bertolini et al. 2015). PCA has been used already in cattle to reduce dimensionality of large SNP data sets and to identify breed informative SNPs (Lewis et al. 2011; Wilkinson et al. 2011b; Bertolini et al. 2015). This pre-filtering PCA step was combined with RF, an approach that can classify and assign individuals. Bertolini et al (2018) demonstrated the usefulness of RF in combination with other SNP reduction techniques to identify breed informative SNPs and that PCA is the best technique to combine with RF in order to classify and assign individuals to breeds. From tests selecting different numbers of informative SNPs (data not shown) the selection of 1000 informative SNPs through PCA and out of these 1000 the 100 most informative SNPs found by RF was large enough to distinguished between the Dutch cattle breeds. However, for the closely related breeds Dutch Red and White Friesian and Dutch Friesian, as Deep Red Cattle and Meuse-Rhine-Yssel the 100 selected SNPs were not sufficient. Therefore, we added additional SNPs based on allele frequency between Dutch Red and White Friesian and Dutch Friesian and between Deep Red Cattle and Meuse-Rhine-Yssel, resulting in a total of 133 selected SNPs. The closely related breeds Dutch Red and White Friesian and Dutch Friesian showed overlap in the results of PCA and RF. This overlap is partial and does not hold for all animals of the Dutch Red and White Friesian population. Hulsegge et al. (2017) stated that Dutch Friesian and Dutch Red and White Friesian are closely related, but that some of the breeding lines

in the Dutch Red and White Friesian population are genetically distinct from each other, from Dutch Friesian and the other breeds. A similar challenge occurred with the differentiation between Deep Red Cattle and Meuse-Rhine-Yssel, which also had a slight overlap between the populations in the PCA and RF results. This overlap can be traced back to the common history of both breeds. As well as the fact that there are still some (crossbred) Meuse-Rhine-Yssel bulls used in the Deep Red Cattle breeding program.

The number of selected informative SNPs depends on the breeds under consideration in the reference population and their respective levels of genetic heterogeneity.

The 133 identified SNPs were useful to discriminate among all the cattle breeds under study. These markers are probably not useful to discriminate among other cattle breeds or even same breeds but from different countries. However, the used strategy can be reproduced to develop marker sets to discriminate other breeds.

6.4.4. Breed assignment

Several studies have proven the software of STRUCTURE to be efficient in assigning animals to their breed of origin (Padilla et al. 2009; Rogberg-Munoz et al. 2014). Although the genealogical purity of animals used in the reference populations was known based on pedigree information, we followed the suggestion of Pritchard et al. (2000) and applied for cattle by Padilla et al. (2009) of assigning animals. That is, before making use of population information, clustering the data without using prior population information should be performed, to check that the genetically defined cluster does agree with population labels. STRUCTURE showed that the reference population split in seven clusters (K = 7) each corresponding to a breed. The genetically defined clusters agreed with the original breeds. Padilla et al. (2009) showed that posterior use of population information improved the accuracy of assigning animals to clusters and the estimates of the probabilities of membership for each animal in each cluster, giving a greater precision in the assignment of individuals lacking genealogical information. Therefore, we activated the PopFlag option in STRUCTURE. In this way, animals of the reference populations were a priori assigned to their predefined clusters (PopFlag = 1), while the animals of the test population (PopFlag = 0) were probabilistically assigned to breeds without using prior knowledge.

6.4.5. Assignment testing

The number of animals of some breeds for assignment testing was very limited, due to lack of more genotype data.

Breed assignment was performed for animals whose listed breed composition is comprised of one of six local breeds in the reference populations. Animals that were composed of breeds not in the reference population got predicted as a seemingly random mixture of the reference populations.

Using the threshold value for the proportion of membership of ≥ 0.775 purebred animals from the test population (based on pedigree) were correctly assigned and crossbreds (again based on pedigree) were identified. There are no firm guidelines for acceptable false positive and false negative results. According to Miciak et al. (2015) the criteria can be ultimately pragmatic, using an optimal balance between false positives and false negatives. The proportion correctly assigned for the purebred test animals differed between breeds, with the highest proportion for Groningen White Headed and lowest for Dutch Red and White Friesian. As mentioned earlier, Meuse-Rhine-Yssel and Deep Red Cattle breeds separated in the recent past, while for Dutch Friesian and Dutch Red and White Friesian recent mixing occurred. For these breeds we suggest that if an animal has a percentage for its own breed < 0.775, but the combined percentage of the two mentioned breeds (Meuse-Rhine-Yssel and Deep Red Cattle or Dutch Friesian and Dutch Red and White Friesian) is ≥ 0.775, the animal can be considered as purebred, provided that the phenotype, colour and/or pattern, meets the requirements for the breed as determined by the herdbook. However, for Groningen White Headed almost half of the crossbred animals were assigned as purebred Groningen White Headed using this threshold value. The threshold value for this breed may have to be set differently.

Altogether, in general the animals of the test population were very well assigned to the correct breed in question, and crossbred animals and the animals from other breeds were identified as well. This latter is beneficial in the way that animals which are not actual purebred for one of the Dutch local cattle breeds, would not be classified as such. And even though the average proportion of membership differed between the breeds, the proportion of membership represented an accurate indication about whether or not an animal is purebred.

6.5. Conclusion

Although tens of thousands of SNP markers are now available, only a small set of SNPs (n = 133), when accurately chosen, was needed to differentiate among the Dutch local cattle breeds. The reference population of purebred animals showed genetic clusters that corresponded to their breed designations and its usefulness for assignment of future unknowns. Although the reference populations could still be improved on numbers and representation of the total genetic diversity. The breed assignment of the test population using STRUCTURE software, the current reference population and the selected SNPs was successful. Therefore, this test was implemented in practice to identify (partly) unregistered individuals as being purebred (or not) for one of the Dutch local cattle breeds.

Acknowledgements

Funding: This work was funded by the Dutch Ministry of Agriculture, Nature and Food Quality.

The text represents the author's views and does not necessary represent a position of the Ministry who will not be liable for the use made of such information.

Conflict of interest

All the authors have no conflict interest.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.livsci.2019.03.002.

7

General discussion

7.1 Introduction

Nowadays, genomic information is available for choices to be made in conservation. Genomics can estimate genetic diversity more accurately than is possible with pedigree information only. Moreover, it can describe in great detail the genetic diversity across the genome. Therefore, it can substantially improve the selection and prioritisation of animals for conservation of genetic resources. Bruford et al. (2015) stated that the livestock sector needs to make a concerted effort to use the powerful genomic tools that are now at its disposal, and to apply them to breed conservation and breed development. According to Oldenbroek et al. (2022), genomics is of importance for various stakeholders – gene bank managers, rare breed associations and commercial breeding companies – in order to improve their efforts to conserve and use genetic diversity.

Gene bank collections are important for three reasons: (1) they are an insurance policy against changes in market or environmental conditions; (2) they are a safeguard against emerging diseases, political instability, and natural disasters; and (3) they provide opportunities for research (Gandini and Oldenbroek 2007; Blackburn et al. 2022). Genomic characterisation of gene bank collections is important to unlock their genetic potential, to assess the genetic diversity captured in gene bank collections, and to understand genetic changes over time better. Examples of the use of genomics to characterise populations and gene bank collections include the characterisation of French local chicken breeds (Restoux et al. 2018) and Dutch cattle breeds (van Breukelen et al. 2019). In this thesis I have used a range of techniques for the analysis of genomic data to better conserve Dutch livestock breeds. In **Chapter 2**, I demonstrated that both the population structure of a breed and its relationship with other breeds should be taken into account in the conservation decisions for a breed. **Chapters 3** and **4** illustrated the value of gene bank collections for the conservation of genetic diversity. These chapters showed that genomic characterisation allows us to get a much richer picture of the content of the Dutch gene bank. In **Chapters 5** and **6**, I demonstrated that only a small set of informative SNPs is needed to differentiate among Dutch local cattle breeds. Using such a small set of informative SNPs, a genetic tool (DNA test) was developed for determining the breed purity of cattle. Genomic tools can be used to identify individual uniqueness, to identify genome regions or even specific markers of importance (i.e., signals of selection in **Chapter 4**), and to accurately estimate relationships between breeds and individuals. These applications show that genomic information should be routinely available for choices to be made in conservation, because it is more accurate than a combination of pedigree and phenotype.

In this general discussion I will elaborate on this statement. I will address questions and opportunities related to conservation of genetic diversity based on genomic information. In addition, I will discuss the upcoming developments in genomics and how they can best be used for genetic conservation, and in particular, how gene banks can benefit from these developments, and in addition, how gene banks should be set up to maximise their potential from a genomic viewpoint.

7.2 From Single Nucleotide Polymorphisms arrays to Whole Genome Sequencing

7.2.1 Single Nucleotide Polymorphisms

New technology and market/research requirements drove the development of SNP arrays for various livestock species with a range of marker densities. For example, in 2015 there were 11 commercial SNP arrays available for cattle, with the number of SNPs ranging from <3K (low-density) to >500K (high-density) (Nicolazzi et al. 2015). In addition, there is a constantly growing number of custom-made SNP arrays which are not commercially available for third parties, and are developed by consortia (e.g. EuroGenomics) (Nicolazzi et al. 2015; Boichard et al. 2018). Generic SNP arrays are known to lack a substantial proportion of globally rare variants and tend to be biased towards variants present in the commercial breeds (such as Holstein Friesian) that were involved in the development process of the SNP arrays (Perez-Enciso et al. 2015; Geibel et al. 2021). This is known as SNP ascertainment bias. As a consequence, informative and breed-specific variants segregating in various local breeds (such as Dutch Friesian Cattle breed) have not been considered. Analysis of local breeds based on commercial SNPs arrays are therefore prone to this ascertainment bias (Neto and Barendse 2010). For this reason, custom SNP arrays may be required for local breeds. Specific variants on customized SNP arrays are particularly interesting for the maintenance of breed-specific genomic variants and properties of different small populations with a specific genealogy (Neumann et al. 2021). For that reason and to facilitate the genotyping of animals in gene banks, two multi-species SNP arrays have been developed to facilitate genotyping of animals in gene banks within the H2020 project “Innovative Management of Animal Genetic resources” (IMAGE) (IMAGE 2020; FAO 2021; Tixier-Boichard et al. 2022). These two arrays (IMAGE001 and IMAGE002) capture genetic diversity of traditional breeds across Europe. Another example, which focuses on managing and maintaining the endangered German Black Pied cattle (DSN) population, is the developed and customised SNP array (DSN200k) that considers genetic variants unique to DSN in addition to informative SNPs from the Illumina BovineSNP50 and genetic information from additional breeds (Neumann et al. 2021). SNPs for these arrays need to be identified

by WGS data from a diverse set of individuals representing the breed. Gene banks contain the required genetic material for sequencing and are therefore a valuable resources for developing custom SNP arrays for local breeds.

7.2.2 Whole Genome Sequence

With the ongoing developments of WGS, especially the decreasing costs, it is expected that WGS will replace SNP arrays over time. Compared to SNP arrays, the use of WGS provides various advantages in assessing genetic diversity, such as the avoidance of ascertainment bias and a significantly higher information content. WGS delivers information on rare variants and it maps genomic regions highly affected by selection pressure. For example, a mutation that leads to the typical dwarfism phenotype in Friesian horses (Leegwater et al. 2016) and mutations associated with polledness in cattle (Medugorac et al. 2017; Aldersey et al. 2020) could only be discovered with the help of WGS. Conservation of rare variants has received little attention due to their inaccessibility through SNP arrays (Eynard et al. 2016). As WGS data capture both common and rare variants, it better meets one of the major goals of genetic diversity management in livestock species, the conservation of rare variants for both long- and short-term perspectives (Eynard et al. 2016; Eusebi et al. 2019). WGS may help to identify unique genetic variation in breeds, which is an important characteristic for the prioritisation of breeds for conservation. For example, in **Chapter 4**, I identified a number of genetic variants in the Dutch gene bank not present in the living Dutch Friesian population, indicating in detail that the genetic diversity within the breed has changed in approximately 5–10 generations. This knowledge can be of great value for the maintenance of the Dutch Friesian breed, especially when inbreeding threatens to increase, which made the conservation of sires from 50 years ago meaningful. Thus, WGS is the preferred technology to identify rare variants and genetic variation.

7.2.3 Number of Single Nucleotide Polymorphisms to use in analysing genetic diversity

Despite the advantages of WGS over SNP arrays, it may not always be the method of choice to analyse genetic variation. Genotyping methods differ with respect to the number of markers and cost per sample. Given this premise, choosing the best method for analysing genetic diversity requires clarity on two key aspects, i.e., the number of markers sufficient to fulfil the goals and the cost associated with each alternative. Despite the large reduction in the cost of WGS, sequencing large number of individuals remains expensive. In addition, routine analyses of WGS data from a large number of individuals still face serious challenges, such as their analyses being

labour intensive and requiring expertise, and the requirement of expensive hardware equipment for the scale needed in breeding programmes. The expected value of the information gained by genotyping/sequencing for breeding and conservation should be reasonably in line with the cost of obtaining the genotypes, especially in practice. Therefore, it is good to realise that WGS is not required for every purpose related to conservation. SNP arrays used in livestock species are available with different marker densities: 1) low-density, below 20k SNPs, 2) medium-density, ~50KSNPs, and 3) high-density, >500K SNPs. Low-density arrays have been developed for purposes such as: breed assignment, breed traceability of animals and animal products, and parentage verification and reconstruction (see e.g. Wilkinson et al. 2011b; Dimauro et al. 2013; Flanagan and Jones 2019; Wilmot et al. 2022a; Wilmot et al. 2022b). In **Chapter 6**, I demonstrated that only a small set of informative SNPs is needed to differentiate among Dutch local cattle breeds. In order to achieve these informative SNPs, it was necessary to genotype gene bank animals in much greater detail with a medium-density array. Using this small set of informative SNPs a genetic tool (DNA test) was developed for determining breed purity of cattle. Out of tens of thousands of SNP markers, only 133 SNPs were needed to assign animals correctly to the different Dutch cattle breeds. In 2018, the DNA test has been implemented in practice in The Netherlands and herdbooks are increasing their breeding population by registering animals without pedigrees as purebred based on the DNA test results. Until now (27-07-2023), 267 non-registered animals were tested for their purebred status. Out of these 267 animals, 211 animal were tested as purebred and could be incorporated in one of the herdbooks. An increase in population size is very important to avoid inbreeding in these small populations. As indicated, a small set of SNPs, if accurately chosen, is sufficient to differentiate the genetic origin of a group of specific breeds. However, more genetic markers will be required to successfully assign closely related breeds and far fewer for distantly related breeds.

Medium-density arrays are useful for e.g. genomic prediction, precision livestock farming, controlling inbreeding, and marker imputation (see e.g. Daetwyler et al. 2012; Hoze et al. 2013; Zhang et al. 2015; Iheshiulor et al. 2016). High-density arrays are, aside from these applications, also used for the detection of genetic associations with complex traits, QTL mapping, detecting signatures of selection, and increasingly, for implementing genomic selection to farm animal species (Kranis et al. 2013). In contrast, there are some other applications, for which having millions of SNPs provided by WGS is a clear benefit. This includes the detection of lethal recessive and deleterious alleles in general, the identification of rare alleles or even

de novo variants, QTL mapping, the calling of structural variants beyond SNPs, such as copy number variants and reciprocal translocations (see e.g. Harland et al. 2017; Zhang et al. 2017; Derks et al. 2018; Eusebi et al. 2019; Bouwman et al. 2020; Reynolds et al. 2021). For gene banks, collections of WGS data will be valuable, since it will help, for example as discussed in 7.2.2, to identify rare genetic variants that were lost over time or variants that are unique to a specific breed. However, the budgets of gene banks for sequencing samples of their collection is limited. As a result, it is necessary to balance between the cost and accuracy of the genomics information. Taking all this into consideration, I propose a strategy for genetic conservation, 1) genotype all animals with a medium-density array and 2) sequence a limited number of animals periodically. The animals that will be selected for sequencing (point 2) should differ sufficiently in diversity from the animals already sequenced. The selection of animals could be done by screening the medium-density genotype information of individuals (point 1). Previous studies have described methods and opportunities of genomics for the conservation of genetic diversity (e.g. Eding et al. 2002; Engelsma 2012; Eynard 2018).

7.3 Dynamics of genetic diversity

Genetic diversity of populations is always changing. Breeds undergo constant changes in genetic makeup, more evident in local breeds due to their small sizes and improper genetic management, which leads to serious genetic drift. Three chapters in this thesis demonstrated that genomics helps to gain insight into genetic diversity and to better understand the dynamics of genetic diversity in populations through time. In **Chapter 2,** I showed that a Dutch Red and White Friesian cattle line for a considerable time in isolation from other lines of the Dutch Red and White Friesian cattle breed has apparently conserved genetic diversity not present anymore in the rest of the population. **Chapters 3** and **4** deal with changes in genetic diversity over time as well. **Chapter 3** showed that the merging of commercial Dutch Landrace pig lines over time has reduced the genetic diversity of the Landrace population in the Netherlands, while at the same time, has produced new combinations of genetic diversity. For poultry, Besbes et al. (2007) also reported that the merging of lines leads to a decrease in the genetic diversity of the available gene pool. The WGS study in **Chapter 4** showed that the genetic diversity of the Dutch Friesians has reduced over time probably due to the decline in population size. In addition, I identified highly differentiated genomic regions across autosomal chromosomes in the recently bred Dutch Friesian bulls. This latter was in agreement with Moscarelli et al. (2021), who reported the presence of several genomic regions that vary between

original and modern Brown cattle populations, in line with their different breeding histories, with respect to the breeding goal.

All breeds tend to be subject to selection and genetic drift; as a result, after several generations, differences will emerge, in allele and haplotype frequencies, between the gene bank collection and the in situ population. This necessitates the resampling of breeds over time. As a result of periodic resampling, gene banks pools sometimes capture more genetic diversity than is present in the current population (Blackburn 2012; Paiva et al. 2016; Boitard et al. 2021). The frequency of resampling of a breed depends upon the genetic change that is occurring. Estimates of how often this should be done have been made, and indicate roughly every 4 to 7 generations (Blackburn 2018; FAO 2021). Similarly, environmental influences can exert selection pressure within and among breeds. The existence of subpopulations within a breed suggests that breeds can adapt to varying environmental factors. Such subpopulations contain potentially useful genetic resources for future use and should therefore be sampled for gene bank collections (Blackburn 2018). In conclusion, periodic resampling of breeds is necessary to keep the gene bank collection up to date.

7.4 The future of gene banks in the genomic era

In this section, I will argue that “digital gene banks” are the future. Gene banks should expand from simply germplasm collections to physical and digital resources. This means that, in addition to the actual germplasm collection, gene banks also contain large amounts of associated information from various data domains, such as phenotypic, molecular, morphologic and geographic data. These will provide better and more targeted access to the material, and increase its use. Consequently, in my view, in the genomic era where bioinformatics plays an important role, this is the time for gene banks to review and revise their approach. Currently, conservation budget in The Netherlands is allocated for both in situ and ex situ conservation. My findings suggest that gene banks should also invest in collecting detailed information on the stored material in the gene bank, in particular genomic data, and to make this information accessible. This should become an additional effort from gene banks besides the important conservation work they already do, hence it will require an increase in the conservation budget.

Modern genomic tools and methods, such as high-density genotyping, whole genome sequencing, and bioinformatics, have been developed since the gene banks emerged. It can be expected that the technologies will continue to improve, such

that obtaining genotypes, or even whole-genome sequences of all material in gene banks, may be an option that becomes available to many gene bank collections in the near future. By implementing the shift to genomic tools and methods, traditional gene banks, which focus on the preservation of collections, will be able to transform into digital resource centres, which combine the conservation of materials with their genomic and phenotypic characterisation. The creation of "digital gene banks" based on genomic information will help gene banks become more efficient, cost-effective, and informative as collectors, conservers, and providers of samples and information. This information should facilitate the use of gene bank material in the genomic management of small as well as of main stream commercial populations. Already in 2014, Van Treuren and Van Hintum (2014) indicated that gene banks should start re-thinking their mission, especially considering users' needs, as well as presenting information in an accessible and useful way. The greatest challenges in creating "digital gene banks" revolve around cost, funding, availability of genomic resources, technical and infrastructural capacity, and expertise in different domain areas. Phenotypic information of individual animals in gene banks is often limited. It would be very helpful if genomic information and bioinformatics could overcome this omission, e.g. by predictions of the phenotypes. Recently, plant genetic resources took the first steps towards digital gene banks. In a study, González et al. (2018) examined strategies to unlock historical research data as a first step towards extending the gene bank into a bio-digital resource centre facilitating an educated choice of barley genetic resources for research and breeding. An important next step was reached with the genomic characterisation of gene bank collections, such as the genomic characterisation of the barley collection comprising more than 22,000 accessions (Milner et al. 2019). For some crops, massive sequencing has been undertaken, as shown in rice, wheat and barley germplasm collections (Wang et al. 2018; Milner et al. 2019; Sansaloni et al. 2020). Gene banks for animal genetic resources lag behind those for plant resources in thinking about digital gene banks that provide material and information that meet the needs of users. However, some first steps for animal gene banks have been taken through the IMAGE project (https://www.imageh2020.eu/ [cited 8-8-2022]). Based on the results of the IMAGE project, the Food and Agriculture Organization of the United Nations (FAO) has drawn up new guidelines that include a section on information of critical importance for gene bank management, new types of information available for improved management of gene bank collections and for placing of gene bank data in the public domain (FAO 2021).

The challenges for animal gene banks in the future will be to raise global awareness of the value of their collections for research and breeding, as well as to further strengthen, implement and optimize the ex situ conservation strategies (Blesbois et al. 2022). Finally, for nearly all aspects of daily life, the quantity and importance of information increased in recent years. Gene banking is no exception. Data about stored samples should be considered an integral aspect of the collection. Modern systems and tools for management of these data, including their integration with other sources of complementary information, and sharing it with stakeholders are becoming more and more fundamental features of gene banks (Blesbois et al. 2022).

7.5 Further developments

This thesis demonstrates that the use of genomic data and methods leads to a more detailed understanding of the genetic diversity conserved in gene banks and in current populations of numerically small breeds. Both approaches, in situ and ex situ, are generally considered complementary to each other (FAO 2019). The goal of almost all management and conservations actions is to obtain comprehensive knowledge of the species/breeds to inform decision making. Many management and conservation questions can be much better addressed using genomic information from SNP arrays, for example, the question posed in **Chapter 2**: What is the relationship of DFR with other Dutch dairy breeds and the contribution of the DFR to the total genetic diversity in Dutch dairy cattle breeds? Questions that cannot be fully resolved with SNP arrays or require sophisticated insights may be answered by WGS using next generation sequencing (NGS). Developments in genomic data and methods may also contribute to management and conservation of genetic diversity. Two of these developments are worth discussing here: new sequencing technologies and pan-genomes.

7.5.1 From first to third generation sequencing

Continuous improvement in sequencing technology implies that the whole genome can be sequenced faster, more easily, and with a higher accuracy since the start of Sanger sequencing (mid-1970s), the so-called first-generation sequencing. At present, the NGS (or second-generation sequencing or short-read sequencing) technology is widely used (Dunisławska et al. 2017; Slatko et al. 2018; Hu et al. 2021). NGS platforms generate relatively short reads (up to ~600 nucleotides). NGS has made it possible to explore genetic diversity with a higher level of detail. However, the short read lengths of NGS methods pose a limitation for the identification of structural variants, sequencing repetitive regions, phasing of alleles and distinguishing highly homologous genomic regions (Mantere et al. 2019).

Furthermore, short-read data analysis is highly dependent on the reference genomes, which are known to be imperfect (Mantere et al. 2019) and represent only a single genome. In response to these limitations, long-read sequencing (or third generation sequencing) platforms have been developed, which are characterised by long reads with an average length of more than 10 kb. However, relative to short read sequencing, long read sequencing suffers from higher nucleotide calling error rates, higher costs, and more limited throughput (Goodwin et al. 2016). New methodologies are focusing on generating synthetic long reads by taking advantage of the benefits of short-read technology but incorporating information from long strands of DNA. This allows for the barcoded short reads to be associated with their original long molecules producing a novel data type known as "Linked-Reads" (i.e., synthetic long reads) (Marks et al. 2019; Stervander and Cresko 2021). After the development of three generations, DNA sequencing technology is now entering the era of single molecule nanopore technology (Feng et al. 2015). The significant advantages of nanopores include label-free, ultra-long reads high throughput, and low material requirement (Feng et al. 2015). Although the nanopore-based sequencing technology has emerged to be a promising tool, several problems remain to be solved. The main bottlenecks are sample imaging, the relatively low efficiency of molecular processes, data handling, and interpretation (Ke et al. 2016). In the coming years, new sequencing platforms will probably appear producing a larger amount of data (in Terabyte), which in turn requires the development of new approaches and applications capable of analysing this large amount of data.

7.5.2 Pan-genome

In **Chapter 4**, I aligned the DNA sequences from the animals to the cattle reference genome ARS-UCD1.2 in order to investigate genome-wide genetic diversity between a group of historic Dutch Friesians bulls, a group of recent ones, and a group of recently used Holstein Friesian bulls. This reference genome is derived from a single European Hereford cow (Rosen et al. 2020). A complete and accurate reference genome is fundamental for read alignment and subsequent comprehensive discovery of genomic variants. However, there is increasing awareness that a reference genome from a single individual cannot fully represent the genomic diversity of one species since many sequences could be absent in the reference genome (Li et al. 2019; Sherman and Salzberg 2020; Derks et al. 2022). For example, regions (i.e. structural variations) in the DNA that are unique to a specific breed will be completely missed if mapped to a single reference genome, leading to a so called "reference bias". Hence, numerous potentially interesting sites of variation will be missed, affecting downstream analysis. Hence, downstream analyses are biased

towards the alleles and haplotypes present in the reference sequence. To overcome this reference bias, a new concept called pan-genomics has gained strong interest in plant, human and livestock genetics ((Sherman and Salzberg 2020; Miga and Wang 2021; Wang et al. 2022) (human); (Golicz et al. 2016; Bayer et al. 2020; Della Coletta et al. 2021) (plant); (Li et al. 2019; Tian et al. 2020; Derks et al. 2022; Talenti et al. 2022) (livestock)). A pan-genome is a collection of sequences and genes within a species, consisting of a core genome and a variable genome (Li et al. 2019). The pan-sequences can aid population genomics research to understand population stratification, the fine mapping of causal genes and variants to better understand the molecular mechanisms underlying (complex) traits (Wong et al. 2018). Therefore, it is necessary to build a pan-genome by uncovering the pan-sequences that are absent from the reference genome to maximally represent the genetic diversity within one species. Talenti et al. (2022) illustrated this by aligning five African cattle assemblies, that a substantial portion (4.2 %) of the cattle pan-genome is likely missing from the Hereford reference. By sampling a diverse set of individuals, one can begin to assemble a pan-genome: a collection of all the DNA sequences that occur in a species. Pan-genomes are usually set up in a graph structure and tools ('pantools') have been developed to map genomic information to these graph structures and to assess structural differences ('bubbles') between genomes (Sheikhizadeh Anari 2020).

The importance of pan-genomes has been widely accepted in the field of plant genomics (Golicz et al. 2016; Zhao et al. 2018; Della Coletta et al. 2021). Animal genomes are much more conserved, as generally, only intergenic or fragmented genic regions are involved in the gain/loss of genomic sequences in animals. This makes it more challenging to assess the impact of such variations. Nonetheless, with increasing quantities of (third generation) sequence information, pan-genomics is becoming more popular in animal genomics. To better assess structural variation between breeds or samples, the pan-genome is of great value, because structural variation calling with short-read whole-genome sequence data still poses challenges (especially in highly complex regions) (Tian et al. 2020). Putting together a pan-genome for a complex genome, like genomes of farm animal species, is facilitated by improvements in genome sequencing technologies, particularly long-read sequencing (third generation sequencing). Larger genomes contain higher proportions of repetitive sequences, which are more difficult to analyse using short reads (next generation sequencing). By sampling a diverse set of individuals, one can begin to assemble a pan-genome: a collection of all the DNA sequences that occur in a species. In pan-genomics gene banks can be important, not only because they

contain many different breeds, but also because they contain animals from long ago that often contain diversity no longer present in the living animals.

7.6 Final words

Overall, genetic diversity is essential to ensure that livestock can adapt to (un)expected changes in breeding goals. In this thesis, I demonstrated that the use of genomic data and methods leads to a more detailed understanding of the genetic diversity conserved in gene banks and in current populations of numerically small breeds. Using genomics, it is possible to quantify genetic diversity within and between breeds, the genetic distance between breeds and relationships between animals within breeds to a high degree of accuracy. Furthermore, genomics provides detailed information about inbreeding and genetic drift, as well as the genomic regions under selection, and can be used to detect potentially valuable rare alleles and haplotypes and their carriers in breeds. Genomics enables us to choose candidates for conservation based on specific genetic diversity. Genomics is of high value as a practical application for conserving the genetic diversity of Dutch livestock (in situ and ex situ).

Genomic technologies, methods and analyses will continue to develop and populations will always change. Therefore, gene banks have to adapt as well by continuously updating collections through adding samples. Furthermore, gene banks have to stay alert as to whether new genomic techniques can provide new insights to conserve populations better and, where necessary, genotype their collections using the latest techniques.

References

References

Abd El Naby WS, Hagos TH, Hossain MM, Salilew-Wondim D, Gad AY, Rings F, Cinar MU, Tholen E, Looft C, Schellander K et al. 2013. Expression analysis of regulatory microRNAs in bovine cumulus oocyte complex and preimplantation embryos. *Zygote* **21**: 31-51. doi:10.1017/S0967199411000566.

Abi-Rached L, Gouret P, Yeh JH, Di Cristofaro J, Pontarotti P, Picard C, Paganini J. 2018. Immune diversity sheds light on missing variation in worldwide genetic diversity panels. *PLoS One* **13**: e0206512. doi:10.1371/journal.pone.0206512.

Ahrens CW, Rymer PD, Stow A, Bragg J, Dillon S, Umbers KDL, Dudaniec RY. 2018. The search for loci under selection: trends, biases and progress. *Mol Ecol* **27**: 1342-1356. doi:10.1111/mec.14549.

Albrechtsen A, Nielsen FC, Nielsen R. 2010. Ascertainment biases in SNP chips affect measures of population divergence. *Mol Biol Evol* **27**: 2534-2547. doi:10.1093/molbev/msq148.

Aldersey JE, Sonstegard TS, Williams JL, Bottema CDK. 2020. Understanding the effects of the bovine POLLED variants. *Anim Genet* **51**: 166-176. doi:10.1111/age.12915.

An B, Xia J, Chang T, Wang X, Xu L, Zhang L, Gao X, Chen Y, Li J, Gao H. 2019. Genome-wide association study reveals candidate genes associated with body measurement traits in Chinese Wagyu beef cattle. *Anim Genet* **50**: 386-390. doi:10.1111/age.12805.

Andrews S. 2010. FASTQC. A quality control tool for high throughput sequence data.

Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D. 2020. Plant pan-genomes are the new reference. *Nat Plants* **6**: 914-920. doi:10.1038/s41477-020-0733-0.

Bennewitz J, Eding H, Ruane J, Simianer H. 2007. Selection of breeds for conservation. In *Utilisation and Conservation of Farm Animal Genetic Resources*, doi:10.3920/978-90-8686-592-5, pp. 131-146. Wageningen Academic Publishers.

Berg P, Windig JJ. 2017. Management of cryo-collections with genomic tools. *Genomic management of animal genetic diversity Wageningen: Wageningen Academic Publishers*: 155-178.

Bertolini F, Galimberti G, Calo DG, Schiavo G, Matassino D, Fontanesi L. 2015. Combined use of principal component analysis and random forests identify population-informative single nucleotide polymorphisms: application in cattle breeds. *J Anim Breed Genet* **132**: 346-356. doi:10.1111/jbg.12155.

Bertolini F, Galimberti G, Schiavo G, Mastrangelo S, Di Gerlando R, Strillacci MG, Bagnato A, Portolano B, Fontanesi L. 2018. Preselection statistics and Random Forest classification identify population informative single nucleotide polymorphisms in cosmopolitan and autochthonous cattle breeds. *Animal* **12**: 12-19. doi:10.1017/S1751731117001355.

Besbes B, Tixier-Boichard M, Hoffmann I, Jain GL. 2007. Future trends for poultry genetic resources. In *Proceedings of the International conference of poultry in the 21st century: Avian influenza and beyond*, pp. 5-7, Bangkok, Thailand.

Biscarini F, Cozzi P, Gaspa G, Marras G. 2018. detectRUNS: Detect runs of homozygosity and runs of heterozygosity in diploid genomes. In *The Comprehensive R Archive Network*.

Bjornstad G, Roed KH. 2002. Evaluation of factors affecting individual assignment precision using microsatellite data from horse breeds and simulated breed crosses. *Anim Genet* **33**: 264-270. doi:10.1046/j.1365-2052.2002.00868.x.

Blackburn HD. 2012. Genetic selection and conservation of genetic diversity*. *Reprod Domest Anim* **47 Suppl 4**: 249-254. doi:10.1111/j.1439-0531.2012.02083.x.

Blackburn HD. 2018. Biobanking Genetic Material for Agricultural Animal Species. *Annu Rev Anim Biosci* **6**: 69-82. doi:10.1146/annurev-animal-030117-014603.

Blackburn HD, Hiemstra SJ, Tixier-Boichard M. 2022. Building a gene banking strategy. *Innovations in Cryoconservation of Animal Genetic Resources FAO*.

Blesbois E, Santiago-Moreno J, Hopkins R, Rajamohan A, Magistrini M, Purdy P. 2022. Guidelines for semen cryopreservation. *Innovations in Cryoconservation of Animal Genetic Resources FAO*.

Boettcher PJ, Hoffmann I, Baumung R, Drucker AG, McManus C, Berg P, Stella A, Nilsen LB, Moran D, Naves M et al. 2014. Genetic resources and genomics for adaptation of livestock to climate change. *Front Genet* **5**: 461. doi:10.3389/fgene.2014.00461.

Boettcher PJ, Tixier-Boichard M, Toro MA, Simianer H, Eding H, Gandini G, Joost S, Garcia D, Colli L, Ajmone-Marsan P et al. 2010. Objectives, criteria and methods for using molecular genetic data in priority setting for conservation of animal genetic resources. *Anim Genet* **41 Suppl 1**: 64-77. doi:10.1111/j.1365-2052.2010.02050.x.

Boichard D, Boussaha M, Capitan A, Rocha D, Hozé C, Sanchez MP, Tribout T, Letaief R, Croiseau P, Grohs C et al. 2018. Experience from large scale use of the EuroGenomics custom SNP chip in cattle. In *11th World Congress of Genetics Applied to Livestock Production*, Auckland, New Zealand.

Boitard S, Paris C, Sevane N, Servin B, Bazi-Kabbaj K, Dunner S. 2021. Gene Banks as Reservoirs to Detect Recent Selection: The Example of the Asturiana de los Valles Bovine Breed. *Front Genet* **12**: 575405. doi:10.3389/fgene.2021.575405.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114-2120. doi:10.1093/bioinformatics/btu170.

Bongiorni S, Gruber CE, Bueno S, Chillemi G, Ferre F, Failla S, Moioli B, Valentini A. 2016. Transcriptomic investigation of meat tenderness in two Italian cattle breeds. *Anim Genet* **47**: 273-287. doi:10.1111/age.12418.

Boulesteix AL, Bender A, Lorenzo Bermejo J, Strobl C. 2012. Random forest Gini importance favours SNPs with large minor allele frequency: impact, sources and recommendations. *Brief Bioinform* **13**: 292-304. doi:10.1093/bib/bbr053.

Bouwman AC, Daetwyler HD, Chamberlain AJ, Ponce CH, Sargolzaei M, Schenkel FS, Sahana G, Govignon-Gion A, Boitard S, Dolezal M et al. 2018. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nat Genet* **50**: 362-367. doi:10.1038/s41588-018-0056-5.

Bouwman AC, Derks MFL, Broekhuijse M, Harlizius B, Veerkamp RF. 2020. Using short read sequencing to characterise balanced reciprocal translocations in pigs. *BMC Genomics* **21**: 576. doi:10.1186/s12864-020-06989-x.

Browning BL, Browning SR. 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* **84**: 210-223. doi:10.1016/j.ajhg.2009.01.005.

Bruford MW, Ginja C, Hoffmann I, Joost S, Orozco-terWengel P, Alberto FJ, Amaral AJ, Barbato M, Biscarini F, Colli L et al. 2015. Prospects and challenges for the conservation of farm animal genomic resources, 2015-2025. *Front Genet* **6**: 314. doi:10.3389/fgene.2015.00314.

Caballero A, Toro MA. 2000. Interrelations between effective population size and other pedigree tools for the management of conserved populations. *Genet Res* **75**: 331-343. doi:10.1017/s0016672399004449.

Caballero A, Toro MA. 2002. Analysis of genetic diversity for the management of conserved subdivided populations. *Conservation Genetics* **3**: 289-299. doi:10.1023/a:1019956205473.

Calus M. 2013. Calc_grm – a Program to Compute Pedigree, Genomic, and Combined Relationship Matrices. *ABGC, Wageningen UR Livestock Research, Wageningen, The Netherlands*.

Cerri RL, Thompson IM, Kim IH, Ealy AD, Hansen PJ, Staples CR, Li JL, Santos JE, Thatcher WW. 2012. Effects of lactation and pregnancy on gene expression of endometrium of Holstein cows at day 17 of the estrous cycle or pregnancy. *J Dairy Sci* **95**: 5657-5675. doi:10.3168/jds.2011-5114.

Chan E. 2012. Calc_wcFstats. http://evachan.org/rscripts.html. (Accessed 21 November 2012.).

Cheng J, Zhao H, Chen N, Cao X, Hanif Q, Pi L, Hu L, Chaogetu B, Huang Y, Lan X et al. 2020. Population structure, genetic diversity, and selective signature of Chaka sheep revealed by whole genome sequencing. *BMC Genomics* **21**: 520. doi:10.1186/s12864-020-06925-z.

Christensen GL, Ivanov IP, Atkins JF, Mielnik A, Schlegel PN, Carrell DT. 2005. Screening the SPO11 and EIF5A2 genes in a population of infertile men. *Fertil Steril* **84**: 758-760. doi:10.1016/j.fertnstert.2005.03.053.

Chung HY, Lee KT, Jang GW, Choi JG, Hong JG, Kim TH. 2015. A genome-wide analysis of the ultimate pH in swine. *Genet Mol Res* **14**: 15668-15682. doi:10.4238/2015.December.1.19.

Cole JB, Waurich B, Wensch-Dorendorf M, Bickhart DM, Swalve HH. 2014. A genome-wide association study of calf birth weight in Holstein cattle using single nucleotide polymorphisms and phenotypes predicted from auxiliary traits. *J Dairy Sci* **97**: 3156-3172. doi:10.3168/jds.2013-7409.

Connolly S, Fortes M, Piper E, Seddon J, Kelly M. 2014. Determining the number of animals required to accurately determine breed composition using genomic data. In *Proceedings of the 10th world congress of genetics applied to livestock production, Vancouver*, pp. 17-22.

Core Team R. 2016. A language and environment for statistical computing. *http://www R-project org*.

CRV. 2021a. Genetische trends van koeien in Nederland. *https://cooperatiecrv-be6kxcdncom/wp-content/uploads/2021/08/gen_trend_koe_nl_20210819pdf* **accessed 22-10-2021**.

CRV. 2021b. Inbreeding in Dutch cattle. *https://wwwcooperatie-crvnl/wp-content/uploads/2021/01/Inbreeding-in-Dutch-cattle-1pdf* **accessed 05-11-2021**.

Daetwyler HD, Kemper KE, van der Werf JH, Hayes BJ. 2012. Components of the accuracy of genomic prediction in a multi-breed sheep population. *J Anim Sci* **90**: 3375-3384. doi:10.2527/jas.2011-4557.

Dalvit C, De Marchi M, Dal Zotto R, Gervaso M, Meuwissen T, Cassandro M. 2008. Breed assignment test in four Italian beef cattle breeds. *Meat Sci* **80**: 389-395. doi:10.1016/j.meatsci.2008.01.001.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156-2158. doi:10.1093/bioinformatics/btr330.

de Cara MA, Fernandez J, Toro MA, Villanueva B. 2011. Using genome-wide information to minimize the loss of diversity in conservation programmes. *J Anim Breed Genet* **128**: 456-464. doi:10.1111/j.1439-0388.2011.00971.x.

de Haas Y, Hoving AH, Maurice - Van Eijndhoven MHT, Bohte-Wilhelmus DI, Sulkers H, Hiemstra SJ. 2009. Deep Red Cattle.

De Man AP. 2008. *Knowledge management and innovation in networks*. Edward Elgar Publishing Ltd.

de Roos AP, Hayes BJ, Spelman RJ, Goddard ME. 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. *Genetics* **179**: 1503-1512. doi:10.1534/genetics.107.084301.

Della Coletta R, Qiu Y, Ou S, Hufford MB, Hirsch CN. 2021. How the pan-genome is changing crop genomics and improvement. *Genome Biol* **22**: 3. doi:10.1186/s13059-020-02224-8.

Derks MFL, Boshove A, Harlizius B, Sell-Kubiak E, Lopes MS, Grindflek E, Knol E, Groenen MAM, Gjuvsland AB. 2022. A pan-genome of commercial pig breeds. In *12th World Congress on Genetics Applied to Livestock Production*

(WCGALP), doi:10.3920/978-90-8686-940-4 (ed. RF Veerkamp, Y de Haas). Wageningen Academic Publishers, Rotterdam.

Derks MFL, Megens HJ, Bosse M, Visscher J, Peeters K, Bink M, Vereijken A, Gross C, de Ridder D, Reinders MJT et al. 2018. A survey of functional genomic variation in domesticated chickens. *Genet Sel Evol* **50**: 17. doi:10.1186/s12711-018-0390-1.

Dimauro C, Cellesi M, Steri R, Gaspa G, Sorbolini S, Stella A, Macciotta NP. 2013. Use of the canonical discriminant analysis to select SNP markers for bovine breed assignment and traceability purposes. *Anim Genet* **44**: 377-382. doi:10.1111/age.12021.

Ding L, Wiener H, Abebe T, Altaye M, Go RC, Kercsmar C, Grabowski G, Martin LJ, Khurana Hershey GK, Chakorborty R et al. 2011. Comparison of measures of marker informativeness for ancestry and admixture mapping. *BMC Genomics* **12**: 622. doi:10.1186/1471-2164-12-622.

Doekes HP. 2020. Genomic characterization and conservation of genetic diversity in cattle. Wageningen University, Wageningen, The Netherlands.

Dunisławska A, Łachmańska J, Sławińska A, Siwek M. 2017. Next generation sequencing in animal science - A review. *Anim Sci Pap Rep* **35**: 205-224.

Eding H, Crooijmans RP, Groenen MA, Meuwissen TH. 2002. Assessing the contribution of breeds to genetic diversity in conservation schemes. *Genet Sel Evol* **34**: 613-633. doi:10.1186/1297-9686-34-5-613.

Engelsma KA. 2012. Use of SNP Markers to Conserve Genome-wide Genetic Diversity in Livestock. Wageningen University, Wageningen, The Netherlands.

Engelsma KA, Veerkamp RF, Calus MP, Windig JJ. 2011. Consequences for diversity when prioritizing animals for conservation with pedigree or genomic information. *J Anim Breed Genet* **128**: 473-481. doi:10.1111/j.1439-0388.2011.00936.x.

EU. 2016a. Regulation (EU) 2016/429 of the European Parliament and of the Council of 9 March 2016 on transmissible animal diseases and amending and repealing certain acts in the area of animal health ('Animal Health Law') (Text with EEA relevance). pp. 1-208.

EU. 2016b. Regulation (EU) 2016/1012 of the European Parliament and of the Council of 8 June 2016 on zootechnical and genealogical conditions for the breeding, trade in and entry into the Union of purebred breeding animals, hybrid breeding pigs and the germinal products thereof and amending Regulation (EU) No 652/2014, Council Directives 89/608/EEC and 90/425/EEC and repealing certain acts in the area of animal breeding ('Animal Breeding Regulation') (Text with EEA relevance). pp. 66-143.

European Cattle Genetic Diversity C. 2006. Marker-assisted conservation of European cattle breeds: An evaluation. *Anim Genet* **37**: 475-481. doi:10.1111/j.1365-2052.2006.01511.x.

Eusebi PG, Martinez A, Cortes O. 2019. Genomic Tools for Effective Conservation of Livestock Breed Diversity. *Diversity* **12**. doi:10.3390/d12010008.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol* **14**: 2611-2620. doi:10.1111/j.1365-294X.2005.02553.x.

Eynard SE. 2018. Using genomic information to conserve genetic diversity in livestock. Vol Phd, Wageningen.

Eynard SE, Windig JJ, Hiemstra SJ, Calus MP. 2016. Whole-genome sequence data uncover loss of genetic diversity due to selection. *Genet Sel Evol* **48**: 33. doi:10.1186/s12711-016-0210-4.

FABRE Technology Platform 2008. Sustainable Farm Animal Breeding & Reproduction Technology Platform p. 32.

Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**: 1567-1587. doi:10.1093/genetics/164.4.1567.

FAO. 2006. *Livestock's Long Shadow*. FAO, Rome, Italy.

FAO. 2007a. Global plan of action for animal genetic resources and the Interlaken Declaration. *FAO Commission on Genetic Resources for Food and Agriculture, FAO: Rome, Italy*.

FAO. 2007b. The state of the world's animal genetic resources for food and agriculture. (ed. BP Rischkowsky, D.).

FAO. 2015. *The Second Report on the State of the World's Animal Genetic Resources for Food and Agriculture*, Rome.

FAO. 2016. MEAT AND MEAT PRODUCTS. In *http://wwwfaoorg/fileadmin/templates/est/COMM_MARKETS_MONITORING/Meat/Documents/FO_Meat_June_2016pdf*.

FAO. 2019. *The State of the World's Biodiversity for Food and Agriculture. J. Bélanger & D. Pilling (eds.). FAO Commission on Genetic Resources for Food and Agriculture Assessments*, Rome.

FAO. 2021. Innovations in cryoconservation of animal genetic resources. https://www.fao.org/fileadmin/user_upload/animal_genetics/docs/CGRFA-18-21-10_2_Inf1_forPDF.pdf. Accessed 16 Aug 2022.

Felius M, Koolmees PA, Theunissen B, European Cattle Genetic Diversity C, Lenstra JA. 2011. On the Breeds of Cattle—Historic and Current Classifications. *Diversity* **3**: 660-692. doi:10.3390/d3040660.

Feng Y, Zhang Y, Ying C, Wang D, Du C. 2015. Nanopore-based fourth-generation DNA sequencing technology. *Genomics Proteomics Bioinformatics* **13**: 4-16. doi:10.1016/j.gpb.2015.01.009.

Fernandez J, Meuwissen TH, Toro MA, Maki-Tanila A. 2011. Management of genetic diversity in small farm animal populations. *Animal* **5**: 1684-1698. doi:10.1017/S1751731111000930.

Fernandez J, Toro MA, Caballero A. 2008. Management of subdivided populations in conservation programs: development of a novel dynamic system. *Genetics* **179**: 683-692. doi:10.1534/genetics.107.083816.

Fimland E, Oldenbroek K. 2007. Practical implications of utilisation and management. In *Utilisation and Conservation of Farm Animal Genetic Resources*,

doi:10.3920/978-90-8686-592-5, pp. 195-213. Wageningen Academic Publishers.

Flanagan SP, Jones AG. 2019. The future of parentage analysis: From microsatellites to SNPs and beyond. *Mol Ecol* **28**: 544-567. doi:10.1111/mec.14988.

Foll M, Gaggiotti O. 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* **180**: 977-993.

Fontanesi L, Schiavo G, Gallo M, Baiocco C, Galimberti G, Bovo S, Russo V, Buttazzoni L. 2017. Genome-wide association study for ham weight loss at first salting in Italian Large White pigs: towards the genetic dissection of a key trait for dry-cured ham production. *Anim Genet* **48**: 103-107. doi:10.1111/age.12491.

Fowler KE, Pong-Wong R, Bauer J, Clemente EJ, Reitter CP, Affara NA, Waite S, Walling GA, Griffin DK. 2013. Genome wide analysis reveals single nucleotide polymorphisms associated with fatness and putative novel copy number variants in three pig breeds. *BMC Genomics* **14**: 784. doi:10.1186/1471-2164-14-784.

Francis RM. 2017. pophelper: an R package and web app to analyse and visualize population structure. *Mol Ecol Resour* **17**: 27-32. doi:10.1111/1755-0998.12509.

Francois L, Wijnrocx K, Colinet FG, Gengler N, Hulsegge B, Windig JJ, Buys N, Janssens S. 2017. Genomics of a revived breed: Case study of the Belgian campine cattle. *PLoS One* **12**: e0175916. doi:10.1371/journal.pone.0175916.

Frankham R. 1995. Conservation genetics. In *Annual Review of Genetics*, Vol 29, pp. 305-327. Annual Reviews Inc.

Frankham R, Ballou JD, Briscoe DA. 2012. *Introduction to Conservation Genetics*. Cambridge University Press, Cambridge.

Franz M, Rolfsmeier S. 2016. Brands, trust and quality in agro-food production networks: The vase of layer hens. *Geogr Ann Ser B Hum Geogr* **98**: 271-286. doi:10.1111/geob.12103.

Frkonja A, Gredler B, Schnyder U, Curik I, Solkner J. 2012. Prediction of breed composition in an admixed cattle population. *Anim Genet* **43**: 696-703. doi:10.1111/j.1365-2052.2012.02345.x.

Funkhouser SA, Bates RO, Ernst CW, Newcom D, Steibel JP. 2017. Estimation of genome-wide and locus-specific breed composition in pigs. *Transl Anim Sci* **1**: 36-44. doi:10.2527/tas2016.0003.

Galla SJ, Brown L, Couch-Lewis Ngāi Tahu Te Hapū O Ngāti Wheke Ngāti Waewae Y, Cubrinovska I, Eason D, Gooley RM, Hamilton JA, Heath JA, Hauser SS, Latch EK et al. 2022. The relevance of pedigrees in the conservation genomics era. *Mol Ecol* **31**: 41-54. doi:10.1111/mec.16192.

Gandini G, Oldenbroek K. 2007. Strategies for moving from conservation to utilisation. In *Utilisation and Conservation of Farm Animal Genetic Resources*, doi:10.3920/978-90-8686-592-5, pp. 29-54. Wageningen Academic Publishers.

Gautier M, Faraut T, Moazami-Goudarzi K, Navratil V, Foglio M, Grohs C, Boland A, Garnier JG, Boichard D, Lathrop GM et al. 2007. Genetic and haplotypic structure in 14 European and African cattle breeds. *Genetics* **177**: 1059-1070. doi:10.1534/genetics.107.075804.

Geibel J, Reimer C, Weigend S, Weigend A, Pook T, Simianer H. 2021. How array design creates SNP ascertainment bias. *PLoS One* **16**: e0245178. doi:10.1371/journal.pone.0245178.

Golicz AA, Bayer PE, Barker GC, Edger PP, Kim H, Martinez PA, Chan CK, Severn-Ellis A, McCombie WR, Parkin IA et al. 2016. The pangenome of an agronomically important crop plant Brassica oleracea. *Nat Commun* **7**: 13390. doi:10.1038/ncomms13390.

Gonzalez MY, Philipp N, Schulthess AW, Weise S, Zhao Y, Borner A, Oppermann M, Graner A, Reif JC. 2018. Unlocking historical phenotypic data from an ex situ collection to enhance the informed utilization of genetic resources of barley (Hordeum sp.). *Theor Appl Genet* **131**: 2009-2019. doi:10.1007/s00122-018-3129-z.

Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17**: 333-351. doi:10.1038/nrg.2016.49.

Gorssen W, Meyermans R, Buys N, Janssens S. 2020. SNP genotypes reveal breed substructure, selection signatures and highly inbred regions in Pietrain pigs. *Anim Genet* **51**: 32-42. doi:10.1111/age.12888.

Goudet J. 2005. hierfstat, a package for r to compute and test hierarchical F-statistics. *Mol Ecol Notes* **5**: 184-186. doi:10.1111/j.1471-8286.2004.00828.x.

Gouveia JJdS, Silva MVGBd, Paiva SR, Oliveira SMPd. 2014. Identification of selection signatures in livestock species. *Genetics and molecular biology* **37**: 330-342.

Gura S. 2007. Das Tierzucht-Monopoly. Konzentration und Aneignungsstrategien einer aufstrebenden Macht in der globalen Ernährungswirtschaft. Liga für Hirtenvölker und Nachhaltige Viehwirtschaft, Ober-Ramstadt, Deutschland.

Gutierrez JP, Altarriba J, Diaz C, Quintanilla R, Canon J, Piedrafita J. 2003. Pedigree analysis of eight Spanish beef cattle breeds. *Genet Sel Evol* **35**: 43-63. doi:10.1186/1297-9686-35-1-43.

Hardie LC, VandeHaar MJ, Tempelman RJ, Weigel KA, Armentano LE, Wiggans GR, Veerkamp RF, de Haas Y, Coffey MP, Connor EE et al. 2017. The genetic and biological basis of feed efficiency in mid-lactation Holstein dairy cows. *J Dairy Sci* **100**: 9061-9075. doi:10.3168/jds.2017-12604.

Haring F. 1961. Schweinerassen in den übrigen Ländern West- und Südeuropas. In *Handbuch der Tierzüchtung Bd 3 Rassenkunde/Halbbd 2 (Schweine-, Schaf-, Ziegen-, Geflügelrassen, Pelztiere, Kaninchen)* (ed. JIJ Hammond, i.).

Harland C, Charlier C, Karim L, Cambisano N, Deckers M, Mni M, Mullaart E, Coppieters W, Georges M. 2017. Frequency of mosaicism points towards mutation-prone early cleavage cell divisions in cattle. *bioRxiv* doi:10.1101/079863: 079863. doi:10.1101/079863.

Hayes BJ, Daetwyler HD. 2019. 1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes. *Annu Rev Anim Biosci* **7**: 89-102. doi:10.1146/annurev-animal-020518-115024.

Hernández FA, Parker BM, Pylant CL, Smyser TJ, Piaggio AJ, Lance SL, Milleson MP, Austin JD, Wisely SM. 2018. Invasion ecology of wild pigs (Sus scrofa) in Florida, USA: the role of humans in the expansion and colonization of an invasive wild ungulate. *Biological Invasions* **20**: 1865-1880.

Herrero-Medrano JM, Megens HJ, Groenen MA, Bosse M, Perez-Enciso M, Crooijmans RP. 2014. Whole-genome sequence analysis reveals differences in population management and selection of European low-input pig breeds. *BMC Genomics* **15**: 601. doi:10.1186/1471-2164-15-601.

Hiemstra SJ, de Haas Y, Mäki-Tanila A, Gandini G. 2010. *Local cattle breeds in Europe: Development of policies and strategies for self-sustaining breeds*. Wageningen Academic Publishers.

Hill WG, Robertson A. 1968. Linkage disequilibrium in finite populations. *Theor Appl Genet* **38**: 226-231. doi:10.1007/BF01245622.

Hillel J, Groenen MA, Tixier-Boichard M, Korol AB, David L, Kirzhner VM, Burke T, Barre-Dirie A, Crooijmans RP, Elo K et al. 2003. Biodiversity of 52 chicken populations assessed by microsatellite typing of DNA pools. *Genet Sel Evol* **35**: 533-557. doi:10.1186/1297-9686-35-6-533.

Hlongwane NL, Hadebe K, Soma P, Dzomba EF, Muchadeyi FC. 2020. Genome Wide Assessment of Genetic Variation and Population Distinctiveness of the Pig Family in South Africa. *Front Genet* **11**: 344. doi:10.3389/fgene.2020.00344.

Hoban S, Bruford MW, Funk WC, Galbusera P, Griffith MP, Grueber CE, Heuertz M, Hunter ME, Hvilsom C, Stroil BK et al. 2021. Global Commitments to Conserving and Monitoring Genetic Diversity Are Now Necessary and Feasible. *BioScience* **71**: 964-976. doi:10.1093/biosci/biab054.

Hoving R, Hulsegge I, Hiemstra SJ. 2017. Varkensrassen in de genenbank: Beschrijving van de rassen en de ontwikkelingen in de varkensfokkerij. Centrum voor Genetische Bronnen, Nederland van Wageningen University & Research.

Hoze C, Fouilloux MN, Venot E, Guillaume F, Dassonneville R, Fritz S, Ducrocq V, Phocas F, Boichard D, Croiseau P. 2013. High-density marker imputation accuracy in sixteen French cattle breeds. *Genet Sel Evol* **45**: 33. doi:10.1186/1297-9686-45-33.

Hu T, Chitnis N, Monos D, Dinh A. 2021. Next-generation sequencing technologies: An overview. *Hum Immunol* **82**: 801-811. doi:10.1016/j.humimm.2021.02.012.

Hulsegge B, Calus MP, Oldenbroek JK, Windig JJ. 2017. Conservation priorities for the different lines of Dutch Red and White Friesian cattle change when relationships with other breeds are taken into account. *J Anim Breed Genet* **134**: 69-77. doi:10.1111/jbg.12233.

Hulsegge B, Calus MP, Windig JJ, Hoving-Bolink AH, Maurice-van Eijndhoven MH, Hiemstra SJ. 2013. Selection of SNP from 50K and 777K arrays to predict

breed of origin in cattle. *J Anim Sci* **91**: 5128-5134. doi:10.2527/jas.2013-6678.

Hulsegge I, Calus M, Hoving-Bolink R, Lopes M, Megens HJ, Oldenbroek K. 2019a. Impact of merging commercial breeding lines on the genetic diversity of Landrace pigs. *Genet Sel Evol* **51**: 60. doi:10.1186/s12711-019-0502-6.

Hulsegge I, Schoon M, Windig J, Neuteboom M, Hiemstra SJ, Schurink A. 2019b. Development of a genetic tool for determining breed purity of cattle. *Livestock Science* **223**: 60-67. doi:10.1016/j.livsci.2019.03.002.

Iheshiulor OO, Woolliams JA, Yu X, Wellmann R, Meuwissen TH. 2016. Within- and across-breed genomic prediction using whole-genome sequence and single nucleotide polymorphism panels. *Genet Sel Evol* **48**: 15. doi:10.1186/s12711-016-0193-1.

Illumina I. 2005. Illumina GenCall Data Analysis Software. https://www.illumina.com/Documents/products/technotes/technote_gencall_data_analysis_software.pdf (Accessed 19 July 2013.).

IMAGE IMoAGR. 2020. DELIVERABLE D4.5: A standard multi-species chip for genomic assessment of collections. https://www.imageh2020.eu/deliverable/D4.5_resubmitted_final.pdf. Accessed 16 Aug 2022.

Jagt CV, Chamberlain A, Schnabel RD, Hayes B, Daetwyler H. 2018. Which is the best variant caller for large whole-genome sequencing datasets? In *In Proceedings of the 11th world congress on genetics applied to livestock production: 11–16 February 2018*, Vol Proceedings of the 11th world congress on genetics applied to livestock production, Auckland, New Zealand.

Jakobsson M, Rosenberg NA. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**: 1801-1806. doi:10.1093/bioinformatics/btm233.

Jang J, Terefe E, Kim K, Lee YH, Belay G, Tijjani A, Han JL, Hanotte O, Kim H. 2021. Population differentiated copy number variation of Bos taurus, Bos indicus and their African hybrids. *BMC Genomics* **22**: 531. doi:10.1186/s12864-021-07808-7.

Junior GA, Costa RB, de Camargo GM, Carvalheiro R, Rosa GJ, Baldi F, Garcia DA, Gordo DG, Espigolan R, Takada L et al. 2016. Genome scan for postmortem carcass traits in Nellore cattle. *J Anim Sci* **94**: 4087-4095. doi:10.2527/jas.2016-0632.

Kassambara A. 2017. *Practical guide to principal component methods in R: PCA, M (CA), FAMD, MFA, HCPC, factoextra*. Sthda.

Ke R, Mignardi M, Hauling T, Nilsson M. 2016. Fourth Generation of Next-Generation Sequencing Technologies: Promise and Consequences. *Hum Mutat* **37**: 1363-1367. doi:10.1002/humu.23051.

Kersbergen P, van Duijn K, Kloosterman AD, den Dunnen JT, Kayser M, de Knijff P. 2009. Developing a set of ancestry-sensitive DNA markers reflecting

continental origins of humans. *BMC Genet* **10**: 69. doi:10.1186/1471-2156-10-69.

Kranis A, Gheyas AA, Boschiero C, Turner F, Yu L, Smith S, Talbot R, Pirani A, Brew F, Kaiser P et al. 2013. Development of a high density 600K SNP genotyping array for chicken. *BMC Genomics* **14**: 59. doi:10.1186/1471-2164-14-59.

Kuehn LA, Keele JW, Bennett GL, McDaneld TG, Smith TP, Snelling WM, Sonstegard TS, Thallman RM. 2011. Predicting breed composition using breed frequencies of 50,000 markers from the US Meat Animal Research Center 2,000 Bull Project. *J Anim Sci* **89**: 1742-1750. doi:10.2527/jas.2010-3530.

Kumar KG, Poole AC, York B, Volaufova J, Zuberi A, Richards BK. 2007. Quantitative trait loci for carbohydrate and total energy intake on mouse chromosome 17: congenic strain confirmation and candidate gene analyses (Glo1, Glp1r). *Am J Physiol Regul Integr Comp Physiol* **292**: R207-216. doi:10.1152/ajpregu.00491.2006.

Leegwater PA, Vos-Loohuis M, Ducro BJ, Boegheim IJ, van Steenbeek FG, Nijman IJ, Monroe GR, Bastiaansen JW, Dibbits BW, van de Goor LH et al. 2016. Dwarfism with joint laxity in Friesian horses is associated with a splice site mutation in B4GALT7. *BMC Genomics* **17**: 839. doi:10.1186/s12864-016-3186-0.

Lewis J, Abas Z, Dadousis C, Lykidis D, Paschou P, Drineas P. 2011. Tracing cattle breeds with principal components analysis ancestry informative SNPs. *PLoS One* **6**: e18007. doi:10.1371/journal.pone.0018007.

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

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

Li R, Fu W, Su R, Tian X, Du D, Zhao Y, Zheng Z, Chen Q, Gao S, Cai Y et al. 2019. Towards the Complete Goat Pan-Genome by Recovering Missing Genomic Segments From the Reference Genome. *Front Genet* **10**: 1169. doi:10.3389/fgene.2019.01169.

Liaw A, Wiener M. 2002. Classification and regression by randomForest. *R news* **2**: 18-22.

Liu X, Du Y, Trakooljul N, Brand B, Murani E, Krischek C, Wicke M, Schwerin M, Wimmers K, Ponsuksili S. 2015. Muscle Transcriptional Profile Based on Muscle Fiber, Mitochondrial Respiratory Activity, and Metabolic Enzymes. *Int J Biol Sci* **11**: 1348-1362. doi:10.7150/ijbs.13132.

Lloyd SS, Steele EJ, Valenzuela JL, Dawkins RL. 2017. Haplotypes for Type, Degree, and Rate of Marbling in Cattle Are Syntenic with Human Muscular Dystrophy. *Int J Genomics* **2017**: 6532837. doi:10.1155/2017/6532837.

Lonergan P, Fair T, Forde N, Rizos D. 2016. Embryo development in dairy cattle. *Theriogenology* **86**: 270-277. doi:10.1016/j.theriogenology.2016.04.040.

Manel S, Gaggiotti OE, Waples RS. 2005. Assignment methods: matching biological questions with appropriate techniques. *Trends Ecol Evol* **20**: 136-142. doi:10.1016/j.tree.2004.12.004.

Mantere T, Kersten S, Hoischen A. 2019. Long-Read Sequencing Emerging in Medical Genetics. *Front Genet* **10**: 426. doi:10.3389/fgene.2019.00426.

Manzanilla-Pech CIV, Veerkamp RF, de Haas Y, Calus MPL, Ten Napel J. 2017. Accuracies of breeding values for dry matter intake using nongenotyped animals and predictor traits in different lactations. *J Dairy Sci* **100**: 9103-9114. doi:10.3168/jds.2017-12741.

Marchitelli C, Consortium E. 2006. Marker-assisted conservation of European cattle breeds: an evaluation. *Animal Genetics* **37**: 475.

Marks P, Garcia S, Barrio AM, Belhocine K, Bernate J, Bharadwaj R, Bjornson K, Catalanotti C, Delaney J, Fehr A et al. 2019. Resolving the full spectrum of human genome variation using Linked-Reads. *Genome Res* **29**: 635-645. doi:10.1101/gr.234443.118.

Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, O'Connell J, Moore SS, Smith TP, Sonstegard TS et al. 2009. Development and characterization of a high density SNP genotyping assay for cattle. *PLoS One* **4**: e5350. doi:10.1371/journal.pone.0005350.

Maurice-Van Eijndhoven MH, Bovenhuis H, Veerkamp RF, Calus MP. 2015. Overlap in genomic variation associated with milk fat composition in Holstein Friesian and Dutch native dual-purpose breeds. *J Dairy Sci* **98**: 6510-6521. doi:10.3168/jds.2014-9196.

Maurice - Van Eijndhoven MHT. 2014. Genetic variation of milk fatty acid composition between and within dairy cattle breeds. Wageningen University, Wageningen.

Medugorac I, Graf A, Grohs C, Rothammer S, Zagdsuren Y, Gladyr E, Zinovieva N, Barbieri J, Seichter D, Russ I et al. 2017. Whole-genome analysis of introgressive hybridization and characterization of the bovine legacy of Mongolian yaks. *Nat Genet* **49**: 470-475. doi:10.1038/ng.3775.

Medugorac I, Medugorac A, Russ I, Veit-Kensch CE, Taberlet P, Luntz B, Mix HM, Forster M. 2009. Genetic diversity of European cattle breeds highlights the conservation value of traditional unselected breeds with high effective population size. *Mol Ecol* **18**: 3394-3410. doi:10.1111/j.1365-294X.2009.04286.x.

Meuwissen T. 2002. GENCONT: an operational tool for controlling inbreeding in selection and conservation schemes. In *Proceedings of the 7th Congress on Genetics Applied to Livestock Production*, pp. 19-23.

Meuwissen T. 2009. Genetic management of small populations: A review. *Acta Agriculturae Scandinavica, Section A - Animal Science* **59**: 71-79. doi:10.1080/09064700903118148.

Meuwissen TH. 1997. Maximizing the response of selection with a predefined rate of inbreeding. *J Anim Sci* **75**: 934-940. doi:10.2527/1997.754934x.

Meuwissen THE, Luo Z. 1992. Computing inbreeding coefficients in large populations. *Genetics Selection Evolution* **24**: 305-313. doi:10.1186/1297-9686-24-4-305.

Mi H, Muruganujan A, Huang X, Ebert D, Mills C, Guo X, Thomas PD. 2019. Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nat Protoc* **14**: 703-721. doi:10.1038/s41596-019-0128-8.

Miciak J, Fletcher JM, Stuebing KK. 2015. Accuracy and validity of methods for identifying learning disabilities in a response-to-intervention service delivery framework. In *Handbook of response to intervention: The science and practice of multi-tiered systems of support*, pp. 421-440. Springer.

Miga KH, Wang T. 2021. The Need for a Human Pangenome Reference Sequence. In *Annual Review of Genomics and Human Genetics*, Vol 22, pp. 81-102. Annual Reviews Inc.

Mill R, Nauta W. 2010. Fundamentfokkerij Fries Hollands vee. Louis Bolk Instituut, Driebergen, The Netherlands.

Milner SG, Jost M, Taketa S, Mazon ER, Himmelbach A, Oppermann M, Weise S, Knupffer H, Basterrechea M, Konig P et al. 2019. Genebank genomics highlights the diversity of a global barley collection. *Nat Genet* **51**: 319-326. doi:10.1038/s41588-018-0266-x.

Moscarelli A, Sardina MT, Cassandro M, Ciani E, Pilla F, Senczuk G, Portolano B, Mastrangelo S. 2021. Genome-wide assessment of diversity and differentiation between original and modern Brown cattle populations. *Anim Genet* **52**: 21-31. doi:10.1111/age.13019.

Muir WM, Wong GK, Zhang Y, Wang J, Groenen MA, Crooijmans RP, Megens HJ, Zhang H, Okimoto R, Vereijken A et al. 2008. Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. *Proc Natl Acad Sci U S A* **105**: 17312-17317. doi:10.1073/pnas.0806569105.

Nazareno AG, Bemmels JB, Dick CW, Lohmann LG. 2017. Minimum sample sizes for population genomics: an empirical study from an Amazonian plant species. *Mol Ecol Resour* **17**: 1136-1147. doi:10.1111/1755-0998.12654.

Neto LRP, Barendse W. 2010. Effect of SNP origin on analyses of genetic diversity in cattle. *Ani Prod Sci* **50**: 792-800. doi:10.1071/an10073.

Neuditschko M. 2011. *Eine Genomweite Populationsstrukturanalyse in Rinderrassen*. Citeseer.

Neumann GB, Korkuc P, Arends D, Wolf MJ, May K, Reissmann M, Elzaki S, Konig S, Brockmann GA. 2021. Design and performance of a bovine 200 k SNP chip developed for endangered German Black Pied cattle (DSN). *BMC Genomics* **22**: 905. doi:10.1186/s12864-021-08237-2.

Nicolazzi EL, Biffani S, Biscarini F, Orozco Ter Wengel P, Caprera A, Nazzicari N, Stella A. 2015. Software solutions for the livestock genomics SNP array revolution. *Anim Genet* **46**: 343-353. doi:10.1111/age.12295.

Nielsen R. 2004. Population genetic analysis of ascertained SNP data. *Hum Genomics* **1**: 218-224. doi:10.1186/1479-7364-1-3-218.

Nishimura S, Watanabe T, Mizoshita K, Tatsuda K, Fujita T, Watanabe N, Sugimoto Y, Takasuga A. 2012. Genome-wide association study identified three major QTL for carcass weight including the PLAG1-CHCHD7 QTN for stature in Japanese Black cattle. *BMC Genet* **13**: 40. doi:10.1186/1471-2156-13-40.

Notter DR. 1999. The importance of genetic diversity in livestock populations of the future. *J Anim Sci* **77**: 61-69. doi:10.2527/1999.77161x.

Okonechnikov K, Conesa A, Garcia-Alcalde F. 2016. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**: 292-294. doi:10.1093/bioinformatics/btv566.

Oldenbroek JK. 2019. Genetic diversity in dairy cattle: variation within and between breeds. In *Advances in breeding of dairy cattle*, pp. 39-70. Burleigh Dodds Science Publishing.

Oldenbroek JK, Windig JJ. 2022. Opportunities of Genomics for the Use of Semen Cryo-Conserved in Gene Banks. *Front Genet* **13**: 907411. doi:10.3389/fgene.2022.907411.

Oldenbroek K. 2007. *Utilisation and conservation of farm animal genetic resources*. Wageningen Academic Publishers.

Padilla JÁ, Sansinforiano E, Parejo JC, Rabasco A, Martínez-Trancón M. 2009. Inference of admixture in the endangered Blanca Cacereña bovine breed by microsatellite analyses. *Livestock science* **122**: 314-322.

Paetkau D, Calvert W, Stirling I, Strobeck C. 1995. Microsatellite analysis of population structure in Canadian polar bears. *Mol Ecol* **4**: 347-354. doi:10.1111/j.1365-294x.1995.tb00227.x.

Paiva SR, McManus CM, Blackburn H. 2016. Conservation of animal genetic resources – A new tact. *Livestock Science* **193**: 32-38. doi:10.1016/j.livsci.2016.09.010.

Pant SD, Karlskov-Mortensen P, Cirera Salicio S, Kogelman L, Jacobsen MJ, Bruun CVS, Jørgensen CB, Meuwissen THE, Kadarmideen H, Fredholm M. 2014. Genome-wide Linkage Disequilibrium Linkage Analysis (LDLA) of Body Fat Traits in an F2 Porcine Model for Human Obesity.

Pant SD, Schenkel FS, Verschoor CP, Karrow NA. 2012. Use of breed-specific single nucleotide polymorphisms to discriminate between Holstein and Jersey dairy cattle breeds. *Animal biotechnology* **23**: 1-10.

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**: 289-290. doi:10.1093/bioinformatics/btg412.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet* **2**: e190. doi:10.1371/journal.pgen.0020190.

Paul N, Kumaresan A, Das Gupta M, Nag P, Guvvala PR, Kuntareddi C, Sharma A, Selvaraju S, Datta TK. 2020. Transcriptomic Profiling of Buffalo Spermatozoa Reveals Dysregulation of Functionally Relevant mRNAs in Low-Fertile Bulls. *Front Vet Sci* **7**: 609518. doi:10.3389/fvets.2020.609518.

Perez-Enciso M, Rincon JC, Legarra A. 2015. Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. *Genet Sel Evol* **47**: 43. doi:10.1186/s12711-015-0117-5.

Pirooznia M, Kramer M, Parla J, Goes FS, Potash JB, McCombie WR, Zandi PP. 2014. Validation and assessment of variant calling pipelines for next-generation sequencing. *Hum Genomics* **8**: 14. doi:10.1186/1479-7364-8-14.

Porras-Hurtado L, Ruiz Y, Santos C, Phillips C, Carracedo A, Lareu MV. 2013. An overview of STRUCTURE: applications, parameter settings, and supporting software. *Front Genet* **4**: 98. doi:10.3389/fgene.2013.00098.

Porter V. 2002. *Mason's World Dictionary of Livestock Breeds, Types and Varieties*. CABI Pub.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**: 904-909. doi:10.1038/ng1847.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**: 945-959. doi:10.1093/genetics/155.2.945.

Pryce JE, Johnston J, Hayes BJ, Sahana G, Weigel KA, 2nd, McParland S, Spurlock D, Krattenmacher N, Spelman RJ, Wall E et al. 2014. Imputation of genotypes from low density (50,000 markers) to high density (700,000 markers) of cows from research herds in Europe, North America, and Australasia using 2 reference populations. *J Dairy Sci* **97**: 1799-1811. doi:10.3168/jds.2013-7368.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**: 559-575. doi:10.1086/519795.

R Core Team R. 2013. R: A language and environment for statistical computing.

Rafiepour M, Ebrahimie E, Vahidi MF, Salekdeh GH, Niazi A, Dadpasand M, Liang D, Si J, Ding X, Han J et al. 2021. Whole-Genome Resequencing Reveals Adaptation Prior to the Divergence of Buffalo Subspecies. *Genome Biol Evol* **13**: 1-14. doi:10.1093/gbe/evaa231.

Ramos AM, Megens HJ, Crooijmans RP, Schook LB, Groenen MA. 2011. Identification of high utility SNPs for population assignment and traceability purposes in the pig using high-throughput sequencing. *Anim Genet* **42**: 613-620. doi:10.1111/j.1365-2052.2011.02198.x.

Restoux G, Rognon XX, Vieaud A, Guéméné D, Chiron G, Petitjean F, Seigneurin F, Vasilescu A, Tixier-Boichard M. 2018. Genetic characterization of french local chicken breeds. In *11 World Congress on Genetics Applied to Livestock Production*, Auckland, New Zealand.

Reynolds EGM, Neeley C, Lopdell TJ, Keehan M, Dittmer K, Harland CS, Couldrey C, Johnson TJJ, Tiplady K, Worth G et al. 2021. Non-additive association analysis using proxy phenotypes identifies novel cattle syndromes. *Nat Genet* **53**: 949-954. doi:10.1038/s41588-021-00872-5.

Roberts KS, Lamberson WR. 2015. Relationships among and variation within rare breeds of swine. *J Anim Sci* **93**: 3810-3813. doi:10.2527/jas.2015-9001.

Rogberg-Munoz A, Wei S, Ripoli MV, Guo BL, Carino MH, Castillo N, Villegas Castagnaso EE, Liron JP, Morales Durand HF, Melucci L et al. 2014. Foreign meat identification by DNA breed assignment for the Chinese market. *Meat Sci* **98**: 822-827. doi:10.1016/j.meatsci.2014.07.028.

Ros-Freixedes R, Gol S, Pena RN, Tor M, Ibanez-Escriche N, Dekkers JC, Estany J. 2016. Genome-Wide Association Study Singles Out SCD and LEPR as the Two Main Loci Influencing Intramuscular Fat Content and Fatty Acid Composition in Duroc Pigs. *PLoS One* **11**: e0152496. doi:10.1371/journal.pone.0152496.

Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, Rowan TN, Low WY, Zimin A, Couldrey C et al. 2020. De novo assembly of the cattle reference genome with single-molecule sequencing. *GigaScience* **9**. doi:10.1093/gigascience/giaa021.

Rosenberg NA, Burke T, Elo K, Feldman MW, Freidlin PJ, Groenen MA, Hillel J, Maki-Tanila A, Tixier-Boichard M, Vignal A et al. 2001. Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. *Genetics* **159**: 699-713. doi:10.1093/genetics/159.2.699.

Rosenberg NA, Li LM, Ward R, Pritchard JK. 2003. Informativeness of genetic markers for inference of ancestry. *Am J Hum Genet* **73**: 1402-1422. doi:10.1086/380416.

Rubin CJ, Zody MC, Eriksson J, Meadows JR, Sherwood E, Webster MT, Jiang L, Ingman M, Sharpe T, Ka S et al. 2010. Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**: 587-591. doi:10.1038/nature08832.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**: 406-425. doi:10.1093/oxfordjournals.molbev.a040454.

Sansaloni C, Franco J, Santos B, Percival-Alwyn L, Singh S, Petroli C, Campos J, Dreher K, Payne T, Marshall D et al. 2020. Diversity analysis of 80,000 wheat accessions reveals consequences and opportunities of selection footprints. *Nat Commun* **11**: 4572. doi:10.1038/s41467-020-18404-w.

Schmidt TL, Jasper M, Weeks AR, Hoffmann AA. 2021. Unbiased population heterozygosity estimates from genome-wide sequence data. *bioRxiv* doi:10.1101/2020.12.20.423694. doi:10.1101/2020.12.20.423694.

Sheikhizadeh Anari S. 2020. Towards comparative pan-genomics. Wageningen University, Wageningen.

Sherman RM, Salzberg SL. 2020. Pan-genomics in the human genome era. *Nat Rev Genet* **21**: 243-254. doi:10.1038/s41576-020-0210-7.

Simianer H, Marti SB, Gibson J, Hanotte O, Rege JEO. 2003. An approach to the optimal allocation of conservation funds to minimize loss of genetic diversity between livestock breeds. *Ecol Econ* **45**: 377-392. doi:10.1016/s0921-8009(03)00092-2.

Slaghuis H. 2009. De geschiedenis van de varkensfokkerij in Nederland. *Argos: bulletin van het Veterinair Historische Genootschap*: 14-19.

Slatko BE, Gardner AF, Ausubel FM. 2018. Overview of Next-Generation Sequencing Technologies. *Curr Protoc Mol Biol* **122**: e59. doi:10.1002/cpmb.59.

Smith JL, Wilson ML, Nilson SM, Rowan TN, Oldeschulte DL, Schnabel RD, Decker JE, Seabury CM. 2019. Genome-wide association and genotype by environment interactions for growth traits in U.S. Gelbvieh cattle. *BMC Genomics* **20**: 926. doi:10.1186/s12864-019-6231-y.

Sölkner J, Ferencakovic M, Karimi Z, OBrien AMP, Mészáros G, Eaglen SAE, Boison SA, Curik I. 2014. Extremely Non-uniform: Patterns of Runs of Homozygosity in Bovine Populations.

Stervander M, Cresko WA. 2021. A highly contiguous nuclear genome assembly of the mandarinfish Synchiropus splendidus (Syngnathiformes: Callionymidae). *G3 (Bethesda)* **11**. doi:10.1093/g3journal/jkab306.

Talenti A, Powell J, Hemmink JD, Cook EAJ, Wragg D, Jayaraman S, Paxton E, Ezeasor C, Obishakin ET, Agusi ER et al. 2022. A cattle graph genome incorporating global breed diversity. *Nat Commun* **13**: 910. doi:10.1038/s41467-022-28605-0.

Talle SB, Fimland E, Syrstad O, Meuwissen T, Klungland H. 2005. Comparison of individual assignment methods and factors affecting assignment success in cattle breeds using microsatellites. *Acta Agriculturae Scandinavica, Section A-Animal Science* **55**: 74-79.

Te Pas MF, Madsen O, Calus MP, Smits MA. 2017. The Importance of Endophenotypes to Evaluate the Relationship between Genotype and External Phenotype. *Int J Mol Sci* **18**. doi:10.3390/ijms18020472.

Terenina E, Bazovkina D, Rousseau S, Salin F, D'Eath R, Turner S, Kulikov A, Mormede P. 2012. Gene polymorphisms associated with aggression in pigs. *Journurn de la Recherche Porcine en France* **44**: 45-46.

Theunissen B. 2008. Een mooie koe is een goede koe. Wetenschappers en practici over de Nederlandse rundveefokkerij, 1900-1950. *Studium* **1**: 47-61. doi:10.18352/studium.1454.

Theunissen B. 2012. Breeding for nobility or for production? Cultures of dairy cattle breeding in the Netherlands, 1945-1995. *ISIS* **103**: 278-309. doi:10.1086/666356.

Tian X, Li R, Fu W, Li Y, Wang X, Li M, Du D, Tang Q, Cai Y, Long Y et al. 2020. Building a sequence map of the pig pan-genome from multiple de novo assemblies and Hi-C data. *Sci China Life Sci* **63**: 750-763. doi:10.1007/s11427-019-9551-7.

Tixier-Boichard M, Peynot N, Duclos D, Weigend S, Monteagudo L, Martinez Martinez A, Delgado JV, Gonzalez Prendes R, Crooijmans R, Restoux G. 2022. Mapping genetic diversity in European gene banks: preliminary results on chickens for the validation of IMAGE001 array In *World Congress on Genetics Applied to Livestock Production (WCGALP)*, p. 4, Rotterdam. The Netherlands.

Turner SD. 2014. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *bioRxiv* doi:10.1101/005165: 005165. doi:10.1101/005165.

Upadhyay M, Eriksson S, Mikko S, Strandberg E, Stalhammar H, Groenen MAM, Crooijmans R, Andersson G, Johansson AM. 2019. Genomic relatedness and diversity of Swedish native cattle breeds. *Genet Sel Evol* **51**: 56. doi:10.1186/s12711-019-0496-0.

van Breukelen AE, Doekes HP, Windig JJ, Oldenbroek K. 2019. Characterization of Genetic Diversity Conserved in the Gene Bank for Dutch Cattle Breeds. *Diversity* **11**: 1-13. doi:10.3390/d11120229.

van Treuren R, van Hintum TJL. 2014. Next-generation genebanking: plant genetic resources management and utilization in the sequencing era. *Plant Genetic Resources* **12**: 298-307. doi:10.1017/s1479262114000082.

Vohra V, Chhotaray S, Gowane G, Alex R, Mukherjee A, Verma A, Deb SM. 2021. Genome-Wide Association Studies in Indian Buffalo Revealed Genomic Regions for Lactation and Fertility. *Front Genet* **12**: 696109. doi:10.3389/fgene.2021.696109.

Wang T, Antonacci-Fulton L, Howe K, Lawson HA, Lucas JK, Phillippy AM, Popejoy AB, Asri M, Carson C, Chaisson MJP et al. 2022. The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**: 437-446. doi:10.1038/s41586-022-04601-8.

Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR, Zhang F et al. 2018. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* **557**: 43-49. doi:10.1038/s41586-018-0063-9.

Wang X, Liu X, Deng D, Yu M, Li X. 2016. Genetic determinants of pig birth weight variability. *BMC Genet* **17 Suppl 1**: 15. doi:10.1186/s12863-015-0309-6.

Wang Y, Nielsen R. 2012. Estimating population divergence time and phylogeny from single-nucleotide polymorphisms data with outgroup ascertainment bias. *Mol Ecol* **21**: 974-986. doi:10.1111/j.1365-294X.2011.05413.x.

Wang Y, Ning C, Wang C, Guo J, Wang J, Wu Y. 2019. Genome-wide association study for intramuscular fat content in Chinese Lulai black pigs. *Asian-Australas J Anim Sci* **32**: 607-613. doi:10.5713/ajas.18.0483.

Warr A, Affara N, Aken B, Beiki H, Bickhart DM, Billis K, Chow W, Eory L, Finlayson HA, Flicek P et al. 2020. An improved pig reference genome sequence to enable pig genetics and genomics research. *GigaScience* **9**. doi:10.1093/gigascience/giaa051.

Weir BS, Cockerham CC. 1984. Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38**: 1358-1370. doi:10.1111/j.1558-5646.1984.tb05657.x.

Wilkinson S, Haley C, Alderson L, Wiener P. 2011a. An empirical assessment of individual-based population genetic statistical techniques: application to British pig breeds. *Heredity (Edinb)* **106**: 261-269. doi:10.1038/hdy.2010.80.

Wilkinson S, Wiener P, Archibald AL, Law A, Schnabel RD, McKay SD, Taylor JF, Ogden R. 2011b. Evaluation of approaches for identifying population informative

markers from high density SNP chips. *BMC Genet* **12**: 45. doi:10.1186/1471-2156-12-45.

Willing EM, Dreyer C, van Oosterhout C. 2012. Estimates of genetic differentiation measured by F(ST) do not necessarily require large sample sizes when using many SNP markers. *PLoS One* **7**: e42649. doi:10.1371/journal.pone.0042649.

Wilmot H, Bormann J, Soyeurt H, Hubin X, Glorieux G, Mayeres P, Bertozzi C, Gengler N. 2022a. Development of a genomic tool for breed assignment by comparison of different classification models: Application to three local cattle breeds. *J Anim Breed Genet* **139**: 40-61. doi:10.1111/jbg.12643.

Wilmot H, Glorieux G, Hubin X, Gengler N. 2022b. A genomic breed assignment test for traceability of meat of Dual-Purpose Blue. *Livestock Science* **263**. doi:10.1016/j.livsci.2022.104996.

Wong KHY, Levy-Sakin M, Kwok PY. 2018. De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. *Nat Commun* **9**: 3040. doi:10.1038/s41467-018-05513-w.

Woolliams J, Toro M. 2007. What is genetic diversity? In *Utilisation and Conservation of Farm Animal Genetic Resources*, doi:10.3920/978-90-8686-592-5, pp. 55-74. Wageningen Academic Publishers.

Woolliams JA, Oldenbroek JK. 2017. Chapter 1. Genetic diversity issues in animal populations in the genomic era. In *Genomic management of animal genetic diversity*, doi:10.3920/978-90-8686-850-6_1, pp. 13-47.

Xu L, Zhang WG, Shen HX, Zhang Y, Zhao YM, Jia YT, Gao X, Zhu B, Xu LY, Zhang LP et al. 2018. Genome-wide scanning reveals genetic diversity and signatures of selection in Chinese indigenous cattle breeds. *Livestock Science* **216**: 100-108. doi:10.1016/j.livsci.2018.08.005.

Xu L, Zhao G, Yang L, Zhu B, Chen Y, Zhang L, Gao X, Gao H, Liu GE, Li J. 2019. Genomic Patterns of Homozygosity in Chinese Local Cattle. *Sci Rep* **9**: 16977. doi:10.1038/s41598-019-53274-3.

Yang BZ, Zhao H, Kranzler HR, Gelernter J. 2005. Practical population group assignment with selected informative markers: characteristics and properties of Bayesian clustering via STRUCTURE. *Genet Epidemiol* **28**: 302-312. doi:10.1002/gepi.20070.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW et al. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**: 565-569. doi:10.1038/ng.608.

Yu Y, Zhang Y, Song X, Jin M, Guan Q, Zhang Q, Li S, Wei C, Lu G, Zhang J et al. 2010. Alternative splicing and tissue expression of CIB4 gene in sheep testis. *Anim Reprod Sci* **120**: 1-9. doi:10.1016/j.anireprosci.2010.01.004.

Zaborski D, Grzesiak W, Pilarczyk R. 2014. Detection of difficult calvings in the Polish Holstein-Friesian Black-and-White heifers. *J Appl Anim Res* **44**: 42-53. doi:10.1080/09712119.2014.987293.

Zhang Q, Calus MP, Guldbrandtsen B, Lund MS, Sahana G. 2015. Estimation of inbreeding using pedigree, 50k SNP chip genotypes and full sequence data in three cattle breeds. *BMC Genet* **16**: 88. doi:10.1186/s12863-015-0227-7.

Zhang Q, Calus MPL, Guldbrandtsen B, Lund MS, Sahana G. 2017. Contribution of rare and low-frequency whole-genome sequence variants to complex traits variation in dairy cattle. *Genet Sel Evol* **49**: 60. doi:10.1186/s12711-017-0336-z.

Zhao Q, Feng Q, Lu H, Li Y, Wang A, Tian Q, Zhan Q, Lu Y, Zhang L, Huang T et al. 2018. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat Genet* **50**: 278-284. doi:10.1038/s41588-018-0041-z.

Zhao X, Zhao K, Ren J, Zhang F, Jiang C, Hong Y, Jiang K, Yang Q, Wang C, Ding N et al. 2016. An imputation-based genome-wide association study on traits related to male reproduction in a White Duroc x Erhualian F2 population. *Anim Sci J* **87**: 646-654. doi:10.1111/asj.12468.

Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**: 3326-3328. doi:10.1093/bioinformatics/bts606.

Zhi D, Da L, Liu M, Cheng C, Zhang Y, Wang X, Li X, Tian Z, Yang Y, He T et al. 2018. Whole Genome Sequencing of Hulunbuir Short-Tailed Sheep for Identifying Candidate Genes Related to the Short-Tail Phenotype. *G3 (Bethesda)* **8**: 377-383. doi:10.1534/g3.117.300307.

Zinovieva NA, Dotsev AV, Sermyagin AA, Deniskova TE, Abdelmanova AS, Kharzinova VR, Solkner J, Reyer H, Wimmers K, Brem G. 2020. Selection signatures in two oldest Russian native cattle breeds revealed using high-density single nucleotide polymorphism analysis. *PLoS One* **15**: e0242200. doi:10.1371/journal.pone.0242200.

Summary

In this thesis, I apply genomics into conservation practises. Conserving genetic diversity is essential for the sustainability of populations. In livestock, the amount of genetic diversity should be large enough to enable the adaptation of populations to changing environments and market requirements, and for selection to genetically improve important traits. Unfortunately, the current trend in populations is often for a reduced genetic diversity due to intense selection or random drift. Consequently, breeding methods and gene banks were developed to avoid the risk of losing genetic diversity. As genomic information becomes more accessible, we now have the option to better manage genetic diversity. In this thesis, I used genomic tools and methods to conservation of Dutch livestock breeds and thereby improve the understanding of the genetic diversity conserved in gene banks and in living populations of numerically small breeds.

In **Chapter 2**, an investigation was conducted on how to deal with lines or supposed lines within a numerically small breed and what the consequences may be for conservation. The genetic diversity within the Dutch Red and White Friesian cattle (DFR), a local Dutch breed was evaluated with genomics. I demonstrated that the several different lines within the DFR all contribute to the diversity of the breed. Moreover, the results of this study revealed a high level of admixture between 5 of the 7 lines. This reflects the similar origin of these lines. Consequently, there seems to be no necessity to conserve these 5 lines separately, because their level of differentiation is very low. The other two lines contained unique diversity. However, in one line this was mainly due to introgressed Holstein Friesian blood and was therefore of lower conservation value. The other line was bred for several generations in isolation from other lines. Although it was highly inbred and had a low level of diversity within the line, it had apparently conserved genetic diversity not present anymore in the rest of the population. This study also illustrates that, when taking conservation decisions for a breed, it is worthwhile to take into account both the population structure of the breed itself and the relationships with other breeds.

In **Chapter 3**, a genomics analysis was performed to investigate the implications for the genetic diversity of merging and terminating lines of the Dutch Landrace pig breed. Over time, lines have been merged and some discontinued, due to consolidation in the pig breeding industry. The merging of lines did reduce the overall genetic diversity in the Landrace population in the Netherlands, albeit a large proportion of the original variation is maintained in the current population. The original lines have been conserved in the gene bank and are now used by hobby

breeders to recreate the original Dutch landrace pig. This shows the important role gene banks can have for commercial breeds as well. Especially when large changes, such as merging selection lines are about to occur, it is important to conserve current genetic diversity in a gene bank.

In **Chapter 4**, Whole Genome Sequencing (WGS) was used to evaluate the consequences for the whole genome and especially for its rare allelic variants of the replacement of the Dutch Friesian cattle by the Holstein Friesian breed. I compared genome-wide genetic diversity in three groups of bulls, chosen from the historic (1961–1989) (hDF) and recent (2003–2015) Dutch Friesian cattle (rDF) population and the recent Holstein Friesian (rHF) (1998–2014) population. The Dutch Friesian cattle was the dominant cattle breed in the Netherlands before the 1990s. Since then it has gradually been replaced by the HF and currently is a rare breed. Over the last century, genetic diversity in the cattle species has been affected by the replacement of many local, dual-purpose breeds with one or a few specialized, high input-high output dairy breeds. Analysis of the WGS data indicated a large overlap of genetic diversity between the three groups due to their common history. However, each of the three groups has a number of group-specific SNPs. The two DF groups are genetically clearly different from the rHF group. The genetic difference between the rDF and rHF is slightly larger than that between the hDF and rHF. The genetic diversity of the DF breed reduced over time, but this did not lead to higher inbreeding levels—especially, inbreeding due to recent ancestors has not increased. The results also highlighted the presence of several genomic regions that differentiated between the groups. The DNA regions that clearly differ are related to traits such as fertility and weight.

In **Chapter 5**, I evaluated different methods of SNP selection, using either the BovineSNP50 BeadChip or the BovineHD 777K BeadChip, in terms of the minimum required number of informative SNP to differentiate among local Dutch cattle breeds. In this study, I evaluated Delta, Wright's F_{ST}, and Weir and Cockerham's F_{ST}, and extended these methods by adding a rule to avoid selection of sets of SNPs in high linkage disequilibrium providing the same information. Only a small set of SNPs (≤37) is needed to differentiate among four Dutch cattle breeds regardless of the selection methods, and selection methods showed only small differences. The 777K BeadChip performed marginally better than the 50K BeadChip. The Global Weir and Cockerham's F_{ST} performed marginally better than other selection methods. The rule

to avoid selection of SNPs in high LD further reduced the required number of SNPs to achieve correct breed assignment.

In **Chapter 6**, the development of a genetic tool for determining breed purity of Dutch cattle breeds is described. Breed registries have been established for livestock species to maintain the purity of breeds and to document the ancestry of animals. However, a significant number of animals can be unregistered with no or incomplete pedigree data and an uncertain ancestral breed origin. A genetic test was developed to unequivocally determine the breed origin of cattle without pedigree data. Reference populations for the six Dutch cattle breeds were constructed based on genotype data (50K SNP array). A combine approach of Principal Component Analysis and Random Forest was used to perform SNP selection, using the genotype data of the reference populations. A total of 133 informative SNPs were selected to determine breed composition of individual animals. The developed test was successful and is implemented in practice to identify (partly) unregistered individuals as being purebred (or not) for one of the Dutch cattle breeds.

In the general discussion (**Chapter 7**), I addressed the recent developments in genomics and how they can be used effectively for genetic conservation, and in particular how gene banks can benefit from these developments, and I outline possible future directions for (a more effective) conservation of breeds using genomic methods. The discussion included the applications from SNP arrays to WGS, and from first to third generation sequencing, as well as the new concept of pan-genomics. I also argued the need for gene banks to transform from “traditional gene banks” to “digital gene banks”.

Conclusive remarks: Several SNP arrays including low- to high-density have been developed. However, SNP arrays are known to lack a substantial proportion of globally rare variants and tend to be biased towards variants present in the breeds that were involved in the development process of the SNP arrays. Inherently, informative, and breed-specific variants segregating in various local breeds have not been considered through this ascertainment bias. Compared to SNP arrays, the use of WGS provides various advantages in assessing genetic diversity, such as the avoidance of ascertainment bias and a significantly higher information content. WGS may help to identify unique genetic variation in breeds, which is an important characteristic in the prioritization of breeds for conservation. Continuous improvement in sequencing technology implies, that the whole genome can be

sequenced faster, easier and with a higher accuracy since the start of Sanger sequencing (mid-1970s). In the coming years, new sequencing platforms will probably appear producing a larger amount of data which requires the development of new approaches and applications capable of analysing this large amount of data. Sequencing a large number of individuals is still expensive and analyses require expertise and expensive hardware. Therefore, it is good to realise that WGS is not necessary for every purpose related to the use and conservation of genetic variation. For example, only a small set of informative SNPs is needed to differentiate among Dutch local cattle breeds.

Genetic diversity of populations is constantly changing. All breeds tend to be under selection and genetic drift; as a result, after several generations, differences will emerge, in allele and haplotype frequencies. As a consequence, differences arise constantly between the gene bank collection and the in situ population. This necessitates the resampling of breeds over time. As a result of periodic resampling, gene banks pools capture more genetic diversity than in situ populations that are under selection and drift. Genomics helps to gain insight into genetic diversity and to better understand the dynamics of genetic diversity in populations through time. Therefore, gene banks, which focus on preserving collections, should also invest in collecting genomic and relevant phenotypic information on the stored material in the gene bank, and to make this information accessible. Efforts should be made by gene banks to transform from “traditional gene banks” to “digital gene banks”, which complement the conservation of materials with associated information from various data domains. This will provide better and more targeted access to the material and increase its use for the genomic management of small and large populations.

Curriculum Vitae

About the author

Ina Hulsegge was born on 10 December 1961 in Raalte, The Netherlands and grew up on a farm. She received in 1981 her HAVO degree from the Florens Radewijns College in Raalte. She studied at the Rijks Hogere Landbouwschool (National Agricultural College, predecessor of University of Applied Sciences Van Hall Larenstein) in Deventer. During which she did an internship at the Department of Animal Husbandry of Wageningen UR. It was during this internship where she began to develop an interest in agricultural research. In 1986, she started working at the Research Institute for Animal Production (Instituut voor Veeteeltkundig Onderzoek) in Zeist, a predecessor of Wageningen Livestock Research (WLR). During her working career, Ina worked on different topics: such as animal transport, meat and carcass quality, animal welfare, genomics and machine learning and computer vision. Over time, Ina has become more and more specialized in data analyses and bioinformatics. This motivated Ina to combine her work at WLR with a MSc Bioinformatics study at Wageningen University during 2007 and 2010. Ina started her PhD in Animal Breeding and Genomics at Wageningen University in 2020, which resulted in this thesis.

Publications

Ina Hulsegge has published more than 45 international peer reviewed publications on different topics, such as animal transport, carcass quality, animal welfare and genomics. Next to the publications that form the basis of this thesis Ina (co-)authored the following publications in the field of genomics and rare breeds:

Wilmot H, Druet T, Hulsegge I, Gengler N, Calus, M.P.L. 2023. Estimation of inbreeding, between-breed genomic relatedness and definition of sub-populations in red-pied cattle breeds. *Animal* **17**: 100793 .

Windig JJ, Hulsegge I. 2021. Retriever and Pointer: Software to Evaluate Inbreeding and Genetic Management in Captive Populations. *Animals* 11(5):1332.

Marjanovic J, Hulsegge B, Calus M.P.L. 2021. Relatedness between numerically small Dutch Red dairy cattle populations and possibilities for multibreed genomic prediction. *Journal of Dairy Science* **104**: 4498-4506.

Calus M.P.L, Vandenplas J, Hulsegge I, Borg R, Henshall J.M, Hawken R. 2019. Assessment of sire contribution and breed-of-origin of alleles in a three-way crossbred broiler dataset. *Poultry Science* **98**(12):6270-6280.

Eynard S.E, Windig JJ, Hulsegge I, Hiemstra S.J, Calus M.P.L. 2018. The impact of using old germplasm on genetic merit and diversity-A cattle breed case study. *Journal of Animal Breeding and Genetics* **135**: 311-322

François L, Wijnrocx K, Colinet F.G, Gengler N, Hulsegge B, Windig J.J, Buys N, Janssens S. 2017. Genomics of a revived breed: Case study of the Belgian campine cattle. *PLoS One* **12**: e0175916.

Daetwyler, H.D. Capitan, A. Pausch, H. Stothard, P. van Binsbergen, R. Brøndum, R.F., Liao, X. Djari, A. Rodriguez, S.C. Grohs, C. Esquerré, D. Bouchez, O. Rossignol, M.N, Klopp, C. Rocha, D. Fritz, S. Eggen, A., Bowman, P.J, Coote, D, Chamberlain, A.J, Anderson, C, VanTassell, C.P, Hulsegge, I, Goddard, M.E, Guldbrandtsen, B, Lund, M.S, Veerkamp, R.F, Boichard, D.A, Fries, R, Hayes, B.J. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nature Genetics 46:858-65.

van Binsbergen R, Bink M.C, Calus M.P, van Eeuwijk F.A, Hayes B.J, Hulsegge I, Veerkamp R.F. 2014. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. *Genetics Selection Evolution* **46**(1):41.

Training and Education

The Basic Package (2.9 credits)	
WIAS Introduction Day	2020
WGS Scientific Integrity course	2020
WGS Ethics and Animal Sciences course	2020
WIAS Course: Introduction to Personal Effectiveness	2020

Disciplinary Competences (15.6 credits)	
WIAS research proposal	2019-2020
Physalia-course: Adaptation Genomics	2022
WIAS Image Analysis Course	2021
Physalia-course: Introductory Population Genomics	2017
ALLBIO and DEANN Training course: Bioinformatics approaches to identify causative sequence variants in farm animals	2014
WIAS course: Genetic analysis using ASRemL 4.0	2014
WIAS course: Course on characterization, management and exploitation of genomic diversity in animals	2019

Professional Competences (4.0 credits)	
Wageningen in'to Languages course: Scientific Writing	2021
WIAS Course: The Final Touch: Writing the General Introduction and Discussion	2021
WGS course: Supervising BSc and MSc thesis students	2021
WGS course: Mindful Productivity for scientists	2022
WGS course: Mobilising your - scientific - network	2021

Societal Relevance (1.5 credits)	
WIAS course: Societal Impact of your research	2021

Presentation Skills (4.0 credits)	
European association of Animal Production (EAAP) (Presentation)	2016
European association of Animal Production (EAAP) (Presentation)	2018
Biodiversity Genomics Meeting 2020 (Poster)	2020
European association of Animal Production (EAAP) (Poster)	2021
World Congress on Genetics Applied to Livestock Production (pitch)	2022

Teaching competences (4.5 credits)	
Animal Breeding and Genomics, Msc student Minor	2018
Animal Breeding and Genomics, Msc student Major	2017
Animal Breeding and Genomics, Bsc student	2019

Total credits	**32.5**

Acknowledgements

The journey that led to this thesis today took over 37 years. I have met and worked with many people during these years at Wageningen Livestock Research and it's precursors. It is only now that I start to realize that all of them were important for finishing this thesis, and even more for the development and growth of myself as a person. It would be impossible to thank all these people individually. Therefore, to all of you who joined me on this ride, whether it was just for a brief moment or for a longer period of time: 'thank you'. Thank you for your help, support, encouragement, guidance, and advice. Also, thank you for the morning coffee moments, the "bel-dates", and the lunch walks. Please know that I have the deepest appreciation and respect for everything that you have done for me. It is because of you that I am where I am today.

As mentioned, I cannot thank all of you individually, but there are a few people that I would like to name specifically. These are my PhD supervisors Roel Veerkamp, Jack Windig, Aniek Bouwman and Kor Oldenbroek. Their support, insightful feedback, and guidance have been invaluable throughout the entire process. I am profoundly grateful for the contributions they made to my development. I am incredibly thankful for the opportunity I have been given to do this.

Special thanks to my paranymphs, Yvette de Haas and Claudia Kamphuis, for agreeing to be my paranymphs and for their support at my defence.

Finally, and most importantly, I owe my deepest gratitude to my family for their love and support.

Ina

Colophon

The research described in this thesis was financially supported by the Ministry of Agriculture, Nature and Food Quality, program ‘Kennisbasis Dier’, code KB-34-013-002, The Centre for Genetic Resources, The Netherlands (CGN) of Wageningen University and Research, and Wageningen Livestock Research of Wageningen University.

Printed by DigiForce | Proefschriftmaken.nl