Livestock Science 223 (2019) 60–67

Contents lists available at ScienceDirect

Livestock Science

journal homepage: www.elsevier.com/locate/livsci

Development of a genetic tool for determining breed purity of cattle

Ina Hulsegge[a,b,*], Mira Schoon[a], Jack Windig[a,b], Marjolein Neuteboom[c], Sipke Joost Hiemstra[a,b], Anouk Schurink[a,b]

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

ARTICLE INFO

Keywords:
Genetic test
Breed purity
Assignment
SNP
Cattle breeds

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 herd books 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 criteria 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.

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 herd book. 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

* Corresponding author to: Animal Breeding and Genomics, Wageningen Livestock Research, P.O. Box 338, 6700 AH, Wageningen, The Netherlands.
E-mail addresses: ina.hulsegge@wur.nl (I. Hulsegge), jack.windig@wur.nl (J. Windig), marjolein.neuteboom@wur.nl (M. Neuteboom), sipkejoost.hiemstra@wur.nl (S.J. Hiemstra), anouk.schurink@wur.nl (A. Schurink).

https://doi.org/10.1016/j.livsci.2019.03.002
Received 18 October 2018; Received in revised form 28 January 2019; Accepted 6 March 2019

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.

2. Materials and methods

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; François et al., 2017; Hulsegge et al., 2017; Manzanilla-Pech et al., 2017) and the recently available genotype data from bulls in the Dutch genebank, 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 genebank 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).

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, 2008). Imputation was carried out for each breed and chromosome independently, except for the Holstein Friesian samples which were already imputated. 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.

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 1). The remaining animals formed the reference population (Table 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 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 1).

Table 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

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 (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.

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.

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.

3. Results

3.1. Selection of informative SNPs

The first three PCs separated the 572 individual animals from the reference population according to their breed (Fig. 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 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 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.

Fig. 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 1. The selected 133 informative SNPs were located across all chromosomes, where the number of SNPs per chromosome ranged from one to 12.

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 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 3). A considerable average proportion of membership (0.227; Table 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 3). The

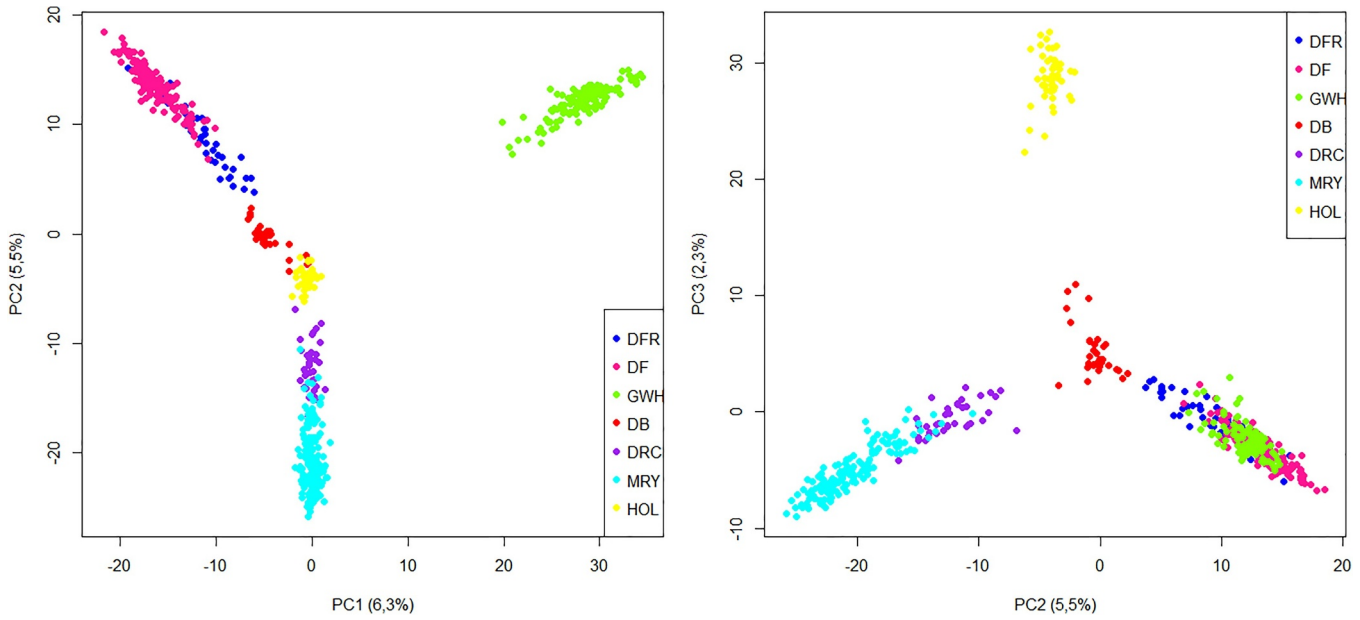

Fig. 1. PCA results visualizing individuals of various breeds within the reference population using 10,449 SNPS, with the percentage of variance explained in brackets.

Table 2
Assignment of reference population animals to breeds based on Random Forest classification using 976 and 133 SNPs.

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

*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.

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.

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 (Fig. 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 (Fig. 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 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 $\geq$0.775, the animal can be considered as purebred, provided that the phenotype, colour and/or pattern and meets the requirements for the breed, as determined by the herdbook. Using this criteria 88.9% (104 out of 117) of the animals in the test population is correctly assigned. Of the 34 purebred animals ($\geq$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

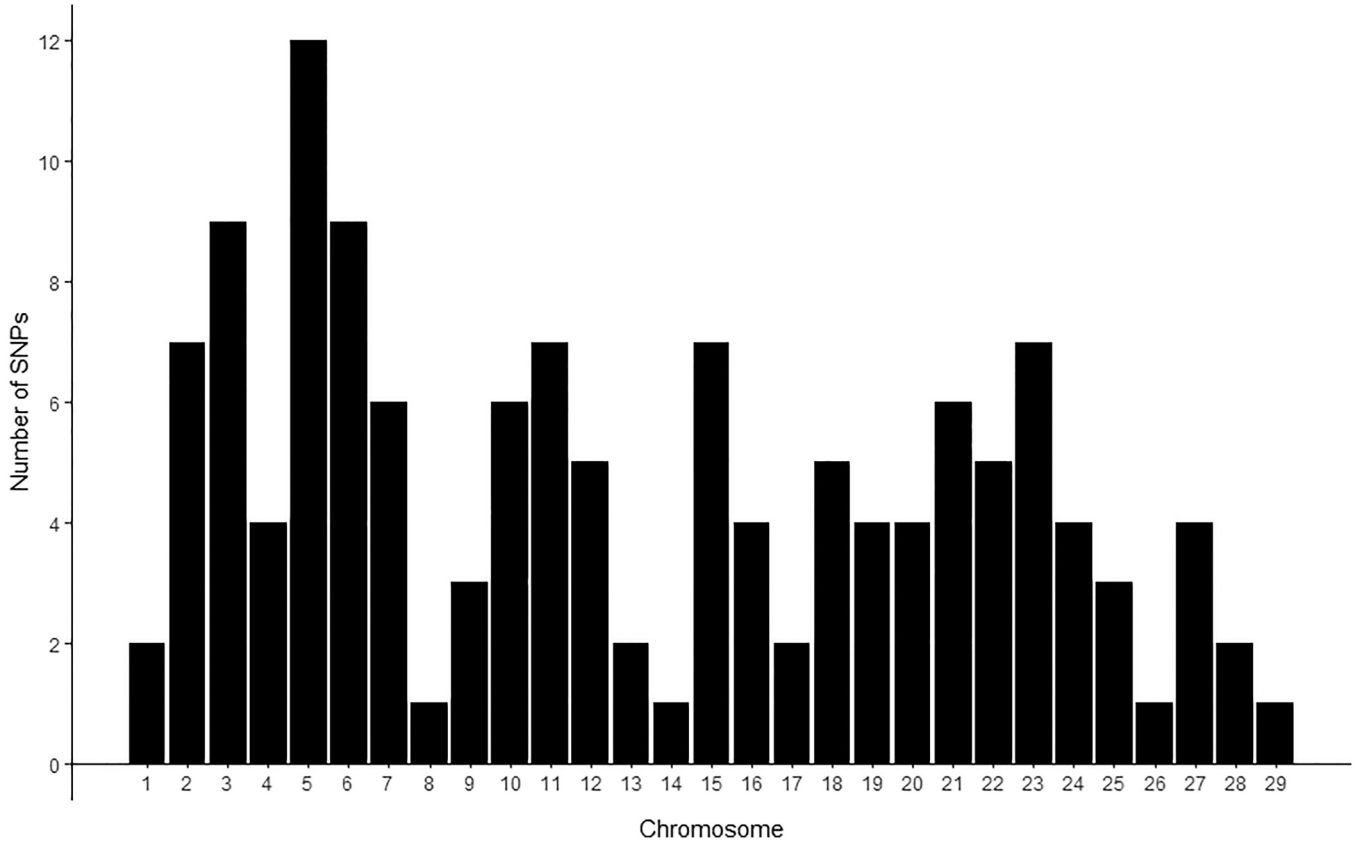

Fig. 2. Distribution of the 133 SNPs over the chromosomes.

Table 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							
	1	2	3	4	5	6	7	Number of animals
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

* 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.

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 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 5). It is very plausible that the unknown breed Dutch Red and White Friesian breed or Dutch Friesian was.

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.

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.

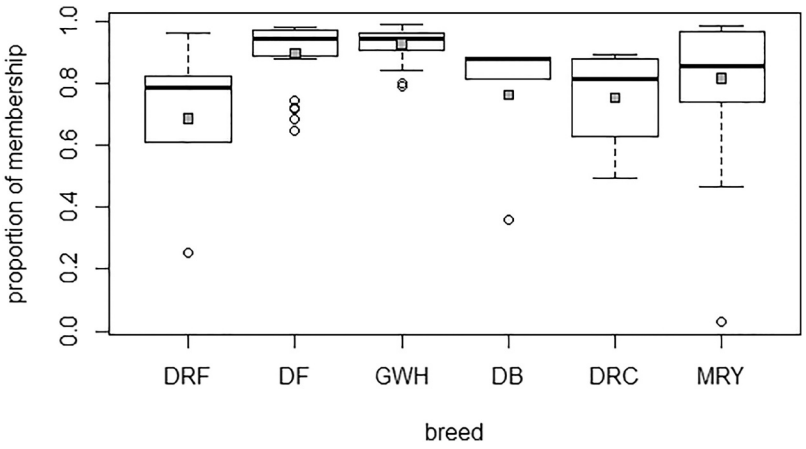

Fig. 3. Boxplot of the breed proportion of membership for the six local cattle breeds. (grey square = average proportion of membership; Dutch Red and White Friesian (DRF) $n = 5$; Dutch Friesian (DF) $n = 23$; Groningen White Headed (GWH) $n = 25$; Dutch Belted (DB) $n = 5$; Deep Red Cattle (DRC) $n = 4$ and Meuse-Rhine-Yssel (MRY) $n = 34$).

Table 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.
** 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.
*** 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 HOL = Holstein Friesian.

Table 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

* 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.

4.2. Reference population

The genotype data available for this study was not specifically gathered to build a reference populations 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; François 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 genebank 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.

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., 2011; 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., 2011; 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.

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-Muñoz 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.

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 breed 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.

5. Conclusion

Although tens of thousands of SNP markers are now available, only a small set of SNP (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 represent 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.

References

Berg, P., Windig, J.J., 2017. Management of cryo-collections with genomic tools.

Bertolini, F., Galimberti, G., Calò, D.G., 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. Animal Breed. Genet. 132, 346–356.

Bertolini, F., Galimberti, G., Schiavo, G., Mastrangelo, S., Di Gerlando, R., Strillacci, M.G., 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.

Boulesteix, A.L., Bender, A., Bermejo, J.L., Strobl, C., 2012. Random forest Gini importance favours SNPs with large minor allele frequency: impact, sources and recommendations. Brief. Bioinform. 13, 292–304.

Browning, B.L., Browning, S.R., 2008. 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.

Connolly, S., Fortes, M.R.S., Piper, E.K., Seddon, J.M., Kelly, M.J., 2014. Determining the number of animals required to accurately determine breed composition using genomic data. In: 10th World Congress of Genetics Applied to Livestock Production. Vancouver, BC, Canada.

Core Team, R, 2016. A Language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/.

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.

de Haas, Y., Hoving-Bolink, R., Maurice-Van Eijndhoven, M.H.T., Bohte-Wilhelmus, D., Sulkers, H., Hiemstra, S.J., 2009. Deep Red Cattle.

Ding, L., Wiener, H., Abebe, T., Altaye, M., Go, R.C.P., Kercsmar, C., Grabowski, G., Martin, L.J., Khurana Hershey, G.K, Chakorborty, R., Baye, T.M., 2011. Comparison of measures of marker informativeness for ancestry and admixture mapping. BMC Genomics 12.

EU, 2016. 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. Off. J. Eur. Union 59, 208 L84.

EU, 2016. 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. Off. J. Eur. Union 59, 66–142 Volume 59.

Falush, D., Stephens, M., Pritchard, J.K., 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164, 1567–1587.

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.

Frkonja, A., Gredler, B., Schnyder, U., Curik, I., Sölkner, J., 2012. Prediction of breed composition in an admixed cattle population. Animal Genet. 43, 696–703.

Funkhouser, S.A., Bates, R.O., Ernst, C.W., Newcom, D., Steibel, J.P., 2017. Estimation of genome-wide and locus-specific breed composition in pigs1. Transl. Animal Sci. 1, 36–44.

Hulsegge, B., Calus, M.P.L., Oldenbroek, J.K., Windig, J.J., 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. Animal Breed. Genet. 134, 69–77.

Hulsegge, B., Calus, M.P.L., Windig, J.J., Hoving-Bolink, A.H., Maurice-van Eijndhoven, M.H.T., Hiemstra, S.J., 2013. Selection of SNP from 50 K and 777 K arrays to predict breed of origin in cattle. J. Animal Sci. 91, 5128–5134.

Kassambara, A., 2017. Practical Guide To Principal Component Methods in R: PCA, M (CA), FAMD, MFA, HCPC, Factoextra. CreateSpace Independent Publishing Platform.

Kuehn, L.A., Keele, J.W., Bennett, G.L., McDaneld, T.G., Smith, T.P.L., Snelling, W.M., Sonstegard, T.S., Thallman, R.M., 2011. Predicting breed composition using breed frequencies of 50,000 markers from the US Meat Animal Research Center 2,000 bull project. J. Animal Sci. 89, 1742–1750.

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.

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News 2, 18–22.

Manel, S., Gaggiotti, O.E., Waples, R.S., 2005. Assignment methods: matching biological questions with appropriate techniques. Trend. Ecol. Evolut. 20, 136–142.

Manzanilla-Pech, C.I.V., Veerkamp, R.F., de Haas, Y., Calus, M.P.L., 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.

Matukumalli, L.K., Lawley, C.T., Schnabel, R.D., Taylor, J.F., Allan, M.F., Heaton, M.P., O'Connell, J., Moore, S.S., Smith, T.P.L., Sonstegard, T.S., Van Tassell, C.P., 2009. Development and characterization of a high density SNP genotyping assay for cattle. PLoS ONE 4 (4), e5350.

Maurice-Van Eijndhoven, M.H.T., Bovenhuis, H., Veerkamp, R.F., Calus, M.P.L., 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.

Miciak, J., Fletcher, J.M., Stuebing, K.K., 2015. Accuracy and validity of methods for identifying learning disabilities in a response-to-intervention service delivery framework. Handbook of Response to Intervention: the Science and Practice of Multi-Tiered Systems of Support, Second Edition. Springer, US, pp. 421–440.

Padilla, J.Á., Sansinforiano, E., Parejo, J.C., Rabasco, A., Martínez-Trancón, M., 2009. Inference of admixture in the endangered Blanca Cacereña bovine breed by microsatellite analyses. Livest. Sci. 122, 314–322.

Porras-Hurtado, L., Ruiz, Y., Santos, C., Phillips, C., Carracedo, A., Lareu, M.V., 2013. An overview of structure: Applications, parameter settings, and supporting software. Front. Genet. 4, 98.

Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155, 945–959.

Rogberg-Muñoz, A., Wei, S., Ripoli, M.V., Guo, B.L., Carino, M.H., Castillo, N., Villegas Castagnaso, E.E., Lirón, J.P., Morales Durand, H.F., Melucci, L., Villarreal, E., Peral-García, P., Wei, Y.M., Giovambattista, G., 2014. Foreign meat identification by DNA breed assignment for the Chinese market. Meat Sci. 98, 822–827.

Rosenberg, N.A., Burke, T., Elo, K., Feldman, M.W., Freidlin, P.J., Groenen, M.A.M., Hillel, J., Mäki-Tanila, A., Tixier-Boichard, M., Vignal, A., Wimmers, K., Weigend, S., 2001. Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. Genetics 159, 699–713.

Wilkinson, S., Wiener, P., Archibald, A.L., Law, A., Schnabel, R.D., McKay, S.D., Taylor, J.F., Ogden, R., 2011. Evaluation of approaches for identifying population informative markers from high density SNP Chips. BMC Genet. 12, 45.

Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C., Weir, B.S., 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28, 3326–3328.