Journal of
Animal Breeding and Genetics

Development of a genomic tool for breed assignment by comparison of different classification models - Application to three local cattle breeds

Journal:	*Journal of Animal Breeding and Genetics*
Manuscript ID	JABG-21-0048
Manuscript Type:	Original Article
Date Submitted by the Author:	11-Mar-2021
Complete List of Authors:	Wilmot, Hélène; Université de Liège Gembloux Agro-Bio Tech, Bormann, Jeanne Soyeurt, Hélène; GxABT - ULg, Animal Science Unit Hubin, Xavier Glorieux, Géry Mayeres, Patrick; Walloon Breeding Association, Research and Development; Bertozzi, Carlo; Walloon Breeding Association, Research and Development Gengler, Nicolas; GxABT - ULg, Animal Science Unit
Subject Area:	diversity, animal breeding, SNP

Development of a genomic tool for breed assignment by comparison of different classification models - Application to three local cattle breeds

H. Wilmot[a,b*], J. Bormann[c], H. Soyeurt[b], X. Hubin[d], G. Glorieux[d], P. Mayeres[d], C. Bertozzi[d] and N. Gengler[b]

[a]National Fund for Scientific Research (F.R.S.-FNRS), Rue d'Egmont 5, B-1000 Brussels, Belgium

[b] TERRA Teaching and Research Centre, Gembloux Agro-Bio Tech, University of Liège, B-5030 Gembloux, Belgium

[c]Administration of Technical Agricultural Services (ASTA), L-1470 Luxembourg, Grand Duchy of Luxembourg

[d]Walloon Association of Breeders, B-5590 Ciney, Belgium

[*] Corresponding author, helene.wilmot@uliege.be, + 32 81 62 23 51.

Abstract

Assignments of individual cattle to a specific breed can often not rely on pedigree information. This is especially the case for local breeds for which the development of genomic assignment tools is required to allow more individuals of unknown origin to be included to their herdbooks. A breed assignment model can be based on two specific stages 1) the selection of breed-informative markers and 2) the assignment of individuals to a breed with a classification method. However, the performance of combination of methods used in these two stages have been rarely studied until now. In this study, the combination of 16 different SNP panels with four classification methods was developed on 562 reference genotypes from 12 cattle breeds. Based on their performances, best models were validated on three local breeds of interest. In cross-validation, 14 models had a global cross-validation accuracy higher than 90%, with a maximum of 98.22%. In validation, best models used 7,153 or 2,005 SNPs, based on a partial least squares-discriminant analysis (PLS-DA), and assigned individuals to breeds based on nearest shrunken centroids. The average validation sensitivity of the first two best models for the three local breeds of interest were, respectively, 98.83% and 97.5%. Moreover, results reported in this study suggest that further studies should consider the PLS-DA method when selecting breed-informative SNPs.

KEYWORDS

Breed assignment; classification; informative SNPs; local breeds; partial least squares; SNP panel

1. INTRODUCTION

Interest in local breeds is increasing because they represent a reservoir of unique phenotypes and genetic material, potentially increasing the resilience of animal production systems to economic and ecological challenges. Because of the hyper-specialization of agriculture, local breeds were often left behind for several decades leading to very incomplete (or even completely missing) pedigree information. However, it has been shown that in situ conservation is a powerful tool to keep local breeds in their natural environment by supporting their social setting and their traditional use (Henson, 1992). According to the article 19 of the EU Regulation 2016/1012 on Animal breeding, a special derogation can be allocated to include animals without pedigree to enter the main section of the herdbook of an endangered breed. From this, the question of how locally subsisting populations can be recognized as members of a given endangered breed arises. The development of a tool based on genomic data that is able to correctly assign animals from the endangered breed and to exclude animals from other similar looking breeds can be the solution. Several studies have already focused on this specific topic (e.g. Baumung, Cubric-Curik, Schwend, Achmann, & Sölkner, 2006; Bertolini et al., 2018; I. Hulsegge et al., 2019; Padilla, Sansinforiano, Parejo, Rabasco, & Martínez-Trancón, 2009). Padilla, Sansinforiano, Parejo, Rabasco & Martinez-Trancón (2009) particularly highlighted the need to find a balance between including more individuals with unknown pedigree but appearing to be members of an endangered cattle breed to the herdbook, while excluding animals that could have a similar phenotype.

However, there is no consensus in studies about breed assignment concerning the methodology of selection of SNP markers or the assignment method to use, which are the main stages to follow to develop a model of this kind, stages that are also not always clearly distinguished. Concerning the methods to choose the best markers, the use of F_{ST} (Weir & Cockerham, 1984; Wright, 1951) is particularly common in studies (e.g. Dalvit et al., 2008;

Ding et al., 2011; Frkonja, Gredler, Schnyder, Curik, & Sölkner, 2012; He et al., 2018; B. Hulsegge et al., 2013; Judge, Kelleher, Kearney, Sleator, & Berry, 2017; Wilkinson et al., 2011). Most likely this is due to the fact that this statistic can be easily adapted (e.g. global vs. pairwise) to make them suitable for the selection of markers for breed assignment. Allelic frequencies are another common methodology for selection of markers for breed assignment (e.g. He et al., 2018; Kuehn et al., 2011; Wilkinson et al., 2011). Several studies also used principal component analysis (PCA), based on different types of data input, to select the best SNP markers for discriminating breeds (e.g. Wilkinson et al., 2011). Recently, random forests (RF) were combined with PCA or F_{ST} for this purpose (Bertolini et al., 2015, 2018; I. Hulsegge et al., 2019). However, the use of PCA for selecting breed-informative SNPs could potentially be optimized because even if the PCA allows to reduce the number of dimensions by linear combination of variables in components that are independent to each other, these components do not necessarily explain the answer i.e., the breed (Jolliffe, 2002).

Moreover, breed assignment methods reported in literature are also diverse. They are often used in a second stage, after selecting SNPs, and can be based on the same statistical methods used for selecting SNPs. Besides the use of RF that has already been used for selecting SNPs and as a breed assignment method (Bertolini et al., 2015; I. Hulsegge et al., 2019; Schiavo et al., 2019), other assignment methods were reported in the literature. Among the used methods, one can cite very different approaches as five-nearest-neighbors classification (Lewis et al., 2011), artificial neural network approach (Iquebal et al., 2014), regression including the partial least squares method (PLS) (e.g. Funkhouser, Bates, Ernst, Newcom, & Steibel, 2017) or clustering with Bayesian models (e.g. Frkonja et al., 2012; Gobena, Elzo, & Mateescu, 2018; He et al., 2018; B. Hulsegge et al., 2013; Judge et al., 2017). However, until now, there has been little investigation on the impact of the combination of 1) different selection methods of SNPs, leading to different SNP panels, and 2) different assignment methods.

For all these reasons, there is still a need to organize and compare methods to select different SNP panels interacting with different assignment methods. In this study, five methods for the selection of breed-informative SNPs were tested: pairwise F_{ST} combined with RF; three PCAs, based on different input data, combined with RF (these methods are commonly used as aforementioned); and the partial least squares-discriminant analysis (PLS-DA). The resulting SNP panels were then used as inputs for four classification methods: the PLS-DA, nearest shrunken centroids (NSC), RF and linear support vector machine (SVM). The main objective of this study was therefore to develop a genomic tool for breed assignment by comparison of these different approaches. The specific activities to fulfill this objective were: 1) based on their performances, to compare different methods for selection of breed-informative SNPs and for classification of cattle breeds and the interactions between both; and 2) to validate the best model in the context of three local breeds of interest.

2. MATERIAL AND METHODS

Figure 1 summarizes Material and Methods in a flowchart. The combination of the SNP selection stage (1.1 to 6.0) and assignment stage (A. to D.) are coded to ease the following of the study. Quality control (QC), selection of breed-informative SNPs, classification methods and validation were performed with PLINK v.1.9 (Chang et al., 2015; Purcell & Chang, 2019; Purcell et al., 2007), R v. 3.6.3 (R Core Team, 2013) and visualized through Rstudio (Rstudio Team, 2020). All the methodology developed below was also applied on a dataset with no deviation of Hardy-Weinberg equilibrium per breed (P-value > 10^{-6}).

2.1. Dataset

This study focused on three endangered local breeds of interest: Dual-Purpose Belgian Blue (DPBB), East Belgian Red and White (EBRW) and Red-Pied of Ösling (RPO), this latter being from Grand Duchy of Luxembourg. These three breeds were lacking breeding structures for a few decades even if this is less marked in DPBB that has a relatively complete pedigree following efforts in the last years to stabilize the breed. Following the European Common Agricultural Policy, DPBB, EBRW and RPO can benefit from subsidies through agri-environment measures that provide direct payments to breeders. Currently, entries to the herdbooks of EBRW and RPO are based on phenotypes of all individuals but also, except for EBRW females, on the visual appraisal of the position of individuals' genotypes to seven principal components (PCs). As this visual appraisal is made by a specific person, it can be subjective and induce some bias in the decision to include the animal to the respective herdbook.

Moreover, the EBRW and RPO are geographically (i.e., the regions border each-other) and genomically close as can be seen in Figure 2. These two breeds overlap and are included in the continuum of Red-Pied breeds composed of several breeds as Dutch Improved Red Pied (DIRP), Belgian Campine (CAM), EBRW, RPO, Rotbunte DN (RDN) and Meuse-Rhine-Yssel (MRY). As usual in Red-Pied breeds a continuous, more or less strong gene flow originating from (Red-)Holstein (HOL) is expected. It is also known that Simmental-type cattle were used in mating of EBRW and RPO. Similarly, DIRP bulls were used in the EBRW population.

From the same Figure 2, the overlap between the DPBB and the Beef Belgian Blue (BBB), two breeds that diverged during the seventies in Belgium and that originated from the Shorthorn (SHO) breed can also be seen. Another DPBB-genomically close breed currently potentially used in Belgium is the Rouge des Prés (RDP) breed.

The overlaps (**Fig. 2**), hypotheses of previous and recent use, and the continuum of breeds explain why 9 other breeds were also added to the reference population for the development of the assignment model; it is of main importance to distinguish DPBB, EBRW and RPO individuals from these breeds. Therefore, genotypes of 562 individuals belonging to one of the following 12 breeds: **DPBB, EBRW, RPO,** BBB, CAM, DIRP, HOL, MRY, RDP, RDN, SHO and SIM, were used. Table 1 shows the number of reference individuals used per breed as well as abbreviations used for each breed in this article. All of these 562 individuals were used as reference animals for tests already implemented in Wallonia, which were based on visual appraisal of the genotype of each individual based on different PCs, similarly to those reported in Figure 2. This should allow a certain global continuity of breed assignment in our system.

2.2. Quality control

Genotypes of the reference population were coded 0 for homozygosity of A allele, 1 for heterozygosity and 2 for homozygosity of B allele. Seven different SNP chips were used in this study: BovineSNP50 Beadchip v1 to 3, BovineHD Beadchip v12, EuroG 10k (imputed to BovineSNP50 Beadchip) and EuroG MD (SI) v9. The EuroG MD (SI) v9 chip was not used for genotyping reference individuals but were added when defining the overlap of the different chips because this chip is currently used for genotyping most new DPBB, EBRW and RPO individuals. This strategy was used because it allowed projecting the use of the developed tool into the next years as most likely future chip designs should include a large majority of these common SNPs. The number of genotyped individuals per chip and version of chip can be found in Appendix 1. A total of 17,667 SNPs, common between all the seven chips, passed the following filters on: no non-mapped SNPs, no SNPs located on sexual chromosomes, no triallelic SNPs, minimum GenCall Score of 0.15, minimum GenTrain Score of 0.55, individual Call-Rate higher than 0.98, minimum genotype Call-Rate per chip of 0.95

and minor allele frequency (MAF) higher than 0.01. Minimum genotype Call-Rate per chip was applied to avoid SNPs that were less well genotyped by certain chips and, thus, less accurate for discriminating breeds. An MAF filter of 0.01 was applied on the whole dataset as several studies (Bertolini et al., 2015; I. Hulsegge et al., 2019) suggested that SNPs with high MAF were beneficial for discriminating breeds. However, private alleles can sometimes help differentiating breeds as explained by several authors (Bertolini et al., 2015; Dalvit, De Marchi, et al., 2008; Ding et al., 2011); that is why the MAF filter was not applied per breed.

Moreover, several methods used in this study for selection of SNPs or classification do not allow any missing values. Therefore, we replaced missing values per SNP by their mean and finally, iteratively, we performed a PCA for the available 17,667 SNPs and imputed missing values until convergence, as implemented in the imputePCA function from the missMDA v1.14 R package (Josse & Husson, 2012). Imputed values were kept as real numbers (i.e., not only integers 0, 1 or 2) which provided a continuous estimate of allele (also called gene) content instead of deciding on one genotype. The number of PCs to perform the imputation was evaluated by cross-validation, using the estim_ncpPCA function from the missMDA v1.14 R package (Josse & Husson, 2012). Twenty-two PCs were chosen to parameterize the imputePCA function as it minimizes the mean squared error of prediction (MSEP=0.3482). The objective was not to have a completely accurate value but a more plausible value than the mean. Indeed, it is difficult to have an accurate imputation for limited-sized/local breeds, in part because of their intrinsic feature of being fewer than main breeds. Moreover, local breeds (especially the red-pied breeds like EBRW and RPO) were lacking structured breeding schemes for a few decades and were admixed with similar other local (or even mainstream) breeds which makes them less genetically differentiated from each other. This could also explain the value of MSEP obtained. In addition, it was expected that the selection of breed-informative SNPs would eliminate SNPs that were less accurate as they would not allow to make a clear discrimination of breeds.

2.3. Methods for selection of breed-informative SNPs

Different methods were tested for the selection of best SNPs (called breed-informative SNPs). Following the idea of Bertolini et al. (2015, 2018) and I. Hulsegge et al. (2019), three different PCAs (i.e., PCA performed on genotypes (classical-PCA), on the mean values of genotypes by breed (mean-PCA) and on genotypes of each autosome separately (chrom-PCA)) were combined with a RF for the selection of breed-informative SNPs. The mean-PCA was equivalent to the use of allelic frequencies found in several studies (He et al., 2018; Kuehn et al., 2011; Wilkinson et al., 2011) because means of genotypes by breed were equal to twice the allelic frequencies as genotypes were coded as 0, 1, or 2. As SNPs used in this study were for practical reasons based on the overlap of seven chips, it was expected that, by chance, several SNPs could be in linkage disequilibrium (LD) whereas other regions of the genome would unfortunately not be represented. Therefore, the chrom-PCA would break LD by the selection of best SNPs at each of the autosomes. Similarly as Bertolini et al. (2018), we also combined the selection of breed-informative SNPs by pairwise Weir & Cockerham's F_{ST} values (Weir & Cockerham, 1984) with RF.

A last method of selection was based on PLS-DA, the adapted form of PLS to classification problems. The PLS is based on a PCA while maximizing the covariance with the response (Despagne, Massart, & Chabot, 2000). It thus ensures that components will be correlated with the answer, which is especially desired for selecting breed-informative SNPs. Moreover, the PLS is particularly fitted for situations where the number of variables highly exceeds the number of samples and when there is a high level of collinearity between variables (Kuhn & Johnson, 2013), which could happen because of LD. The PLS-DA-based selection method has the advantage to be performed in one step whereas the other described methods are decomposed into two steps.

Even if several studies have already used PLS-based models for breed assignment/composition (e.g. Frkonja, Gredler, Schnyder, Curik, & Sölkner, 2012), it is however the first time, to our knowledge, that the PLS-DA is used as a tool for selecting breed-informative SNPs in the context of breed assignment. In another context, Soyeurt et al. (2020) used PLS for selection of best wavelengths in mid-infrared spectrum-based prediction of milk lactoferrin content, which inspired the methodology of the current study.

Finally, some classification methods were also tested on the entire SNP panel at the overlap of the seven chips i.e., on 17,667 SNPs. A total of six different SNP methods for selecting breed-informative SNPs were therefore explored in this study. Figure 1 illustrates these six different methods of selection of breed-informative SNPs in a flowchart.

2.4. Computation of thresholds for selection of breed-informative SNPs

The different thresholds, their code and the ranking measure used for each panel of breed-informative SNPs can be found in Table 2.

2.4.1. PCA

The PCA scores were computed as followed: for each SNP, loadings corresponding respectively to the first seven, five and a range of the first three to nine PCs were squared and summed, as proposed initially by Paschou et al. (2007) and used e.g. by Bertolini et al. (2015, 2018) and Wilkinson et al., (2011). The number of PCs to take into account was evaluated considering the PC after which there is a stabilization of eigenvalue and percentage of total variance explained. As highlighted by Bertolini et al. (2015), it is important to recover the variance due to breed differentiation to fulfill our objective and not to know which percentage of the total variance is explained by the PCs considered.

Thresholds were then defined as the mean of these scores plus one, two or three times the standard deviation (SD) of scores. The SNPs corresponding to scores' values that were higher than these thresholds were kept for the second step of selection of breed-informative markers i.e., for RF. The three PCA variants were performed using the FactomineR v.2.3 R package (Lê, Josse, & Husson, 2008) on the matrix of covariances (option "scale.unit=FALSE" of the PCA function) (Bertolini et al., 2015; Paschou et al., 2007; Wilkinson et al., 2011). The use of the matrix of covariances, and therefore the choice to not scale SNP values, would allow to determine directions of maximal variability in the PCA, as explained by Price et al. (2006).

2.4.2. F_{ST}

The F_{ST} values were computed using the formula by Weir & Cockerham (1984) as implemented in Plink v.1.9 (Purcell & Chang, 2019):

$$F_{ST} = \frac{s^2}{\overline{p}(1-\overline{p})}$$

where s^2 and $\overline{p}$ are the variance and the mean of allelic frequencies, respectively.

Thresholds were defined, for each pair of breeds, as the mean of their pairwise F_{ST} values plus one, two or three times their SD. Therefore, for each pair of breeds, SNPs corresponding to F_{ST} values that were higher than these thresholds were kept for the second step of selection of breed-informative markers i.e., for RF.

2.4.3. RF

It should be highlighted that the selection of breed-informative SNPs through RF was combined with each of the aforementioned selection methods i.e., selection of SNPs based on the three PCAs and F_{ST} values.

For the selection of breed-informative SNPs based on RF, values of each SNP were standardized (i.e., each SNP column mean centered and divided by the SD). This standardization was applied to avoid the effect of discriminating SNPs with low MAF to be hidden by the effect of SNPs with higher MAF. The predictive performance of RF is based on the prediction of the out-of-bag (OOB) sample which is not used for the elaboration of the tree (Hastie, Tibshirani, & Friedman, 2009). Internal validation of RF is therefore based on the average error of OOB samples of all trees, called the OOB error rate. For each panel of SNPs (based on one of the three PCAs or on F_{ST} values), RF was optimized for the number of trees (maximum of 5,000 trees tested) and the minimum node size (values of 1 to 50 tested) as implemented by the randomForest function of the randomForest v.4.6-14 R package (Breiman, 2001). The number of tested predictors at each tree node (*mtry*) was optimized by, first using the default value, and then, inflating or deflating this value by steps of one to verify if the OOB error estimate was improved or not. This was implemented by the tuneRF function of the same R package.

Thresholds were defined the same way as for PCA and F_{ST} values but were based on the mean decrease of the Gini Index (MDGI), a measure of the importance of variables (Hastie et al., 2009; Kuhn & Johnson, 2013):

$$Gini\ index = 1 - \sum_{i=1}^{C} (P_i)^2$$

with C is the number of classes and P the observed class probabilities induced by the split. The Gini Index can therefore be seen as an indication of the purity of the nodes. It is minimized when the probability to belong to one class is maximized. If a SNP allows an important decrease of the Gini Index, it means that it increases the purity of each node. Moreover, the MDGI was demonstrated to be efficient for the selection of breed-informative SNPs (Bertolini et al., 2015, 2018; Boulesteix, Bender, Bermejo, & Strobl, 2012; I. Hulsegge et al., 2019).

2.4.4. PLS-DA

The last method of selection of breed-informative SNPs is the PLS-DA, as implemented by the caret v.6.0-85 R package (Kuhn, 2008). As for RF, SNPs were centered and scaled. Again, this standardization was applied to avoid the effect of discriminating SNPs with low MAF to be hidden by the effect of SNPs with higher MAF. A number of 50 components were tested to optimize the accuracy using a 10-folds cross-validation. It means that the sample was divided in 10 parts, nine being used for the elaboration of the classification model and the last part for internal validation, and this is done 10 times, one for each tenth of the sample. In our case, 15 components provided the best accuracy. The maximum number of iterations allowed for convergence of the model was 20,000. Thresholds were then defined as for PCA, F_{ST} and RF, but were based on the importance of each variable for each of the twelve models (one per breed) i.e., based on the absolute value of coefficients computed for each SNP for each model. As selection of SNPs by the twelve models can partially overlap, we also determined the number of SNPs that passed the threshold for one to 12 models.

2.5. Classification methods

Four methods were trained on the standardized genotypes of the reference set (n=562) for assignment models: RF, PLS-DA, NSC and linear SVM. The RF was not tested on the non-selected panel of SNPs. For the other 15 panels of SNPs, the four aforementioned methods were tested. As for selection of breed-informative SNPs, RF was optimized for the number of trees, the *mtry* and the minimum node size as implemented in the randomForest v.4.6-14 R package (Breiman, 2001). The same parameter values as for selection of breed-informative SNPs were tested for their optimization. Other classification methods were implemented by the caret v.6.0-85 R package (Kuhn, 2008). For the PLS-DA, a maximum number of 30 components were tested to optimize the accuracy and the maximum number of iterations allowed for convergence of models was 20,000.

The NSC method is based on distance of the sample to overall and class centroids and is therefore a linear classification method. Some advantages of NSC are its suitability for a large number of variables and low number of samples, which is the case in this study (i.e., 17,667 SNPs and 562 individuals). The particularity of NSC compared to the classical nearest centroid method is to shrink class centroids toward the overall centroid. Before the shrinkage, the within-class SD of each variable is used for standardization as it gives more weight to variables that are stable within class. Therefore, it should give more weight to private alleles if they exist. Class variables that confounded with the overall centroid are not used by the model because they do not allow differentiation. It highlights another benefit from NSC: selection of variables, which are not necessarily the same for each class. Moreover, NSC targets misclassification errors (Kuhn & Johnson, 2013). One parameter has to be optimized for NSC: the level of shrinkage called Δ. The higher it is, the higher the shrinkage to the overall centroid is and so less variables are used by the model. In this study, a maximum 20 different levels of shrinkage were tested by the caret v.6.0-85 R package (Kuhn, 2008). More information about the NSC method and computation can be reached through Tibshirani, Hastie, Narasimhan, & Chu (2002).

The linear SVM is a method that builds linear hyperplanes as boundaries between classes. In this method, boundaries are defined to maximize their margins i.e., their distance with the closest training set points called support vectors. The particularity of the SVM method is that the prediction equation is only a function of these support vectors i.e., on samples that are predicted with the least accuracy and that are the most extreme. For the linear SVM, the cost (*C*) is the only parameter to tune; the higher it is, the more complex is the model and closer to overfitting. In this study, several values of *C* have been tested: 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2 and 5, and implemented by the caret v.6.0-85 R package (Kuhn, 2008).

2.6. Performance of classification models in cross-validation

As explained previously, the internal validation of RF is based on the OOB error rate. For all other models, the 10-folds CV was used. However, Kuhn & Johnson (2013) reported that cross-validation and OOB error rate give a similar insight of the predictive performance. Ranking of the classification models was thus based on 1) the values of global cross-validation accuracy which is the proportion of right assignments or 2) for RF, on 100 minus the OOB error rate. To avoid possible overfitting, the tuning parameters of the simplest model within one standard deviation (SD) of the best model (based on global cross-validation accuracy) were chosen ("oneSE" function of caret v.6.0-85 R package; Kuhn, 2008). This was applied for PLS-DA, NSC and linear SVM as it is accepted that RF avoids overfitting (Kuhn & Johnson, 2013). Models with a cross-validation accuracy higher than 90% were further validated on the validation set.

2.7. Validation set

A balanced independent validation set made of 200 animals of which 40 BBB, 40 DPBB, 40 EBRW, 40 HOL and 40 RPO was used. The validation set comprised animals that were included in the breed herdbook and for which genotypes were available. For EBRW males and RPO animals, this inclusion is based on visual appraisal of phenotypes, and of genotypes on seven PCs. For BBB, DPBB and HOL, the animals corresponding to the breed standards are included in the herdbook based on their pedigree. The concordance of phenotypes with breed standards is checked on farm for BBB, DPBB, EBRW and RPO.

Assigning EBRW, DPBB and RPO individuals to the right breed being the main objective, other breeds in cross-validation were used to control if animals from these breeds could be identified as DPBB, EBRW and RPO. Moreover, HOL and BBB are common breeds in Belgium and it should be expected from the models to correctly assign animals from these breeds. These two breeds were therefore used as a “control” of the validation test. We would also ensure that BBB were not classified as DPBB, as both breeds genetically overlap (**Fig. 2**). Imputation of missing values was performed iteratively on each of the validation animal by adding it to the imputed reference population and following the same algorithm described above. Then, the mean and variance per SNP of the reference population were used to standardize SNP values of validation animals.

To determine the best model in validation, different performance measures were looked at: the global validation accuracy, the average validation sensitivity of DPBB, EBRW and RPO, the average validation specificity of BBB and HOL and probabilities of an animal to belong to its predicted breed. The sensitivity of a model is defined as the proportion of animals of a specific breed correctly assigned to this breed by the assignment model. If the average validation sensitivity of DPBB, EBRW and RPO is high, it means that animals effectively belonging to one of these three breeds are correctly assigned to their breed. On the opposite, the specificity of a model is the proportion of animals not belonging to a specific breed that are not assigned to this specific breed. If the average validation specificity of BBB and HOL is high, it means that DPBB, EBRW and RPO are not assigned to these breeds. This is of interest as DPBB are genetically close to BBB (same history for both breeds until the seventies) and as EBRW and RPO were sometimes crossed with red-pied Holsteins. Therefore, if these breeds are not confused by the model, it will ensure a better definition of the three local breeds of interest.

3. RESULTS

In this study, the importance of using a HW filter was tested. However, the performances only slightly differed between both datasets. Therefore, more information on the results obtained for the dataset with no HW equilibrium deviation can be found in Appendixes 2B and 3B.

3.1. Selection of breed-informative SNPs

Five different methods of selection of breed-informative SNPs, each associated to three thresholds for selecting SNPs, and a panel with all the SNPs that passed QC were used, leading to 16 different SNP panels in total (**Figure 1; Table 2**). In Table 3, the number of SNPs for each panel is shown. When the less stringent threshold is applied (mean + SD), the number of selected SNPs ranged from 205, for classical-PCA combined with RF (3.1.), to 15,102 for PLS-DA (1.1.). When the most stringent threshold is applied (mean + three SD), the number of selected SNPs ranged from three, for classical-PCA combined with RF (3.3.), to 2,005 for PLS-DA (1.3). Three panels included only three, five and six SNPs (3.3./5.3./4.3.). These numbers of SNPs were smaller than the number of breeds to discriminate. Therefore, it was expected that these panels could not be able to perform correctly and they were discarded from further use in this study.

Table 4 shows the total number of SNPs selected by each threshold of the PLS-DA (1.1./1.2./1.3.) and, inside each threshold, by how many of the 12 models these SNPs were selected. It can be observed that the number of SNPs selected by several models decreased with the stringency of the thresholds. Moreover, with the lowest level of stringency of the threshold (1.1.), the maximum number of models that selected the same SNP was nine. With the intermediary and higher levels of stringency (1.2./1.3.), the maximum number of models that selected the same SNPs dropped to five and three models, respectively. It was expected that PLS-DA would give a higher number of selected SNPs than other methods of selection of SNPs because this is a one-step method that selects best SNPs for each of the 12 models (each model predicting one specific breed).

On the opposite, other methods were two-step, which allows to limit the number of selected SNPs a second time. The F_{ST} combined with RF gave higher numbers of selected SNPs, compared to PCA-based methods, as F_{ST} is based on selection of the best SNPs for discriminating each pair of breeds, leading to 66 combinations of breeds. The PCA-based methods used only a small number of PCs (from three to nine PCs out of 17,667 PCs in total) to compute the loadings, which explains the lower number of selected SNPs with PCA-based methods.

3.2. Cross-validation

Each of the 16 different SNP panels developed before was tested on several classification methods which are PLS-DA (A), NSC (B), RF (C)(this latter was not tested on the panel without selection of SNPs) and linear SVM (D). This led to a total of 63 different models. For simplification, only models with a cross-validation accuracy greater than 90% are shown in Table 5. All the 63 different models and their performances can be found in Appendixes 2A and 3A. Among them, results obtained with panels of less than 12 SNPs (i.e., less than the number of breeds to discriminate) are only available for an informative purpose as it was obvious that they could not perform correctly.

From Table 5, it can be observed that 14 models had a cross-validation accuracy greater than 90%. The maximum global cross-validation accuracy of 98.22% was obtained with the panel of 2,005 SNPs and the PLS-DA classification method (1.3.A). The minimum global cross-validation accuracy of 90.39% was obtained with the panel of 228 SNPs and the NSC classification method (5.1.B). Moreover, only the NSC (B) and PLS-DA (A) classification methods allowed global cross-validation accuracy greater than 90%. These classifications methods seemed therefore more appropriate than RF (C) and linear SVM (D) for discriminating the twelve breeds under study.

It should also be highlighted that the most performant methods of selection of breed-informative SNPs seemed to be PLS-DA (1.1./1.2./1.3.) and F_{ST} combined with RF (2.1./2.2.). No selection of SNPs (6.0.) also allowed good assignment even if it is probably related to the high number of breeds to discriminate. If there are more breeds to discriminate, it means that more SNPs would be selected to discriminate each breed from other breeds. Thus, it can be thought that the total panel of 17,667 SNPs is carrier of less noise for discriminating twelve breeds than it would be for discriminating a lower number of breeds.

The PCA-based methods for selecting SNPs seemed less relevant in the context of this study. However, the smallest SNP panel that allowed global cross-validation accuracy greater than 90% was composed of 221 SNPs and was based on mean-PCA combined with RF (4.1.). Therefore, it can be suggested not to use too stringent thresholds when selection of breed-informative SNPs is based on a PCA combined with the RF method.

3.3. Validation tests

Validation tests were performed on a set based on the three breeds of interest (DPBB, EBRW and RPO) as well as on “control breeds” (BBB and HOL). Table 6 shows the results of global validation accuracy, average validation sensitivity of DPBB, EBRW and RPO and average validation specificity of BBB and HOL. It should be noticed that changes in ranking and in global accuracies of models from cross-validation to validation could not be strictly compared because the reference set used for establishing models and the validation set differed. The objective of this study was to assign properly three breeds of interest (DPBB, EBRW and RPO) and to distinguish them from nine other “close” or sister breeds. To fulfill this objective, it was necessary to use the other nine breeds in the reference for developing models. However, it seemed relevant to focus on DPBB, EBRW and RPO in validation.

The 7,153 panel of SNPs followed by the NSC (1.2.B) classification method was the model that provided the best global validation accuracy (99%), average validation sensitivity of DPBB, EBRW and RPO (98.33%) and average validation specificity of BBB and HOL (100%). This model was in fourth position in cross-validation. However, this model is less parsimonious than the best model found in cross-validation (1.3.A.). It should also be noticed that the second best model obtained in validation (1.3.B.) performed very similarly than the best model with much less SNPs i.e., with 2,005 SNPs (98.5% *vs.* 99%). The model based on the less stringent threshold of F_{ST} followed by RF for selection of SNPs and then by NSC for classification (2.1.B.) performed remarkably well (global validation accuracy of 97.5%) with only 1,014 SNPs. The panels with 2,005 (1.3.) and 1,014 SNPs (2.1.) could therefore be a compromise between using less SNPs and having a correct assignment.

A confusion matrix of the best validation model (1.2.B.) is shown in Table 7. In this confusion matrix, it can be seen that the mainstream breeds ("control") as well as DPBB and EBRW seemed to be perfectly predicted. There were two RPO animals that were predicted as EBRW. This was expected since exchanges of animals between both breeds have been existing for many years and as they are geographically close, considered as sister breeds. It is also known from tests already implemented in Wallonia (Southern region of Belgium) that they could genomically overlap (**Fig. 2**). These elements should be considered when interpreting results.

4. DISCUSSION

A lot of studies have already targeted the topic of breed assignment/composition in animal productions, based on SNPs or other markers. Generally, they focused only on comparison of methods of selection of breed-informative markers and then applied the resulting selected markers on one or two assignment methods (among other examples Bertolini et al., 2015, 2018; Dalvit, De Marchi, et al., 2008; Ding et al., 2011; He et al., 2018; B. Hulsegge et al., 2013; I. Hulsegge et al., 2019; Judge et al., 2017; Wilkinson et al., 2011). Some other studies focused more on comparison of classification methods themselves (e.g. Iquebal et al., 2014; Nikolic, Park, Sancristobal, Lek, & Chevalet, 2009). However, studies focusing on comparisons of the combination of different SNP panels and classification methods seemed not common.

Even if their objective was slightly different i.e., determining breed composition, Frkonja et al. (2012) compared 23 panels of SNPs (all one-step, whereas some were two-step in this study) combined with four clustering/regression methods and they computed the correlations between these models and ADMIXTURE (Alexander, Novembre, & Lange, 2009) breed composition. In our study moreover, SNPs were all selected based on a measure of informativeness (**Table 2**) whereas Frkonja et al. (2012), aside the panels based on F_{ST}, selected equally spaced panels, panels of one/a few chromosome(s) or full panels. The models chosen by Frkonja et al. (2012) were all based on clustering or regression methods because their objective was slightly different. In our study, classification methods (RF, linear SVM, NSC and PLS-DA) were then preferred.

Moreover, our study has the specificity to combine data from 1) 12 breeds (of which at least eight are local) with several of them being relatively, or even tightly, historically and genetically connected (“sister breeds”) and 2) seven chips. In most studies (e.g. Bertolini et al., 2015; Dalvit et al., 2008; Funkhouser et al., 2017; I. Hulsegge et al., 2019; Judge et al., 2017; Padilla et al., 2009), the number of breeds to discriminate ranged from two to six and they generally distinctly cluster from one to each other even if they can also be admixed (e.g. Frkonja et al., 2012; Gobena et al., 2018). Similarly to our study, He et al. (2018) discriminated against ten cattle breeds using the overlap of five SNP chips. However, they used completely different methods to assign animals i.e., selection of breed-informative markers was based on mean Euclidean distances of allelic frequencies and assignment of animals on the ADMIXTURE and linear regression models. Results from their study and ours could therefore hardly be compared. Iquebal et al. (2014) used different artificial neural networks to discriminate between 22 goat breeds. However, their study was based on microsatellites. Kuehn et al. (2011) also discriminate against a high number of cattle breeds (16 low related/unrelated breeds) but they only used one chip for this purpose. Wilkinson et al. (2011) also used only one chip when assigning animals to one of the 17 breeds under study. To achieve 98% of correct assignment, they estimated that around 242 SNPs were necessary with a pairwise Weir & Cockerham's F_{ST} based panel and a log-likelihood ratio of three. For a similar level of correct assignment, our study demonstrated that 2,005 SNPs (1.3.B.) were necessary. However, the breeds to discriminate in the study of Wilkinson et al. (2011) were generally weakly genetically related. They also had the choice for selecting SNPs from the entirety of the BovineSNP50 Beadchip while we considered the overlap of 7 chips. The effect of the number of chips, the number of breeds, their level of differentiation, the number of individuals from each breed in the reference population, the genomic representativeness of these reference individuals to their breed and the combination of these elements should therefore be investigated.

Because it is known, among other advantages, to eliminate SNPs that were less well genotyped (Pongpanich, Sullivan, & Tzeng, 2010), a dataset with SNPs that do not deviate from HW equilibrium per breed (P-value > 10^{-6}) was also evaluated in this study. Compared to the dataset with no HW equilibrium filter, this dataset did not demonstrate any major differences in performances. To our knowledge, nobody has studied the impact of HW filter on performances of breed assignment models before. However, there were some studies that highlighted the impact of MAF and LD on the selection of breed-informative markers, mostly in a retrospective manner (e.g. Dalvit et al., 2008; Ding et al., 2011; I. Hulsegge et al., 2019). The influence of QC on the performance of breed assignment models should be targeted by following studies as this could strengthen the power of discrimination of models developed. In this study, it was also chosen to standardize data to be handled by the different models (i.e., mean centering and standard deviation division for each SNP value). It might be interesting to study to which level and how standardization is beneficial to develop models to assign breeds.

Some studies (e.g. Frkonja et al., 2012; Kuehn et al., 2011) computed the correlation between pedigree breed composition and estimated breed composition. As already explained, it could not be excluded that other red-pied breeds have been used to ensure the viability of EBRW and RPO. Moreover, local breeds on which this study focused have been pedigree recorded for only a few years and genotype recorded for even less years. Therefore, it was not possible to precisely correlate our results with the available pedigree breed composition. The fact that these limited-sized breeds were only recently registered also explain why it was not possible to select animals to be in the reference population based on their relationship as other studies did (e.g. Funkhouser et al., 2017; Gobena et al., 2018). For the same reason, it was not relevant to eliminate possible outliers (He et al., 2018) as it was desired to consider the maximum diversity of these limited-sized breeds. When pedigree and genotypes are available for a certain time, it can therefore be advised for following studies to compare different methods for defining reference populations (e.g. methods for eliminating outliers or animals from the same family) and the suitability of this reference population for developing a model that can handle genetically diverse animals from the same breed while excluding animals from other breeds.

Most of the studies used Delta, i.e., absolute allele frequency differences (e.g., Dalvit, De Marchi, et al., 2008; Ding et al., 2011; Frkonja et al., 2012; Gebrehiwot, Strucken, Marshall, Aliloo, & Gibson, 2021; B. Hulsegge et al., 2013; Wilkinson et al., 2011), F_{ST} based methods (e.g., Dalvit et al., 2008; Ding et al., 2011; Frkonja et al., 2012; B. Hulsegge et al., 2013; Wilkinson et al., 2011) or PCA based methods (e.g., Bertolini et al., 2015, 2018; I. Hulsegge et al., 2019; Wilkinson et al., 2011) as a measure of breed informativeness to select markers, even if other methods of selection of markers can be used (e.g., Ding et al., 2011; He et al., 2018).

The power of F_{ST} values for selecting efficient markers for discriminating breeds is obviously related to their intrinsic ability to provide an index of differentiation between breeds. Moreover, the use of F_{ST} can also be declined in several forms: global vs. pairwise, Wright vs. Weir & Cockerham, alone vs. combined with one method, etc. Several studies pointed out that pairwise F_{ST} were more appropriate than global F_{ST} when discriminating more than two breeds (or populations) (e.g., Ding et al., 2011; Kersbergen et al., 2009; Wilkinson et al., 2011) which explains why it was decided to test only pairwise F_{ST} in this study. Bertolini et al. (2018) combined selection of SNPs based on F_{ST} values with RF and compared this method with PCA-based methods combined with RF. As it was already noticed in our study, models based on PCA-based panels (3.1./3.2./3.3./4.1./4.2./4.3./5.1./5.2./5.3.) performed relatively poorly compared to models based on other panels. This was expected since PCA-based methods do not only select SNPs that explain breed differentiation but also other sources of variation observed in the reference set. On the contrary, the PLS-DA method for selection of SNPs considers the correlation of SNPs with the different breeds to discriminate. To our knowledge, this study is the first one to use PLS-DA for selection of SNPs in a breed assignment context. This was inspired by the article of Soyeurt et al. (2020) who used PLS for selecting mid-infrared spectral points of interest for predicting milk lactoferrin content.

The models based on PLS-DA used for selection of SNPs (1.1./1.2./1.3.) performed very well on the reference and validation sets. One of the advantages of using the PLS-DA for selecting breed-informative SNPs is that this method was one-step whereas other methods used in this study were two-step. Consequently, they tended to use more SNPs than F_{ST} based panels (2.1./2.2./2.3.), which performed well with only 1,014 SNPs (2.1.), instead of 7,153 and 2,005 SNPs (1.2./1.3.) for the two best models obtained in validation. Therefore, the power of discrimination of panels based on the PLS-DA method with adapted thresholds should be investigated. It can be asked why it may be necessary to use the PLS-DA to select most informative SNPs and then apply it again on the resulting panel for assignment purposes (1.1./1.2./1.3.A). This preselection stage strengthens the signal of most informative SNPs by removing the "noise" and the collinearity. This explanation is also applicable when RF is used for selection of breed-informative SNPs and then to assign animals (2.1./2.2./2.3./3.1./3.2./3.3./4.1./4.2./4.3./5.1./5.2./5.3.C.). Indeed, when applied on the reduced panel, the classification method estimates new weights (or importance) for each of the SNP. Consequently, the importance of SNPs outside the selected panel is forced to 0. This iterative way of working can be seen as pseudo-bayesian and is similar to the heuristic approaches applied for selection of SNPs and estimation of their weights in genomic selection (e.g., VanRaden, 2008).

Besides their power of discrimination of breeds, another key point about SNPs selected by the best model is that they should be expected to appear in new chips, as highlighted in He et al. (2018). In our study, the overlap of seven chips was used to select the most informative SNPs. It could therefore be expected that new chips will also contain the selected SNPs and there is a higher chance of this if assignment models need less SNPs to perform properly. This fact also explains why the second best model (1.3.B.) could be preferred to the best model (1.2.B.) obtained in validation (2,005 *vs*. 7,153 SNPs). As already mentioned, the use of 2,005 SNPs would be a good compromise between limiting the number of SNPs used and discriminating correctly.

Another issue, that was not targeted in this study, is the correlation between the different SNP panels by determining their overlap (Bertolini et al., 2015; Ding et al., 2011; B. Hulsegge et al., 2013; Paschou et al., 2007). This could also be of interest when determining the best model to use for breed assignment. For example, some panels could lead to similar performances but not using the same SNPs at all to fulfill this objective. There could maybe also have SNP panels with a different number of SNPs, one being almost entirely contained in the other, with similar or highly different global accuracies. Some panels may be also more appropriate to specific methods of assignment, meaning that some methods are more able to handle specific measures of informativeness.

The definition of thresholds for selecting SNPs that should allow the differentiation amongst breeds is also a burning question. However, in several studies, the number of best selected SNPs seemed arbitrary (e.g., Bertolini et al., 2018; He et al., 2018; I. Hulsegge et al., 2019; Judge et al., 2017) while others focused on regularly spaced SNPs or full sets (Frkonja et al., 2012; Funkhouser et al., 2017; Kuehn et al., 2011). A more accurate manner to determine the number of best SNPs to be chosen is to have a look at the log-likelihood ratio (B. Hulsegge et al., 2013) or to define a threshold either of the global accuracy needed (as in Wilkinson et al., 2011) either of the measure of informativeness (as in our study or e.g., in Frkonja et al., 2012). It seems less accurate to compare models that did not contain the same level of informativeness (in case of same number of SNPs for different methods) than different models with different number of SNPs that are supposed to share similar levels of informativeness even if this informativeness could be expressed with different measures.

The fact that some studies used a log-likelihood ratio for determining the number of necessary SNPs highlighted the need for accounting for probabilities when assigning animals to a breed. I. Hulsegge et al. (2019) defined thresholds of probabilities for defining if the animal is pure or crossbred. Even though this topic is also of importance, it was not the objective of this study. For two breeds under study (EBRW and RPO), the herdbook, following the EU Animal Breeding regulation for endangered breeds, only allows animals that fit with the genomic (and phenotypic) breed standards, which means that animals included should be considered as “pure” or excluded.

Moreover, probabilities of animals to belong to a certain breed were higher in NSC (B) than in PLS-DA models (A) (data not shown), which probably ties to the intrinsic way those methods handle data. In general, NSC models (B) performed better than PLS-DA models (A) in validation. It therefore seems that NSC models (B) are less susceptible to overfitting than PLS-DA models (A) in the specific case of this study. The PLS-DA models (A) may be more appropriate to discriminate amongst a smaller number of breeds, that are less genomically related, than the twelve breeds of this study. On the contrary, hyperplanes dedicated to each breed can overlap with the NSC method because it is based on the distance of the animal to assign to each of the class shrunken centroids. Another advantage of the NSC model (B) is that it should easily adapt to dynamic reference populations by modifications of the position of the centroids. A breed should not be considered as static and the reference population should change accordingly.

This study did not demonstrate high cross-validation accuracy of assignment when using the linear SVM method, maybe because it designs margins based on most extreme animals from each breed. Therefore, the lack of performance can be due to the genomic relatedness of the 12 breeds involved i.e., the 12 breeds did not clearly discriminate from each other and formed a continuum as shown in Figure 2. With a model based on a SVM, Pasupa, Rathasamuth, & Tongsima (2020) obtained an accuracy of 95.12% with only 164 SNPs to discriminate 21 pig breeds that seemed less genomically related than the 12 cattle breeds under study. Moreover, they used an iterative combination of algorithms to select breed-informative SNPs and they tuned the SVM to be radial, which may take time to parameterize as they highlighted. This can explain the differences of performances between their study and ours. Therefore, it can be suggested to further studies to use the radial SVM method instead of the linear SVM.

The RF method gave intermediary cross-validation accuracies even if it was previously shown to be efficient. For example, I. Hulsegge et al. (2019) obtained a global accuracy of 86.13% with 976 SNPs to discriminate seven breeds. This result is similar to the best cross-validation accuracy found with the RF classification method (1.3.C.) i.e., to 88.79%. Another example is provided by Bertolini et al. (2018) that obtained high global accuracies of minimum 98.62% with only 48 SNPs when comparing different methods of selection of breed-informative SNP,. However, again, the number of breeds they studied was lower (n=5) and they were more genomically differentiated than the 12 breeds under study.

It should be highlighted that the SNPs selected, their number, the best models and their parameters were specific to the number of breeds under study, their genomic diversity and relatedness and the SNPs available. The best model (1.2.B.) found in this study may not be the best in other cases. However, methods for selection of SNPs and/or assignment of breeds that were proven to perform well should be expected to perform well also in other cases (e.g., selection of SNPs based on F_{ST} combined with RF (2) /PLS-DA (1) and classification methods based on NSC (B) /PLS-DA (A)). The results obtained in this study (99% and 98.5% of global accuracy in validation, with 7,153 (1.2.B.) and 2,005 SNPs (1.3.B.) respectively) are encouraging and should pave the way for other studies concerned about this topic.

The results obtained can also potentially be used in the process of certification of labelled breed-products. One usual strategy that can be relatively easily implemented for preserving endangered breeds is the development of labelled products, for example based on meat or cheese. In Belgium and France, for example, the transboundary BlueSter (2020) project was launched to preserve and enhance the DPBB breed. Among other outcomes, a cheese of DPBB was created and will be followed by other certified products like meat. One of the objectives of the BlueSter (2020) project is to develop a certification tool to ensure to consumers that these local products are actually derived from the DPBB breed and not, for example, from the BBB breed, which only diverged from DPBB a few decades ago (BlueSter, 2020). This situation highlights the need for a breed assignment model that can be applied on breed-derived products themselves. Several studies have already developed models for meat breed certification purposes (e.g., Dimauro et al., 2013; Judge et al., 2017). Moreover, genomic DNA can now be extracted from small samples of milk (Pokorska, Kułaj, Dusza, Żychlińska-Buczek, & Makulska, 2016). This technological advance will possibly ensure a non-invasive manner of determining the breed of origin of labelled dairy products.

5. CONCLUSIONS

This study demonstrated that PLS-DA and NSC are effective for selection of breed-informative SNPs and breed assignment, respectively. The results obtained are promising, especially as models were 1) developed on a high number of breeds (n=12); 2) based on the overlap of seven chips; and 3) validated on three local endangered cattle breeds of interest. The best model (1.2.B.) found in this study used 7,153 SNPs, had a global validation accuracy of 99% and an average validation sensitivity for the three local breeds (DPBB, EBRW and RPO) of interest of 98.83%. However, the second best model (1.3.B.) performed almost equally with only 2,005 SNPs, a global validation accuracy of 98.5% and an average validation sensitivity for DPBB, EBRW and RPO of 97.5%. This second model (1.3.B.) may be preferred in application to limit the number of SNPs to be used and then ensure the continued use of the model for next years. Future breed assignments for EBRW and RPO will be based on these results. Also, these results indicate, that at least for the studied breeds (e.g., DPBB), certification of breed-derived products can be considered a feasible option. Finally, to our knowledge, this is the first time that the PLS-DA is used in the context of breed assignment for selection of breed-informative markers. This method of selection of SNPs should further be investigated and potentially be compared with other strategies, especially those not tested in this study.

ACKNOWLEDGEMENTS

H. Wilmot, as a current Research Fellow, and N. Gengler, as a former senior research associate, acknowledge the support of the Fonds de la Recherche Scientifique – FNRS (Brussels, Belgium). The authors thank the INTERREG V France-Wallonie-Vlaanderen program and the Walloon Government (Service Public de Wallonie – Direction Générale Opérationnelle Agriculture, Ressources Naturelles et Environnement, SPW-DGARNE; Namur, Belgium) for their financial support through the BlueSter project and previous projects. The authors also acknowledge the financial support of the Administration of Technical Agricultural Services (ASTA) of the Ministry of Agriculture, Viticulture and Rural Development of the Grand Duchy of Luxembourg, in particular through the CaLiCCo Project, and the technical support by the Walloon Breeders Association (elevéo) and CONVIS, a breeders' cooperative from the Grand Duchy of Luxembourg. The University of Liège – Gembloux Agro-Bio Tech (Liège, Belgium) supported computations through the technical platform Calcul et Modélisation Informatique (CAMI) of the TERRA Teaching and Research Centre, partly supported by the Fonds de la Recherche Scientifique – FNRS under grant no. T.0095.19 (PDR "DEEPSELECT"). Genotyping was facilitated through the support of the Fonds de la Recherche Scientifique – FNRS under grant no. J.0174.18 (CDR "PREDICT-2").

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

AUTHOR CONTRIBUTIONS

H. Wilmot, J. Bormann, X. Hubin, H. Soyeurt and N. Gengler conceived and designed the study. G. Glorieux and X. Hubin provided data from the Walloon part of Belgium and J.

Bormann provided data from Luxembourg. G. Glorieux and J. Bormann contributed to the understanding of current breeding situations in the studied breeds and X. Hubin contributed to information on the used chipsets. H. Wilmot performed the experiment. H. Wilmot and N. Gengler interpreted the results. H. Wilmot wrote the paper. N. Gengler edited and reviewed the manuscript. All authors read and approved the final manuscript.

ETHICAL APPROVAL

The SNP data for the animals included in this study were previously obtained on samples collected by concerned breeder associations based on relevant authorization by the different local authorities.

CONSENT FOR PUBLICATION

Not applicable.

DATA AVAILABILITY STATEMENT

The data supporting the findings of this study cannot be made available as a whole. The corresponding author, upon reasonable request, will forward request to relevant data owners.

ORCID

Hélène Wilmot https://orcid.org/0000-0002-1075-088X

Hélène Soyeurt https://orcid.org/0000-0001-9883-9047

Géry Glorieux https://orcid.org/0000-0003-0380-2837

Carlo Bertozzi https://orcid.org/0000-0003-4602-3868

Nicolas Gengler https://orcid.org/0000-0002-5981-5509

REFERENCES

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Baumung, R., Cubric-Curik, V., Schwend, K., Achmann, R., & Sölkner, J. (2006). Genetic characterisation and breed assignment in Austrian sheep breeds using microsatellite marker information. *Journal of Animal Breeding and Genetics*, *123*(4), 265–271. https://doi.org/10.1111/j.1439-0388.2006.00583.x

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. *Journal of Animal Breeding and Genetics*, *132*(5), 346–356. https://doi.org/10.1111/jbg.12155

Bertolini, F., Galimberti, G., Schiavo, G., Mastrangelo, S., Di Gerlando, R., Strillacci, M. G., … Fontanesi, L. (2018). Preselection statistics and Random Forest classification identify population informative single nucleotide polymorphisms in cosmopolitan and autochthonous cattle breeds. *Animal*, *12*(1), 12–19. https://doi.org/10.1017/S1751731117001355

BlueSter. (2020). BlueSter. Retrieved April 27, 2020, from https://www.projet-bluester.eu/

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. *Briefings in Bioinformatics*, *13*(3), 292–304. https://doi.org/10.1093/bib/bbr053

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1017/CBO9781107415324.004

Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., & Lee, J. J.

(2015). Second-generation PLINK : rising to the challenge of larger and richer datasets. *GigaScience*, *4*(7), 1–16. https://doi.org/10.1186/s13742-015-0047-8

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 Science*, *80*(2), 389–395. https://doi.org/10.1016/j.meatsci.2008.01.001

Dalvit, C., Marchi, M. De, Targhetta, C., Gervaso, M., & Cassandro, M. (2008). Genetic traceability of meat using microsatellite markers. *Food Research International*, *41*, 301–307. https://doi.org/10.1016/j.foodres.2007.12.010

Despagne, F., Massart, L. D., & Chabot, P. (2000). Development of a robust calibration model for nonlinear in-line process data. *Analytical Chemistry*, *72*(7), 1657–1665. https://doi.org/10.1021/ac991076k

Dimauro, C., Cellesi, M., Steri, R., Gaspa, G., Sorbolini, S., Stella, A., & Macciotta, N. P. P. (2013). Use of the canonical discriminant analysis to select SNP markers for bovine breed assignment and traceability purposes. *Animal Genetics*, *44*, 377–382. https://doi.org/10.1111/age.12021

Ding, L., Wiener, H., Abebe, T., Altaye, M., Go, R. C. P., Kercsmar, C., … Baye, T. M. (2011). Comparison of measures of marker informativeness for ancestry and admixture mapping. *BMC Genomics*, *12*. https://doi.org/10.1186/1471-2164-12-622

Frkonja, A., Gredler, B., Schnyder, U., Curik, I., & Sölkner, J. (2012). Prediction of breed composition in an admixed cattle population. *Animal Genetics*, *43*(6), 696–703. https://doi.org/10.1111/j.1365-2052.2012.02345.x

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 pigs. *Translational Animal Science*, *1*(1), 36–44. https://doi.org/10.2527/tas2016.0003

Gebrehiwot, N. Z., Strucken, E. M., Marshall, K., Aliloo, H., & Gibson, J. P. (2021). SNP panels for the estimation of dairy breed proportion and parentage assignment in African crossbred dairy cattle. *Genetics Selection Evolution*, *53*(21), 1–18. https://doi.org/10.1186/s12711-021-00615-4

Gobena, M., Elzo, M. A., & Mateescu, R. G. (2018). Population structure and genomic breed composition in an Angus-Brahman crossbred cattle population. *Frontiers in Genetics*, *9*. https://doi.org/10.3389/fgene.2018.00090

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (Second Edi; Springer, Ed.). New York: Springer. https://doi.org/10.1007/978-1-4419-9863-7_941

He, J., Guo, Y., Xu, J., Li, H., Fuller, A., Tait, R. G., … Bauck, S. (2018). Comparing SNP panels and statistical methods for estimating genomic breed composition of individual animals in ten cattle breeds. *BMC Genetics*, *19*. https://doi.org/10.1186/s12863-018-0654-3

Henson, E. L. (FAO). (1992). The Need For Conservation. In *In situ conservation of livestock and poultry*. Rome, Italy: FAO and UNEP.

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 50K and 777K arrays to predict breed of origin in cattle. *Journal of Animal Science*, *91*(11), 5128–5134. https://doi.org/10.2527/jas.2013-6678

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

Iquebal, M. A., Ansari, M. S., Sarika, S., Dixit, S. P., Verma, N. K., Aggarwal, R. A. K., … Kumar, D. (2014). Locus minimization in breed prediction using artificial neural network approach. *Animal Genetics*, *45*(6), 898–902. https://doi.org/10.1111/age.12208

Jolliffe, I. T. (2002). Principal components analysis. In *Springer Series in Statistics* (2nd Editio). New York (USA): Springer-Verlag. https://doi.org/10.1016/B978-0-08-044894-7.01358-0

Josse, J., & Husson, F. (2012). Handling missing values in exploratory multivariate data analysis methods. *Journal de La Société Française de Statistique*, *153*(2), 77–99.

Judge, M. M., Kelleher, M. M., Kearney, J. F., Sleator, R. D., & Berry, D. P. (2017). Ultra-low-density genotype panels for breed assignment of Angus and Hereford cattle. *Animal*, *11*(6), 938–947. https://doi.org/10.1017/S1751731116002457

Kersbergen, P., van Duijn, K., Kloosterman, A. D., den Dunnen, J. T., Kayser, M., & de Knijff, P. (2009). Developing a set of ancestry-sensitive DNA markers reflecting continental origins of humans. *BMC Genetics*, *10*, 69. https://doi.org/10.1186/1471-2156-10-69

Kuehn, L. A., Keele, J. W., Bennett, G. L., McDaneld, T. G., Smith, T. P. L., Snelling, W. M., … 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. *Journal of Animal Science*, *89*(6), 1742–1750. https://doi.org/10.2527/jas.2010-3530

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. In S. Imprint (Ed.), *Applied Predictive Modeling*. New York (USA): Spinger Nature. https://doi.org/10.1007/978-1-4614-6849-3

Lê, S., Josse, F., & Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software*, *25*(1), 1–18. https://doi.org/10.18637/jss.v025.i01

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*(4), e18007. https://doi.org/10.1371/journal.pone.0018007

Nikolic, N., Park, Y.-S., Sancristobal, M., Lek, S., & Chevalet, C. (2009). What do artificial neural networks tell us about the genetic structure of populations? The example of European pig populations. *Genetics Research*, *91*(2), 121–132. https://doi.org/10.1017/S0016672309000093

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. *Livestock Science*, *122*(2–3), 314–322. https://doi.org/10.1016/j.livsci.2008.09.016

Paschou, P., Ziv, E., Burchard, E. G., Choudhry, S., Rodriguez-Cintron, W., Mahoney, M. W., & Drineas, P. (2007). PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genetics*, *3*(9), 1672–1686. https://doi.org/10.1371/journal.pgen.0030160

Pasupa, K., Rathasamuth, W., & Tongsima, S. (2020). Discovery of significant porcine SNPs for swine breed identification by a hybrid of information gain, genetic algorithm, and frequency feature selection technique. *BMC Bioinformatics*, *21*(216). https://doi.org/10.1186/s12859-020-3471-4

Pokorska, J., Kułaj, D., Dusza, M., Żychlińska-Buczek, J., & Makulska, J. (2016). New Rapid Method of DNA Isolation from Milk Somatic Cells. *Animal Biotechnology*, *27*(2), 113–117. https://doi.org/10.1080/10495398.2015.1116446

Pongpanich, M., Sullivan, P. F., & Tzeng, J.-Y. (2010). A quality control algorithm for filtering SNPs in genome-wide association studies. *Bioinformatics*, *26*(14), 1731–1737. https://doi.org/10.1093/bioinformatics/btq272

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*, 904–909. https://doi.org/https://doi.org/10.1038/ng1847

Purcell, S., & Chang, C. (2019). *PLINK v1.9*. Retrieved from www.cog-genomics.org/plink/1.9/

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK : A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

R Core Team. (2013). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/

Schiavo, G., Bertolini, F., Galimberti, G., Bovo, S., Dall'olio, S., Nanni Costa, L., … Fontanesi, L. (2019). A machine learning approach for the identification of population-informative markers from high-throughput genotyping data: Application to several pig breeds. *Animal*, 223–232. https://doi.org/10.1017/S1751731119002167

Soyeurt, H., Grelet, C., McParland, S., Calmels, M., Coffey, M., Tedde, A., … Gengler, N. (2020). A comparison of 4 different machine learning algorithms to predict lactoferrin content in bovine milk from mid-infrared spectra. *Journal of Dairy Science*, *103*(12), 11585–11596. https://doi.org/10.3168/jds.2020-18870

Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(10), 6567–6572. https://doi.org/10.1073/pnas.082099299

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*(11), 4414–4423. https://doi.org/10.3168/jds.2007-0980

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, *38*(6), 1358–1370. https://doi.org/10.2307/2408641

Wilkinson, S., Wiener, P., Archibald, A. L., Law, A., Schnabel, R. D., McKay, S. D., … Ogden, R. (2011). Evaluation of approaches for identifying population informative markers from high density SNP Chips. *BMC Genetics*, *12*. https://doi.org/10.1186/1471-2156-12-45

Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics*, *15*, 323–354.

Table 1. Breeds used for assignment models and the number (n) of reference individuals used per breed. **In bold**, breeds on which this study specifically focuses.

Breed	**Abbreviation**	**n**
Beef Belgian Blue	BBB	60
Belgian Campine	CAM	33
Dutch Improved Red Pied	DIRP	25
Dual-Purpose Belgian Blue	**DPBB**	**60**
East Belgian Red and White	**EBRW**	**50**
(Red-)Holstein	HOL	120
Meuse-Rhine-Yssel	MRY	63
Rotbunte DN	RDN	17
Rouge des Prés	RDP	20
Red-Pied of Ösling	**RPO**	**51**
Shorthorn	SHO	30
Simmental	SIM	33

Table 2. Methods used for stage 1: selection of breed-informative informative SNPs, ranking measures and definition of thresholds for each ranking measure. Cases relative to stage 1 are coded following Figure 1.

Method of selection of breed-informative markers	**Ranking measure**	**Thresholds**	**Used in case**
Classical-PCA/Mean-PCA/Chrom-PCA	Scores defined as: $\Sigma_{i=1}^{k}(\textbf{SNP loading to the i}^{\textbf{th}}\textbf{PC})^2$ k = number of PCs considered	$\mu_{scores} + \sigma_{scores}$ $\mu_{scores} + 2*\sigma_{scores}$ $\mu_{scores} + 3*\sigma_{scores}$	First step of 3.1./4.1./5.1. First step of 3.2./4.2./5.2. First step of 3.3./4.3./5.3.
W&C F_{ST}	NA	$\mu_{W\&C\ Fst} + \sigma_{W\&C\ Fst}$ $\mu_{W\&C\ Fst} + 2*\sigma_{W\&C\ Fst}$ $\mu_{W\&C\ Fst} + 3*\sigma_{W\&C\ Fst}$	First step of 2.1. First step of 2.2. First step of 2.3.
Random forest	MDGI	$\mu_{MDGI} + \sigma_{MDGI}$ $\mu_{MDGI} + 2*\sigma_{MDGI}$ $\mu_{MDGI} + 3*\sigma_{MDGI}$	Second step of 2.1./3.1./4.1./5.1 Second step of 2.2./3.2./4.2./5.2. Second step of 2.2./3.3./4.3./5.3.
Partial least square-discriminant analysis	Absolute value of coefficients	$\mu_{coefficient} + \sigma_{coefficient}$[a] $\mu_{coefficient} + 2*\sigma_{coefficient}$[a] $\mu_{coefficient} + 3*\sigma_{coefficient}$[a]	1.1. 1.2. 1.3.
No selection	NA	NA[e]	6.0.

Abbreviations: classical-PCA: principal component analysis on genotypes; mean-PCA: principal component analysis on mean of genotypes by breed; chrom-PCA: principal component analysis per autosome; W&C F_{ST}: pairwise Weir & Cockerham's F_{ST}; PC: principal component; NA: not applicable; MDGI: mean decrease in the Gini Index.

[a]: thresholds were applied to the twelve models (one per breed) to determine breed-informative markers specific to each of these twelve breeds.

Table 3. Number of SNPs selected by each method of selection of breed-informative SNPs. Cases are coded following Figure 1. *In italic,* panels with less SNPs than the number of breeds to discriminate.

Case	**Number of SNPs**
1.1.	15,102
1.2.	7,153
1.3.	2,005
2.1.	1,014
2.2.	396
2.3.	154
3.1.	205
3.2.	30
3.3.	*3*
4.1.	221
4.2.	35
4.3.	*6*
5.1.	228
5.2.	33
5.3.	*5*
6.0.	17,667

Table 4. Total number of SNPs selected by each threshold of the partial least squares-discriminant analysis (PLS-DA) and, within each threshold, number of SNPs detected as informative by one to nine models of the PLS-DA. Cases are coded following figure 1.

Case	Number of models of the PLS-DA[a]									Total number of SNPs
	1	2	3	4	5	6	7	8	9	
1.1.	5,074	4,887	3,035	1,429	520	134	22	0	1	15,102
1.2.	5,387	1,456	277	29	4	0	0	0	0	7,153
1.3.	1,857	138	10	0	0	0	0	0	0	2,005

[a]: the PLS-DA creates a model for each breed. There are therefore 12 models inside each threshold.

Table 5. Number of selected SNPs, classification methods and ranked global accuracy obtained in 10-folds cross-validation. Cases are coded following Figure 1.

Case	**Number of selected SNPs**	**Classification method**	**Global accuracy (%)**
1.3.A.	2,005	PLS-DA	98.22
1.3.B.	2,005	NSC	97.33
1.2.A.	7,153	PLS-DA	96.62
1.2.B.	7,153	NSC	96.26
1.1.B.	15,102	NSC	94.66
2.1.A.	1,014	PLS-DA	95.54
1.1.A.	15,102	PLS-DA	95.19
6.0.A.	17,667	PLS-DA	94.48
6.0.B.	17,667	NSC	93.77
2.1.A.	1,014	NSC	93.41
2.2.A.	396	PLS-DA	92.88
2.2.B.	396	NSC	92.7
4.1.B.	221	NSC	91.46
5.1.B.	228	NSC	90.39

Abbreviations PLS-DA: partial least squares-discriminant analysis; NSC: nearest shrunken centroids.

Table 6. Number of selected SNPs, classification methods, ranked global accuracy, ranked average sensitivity of Dual-Purpose Belgian Blue (DPBB), East Belgian Red and White (EBRW) and Red Pied of Ösling (RPO), average specificity of Beef Belgian Blue (BBB) and Holstein (HOL), obtained in validation. Cases are coded following Figure 1.

Case	Number of selected SNPs	Classification method	Global accuracy (%)	Average sensitivity of BBM, EBRW and RPO (%)	Average specificity of BBB and HOL (%)
1.2.B.	7,153	NSC	99	98.83	100
1.3.B.	2,005	NSC	98.5	97.5	100
1.1.B.	15,102	NSC	98	96.67	100
6.0.B.	17,667	NSC	98	96.67	100
1.1.A.	15,102	PLS-DA	97.5	95.83	100
6.0.A.	17,667	PLS-DA	97.5	95.83	100
2.1.B.	1,014	NSC	97.5	95.83	100
1.2.A.	7,153	PLS-DA	97	95	99.69
2.1.A.	1,014	PLS-DA	97	95	99.69
1.3.A.	2,005	PLS-DA	96	93.33	99.38
2.2.B.	396	NSC	93	88.33	100
4.1.B.	221	NSC	90.5	84.17	99.69
2.2.A.	396	PLS-DA	89.5	82.5	99.38
5.1.B.	228	NSC	51	49.17	100

Abbreviations NSC: nearest shrunken centroids; PLS-DA: partial least squares-discriminant analysis.

Table 7. Confusion matrix of the best model found in validation (7,153 SNPs selected and the nearest shrunken centroids classification method; case 1.2.B.).

Predicted breed	**Breed of origin**				
	BBB	DPBB	EBRW	HOL	RPO
BBB	40	0	0	0	0
DPBB	0	40	0	0	0
EBRW	0	0	40	0	2
HOL	0	0	0	40	0
RPO	0	0	0	0	38

Abbreviations BBB: Beef Belgian Blue; DPBB: Dual Purpose Belgian Blue; EBRW: East Belgian Red and White; HOL: Holstein; RPO: Red Pied of Ösling.

Figure 1. Schematic representation of the Material & Methods followed in this study. Abbreviations MAF: minor allele frequency; PLS-DA: partial least square-discriminant analysis; RF: random forest; Classical-PCA: principal component analysis on genotypes; Mean-PCA: principal component analysis on mean of genotypes by breed; Chrom-PC: principal component analysis per autosome; NSC: nearest shrunken centroids; SVM: support vector machine. ^{a}: for each method of selection of breed-informative SNPs, case "1" represents the less stringent threshold, case "2" represents the intermediary threshold and case "3" represents the most stringent threshold. Case "0" is used when there is no threshold because of no selection.

Figure 2. Distribution of animals from the reference population on the first two components of a principal component analysis. Animals from each breed are represented with a different colour. **In bold**, breeds on which this study specifically focuses. Abbreviations BBB: Beef Belgian Blue; CAM: Belgian Campine; DIRP: Dutch Improved Red Pied; **DPBB: Dual-Purpose Belgian Blue**; **EBRW: East Belgian Red and White**; HOL: Holstein; MRY: Meuse-Rhine-Yssel; RDN: Rotbunte DN; RDP: Rouge des Prés; **RPO: Red Pied of Ösling**; SHO: Shorthorn; SIM: Simmental.

Appendix 1.

Chips used and number (n) of genotyped reference individuals per chip.

Chip	n
BovineSNP50 Beadchip v1	97
BovineSNP50 Beadchip v2	274
BovineSNP50 Beadchip v3	76
BovineHD Beachip v12	110
EuroG 10k	5
EuroG MD v9[a]	0
EuroG MD SI v9[a]	0

[a]: this chip was not used for genotyping reference individuals but for genotyping new individuals.

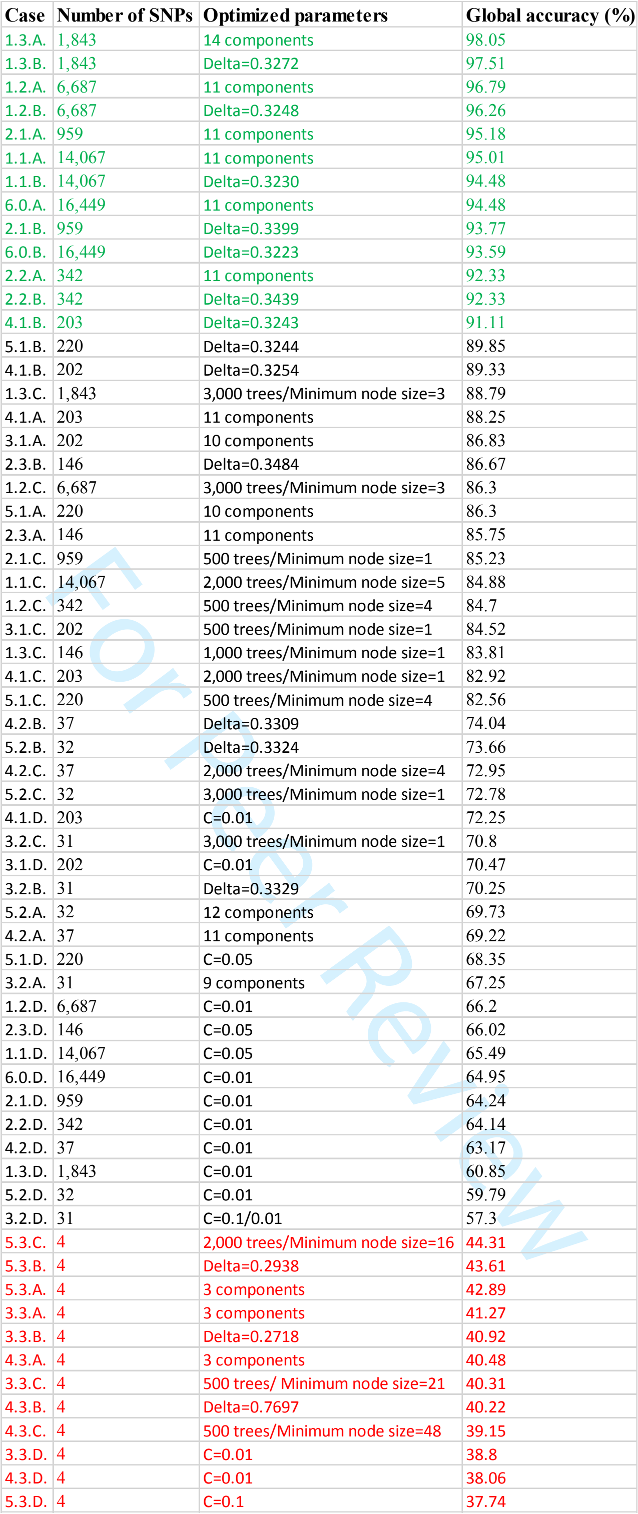
Appendix 2.

Ranked results of cross-validation obtained on a) the dataset without any Hardy-Weinberg filter and b) the dataset filtered out for Hardy-Weinberg equilibrium P-value smaller than 10^{-6}. In green, results with more than 90% of global accuracy. In red, results from panels with less than 12 SNPs. Cases are coded following Figure 1.

a)

Case	Number of SNPs	Optimized parameters	Global accuracy (%)
1.3.A.	2,005	17 components	98.22
1.3.B.	2,005	Delta=0.3273	97.33
1.2.A.	7,153	11 components	96.62
1.2.B.	7,153	Delta=0.3248	96.26
2.1.A.	1,014	11 components	95.54
1.1.A.	15,102	11 components	95.19
1.1.B.	15,102	Delta=0.3230	94.66
6.0.A.	17,667	11 components	94.48
6.0.B.	17,667	Delta=0.3224	93.77
2.1.B.	1,014	Delta=0.3403	93.41
2.2.A.	396	11 components	92.88
2.2.B.	396	Delta=0.3445	92.7
4.1.B.	221	Delta=0.3251	91.46
5.1.B.	228	Delta=0.3242	90.39
3.1.B.	205	Delta=0.3259	89.15
4.1.A.	221	11 components	88.79
1.3.C.	2,005	2,000 trees/Minimum node size=1	88.79
2.3.B.	154	Delta=0.3475	88.24
3.1.A.	205	10 components	87.36
5.1.A.	228	10 components	86.84
1.2.C.	7,153	2,000 trees/Minimum node size=6	86.3
1.1.C.	15,102	2,000 trees/Minimum node size=3	85.77
2.1.C.	1,014	500 trees/Minimum node size=4	85.59
2.3.A.	154	11 components	85.4
2.2.C.	396	500 trees/Minimum node size=1	84.88
3.1.C	205	500 trees/ Minimum node size=3	83.63
2.3.C.	154	3,000 trees/Minimum node size=1	83.27
4.1.C.	221	500 trees/Minimum node size=1	83.27
5.1.C.	228	500 trees/Minimum node size=4	82.56
4.1.D.	221	C=0.01	75.08
4.2.B.	35	Delta=0.3309	74.78
5.2.C.	33	2,000 trees/ Minimim node size=1	72.95
5.2.B.	33	Delta=0.3323	71.89
4.2.C.	35	500 trees/Minimum node size=2	71.35
3.1.D.	205	C=0.01	70.8
2.3.D.	154	C=0.01	70.65
3.2.B.	30	Delta=0.3318	70.62
5.1.D.	228	C=0.01	70.49
4.2.A.	35	8 components	70.29
5.2.A.	33	15 components	70.12
3.2.C.	30	500 trees/Minimum node size=1	68.86
2.2.A.	30	7 components	66.72
2.1.C.	1,014	C=0.01	66.02
1.2.D.	7,153	C=0.01	66.02
1.1.D.	15,102	C=0.05	65.13
6.0.D.	17,667	C=0.01	64.78
4.2.D.	35	C=0.01	64.6
2.2.D.	396	C=0.1	64.23
1.3.D.	2,005	C=0.01	61.03
3.2.D.	30	C=0.01	59.45
5.2.D.	33	C=0.01	58.71
5.3.B.	5	Delta=0.3085	46.28
5.3.C.	5	2,000 trees/ Minimum node size=26	45.73
4.3.C.	6	500 trees/ Minimum node size=18	43.5
5.3.A.	5	4 components	43.25
3.3.C.	3	2000 trees/ Minimum node size=3	42.7
4.3.A.	6	4 components	42.51
4.3.B.	6	Delta=0.2805	42.35
5.3.D.	5	C=0.05	38.44
3.3.B.	3	Delta=0.3106	38.43
4.3.D.	6	C=0.05	38.42
3.3.A.	3	2 components	37.72
3.3.D.	3	C=0.25	33.99

b)



Case	Number of SNPs	Optimized parameters	Global accuracy (%)
1.3.A.	1,843	14 components	98.05
1.3.B.	1,843	Delta=0.3272	97.51
1.2.A.	6,687	11 components	96.79
1.2.B.	6,687	Delta=0.3248	96.26
2.1.A.	959	11 components	95.18
1.1.A.	14,067	11 components	95.01
1.1.B.	14,067	Delta=0.3230	94.48
6.0.A.	16,449	11 components	94.48
2.1.B.	959	Delta=0.3399	93.77
6.0.B.	16,449	Delta=0.3223	93.59
2.2.A.	342	11 components	92.33
2.2.B.	342	Delta=0.3439	92.33
4.1.B.	203	Delta=0.3243	91.11
5.1.B.	220	Delta=0.3244	89.85
4.1.B.	202	Delta=0.3254	89.33
1.3.C.	1,843	3,000 trees/Minimum node size=3	88.79
4.1.A.	203	11 components	88.25
3.1.A.	202	10 components	86.83
2.3.B.	146	Delta=0.3484	86.67
1.2.C.	6,687	3,000 trees/Minimum node size=3	86.3
5.1.A.	220	10 components	86.3
2.3.A.	146	11 components	85.75
2.1.C.	959	500 trees/Minimum node size=1	85.23
1.1.C.	14,067	2,000 trees/Minimum node size=5	84.88
1.2.C.	342	500 trees/Minimum node size=4	84.7
3.1.C.	202	500 trees/Minimum node size=1	84.52
1.3.C.	146	1,000 trees/Minimum node size=1	83.81
4.1.C.	203	2,000 trees/Minimum node size=1	82.92
5.1.C.	220	500 trees/Minimum node size=4	82.56
4.2.B.	37	Delta=0.3309	74.04
5.2.B.	32	Delta=0.3324	73.66
4.2.C.	37	2,000 trees/Minimum node size=4	72.95
5.2.C.	32	3,000 trees/Minimum node size=1	72.78
4.1.D.	203	C=0.01	72.25
3.2.C.	31	3,000 trees/Minimum node size=1	70.8
3.1.D.	202	C=0.01	70.47
3.2.B.	31	Delta=0.3329	70.25
5.2.A.	32	12 components	69.73
4.2.A.	37	11 components	69.22
5.1.D.	220	C=0.05	68.35
3.2.A.	31	9 components	67.25
1.2.D.	6,687	C=0.01	66.2
2.3.D.	146	C=0.05	66.02
1.1.D.	14,067	C=0.05	65.49
6.0.D.	16,449	C=0.01	64.95
2.1.D.	959	C=0.01	64.24
2.2.D.	342	C=0.01	64.14
4.2.D.	37	C=0.01	63.17
1.3.D.	1,843	C=0.01	60.85
5.2.D.	32	C=0.01	59.79
3.2.D.	31	C=0.1/0.01	57.3
5.3.C.	4	2,000 trees/Minimum node size=16	44.31
5.3.B.	4	Delta=0.2938	43.61
5.3.A.	4	3 components	42.89
3.3.A.	4	3 components	41.27
3.3.B.	4	Delta=0.2718	40.92
4.3.A.	4	3 components	40.48
3.3.C.	4	500 trees/ Minimum node size=21	40.31
4.3.B.	4	Delta=0.7697	40.22
4.3.C.	4	500 trees/Minimum node size=48	39.15
3.3.D.	4	C=0.01	38.8
4.3.D.	4	C=0.01	38.06
5.3.D.	4	C=0.1	37.74

Appendix 3.

Ranked results of validation obtained on a) the dataset without any Hardy-Weinberg filter and b) the dataset filtered out for Hardy-Weinberg equilibrium P-value smaller than 10^{-6}. Only models with global accuracy greater than 90% in cross-validation were tested on validation.

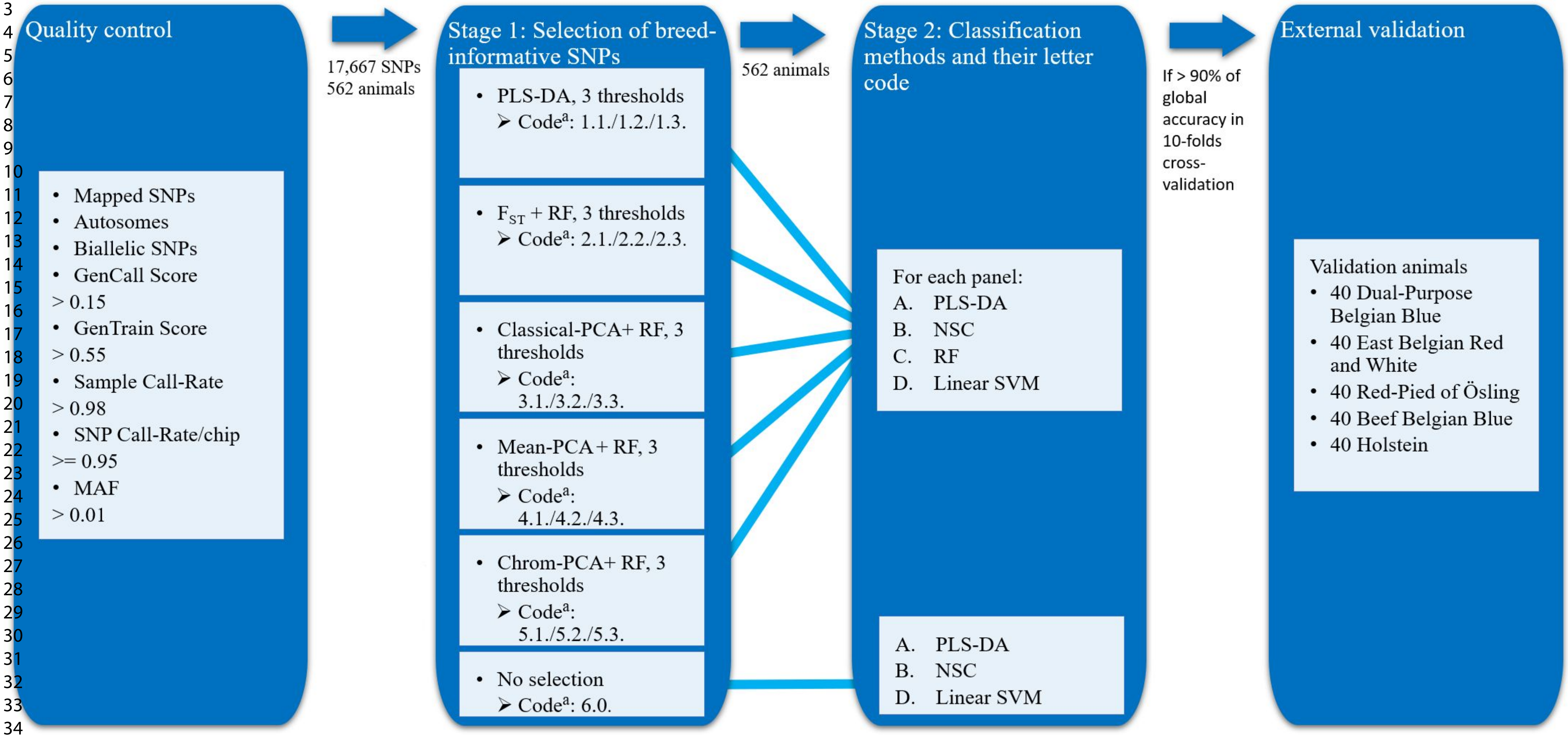
a)

Case	Number of SNPs	Global accuracy (%)	Sensitivity (%)	Specificity (%)
1.2.B.	7,153	99	BBB 100/DPBB 100/EBRW 100/HOL 100/RPO 95	BBB 100/CAM 100/DIRP 100/DPBB 100/EBRW 98.75/HOL 100/MRY 100/RDN 100/RDP 100/RPO 100/SHO 100/SIM 100
1.3.B.	2,005	98.5	BBB 100/DPBB 100/EBRW 100/HOL 100/RPO 92.5	BBB 100/CAM 100/DIRP 100/DPBB 100/EBRW 98.12/HOL 100/MRY 100/RDN 100/RDP 100/RPO 100/SHO 100/SIM 100
1.1.B.	15,102	98	BBB 100/DPBB 100/EBRW 100/HOL 100/RPO 90	BBB 100/CAM 100/DIRP 100/DPBB 100/EBRW 99.38/HOL 100/MRY 100/RDN 98.5/RDP 100/RPO 100/SHO 100/SIM 100
6.0.B.	17,667	98	BBB 100/DPBB 100/EBRW 100/HOL 100/RPO 90	BBB 100/CAM 100/DIRP 100/DPBB 100/EBRW 99.38/HOL 100/CAM 100/MRY 100/RDN 98.5/RDP 100/RPO 100/SHO 100/SIM 100
1.1.A.	15,102	97.5	BBB 100/DPBB 100/EBRW 95/HOL 100/RPO 92.5	BBB 99.38/CAM 100/DIRP 100/DPBB 100/EBRW 100/HOL 96.88/MRY 98.5/RDN 100/RDP 100/RPO 100/SHO 100/SIM 100
6.0.A.	17,667	97.5	BBB 100/DPBB 100/EBRW 95/HOL 100/RPO 92.5	BBB 100/CAM 99.38/DIRP 100/DPBB 99.38/EBRW 100/HOL 100/MRY 98.5/RDN 100/RDP 100/RPO 100/SHO 100/SIM 100
2.1.B.	1,014	97.5	BBB 100/DPBB 100/EBRW 92.5/HOL 100/RPO 95	BBB 100/CAM 99.5/DIRP 99.5/DPBB 100/EBRW 98.75/HOL 100/MRY 100/RDN 100/RDP 100/RPO 99.38/SHO 100/SIM 100
1.2.A.	7,153	97	BBB 100/DPBB 100/EBRW 92.5/HOL 100/RPO 92.5	BBB 99.38/CAM 100/DIRP 100/DPBB 100/EBRW 100/HOL 100/MRY 97.5/RDN 100/RDP 100/RPO 100/SHO 100/SIM 100
2.1.A.	1,014	97	BBB 100/DPBB 100/EBRW 90/HOL 100/RPO 95	BBB 100/CAM 99.5/DIRP 100/DPBB 100/EBRW 100/HOL 99.38/MRY 98.50/RDN 100/RDP 100/RPO 99.38/SHO 100/SIM 100
1.3.A.	2,005	96	BBB 100/DPBB 100/EBRW 85/HOL 100/RPO 95	BBB 99.38/CAM 99.5/DPBB 100/EBRW 100/HOL 99.38/MRY 98.5/RDN 100/RDP 99.5/RPO 99.38/SHO 100/SIM 100
2.2.B.	396	93	BBB 100/DPBB 100/EBRW 72.5/HOL 100/RPO 92.5	BBB 100/CAM 97.5/DIRP 99/DPBB 100/EBRW 98.12/HOL 100/MRY 100/RDN 100/RDP 100/RPO 97.5/SHO 100/SIM 100
4.1.B.	221	90.5	BBB 100/DPBB 97.5/EBRW 65/HOL 100/RPO 90	BBB 99.38/CAM 95.5/DIRP 99.5/DPBB 100/EBRW 97.5/HOL 100/MRY 100/RDN 100/RDP 100/RPO 97.5/SHO 100/SIM 100
2.2.A.	396	89.5	BBB 100/DPBB 97.5/EBRW 70/HOL 100/RPO 80	BBB 98.75/CAM 99.5/DIRP 99.5/DPBB 99.38/EBRW 100/HOL 100/MRY 95/RDN 99.5/RDP 100/RPO 96.88/SHO 100/SIM 100
5.1.B.	228	51	BBB 60/DPBB 40/EBRW 60/HOL 47.5/RPO 47.5	BBB 100/CAM 84/DIRP 85/DPBB 96.88/EBRW 81.25/HOL 100/MRY 100/RDN 99.5/RDP 100/RPO 100/SHO 100/SIM 100

b)

Case	Number of SNPs	Global accuracy (%)	Sensitivity (%)	Specificity (%)
1.2.B.	6,687	99	BBB 100/DPBB 100/EBRW 100/HOL 100/RPO 95	BBB 100/CAM 100/DIRP 100/DPBB 100/EBRW 99.38/HOL 100/MRY 100/RDN 99.5/RDP 100/RPO 100/SHO 100/SIM 100
1.1.B.	14,067	98	BBB 100/DPBB 100/EBRW 100/HOL 100/RPO 90	BBB 100/CAM 100/DIRP 100/DPBB 100/EBRW 99.38/HOL 100/MRY 100/RDN 98.5/RDP 100/RPO 100/SHO 100/SIM 100
6.0.B.	16,449	98	BBB 100/DPBB 100/EBRW 100/HOL 100/RPO 90	BBB 100/CAM 100/DIRP 100/DPBB 100/EBRW 99.38/HOL 100/MRY 100/RDN 98.5/RDP 100/RPO 100/SHO 100/SIM 100
1.3.B.	1,843	97.5	BBB 100/DPBB 100/EBRW 95/HOL 100/RPO 92.5	BBB 100/CAM 99/DIRP 100/DPBB 100/EBRW 98.12/HOL 100/MRY 100/RDN 100/RDP 100/RPO 100/SHO 100/SIM 100
1.1.A.	14,067	97.5	BBB 100/DPBB 100/EBRW 95/HOL 100/RPO 92.5	BBB 99.38/CAM 100/DIRP 100/DPBB 100/EBRW 100/HOL 100/MRY 98/RDN 100/RDP 100/RPO 100/SHO 100/SIM 100
1.2.A.	6,687	97	BBB 100/DPBB 100/EBRW 92.5/HOL 100/RPO 92.5	BBB 99.38/CAM 100/DIRP 100/DPBB 100/EBRW 100/HOL 100/MRY 97.5/RDN 100/RDP 100/RPO 100/SHO 100/SIM 100
6.0.A.	16,449	97	BBB 100/DPBB 100/EBRW 95/HOL 100/RPO 90	BBB 99.38/CAM 100/DIRP 100/DPBB 100/EBRW 99.38/HOL 100/MRY 98/RDN 100/RDP 100/RPO 100/SHO 100/SIM 100
1.3.A.	1,843	96.5	BBB 100/DPBB 100/EBRW 87.5/HOL 100/RPO 95	BBB 99.38/CAM 99.5/DIRP 100/DPBB 100/EBRW 100/HOL 99.38/MRY 98.5/RDN 100/RDP 100/RPO 99.38/SHO 100/SIM 100
2.1.B.	959	96.5	BBB 100/DPBB 100/EBRW 92.5/HOL 100/RPO 90	BBB 100/CAM 99/DIRP 99.5/DPBB 100/EBRW 97.5/HOL 100/MRY 100/RDN 100/RDP 100/RPO 100/SHO 100/SIM 100
2.1.A.	959	95	BBB 100/DPBB 95/EBRW 85/HOL 100/RPO 95	BBB 98.75/CAM 99.5/DIRP 99.5/DPBB 100/EBRW 100/HOL 100/MRY 98/RDN 100/RDP 100/RPO 98.75/SHO 100/SIM 100
2.2.B.	342	93.5	BBB 100/DPBB 100/EBRW 75/HOL 100/RPO 92.5	BBB 100/CAM 97/DIRP 99/DPBB 100/EBRW 98.75/HOL 100/MRY 99.5/RDN 100/RDP 100/RPO 98.75/SHO 100/SIM 100
4.1.B.	203	93	BBB 100/DPBB 97.5/EBRW 77.5/HOL 100/RPO 90	BBB 99.38/CAM 96.50/DIRP 99.5/DPBB 100/EBRW 97.5/HOL 100/MRY 100/RDN 100/RDP 100/RPO 99.38/SHO 100/SIM 100
2.2.B.	342	89.5	BBB 100/DPBB 100/EBRW 65/HOL 100/RPO 82.25	BBB 99.38/CAM 97/DIRP 99/DPBB 99.38/EBRW 99.38/HOL 100/MRY 96/RDN 100/RDP 100/RPO 98.75/SHO 100/SIM 100

Abbreviations NSC: Nearest Shrunken Centroids; PLS-DA: Partial Least Squares-Discriminant Analysis; BBB: Beef Belgian Blue; DPBB: Dual-Purpose Belgian Blue; EBRW: East Belgian Red and White; HOL: Holstein; RPO: Red-Pied of Ösling; CAM: Belgian Campine; DIRP: Dutch improved Red Pied; MRY: Meuse-Rhine-Yssel; RDN: Rotbunte DN; RDP: Rouge des Prés; SHO: Shorthorn; SIM: Simmental.



Quality control

- Mapped SNPs
- Autosomes
- Biallelic SNPs
- GenCall Score > 0.15
- GenTrain Score > 0.55
- Sample Call-Rate > 0.98
- SNP Call-Rate/chip >= 0.95
- MAF > 0.01

17,667 SNPs
562 animals

Stage 1: Selection of breed-informative SNPs

- PLS-DA, 3 thresholds
  - Code[a]: 1.1./1.2./1.3.
- F_{ST} + RF, 3 thresholds
  - Code[a]: 2.1./2.2./2.3.
- Classical-PCA+ RF, 3 thresholds
  - Code[a]: 3.1./3.2./3.3.
- Mean-PCA + RF, 3 thresholds
  - Code[a]: 4.1./4.2./4.3.
- Chrom-PCA+ RF, 3 thresholds
  - Code[a]: 5.1./5.2./5.3.
- No selection
  - Code[a]: 6.0.

562 animals

Stage 2: Classification methods and their letter code

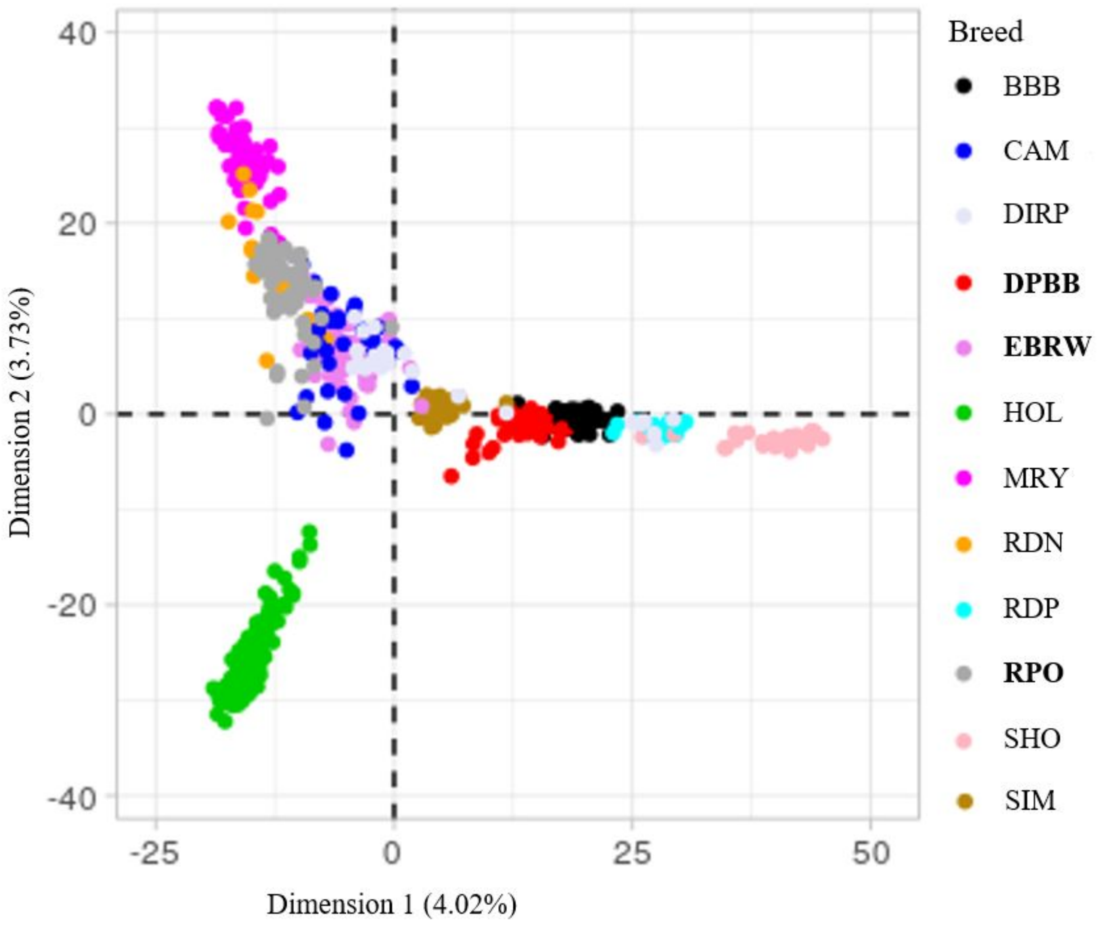
For each panel:
A. PLS-DA
B. NSC
C. RF
D. Linear SVM

A. PLS-DA
B. NSC
D. Linear SVM

If > 90% of global accuracy in 10-folds cross-validation

External validation

Validation animals
- 40 Dual-Purpose Belgian Blue
- 40 East Belgian Red and White
- 40 Red-Pied of Ösling
- 40 Beef Belgian Blue
- 40 Holstein


Breed
BBB
CAM
DIRP
DPBB
EBRW
HOL
MRY
RDN
RDP
RPO
SHO
SIM
Dimension 2 (3.73%)
Dimension 1 (4.02%)
40
20
0
-20
-40
-25
0
25
50