Original scientific paper
10.7251/AGRENG1802041K
UDC 577.2

GENETIC MARKERS AND BIOSTATISTICAL METHODS AS APPROPRIATE TOOLS TO PRESERVE GENETIC RESOURCES

Veronika KUKUČKOVÁ*, Radovan KASARDA, Július ŽITNÝ, Nina MORAVČÍKOVÁ

Slovak University of Agriculture in Nitra, Faculty of Agrobiology and Food Resources, Department of Animal Genetics and Breeding Biology, Tr. A. Hlinku 2, 949 76 Nitra, Slovakia
*Corresponding author: veron.kukuckova@gmail.com

ABSTRACT

The aim of presented study was to assess the most suitable way how to distinguish different breeds based on molecular markers. One of the most difficult aspects of quality assurance schemes is their reliability. The verification of fraud needs great efforts in control strategies. The use of DNA markers has been shown to be a useful tool for individual identification. It is necessary to use modern statistical method based on data mining and supervised learning. Supervised pattern recognition techniques use the information about the class membership of the samples to a certain group (class or category) in order to classify new unknown samples in one of the known classes on the basis of its pattern of measurements. Large scale of supervised learning oriented method was used for traceability and identification on individual level. A result of provided study shows the possibility to classify unknown samples according to genetic data. Model is also useful for classification on many logical levels as brand, region and many others. If we take in the account only Slovak and Austrian Pinzgau cattle, based on SNP chip data, it is not possible to separate them using Bayesian approach. Once we considered with the admixture of breeds involved in the historical development as well as inbreeding, selection signatures and migration, we were able to separate even genetically similar breeds. It is possible distinguish between closely related populations based on different markers. We just need to select the appropriate type of analysis.

Keywords: *cattle, markers, supervised learning, structure assessment.*

INTRODUCTION

Research of cattle breeding is a complex contemporary issue of interdisciplinary scientific interest, including research of agricultural landscapes and resilient urban food systems (Tóth et al, 2016). Considering that wild cattle no longer exist and that all surviving genetic diversity is now present in domestic animals, a better understanding of cattle genetics could help us to reduce some of these undesirable effects (Canavez et al., 2012). Molecular markers have been comprehensively exploited to access genetic variability as they contribute information on every region of the genome, regardless of the level of gene expression. Employment of microsatellite markers is one of the most powerful means for studying the genetic diversity, calculation of genetic distances, detection of bottlenecks and admixture because of high degree of polymorphism, random distribution across the genome, codominance and neutrality with respect to selection (Putman and Carbone, 2014).

Machine learning (ML) is the science of building systems that automatically learn from data (Swan et al, 2013). The ML represents a set of topics dealing with the creation and evaluation of algorithms that facilitate pattern recognition, classification, and prediction, based on models derived from existing data. The data can present identification patterns which are used to classify into groups. The result of the analysis is the pattern which can be used for identification of data set without the need to obtain input data used for creation of this pattern. An important requirement in this process is careful data preparation validation of model used and its suitable interpretation (Židek et al., 2014). Tarca et al. (2007) described supervised as well as unsupervised learning methods in their study. In supervised learning, objects in a given collection are classified using a set of attributes, or features. The result of the classification process is a set of rules that prescribe assignments of objects to classes based solely on values of features. Supervised learning algorithms induce models from these training data and these models can be used to classify other unlabelled data (Židek et al., 2014). During the last twenty years, supervised learning has been a tool of choice to analyse the always increasing and complexifying data generated in the context of molecular biology, with successful applications in genome annotation, function prediction, or biomarker discovery (Guerts et al., 2009).

However, recent advances in genome sequencing and high-throughput DNA techniques has led to the development of single nucleotide polymorphism (SNP) genotyping arrays as a new molecular tool. Single nucleotide polymorphism arrays provide information on a large number of markers distributed over the whole genome at an affordable price. Consequently, this improvement enables a more realistic estimation of genetic diversity, population structure and admixture level (Kukučková et al., 2017). High-throughput technologies have already been used in many areas, as a genomic inbreeding measure (Ferenčaković et al., 2013), genetic and population structure (Mastrangelo et al., 2014).

The aim of this study was to classify the observed animals into Slovak and Austrian cattle using supervised and unsupervised learning models based on different molecular data.

MATERIALS AND METHODS

DNA of 412 selected Slovak (346) and Austrian (66) Pinzgau cows was isolated from hair roots and amplified in one multiplex PCR with 8 microsatellites (TGLA227, SPS115, ETH3, BM1824, CSRM60, CSSM66, TGLA122 and INRA23) localized on 8 chromosomes (18, 15, 19, 1, 10, 14, 21 and 3, respectively). The polymorphism of microsatellite sequences was determined by fluorescent fragmentation analysis using capillary electrophoresis and the sizes of alleles were evaluated. All observed animals were divided into 2 logical groups based on country of origin. The classification models for identity verification of animals was developed. Statistical analysis was performed using Tanagra software (Rakotomalala, 2005).

Data mining statistical approaches using supervised classification were used in the learning phase. In total, 20 different methods of supervised machine learning and their ability to classify examined data were analysed. The basic output of supervised learning methods was "confusion matrix" representing the number of classified individuals using statistical method to some logical group. Bootstrapping and cross validation have been applied to minimize the model error. For construction of the algorithm in the using phase 75% of the data were used and remaining 25% were presented to algorithm as unknown classification.

The software Tanagra 1.4 was used for analysis of relatedness and principal component analysis (PCA) of microsatellite data (Rakotomalala, 2005). PCA is used to characterize how different multiple populations are, often using only the two first principal components (Albrechtsen et al., 2010). Mixture partition based on unsupervised clustering using Bayesian Analysis of Population Structure (BAPS v. 6.0) software was executed, further described in Cheng et al. (2013). The interpretation of the optimal number of clusters is directly inferred by the implemented algorithm in BAPS. The maximum number of clusters was set to 1, 2 or 4, repeated ten times. Each run has led to the same results.

Slovak Pinzgau cattle were genotyped by the Illumina BovineSNP50 v2 BeadChip (Illumina Inc., San Diego, CA). Ten active breeding bulls of Pinzgau cattle from Slovakia used in breed management were analysed. Genotyping information (BovineSNP50 v1 BeadChip) for 33 Austrian Pinzgau sires described in Ferenčaković et al. (2013) was used. The consensus map with the same number of autosomal SNPs for both breeds considered in the further analysis was firstly created. The population structure and the admixture level were inferred by the program BAPS version 6.0 (Corander et al. 2004) where the interpretation of the optimal number of clusters is directly inferred by the implemented algorithm. The maximum number of clusters was set to 5, because this number it is recommended to be higher than expected number of populations (Corander et al. 2004). An admixture analysis conditional on the optimal genetic mixture estimated from the individual level analysis was performed. Results were based on 5000 simulations from the posterior allele frequencies. The number of clusters containing more than 10 individuals as a point estimate of K was used, since the lowest population size was 10. Furthermore, to assess the significance of the admixture estimates, 200 individuals were generated from each identified ancestral source to provide an approximation to the distribution of the estimates under the hypothesis of no admixture. Ten iterations for the reference individuals were run.

RESULTS AND DISCUSSION

A model for animal identity verification was developed using microsatellite panel and machine learning methods. The reliability of individual methods was observed by application of all available models of supervised learning for data set preparation. Three of 20 tested methods have been selected with highest value of reliability (Table 1). The method with the lowest algorithm error in direct classification was Rnd Tree, applying decision trees techniques. Methods C4.5 and CS-MC4 appeared as preferred due to memorization phenomenon of Rnd Tree method. Although both methods recorded the higher value of the algorithm error in the phase of direct learning, after verifying the reliability using bootstrapping and cross validation lower error rate was recorded. Using CS-MC4 (C4.5) method 99.6% (98.6%) of animals can be correctly assigned to Slovak population and excluded from Austrian population only with 3.6% (7.9%) error rate.

Modern biology can benefit from the advancements made in the area of machine learning. Caution should be taken when judging the superiority of some machine learning approaches over other categories of methods. Of special concern with supervised applications is that all steps involved in the classifier design (selection of input variables, model training, etc.) should be cross-validated to obtain an unbiased estimate for classifier accuracy. For instance, selecting the features using all available data and subsequently cross-validating the classifier training will produce an optimistically biased error estimate (Tarca et al., 2007).

Table 1 Reliability of learning process, validation reliability (bootstrapping and cross validation) and reliability of using process expressed as a percentage

Method		Recall	Precision	Algorithm error	Bootstrap .632+	CV error	Recall	Precision	Method error
C4.5	A	87.9	92.1	3.16	**7.4**	8.1	84.2	88.9	**4.8**
	S	98.6	97.7				97.6	96.5	
CS-MC4	A	57.5	96.4	6.8	8.1	**7.8**	79.0	93.7	**4.8**
	S	99.6	92.9				98.8	95.4	
Rnd Tree	A	100	100	**0.0**	9.0	12.0	73.7	70.0	10.7
	S	100	100				92.9	94.0	

The test set (25% of individuals) is used for the generalization error assessment of the final chosen models (Table 1). The methods C4.5 and CS-MC4 have been confirmed as the most reliable for classification of animals by country origin ($p < 0.05$). Algorithm is able to mark animals with specific pattern typical only for Slovak population. Precision of assessment is 94-96.5%. Similarly is possible to mark animals which do not belong to pure Slovak Pinzgau with recall probability 92.9-98.8%. The correct classification rate obtained with the reliability validation of the model were sufficient for identification of animals. Evaluation of classification models is essential to determine their ability and accuracy; ideally this would be performed by producing the model on a training set and testing it on an independent test set (Swan et al., 2013). Although in learning process appeared the Rnd tree method as the most appropriate, after verification were all three observed models very balanced. In using process have proven methods C4.5 and CS-MC4 as the most accurate and therefore most suitable for this type of analysis.

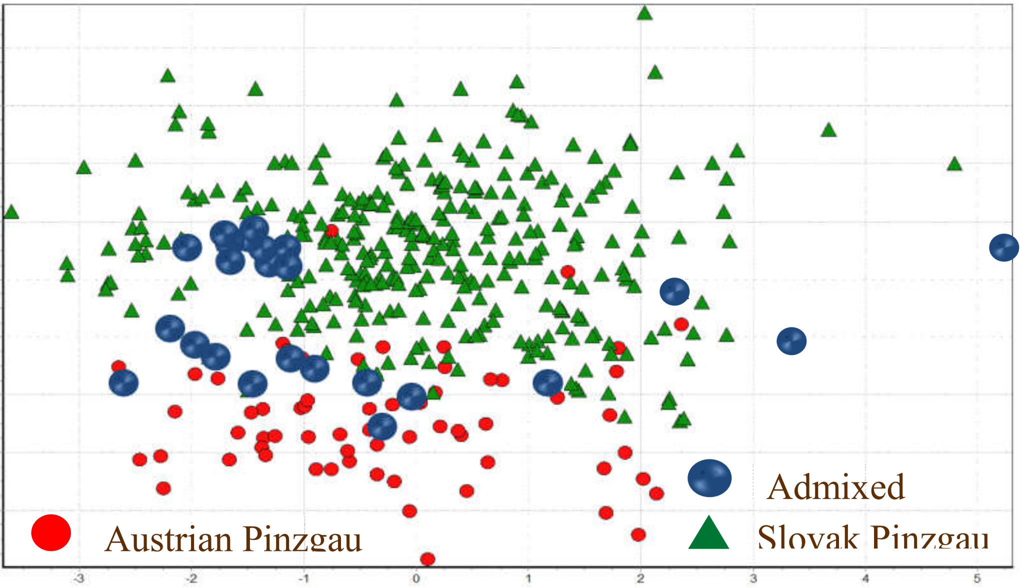

Figure 1. Animals of Slovak Pinzgau admixed with Austrian Pinzgau

The PCA and ancestry model were used to cluster animals, to explore the relationships among and within breeds, and to place the Slovak Pinzgau in a context with closely related Austrian Pinzgau. The PCA of 3 Pinzgau groups is visualised in figure 1. Slovak and Austrian Pinzgau created separate clusters although populations are very close. Admixed animals were between those 2 groups regardless of whether it was the Slovak and Austrian individual. Mixture partition based on unsupervised clustering using Bayesian approach clearly distinguished even genetically similar breeds (Figure 2). The approach used for populations' structure assessment is characterized as unsupervised learning methods with specific computation algorithm. It is possible to use or do not use a priori information about population unlike supervised learning.

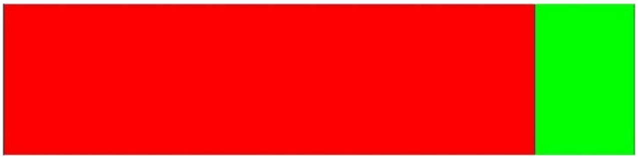

Figure 2 Stacked bar plot of the cluster membership suggested by the BAPS algorithm ("unsupervised") presenting Slovak cattle in green and Austrian cattle in red

In comparison to microsatellite analysis the high-throughput genotyping data was used in subsequent analysis. Using genomic information estimated from 43 animals and 41,135 SNPs, the population structure of 2 cattle breeds was evaluated. A detailed analysis of genetic structure at both the individual and population level was performed based on the Bayesian clustering method adopted in BAPS. Comparing only Slovak and Pinzgau population based on SNP chip data it is not possible to separate them. Since both populations of Pinzgau cattle have the same origin and thus they are genetically similar, the Bayesian approach considered both populations as one cluster. But finally we were able to separate even these closely related breeds (S and A) since they incorporated the admixture with breeds involved in the historical development as well as inbreeding, selection signatures and migration. Each of the 15 clusters presented one exact population from the metapopulation of 15 European breeds (Figure 3). It is possible distinguish between closely related populations based on different markers. We just need to select the appropriate type of analysis.

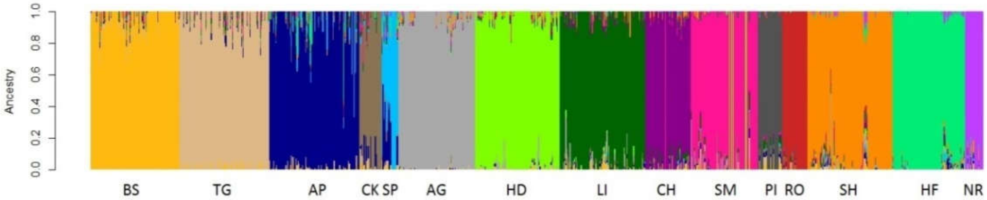

Figure 3. Posterior admixture analysis for 15 European cattle breeds based on the optimal genetic mixture estimate with 15 clusters using the BAPS uniform prior clustering model for individuals. Brown Swiss (BS), Tyrol Grey (TG), Austrian Pinzgau (AP), Cika (CK), Slovak Pinzgau (SP), Angus (AG), Hereford (HD, Limousine (LI), Charolais (CH), Simmental (SM), Piedmontese (PI), Romagnola (RO), Shorthorn (SH), Holstein (HF), Norwegian Red (NR).

The intensive selection of Slovak and Austrian Pinzgau cattle due to mass artificial insemination could increase the similarity among animals. The change in breeding goals to preserve the dual-purpose character of the Slovak Pinzgau was proposed for the long time, and consequently, a positive impact on population structure is expected. According to Jemma et al. (2015), the presence of purebred local individuals has become rare and thus highlights the need to implement a national conservation strategy. There is clearly a race between the characterization of genetic resources and their loss. In the same way, the development of genomic tools will allow to optimize the breeding strategies for ensuring the improvement of performance together with the preservation of genetic diversity. For breeders, it is important to know the origin of animals from the point of the genetic diversity. In case of missing pedigree information, other methods can be used for traceability of animal´s origin. Genetic diversity written in genetic data is holding relatively useful information to identify animals originated from individual countries (Židek et al., 2014).

CONCLUSION

Many of the local farm animal breeds substituted by more efficient breeds in the past are now endangered and preserved in situ as small populations in some regions. The possible extinction of these breeds would also mean irrecoverable loss of the genetic variability and so the damage of unique gene and allele combinations that would be very useful in the future for the generation of new farm animal genotypes. The global breeding program including very close populations will be more efficient providing higher genetic progress and diversity. Classification of individuals on the level of DNA is a valuable tool for origin traceability. The use of supervised learning allowed apparent distinction of closely related animals with Austrian and Slovak origin based on microsatellite markers. We can conclude the correct classification rate obtained with the reliability validation of the model were sufficient for identifying of animals. Datamining techniques based on genetic data are applicable in protection of Pinzgau cattle, breeding management and herdbook core conservation. Using high-throughput molecular information based on the method with linked markers, including inbreeding, gene flow, mutation, and thus introgression of other breeds, the more accurate view on the genetic structure of the observed breed was successfully performed. Presented methodology for differentiation of genealogically close breeds (Slovak and Austrian Pinzgau) based on various molecular markers can be proposed as general, how to distinguish among all highly related breeds.

ACKNOWLEDGEMENT

This study was supported by the Slovak Research and Development Agency under the Contract no. APVV-14-0054.

REFERENCES

Albrechtsen A., Nielsen F.C., Nielsen, R. (2010). Ascertainment Biases in SNP Chips Affect Measures of Population Divergence. Molecular Biology and Evolution, 27(11), 2534-47.

Canavez F.C., Luche D.D., Stothard P., Leite K.R.M., Sousa-Canavez J.M., Plastow G., Meidanis J., Souza M.A., Feijao P., Moore S.S., Camara-Lopes L.H. (2012). Genome Sequence and Assembly of Bos indicus. Journal of Heredity, 103(3), 342-348.

Corander J., Waldmann P., Marttinen P., Sillanpä M.J. (2004). BAPS 2: enhanced possibilities for the analysis of genetic population structure. Bioinformatics, 20, 2363-2369.

Cheng L., Connor T.R., Sirén J., Aanensen D.M., Corander J. (2013). Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. Molecular Biology and Evolution, 30(5), 1224-1228.

Ferenčaković M., Solkner J., Curik, I. (2013). Estimating autozygosity from high-throughput information: effects of SNP density and genotyping errors. Genetics Selection Evolution, 45(1), 42.

Guerts P., Irrthum A., Wehenkel L. (2009). Supervised learning with decision tree-based methods in computational and systems biology. Molecular BioSystems. 5(12), 1593-1605.

Kasarda R., Mészáros G., Kadlečík O., Hazuchová E., Šidlová V., Pavlík I. (2014). Influence of mating systems and selection intensity on the extent of inbreeding and genetic gain in the Slovak Pinzgau cattle. Czech Journal of Animal Science, 59(5), 219-226.

Kukučková V., Moravčíková N., Ferenčaković M., Simčič M., Mészáros G., Sölkner J., Trakovická A., Kadlečík O., Curik I., Kasarda R. (2017). Genomic characterization of Pinzgau cattle: genetic conservation and breeding perspectives. Conservations Genetics.

Jemma S.M., Boussaha M., Mehdi M.B., Lee J.H., Lee S.-H. (2015). Genome-wide insights into population structure and genetic history of tunisian local cattle using the illumina bovinesnp50 beadchip. BMC Genomics, 16, 677.

Mastrangelo S., Saura M., Tolone M., Salces-Ortiz J., Di Gerlando R., Bertolini F., Fontanesi L., Sardina M.T., Serrano M., Portolano B. (2014). The genome-wide structure of two economically important indigenous Sicilian cattle breeds. Journal of Animal Science, 92, 4833-4842.

Putman A.I., Carbone I. (2014). Challenges in analysis and interpretation of microsatellite data for population genetic studies. Ecology and Evolution, 4(22), 4399-4428.

Swan A.L., Mobasheri A., Allaway D., Liddell S., Bacardit J. (2013). Application of Machine Learning to Proteomics Data: Classification and Biomarker Identification in Postgenomics Biology, OMICS, 17(12), 595-610.

Tarca A.L., Carey V.J., Chen X.W., Romero R., Drăghici S. (2007). Machine Learning and Its Applications to Biology. Plos Computational Biology, 3(6), e116.

Tóth A., Rendall S., Reitsma F. (2016). Resilient food systems: A qualitative tool for measuring food resilience. Urban Ecosystems, 19(1), 19-43.

Židek R., Šidlová V., Kasarda R, Fuerst-Waltl B. (2014). Methods for Distinction of Cattle Using Supervised Learning. International Journal of Biological, Veterinary, Agricultural and Food Engineering, 8(5), 500-502.