GENOME-WIDE SELECTION OF DISCRIMINANT SNP MARKERS FOR BREED ASSIGNMENT IN INDIGENOUS SHEEP BREEDS*

Mohammad Hossein Moradi^{1*}, Amir Hossein Khaltabadi-Farahani¹, Mahdi Khodaei-Motlagh¹, Mehdi Kazemi-Bonchenari¹, John McEwan²

¹Department of Animal Science, Faculty of Agriculture and Natural Resources, Arak University, Arak 38156-8-8349, Iran

²Centre for Reproduction and Genomics, AgResearch, Invermay, Mosgiel, New Zealand

*Corresponding author: Hosenmoradi@ut.ac.ir; h-moradi@araku.ac.ir

Abstract

The assignment of an individual to the true population of origin is one of the most important applications of genomic data for practical use in animal breeding. The aim of this study was to develop a statistical method and then, to identify the minimum number of informative SNP markers from high-throughput genotyping data that would be able to trace the true breed of unknown samples in indigenous sheep breeds. The total numbers of 217 animals were genotyped using Illumina OvineSNP50K BeadChip in Zel, Lori-Bakhtiari, Afshari, Moqani, Qezel and a wild-type Iranian sheep breed. After SNP quality check, the principal component analysis (PCA) was used to determine how the animals allocated to the groups using all genotyped markers. The results revealed that the first principal component (PC₁) separated out the two domestic and wild sheep breeds, and all domestic breeds were separated from each other for PC₂. The genetic distance between different breeds was calculated using F_{ST} and Reynold methods and the results showed that the breeds were well differentiated. A statistical method was developed using the stepwise discriminant analysis (SDA) and the linear discriminant analysis (LDA) to reduce the number of SNPs for discriminating 6 different Iranian sheep populations and K-fold cross-validation technique was employed to evaluate the potential of a selected subset of SNPs in assignment success rate. The procedure selected reduced pools of markers into 201 SNPs that were able to exactly discriminate all sheep populations with 100% accuracy. Moreover, a discriminate analysis of principal components (DAPC) developed using 201 linearly independent SNPs revealed that these markers were able to assign all individuals into true breed. Finally, these 201 identified SNPs were successfully used in an independent out-group breed consisting of 96 samples of Baluchi sheep breed and the results indicated that these markers are able to correctly allocate all unknown samples to true population of origin. In general, the results of this study indicated that the combined use of the SDA and LDA techniques represents an efficient strategy for selecting a reduced pool of highly discriminant markers.

Key words: indigenous sheep breeds, assignment methods, discriminant SNP markers, principal component analysis (PCA), linear discriminant analysis (LDA)

*The study was financed from funds of the project: "Identification of discriminant SNP markers and development of technical knowledge for diagnostic kits in indigenous sheep breeds" co-financed by the Arak University (Iran) and AgResearch (New Zealand) under the Research's project, Project no: 95.11547.

Pedigree information is essential for accurate genetic evaluation (Heaton et al., 2014). However, in the most of sheep populations reared in the world, the accuracy of pedigree information is limited and varies widely (Tortereau et al., 2017). The assignment of an individual to the true population of origin is one of the most appealing opportunities to exploit genomic data for practical use (Dimauro et al., 2013; Panetto et al., 2017). Assigning animals to their true populations, where training data (known population membership) are available, is known as discriminant analysis in the statistical literature or supervised learning in the machine learning literature (Dodds et al., 2014). An assignment method relies on the use of genetic information and statistical procedure to determine the origin of unknown individuals (for a review see Manel et al., 2005; Negrini et al., 2009). In livestock species, DNA-based assignment tests can be used in breed assignment, improvement, conservation and are generally used to trace the origin of animal products (Hulsegege et al., 2019; Schwägele, 2005). Recently, the assignment analysis has been mostly focused on cattle breeding (eg. Negrini et al., 2009; Bertolini et al., 2018); while, given the weakness of pedigree information and consequence limitations of sheep breeding, the development of these methods in sheep, especially indigenous breeds can be on priority of animal breeding researches.

Assignment tests were traditionally carried out by using microsatellites (e.g., Dalvit et al., 2008) or AFLP markers (e.g., Negrini et al., 2007). Compared to these markers, single nucleotide polymorphisms (SNP) offer the advantage such as lower rates of genotyping errors (Weller et al., 2006) and very abundant over the genome (Heaton et al., 2014). Numerous SNPs have been identified in the genome of domestic animals, for example, in the dog (>14 million) (Plassais et al., 2019), chicken (>10.4 million) (Ni et al., 2017) and cattle (>10 million) (Choi et al., 2014). This has led to the technological development of standard products commonly termed 'SNP Chips', which enable the rapid automated large-scale production of genomic data. The high throughput SNP Chips are now available for several species including the Illumina Ovine 50kSNP BeadChip (Illumina Inc., San Diego, CA) for sheep. Although these high density (HD) SNP chips provide the best coverage of the genomes for genomic evaluations, they may not be cost-effective for assigning goals and currently, HD SNP chips are still too expensive for most agricultural genomic applications. Consequently, interest has increased in the use of low density (LD) SNP chips as a cost-effective solution (Clarke et al., 2014). The practical application of SNP markers for assigning of unknown individuals requires the use of smaller, cost effective and flexible multiplex panels for genotyping (Heaton et al., 2014).

A few statistical approaches have been proposed for assignment goals yet. One of the simplest methods uses the Delta values that measure allele frequency differences, and has been explored by Wilkinson et al. (2011) for cattle. Bowcock et al. (1994) suggested that informative genetic markers may be identified using Wright's F_{ST} (Wright, 1950) and its derivatives (Weir and Cockerham, 1984). However, these approaches are computationally intensive and their execution may be prohibitively slow with large datasets (Wilkinson et al., 2011). Principal component analysis (PCA) has also been more recently proposed as an alternative method to determine population informative SNP markers. This method has been already used in human

populations to characterize their structure based on SNP genotyping data (Paschou et al., 2007) and in cattle to identify breed informative SNPs (Bertolini et al., 2018). A drawback of this method is that it does not easily translate into estimated breed proportions; and its main use would be to verify that an animal is (or is not) allocated to the recorded breed (Dodds et al., 2014). A popular method for understanding population structure is the model-based clustering method implemented in the program STRUCTURE (Pritchard et al., 2000). This method has been applied to surveys of breed variation in sheep (Kijas et al., 2012). However, it does not produce a prediction equation and requires a re-analysis when new data is added (Dodds et al., 2014). Discriminant analysis (DA) is a powerful classification technique that is suited when classes are well-separated and predictors have a common covariance structure (Hastie et al., 2009). DA is used when groups are known a priori (unlike in cluster analysis) and has been successfully applied to bovine breed assignment for traceability purposes (Dimauro et al., 2013).

Iran is recognized as the reservoir of genetically unique breeds playing crucial roles in the livelihood of the human populations of this area (Zeder, 2012). Archaeological evidence suggests that Iran has been one of the main origins of sheep domestication (Zeder, 2008) and due to its geo-climatic conditions, represents a biodiversity hotspot, with approximately 30 million heads of sheep including 27 breeds and ecotypes well adapted to a patchwork of extremely heterogeneous harsh habitats (Moradi et al., 2012). Livestock has always played a vital role in the agricultural economies of Iran, where commercial livestock enterprises, smallholder and communal farming contributes to food production, and the general wellbeing of rural households. Pedigree information is essential for accurate genetic evaluation; however, in Iranian sheep populations the rate of known sires can vary widely, from a few percent in hardy breeds reared in the rural areas up to 100% in specialized industrial reared breeds (Vatankhah et al., 2004). The lack of complete pedigrees and misidentification of sires affects the accuracy of genetic evaluation and the efficiency of breeding programs (Tortereau et al., 2017). In this situation developing of a statistical method for selecting the discriminant SNPs for breed assigning in Iranian sheep breeds could be crucial, where no report is available yet.

The aim of this study was to identify the minimum number of informative SNPs from high throughput genotyping data for assigning unknown individuals to the true population of origins in Iranian sheep breeds. In this study, two complementary DA statistical techniques including the stepwise discriminant analysis (SDA) and the linear discriminant analysis (LDA) were used to assign independent individuals to the proper group using genomic data. The results of this study could be further used to develop a low cost customized assay to trace the breed or derived foodstuffs.

Material and methods

Animal samples

The genotypic data of 217 animals, coming from distinct geographical conditions of Iran, were collected for this study. They consisted of Zel (47 head), Lori Bakhtiari

(47 head), Afshari (41 head), Qezel (35 head), Moghani (35 head), and a wild-type Iranian sheep breed located in Markazi province (12 head). The geographic distribution of these breeds has been shown in Figure 1. The breeds used in this study are the most common and main indigenous sheep breeds reared in Iran that have been prioritized over the past few years by animal breeding researches and private investment sectors (Moradi et al., 2017).

Figure 1. The studied breeds and their geographic distribution in different regions of Iran

The samples of Zel and Lori-Bakhtiari sheep breeds were obtained from the herds that have been under the pedigree registration system of Animal Breeding Center of Iran (ABCI) during recent years. Sampling in Zel breed was performed from 10 flocks of population in Mazandaran province, near the border between Alburz Mountain and the Caspian Sea, and the samples of Lori-Bakhtiari breed collected from around 3 flocks in Shahr-e Kord city close to Zagros mountainous regions of central part of Iran (Moradi et al., 2012). The genotype data from other Iranian domestic sheep breeds of Afshari, Ghezel and Moghani were obtained from the international sheep HapMap project that have already described by Kijas et al. (2009). Data on Iranian wild-type sheep breeds were also collected from Haftad Qolleh pro-

tected area located in the central region of Iran, in Markazi province, Arak. Extraction of DNA in Zel, Lori-Bakhtiari and wild-type sheep breeds was performed using standard protocols. Genotyping was carried out in all animals using Illumina Ovine SNP 50K BeadChip in the Centre for Genomics and Reproduction, Invermay, New Zealand.

SNP quality check

Several steps of quality control were performed by the PLINK v1.07 software (Purcell et al., 2007) to remove animals or SNPs not useful for the analysis in this study. First, all animals with less than 95% of genotyping call rate were discarded from next analysis. It should be noted that the implementation of this step in Zel, Lori-Bakhtiari, and wild-type breeds was carried out in this study, and for Afshari, Moghani and Qezel breeds were earlier performed in the Sheep HapMap project. Then, according to the standards defined by the Wellcome Trust Case Control Consortium, SNPs with call rate less than 95% within each breed and minor allele frequency (MAF) less than 2% in all samples were excluded (Moradi et al., 2012). For the remaining SNPs, the Hardy-Weinberg equilibrium was tested within each of the six populations and SNPs that were not at equilibrium in at least one population ($P\text{-value} < 10^{-6}$) were discarded as genotype errors (Moradi et al., 2012). The SNP locations were mapped for ovine genome assembly version 3.1 using information from the Sheep HapMap dataset (<http://www.sheephapmap.org>) and finally, the SNP markers whose genomic location was unknown were removed from the total markers.

Statistical analysis

Genetic distances

In this research, the distances between breeds were evaluated using two important methods: Reynolds' distance (Reynolds et al., 1983) and the degree of differentiation F_{ST} (Weir and Cockerham, 1984). The Reynolds assumes that the genetic differentiation occurs only by genetic drift and it is based on coancestry coefficients constructed for multi loci data (Dimauro et al., 2015). While Weir and Cockerham's F_{ST} estimator is widely used because it is asymptotically unbiased with respect to sample size, and can compensate for overestimates particularly at low levels of genetic differentiation (Willing et al., 2012).

The Reynolds's distance was calculated by using the following formula in R version 3.4.1 software:

$$\theta_w = \sum_i \sum_u \frac{(X_u - Y_u)^2}{\sum_i (1 - \sum_u [(X_u Y_u)])}$$

Where X_u and Y_u , are standing for the frequency of u^{th} alleles in the i^{th} position of X and Y populations, respectively (Reynolds et al., 1983).

Wright (1950) introduced F-statistics to describe the proportion of genetic diversity within and among populations. Unbiased estimates of Weir and Cockerham F_{ST} were estimated as functions of variance components and fixation indices as described by Weir and Cockerham (1984) and were calculated using R version 3.4.1 software in this study as:

$$\hat{\theta} = \frac{a}{(a + b + c)}$$

Where a is standing for the variance of allele frequencies among populations, b is the variance of allele frequencies among individuals within populations, and c showing the variance of allele frequencies among gametes within individuals.

Principal component analysis (PCA)

In this study, to determine how the animals allocated to the groups using all genotyped markers, a principal component analysis (PCA) was performed using the `prcomp` function in the R version 3.4.1 software. For summarizing data from many variables into a few variables which describe as much of the variation in the data as possible, the variance-covariance matrix of independent variables was first calculated and principal components were extracted. Each new variable has an associated eigenvalue that measures the respective amount of explained variance (Dodds et al., 2014).

Discriminant analysis (DA) and validation

To detect a pool of highly discriminant markers based on their correlation structure using a multivariate technique, two complementary approaches including the stepwise discriminant analysis (SDA) and the linear discriminant analysis (LDA) were used based on Illumina's OvineSNP50 BeadChip genotyping data in Iranian sheep breeds. The DA methods have previously been implemented in some other animal species (eg. Dimauro et al., 2013), however, some modifications have been applied in the current study to decrease the false discovery rate (FDR). Since the number of animals (207) was much lower than the number of markers (~33000) in the current study, the markers located in each chromosome were divided into the subsets according to their locations firstly, so that the number of markers in each subset was lower than the number of animals. The following steps were then used to select markers with the most information for assigning of unknown animals into the true population of origins: 1 – the first subset of markers was used for assigning of animals by using SDA and simultaneously the k-fold cross validation was applied. The SDA is a statistical technique specifically conceived to select the subset of variables that better discriminate groups (Hastie et al., 2009). In particular, to determine which variables will mostly contribute in discriminating groups, all variables are reviewed and evaluated in the stepwise process. In this research, the adopted criterion for selecting variables was the squared partial correlation R^2 (Baba et al., 2004). 2 – The second subset of markers, along with the discriminant markers identified in the first step, was subjected to the same analysis, to recognize the new markers that

are able to assign animals to their true populations. The stepwise selection process stops when all variables included in the model meet the criterion to stay and none of the other variables meets the criterion to enter. 3 – The third subset of markers was used along with the markers identified in the last steps to perform subsequent analysis as described earlier. These steps were performed for all subsets and applied for all chromosomes to finally identify the SNP markers that are able to assign animals with higher accuracy. One of the benefits of this algorithm was that the discriminant markers identified in each step, were used and analyzed again in the subsequent step, so some of them could be removed. This algorithm is expected to reduce the FDR, or expected proportion of discoveries which are falsely rejected (Korthauer et al., 2019) and increase the ability of markers to assign unknown animals.

The LDA was then used to assign animals to their groups using the largest number of linearly independent markers previously selected by the SDA. In practice, these functions applied to each animal to produce a discriminant score as described by Hastie et al. (2009). An individual was assigned to a particular group if its discriminant score was lower than the cutoff value obtained by calculating the weighted mean distance among group centroids (Hastie et al., 2009). Finally, a new run of the SDA was developed to obtain the minimum number of markers able to significantly separate breeds and to correctly assign individuals. KLaR package in R v.3.4.1 (Weihs et al., 2005) was used to develop the SDA and LDA analysis in the current study, and the modifications have been also applied to decrease the FDR.

To test the ability of the selected SNPs in assigning new animals to the proper breed and to increase the confidence level at the selection step, K-fold cross-validation technique was employed to evaluate the potential of a selected subset of SNPs in assignment success rate (Dimauro et al., 2013). Genotyped animals that passed the quality control steps (207) were allocated to ten mutually exclusive subsets for cross-validation. Of the ten subsets, nine were used as training data, and the remaining subset was considered as validation data; hence, each subset was predicted once from the other subsets. The correlation between predicted and real population of all animals in the validation population was defined as the predictive ability of discriminant SNPs for each fold. The accuracy was obtained as the mean of accuracies for ten-fold cross-validations.

In order to have a general view of the demographic composition of the animals and studied breeds using the markers identified during the discriminant analysis (DA), the analysis of the discriminant analysis of principal components (DAPC) was also performed by these markers alone using the adegenet package in R v 3.4.1 software (Jombart et al., 2008). DAPC was conducted without *a priori* group assignment by inferring the most likely number of genetic clusters (K) using the *find.clusters* function in the adegenet package. This function utilizes K -means clustering to calculate a Bayesian information criterion (BIC) value for each potential value of K (the most likely K has the lowest BIC value) and delineates individual group assignments for DAPC.

Finally, to test the ability of selected markers in assigning unknown individuals to their true population of origin in Iranian indigenous sheep breeds, an additional testing has also been implemented for assessing the traceability of predictive model

at the out-group breed. In particular, 96 animals belonging to Baluchi sheep breed were used for this goal, which were reared in eastern parts of Iran and genotyped using the Illumina Ovine SNP50k Beadchip fully described previously by Gholizadeh et al. (2014).

The presence of the genes close to the discriminant markers identified in this study, were acquired by the use of the data mining tool Biomart (Hubbard et al., 2009), by exploring 20kb upstream and downstream from the marker (Carignano et al., 2018).

Results

Data editing

Different steps of data editing were performed on 49,018 SNP markers in 217 samples of Iranian sheep breeds (Table 1). The results indicated that the highest and lowest number of remaining SNP markers were observed in Qezel and wild-type sheep breeds, respectively. However, it should be noted that in this study, the SNPs that were shared in all breeds and successfully completed the quality control steps (33342 SNP markers), were used to carry out subsequent analyses. The average of expected and observed heterozygosity in this common panel of SNPs ranged from 0.362-0.378 and 0.374-0.434 in different Iranian sheep breeds, respectively.

Table 1. Summary of data mining steps performed on genotyping data of Iranian sheep breeds

Step	Zel	Lori-Bakhtiari	Afshari	Moghani	Qezel	Wild-type
Number of animals	47	47	41	35	35	12
Excluding animals with >5% genotyping failure	2	2	4	1	0	1
Remaining animals	45	45	37	34	35	11
Number of SNPs	49018	49018	49018	49018	49018	49018
Excluding SNPs with Call rate<95% or MAF<2%	1558	1883	2613	1911	1555	14948
Excluding SNPs with deviation from HWE (<10–6)	41	45	13	22	31	0
Excluding unmapped SNPs	318	307	307	305	301	215
SNPs remaining	47101	46783	46085	46780	47103	33855

Genetic distances

One of the common questions in population genetics and assignment analysis is the degree of differentiation between populations (Dimauro et al., 2015). The genetic distances between different populations using Reynold’s distance and the degree of differentiation F_{ST} have been shown in Table 2.

Table 2. Unbiased F_{ST} (over diagonal) and Reynolds (low diagonal) genetic distances between different Iranian sheep breeds

Breeds	Zel	Lori-Bakhtiari	Afshari	Moghani	Qezel	Wild-type
Zel		0.121	0.233	0.186	0.202	0.185*
Lori-Bakhtiari	0.464		0.240	0.208	0.196	0.238
Afshari	0.619	0.624		0.099	0.077	0.240
Moghani	0.564	0.588	0.425		0.058	0.207
Qezel	0.584	0.574	0.379	0.331		0.209
Wild-type	0.673	0.721	0.701	0.644	0.677	

*The minimum genetic distances between domesticated breeds with each other as well as wild breed with other breeds are highlighted in the table.

Principal component analysis (PCA)

Scatterplot of the first four principal components using genotypes information was shown in Figure 2. The results of PCA revealed that all animals allocated to the groups that they sampled and the principal components of PC_1 separated out the two domestic and wild sheep breeds and all domestic breeds were separated from each other for PC_2 , although there is some limitation for the separation of the Moghani and Qezel sheep breeds (Figure 2). In this analysis, PC_1 , PC_2 , PC_3 and PC_4 accounted for 35%, 17%, 12%, and 8% of the total variance, respectively.

Figure 2. Scatterplot of the first four principal components using genotypes information. In this figure, green, black, red, turquoise, blue and pink are standing for Zel, Lori-Bakhtiari, Afshari, Moghani, Qezel and the Wild-type sheep breeds, respectively

Breed assignment based on DA approaches and identification of discriminant SNP markers

The DA approaches used in the current study identified a total number of 201 SNP markers that were able to exactly discriminate all sheep populations (even Qezel and Moghani animal samples) with 100% accuracy. Our results revealed that no further gain in power of the assignment was achieved by sampling in excess of 201 SNP markers (data not shown). Figure 3 shows the distinction and grouping of different Iranian sheep breeds using these markers based on the DA plots. The complete list of the 201 linearly independent SNP markers are reported in Supplementary Table 1.

Figure 3. Classification of animals from different Iranian sheep breeds based on linear discriminant (LD) analysis using genotypes information of 201 SNP markers: In this figure, 1 (Black), 2 (Red), 3 (Green), 4 (Blue), 5 (Turquoise) and 6 (Pink) are standing for Zel, Lori-Bakhtiari, Afshari, Moghani, Qezel and the Wild-type sheep breeds, respectively and the first five linear discriminants (LD) were plotted against each other

Figure 4 shows the distribution of discriminant SNP markers obtained in this study on the sheep genome. Also, the presence of the genes close to the discriminant markers identified in this study, was shown in Supplementary Table 2.

Figure 4. Genomic distribution of single nucleotide polymorphism (SNP) markers identified for assigning of different Iranian sheep breeds: In this figure, the X-axis is showing the various chromosomes and Y axis represents the discriminant markers proportions located on each chromosome

The population clustering of the animals and breeds by using 201 discriminant SNP markers identified in this study, was evaluated by discriminant analysis of principal components (DAPC) and the DAPC chart is shown in Figure 5. Based on the Bayesian information criterion (BIC) statistic, the optimal number of clusters in the data set with the lowest BIC value was $K=4$, showing four clusters generated by DAPC.

Figure 5. Discriminate analysis of principal components (DAPC) plot using discriminant SNP markers in Iranian sheep breeds: In this figure, 1, 2, 3, 4, 5 and 6 are standing for Zel, Lori-Bakhtiari, Afshari, Moghani, Qezel and the Wild-type sheep breeds, respectively

Supplementary Figure 1. Dendrogram constructed from unbiased F_{ST} genetic distances between different sheep breeds

Finally, the statistical methods used in this study, were applied in an external dataset including individuals of the Baluchi sheep breed. The results indicated that by only using these 201 linearly independent markers, all sheep in the external validation datasets were correctly assigned (Supplementary Figure 2).

Discussion

Several studies have shown the usefulness of SNP data for identifying breed informative SNPs for discrimination among breeds in a variety of species like cattle (e.g., Negrini et al., 2009), pigs (Ramos et al., 2011), goats (Talenti et al., 2016) and bees (Muñoz et al., 2015). However, a few studies have already been set up for developing the SNP panels in sheep at the international level (Tortereau et al., 2017), and no report is available for selecting the discriminant SNPs for breed assigning in Iranian sheep breeds.

As shown in Table 1, the highest and lowest number of remaining SNP markers, after data editing, was observed in Qezel and wild-type sheep breeds, respectively. This can be explained by ascertainment bias due to the selection of SNPs for arrays designing. This bias could occur considering that Qezel genomic information has been used in the designing of SNP Chip arrays in sheep, while a small number

of wild-type sheep breeds have been used and Iranian wild breeds have not been considered (<http://www.sheephapmap.org>). Similar results have been reported in the previous studies in various species (eg. Lachance and Tishkoff, 2013).

The results of genetic distances indicated that the lowest genetic distance was observed between Qezel-Moghani. These two sheep breeds are commonly distributed in the same geographical regions in northwestern Iran, located in Ardebil and East Azerbaijan provinces. The lowest genetic distance observed between these breeds corresponded to their geographical distribution. Similar results were obtained by Ciani et al. (2014) who observed a phylogeographic gradient, almost completely overlapping the areas of origin of breeds in the genetic variability of Italian breeds. Dimauro et al. (2015) also reported that the genetic distances between Italian sheep breeds are reflecting the geographic distributions using Reynolds' distances between groups confirming the ability of this method for representing the genetic differences between groups.

Another interesting result of the F_{ST} test was observed for comparing the genetic distance of wild-type Iranian sheep breeds with other domesticated breeds. As shown in Table 2 and Supplementary Figure 1, the lowest genetic distance between the wild-type sheep breed was obtained with the Zel breed (0.185). Zel is the only Iranian thin-tail sheep breed (Moradi et al., 2012). This breed is also known as the Aryan breed since the historical evidence shows that the Aryans, who were living in these areas, attempted to domesticate these animals. Hence, there is a general belief that Zel is the contour of wild and domesticated strains in Iran. The nearest distance observed in this study between Zel and wild-type sheep breeds may also be used to confirm this hypothesis.

PCA is an unsupervised linear technique for dimension reduction and allows extracting axes of maximal variation from data sets (Jolliffe and Cadima, 2016). This method is often used for clustering analysis when classes are not well-known and animals of the same breed tend to be located close together in plots of the PCA (Manel et al., 2005). Since it was assumed that some of the animals used in the current study may not be pure and unmanaged crossbreeding could be occurring in their herds, the PCA was used to determine how the animals allocated to the groups using all genotyped markers. The results of this study indicated that the PCA approach was able to separate the breeds from each other, but faced some limitation for the separation of the Moghani and Qezel breeds where closest genetic distances were shown using F_{ST} and Reynolds methods (Figure 2).

Discriminant analysis (DA) is a powerful classification and a supervised statistical learning technique that is suited when classes are well-separated and predictors have a common covariance structure (Manel et al., 2005). Then DA, unlike clustering analysis, uses a priori knowledge on which observation belongs to which group. Our results indicated that a total number of 201 SNP markers were able to exactly discriminate all sheep populations (even Qezel and Moghani animal samples) with 100% accuracy. In general, it is difficult to compare the results obtained here to other studies, since first, most previous studies used microsatellite or AFLP markers (e.g., Dalvit et al., 2008; Negrini et al., 2007) and, second, these studies had only a limited number of markers or have used different statistical approaches. These studies also pri-

marily focused on the practicality of assigning individuals among different breeds with the available markers and were not concerned with how many markers would be required to achieve confident assignment of individual genotypes (Wilkinson et al., 2011).

Maudet et al. (2002) reported 23 microsatellite loci that were able to assign 93% of individuals to their breed of origins in French cattle breeds. Assignment tests based on genotypes derived from microsatellite markers (Dalvit et al., 2008) and AFLP (Negrini et al., 2007) have also been reported in Italian cows. Negrini et al. (2009) used a panel of 90 SNPs assayed on 24 European breeds. Two methods including Bayesian and classical approaches were used and the results showed that Bayesian methods were more accurate (96% vs. 85%) in the allocation of animals of known origin to their populations. However, the low level of accuracy of these methods can be due to the inadequate number of markers with appropriate information. The development of high throughput SNP platforms for livestock genomes has opened new perspectives for breed traceability.

Supplementary Figure 2. Classification of animals from different Iranian sheep breeds and the external independent dataset (Baluchi sheep breed) based on LDA analysis using genotypes information of 201 SNP markers: In this figure, 1, 2, 3, 4, 5, 6 and 7 are standing for Zel, Lori-Bakhtiari, Afshari, Moghani, Qezel, the Wild-type and Baluchi sheep breeds, respectively and first linear discriminants (LD) plotted against each other

Several selection techniques are suggested to determine which SNP panel contains most information to predict breed of origin specially in cattle breeds. Wilkinson et al. (2011) evaluated the methods of marker selection and determining the

minimum number of markers from the BovineSNP50 BeadChip required verifying the origin of individuals in 17 European cattle breeds. Delta, Wright's F_{ST} , Weir and Cockerham's F_{ST} and PCA methods were compared in this study. The results revealed that 95% assignment success rate was achieved with 200 SNP markers across all the examined methods for marker selection and no further gain in power of assignment was achieved by sampling in excess of 200 SNP markers. Dimauro et al. (2013) studied the bulls of three cattle breeds farmed in Italy (Holstein, Brown, and Simmental) using the canonical discriminant analysis (CDA) functions and finally reported a set of 48 SNPs that was able to significantly identify the three breeds and to correctly assign new observations.

It should be considered that the number of markers required for population assignment will depend on the species, the level of genetic differentiation and the statistical methods under consideration (Wilkinson et al., 2011). For instance, within dogs 27% of the genetic variation is found between breeds, whereas for humans the level between populations is only 5%–10% (Parker and Ostrander, 2005). Meadows et al. (2005) studied the mitochondrial genome diversity in different Asian and European sheep breeds, and reported that sheep has largest gene flow and weakest population structure in domestic animals, so that only 2.7% of the total genetic diversity is intercontinental diversity, whereas it is 10 and 50 percent for goat and cow, respectively. Moradi et al. (2017) using mitochondrial markers in Iranian indigenous sheep breeds reported that only 3% of the observed genetic variation is found between breeds, compared to 97% variation within breeds. As a result, the number of SNP markers required for individual assignment and discrimination amongst populations (breeds) will differ between species under consideration and it seems it would be encountered with more restrictions in sheep than many other domestic animals.

As mentioned earlier, a few studies applied assignment techniques in sheep breeds. Heaton et al. (2014) applied the International Sheep Genomics Consortium (ISGC) data to identify parentage SNPs for use in globally diverse breeds and to develop a subset for use in North American sheep. The results of this study showed that, using 163 SNP markers, these breeds can be distinguished from each other. Dimauro et al. (2015) used three statistical techniques to select the most informative markers for discriminating 21 different Italian sheep populations and their 5 geographic areas of origin. They reported that 110 and 108 were able to exactly discriminate the 5 geographic areas and the 21 sheep populations, respectively. The average probability of assignment was around 99%. Tortereau et al. (2017) also designed a set of 249 SNPs for assignment in French sheep breeds. The set was derived from the 54K SNP chip that was used to genotype the thirty main French sheep populations. The panel of 249 SNPs was successfully used in an on-farm test in the BMC breed and resulted in more than 95% of lambs being assigned to a unique sire.

As shown in figure 4, around 40% of breed selection markers identified in the current study, were located in the first three chromosomes, whereas the others were arranged in the remaining portion of the genome. These results were in agreement with Dimauro et al. (2015) who reported that more than 40% linearly independent markers able to significantly discriminate 21 Italian sheep breeds based on their geographic distribution were located in the first three chromosomes.

In the DAPC method, the researcher does not suggest the number of populations, but it is suggested by statistical method. This approach is based on the assumption that the markers are not related and the mating in the populations is randomized (Jombart et al., 2010). Furthermore, the model independent of DAPC method is better than PCA for population stratification correction, because it increases the variance between groups and decreases the variance within the groups. The correct assignment rate or accuracy of the DAPC method in assigning individuals to groups varies from 80% to 97% depending on the number of replicates (Jombart et al., 2010). With this in mind, the results of the DAPC analysis of Iranian sheep breeds (Figure 5) indicated that these breeds are clustered in at least four different groups, which have a fair distance to each other in terms of genetic variation, although the relation between the Afshari, Maghoni and Qezel breeds is closer. The results of this analysis are also consistent with the geographical distribution of these breeds, with the least distance between the Qezel and Moghani breeds, and also the Zel breed (the only Iranian thin tailed sheep breed) can be seen between the wild type sheep breed and the domesticated fat tailed breeds (Figure 5).

Finally, to assess the traceability of predictive model at the out-group breed, the method developed in this study was applied in an external dataset including individuals of Baluchi sheep breed by only using these 201 independent SNP markers. This breed has the largest population of all breeds of sheep in Iran (about 30% sheep population); it is a fat-tailed and dual-purpose breed (wooly and meaty), which is well adapted to the dry and hot climate conditions of eastern Iran (Gholizadeh et al., 2014). Our results indicated that by only using these 201 linearly independent markers, all sheep in the external validation datasets were correctly assigned (Supplementary Figure 2).

In general, the results of our study revealed that the statistical method used in this study was able to identify a panel of 201 SNPs markers that exactly discriminate all Iranian sheep populations in the breeds used to select these markers, and also in the new data from external out-group breed. Likewise, it was able to cover some drawbacks encountered in other standard statistical methods, for example the slow execution with large datasets in F_{ST} (Wilkinson et al., 2011), the limitations for discriminating of closely related breeds in PCA (Dodds et al., 2014), and the requirement of re-analysis for adding new data in Structure (Dodds et al., 2014). The results of our study can provide valuable information for developing a method in designing a low-cost SNP assay for assigning unknown animals to their true population of origins in other sheep breeds. It is expected that the genotyping of 201 discriminant SNP loci selected from ~50000 SNPs, available on Illumina Ovine 50KSNP BeadChip, will commercially decrease the cost of the genotyping for assignment goals using assays such as Sequenom multiplex MassARRAY (Heaton et al., 2014). For comparative purposes, a commercial reduced SNP panel that included 84 SNPs suggested by Clarke et al. (2014), was generated for parentage assignment in New Zealand sheep by using the Sequenom iPLEX Chemistry Assay. Similarly, a cost effective panel comprising the 163 SNPs was developed by Heaton et al. (2014) to be useful for parentage testing across diverse breeds of sheep. It should be considered that the number of samples may affect the discriminant SNP marker list identified in this study, however, most of the similar published reports have been performed with the same or lower sample

size in different indigenous sheep breeds (eg. Dimauro et al., 2015; Tortereau et al., 2017; Sottile et al., 2018). Also, as showed in Supplementary Table 2, a small proportion of the discriminant SNPs identified in current study, are located near the coding genes (56 SNPs out of all discriminant SNPs). It means that the majority of the SNPs are in non-coding/ intergenic regions of the sheep genome which is ideal for identification and assignment purpose since these regions/SNPs should be less influenced by natural or artificial selection (Allen et al., 2010). This result was in agreement with previously published manuscript (Sottile et al., 2018).

Conclusion

In this study, the genomic information of Iranian indigenous sheep breeds was used to develop a statistical method and therefore, to identify a minimal set of informative SNPs for assignment of breeds. At first, the results of genetic distance statistics showed that the different breeds were well differentiated and their genomic information could be used for determining of discriminant SNP markers. In order to have a general view of the population structure of the studied breeds, the principal components analysis (PCA) was performed by all SNP markers that passed the quality control steps. The results of this analysis showed that all sheep breeds differentiated from each other on the PC_1 and PC_2 . Next, to identify the SNP markers with maximum information for assigning animals to their true populations, discriminant analysis functions were applied. The results showed that using the total number of 201 SNP markers from the total markers available on the Ovine 50k SNP Chip, we are able to completely separate all animals with 100% accuracy, that about 40% of the selected SNP markers were located on the first three chromosomes. The DAPC analysis using 201 independent discriminant SNP markers shows that Iranian sheep breeds are clustered in at least four different groups, which are distinguished from each other in terms of genetic distances. Finally, to verify the effectiveness of the selected SNPs, these markers were used for assigning the samples of an independent Iranian sheep breed namely Baluchi, and the results showed that we can distinguish and assign all new samples to their true breed of origin by only using these 201 markers. In general, the statistical methods developed in this study and SNP markers identified can provide valuable information to facilitate the design of accurate and low-cost SNP assay for assigning unknown animals to their true population of origins in indigenous sheep breeds.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgment

The authors gratefully acknowledge the International Sheep Genomics Consortium for accessing to the Ovine HapMap genotypes and also Animal Breeding Center of Iran (ABCI) for accessing to the records and animals of the Iranian breeds. Thanks to the staffs from university of Tehran and AgResearch, Dr. Nejati-Javaremi, Dr. Moradi-Shahrbabak, Dr. Gholizadeh and Rayna Anderson who helped and supported this research. The authors also acknowledge the financial contributions of Animal Science Research Institute of Iran, Mobarakandish Institute and AgResearch, New Zealand.

Supplementary Table 1. The list of discriminant SNP markers identified in this study

Coordinate	Chr	SNP Names	Coordinate	Chr	SNP Names	Coordinate	Chr	SNP Names
1	2	3	4	5	6	7	8	9
36159676	11	s16859.1	123435789	3	OAR3_131536511.1	6480529	1	s23059.1
44753284	11	OAR11_47584423.1	125249527	3	s40131.1	17038771	1	s23433.1
55384890	11	OAR11_59401735.1	129245465	3	OAR3_137891445.1	38633519	1	OAR1_39850200.1
2314979	12	s52156.1	131663792	3	OAR3_140632356.1	39280480	1	OAR1_40517497.1
44138385	12	s64890.1	136398703	3	OAR3_136398703.1	58695038	1	OAR1_61931829.1
56965004	12	OAR12_63375554.1	137101201	3	s61782.1	58985824	1	OAR1_62185766.1
15477710	13	OAR13_13130994.1	152971492	3	OAR3_163532819.1	68996791	1	OAR1_73583793.1
47855248	13	OAR13_51134025_X.1	169417517	3	OAR3_182027991.1	69004560	1	OAR1_73591836.1
48338790	13	s14009.1	172893394	3	OAR3_185589716_X.1	69327178	1	OAR1_73950077.1
54219758	13	OAR13_58977999.1	179832455	3	OAR3_193567675.1	69799470	1	OAR1_74421103.1
56327133	13	OAR13_61210720.1	180817708	3	OAR3_194730311.1	91794880	1	OAR1_97455754.1
56349805	13	OAR13_61232811.1	185881136	3	OAR3_200125923.1	99667634	1	OAR1_106946819.1
56393306	13	s13640.1	192327941	3	OAR3_206996427.1	103763637	1	s32825.1
56428019	13	s14721.1	193399190	3	OAR3_208088102.1	109210933	1	OAR1_117978374.1
72821453	13	OAR13_78419787.1	194576286	3	OAR3_209315205_X.1	111609419	1	OAR1_120550895.1
4382354	14	OAR14_4313360.1	200083592	3	OAR3_215267360.1	119491972	1	s34722.1
15700496	14	OAR14_15968361.1	208191238	3	OAR3_225860485.1	127175029	1	OAR1_137791545.1
32058213	14	OAR14_33354501.1	209590314	3	OAR3_227242428.1	133060150	1	OAR1_143967121.1
34733412	14	s57849.1	213524510	3	s56918.1	153606843	1	OAR1_165359755.1
61008970	14	s52668.1	214478744	3	OAR3_232475527.1	160449086	1	s37063.1
2008093	16	s21336.1	218125347	3	s41886.1	164709786	1	OAR1_177674843.1

26612524	16	OAR16_28856625.1	219583554	3	OAR3_238043094.1	170387319	1	OAR1_183630759.1
33545078	16	OAR16_36452841.1	17038787	4	OAR4_17299398.1	177180380	1	s48584.1
35781385	16	OAR16_38988953.1	19607622	4	OAR4_20257612.1	195830343	1	s19660.1
42939451	16	OAR16_46667869.1	48669962	4	OAR4_51489408.1	215511525	1	s51892.1
56577830	16	s69802.1	68754789	4	OAR4_72720594.1	235877668	1	OAR1_254806541.1
25018550	17	OAR17_27639885.1	68802676	4	s25289.1	245712422	1	OAR1_265350597.1
34121657	17	OAR17_37138062.1	96279291	4	OAR4_102060610.1	256016538	1	OAR1_276613221.1
52358376	17	OAR17_56996543.1	97533598	4	OAR4_103580779.1	259556965	1	OAR1_280813799.1
52959387	17	OAR17_57611603.1	100255840	4	OAR4_106683143.1	260272620	1	s11035.1
67956856	17	OAR17_74103903.1	101410245	4	OAR4_107881931.1	270698357	1	OAR1_293214958.1
1071516	18	s30628.1	111517280	4	OAR4_119012773.1	275427922	1	s63549.1
12985346	18	s61259.1	111789721	4	OAR4_119285102.1	1038580	2	s18389.1
13299667	18	OAR18_13189891.1	14686309	5	s57684.1	11979482	2	OAR2_11121541.1
19342316	18	s31152.1	43128620	5	s55322.1	15141339	2	OAR2_14584156.1
19399095	18	s62564.1	85998123	5	OAR5_94211275.1	24092677	2	s55788.1
38054675	18	s45597.1	21063724	6	OAR6_24121723_X.1	39387199	2	s12814.1
65733758	18	s25545.1	22362351	6	OAR6_25605499.1	45552152	2	OAR2_48363557.1
3943735	19	OAR19_4188125.1	52034632	6	OAR6_57405903.1	52409379	2	s20468.1
15501419	19	s51457.1	94639943	6	s21552.1	58497237	2	OAR2_62874966.1
21167520	19	OAR19_22249944.1	115207925	6	OAR6_116658909.1	69145328	2	OAR2_73858846.1
37596202	19	OAR19_39442752.1	27924610	7	OAR7_31647698.1	84530797	2	OAR2_89841590.1
17460319	20	s30056.1	32371875	7	OAR7_36539881.1	89239384	2	OAR2_95733865.1
40240144	20	OAR20_43863058.1	41915624	7	OAR7_46498332.1	118180362	2	OAR2_126354465.1
16621210	21	s47135.1	52548051	7	OAR7_58068949.1	164516447	2	OAR2_174148415.1

Supplementary Table 1 – contd.

1	2	3	4	5	6	7	8	9
40240144	20	OAR20_43863058.1	41915624	7	OAR7_46498332.1	118180362	2	OAR2_126354465.1
16621210	21	s47135.1	52548051	7	OAR7_58068949.1	164516447	2	OAR2_174148415.1
28853040	21	OAR21_32292765_X.1	60226641	7	OAR7_66119984.1	165486783	2	OAR2_175122529.1
6065909	22	OAR22_7329478.1	76991147	7	s52907.1	165667917	2	s37054.1
9162056	22	OAR22_10936495.1	6153266	8	OAR8_6430836.1	178922457	2	OAR2_189750050.1
9181371	22	OAR22_10973184.1	6926523	8	s07254.1	182607165	2	OAR2_193444466_X.1
28813203	22	OAR22_33115630.1	12803219	8	OAR8_14347534.1	182923305	2	s05571.1
59686993	23	OAR23_63703760.1	13172803	8	OAR8_14718655.1	186193337	2	s56826.1
1303692	24	s09648.1	21344959	8	OAR8_23796639.1	188109777	2	OAR2_199435945.1
1377045	24	s32804.1	52246161	8	OAR8_56166924.1	198629088	2	OAR2_210295546.1
28785295	24	s74901.1	54280168	8	OAR8_58181648_X.1	203129937	2	OAR2_214959722.1
34689602	25	s40440.1	61031105	8	OAR8_65770103.1	209015882	2	s20634.1
4730650	26	OAR26_5333971.1	83916893	8	s52715.1	216769344	2	s08262.1
13909530	X	OARX_25160755.1	39044229	9	OAR9_40990389.1	229040784	2	s18373.1
50852578	X	OARX_6082722_X.1	45161740	9	OAR9_47384888.1	240756928	2	s64586.1
54296800	X	s08442.1	3881451	10	OAR10_1702026.1	17975672	3	s03702.1
55909305	X	OARX_59327581.1	11920339	10	s52535.1	22480041	3	OAR3_24232824.1
55937193	X	OARX_59383635.1	22592412	10	OAR10_22217205.1	105585637	3	OAR3_112264725.1
56153635	X	OARX_59571364.1	30618253	10	s68983.1	109474803	3	OAR3_116664764.1
56165053	X	OARX_59578440.1	31067620	10	s06792.1	110196723	3	OAR3_117378944.1
56538556	X	OARX_59912586.1	45404298	10	OAR10_45961582.1	110431071	3	OAR3_117626019.1
58160209	X	s41440.1	78238799	10	s65075.1	117406030	3	s31620.1
58161365	X	s15970	26571629	11	s56248.1	118482516	3	s23959.1
58192278	X	s61445.1	31701852	11	s64527.1	120312152	3	OAR3_128119548.1

Supplementary Table 2. The list of the discriminant SNPs consisting of coding genes in this study

Gene Type*	Gene names	Coordinate	Chr	SNP Names
1	2	3	4	5
Protein coding	<i>C1orf50, TMEM269</i>	17038771	1	s23433.1
Protein coding	<i>PGM1</i>	38633519	1	OAR1_39850200.1
Protein coding	<i>FBNP1L</i>	69799470	1	OAR1_74421103.1
Protein coding	<i>CKS1B, FLAD1</i>	103763637	1	s32825.1
Protein coding	<i>BACE2</i>	259556965	1	OAR1_280813799.1
Protein coding	<i>BTB</i>	270698357	1	OAR1_293214958.1
Protein coding	<i>KAT2B</i>	275427922	1	s63549.1
Protein coding	<i>ECPAS</i>	11979482	2	OAR2_11121541.1
snRNA	<i>U6</i>	15141339	2	OAR2_14584156.1
Protein coding	<i>EBF2</i>	39387199	2	s12814.1
Protein coding	<i>FAM221B, HINT2</i>	52409379	2	s20468.1
Protein coding	<i>GNAQ</i>	58497237	2	OAR2_62874966.1
snoRNA	<i>SNORA70</i>	198629088	2	OAR2_210295546.1
Protein coding	<i>CCL20</i>	229040784	2	s18373.1
rRNA	<i>5S_rRNA</i>	118482516	3	s23959.1
Protein coding	<i>FMNL3</i>	136398703	3	OAR3_136398703.1
Protein coding	<i>FKBP11</i>	137101201	3	s61782.1
Protein coding	<i>GYS2</i>	193399190	3	OAR3_208088102.1
snRNA	<i>U6</i>	194576286	3	OAR3_209315205_X.1
Protein coding	<i>GUCY2C</i>	200083592	3	OAR3_215267360.1
Protein coding	<i>VWF</i>	208191238	3	OAR3_225860485.1
Protein coding	<i>TRIOBP</i>	213524510	3	s56918.1
Protein coding	<i>BCAP29</i>	48669962	4	OAR4_51489408.1
lincRNA	<i>Y_RNA</i>	111517280	4	OAR4_119012773.1
Protein coding	<i>ANGPTL4, RAB11B</i>	14686309	5	s57684.1
Protein coding	<i>TCF7</i>	43128620	5	s55322.1
Protein coding	<i>NOPI4</i>	115207925	6	OAR6_116658909.1
Protein coding	<i>THBS1</i>	32371875	7	OAR7_36539881.1
Protein coding	<i>RTRAF</i>	41915624	7	OAR7_46498332.1
Protein coding	<i>TTK</i>	6926523	8	s07254.1
Protein coding	<i>RAB2A</i>	39044229	9	OAR9_40990389.1
Protein coding	<i>PREX2</i>	45161740	9	OAR9_47384888.1
Protein coding	<i>NAA16</i>	11920339	10	s52535.1
Protein coding	<i>UBL3</i>	31067620	10	s06792.1
Protein coding	<i>TPP2</i>	78238799	10	s65075.1
Protein coding	<i>ACADVL</i>	26571629	11	s56248.1
Protein coding	<i>COX10</i>	31701852	11	s64527.1

Supplementary Table 2 – contd.

1	2	3	4	5
Protein coding	<i>ATP5PD</i>	55384890	11	OAR11_59401735.1
Protein coding	<i>TMCC2</i>	2314979	12	s52156.1
snRNA	<i>U6</i>	15477710	13	OAR13_13130994.1
Protein coding	<i>LAMA5</i>	54219758	13	OAR13_58977999.1
Protein coding	<i>EDC4, PSKH1</i>	34733412	14	s57849.1
Protein coding	<i>DOCK2</i>	2008093	16	s21336.1
Protein coding	<i>RETREG1, ZNF622</i>	56577830	16	s69802.1
Protein coding	<i>HCAR1</i>	52358376	17	OAR17_56996543.1
Protein coding	<i>KREMEN1</i>	67956856	17	OAR17_74103903.1
snoRNA	<i>SNORD116</i>	1071516	18	s30628.1
Protein coding	<i>WDR20</i>	65733758	18	s25545.1
Protein coding	<i>RBMS3</i>	3943735	19	OAR19_4188125.1
Protein coding, snRNA	<i>ABHD5, U6</i>	15501419	19	s51457.1
Protein coding	<i>TENM4</i>	16621210	21	s47135.1
Protein coding	<i>FAM118B, FOXRED1</i>	28853040	21	OAR21_32292765_X.1
Protein coding	<i>PAPSS2</i>	9181371	22	OAR22_10973184.1
Protein coding	<i>JPT2, MAPK8IP3</i>	1303692	24	s09648.1
Protein coding	<i>CALN1</i>	28785295	24	s74901.1
Protein coding	<i>STARD8</i>	58160209	X	s41440.1

*snRNA, snoRNA, rRNA and lincRNA are standing for small nuclear RNA, small nucleolar RNA, ribosomal RNA and long non coding RNA respectively.

References

- Allen A.R., Taylor M., McKeown B., Curry A.I., Lavery J.F., Mitchell A., Harts-horne D., Fries R., Skuce R.A. (2010). Compilation of a panel of informative single nucleotide polymorphisms for bovine identification in the northern Irish cattle population. *BMC Genet.*, 11: 1–8.
- Baba K., Shibata R., Sibuya M. (2004). Partial correlation and conditional correlation as measures of conditional independence. *Aust. New Zeal. J. Stat.*, 46: 657–664.
- Bertolini F., Galimberti G., Schiavo G., Mastrangelo S., Di Gerlando R., Stril-lacci M.G., Bagnato A., Portolano B., Fontanesi L. (2018). Preselection statistics and Random Forest classification identify population informative single nucleotide polymorphisms in cosmopolitan and autochthonous cattle breeds. *Animal*, 12: 12–19.
- Bowcock A.M., Ruiz-Linares A., Tomfohrde J., Minch E., Kidd J.R., Cavalli-Sforza L.L. (1994). High resolution of human evolutionary trees with polymorphic microsatel-lites. *Nature*, 368: 455–457.
- Carignano H.A., Roldan D.L., Beribe M.J., Raschia M.A., Amadio A., Nani J.P., Gutierrez G., Alvarez I., Trono K., Poli M.A., Miretti M.M. (2018). Genome-wide scan for commons SNPs affecting bovine leukemia virus infection level in dairy cattle. *BMC Ge-nomics*, 19: 142.
- Choi J.W., Liao X., Stothard P., Chung W.H., Jeon H.J., Miller S.P., Choi S.Y., Lee J.K., Yang B., Lee K.T., Han K.J., Kim H.C., Jeong D., Oh J.D., Kim N., Kim

- T.H., Lee H.K., Lee S.J. (2014). Whole-genome analyses of Korean native and Holstein cattle breeds by massively parallel sequencing. *PLoS One*, 9: e101127.
- Ciani E., Crepaldi P., Nicoloso L., Lasagna E., Sarti F.M., Moioli B., Napolitano F., Carta A., Usai G., D'Andrea M., Marletta D., Ciampolini R., Riggio V., Occidente M., Matassino D., Kompan D., Modesto P., Macciotta N., Ajmone-Marsan P., Pilla F. (2014). Genome-wide analysis of Italian sheep diversity reveals a strong geographic pattern and cryptic relationships between breeds. *Anim. Genet.*, 45: 256–266.
- Clarke S.M., Henry H.M., Dodds K.G., Jowett T.W., Manley T.R., Anderson R.M., McEwan J.C. (2014). A high throughput single nucleotide polymorphism multiplex assay for parentage assignment in New Zealand sheep. *PLoS One*, 9: e93392.
- Dalvit C., De Marchi M., Targhetta C., Gervaso M., Cassandro M. (2008). Genetic traceability of meat using microsatellite markers. *Food Res. Int.*, 41: 301–307.
- 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. *Anim. Genet.*, 44: 377–382.
- Dimauro C., Nicoloso L., Cellesi M., Macciotta N.P.P., Ciani E., Moioli B., Pilla F., Crepaldi P. (2015). Selection of discriminant SNP markers for breed and geographic assignment of Italian sheep. *Small Rumin. Res.*, 128: 27–33.
- Dodds K.G., Auvray B., Newman S.A.N., McEwan J.C. (2014). Genomic breed prediction in New Zealand sheep. *BMC Genet.*, 15: 92–101.
- Gholizadeh M., Rahimi-Mianji G., Nejati-Javaremi A., De Koning D.J., Jonas E. (2014). Genomewide association study to detect QTL for twinning rate in Baluchi sheep. *J. Genet.*, 93: 489–493.
- Hastie T., Tibshirani R., Friedman J. (2009). The elements of statistical learning the elements of statistical learning data mining, inference, and prediction. Springer Press, 2nd ed., 545 pp.
- Heaton M.P., Kalbfleisch T.S., Basnayake V., Clarke S.M., Kijas J.W., Leymaster K.A., Chitko-McKown C.G., Simpson B., McEwan J., Petrik D.T., Maddox J.F., Smith T.P.L. (2014). SNPs for parentage testing and traceability in globally diverse breeds of sheep. *PLoS One*, 9: e94851.
- Hubbard T.J.P., Aken B.L., Ayling S., Ballester B., Beal K., Bragin E., Brent S., Chen Y., Clapham P., Clarke L., Coates G., Fairley S., Fitzgerald S., Fernandez-Banet J., Gordon L., Graf S., Haider S., Hammond M., Holland R., Howe K., Jenkinson A., Johnson N., Kahari A., Keefe D., Keenan S., Kinsella R., Kokocinski F., Kulesha E., Lawton D., Longden I., Megy K., Meidl P., Overduin B., Parker A., Pritchard B., Rios D. (2009). Ensemble BioMart: Ensemble online genome data base BioMart tool. *Nucleic Acids Res.*, 37: 690–697.
- 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. *Livest. Sci.*, 223: 60–67.
- Jolliffe I.T., Cadima J. (2016). Principal component analysis: A review and recent developments. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, 13: 374–381.
- Jombart T. (2008). Adegnet: A R package for the multivariate analysis of genetic markers. *Bioinformatics*, 24: 1403–1405.
- Jombart T., Devillard S., Balloux F. (2010). Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genet.*, 11: 94.
- Kijas J.W., Townley D., Dalrymple B.P., Heaton M.P., Maddox J.F., McGrath A., Wilson P., Ingersoll R.G., McCulloch R., McWilliam S., Tang D., McEwan J., Cockett N., Oddy V.H., Nicholas F.W., Raadsma H., Barris W., Bishop S.C., Coltman D., Crawford A., Eggen A., Erhardt G., Forage R., Hanotte O., Hunt P., Jianlin H., Li K., Marsan P.A., Miller J.E., Pemberton J., Schibler L. (2009). A genome wide survey of SNP variation reveals the genetic structure of sheep breeds. *PLoS One*, 11: 94.
- Kijas J.W., Lenstra J.A., Hayes B., Boitard S., Neto L.R., Cristobal M.S., Servin B., McCulloch R., Whan V., Gietzen K., Paiva S., Barendse W., Ciani E., Raadsma H., McEwan J., Dalrymple B. (2012). Genome-wide analysis of the world's

- sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biol.*, 10: e1001258.
- Korthauer K., Kimes P.K., Duvallet C., Reyes A., Subramanian A., Teng M., Shukla C., Alm E.J., Hicks S.C. (2019). A practical guide to methods controlling false discoveries in computational biology. *Genome Biol.*, 20: 118.
- Lachance J., Tishkoff S.A. (2013). SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. *BioEssays*, 35: 780–786.
- Manel S., Gaggiotti O.E., Waples R.S. (2005). Assignment methods: Matching biological questions with appropriate techniques. *Trends Ecol. Evol.*, 20: 3.
- Maudet C., Luikart G., Taberlet P. (2002). Genetic diversity and assignment tests among seven French cattle breeds based on microsatellite DNA analysis. *J. Anim. Sci.*, 80: 942–950.
- Meadows J.R.S., Li K., Kantanen J., Tapio M., Sipos W., Pardeshi V., Gupta V., Calvo J.H., Whan V., Norris B., Kijas J.W. (2005). Mitochondrial sequence reveals high levels of gene flow between breeds of domestic sheep from Asia and Europe. *J. Hered.*, 96: 494–501.
- Moradi M.H., Nejati-Javaremi A., Moradi-Shahrabak M., Dodds K.G., McEwan J.C. (2012). Genomic scan of selective sweeps in thin and fat tail sheep breeds for identifying of candidate regions associated with fat deposition. *BMC Genet.*, 13: 10.
- Moradi M.H., Phua S.H., Hedayat N., Khodaei-Motlagh M., Razmkabir M. (2017). Haplotype and genetic diversity of mtDNA in indigenous Iranian sheep and an insight into the history of sheep domestication. *J. Agric. Sci. Technol.*, 19: 591–601.
- Muñoz I., Henriques D., Johnston J.S., Chávez-Galarza J., Kryger P., Pinto M.A. (2015). Reduced SNP panels for genetic identification and introgression analysis in the dark honey bee (*Apis mellifera mellifera*). *PLoS One*, 10: e0124365.
- Negrini R., Milanesi E., Colli L., Pellecchia M., Nicoloso L., Crepaldi P., Lenstra J.A., Ajmone-Marsan P. (2007). Breed assignment of Italian cattle using biallelic AFLP® markers. *Anim. Genet.*, 38: 147–153.
- Negrini R., Nicoloso L., Crepaldi P., Milanesi E., Colli L., Chegdani F., Pariset L., Dunner S., Leveziel H., Williams J.L., Ajmone Marsan P. (2009). Assessing SNP markers for assigning individuals to cattle populations. *Anim. Genet.*, 40: 18–26.
- Ni G., Caverio D., Fangmann A., Erbe M., Simianet H. (2017). Whole-genome sequence-based genomic prediction in laying chickens with different genomic relationship matrices to account for genetic architecture. *Genet Sel Evol.*, 49: 8.
- Panetto J.C. do C., Machado M.A., da Silva M.V.G.B., Barbosa R.S., dos Santos G.G., Leite R. de M.H., Peixoto M.G.C.D. (2017). Parentage assignment using SNP markers, inbreeding and population size for the Brazilian Red Sindhi cattle. *Livest. Sci.*, 204: 33–38.
- Parker H.G., Ostrander E.A. (2005). Canine genomics and genetics: running with the pack. *PLoS Genet.*, 1: 507–513.
- 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 Genet.*, 3: e160.
- Plassais J., Kim J., Davis B.W., Karyadi B.W., Hogan A.N., Harris A.N., Decker B., Parker H.G., Ostrander E.A. (2019). Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. *Nat. Commun.*, 10: 1489.
- Pritchard J.K., Stephens M., Donnelly P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155: 945–959.
- Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A., Bender D., Maller J., Sklar P., De Bakker P.I., Daly M.J., Sham P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, 81: 559–575.
- Ramos A.M., Megens H.J., Crooijmans R.P.M.A., Schook L.B., Groenen M.A.M. (2011). Identification of high utility SNPs for population assignment and traceability purposes in the pig using high-throughput sequencing. *Anim. Genet.*, 42: 613–620.
- Reynolds J., Weir B.S., Cockerham C.C. (1983). Estimation of the coancestry coefficient: Basis for a short-term genetic distance. *Genetics*, 105: 767–779.
- Schwäglele F. (2005). Traceability from a European perspective. *Meat Sci.*, 71: 164–173.

- Sottile G., Sardina M.T., Mastrangelo S., Di Gerlando R., Tolone M., Chiodi M., Portolano B. (2018). Penalized classification for optimal statistical selection of markers from high-throughput genotyping: application in sheep breeds. *Animal*, 12: 1118–1125.
- Talenti A., Nicolazzi E.L., Chessa S., Frattini S., Moretti R., Coizet B., Nicolo L., Colli L., Pagnacco G., Stella A., Ajmone-Marsan P., Ptak G., Crepaldi P. (2016). A method for single nucleotide polymorphism selection for parentage assessment in goats. *J. Dairy Sci.*, 99: 1–8.
- Tortereau F., Moreno C.R., Tosser-Klopp G., Servin B., Raoul J. (2017). Development of a SNP panel dedicated to parentage assignment in French sheep populations. *BMC Genet.*, 18: 50.
- Vatankhah M., Moradi-Shahrabak M., Nejati-Javaremi A., Mireaei Ashtiani S., Vaez Torshizi R. (2004). A review of sheep breeding in Iran. *Proc. First Congress on Animal and Aquatic Sciences*, University of Tehran, Iran.
- Weihs C., Ligges U., Luebke K., Raabe N. (2005). klaR: analyzing German business cycles. In: *Data analysis and decision support*, Baier D., Decker R., Schmidt-Thieme L. (eds.). Springer-Verlag, pp. 335–343.
- Weir B.S., Cockerham C.C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, 38: 1358–1370.
- Weller J.I., Seroussi E., Ron M. (2006). Estimation of the number of genetic markers required for individual animal identification accounting for genotyping errors. *Anim. Genet.*, 37: 387–389.
- Wilkinson S., Wiener P., Archibald A.L., Law A., Schnabel R.D., McKay S.D., Taylor J.F., Ogden R. (2011). Evaluation of approaches for identifying population informative markers from high density SNP Chips. *BMC Genet.*, 12: 45.
- Willing E.M., Dreyer C., van Oosterhout C. (2012). Estimates of genetic differentiation measured by F_{st} do not necessarily require large sample sizes when using many SNP markers. *PLoS One*, 7: e42649.
- Wright S. (1950). The genetical structure of populations. *Ann. Eugen.*, 15: 323.
- Zeder M.A. (2008). Animal domestication in the Zagros: A review of past and current research. *Paléorient*, 25: 11–25.
- Zeder M.A. (2012). Pathways to animal domestication. In: *Biodiversity in agriculture: domestication, evolution, and sustainability*, 1: 227–259.

Received: 28 II 2020

Accepted: 9 IX 2020