Animal (2018), **12:6**, pp 1118–1125 © The Animal Consortium 2017
doi:10.1017/S175173111700266X

animal

Penalized classification for optimal statistical selection of markers from high-throughput genotyping: application in sheep breeds

G. Sottile[1†], M. T. Sardina[2], S. Mastrangelo[2], R. Di Gerlando[2], M. Tolone[2], M. Chiodi[1] and B. Portolano[2]

[1]*Dipartimento Scienze Economiche, Aziendali e Statistiche, University of Palermo, Palermo, Italy;* [2]*Dipartimento Scienze Agrarie, Alimentari e Forestali, University of Palermo, Palermo, Italy*

(Received 29 March 2017; Accepted 15 September 2017; First published online 24 October 2017)

The identification of individuals' breed of origin has several practical applications in livestock and is useful in different biological contexts such as conservation genetics, breeding and authentication of animal products. In this paper, penalized multinomial regression was applied to identify the minimum number of single nucleotide polymorphisms (SNPs) from high-throughput genotyping data for individual assignment to dairy sheep breeds reared in Sicily. The combined use of penalized multinomial regression and stability selection reduced the number of SNPs required to 48. A final validation step on an independent population was carried out obtaining 100% correctly classified individuals. The results using independent analysis, such as admixture, F_{st}*, principal component analysis and random forest, confirmed the ability of these methods in selecting distinctive markers. The identified SNPs may constitute a starting point for the development of a SNP based identification test as a tool for breed assignment and traceability of animal products.*

Keywords: penalized multinomial regression, stability selection, sheep breeds, livestock genetic resources, single nucleotide polymorphism markers

Implications

The identification of individuals' breed/population of origin offers unbiased tools in livestock and is useful in different biological contexts, such as management of livestock genetic resources for breed confirmation and estimation of hybridization level, and authentication of typical products. The identified single nucleotide polymorphisms (SNPs) may constitute a starting point for the development of a SNP based identification test as a tool for breed assignment and traceability of animal products.

Introduction

Assignment tests using genetic information to establish population membership of individuals provide the most direct methods to determine population of origin of unknown individuals (Negrini *et al.*, 2009). The identification of individuals' breed/population of origin offers has several practical applications in livestock and is useful in different biological contexts, such as management of livestock genetic resources for breed confirmation, estimation of hybridization level and authentication of brand products that are produced using only a few particular breeds or populations (Wilkinson *et al.*, 2011; Bertolini *et al.*, 2015). Moreover, assignment of individuals to a specific breed is very important both for biodiversity purposes and products traceability, especially when the phenotypic differentiation among breeds is difficult (Tolone *et al.*, 2012).

Recently developed genomic technologies, such as medium and high-density SNP arrays, are important tools that can be used for these purposes. Dense genome-wide data is valuable but is relatively costly and time-consuming or computationally expensive to analyse. However, some methods are tractable and capable to efficiently predict breed composition using breed frequencies of thousands of markers (Kuehn *et al.*, 2011). Therefore, it is often desirable to reduce the number of markers according to their information content, in order to create reduced panels for population genetic analysis (Paschou *et al.*, 2007). Many clustering algorithms have been developed employing population genetic data to assign individuals to clusters (Jakobsson and Rosenberg, 2007). Several statistical methods were used to determine which genetic markers contain the most information to discriminate among populations (Rosenberg, 2005; Wilkinson *et al.*, 2011), such as the combined approach of principal component analysis (PCA)

† E-mail: gianluca.sottile@unipa.it

and random forest (RF) (Bertolini *et al.*, 2015), multivariate canonical discriminant analysis (Dimauro *et al.*, 2013), the statistic delta (Shriver *et al.*, 1997), and Wright's F_{st} (Bowcock *et al.*, 1994). While all these methodologies yielded reduced marker panels useful for breed identification, the power of assignment varied among analysis methods.

In the present study, starting from available high-throughput SNP data, penalized multinomial regression (PMR) and stability selection (SS) were applied to identify the optimal set of informative SNPs useful to discriminate among five Sicilian dairy sheep breeds.

Material and methods

Data

A total of 236 animals, randomly collected from several farms in different areas of Sicily, were used for the analysis. Samples consisted of 30 Barbaresca (Bar), 51 Comisana (Com), 77 Pinzirita (Pin), 30 Sarda (Sar) and 48 Valle del Belice (VdB) individuals. The procedures involving animal samples collection followed the recommendation of Directive 2010/63/EU. All animals were genotyped for 54 241 SNPs using the Illumina OvineSNP50K Genotyping BeadChip. Genotyping was performed by Dipartimento Scienze Agrarie e Forestali, University of Palermo. Input data were genotyping data of 54 241 SNPs, that is GType data in Illumina AB format exported from GenomeStudio v1.0 (Illumina Inc., San Diego, CA). We excluded all SNPs not assigned to chromosome (OAR) or assigned to X and Y chromosomes. Markers were filtered according to the following quality criteria: (i) call frequency ($\geqslant$95%), (ii) minor allele frequency (MAF $\geqslant$ 0.01). SNPs that did not satisfy these quality criteria were excluded. A total of 48 068 SNPs were retained for subsequent analyses.

We transformed the genotyping data to numeric values, without any loss of information, in order to apply PMR. The initial data table $\boldsymbol{X}$ consisted of N rows, one per animal, and p columns, one per SNP. Each entry of $\boldsymbol{X}$, AA, AB and BB, was scored as −1, 0, 1, or empty. SNPs with missing genotypes were randomly imputed within each breed according to the corresponding genotype frequency.

Statistical analysis of single nucleotide polymorphisms and variable reduction

Each sheep breed was divided into a test population and a validation population. The validation population, generated by randomly sampling 15% of animals within each breed, was used for the final validation procedure of breed assignment. The test population consisted of the remaining animals.

Suppose to have a set of N individuals and p SNPs, with $p \gg N$, divided in K groups, the main goal in a high-dimensional setting was the selection of a limited number of SNPs with a high discrimination power among groups. To achieve this aim some authors proposed the LASSO method (Tibshirani, 1996) or L_1-penalty, a shrinkage and selection method for linear regression. In statistics, least absolute shrinkage and selection operator (LASSO) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. LASSO is able to achieve these goals by forcing the sum of the absolute value of the regression coefficients to be less than a fixed value, which forces certain coefficients to be set to zero. LASSO was originally introduced in the context of least squares. Consider a sample consisting of N observations, each of which consists of p covariates and a single outcome. Let y_i be the outcome and $\boldsymbol{x}_i = (x_1, x_2, \ldots, x_p)^T$ be the covariate vector for the ith observation. Than the objective of the LASSO is to solve

$$\min_{\beta_0, \boldsymbol{\beta}} \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta_0 - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 \right\} \text{subject to} \sum_{j=1}^{p} |\beta_j| \leq t,$$

where t is a prespecified free parameter that determines the amount of regularization, β_0 is the intercept of the model and $\boldsymbol{\beta}$ is the p-variate vector of regression coefficients. Letting $\boldsymbol{X}$ be the covariate matrix, so that $\boldsymbol{X}_{ij} = (\boldsymbol{x}_i)_j$ and $\boldsymbol{x}_i^T$ is the ith row of $\boldsymbol{X}$ we can write this in the so-called Lagrangian form

$$\min_{\beta \in \mathbb{R}^{p+1}} \left\{ \frac{1}{N} ||y - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_1 \right\}$$

where the relationship between t and λ is $t \approx \lambda^{-1}$, that is t is approximately the multiplicative inverse of λ. The L_1 norm of $\boldsymbol{\beta}$ is a constraint on the regression coefficients that strictly depends on the tuning parameter λ.

LASSO regression originally defined for least squares, is easily extended to a wide variety of statistical models including generalized linear models. In our framework, given the nature of the outcome y, that is a categorical variable with $K > 2$ levels, we used a penalized multinomial regression. Here, we model the probability to belong to breed k given the SNPs' matrix $\boldsymbol{X}$ of dimension $N \times p$

$$\Pr(y = k \mid \boldsymbol{X}) = \frac{e^{\boldsymbol{X}\beta_k}}{\sum_{l=1}^{K} e^{\boldsymbol{X}\beta_l}}, \; k = 1, \ldots, K.$$

Let the outcome $\boldsymbol{y}$ be the $N \times K$ indicator response matrix, with elements $y_{il} = I(y_i = l), l = 1, \ldots, K$ and $i = 1, \ldots, N$. Then the regression coefficients are obtained as the solution of the following optimization problem

$$\max_{\boldsymbol{\beta} \in \mathbb{R}^{k(p+1)}} \left\{ -\frac{1}{N} \sum_{i=1}^{N} \left(\sum_{l=1}^{K} y_{il} \boldsymbol{X}\boldsymbol{\beta}_l - \log \left(\sum_{l=1}^{K} e^{\boldsymbol{X}\beta_l} \right) \right) + \lambda \sum_{j=1}^{p} \sum_{l=1}^{K} |\beta_{jl}| \right\}$$

where $\boldsymbol{\beta}$ is a $p \times K$ matrix coefficients, $\boldsymbol{\beta}_k$ refers to the k-th column (for outcome breed k), and $\boldsymbol{\beta}_j$ the j-th row (vector of K coefficients for variable j).

In this step, as usual for the supervised classification approach y, that is the true vector of labels of breeds, is set as response variable.

All the following analysis was computed using *glmnet* package (Friedman *et al.*, 2010) of the R software 3.3.0 (R Core Team, 2016).

The stability selection method (Meinshausen and Bühlmann, 2010) was used to discover the best subset of variables S^{stable} that will have nonzero weight in the model. Let's assume that we have a generic structure estimation algorithm (i.e. the LASSO) that takes a data set $\boldsymbol{X}$ and a regularization parameter λ, it returns a selection set S^{λ}. The *j*-th covariate belongs to S^{λ} if the regression coefficient $\boldsymbol{\beta}_j \neq 0$.

The SS algorithm then runs as follows:

1. Define a candidate set of dimension *m* of regularization parameters $\Lambda=\{\lambda_1, \ldots, \lambda_m\}$ and a number *n* of subsample.
2. For each value of $\lambda \in \Lambda$, do:

   a. Start with the full data set $\boldsymbol{X}$
   b. For each $i=1, \ldots, n$ do:

      i. Subsample from $\boldsymbol{X}$ without replacement to generate a smaller data set of size *N*/2, namely $\boldsymbol{Z}_i$.
      ii. Run the selection algorithm on dataset $\boldsymbol{Z}_i$ with parameter λ to obtain a selection set S_i^{λ}.

   c. Given the *n* selection sets from each subsample, calculate the empirical selection probability for each covariate:

$$\Pi_{\mathrm{j}}^{\lambda} = P\left\{j \in S_i^{\lambda}\right\} = \frac{1}{n}\sum_{i=1}^{n} I\left\{j \in S_i^{\lambda}\right\}, j = 1, \ldots, p$$

   The selection probability for covariate *j* is its probability of being selected by the algorithm.

3. Given the selection probabilities for each covariate and for each value of $\lambda \in \Lambda$, construct the stable set according to the following definition:

$$\hat{S}^{\text{stable}} = \left\{ j : \max_{\lambda \in \Lambda} \Pi_{\mathrm{j}}^{\lambda} \geq \pi_{thr} \right\}$$

   where π_{thr} is a predefined threshold.

In our study, fixing a sequence of 100 values of λ, step 2. (b) was repeated *B* times by randomly splitting, within each breed, the test population. After calculating the empirical selection probability for each SNP and fixing the threshold value, a final set S^{stable} of p_1 SNPs was selected. For this reduced panel, a new multinomial regression model was then fitted. To assess the classification performance of this set of p_1 SNPs we tested the discrimination rule using the validation population which is considered to be an independent subset of samples.

Other statistical and genetic methods

To better understand the potential of our strategy and the strength of a reduced panel of SNPs we decided to use the *k*-means approach, which is an unsupervised technique. In particular, we used all the principal components up to 70% of explained variance of the model matrix $\boldsymbol{X}$. We applied this technique once to the whole set of SNPs and once to the reduced set p_1.

The efficiency of the selected markers to cluster individuals was also tested using model-based clustering algorithm implemented in the admixture software 1.3.0 (Alexander and Lange, 2011) which used unsupervised classification approaches and Genepop software 4.1.4 (Rousset, 2008) to calculate F_{st}. The most probable number of populations in the data set (*K*) was estimated using the default (fivefold) admixture's cross-validation procedure, by which estimated prediction errors are obtained, for each *K* value, by adopting a kind of 'leave-one-out' approach through which an estimation of prediction errors can be assumed to be the most suitable one. Genepop was also used to estimate population relatedness using pairwise estimates of F_{st} among breeds. The reduced panel was analysed using SNPchiMp (Nicolazzi *et al.*, 2015) to obtain information on the genomic distribution of SNPs.

In order to compare our approach to those previously reported, another mixed strategy was considered (Bertolini *et al.*, 2015). In particular, PCA and RF was used to discover a new SNP panel able to discriminate among the breeds For each autosome, the top 20 SNPs were selected and merged together, leading to a final panel of 520 markers. Random forest based on the selected 520 SNPs were built on the test population. The mean decrease in the Gini index (MDGI) or the mean accuracy decrease (MAD) were used in order to select the most discriminant SNPs. Four different SNP panels were created selecting the first 48 and 96 SNPs from the MDGI and the first 48 and 96 from the MAD, respectively. This SNP panels' size was chosen considering the practical possibilities to develop multiplex SNP panels containing a reduced number of SNPs for field applications (Bertolini *et al.*, 2015). For each of the four reduced panels, a new RF was fitted and the corresponding out-of-bag (OOB) error rate was calculated. Classification performance of these four RFs was assessed also using the validation population.

A simulation study has been done to compare the performance of our proposed strategy and the PCA-RF strategy. A group of genetic variants has been randomly generated by using the real data set. In general, we sampled with replacement *n* observations of the real data set, so to maintain the same structure and association between SNPs. Moreover, we built $\tilde{\boldsymbol{X}}_{\text{test}}$ which is the simulated test population and $\tilde{\boldsymbol{X}}_{\text{val}}$ which is the simulated validation population (15% of the sample size). The response variable $\tilde{y}$ was the label vector of length *n*, indicating the membership of each animal to their own breed in the simulated data. $\tilde{\boldsymbol{X}}_{\text{test}}$ and $\tilde{\boldsymbol{X}}_{\text{val}}$ were used to evaluate the out-of-bag error and misclassification error rate for both strategies.

Results

Penalized multinomial regression and stability selection

Out of a total of 54 241 genotyped SNPs, 378 unmapped and 1450 were located on sex chromosomes. Thus, 52 413 SNPs

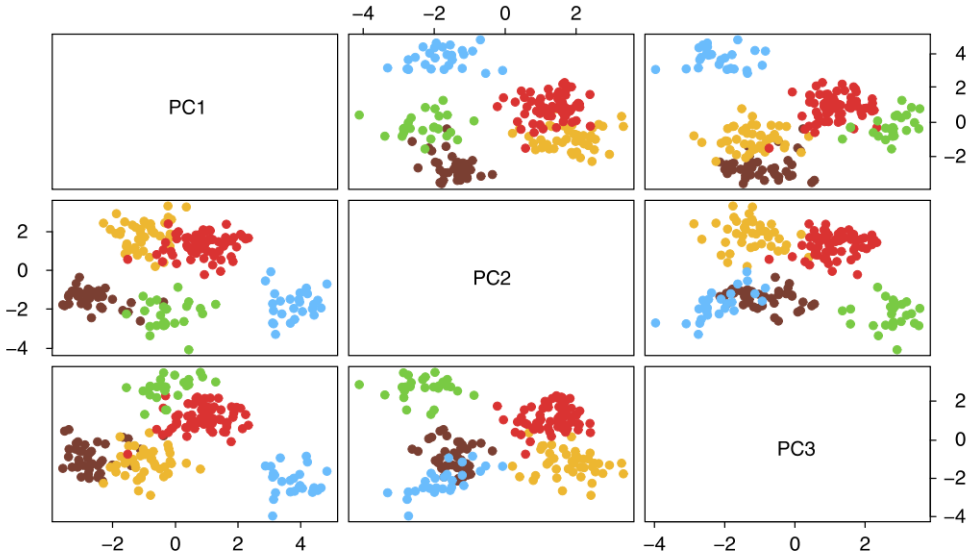

Figure 1 Plot of the first three principal components obtained using the panel of 48 single nucleotide polymorphisms (coded as genotype), selected after the first step with the penalized multinomial regression and stability selection procedure. ● = Valle del Belice (VdB), ● = Comisana (Com), ● = Pinzirita (Pin), ● = Barbaresca (Bar), ● = Sarda (Sar).

mapped onto 26 sheep autosomes were used, and after filtering (see Material and Methods), the final number of common SNPs was 48 068. On these SNPs, PMR and SS procedure, with $B = 500$, have been performed to select the most informative markers obtaining a final small set of 48 SNPs.

Figure 1 shows the 201 animals of the test population in the subspace defined by the first three principal components calculated on these 48 SNPs. A plot of the 201 animals of the test population in the subspace defined by the first three principal components calculated on 48 068 SNPs is shown in Supplementary Material Figure S1. This allowed an assessment of whether the reduced SNP panel leads to loss of important genetic information which is relevant to explain the differences among breeds. Figure 1 shows partial overlaps of historically and phylogenetically related breeds (Mastrangelo *et al.*, 2012 and 2014; Tolone *et al.*, 2012) and the difficulty in separating them. To analyse the structure of each cluster, we used the standard deviations as a measure of spread within each breed in the first three principal components. We observed a standard deviation average of about 0.65 for each principal component. Using this SNP panel, the corresponding misclassification error rate both for test and validation population was equal to 0%.

The unsupervised strategy which consists of the combination of PCA and *k*-means also provide excellent results. Using the first 15 principal components, which are highly correlated (>0.50) with 31 SNPs, and explaining 70% of total variability, we miss only one individual to perfectly discriminate the five breeds involved in the study. A cluster plot is shown in Supplementary Material Figure S2.

Moreover, using the whole set of SNPs and applying the same unsupervised strategy we again missed only one individual. In this case, 112 principal components, which are highly correlated (>0.50) with 608 SNPs, are used in the *k*-means step.

Table 1 *Simulation results of accuracy for classification both in test and validation population (based on 500 runs) after randomly sampled different number of single nucleotide polymorphisms from the whole set*

p	Average % test	Average % val	Kruskal–Wallis test	*P*-values
48	99.9%	60.6%	–	–
50	99.9%	63.3%	27.66	<0.0001
100	100%	75.2%	376.50	<0.0001
200	100%	83.1%	254.08	<0.0001
400	100%	87.7%	126.38	<0.0001
800	100%	90.8%	82.10	<0.0001
1600	100%	92.7%	41.67	<0.0001
3200	100%	94.5%	265.60	<0.0001
6400	100%	95.2%	16.63	<0.0001

The last two columns are Kruskal–Wallis test and *P* values.

Random sampling of single nucleotide polymorphisms

In order to assess the ability of these 48 selected SNPs to efficiently discriminate among the sheep breeds, a simulation was performed. Another sets of 48 SNPs were randomly sampled 500 times from the whole set of SNPs and the average of accuracy for classification in the validation population was about 60%. Repeating this procedure, sampling different numbers of SNPs (i.e. 50, 100, 200, 400, 800, 1600, 3200, 6400), the accuracy for classification was tested using a Kruskal–Wallis rank sum test (Kruskal and Wallis, 1952), to evaluate if any increment in accuracy was significant. These results are shown in Table 1.

Figure 2 shows the strength of the selected panel of 48 SNPs to discriminate across all the breeds, and the difficulty to perfectly discriminate among them using a large set of SNPs. Moreover, the Kruskal–Wallis test results were significant for each increment after sampling even more SNPs (Table 1).

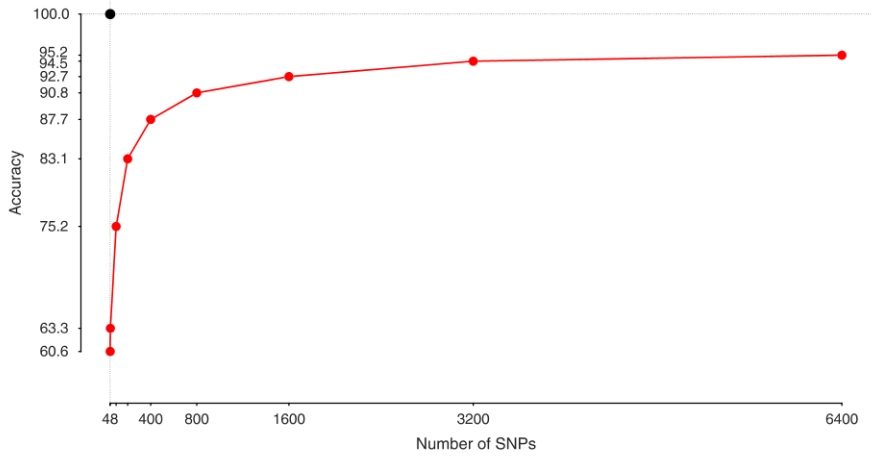

Figure 2 Plot of the mean accuracy to classify among the five breeds after repeating, for different number of single nucleotide polymorphisms (SNPs), a random sampling procedure from the whole set of available SNPs. Black dot is the accuracy level of the selected 48 SNP panel.

Table 2 *Out-of-bag (OOB) errors on the test population and misclassification error rates on the validation population for the two single nucleotide polymorphism panels and the two rankings, mean decrease in the Gini index (MDGI) and mean accuracy decrease (MAD) and for the mixed strategy penalized multinomial regression and stability selection (PMR–SS)*

Rankings	No. of SNPs	OOB	Misclassification
MDGI	48	8.21/201	1/35
	96	4.08/201	1/35
MAD	48	5.12/201	2/35
	96	5.12/201	1/35
PMR–SS	48	0/201	0/35

Penalized multinomial regression and stability selection v. *principal component analysis and random forest*

Penalized multinomial regression and stability selection procedure is a new strategy used for assigning animals to a breed. In order to compare our approach with other previously reported strategies and to test its efficiency in assigning individuals, PCA and RF strategy (Bertolini *et al.*, 2015) were also used with the real data. With respect to the two first ranking SNP panels (MDGI and MAD for 48 and 96 SNPs), the OOB errors in the test population were 4.09% and 2.03%, respectively, while the misclassification error rates for the validation population were both 2.86%. In the second ranking, the OOB errors for the test population were 2.55% for the 48 SNP panel and 2.55% for the 96 SNP panel, whilst the misclassification error rates are 5.71% and 2.86%, respectively. These results were summarized in Table 2. Figure 3 shows the distribution of the 48 selected SNPs along the 26 chromosomes, and the four SNPs panels obtained through PCA and RF procedure; 15 and 13 SNPs out of 48 are the same as in two 48 rankings MDGI and MAD, respectively.

To compare PMR–SS and PCA-RF strategies in more depth, we performed a simulation study. We artificially built, 300 times, test ($\tilde{\boldsymbol{X}}_{\text{test}}$) and validation ($\tilde{\boldsymbol{X}}_{\text{val}}$) population sampling with replacement the observation of the real data set ($\boldsymbol{X}$). For each replicate, OOB and misclassification error rates were calculated according to a new reselected SNP panel. The results are summarized in Table 3.

Breed assignment

The performance of the selected informative SNP markers in individual assignment test was evaluated using traditional genetic statistics such as model-based clustering algorithm and Wright's fixation index. These analyses were conducted using the whole set of SNPs (48 068) and the final number of selected SNPs (48). Results from within population substructure, using admixture analysis and considering a range of 2 through 10 potential clusters (K), indicated that the most probable number of inferred populations was $K = 5$. A graphic representation of the estimated membership coefficients, using the whole set of SNPs and the final number of selected SNPs is shown in Figure 4, where model-based clustering partitioned the genome of each sample into a predefined number of components. Some breeds tend to have their own distinct cluster (Bar, Sar and VdB), whereas other breeds, such as Pin and Com, showed a complex admixture-like pattern. These results support the findings on the basis of PCA.

The degree of genetic differentiation between pairs of breeds is reported in Table 4. The highest F_{st} value, for both SNPs panels, is seen between Bar and Sar and the lowest value was for Com *v.* Pin. Based upon the reference population, the average pairwise breeds F_{st} showed a higher value using the 48 SNPs, confirming the ability of this method to select discriminating markers.

Discussion

The aim of this study was to apply a new strategy to identify the minimum number of informative SNPs from high-throughput genotyping data in sheep breeds reared in Sicily and to investigate their usefulness for breed assignment purposes. Generally, the selection of genetic markers useful

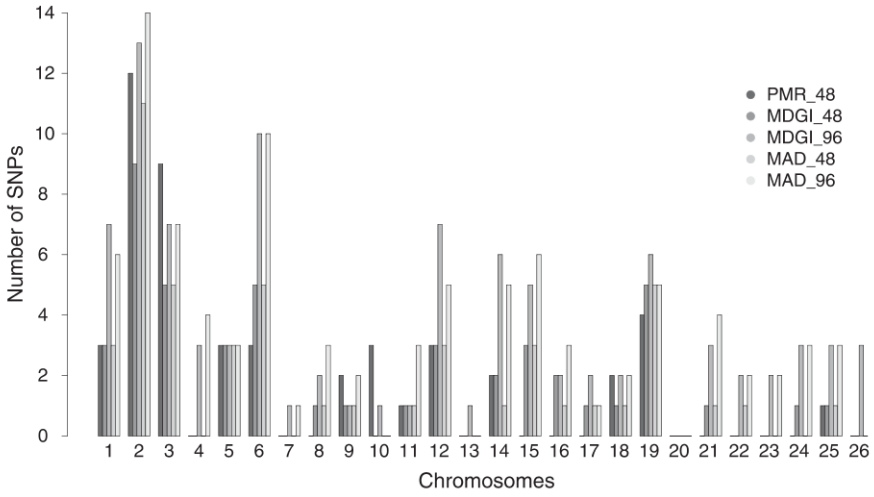

Figure 3 Chromosome distribution of the single nucleotide polymorphisms (SNPs) selected according to the proposed strategy, penalized multinomial regression and stability selection (PMR–SS), and whom according the two panels of mean decrease in the Gini index (MDGI) and the two panels of mean accuracy decrease (MAD).

Table 3 *Out-of-bag (OOB%) error and misclassification error rate (MER%) on the test and validation population for both strategies, penalized multinomial regression and stability selection and principal component analysis-random forest*

	MDGI 48	MDGI 96	MAD 48	MAD 96	PMR–SS 48
OOB%	1.63 (0.81)	1.29 (0.80)	1.49 (0.77)	1.25 (0.83)	0.00 (0.00)
MER%	2.11 (2.60)	1.49 (1.89)	1.91 (2.49)	1.49 (1.97)	1.46 (1.82)

MDGI = mean decrease in the Gini index; MAD = mean accuracy decrease; PMR–SS = penalized multinomial regression and stability selection.
Between brackets standard deviations. Results based on 300 simulation runs.

for these purposes is based on two approaches: a deterministic one, in which markers with different allelic variants fixed in the compared breeds are used, and the probabilistic one, in which selected markers present typical allelic frequencies in different breeds (Negrini *et al.*, 2009).

Several strategies have been already proposed to identify breed-informative SNPs derived from high-throughput genotyping platforms. These systems usually include a first step in which SNPs are preselected and a second step in which different assignment methods are applied (Bertolini *et al.*, 2015). For example, Allen *et al.* (2010) in a study on Irish cattle, reported a set of 43 SNPs for breed identification on the basis of allele frequency. Heaton *et al.* (2014) identified 163 SNPs for use in parentage testing and traceability in sheep, using the minor allele frequency (>0.3). In Mastrangelo *et al.* (2014), a subset of 119 SNPs was tested to evaluate their ability to assign individuals to the same groups that have been used in the present study. These SNPs were selected according to their informativeness in breed pair comparisons meaning SNPs with the largest allele frequency differences between pairs of breeds were chosen (fixed alleles in one breed and MAF > 0.25 in the other ones). Principal component analysis and *k*-means using this subset of SNPs showed a lack of ability to discriminate among the breeds and the presence of overlapped areas. Recently, Dimauro *et al.* (2015) using three complementary multivariate statistical techniques (discriminant analysis) and using two reduced pools of 110 and 108 SNPs, respectively, obtained a separation among divergent sheep breeds.

In this paper, supervised approaches, penalized multinomial regression and stability selection procedures were applied to identify the minimum number of informative SNPs from high-throughput genotyping data and these were used as a classification method for unknown samples. The method proposed in the present work differs from other studies due to the statistical technique used to reduce the number of SNPs. The main result was the selection of 48 SNPs from a whole set of 48 068 and these contained sufficient genetic information to produce sufficient power for individuals' breed assignment, using a relatively low number of individuals for breed and closely related breeds. The majority of the SNPs are in non-coding/intergenic regions of the sheep genome (Supplementary Material Table S1) 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).

The study proved that the combination of these methods allowed efficient discrimination between individuals of the studied breeds. Of course, the 48 identified SNPs were useful to discriminate among all the sheep breeds under study and these markers are probably not useful to discriminate among other sheep breeds. However, this strategy could be easily reproduced to discriminate among other breeds. Wilkinson *et al.* (2011) reported poor assignment power for breeds with low sample size and closely related individuals, showing that closely related breeds require about 200 markers to achieve >95% assignment success. Bertolini *et al.* (2017) in a study on cosmopolitan and autochthonous cattle breeds showed that a 96 SNP panel was generally more able to discriminate all breeds, whereas for the 48 SNP panel, the error rate increased mainly for autochthonous breeds, probably as a consequence of their admixed origin, lower selection

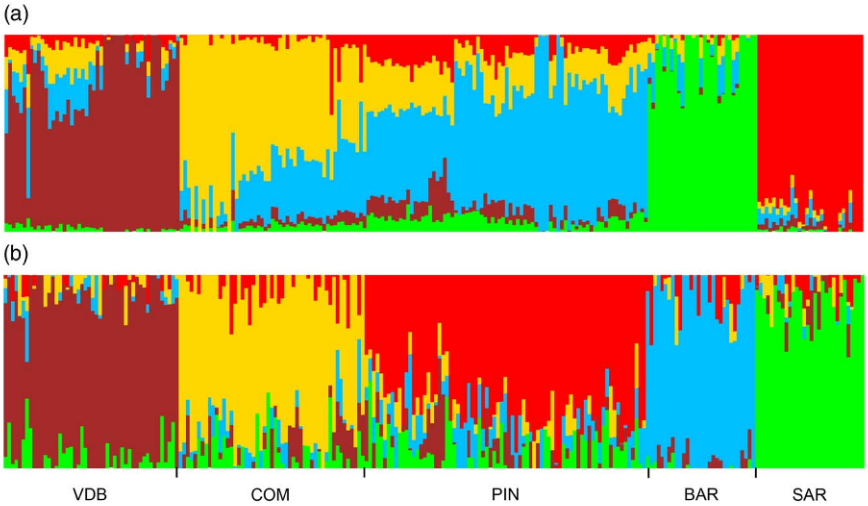

Figure 4 Model-based clustering of the five sheep breeds analysed for the most likely clusters ($K = 5$), using (a) the whole set of single nucleotide polymorphisms (SNPs; 48 068) and (b) the final number of selected SNPs (48).

Table 4 *Population genetic differentiation (F_{st} statistic) among the five sheep breeds using the whole set of single nucleotide polymorphisms (SNPs; 48 068) (above diagonal) and the final number of selected SNPs (48) (below diagonal)*

	VdB	Com	Pin	Bar	Sar
VdB	0	0.05	0.04	0.10	0.07
Com	0.24	0	0.02	0.08	0.06
Pin	0.26	0.18	0	0.07	0.04
Bar	0.43	0.37	0.30	0	0.11
Sar	0.31	0.31	0.25	0.42	0

VdB = Valle del Belice; Com = Comisana; Pin = Pinzirita; Bar = Barbaresca; Sar = Sarda.

pressure and by ascertaining bias in the construction of the SNP chip. In fact, where there is sufficient genetic heterogeneity among populations, a few genetic markers can be easily used to identify and verify the origin of individuals, whereas it becomes more complicated for population with low genetic differentiation, such as the sheep breeds involved in this study (Mastrangelo *et al.*, 2012; Tolone *et al.*, 2012). It is well known that a high number of genotyped animals can capture the whole within population variability reducing the possibility that some individuals would not be assigned correctly due to atypical genotypes (Hulsegge *et al.*, 2013). Considering the high level of admixture among these sheep breeds (Mastrangelo *et al.*, 2017), and the relative low number of analysed individuals, our study reported relevant results. A good separation among breeds was still obtained with high percentages of correct assignment. The applicability of reduced SNP panels with low classification error rate is therefore still possible also for local breeds in which the total or partial lack of selection programs have not shaped the genome as it might be the case for cosmopolite breeds.

The combined use of PCA and RF proposed by Bertolini *et al.* (2015), and applied to our sheep breeds, highlights difficulties to discriminate among them, even when using two different panels of 48 and 96 SNPs. The simulation results have also highlighted that the proposed strategy performed slightly better than PCA–RF strategy, as did the results in the real application. Therefore, the proposed strategy could provide a new tool to get over problems in which breeds are phylogenetically close.

The results reported using independent analyses, such as the model-based clustering algorithm implemented in admixture software (Alexander and Lange, 2011) and F_{st} confirmed the ability of this method in selecting discriminating markers. The reduced SNP panel captured a large proportion of genetic variation between the dairy sheep breeds with estimates of F_{st} exceeding those previously reported using microsatellites (Tolone *et al.*, 2012) and SNPs (Mastrangelo *et al.*, 2014). Moreover, a previous study on Sicilian sheep breeds (Tolone *et al.*, 2012) using a set of 20 microsatellites, reported that the Bayesian assignment test showed a low assignment value for these breeds, and the low robustness of the assignment test made it infeasible for traceability purposes.

Validation analyses will be conducted on the identified SNPs using a wider sample of individuals and other laboratory assay, for example Sanger sequencing. Finally, a multiplexed genotyping-by-sequence assay will be developed highlighting the economic advantage on the use of reduced SNP panels, compared with dense genome-wide assay, for routine use in the management of local populations.

Conclusions

Results for assignment test using the mixed strategy were interesting, because 100% of the individuals were correctly assigned to their breeds of origin. Using genotypic data, a small set of SNPs was identified. The results laid the basis to improve the proposed strategy for the potential use of it to generate panels that may be used for breed assignment or

within an industrial setting for tracing the origin of animal products derived from the five breeds involved in the study.

Acknowledgements

This research was financed by PON02_00451_3133441, CUP: B61C1200076005 funded by MIUR.

Supplementary material

To view supplementary material for this article, please visit https://doi.org/10.1017/10.1017/S175173111700266X

References

Alexander D and Lange K 2011. Enhancements to the admixture algorithm for individual ancestry estimation. BMC Bioinformatics 12, 1–6.

Allen AR, Taylor M, McKeown B, Curry AI, Lavery JF, Mitchell A, Hartshorne D, Fries R and Skuce RA 2010. Compilation of a panel of informative single nucleotide polymorphisms for bovine identification in the northern Irish cattle population. BMC Genetics 11, 1–8.

Bertolini F, Galimberti G, Calò DG, Schiavo G, Matassino D and 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, 346–356.

Bertolini F, Galimberti G, Schiavo G, Mastrangelo S, Di Gerlando R, Strillacci MG, Bagnato A, Portolano B and Fontanesi L 2017. Preselection statistics and random forest classification identify population informative single nucleotide polymorphisms in cosmopolitan and autochthonous cattle breeds. Animal https://doi.org/10.1017/S1751731117001355.

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR and Cavalli-Sforza LL 1994. High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368, 455–457.

Dimauro C, Cellesi M, Steri R, Gaspa G, Sorbolini S, Stella A and Macciotta NPP 2013. Use of the canonical discriminant analysis to select SNP markers for bovine breed assignment and traceability purposes. Animal Genetics 44, 377–382.

Dimauro C, Nicoloso L, Cellesi M, Macciotta NPP, Ciani E, Moioli B, Pilla F and Crepaldi P 2015. Selection of discriminant SNP markers for breed and geographic assignment of Italian sheep. Small Ruminant Research 128, 27–33.

Friedman J, Hastie T and Tibshirani R 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33, 1–22.

Heaton MP, Leymaster KA, Kalbfeisch TS, Kijas JW, Clarke SM, McEwan J, Maddox JF, Basnayake V, Petrik DT, Simpson B, Smith TP and Chitko-McKown CG 2014. SNPs for parentage testing and traceability in globally diverse breeds of sheep. PLoS One 9, e94851.

Hulsegge B, Calus MPL, Windig JJ, Hoving-Bolink AH, Maurice-van Eijndhoven MH and Hiemstra SJ 2013. Selection of SNP from 50K and 777K arrays to predict breed of origin in cattle. Journal of Animal Science 91, 5128–5134.

Jakobsson M and Rosenberg NA 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23, 1801–1806.

Kuehn LA, Keele JW, Bennett GL, McDaneld TG, Smith TP, Snelling WM, Sonstegard TS and Thallman RM 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, 1742–1750.

Kruskal WH and Wallis WA 1952. Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association 47, 583–621.

Mastrangelo S, Di Gerlando R, Tolone M, Tortorici L, Sardina MT and Portolano B 2014. Genome wide linkage disequilibrium and genetic structure in Sicilian dairy sheep breeds. BMC Genetics 15, 108.

Mastrangelo S, Portolano B, Di Gerlando R, Ciampolini R, Tolone M and Sardina MT 2017. Genome-wide analysis in endangered populations: a case study in Barbaresca sheep breed. Animal 12, 1–10.

Mastrangelo S, Sardina MT, Riggio V and Portolano B 2012. Study of polymorphisms in the promoter region of ovine β-lactoglobulin gene and phylogenetic analysis among the Valle del Belice breed and other sheep breeds considered as ancestors. Molecular Biology Reports 39, 745–751.

Meinshausen N and Bühlmann P 2010. Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72, 417–473.

Negrini R, Nicoloso L, Crepaldi P, Milanesi E, Colli L, Chegdani F, Pariset L, Dunner S, Leveziel H, Williams JL and Ajmone Marsan P 2009. Assessing SNP markers for assigning individuals to cattle populations. Animal Genetics 40, 18–26.

Nicolazzi E, Caprera A, Nazzicari N, Cozzi P, Strozzi F, Lawley C, Pirani A, Soans C, Brew F, Jorjani H, Evans G, Simpson B, Tosse-Klopp G, Brauning R, Williams JL and Stella A 2015. SNPchiMp v.3: integrating and standardizing single nucleotide polymorphism data for livestock species. BMC Genomics 16, 283.

Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, Mahoney MW and Drineas P 2007. PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genetics 3, e160.

R Core Team 2016. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rousset F 2008. GENEPOP '007: a complete re-implementation of the GENEPOP software for Windows and Linux. Molecular Ecology Resources 8, 103–106.

Rosenberg NA 2005. Algorithms for selecting informative marker panels for population assignment. Journal of Computational Biology 12, 1183–1201.

Shriver MD, Smith MW, Jin L, Akey JM, Deka R and Ferrell RE 1997. Ethnic-affiliation estimation by use of population-specific DNA markers. American Journal of Human Genetics 60, 957–964.

Tibshirani R 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 58, 267–288.

Tolone M, Mastrangelo S, Rosa AJM and Portolano B 2012. Genetic diversity and population structure of Sicilian sheep breeds using microsatellite markers. Small Ruminant Research 102, 18–25.

Wilkinson S, Wiener P, Archibald AL, Law A, Schnabel RD, McKay SD, Taylor JF and Ogden R 2011. Evaluation of approaches for identifying population informative markers from high density SNP chips. BMC Genetics 12, 45.