


A machine learning approach for the identification of population-informative markers from high-throughput genotyping data: application to several pig breeds

G. Schiavo¹, F. Bertolini², G. Galimberti³, S. Bovo¹, S. Dall'Olio¹, L. Nanni Costa¹, M. Gallo⁴ and L. Fontanesi^{1†} 

¹Department of Agricultural and Food Sciences, Division of Animal Sciences, University of Bologna, Viale G. Fanin 46, Bologna 40127, Italy; ²National Institute of Aquatic Resources, Technical University of Denmark, 2800 Kongens Lyngby, Denmark; ³Department of Statistical Sciences 'Paolo Fortunati', University of Bologna, via delle Belle Arti 41, Bologna 40126, Italy; ⁴Associazione Nazionale Allevatori Suini (ANAS), Via Nizza 53, Roma 00198, Italy

(Received 5 January 2019; Accepted 5 August 2019; First published online 11 October 2019)

Single nucleotide polymorphisms (SNPs) able to describe population differences can be used for important applications in livestock, including breed assignment of individual animals, authentication of mono-breed products and parentage verification among several other applications. To identify the most discriminating SNPs among thousands of markers in the available commercial SNP chip tools, several methods have been used. Random forest (RF) is a machine learning technique that has been proposed for this purpose. In this study, we used RF to analyse PorcineSNP60 BeadChip array genotyping data obtained from a total of 2737 pigs of 7 Italian pig breeds (3 cosmopolitan-derived breeds: Italian Large White, Italian Duroc and Italian Landrace, and 4 autochthonous breeds: Apulo-Calabrese, Casertana, Cinta Senese and Nero Siciliano) to identify breed informative and reduced SNP panels using the mean decrease in the Gini Index and the Mean Decrease in Accuracy parameters with stability evaluation. Other reduced informative SNP panels were obtained using Delta, Fixation index and principal component analysis statistics, and their performances were compared with those obtained using the RF-defined panels using the RF classification method and its derived Out Of Bag rates and correct prediction proportions. Therefore, the performances of a total of six reduced panels were evaluated. The correct assignment of the animals to its breed was close to 100% for all tested approaches. Porcine chromosome 8 harboured the largest number of selected SNPs across all panels. Many SNPs were included in genomic regions in which previous studies identified signatures of selection or genes (e.g. ESR1, KITL and LCORL) that could contribute to explain, at least in part, phenotypically or economically relevant traits that might differentiate cosmopolitan and autochthonous pig breeds. Random forest used as preselection statistics highlighted informative SNPs that were not the same as those identified by other methods. This might be due to specific features of this machine learning methodology. It will be interesting to explore if the adaptation of RF methods for the identification of selection signature regions could be able to describe population-specific features that are not captured by other approaches.

Keywords allocation, random forest, selection signature, single nucleotide polymorphism, *Sus scrofa*

Implications

The identification of breed informative markers in the genome of livestock species can have several practical applications. This study evaluated the performances of a machine learning approach (random forest) to identify informative single nucleotide polymorphisms in seven pig breeds and compared the results with other three methods (Fixation index, Delta and principal component analysis). The random forest approach proposed in this study introduced a useful methodology that can be extended when it is needed to select informative single

nucleotide polymorphisms. Random forest was able to identify markers in genes affecting phenotypic traits that could differentiate the investigated pig breeds.

Introduction

Genetic diversity among livestock breeds and populations derives from many different events that have contributed to shape their peculiar population genetic structures and uniqueness. Breeds are the results of artificial and natural selection or adaptation to different farming and

[†] E-mail: luca.fontanesi@unibo.it

environmental conditions, genetic drift and admixture that have modified allele frequencies and fixed (or almost fixed) genetic variants. Commercial single nucleotide polymorphism (SNP) genotyping tools, developed for the most important livestock species, including the pig, have been used to describe genetic variability and to capture breed or population-informative markers useful for several applications. Preselected and informative SNP panels have been proposed for parentage verification, comparative selection signature analyses, breed assignment of individual animals and breed authentication of mono-breed products as well as for several other applications (e.g. Wilkinson *et al.*, 2011; Bertolini *et al.*, 2015).

To identify the most discriminating SNPs among thousands of markers included in the commercial tools, several statistical measures have been applied. One of the simplest proposed methods is based on the Delta values, which are the absolute allele frequency differences at each polymorphic marker in pairwise comparisons. For example, Delta statistic has been already used to identify informative markers in British pig breeds (Wilkinson *et al.*, 2012). Fixation index (Fst) is another statistic extensively applied to identify population-informative SNPs, population structures and signature of selection in livestock (Wilkinson *et al.*, 2011; Hulsegge *et al.*, 2013). It calculates the standardized variance in allele frequencies among populations. Principal component analysis (PCA) is an unsupervised linear technique for dimension reduction that allows to extract axes of maximal variation from datasets (Jolliffe and Cadima, 2016). Principal component analysis has been first used on SNP data to describe the structure of human populations (Paschou *et al.*, 2007). This multivariate approach has been subsequently used in livestock to reduce dimensionality of large SNP datasets and to identify cattle breed informative SNPs (Wilkinson *et al.*, 2011; Bertolini *et al.*, 2015). Despite all these methods being standard, their application in terms of compared populations (i.e. all considered groups or only pairwise comparisons) and evaluated SNP datasets (i.e. all SNPs across all autosomes, only tag SNPs or only chromosome by chromosome analyses) has not been consistent across studies. Most studies have not considered high levels of linkage disequilibrium (LD) among informative SNPs when developing breed informative panels.

In many cases, these preselection statistics are then coupled with other techniques that can classify or assign individuals to groups and estimate the discriminatory or allocation power or error. Random forest (RF) is a machine learning technique that has been proposed for these purposes (Bertolini *et al.*, 2015 and 2018). Random forests are ensemble techniques that derive prediction rules by combining several binary decision trees obtained after introducing random perturbations in the data. These random perturbations are introduced to reduce correlation among the decision trees, thus leading to ensemble prediction rules with a prediction error lower than those derived from single decision trees (Breiman, 2001). These ensemble prediction rules can be applied to assign an unknown sample to one of the predetermined groups. Random forest has been used in population genomics for several pilot applications that range from

genome-wide association studies (Kijas *et al.*, 2013) to estimation of genomic breeding values (Naderi *et al.*, 2016). A few works used RF for the detection of cryptic population structures and the selection of informative markers without any preselection steps (Jacobs *et al.*, 2018). We recently explored performances of RF to identify breed informative SNPs in cattle breeds (Bertolini *et al.*, 2015 and 2018). As RF is computationally challenging when using thousands of markers and prone to be biased by high LD between markers (Meng *et al.*, 2009), we combined this technique with several SNP prefiltering approaches. These strategies were able to identify SNPs that are located in genes or in genomic regions known to affect cattle breed-specific traits (Bertolini *et al.*, 2015 and 2018).

In this study, we extended the use of RF to analyse SNP chip data of seven Italian pig breeds (which have peculiar production and phenotypic characteristics), including three cosmopolitan-derived breeds (Italian Large White, Italian Duroc and Italian Landrace) and four autochthonous breeds (Apulo-Calabrese, Casertana, Cinta Senese and Nero Siciliano). To reduce computational burden and the problem derived by the potential high LD within selected SNP panels, a tagged SNP dataset was used for the identification of informative SNPs using RF methods and a few other statistics (Delta, Fst and PCA). We then compared the performances of these SNP panels in terms of individual allocation errors to these breeds using RF classifications. We evidenced that many selected SNPs are included in genomic regions in which previous studies identified signatures of selection or that could contribute to explain, at least in part, phenotypically or economically relevant traits that differentiate cosmopolitan and autochthonous pig breeds.

Material and methods

Animals and single nucleotide polymorphism datasets

A total of 2737 pigs from 7 pig breeds (Italian Large White, n. 1983; Italian Duroc, n. 432; Italian Landrace, n. 48; Apulo-Calabrese, n. 92; Casertana, n. 96; Cinta Senese, n. 38; Nero Siciliano, n. 48) were genotyped with the PorcineSNP60 BeadChip array (Illumina, San Diego, CA, USA). The pigs of the cosmopolitan-derived breeds were from performance-tested animals evaluated under the national selection program run by the National Pig Breeder Association (ANAS). Apulo-Calabrese, Casertana, Cinta Senese and Nero Siciliano are Italian local pig breeds under the conservation program managed by ANAS. About 200 to 1000 pigs of these breeds are registered to their respective herd books (ANAS, 2018). Apulo-Calabrese pigs have solid black coat colour. These pigs are raised in the Central-South of Italy. Casertana pigs have dark coat colour (dark grey or black) with a hairless phenotype, mainly raised in Molise, Campania and Puglia regions (Central-South of Italy). Casertana breed is considered the descendant of the Neapolitan pig population that influenced the first British breeds in the 19th century. Cinta Senese (Siena Belted) pigs,

farmed in Toscana region, have a characteristic black coat colour with a white belted phenotype. Nero Siciliano (Sicilian Black) pigs, raised in the Sicily island, have solid black coat colour. All these local breeds are raised in extensive or semi-extensive farming systems. More information on all animals included in this study are reported in Supplementary Material Table S1.

Single nucleotide polymorphisms were mapped on the Sscrofa11.1 (GCA_000003025.6, National Center for Biotechnology Information). Only SNPs located in unique positions and mapped on autosomal chromosomes were analysed. Single nucleotide polymorphisms were discarded if monomorphic in all breeds and if call rate in at least one breed was <0.98 . These basic statistics were computed with PLINK software version 1.9 (Chang *et al.*, 2015). Animals were discarded when individual call rate was <0.90 of all SNPs.

Reference and test populations

Each pig breed was then randomly divided into a reference population and a test population. The reference population included 90% of pigs, whereas the test population included the remaining 10% of animals. The reference population was used for all SNP analyses, including the preselection steps and allocation analyses. The test population was created for the validation step of the subsequent analyses. However, it is worth to mention that RF does not need any cross-validation on a separate test set to get an unbiased estimate of the test set error. Error in the RF classification is estimated internally, directly during the run. However, the use of the test population could provide a further validation of the results. Multidimensional scaling (MDS) plots were obtained using starting reference dataset, and subsequent reduction (96 SNP panels) was obtained using PLINK software version 1.9 (Chang *et al.*, 2015) and plotted with the R package 'scatterplot3D' (Ligges and Mächler, 2003).

Single nucleotide polymorphism selection strategies and breed allocation

Tag SNPs from the whole filtered SNP dataset were identified for each breed separately with the PLINK tagging option '-show-tags all' without any distance threshold as follows: (i) all SNPs with $r^2 < 0.3$ with any other SNPs, within breed, were kept; (ii) when a group of SNPs within breed was tagged by one SNP, only one SNP randomly chosen was kept. The tagged SNP dataset was then analysed to identify breed informative SNPs using Delta, Fst and PCA statistics and two RF approaches that retained in total six 96 SNP panels (see below for details). The RF approach was used to evaluate the best panels among those obtained by the mentioned approaches. The choice of this number of SNPs was due to the practical possibility to develop multiplex SNP panels for field applications (Bertolini *et al.*, 2015).

Delta. Delta values were calculated as the absolute difference between allele frequencies at each considered SNPs in pairwise comparisons: $|p_{Ai} - p_{Aj}|$, where p_{Ai} and p_{Aj} are the frequencies of allele A in the i th and j th breeds,

respectively. Pairwise values were averaged across all pairwise comparisons to obtain and estimate value for each SNP (Wilkinson *et al.*, 2011).

Fixation index. The Fst, as proposed by Weir and Cockerham (1984), was computed with PLINK considering population-specific allele frequency at each SNP.

Principal component analysis. Principal component analysis is an unsupervised learning technique for dimension reduction based on the singular value decomposition of the data matrix containing one row for each pig and one column for each SNP. Axes of maximal variations (sometimes referred to as eigenSNPs in the context of genomic analysis) are obtained by computing linear combinations of the SNPs using the left singular vector of the data matrix. Principal component analysis was computed using the *prcomp* function of the R software 2.12 (<https://www.r-project.org/>), with default parameters after codifying each animal with a vector of values (0, 1 or 2, depending on the number of minor alleles for each SNP). The informativeness for the SNPs was determined by considering the sum of the squares of the six first principal components (PC), according to Paschou *et al.* (2007). The choice of the number of PC was determined by the amount of variance explained as previously defined (Bertolini *et al.*, 2015). The resulting values were used to rank SNPs.

Random forest. Random forest is a supervised learning method to build classification rules based on the aggregation of a number of classification trees. These trees are fitted after introducing random perturbations in the data. The aim of these random perturbations is to reduce correlation among the prediction rules associated with each single tree. More specifically, the recursive algorithm used to fit classification trees is applied to a bootstrap version of the reference population, obtained by randomly selecting (i.e. after randomly selecting pigs from the reference population with replacement). Furthermore, at each step of this recursive algorithm only a random subset of SNPs is considered in order to define the optimal splitting rule to grow the tree. No pruning step is performed after the growing step, so that the prediction rules associated with the classification trees that compose the RF are characterized by a small bias and a large variance. A final aggregation step (typically by majority voting) leads to an ensemble classification rule that preserves the small bias associated with each single tree while reducing the variance, thus leading to a lower prediction error. More details about RF can be found in Breiman (2001), along with a proof showing that the reduction in the prediction error due to the final aggregation step increases as the correlation among the individual trees decreases. It is worth mentioning that the use of bootstrap to create random perturbations in the data leads to the definition of an Out Of Bag (OOB) population for each individual tree in the forest. This OOB population consists of all the pigs that are not included in the bootstrap population used to build a given single tree and that can be used to obtain internal unbiased estimates of the prediction error

and to evaluate variable importance. Random forest analyses were performed on the mentioned vectors using the *randomForest* package in R (Liaw and Wiener, 2002). Classes were weighted depending on the number of animals in each breed, using the parameter 'classwt'. These weights were chosen in order to counterbalance the unbalanced number of the genotyped pigs in the investigated breeds. Single nucleotide polymorphisms were ranked using two different ranking parameters implemented in the function 'importance': the mean decrease in the Gini Index (GI) and the Mean Decrease in Accuracy (MDA). The GI is a variable importance measure that was specifically devised for ensemble of classification trees, such as RFs. It is based on the contribution of each variable in reducing the within-node heterogeneity of a tree. These contributions are averaged over all the trees that compose the RF. A high value means a high contribution of the SNP in shaping the structure of the trees that compose the RF, and hence in determining category assignments. The MDA is the decrease in the accuracy of the prediction rule induced by a random permutation of the values in each feature (for more details, see Hastie *et al.* (2009)). It is worth mentioning that different runs of the RF procedure can lead to different results in terms of GI and MDA. These differences are due to the random perturbation mechanism that is applied to the data. In order to assess the impact of these differences and to evaluate the stability of the RF selection, 100 runs of the analysis were performed. At each run, the SNPs were ranked by importance. Then, two approaches were used to assess stability for the SNP selection procedure: (1) the number of times an SNP was among the first top selected 96 was recorded and the SNPs that occurred more frequently were then listed as the most stable; (2) the importance value of the SNPs was averaged over the 100 runs, then the 96 SNP panels were chosen by selecting the SNPs with the highest importance average value. These two methods applied to assess stability and then to select the SNPs were used by considering the GI and the MDA approaches separately, leading to four 96 SNP panels. However, for the MDA, the 96 SNPs with the highest average importance value were the same which occurred most frequently in the 100 runs; therefore, the two panels determined with the two different stability methods included the same SNPs. For the GI, the two stability methods produced two non-identical 96 SNP panels (even if largely overlapping) that were evaluated separately.

To compare the performances of the different SNP pre-selection methods in the same way, RF was then applied to evaluate the corresponding OOB error rate (a method of measuring the prediction error of the RF classifier) of all other final 96 SNP panels defined using Delta, Fst and PCA statistics. A further evaluation of the informativity of the six identified 96 SNP panels was obtained by running RF prediction to the test population.

Single nucleotide polymorphism annotation

Genes annotated in the Sscrofa11.1 genome version spanning a region of ± 500 kb around all SNPs included in

the final panels were retrieved using Ensembl Biomart tool (<http://www.ensembl.org/biomart/martview/>). The closest genes to the SNPs in the final SNP panels were then reported. Identified genes and corresponding chromosome regions were then compared to selection signature regions described in the *Sus scrofa* genome by previous studies.

Results

Population genetic parameters

The number of pigs for each breed that remained in the final dataset was 1968 for Italian Large White, 432 for Italian Duroc, 46 for Italian Landrace, 92 for Apulo-Calabrese, 96 for Casertana, 38 for Cinta Senese and 48 for Nero Siciliano. One tenth of these animals were used to construct the test population.

A total of 40 680 SNPs was retained after the filtering steps. About 22.5% of these SNPs ($n = 9172$) constituted the tag SNP dataset. Supplementary Material Table S2 reports the distribution of these SNPs in the 2 datasets (i.e. after filtering and then after tagging) for the 18 autosomes. The number of monomorphic SNPs in the tagged dataset ranged from 105 (in the Italian Large White breed) to 1328 (in the Casertana breed). The comparison of these SNPs revealed that a total of 119 SNPs had a private allele in one of the analysed breeds (for the same allele: allele frequency > 0 in one breed and allele frequency $= 0$ in all other breeds; Supplementary Material Table S3). The highest number of private alleles (most of them with minor allele frequency < 0.01) was observed in the Italian Large White breed, which accounted for the largest number of analysed animals. Monomorphic SNPs (including those with private alleles) were used in the subsequent SNP selection methods as some of them could be very informative in the breed comparisons.

Multidimensional scaling plots obtained using the whole and tagged SNP datasets (Figure 1) showed a clear separation of the Italian Large White and the Italian Duroc groups of pigs, whereas all other breeds clustered in a partially overlapping cloud.

Description of the 96 single nucleotide polymorphism panels

A total of six different reduced panels that included 96 SNPs each were selected by the different approaches that were applied: one was derived using Delta, one using Fst, one using PCA statistics (included for comparison) and three using RF by applying GI and MDA ranking methods (two stability methods were applied for each RF approach; for the MDA, the two stability methods obtained the same SNP set). Supplementary Material Table S4 lists all SNPs included in the six panels.

Table 1 reports the number of common SNPs among these panels. If we consider the two panels derived using the two stability methods of the RF GI approach, they shared 93 SNPs. Common SNPs were in general low in all other pairwise-based comparisons (ranging from 0 to 21). This might reflect

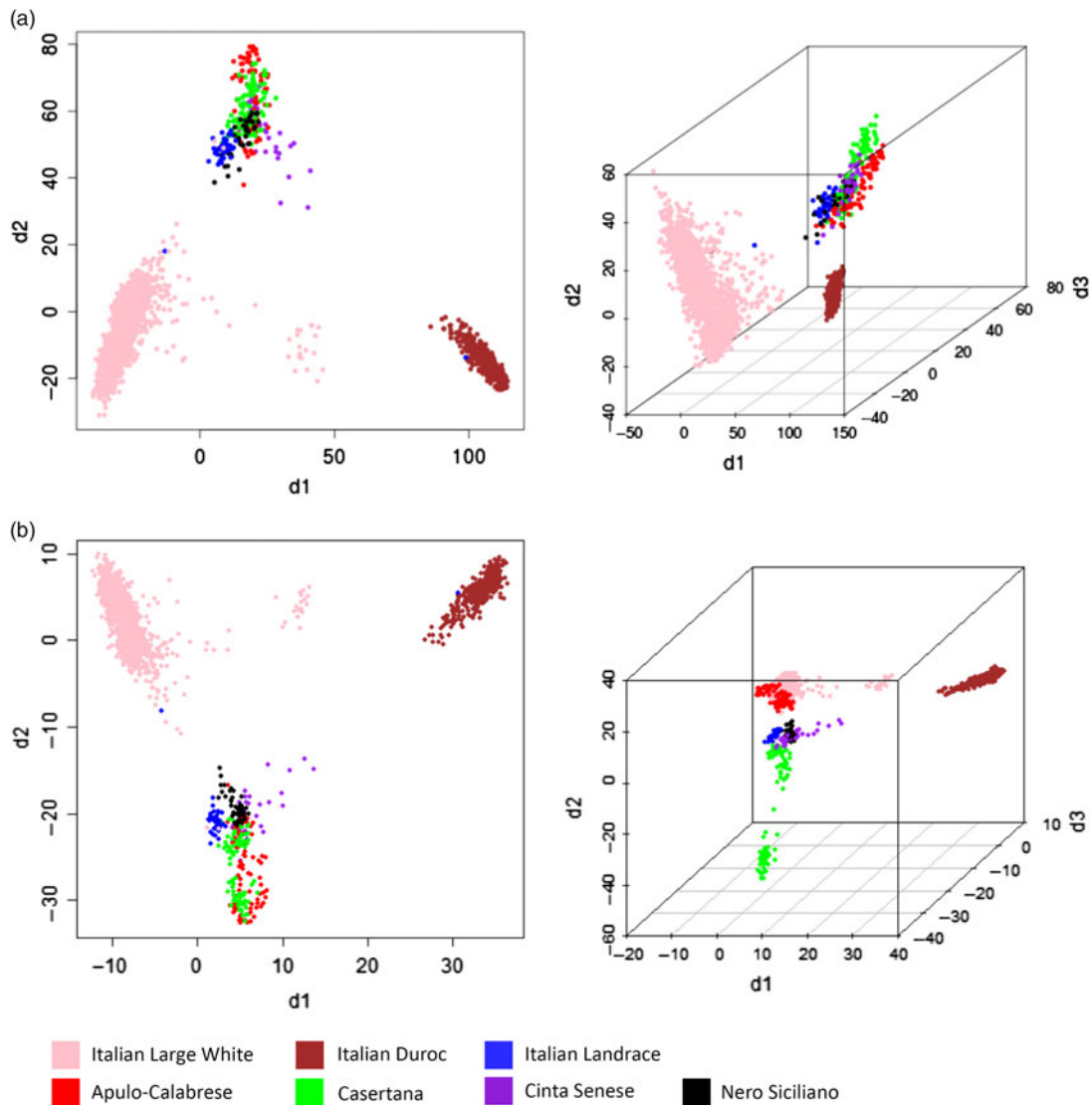


Figure 1 (Colour online) Bidimensional and tridimensional multidimensional scaling (MDS) plots obtained using the untagged (a) and tagged (b) single nucleotide polymorphism (SNP) datasets of the different pig breeds.

Table 1 Number of SNPs shared between pairs of pig SNP panels determined with the six different methods reported in this study (in the diagonal, the 96 SNPs)

Methods	RF Gini index 1 ¹	RF Gini index 2 ²	RF MDA	Delta	Fst	PCA
RF Gini Index 1	96					
RF Gini Index 2	93	96				
RF MDA	13	13	96			
Delta	20	21	15	96		
Fst	6	6	6	5	96	
PCA	1	1	13	17	0	96

SNPs = single nucleotide polymorphisms; RF MDA = random forest Mean Decrease in Accuracy; Fst = Fixation index; PCA = principal component analysis.

¹Random forest (RF) Gini Index 1 = stability mean.

²RF Gini Index 2 = stability occurrence.

the differences among the preselection techniques considered in this study. Being an unsupervised technique, PCA simply exploited the observed variability. As far as the other techniques are concerned, RF is the only one that exploited possible interactions among SNPs. Supplementary Material S1 shows bidimensional and tridimensional MDS plots obtained using the six reduced SNP panels. Cinta Senese pigs are clearly separated in the third dimension of the RF GI panels, whereas Apulo-Calabrese pigs are well separated in the third dimension of the Delta-derived SNP panel. Casertana animals are included in a separated cloud in the two-dimensional RF MDA-derived plot. Italian Large White and Italian Duroc pigs are always well separated in all these plots.

Figure 2 shows the chromosome distribution of the selected SNPs of the six panels. Within panel, SNPs per

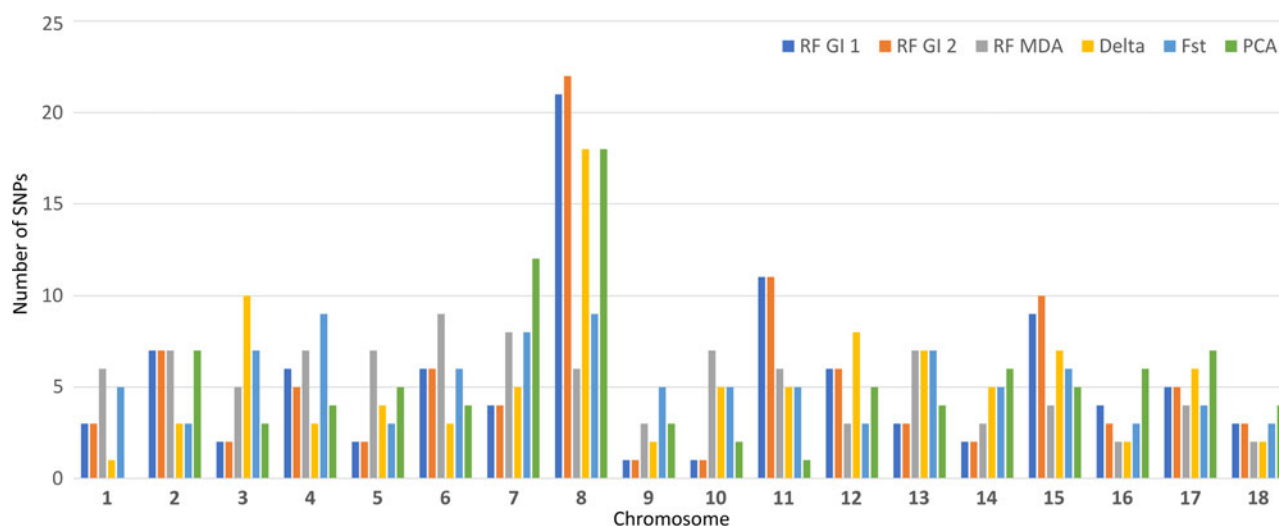


Figure 2 (Colour online) Distribution on the 18 porcine autosomes of the SNPs selected for the 96 SNP panels using the six different methods described in this study (RF GI 1 = random forest Gini Index stability mean; RF GI 2 = random forest Gini Index stability occurrence; RF MDA = random forest Mean Decrease in Accuracy; Delta; Fst = Fixation index; PCA = principal component analysis) for the analysis of different pig breeds.

chromosome ranged from zero (porcine chromosome (SSC) 1 with PCA, which was the only one chromosome/method without any SNP) to 22 (SSC8 with RF GI with stability selection based on occurrence: RF GI 2). Considering panels together, there was a similar distribution per chromosome of the selected SNPs. The sum of all the SNP detected in the six panel ranged from 15 SNPs on SSC9 to 41 SNPs on SSC7 and SSC15; SSC8 was an evident outlier as it harboured the largest number of SNPs in five out of six panels, with an overall total of 94 SNPs (the RF MDA method identified only 6 SNPs, the lowest number among the six panels for this chromosome).

Out Of Bag values of the six single nucleotide polymorphism panels

Random forest analyses were applied to the six panels with the purpose of learning a classification rule to assign animals to the seven pig breeds included in this study. Out Of Bag rates and correct prediction proportions in the reference population dataset are reported in Table 2. The highest OOB was observed for the SNP panel derived using the PCA method (2.15%), the lowest was observed for the Fst method (0.79%) whereas all RF derived panels had OOB values that ranged from 1.16% to 1.28% and the value for the Delta derived panel was 1.45%. The correct prediction proportions were always 100% except for two cases in which Fst and PCA methods mis-assigned just one Italian Duroc pig to the Italian Large White group. A general low error rate was also obtained in the test population (Supplementary Material Table S5): two to three Nero Siciliano pigs were mis-assigned to the Italian Large White population with all methods except for the Fst-derived panel that had the 100% of correct assignment. Finally, the PCA-derived panel assigned two Italian Landrace pigs to the Italian Large White breed and one Italian Landrace pig to the Italian Duroc breed.

Table 2 OOB error rate (%) and the CPP of the reference pig populations (total: considering all breeds together; or separated by breeds) using the six 96 SNP panels obtained using the RF, Delta, Fst and PCA methods

Parameters/ methods	RF Gini index 1 ¹	RF Gini index 2 ²	RF MDA	Delta	Fst	PCA
OOB error rate %	1.16	1.28	1.12	1.45	0.79	2.15
CPP ³ total	1	1	1	1	0.99	0.99
CCP Italian Large White	1	1	1	1	0.99 ⁴	0.99 ⁴
CCP Italian Landrace	1	1	1	1	1	1
CCP Italian Duroc	1	1	1	1	1	1
CCP Apulo-Calabrese	1	1	1	1	1	1
CCP Casertana	1	1	1	1	1	1
CCP Cinta Senese	1	1	1	1	1	1
CCP Nero Siciliano	1	1	1	1	1	1

OOB = Out Of Bag; CPP = correct prediction proportion; SNP = single nucleotide polymorphism; RF = random forest; Fst = Fixation index; PCA = principal component analysis; RF MDA = random forest Mean Decrease in Accuracy; Fst = Fixation index; PCA = principal component analysis.

¹Random forest (RF) Gini Index 1 = stability mean.

²RF Gini Index 2 = stability occurrence.

³CCP = (1 – Classification Error).

⁴One pig was assigned to the Italian Duroc breed.

Marked genes

The closest genes to the SNPs contained in the six reduced SNP panels are listed in Supplementary Material Table S4. A total of 42 SNPs of the two RF GI panels were within 42 different genes. For the other panels (RF MDA, Delta, Fst and PCA panels, respectively) a total of 50, 45, 38 and 48

Table 3 SNPs included in the panels selected by the six different methods used in this study (RF GI1, RF GI2, RF MDA, Delta, Fst and PCA) that are within or close to genes located in SSC regions in which other authors have reported signature of selection

Methods ¹	SSC	SNPs	Position ²	Ensembl ID ³	Gene symbol	Distance ⁴	References
Fst	1	MARC0049965	1208216	ENSSSCG00000004013	<i>SMOC2</i>	0	Wilkinson <i>et al.</i> (2013)
RF MDA	1	DRGA0000162	14280903	ENSSSCG000000025777	<i>ESR1</i>	0	Wilkinson <i>et al.</i> (2013), Li <i>et al.</i> (2013), Yang <i>et al.</i> (2017)
RF MDA	1	H3GA0001444	35058376	ENSSSCG00000004209	<i>PTPRK</i>	0	Wang <i>et al.</i> (2018)
RF MDA	1	ALGA0009192	251437923	ENSSSCG00000005455	<i>SVEP1</i>	0	Li <i>et al.</i> (2013), Wang <i>et al.</i> (2018)
Delta, Fst	3	MARC0043512	5129774	ENSSSCG00000007590	<i>PMS2</i>	0	Li <i>et al.</i> (2013)
PCA	3	ALGA0019028	53207188	ENSSSCG00000008169	<i>TBC1D8</i>	0	Li <i>et al.</i> (2013)
Fst	4	H3GA0011590	6538154	ENSSSCG00000005941	<i>KHDRBS3</i>	91381	Yang <i>et al.</i> (2014)
RF GI1, RF GI2, Fst	4	H3GA0013086	75531190	ENSSSCG00000006243	<i>PENK</i>	0	Li <i>et al.</i> (2013)
RF MDA	4	MARC0073404	107197826	ENSSSCG00000006767	<i>MAGI3</i>	77791	Li <i>et al.</i> (2013)
PCA	4	MARC0073404	107197826	ENSSSCG00000006767	<i>MAGI3</i>	77791	Li <i>et al.</i> (2013)
RF GI1, RF GI2	4	MARC0025311	115921374	ENSSSCG00000006857	<i>COL11A1</i>	80668	Wang <i>et al.</i> (2018)
RF GI1, RF GI2	5	ALGA0033636	93710246	ENSSSCG000000035495	<i>KITLG</i>	306741	Wilkinson <i>et al.</i> (2013), Li <i>et al.</i> (2013)
Fst	7	INRA0023116	1629309	ENSSSCG000000028777	<i>MYLK4</i>	0	Schiavo <i>et al.</i> (2016)
RF GI1, RF GI2	8	H3GA0024312	12664341	ENSSSCG000000026232	<i>FAM184B</i>	0	Yang <i>et al.</i> (2014)
RF GI1, RF GI2	8	ASGA0037865	12703060	ENSSSCG000000026232	<i>FAM184B</i>	0	Yang <i>et al.</i> (2014)
RF GI1, RF GI2, Fst	8	H3GA0024318	13010923	ENSSSCG00000008748	<i>LCORL</i>	40929	Rubin <i>et al.</i> (2012), Li <i>et al.</i> (2013), Wilkinson <i>et al.</i> (2013)
Fst, PCA	8	ASGA0037875	13830411	ENSSSCG00000008748	<i>LCORL</i>	860417	Rubin <i>et al.</i> (2012), Li <i>et al.</i> (2013), Wilkinson <i>et al.</i> (2013)
Delta, PCA	8	ASGA0037899	14973392	ENSSSCG00000008749	<i>SLIT2</i>	16403	Wang <i>et al.</i> (2018)
RF GI1, RF GI2	8	H3GA0024339	15102410	ENSSSCG00000008749	<i>SLIT2</i>	0	Wang <i>et al.</i> (2018)
RF GI1, RF GI2, RF MDA	8	ALGA0048895	102209143	ENSSSCG00000009092	<i>TRPC3</i>	2197	Schiavo <i>et al.</i> (2016)
PCA	8	ALGA0049529	124763157	ENSSSCG000000029621	<i>BMPRI1B</i>	0	Li <i>et al.</i> (2013)
Delta	10	ALGA0117795	50907471	ENSSSCG000000011075	<i>KIAA1217</i>	0	Wang <i>et al.</i> (2018)
Delta	11	ASGA0051087	61168204	ENSSSCG000000031946	<i>GPC5</i>	0	Li <i>et al.</i> (2013)
RF GI1, RF GI2, Delta	13	INRA0039430	4711430	ENSSSCG000000011199	<i>TBC1D5</i>	0	Wang <i>et al.</i> (2018)
Delta	13	ASGA0058976	137020019	ENSSSCG000000027952	<i>ADCY5</i>	0	Rubin <i>et al.</i> (2012)
RF MDA	13	MARC0015751	200576767	ENSSSCG000000012059	<i>HLCS</i>	0	Schiavo <i>et al.</i> (2016)
Fst	14	ASGA0062298	25679165	ENSSSCG000000038915	<i>TMEM132D</i>	0	Ai <i>et al.</i> (2013), Li <i>et al.</i> (2013)
Delta, PCA	14	MARC0060803	25778376	ENSSSCG000000038915	<i>TMEM132D</i>	0	Ai <i>et al.</i> (2013), Li <i>et al.</i> (2013)
RF GI1, RF GI2	15	MARC0006806	24186234	ENSSSCG000000015715	<i>EN1</i>	217137	Zhang <i>et al.</i> (2018)
Delta, PCA	17	ASGA0075694	19679617	ENSSSCG00000007067	<i>JAG1</i>	49532	Li <i>et al.</i> (2013)
RF GI1, RF GI2, Delta	18	ALGA0096892	8348758	ENSSSCG000000016492	<i>AGK</i>	0	Wang <i>et al.</i> (2018)

SNPs = single nucleotide polymorphisms; RF = random forest; GI = Gini index; MDA = Mean Decrease in Accuracy; Fst = Fixation index; PCA = principal component analysis; SSC = porcine chromosome.

¹Method acronyms are defined in the notes to Tables 1 to 3.

²Position of the SNP on the chromosome (coordinate system on the Sscrofa11.1 genome version).

³Ensembl annotated gene identification.

⁴Distance in bp of the indicated gene to the SNP. When '0' is reported, the SNP is within the annotated gene.

SNPs were within annotated genes. Summarizing, 31 SNPs from the six panels marked 26 genes that have been already reported to affect production and morphological traits or that are within selection signature regions reported by other studies in pigs (Table 3). For example, four panels (i.e. the two RF GI panels, Fst and PCA panels) included one or two SNPs close to the *ligand dependent nuclear receptor corepressor like (LCORL)* gene, which has been already reported by Rubin *et al.* (2012) to be in a selection signature region of SSC8. A large number of selected SNPs were located on this

chromosome, confirming the informativity of SSC8 markers to differentiate pig breeds (Rubin *et al.*, 2012; Wilkinson *et al.*, 2013). An SNP within the *estrogen receptor 1 (ESR1)* gene was listed in the RF MDA panel. Variability in this gene was associated with the number of piglets borne and weaned per sow (Rothschild *et al.*, 1996). The *KIT ligand (KITLG)* gene was listed among those annotated for the RF GI SNP panels. Variability in this gene has been associated with different coat colour phenotypes in pigs (Wilkinson *et al.*, 2013).

Discussion

Breed informative SNPs are useful when there is the need to allocate animals and their derived products to a population. For example, brand mono-breed products that have been recently developed in several livestock species, including pork products, can be authenticated using population-informative markers (e.g. Wilkinson *et al.*, 2011; Fontanesi *et al.*, 2016). Other applications of breed informative markers span from their use in parentage testing and conservation genetic programs among several other applications (e.g. Huisman, 2017).

Several approaches have been proposed for the identification of population-informative markers which could have *pros* and *cons* compared to the approaches of this study, depending on the starting hypothesis, number of tested populations, genetic diversity among populations and technical aspects related to the computational time or to the availability of dedicated computational tools (e.g. Wilkinson *et al.*, 2011; Bertolini *et al.*, 2015 and 2018).

Single nucleotide polymorphism data from seven Italian breeds were used in this study. Italian Large White, Italian Duroc and Italian Landrace are cosmopolitan-derived heavy pig breeds under the national selection program which is aimed to select purebred animals useful to obtain crossbred terminal animals whose legs are processed for the production of Protected Designation of Origin (PDO) dry-cured hams. The four analysed autochthonous pig breeds (Apulo-Calabrese, Casertana, Cinta Senese and Nero Siciliano) produce highly appreciated niche processed or fresh pork products that are usually sold at a higher price than other commercial or undifferentiated products. Fresh pork from Cinta Senese has recently obtained the PDO label. All these breeds are slow-growing and less efficient in terms of feed conversion and reproduction performances than cosmopolitan breeds.

To design the strategy for the identification of breed informative SNPs, this study took advantage from previous works that showed that one of the main problems for the identification of fully informative SNP panels is derived by the high level of LD that is present in most livestock populations (Bertolini *et al.*, 2015). This aspect cannot be completely managed by most SNP selection and reduction strategies (based on established or more sophisticated statistics) which tend to co-select SNPs that are in high LD (based on their single informativity, despite they are not independent). A few approaches were then proposed to overcome this problem, including the use of chromosome by chromosome analyses (with limited number of markers that can be selected for each chromosome) or averaged statistics across all populations considered (Wilkinson *et al.*, 2011; Bertolini *et al.*, 2015). Moreover, Bertolini *et al.* (2018) demonstrated that using Delta, Fst and PCA preselection statistics and then including a further selection step based on RF for the final identification of a smaller SNP panel, the problem could be reduced, at least in part, but could not be completely solved. Therefore, this study started from reducing the

redundancy of the whole SNP panel from the beginning, considering only SNPs not in high LD (the prefiltering step used only SNPs having $r^2 < 0.3$ with any other SNPs, within breed). This preselection step made it possible to apply on the remaining dataset, reduced at about 1/5 of the untagged SNP panel, a machine learning approach (i.e. RF methods, based on the mean decrease in the GI and on the MDA) to directly select informative SNPs useful to discriminate pigs of seven breeds. Implementing RF on the whole SNP dataset (resulting only from a first filtering step) would be too computationally demanding, preventing any potential advantages derived from this machine learning methodology. Stability of RF selections was then assessed implementing a method based on iterations (and evaluating the frequencies by which SNPs were selected and the mean values of the ranking parameters) as already proposed for other applications of RF methodologies (e.g. Genuer *et al.*, 2015). Stability statistics ranked SNPs in terms of importance: a large fraction of the selected 96 SNPs had the same occurrence frequency or similar frequencies or values (depending on the considered method; Supplementary Material Table S4) suggesting that the prefiltering step which retained only tag SNPs stabilized the dataset, reducing the randomness in the tree construction. Then, RF selection methods were compared to the performances of other three statistics used to identify informative SNPs in breed comparative analyses (Delta, Fst and PCA) on the same reduced starting SNP dataset. Performances of the final 96 SNP panels obtained using all reductionist statistics were assessed using the same methodology. Again, RF was used for this purpose: it provided the OOB error rate and the correct prediction proportion which estimate the population assignment error rates based on the selected 96 SNP panels. This is one of the advantages of this machine learning methodology that can be applied for both selection and evaluation purposes.

Based on these statistics, all 96 SNP panels performed quite well. The correct prediction proportion for all analysed breeds in the reference dataset was 100% for the three SNP panels defined directly using RF methods and for the Delta-derived panel (Table 3). Only one Italian Large White pig out of 1968 animals of this breed was incorrectly assigned using the Fst and PCA panels. In the test dataset (which included only 10% of the animals of the whole investigated population) a few animals were wrongly assigned to another breed. In particular, a few Nero Siciliano pigs were wrongly assigned to the Italian Large White breed with four out of six panels. This incorrect assignment might reflect the high level of variability that is present in this breed that experienced in the past several admixture events from other breeds (Russo *et al.*, 2004).

It should be however clear that the most informative SNPs might change according to the breeds that are included in the marker selection procedures. Another factor affecting the choice of the informative SNP panels and thus their performances is the size of the reference populations. For the SNP selection procedures, a high number of genotyped animals could take into account the whole within population

variability. This, in turn, might reduce the possibility that a few animals are not assigned correctly due to atypical genotypes (Hulsegge *et al.*, 2013). This aspect might also need some adjustments in terms of numbers of informative SNPs that are selected (i.e. not only the same 96 final number). However, for many practical reasons it is not always possible to use large reference datasets for all considered breeds (some breeds have a small population size or are difficult or too expensive to sample and genotype; e.g. Wilkinson *et al.*, 2011). In this study, the problem related to the different numbers of genotyped animals per breed was managed in part by weighting the RF analyses on the number of animals in each breed. Moreover, the needed scalability of the genotyping tools imposes in practise 96 or multiples of 96 SNPs (Bertolini *et al.*, 2015).

Despite performances of the SNP panels were similar, it was interesting to note that there was a general low SNP overlapping between the tested approaches (excluding the RF panels). That means that different SNPs (or SNP combinations) might be able to provide the same level of informativity. It was also interesting to note that there was an even chromosome distribution of the selected SNPs that does not reflect the chromosome size. Among the porcine autosomes, SSC8 captured the highest number of SNPs in most panels. As the prefiltering step retained only tag SNPs with low level of LD, SSC8 markers were distributed along the whole chromosome (Supplementary Material Table S4). This chromosome contains several selection signature regions reported by previous studies (e.g. Rubin *et al.*, 2012; Wilkinson *et al.*, 2013; Schiavo *et al.*, 2016). The *LCORL* gene, which was identified by SNPs selected in four panels, is located on this chromosome. This gene has been implicated in mechanisms affecting BW, height, size and growth traits in humans, cattle and horses (reviewed in Takasuga, 2016). In pigs, it is located in selection signature regions which might be due to its effects on these traits (Rubin *et al.*, 2012). The compared pig breeds have clearly different size and growth performances that might capture, indirectly, SNPs close to *LCORL*. Other genes that are well known for their effects on economically relevant traits or morphological traits were captured by the selected SNPs (Table 3 and Supplementary Material Table S5). On SSC1, *ESR1* was marked by an SNP in the RF GI panels. A polymorphism in this gene has been associated with several reproduction parameters of the sows (Rothschild *et al.*, 1996). The use of this variant has been one of the first examples of marker-assisted selection in the pig breeding industry to improve reproduction performances. Commercial and autochthonous pig breeds have extreme reproduction efficiency that might have driven the identification of an *ESR1* gene marker. A gene affecting coat colour (*KITLG*) was identified by a marker on SSC5. Variability in this gene has been suggested to be involved in the Berkshire breed coat colour phenotype (Wilkinson *et al.*, 2013). As none of the investigated breeds have a similar phenotype, it could be possible that some variants might be needed to express other coat colour patterns (i.e. solid or belted). Four panels (i.e. RF


MDA, Delta, Fst and PCA panels) included an SNP within the *adenylate cyclase 8 (ADCY8)* gene. This gene was significantly associated with cholesterol blood content in Italian Large White pigs (Bovo *et al.*, 2019).

Among the different approaches used, RF methods captured, on the whole, 16 SNPs that have been already reported to be close or within genes located in selection signature regions, performing quite well compared to other methods that are traditionally used to identify selective sweep regions (i.e. Delta and Fst). It is worth mentioning that the purpose of this study was not that of detecting selection signature regions in the genome of the investigated breeds, but it seems evident that the adopted methodologies could reach this goal.

This study is a further step forward on the application of RF for the identification of population-informative markers derived by high-throughput genotyping platforms. Random forest selection procedures highlighted informative SNPs that were not the same as those identified by other methods (Delta, Fst and PCA). This might be due to specific features of this machine learning methodology and, particularly, to its ability in discovering and exploiting information about SNP interdependences and interactions, which can lead to substantial improvements in case of non-linear class boundaries. It will be interesting to explore if the adaptation of RF methods for the identification of selection signature regions could be able to describe population-specific features that are not captured by other approaches.

Acknowledgements

We thank ConSDABI for the collaboration in sampling Casertana pigs. This study is associated with the PSRN SUIIS project and it received funding from the Italian Ministry of Agriculture, Food and Forestry (MiPAAF) under the project INNOVAGEN and from the University of Bologna RFO 2018 program.

 L. Fontanesi 0000-0001-7050-3760

Declaration of interest

The authors declare that they do not have competing interests.

Ethics statement

No ethical approval was required since only genotyping data were used in the study and data were provided by the research program INNOVAGEN (Italian Ministry of Agriculture, Food and Forestry).

Software and data repository resources

None of the data were deposited in an official repository.

Supplementary material

To view supplementary material for this article, please visit <https://doi.org/10.1017/S1751731119002167>

References

- Ai H, Huang L and Ren J. 2013. Genetic diversity, linkage disequilibrium and selection signatures in Chinese and Western pigs revealed by genome-wide SNP markers. *PLoS ONE* 8, e56001.
- ANAS 2018. Registro Anagrafico. Retrieved on 10 December 2018 from <http://www.anas.it/>
- 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 2018. Preselection statistics and Random Forest classification identify population informative single nucleotide polymorphisms in cosmopolitan and autochthonous cattle breeds. *Animal* 12, 12–19.
- Bovo S, Mazzoni G, Bertolini F, Schiavo G, Galimberti G, Gallo M, Dall'Olio S and Fontanesi L 2019. Genome-wide association studies for 30 haematological and blood clinical-biochemical traits in Large White pigs reveal genomic regions affecting intermediate phenotypes. *Scientific Reports* 9, 7003.
- Breiman L 2001. Random forests. *Machine Learning* 45, 5–32.
- Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM and Lee JJ 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 4, s13742–015–0047–8.
- Fontanesi L, Scotti E, Gallo M, Nanni Costa L and Dall'Olio S 2016. Authentication of “mono-breed” pork products: identification of a coat colour gene marker in Cinta Senese pigs useful to this purpose. *Livestock Science* 184, 71–77.
- Genuer R, Poggi J-M and Tuleau-Malot C 2015. VSURF: an R package for variable selection using random forests. *The R Journal* 7/2, 19–33.
- Hastie T, Tibshirani R and Friedman JH 2009. *The elements of statistical learning*, 2nd edition. Springer, New York, NY, USA.
- Huisman J 2017. Pedigree reconstruction from SNP data: parentage assignment, sibship clustering and beyond. *Molecular Ecology Resources* 17, 1009–1024.
- Hulsegge B, Calus MP, Windig JJ, Hoving-Bolink AH, Eijndhoven MH and Hiemstra SJ 2013. Selection of SNPs from 50K and 777K arrays to predict breed-of-origin in cattle. *Journal of Animal Science* 91, 5128–5134.
- Jacobs A, De Noia M, Praebel K, Kanstad-Hanssen Ø, Paterno M, Jackson D, McGinnity P, Sturm A, Elmer KR and Llewellyn MS 2018. Genetic fingerprinting of salmon louse (*Lepeophtheirus salmonis*) populations in the North-East Atlantic using a random forest classification approach. *Scientific Reports* 8, 1203.
- Jolliffe IT and Cadima J 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A* 374, 20150202.
- Kijas JW, Serrano M, McCulloch R, Li Y, Salces Ortiz J, Calvo JH, Pérez-Guzmán MD and International Sheep Genomics Consortium 2013. Genome wide association for a dominant pigmentation gene in sheep. *Journal of Animal Breeding and Genetics* 130, 468–475.
- Li M, Tian S, Jin L, Zhou G, Li Y, Zhang Y, Wang T, Yeung CK, Chen L, Ma J, Zhang J, Jiang A, Li J, Zhou C, Zhang J, Liu Y, Sun X, Zhao H, Niu Z, Lou P, Xian L, Shen X, Liu S, Zhang S, Zhang M, Zhu L, Shuai S, Bai L, Tang G, Liu H, Jiang Y, Mai M, Xiao J, Wang X, Zhou Q, Wang Z, Stothard P, Xue M, Gao X, Luo Z, Gu Y, Zhu H, Hu X, Zhao Y, Plastow GS, Wang J, Jiang Z, Li K, Li N, Li X and Li R 2013. Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. *Nature Genetics* 45, 1431–1438.
- Liaw A and Wiener M 2002. Classification and regression by random forest. *R News* 2, 18–22.
- Ligges U and Mächler M 2013. Scatterplot3d - an R package for visualizing multi-variate data. *Journal of Statistical Software* 8, 1–20.
- Meng YA, Yu Y, Cupples LA, Farrer LA and Lunetta KL 2009. Performance of random forest when SNPs are in linkage disequilibrium. *BMC Bioinformatics* 10, 78.
- Naderi S, Yin T and König S 2016. Random forest estimation of genomic breeding values for disease susceptibility over different disease incidences and genomic architectures in simulated cow calibration groups. *Journal of Dairy Science* 99, 7261–7273.
- Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, Mahoney MW and Drineas P 2007. PCA-correlated SNPs for structure identification in world-wide human populations. *PLoS Genetics* 9, 1672–1686.
- Rothschild M, Jacobson C, Vaske D, Tuggle C, Wang L, Short T, Eckardt G, Sasaki S, Vincent A, McLaren D, Southwood O, van der Steen H, Mileham A and Plastow G 1996. The estrogen receptor locus is associated with a major gene influencing litter size in pigs. *Proceedings of the National Academy of Sciences of the USA* 93, 201–205.
- Rubin CJ, Megens HJ., Martinez Barrio A, Maqbool K, Sayyab S, Schwochow D, Wang C, Carlborg Ö, Jern P, Jørgensen CB, Archibald AL, Fredholm M, Groenen MA and Andersson L 2012. Strong signatures of selection in the domestic pig genome. *Proceedings of the National Academy of Sciences of the USA* 109, 19529–19536.
- Russo V, Fontanesi L, Davoli R, Chiofalo L, Liotta L and Zumbo A 2004. Analysis of single nucleotide polymorphisms in major and candidate genes for production traits in Nero Siciliano pig breed. *Italian Journal of Animal Science* 3, 19–29.
- Schiavo G, Galimberti G, Calò DG, Samorè AB, Bertolini F, Russo V, Gallo M, Buttazzoni L and Fontanesi L 2016. Twenty years of artificial directional selection have shaped the genome of the Italian Large White pig breed. *Animal Genetics* 47, 181–191.
- Takasuga A 2016. PLAG1 and NCAPG-LCORL in livestock. *Animal Science Journal* 87, 159–167.
- Wang K, Wu P, Yang Q, Chen D, Zhou J, Jiang A, Ma J, Tang Q, Xiao W, Jiang Y, Zhu L, Li X and Tang G 2018. Detection of selection signatures in Chinese Landrace and Yorkshire pigs based on genotyping-by-sequencing data. *Frontiers in Genetics* 9, 119.
- Weir BS and Cockerham CC 1984. Estimating F-statistics for the analysis of population structure. *Evolution* 38, 1358–1370.
- Wilkinson S, Archibald AL, Haley CS, Megens HJ, Crooijmans RP, Groenen MA, Wiener P and Ogden R 2012. Development of a genetic tool for product regulation in the diverse British pig breed market. *BMC Genomics* 13, 580.
- Wilkinson S, Lu ZH, Megens HJ, Archibald AL, Haley C, Jackson IJ, Groenen MA, Crooijmans RP, Ogden R and Wiener P 2013. Signatures of diversifying selection in European pig breeds. *PLoS Genetics* 9, e1003453.
- 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.
- Yang B, Cui L, Perez-Enciso M, Traspov A, Crooijmans RPMA, Zinovieva N, Schook LB, Archibald A, Gatphayak K, Knorr C, Triantafyllidis A, Alexandri P, Semiadi G, Hanotte O, Dias D, Dovč P, Uimari P, Iacolina L, Scandura M, Groenen MAM, Huang L and Megens HJ 2017. Genome-wide SNP data unveils the globalization of domesticated pigs. *Genetics Selection Evolution* 49, 71.
- Yang S, Li X, Li K, Fan B and Tang Z 2014. A genome-wide scan for signatures of selection in Chinese indigenous and commercial pig breeds. *BMC Genetics* 15, 7.
- Zhang Z, Xiao Q, Zhang QQ, Sun H, Chen JC, Li ZC, Xue M, Ma PP, Yang HJ, Xu NY, Wang QS and Pan YC 2018. Genomic analysis reveals genes affecting distinct phenotypes among different Chinese and western pig breeds. *Scientific Reports* 8, 13352.