RESEARCH ARTICLE

A N I M A L   G E N E T I C S WILEY

Identification of population-informative markers from high-density genotyping data through combined feature selection and machine learning algorithms: Application to European autochthonous and cosmopolitan pig breeds

Giuseppina Schiavo[1] | Francesca Bertolini[1] | Samuele Bovo[1] | Giuliano Galimberti[2] | María Muñoz[3] | Riccardo Bozzi[4] | Marjeta Čandek-Potokar[5] | Cristina Óvilo[3] | Luca Fontanesi[1]

[1]Animal and Food Genomics Group, Division of Animal Sciences, Department of Agricultural and Food Sciences, University of Bologna, Bologna, Italy

[2]Department of Statistical Sciences 'Paolo Fortunati', University of Bologna, Bologna, Italy

[3]Departamento Mejora Genética Animal, INIA-CSIC, Madrid, Spain

[4]Animal Science Division, Dipartimento di Scienze e Tecnologie Agrarie, Alimentari, Ambientali e Forestali, Università di Firenze, Firenze, Italy

[5]Kmetijski Inštitut Slovenije, Ljubljana, Slovenia

Correspondence
Luca Fontanesi, Animal and Food Genomics Group, Division of Animal Sciences, Department of Agricultural and Food Sciences, University of Bologna, Viale G. Fanin 46, 40127 Bologna, Italy.
Email: luca.fontanesi@unibo.it

Abstract

Large genotyping datasets, obtained from high-density single nucleotide polymorphism (SNP) arrays, developed for different livestock species, can be used to describe and differentiate breeds or populations. To identify the most discriminating genetic markers among thousands of genotyped SNPs, a few statistical approaches have been proposed. In this study, we applied the Boruta algorithm, a wrapper of the machine learning random forest algorithm, on a database of 23 European pig breeds (20 autochthonous and three cosmopolitan breeds) genotyped with a 70k SNP chip, to pre-select informative SNPs. To identify different sets of SNPs, these pre-selected markers were then ranked with random forest based on their mean decrease accuracy and mean decrease gene indexes. We evaluated the efficiency of these subsets for breed classification and the usefulness of this approach to detect candidate genes affecting breed-specific phenotypes and relevant production traits that might differ among breeds. The lowest overall classification error (2.3%) was reached with a subpanel including only 398 SNPs (ranked based on their mean decrease accuracy), with no classification error in seven breeds using up to 49 SNPs. Several SNPs of these selected subpanels were in genomic regions in which previous studies had identified signatures of selection or genes associated with morphological or production traits that distinguish the analysed breeds. Therefore, even if these approaches have not been originally designed to identify signatures of selection, the obtained results showed that they could potentially be useful for this purpose.

K E Y W O R D S
genome, population genomics, random forest, signatures of selection, SNP, *Sus scrofa*

INTRODUCTION

Whole genome genotyping datasets using commercial and customised single nucleotide polymorphism (SNP) arrays developed for many different species, including all major livestock animals, can be used to describe, at an unprecedented level, population genomic features that are useful for many downstream genetic applications and novel discoveries. For example, high-density SNP data provide new tools to evaluate and monitor the inbreeding level in livestock populations and infer their genetic history and past genetic events (e.g. Schiavo et al., 2021, 2022). Population structure and signatures of selection can be detected by exploiting these datasets with a variety of statistical methodologies and comparative analyses (Dadousis et al., 2022; Mulim et al., 2022; Muñoz et al., 2019). For other applications, these high-density SNP panels provide quite redundant genotyping information. Therefore, preselected subsets of informative SNPs have been proposed for breed assignment of individuals, estimation of breed proportion in crossbred animals, authentication of the breed of origin of breed-branded products, parentage verification and several other evaluations and analyses (Bertolini et al., 2015; Gebrehiwot et al., 2021; Muñoz et al., 2020; Wilkinson et al., 2011, 2012). The identification of informative SNPs for breed assignment can also highlight genomic regions under selection or containing relevant genes involved in determining breed-specific traits (Bertolini et al., 2018; Schiavo et al., 2020).

Different statistical approaches and measures have been used to identify the most informative and discriminating SNPs across few or many breeds by using thousands of markers included in the commercial arrays. One of the simplest methods that has been applied for this purpose relies on the absolute allele frequency difference at each SNPs obtained in the population pairwise comparisons, summarised in the Delta values (Wilkinson et al., 2012). Another frequently used statistic for the identification of breed informative SNPs as well as signatures of selection and population structures is the F_{ST}, the fixation index, which returns the standardised variance in allele frequencies among pairs of populations (Hulsegge et al., 2013; Wilkinson et al., 2011). Principal component analysis, an unsupervised linear technique for dimension reduction that allows to extract axes of maximal variation from datasets (Jollife & Cadima, 2016), has been extensively used to describe population structures and then to reduce dimensionality of high-density SNP datasets and identify breed discriminant markers (Bertolini et al., 2015; Paschou et al., 2007; Wilkinson et al., 2011).

Machine learning approaches have recently been applied in this context by combining feature selection and classification techniques to assign an unknown sample (e.g. an animal) to one of the pre-determined groups (e.g. breeds) using reduced and selected SNP datasets and identifying discriminant SNPs (Bertolini et al., 2015, 2018; Liu et al., 2022; Pasupa et al., 2020; Schiavo et al., 2020). Among the machine learning techniques, random forest (RF) is an ensemble technique that derives prediction rules by combining multiple 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 pre-determined groups. Recently, several authors have tested some RF-based approaches to identify breed informative SNPs in cattle and pig breeds (Bertolini et al., 2015, 2018; Gao et al., 2022; Schiavo et al., 2020). As RF is prone to being biased by high linkage disequilibrium between markers, it might become computationally very demanding when using thousands of markers (Meng et al., 2009). Therefore, it is a common practice to reduce the dataset complexity by reducing the number of variables, in this case, the number of markers. There are, however, no defined rules and guidelines to proceed in this direction and the usual strategies are therefore designed to test performance by applying different approaches and evaluating the final performances of RF (Bertolini et al., 2015, 2018; Schiavo et al., 2020).

The Boruta algorithm is an RF wrapper (Kursa, Jankowski et al., 2010). As RF can be applied to estimate the importance of each feature (SNP) in the classification and ranks all the features in order of importance, Boruta reinforces the estimated statistical importance of the features. This is done by iterating the RF analysis with real and shadow features, labelling the real features as 'Confirmed', 'Tentative' or 'Rejected', depending on their ability to discriminate classes when compared with shadow features. Boruta implements a lighter procedure for decision trees, thus the analysis on large datasets with this wrapper may require lower computational efforts than doing the same directly with RF. However, Boruta does not provide a ranking of the features but only a qualitative value. The combined use of Boruta wrapper and simple RF allows statistically stable results to be obtained (with Boruta) that are also associated with a ranking (RF) and are useful to explore the final classification error of subsets with different sizes by tuning a variable number of informative SNPs based on their ranking.

Boruta-based reduction and classification in livestock has already been implemented in other fields, including for example, image analyses and the classification of behavioural data (Kleanthous et al., 2018), and in the selection of fatty acids with a predictive function for the diagnosis of ketosis in cattle (Fiore et al., 2020). Recently, Boruta has been also tested for the detection of informative SNPs useful for the classification of four pig breeds (Hayah et al., 2021). In several contexts, Boruta has been shown to be one of the most stable and efficient methodologies in panels with high complexity, if compared

with other reduction approaches (Acharjee et al., 2020; Speiser et al., 2019).

The aim of this study was to apply Boruta algorithm and RF approaches to identify and rank different sets of breed-informative SNPs. The combination of Boruta algorithm and RF was tested using high-density SNP datasets obtained from pigs of 23 different breeds. For the selected SNP datasets, the efficiency on the animal classification (i.e. the assignment of an animal to its breed) was evaluated. In addition, the usefulness of these combined approaches to mark genomic regions containing candidate genes affecting breed-specific phenotypes and relevant production traits that might differ among breeds was also evaluated.

MATERIALS AND METHODS

Pig breeds and SNP datasets

The study included a total of 1131 pigs from 20 autochthonous and three cosmopolitan-derived breeds from nine European countries (39–53 pigs for each breed; Table S1): two breeds from Portugal (Alentejana and Bísara); two from Spain (Iberian and Majorcan Black); two from France (Basque and Gascon); six autochthonous (Apulo-Calabrese, Casertana, Cinta Senese, Mora Romagnola, Nero Siciliano and Sarda) and three cosmopolitan-derived breeds (Italian Large White, Italian Landrace and Italian Duroc) from Italy; one from Slovenia (Krškopolje pig, hereafter referred to as Krškopolje); two from Croatia (Black Slavonian and Turopolje); two from Serbia (Moravka and Swallow-Bellied Mangalitsa); one from Germany (Schwäbisch-Hällisches Schwein); two from Lithuania (Lithuanian indigenous wattle and Lithuanian White old type). Selection of the pigs for genotyping was performed so as to avoid highly related animals (no full- or half-sibs), when possible, by balancing between sexes and prioritising adult individuals or, at least, animals with the morphology of an adult. All pigs had standard characteristics of their corresponding breed and were registered in their respective herd books. More information of the investigated pig breeds is reported in Bovo et al. (2020) and in Table S1.

Genotyping and multidimensional scaling analysis

Blood samples were obtained during a general breeding procedure and reused for this work. No animal experiments were performed for this research. DNA was extracted from leukocytes as described by Muñoz et al. (2018). Animals were genotyped with GGP Porcine HD Genomic Profiler following the producers' protocols. The genotyping data have been checked for quality with the software PLINK1.9 (Chang et al., 2015). For each breed, SNPs with call rate >0.9 and Hardy–Weinberg equilibrium P>0.0001 were retained and animals with individual call rate <0.90 were excluded. SNPs have not been filtered for low minor allele frequency (MAF), to consider alleles that could be fixed in few breeds. Only SNPs with MAF equal to zero were removed, considering all breeds together (fixed in the whole dataset). Retained SNPs with missing genotypes were randomly imputed within each breed according to the corresponding genotype frequency with an in-house script used in a previous work (Bertolini et al., 2015).

Multidimensional scaling was calculated for three dimensions using the --cluster function of the software PLINK1.9 (Chang et al., 2015).

Boruta and random forest

Boruta

Boruta analysis consisted of several iterative applications of RF, each obtained by adding to the real features (the SNPs) some shadow features artificially created by randomly permuting the observed ones. At the end of the iterations, the estimated importance of real and shadow features was compared. Features that emerged from the comparison, were labelled as 'Confirmed', 'Tentative' and 'Rejected'. Real features whose estimated importance was less than the estimated importance of one or more shadow features were labelled 'Rejected'. The label 'Tentative' meant that the estimated importance of a real feature was comparable with that of the shadow features and that the number of iterations was not sufficient to reach a conclusion. A 'Confirmed' feature showed an estimated importance that was always better than the estimated importance of any shadow feature.

The Boruta algorithm, implemented in the R package 'Boruta' (Kursa & Rudnicki, 2010, R Core Team, 2021) was applied to the whole filtered dataset for two subsequent actions:

1. Boruta was first run independently on each chromosome with default parameters (namely, a confidence interval of 0.01 and multiple comparisons adjustment of p using the Bonferroni method). The number of iterations was set to 1000. This number allowed the resolution of a higher number of 'Tentative' labels with respect to the default value of 100. For each independent run, all SNPs that were labelled as 'Confirmed' by the algorithm were kept and merged to create a filtered SNP panel, whereas SNPs that were labelled as 'Tentative' or 'Rejected' were not taken into consideration for any subsequent steps.
2. Boruta was then run again using the filtered SNP panel defined above with the same parameters, and the SNPs that were labelled again as 'Confirmed' were
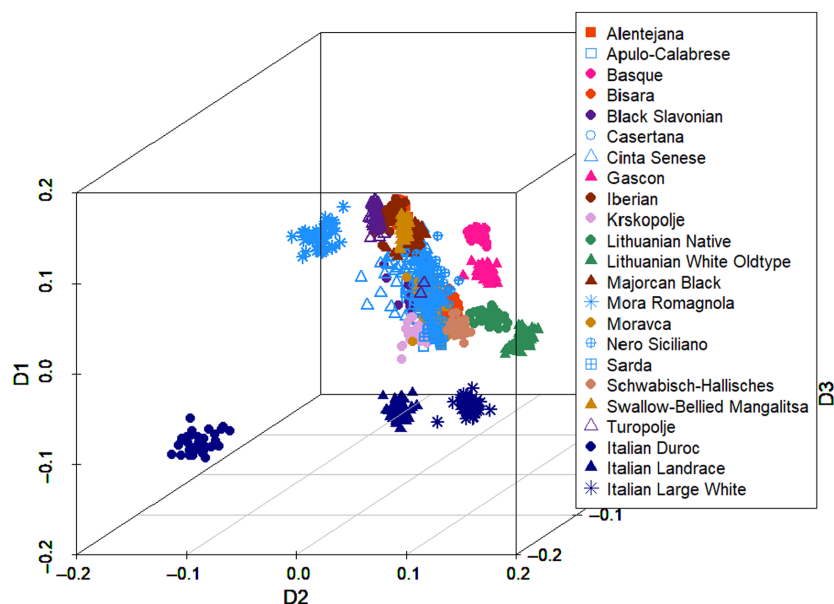
FIGURE 1 Multidimensional scaling plot of the breeds included in the analyses. Each cluster of colour represents a country (purple, Croatia; dark pink, France; light brown, Germany; light blue, Italy, autochthonous; dark blue, Italy, cosmopolitan; green, Lithuania; orange, Portugal; yellow, Serbia; dark brown, Spain; light pink, Slovenia).

retained, whereas the SNPs that resulted as 'Tentative' or 'Rejected' in this second runs were discarded. This second step made possible to further reduce the number of variables and identify a more robust panel of informative SNPs.

Random forest

The SNPs included in this reduced SNP panel selected with Boruta were analysed with the standard RF algorithm implemented in the R package 'randomForests' (Breiman, 2001). The analysis was run with default parameters and iterations. Out Of Bag (OOB) classification error estimates were considered to measure the ability of the SNP panels to correctly assign each animal to its breed. This error estimation, included in RF algorithm and consequently in the Boruta algorithm, is an efficient alternative to cross-validation methods and it allows evaluation of the goodness of classification without the need to set any leave-one-out approaches. Here, for each SNP of the reduced (pre-selected) SNP panel, the mean decrease Gini (MDG) and mean decrease accuracy (MDA) were calculated. These values were then used to rank the SNPs of the reduced SNP panel based on their contribution to the obtained classification: the higher the value, the higher the importance of the variable (i.e. SNP) in the model. These ranking parameters were used to define five subsets of SNPs based on MDG and five based on MDA classifications (namely panels $N/2$, $N/4$, $N/8$, $N/16$ and $N/32$), with the SNP number (N) that was subsequently halved in each panel (therefore dividing N by 2, 4, 8, 16 and 32, and rounding the number), taking the top ranked SNPs in each reduction step. These SNP subsets were used for independent RF analyses with default parameters and iteration but including the 'classwt' option that corrects for the different number of animals

available per breed. Out Of Bag and classification error estimates were then retrieved again.

Annotation of selected single nucleotide polymorphisms

All genes ±500 kb near the SNPs that composed the whole reduced SNP panel were retrieved from the Sscrofa11.1 genome annotation available in NCBI (https://www.ncbi.nlm.nih.gov/) and ENSEMBL (https://www.ensembl.org/index.html) databases using BEDTOOLS software v2.30 (Quinlan & Hall, 2010) and used for comparative analyses with existing literature. Then, genes annotated ±100 kb from the SNPs included in some SNP panels (i.e. $N/2$ MDG and $N/4$ MDA panels) were used for gene enrichment analysis. Two gene enrichment analyses were carried out: one analysis was carried out with R package enrichR (Chen et al., 2013) by interrogating the GWAS catalogue, a comprehensive database of relationships between human phenotypes and genes (Buniello et al., 2019).

RESULTS

Genotyping and multidimensional scaling analysis

The filtering step across all breed datasets retained on average 55 277 SNPs, ranging from 55 087 for the Turopolje to 56 552 SNPs for the Italian Large White breeds. Among the filtered SNPs, 54 797 were commonly present in all breeds and a final number of 52 542 SNPs, of which 48 544 were autosomal, had a total MAF >0. All 1131 animals passed the quality threshold (Table S1). Multidimensional scaling analysis

(Figure 1) showed that individual pigs were generally grouped according to their breed. Considering a second level of information, most autochthonous breeds clustered close to other autochthonous breeds from the same country. The three cosmopolitan breeds (i.e. Italian Large White, Italian Landrace and Italian Duroc) clustered separately from the rest of the autochthonous breeds (Figure 1).

Boruta filtering

The labelling of the SNPs identified after the first round of Boruta analysis is shown in Figure 2a. From the 48 544 autosomal SNPs, a total of 28 713 SNPs were labelled as 'Confirmed', 4241 as 'Tentative' and 19 590 as 'Rejected'. Each chromosome contributed with a number of 'Confirmed' SNPs that ranged from 1193 (SSC18) to 1781 (SSC1), in relation with the chromosome size and the total number of filtered SNPs located on each chromosome (Pearson's correlation, $r=0.83$, $p<0.01$). However, these were also the chromosomes with the highest number of 'Rejected' or 'Tentative' SNPs (SSC1: 3004 and 491, respectively) and the lowest number of 'Rejected' and 'Tentative' (SSC18: 235 and 31, respectively), again in relation to their size and total number of starting SNPs assigned to these chromosomes. The second analysis with Boruta, that was performed only on the 28 713

'Confirmed' SNPs coming from the first analysis, further reduced the number of SNPs by retaining only the SNPs labelled as 'Confirmed', which were in total 1595. Here, less than 10% of the SNPs in each chromosome was retained and more than 80% of the SNPs were rejected (Figure 2b) or labelled as 'Tentative'. Again, correlation between the retained number of SNPs and the length of the chromosomes was very high ($r=0.93$). The list of SNPs that constituted the reduced panel is reported in Table S2.

Random forest classification

Random forest analyses were run with the 1595 SNPs that constituted the reduced panel derived from the Boruta steps and with the subsets of SNPs that were subsequently identified to test the lower number of SNPs to assign the pigs to the correct breed.

Considering the 1595 SNP panel, RF assigned the pigs to their correct breeds with an overall OOB classification error of 2.39% (Figure 3). Among the different breeds, 16 out of 23 had all animals correctly classified with this panel (classification error=0; Table 1). Breeds with a few misclassified animals were Black Slavonian, Cinta Senese, Krskopolje, Lithuanian White Old Type, Moravka, Nero Siciliano and Sarda (Table 1). Sarda
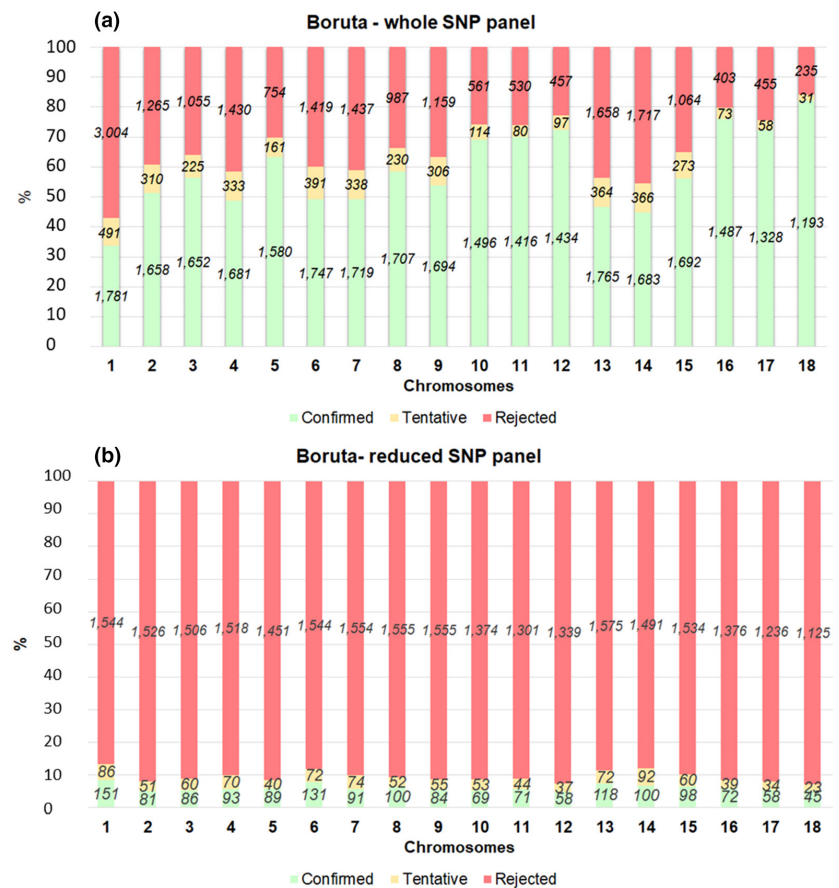

FIGURE 2 Distribution of the Boruta labelled SNPs across the different chromosomes: 'Confirmed' (green), 'Tentative' (yellow) and 'Rejected' (red). (a) The SNP panel derived from the first run of Boruta. (b) The reduced SNP panel derived from the second run of Boruta.
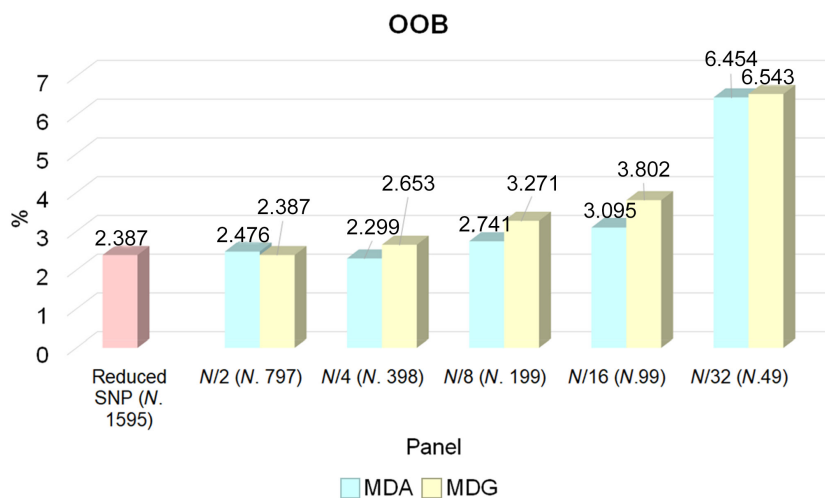
FIGURE 3 Out Of Bag for the reduced SNP panel and of the subsets that were defined by subsequently halving the number of SNPs (N) including the SNPs ranked based on their mean decrease accuracy (MDA) and mean decrease Gini (MDG). For each SNP panel, the number of SNPs is indicated in parentheses.

had the highest classification error (0.143) with this SNP panel.

We then tested the performances of the five SNP subsets (derived by subsequently dividing the number of SNPs of the reduced panel obtained from the Boruta steps) based on MDA ranking values and the performances of the five SNP subsets based on MDG ranking values. The OOB classification errors were almost equal or very close to the value obtained from the 1595 SNP panel when the numbers of SNPs were 797 (N/2 panels) and 398 (N/4 panels) for the MDA and MDG ranking methods (Figure 3). The MDA ranking had the lowest OOB classification error (2.30%) with the N/4 SNP panel whereas the MDG ranking had the lowest OOB classification error (2.39%) with the N/2 SNP panel. The OOB classification error increased progressively for both MDA and MDG panels when the number of SNPs was subsequently reduced. Both N/32 MDA and MDG panels, which contained the lowest number of tested SNPs, were the subsets with the highest OOB classification errors (Figure 3).

From the general overview of the classification error defined for each breed and reported in Table 1, it is worth mentioning that this parameter remained always equal to zero with all different panels (based on different numbers of SNPs and with both ranking systems) in six autochthonous breeds (Basque, Gascon, Majorcan Black, Mora Romagnola, Schwabisch-Hallisches and Swallow-Bellied Mangalitsa) and the Italian Duroc breed. In one breed, the Casertana, the classification error was not equal to zero only in the case of N/32 panels. For Turopolje, the classification error was the same (0.04) across all SNP panels, from the largest to the smallest. For several other breeds, the classification error increased with decreasing number of SNPs in the panels, reaching the highest values of classification errors in Sarda breed with the N/32 panel (0.27 for the MDA ranking and 0.35 for the MDG ranking). This trend was not always consistent as in a few breeds, the SNP panels with higher numbers of SNPs had higher classification errors than

those observed for some SNP panels with lower numbers of markers: for example, in the Lithuanian Native, the MDA N/4, N/8 SNP and N/32 panels had a classification error of 0.02 whereas the MDA N/2 and N/16 had a classification error equal to zero (Table 1).

It was also interesting to check the pigs that were wrongly assigned by the different subpanels and the MDA and MDG ranking methods used (Tables S3–S13). For a quick overview of the classification errors of the panels (i.e. N/4 MDA and N/32 MDG panels) with lower OOB values, Figure 4 reports the number of wrongly assigned animals for each breed. For example, a few Cinta Senese pigs were misclassified as Black Slavonian (and vice versa). With the panel N/32, Sarda pigs were misclassified to several breeds, including Bísara, Italian Landrace, Nero Siciliano and Schwäbisch-Hällisches Schwein (these are the breeds in which at least two Sarda pigs were misplaced; Tables S8 and S13).

Annotation of the top-ranked SNPs and signatures of selection

The complete list of annotated genes ±500 kb near the markers that composed the 1595 SNP panel is reported in Table S14. Some of the genes are known to affect economically relevant traits, including performance and morphological traits, or are included in genomic regions where signatures of selection have already been reported in pigs (Bovo et al., 2020; Rubin et al., 2012; Schiavo et al., 2021). The top-ranked SNP for MDG was located on SSC5 within the *methionine sulfoxide reductase B3* (*MSRB3*) gene, which has been shown to affect ear shape and ear size in pigs (Chen et al., 2013; Zhang et al., 2015). Other SNPs mark genes known to affect the body size (e.g. *PLAG1 zinc finger*, *PLAG1*, on SSC4), which have already been reported in genomic regions harbouring signatures of selection in many pig breeds. Some genes close to the selected SNPs have been associated with growth performance traits, e.g. *leptin* (*LEP*) on SSC18 and

TABLE 1 Classification error of the different breeds utilising the reduced SNP panel ($N=1595$ SNPs) and the subpanels $N/2$ ($N=797$), $N/4$ ($N=398$), $N/8$ ($N=199$), $N/16$ ($N=99$) and $N/32$ ($N=49$) that were defined using the ranking derived from the mean decrease accuracy (MDA) and mean decrease Gini (MDG) values of the SNPs.

Breed	Reduced SNP panel (N)	MDA $N/2$	MDA $N/4$	MDA $N/8$	MDA $N/16$	MDA $N/32$	MDG $N/2$	MDG $N/4$	MDG $N/8$	MDG $N/16$	MDG $N/32$
Alentejana	0.000	0.000	0.000	0.000	0.062	0.125	0.000	0.000	0.021	0.042	0.125
Apulo-Calabrese	0.000	0.000	0.000	0.000	0.038	0.075	0.000	0.019	0.000	0.038	0.057
Basque	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Bísara	0.000	0.000	0.020	0.000	0.000	0.061	0.000	0.020	0.000	0.020	0.020
Black Slavonian	0.122	0.122	0.122	0.143	0.143	0.204	0.122	0.122	0.143	0.143	0.184
Casertana	0.000	0.000	0.000	0.000	0.000	0.018	0.000	0.000	0.000	0.000	0.018
Cinta Senese	0.074	0.074	0.074	0.074	0.074	0.074	0.074	0.074	0.074	0.074	0.093
Gascon	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Iberian	0.000	0.000	0.000	0.063	0.042	0.125	0.000	0.000	0.063	0.042	0.125
Krškopolje	0.019	0.000	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.019	0.058
Lithuanian Indigenous Wattle	0.000	0.000	0.021	0.021	0.000	0.021	0.021	0.000	0.000	0.000	0.021
Lithuanian White Old Type	0.021	0.021	0.021	0.021	0.021	0.042	0.021	0.021	0.021	0.021	0.042
Majorcan Black	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Mora Romagnola	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Moravka	0.080	0.100	0.040	0.060	0.060	0.160	0.080	0.100	0.08	0.08	0.140
Nero Siciliano	0.040	0.020	0.020	0.040	0.040	0.180	0.020	0.040	0.120	0.060	0.120
Sarda	0.143	0.184	0.143	0.143	0.163	0.265	0.102	0.143	0.163	0.224	0.347
Schwäbisch-Hällisches Schwein	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Swallow-bellied Mangalitsa	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Turopolje	0.040	0.040	0.040	0.040	0.040	0.040	0.040	0.040	0.040	0.040	0.040
Italian Duroc	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Italian Landrace	0.000	0.000	0.000	0.000	0.000	0.021	0.000	0.000	0.000	0.021	0.042
Italian Large White	0.000	0.000	0.000	0.000	0.000	0.042	0.000	0.042	0.000	0.042	0.062

Note: The number of SNPs included in the different panels is reported in the legend.

growth hormone releasing hormone (*GHRH*) on SSC17 (De Oliveira Peixoto et al., 2006; Franco et al., 2005; Kennes et al., 2001; Pérez-Montarelo et al., 2012), and reproduction traits (including total number of piglets born), e.g. *Kruppel like factor 3* (*KLF3*) on SSC8 (Wang et al., 2022). Other examples derive from marked genes that have been associated with carcass and meat quality traits: two have been associated with boar taint, namely *CTD small phosphatase 2* (*CTDSP2*) on SSC1 (Botelho et al., 2022) and *hydroxysteroid 17-beta dehydrogenase 13* (*HSD17B13*) on SSC8 (Moe et al., 2008); *EPH receptor A3* (*EPHA3*), on SSC13, has been associated with ham weight loss at first salting (Fontanesi et al., 2017); carnitine O-acetyltransferase (*CRAT*), on SSC1, has been associated with backfat thickness and lipid metabolism (Casiró et al., 2017; Pena et al., 2013).

Carboxypeptidase E (*CPE*) gene on SSC8, which is close to the top-ranked SNP for MDA, is in a signature of selection region, previously identified by Bovo et al. (2020) in some of the investigated autochthonous pig breeds. When matching the filtered SNPs from Boruta with the selection signatures identified in our previous study with the same breeds (Bovo et al., 2020), the selected SNPs overlapped with 18 other selection sweep regions identified in these breeds (Table S15). Five were located in genomic regions with signatures of selection that resulted from the comparison of groups of breeds with different body size (Table S15). Some SNPs were located in regions that have been previously detected in genomic data comparison between belted and spotted pigs. Other SNPs are in signatures of selection that emerged in one breed. For example, an SNP on SSC1 marked a region that emerged in the Alentejana breed on SSC1, including the *MC4R* gene. An additional 11 signatures of selection identified in the Basque, Black Slavonian, Casertana, Cinta Senese, Gascon, Italian Large White,

(a)

breed	Alentejana	Apulo Calabrese	Basque	Bísara	Black Slavonian	Casertana	Cinta Senese	Italian Duroc	Gascon	Iberian	Krškopolje	Italian Landrace	Italian Large White	Lithuanian White Old Type	Lithuanian Indigenous Wattle	Majorcan Black	Mora Romagnola	Moravka	Nero Siciliano	Sarda	Schwabisch_Hallisches	Swallow Bellied Mangalitsa	Turopolje
Alentejana	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Apulo Calabrese	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Basque	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bísara	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Black Slavonian	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Casertana	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Cinta Senese	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Italian Duroc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gascon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Iberian	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Krškopolje	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Italian Landrace	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Italian Large White	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Lithuanian White Old Type	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Lithuanian Indigenous Wattle	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Majorcan Black	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Mora Romagnola	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Moravka	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
Nero Siciliano	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Sarda	0	0	0	2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	3	0	0	0
Schwabisch_Hallisches	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Swallow Bellied Mangalitsa	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Turopolje	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

(b)

breed	Alentejana	Apulo Calabrese	Basque	Bísara	Black Slavonian	Casertana	Cinta Senese	Italian Duroc	Gascon	Iberian	Krškopolje	Italian Landrace	Italian Large White	Lithuanian White Old Type	Lithuanian Indigenous Wattle	Majorcan Black	Mora Romagnola	Moravka	Nero Siciliano	Sarda	Schwabisch_Hallisches	Swallow Bellied Mangalitsa	Turopolje
Alentejana	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	1	0	0	0	0	0	0	0
Apulo Calabrese	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0
Basque	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bísara	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Black Slavonian	0	0	0	0	0	0	4	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	2
Casertana	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Cinta Senese	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Italian Duroc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Gascon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Iberian	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Krškopolje	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2	0	0	0	0	0
Italian Landrace	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
Italian Large White	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	0
Lithuanian White Old Type	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Lithuanian Indigenous Wattle	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
Majorcan Black	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Mora Romagnola	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Moravka	0	0	0	1	1	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	1	1	0
Nero Siciliano	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	2	0	2	0	0	0
Sarda	1	0	0	3	1	0	0	0	0	0	1	2	1	0	1	0	0	0	5	0	2	0	0
Schwabisch_Hallisches	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Swallow Bellied Mangalitsa	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Turopolje	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

FIGURE 4 Matrices of the classification error for (a) panel *N*/4 MDA and (b) panel *N*/32 MDG. Lines include the input breeds; the columns include the number of wrongly predicted animals for each breed.

Majorcan Black, Nero Siciliano and Swallow-Bellied Mangalitsa breeds harboured SNPs that were included in the Boruta reduced SNP panel (Table S15).

Gene enrichment analysis, that included the genes captured by the two best SNP panels according to the lowest OOB values (i.e. MDA $N/4$ and MDG $N/2$), was carried out with two tools (enrichR and NET-GE). Meaningful results were obtained only with enrichR, which identified several terms involved in general biological functions, otherwise, some terms were more specific for exterior phenotypes such as hair thickness and distribution (Tables S16 and S17). In particular, from the GWAS catalogue, the phenotypes 'monobrow', 'beard thickness' and 'male baldness' had p-values of 1.21×10^{-5}, 1.7×10^{-4} and 1.3×10^{-3}, respectively. Other significantly represented terms were related to fatty acid metabolism and to more general biological processes such as circadian entrainment or blood selenium level (Tables S16 and S17).

DISCUSSION

The identification of breed-informative SNPs can provide useful tools for the assignment of individual animals or their products to a particular breed. Applications of this information can have, for example, implications for the correct implementation of breeding and conservation programmes of animal genetic resources that request the verification of the breed to register the animals in the breed herd books (Tinarelli et al., 2021). Authentication of breed-branded (mono-breed) products (like meat and dairy products) can be obtained using this DNA information with impacts on the protection of niche value chains derived from local breeds (Fontanesi et al., 2016; Muñoz et al., 2020; Wilkinson et al., 2012).

We already applied several strategies to identify breed-informative SNPs that ranged from candidate gene approaches, where genes affecting breed-specific traits were targeted (Fontanesi et al., 2016; Tinarelli et al., 2021), to the use of high-density SNP chip data obtained at a population-wide level in different cattle and pig breeds (Bertolini et al., 2015, 2018; Muñoz et al., 2020; Schiavo et al., 2020). In the latter cases, different statistical methodologies and steps with their pros and cons, in comparison with the strategy that we developed in this study, have been proposed (Hulsegge et al., 2013; Kasarda et al., 2023; Miao et al., 2023; Wilkinson et al., 2011; Wilmot et al., 2022). In this context, however, the challenges are quite common and are due to (i) the large number of variables that should be compared across and (ii) several or many different groups of animals (i.e. breeds), (iii) which could be all or some genetically very similar and/or, alternatively, very divergent. Based on these elements, (iv) appropriate statistical tools and approaches should be evaluated also considering, (v) the requested computational time needed to establish ranked lists of

SNPs based on their combined informative features, which in turn are related to the compared populations. Most approaches provide only lists of markers lacking additional elements. Additional information, however, would be useful to understand marker informativeness (which is helpful for marker prioritisation), and to evaluate the combined multimarker classification errors, useful to define the allocation power of marker subsets.

Here, the computation burden that several statistical treatments of the SNP datasets has to face is caused by the high number of markers that are investigated. Therefore, one common strategy is to first reduce the number of SNPs. This reduction, however, should not lose informative SNPs and might deal with the high level of linkage disequilibrium that is usually present in livestock populations (Bertolini et al., 2018). In this study, we tested the use of a wrapper (i.e. Boruta algorithm) to first select features in a high-dimensionality space, characterised by the high-density SNP datasets obtained in 23 pig breeds. Then, based on a reduced number of markers, we exploited RF to rank the SNPs and define the most informative SNP panels, which could correctly classify the pigs to their original breed. The introduction of the Boruta algorithm is an improvement on previously used methods, allowing the selection of statistically more stable markers with good classification performances within many breeds. Two different ranked lists of the informative SNPs were obtained, based on MDA and MDG. In addition, starting from these ranked lists, RF was applied to explore different subsets of the informative SNPs, obtained by halving the number of SNPs five times, from 1595 SNPs (obtained after the Boruta steps) to the lowest tested number of 32 SNPs.

Across the 23 breeds, the lowest OOB classification error (2.299%) was obtained with the MDA $N/4$ panel (which included 398 SNPs). This value was even lower than that of the Boruta reduced panel ($N = 1595$ SNPs). For some breeds (Basque, Gascon, Majorcan Black, Mora Romagnola, Schwabisch-Hallisches, Swallow-Bellied Mangalitsa and Italian Duroc), the classification error was equal to zero for all defined SNP panels (obtained using MDA and MDG parameters to the lowest number of SNPs tested, i.e. 49 SNPs). That means that Boruta was able to capture genetic features that distinguished these breeds well. On the other hand, these captured features might be due to some distinct genetic differences at a few markers that allowed maximisation of the informative value of the selected SNP panels for about one-third of the studied breeds. For several other breeds (Bísara, Italian Landrace and Italian Large White, using the MDA ranking method; Casertana and Lithuanian Indigenous Wattle, using both MDA and MGA ranking methods), the classification error was zero with the next lowest number of SNPs ($N/16$, 99 SNPs). This is quite interesting, considering that the genetic history of some of these breeds is very similar or that, in some

cases, they have undergone mutual introgression (Bovo et al., 2020; Muñoz et al., 2018, 2019; Ojeda et al., 2006; Schiavo et al., 2021). For example, Mora Romagnola (an Italian local breed) was in the past recovered by crossbreeding with Duroc pigs (Tinarelli et al., 2021). Therefore, the fact that there were no classification errors in the Mora Romagnola and Italian Duroc could indicate that this autochthonous breed recovered its distinctiveness and particular genetic characteristics through a well customised conservation programme (Tinarelli et al., 2021), that the Boruta algorithm was able to capture.

On the other hand, for a few breeds, the classification error was higher than zero for all or almost all SNP subsets. In particular, the classification error for Black Slavonian and Sarda ranged from 0.122 to 0.204 and from 0.102 to 0.347 (over the different SNP panels), respectively. These breeds are quite heterogeneous and have been just recently recognised or established and their genetic heterogeneity has already been reported in previous studies (Bovo et al., 2020; Muñoz et al., 2018, 2019; Schiavo et al., 2021). For these breeds and for a few other breeds for which the classification error was always different from zero, it would, therefore, be difficult to establish SNP panels that could allocate animals to the correct breed without any error. This problem could also have implications for the possibility of authenticating the meat coming from these breeds and therefore using these tools to monitor the breed-branded value chains and protect them against the problems of frauds. In these cases, other strategies have been proposed, for example, a breed-specific phenotype, the belted coat colour, associated with a marker in the *KIT* gene, has been used to link all Cinta Senese pigs registered to their herd book to a DNA based authentication system of Cinta Senese meat derived only by this marker (Fontanesi et al., 2016, 2022).

Another very interesting outcome of this study is that highly breed-informative SNPs selected with the Boruta algorithm and then evaluated with the RF analyses can capture breed-specific phenotypes and other genetic characteristics of the breeds that define signatures of selection. We already noted that methods that identify breed-informative markers could be relevant also for the identification of genomic regions that might contain genetic features that affect breed-specific traits (Bertolini et al., 2018; Schiavo et al., 2020). In this study, we further found evidence that it could be possible to extend the application of some statistic methodologies, that have not been originally designed to identify breed-relevant genomic regions, for this purpose. For example, it is worth mentioning that the top MDG informative marker was in the *MSRB3* gene, which is known to affect one of the morphological traits (ear shape and size) that distinguish many pig breeds (Chen et al., 2013; Zhang et al., 2014, 2015). Several pig breeds investigated in our study can be distinguished by their peculiar ear shape and size (Table S1). A few other breed-specific phenotypes have

been captured, through the identification of informative SNPs, by the applied Boruta algorithm and RF ranking methods, including information on genetic factors determining body size and coat colour. Other markers included in the top-ranked lists have been associated with production traits, including carcass and meat quality traits, which might explain breed differences in production efficiency and production aptitudes (Čandek-Potokar & Nieto Linan, 2019; Fontanesi et al., 2013).

Other statistical approaches, including different combinations of statistical tests, will be investigated to identify other markers that could be useful to further improve the classification performances of the high-density SNP datasets for these pig breeds and other breeds and to capture markers that could explain the genetic differences between breeds for morphological and production traits. In this context, other methodological approaches should be tested or developed to directly link selected SNP markers with breed-specific traits.

AUTHOR CONTRIBUTIONS

Giuseppina Schiavo: Conceptualization; data curation; formal analysis; visualization; writing – original draft; writing – review and editing. **Francesca Bertolini:** Conceptualization; data curation; formal analysis; investigation; writing – original draft; writing – review and editing. **Samuele Bovo:** Data curation; formal analysis; investigation; writing – original draft. **Giuliano Galimberti:** Conceptualization; methodology; writing – original draft. **María Muñoz:** Data curation; resources. **Riccardo Bozzi:** Data curation; resources. **Marjeta Čandek-Potokar:** Data curation; funding acquisition; resources. **Cristina Óvilo:** Data curation; resources; writing – original draft. **Luca Fontanesi:** Conceptualization; funding acquisition; project administration; supervision; writing – original draft; writing – review and editing.

ACKNOWLEDGEMENTS

Science, Faculty of Agriculture, University of Zagreb, Croatia), Goran Kušec (Faculty of Agrobiotechnical Sciences, University of Osijek, Croatia), Marie-José. Mercat (IFIP Institut du Porc, France), Raquel Quintanilla (Programa de Genética y Mejora Animal, IRTA, Spain), Čedomir Radović (Department of Pig Breeding and Genetics, Institute for Animal Husbandry, Serbia), Violeta Razmaite (Animal Science Institute, Lithuanian University of Health Sciences, Lithuania) Juliette Riquet (GenPhySE, Université de Toulouse, INRA, France), Radomir Savić (Faculty of Agriculture, University of Belgrade, Serbia), Graziano Usai (AGRIS SARDEGNA, Italy) and Christoph Zimmer (Bäuerliche Erzeugergemeinschaft Schwäbisch Hall, Germany).

CONFLICT OF INTEREST STATEMENT

The authors declare they do not have any competing interests.

DATA AVAILABILITY STATEMENT

Genotyping datasets will be available on reasonable request addressed to the TREASURE Consortium, after a signature of an agreement on their use. Requests can be sent to luca.fontanesi@unibo.it.

ORCID

Samuele Bovo https://orcid.org/0000-0002-5712-8211
Luca Fontanesi https://orcid.org/0000-0001-7050-3760

REFERENCES

Acharjee, A., Larkman, J., Xu, Y., Cardoso, V.R. & Gkoutos, G.V. (2020) A random forest based biomarker discovery and power analysis framework for diagnostics research. *BMC Medical Genomics*, 13, 178. Available from: https://doi.org/10.1186/s12920-020-00826-6

Bertolini, F., Galimberti, G., Calò, D.G., Schiavo, G., Matassino, D. & Fontanesi, L. (2015) Combined use of principal component analysis and random forests identify population-informative single nucleotide polymorphisms: application in cattle breeds. *Journal of Animal Breeding and Genetics*, 132, 346–356. Available from: https://doi.org/10.1111/jbg.12155

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

Botelho, M.E., Lopes, M.S., Mathur, P.K., Knol, E.F., e Silva, F.F., Lopes, P.S. et al. (2022) Weighted genome-wide association study reveals new candidate genes related to boar taint compounds l. *Livestock Science*, 257, 104845. Available from: https://doi.org/10.1016/J.LIVSCI.2022.104845

Bovo, S., Ribani, A., Muñoz, M., Alves, E., Araujo, J.P., Bozzi, R. et al. (2020) Whole-genome sequencing of European autochthonous and commercial pig breeds allows the detection of signatures of selection for adaptation of genetic resources to different breeding and production systems. *Genetics Selection Evolution*, 52, 33. Available from: https://doi.org/10.1186/S12711-020-00553-7

Breiman, L. (2001) Random forests. *Machine Learning*, 45, 1–122. Available from: https://doi.org/10.4324/9781003109396-5

Buniello, A., Macarthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C. et al. (2019) The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47, D1005–D1012. Available from: https://doi.org/10.1093/nar/gky1120

Čandek-Potokar, M. & Nieto Linan, R.M. (2019) *European local pig breeds – diversity and performance: a study of project TREASURE*. London: IntechOpen. Available from: https://doi.org/10.5772/intechopen.83749

Casiró, S., Velez-Irizarry, D., Ernst, C.W., Raney, N.E., Bates, R.O., Charles, M.G. et al. (2017) Genome-wide association study in an F2 Duroc × Pietrain resource population for economically important meat quality and carcass traits. *Journal of Animal Science*, 95, 545–558. Available from: https://doi.org/10.2527/jas.2016.1003

Chang, C.C., Chow, C.C., Tellier, L.C.A.M., Vattikuti, S., Purcell, S.M. & Lee, J.J. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4, 7. Available from: https://doi.org/10.1186/s13742-015-0047-8

Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V. et al. (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14, 128. Available from: https://doi.org/10.1186/1471-2105-14-128

Dadousis, C., Muñoz, M., Óvilo, C., Fabbri, M.C., Araújo, J.P., Bovo, S. et al. (2022) Admixture and breed traceability in European indigenous pig breeds and wild boar using genome-wide SNP data. *Scientific Reports*, 12, 7346. Available from: https://doi.org/10.1038/s41598-022-10698-8

De Oliveira Peixoto, J., Guimarães, S.E.F., Lopes, P.S., Soares, M.A.M., Pires, A.V., Barbosa, M.V.G. et al. (2006) Associations of leptin gene polymorphisms with production traits in pigs. *Journal of Animal Breeding and Genetics*, 123, 378–383. Available from: https://doi.org/10.1111/j.1439-0388.2006.00611.x

Fiore, E., Blasi, F., Morgante, M., Cossignani, L., Badon, T., Gianesella, M. et al. (2020) Changes of milk fatty acid composition in four lipid classes as biomarkers for the diagnosis of bovine ketosis using bioanalytical thin layer chromatography and gas chromatographic techniques (TLC-GC). *Journal of Pharmaceutical and Biomedical Analysis*, 188, 113372. Available from: https://doi.org/10.1016/j.jpba.2020.113372

Fontanesi, L. (2022) Genetics and genomics of pigmentation variability in pigs: a review. *Livestock Science*, 265, 105079. Available from: https://doi.org/10.1016/j.livsci.2022.105079

Fontanesi, L., Buttazzoni, L., Galimberti, G., Calò, D.G., Scotti, E. & Russo, V. (2013) Association between melanocortin 4 receptor (*MC4R*) gene haplotypes and carcass and production traits in Italian large white pigs evaluated with a selective genotyping approach. *Livestock Science*, 157, 48–56. Available from: https://doi.org/10.1016/j.livsci.2013.07.006

Fontanesi, L., Schiavo, G., Gallo, M., Baiocco, C., Galimberti, G., Bovo, S. et al. (2017) Genome-wide association study for ham weight loss at first salting in Italian large white pigs: towards the genetic dissection of a key trait for dry-cured ham production. *Animal Genetics*, 48, 103–107. Available from: https://doi.org/10.1111/age.12491

Fontanesi, L., Scotti, E., Gallo, M., Costa, L.N. & 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. Available from: https://doi.org/10.1016/j.livsci.2015.12.007

Franco, M.M., Antunes, R.C., Silva, H.D. & Goulart, L.R. (2005) Association of PIT1, GH and GHRH polymorphisms with performance and carcass traits in landrace pigs. *Journal of Applied Genetics*, 46, 195–200.

Gao, J., Sun, L., Zhang, S., Xu, J., He, M., Zhang, D. et al. (2022) Screening discriminating SNPs for Chinese indigenous pig breeds identification using a random forests algorithm. *Genes (Basel)*, 13, 2207. Available from: https://doi.org/10.3390/genes13122207

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

Hayah, I., Ababou, M., Botti, S. & Badaoui, B. (2021) Comparison of three statistical approaches for feature selection for fine-scale genetic population assignment in four pig breeds. *Tropical Animal Health and Prodution*, 53, 3. Available from: https://doi.org/10.1007/s11250-021-02824-x

Hulsegge, B., Calus, M.P.L., Windig, J.J., Hoving-Bolink, A.H., Maurice-van Eijndhoven, M.H.T. & Hiemstra, S.J. (2013) Selection of SNP from 50K and 777K arrays to predict breed of origin in cattle. *Journal of Animal Science*, 91, 5128–5134. Available from: https://doi.org/10.2527/jas.2013-6678

Jollife, I.T. & Cadima, J. (2016) Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A – Mathematical Physical and Engineering Sciences*, 374, 20150202. Available from: https://doi.org/10.1098/rsta.2015.0202

Kasarda, R., Moravčíková, N., Mészáros, G., Simčič, M. & Zaborski, D. (2023) Classification of cattle breeds based on the random forest approach. *Livestock Science*, 267, 105143. Available from: https://doi.org/10.1016/j.livsci.2022.105143

Kennes, Y.M., Murphy, B.D., Pothier, F. & Palin, M.F. (2001) Characterization of swine leptin (LEP) polymorphisms and their association with production traits. *Animal Genetics*, 32, 215–218. Available from: https://doi.org/10.1046/j.1365-2052.2001.00768.x

Kleanthous, N., Hussain, A., Mason, A., Sneddon, J., Shaw, A., Fergus, P. et al. (2018) Machine learning techniques for classification of livestock behavior. *Neural Information Processing. ICONIP, Springer.* 11304 LNCS, 304–315. https://doi.org/10.1007/978-3-030-04212-7_26

Kursa, M.B., Jankowski, A. & Rudnicki, W.R. (2010) Boruta – A system for feature selection. *Fundamenta Informaticae*, 101, 271–285. Available from: https://doi.org/10.3233/FI-2010-288

Kursa, M.B. & Rudnicki, W.R. (2010) Feature selection with the Boruta package. *Journal of Statistical Sotware*, 36, 1–13. Available from: https://doi.org/10.18637/jss.v036.i11

Liu, R., Xu, Z., Teng, J., Pan, X., Lin, Q., Cai, X. et al. (2022) Evaluation of six machine learning classification algorithms in pig breed identification using SNPs array data. *Animal Genetics*, 54, 113–122. Available from: https://doi.org/10.1111/age.13279

Meng, Y.A., Yu, Y., Cupples, L.A., Farrer, L.A. & Lunetta, K.L. (2009) Performance of random forest when SNPs are in linkage disequilibrium. *BMC Bioinformatics*, 10, 78. Available from: https://doi.org/10.1186/1471-2105-10-78

Miao, J., Chen, Z., Zhang, Z., Wang, Z., Wang, Q., Zhang, Z. et al. (2023) A web tool for the global identification of pig breeds. *Genetics Selection Evolution*, 55, 18. Available from: https://doi.org/10.1186/s12711-023-00788-0

Moe, M., Lien, S., Bendixen, C., Hedegaard, J., Hornshøj, H., Berget, I. et al. (2008) Gene expression profiles in liver of pigs with extreme high and low levels of androstenone. *BMC Veterinary Research*, 4, 29. Available from: https://doi.org/10.1186/1746-6148-4-29

Mulim, H.A., Brito, L.F., Pinto, L.F.B., Ferraz, J.B.S., Grigoletto, L., Silva, M.R. et al. (2022) Characterization of runs of homozygosity, heterozygosity-enriched regions, and population structure in cattle populations selected for different breeding goals. *BMC Genomics*, 23, 209. Available from: https://doi.org/10.1186/s12864-022-08384-0

Muñoz, M., Bozzi, R., García, F., Núñez, Y., Geraci, C., Crovetti, A. et al. (2018) Diversity across major and candidate genes in European local pig breeds. *PLoS One*, 13, e0207475. Available from: https://doi.org/10.1371/journal.pone.0207475

Muñoz, M., Bozzi, R., García-Casco, J., Núñez, Y., Ribani, A., Franci, O. et al. (2019) Genomic diversity, linkage disequilibrium and selection signatures in European local pig breeds assessed with a high density SNP chip. *Scientific Reports*, 9, 13546. Available from: https://doi.org/10.1038/s41598-019-49830-6

Muñoz, M., García-Casco, J.M., Alves, E., Benítez, R., Barragán, C., Caraballo, C. et al. (2020) Development of a 64 SNV panel for breed authentication in Iberian pigs and their derived meat products. *Meat Science*, 167, 108152. Available from: https://doi.org/10.1016/j.meatsci.2020.108152

Ojeda, A., Rozas, J., Folch, J.M. & Pérez-Enciso, M. (2006) Unexpected high polymorphism at the FABP4 gene unveils a complex history for pig populations. *Genetics*, 174, 2119–2127. Available from: https://doi.org/10.1534/genetics.106.063057

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

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

Pena, R.N., Noguera, J.L., Casellas, J., Díaz, I., Fernández, A.I., Folch, J.M. et al. (2013) Transcriptional analysis of intramuscular fatty acid composition in the longissimus thoracis muscle of Iberian × Landrace back-crossed pigs. *Animal Genetics*, 44, 648–660. Available from: https://doi.org/10.1111/age.12066

Pérez-Montarelo, D., Fernández, A., Folch, J.M., Pena, R.N., Ovilo, C., Rodríguez, C. et al. (2012) Joint effects of porcine leptin and leptin receptor polymorphisms on productivity and quality traits. *Animal Genetics*, 43, 805–809. Available from: https://doi.org/10.1111/j.1365-2052.2012.02338.x

Quinlan, A.R. & Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26, 841–842. Available from: https://doi.org/10.1093/bioinformatics/btq033

R Core Team. (2021) *R: a language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing. Available from: https://www.R-project.org/

Rubin, C.J., Megens, H.J., Barrio, A.M., Maqbool, K., Sayyab, S., Schwochow, D. et al. (2012) Strong signatures of selection in the domestic pig genome. *Proceedings of the National Academy of Sciences of the USA*, 109, 19529–19536. Available from: https://doi.org/10.1073/pnas.1217149109

Schiavo, G., Bertolini, F., Galimberti, G., Bovo, S., Dall'Olio, S., Nanni Costa, L. et al. (2020) A machine learning approach for the identification of population-informative markers from high-throughput genotyping data: application to several pig breeds. *Animal*, 14, 223–232. Available from: https://doi.org/10.1017/S1751731119002167

Schiavo, G., Bovo, S., Muñoz, M., Ribani, A., Alves, E., Araújo, J.P. et al. (2021) Runs of homozygosity provide a genome landscape picture of inbreeding and genetic history of European autochthonous and commercial pig breeds. *Animal Genetics*, 52, 155–170. Available from: https://doi.org/10.1111/age.13045

Schiavo, G., Bovo, S., Ribani, A., Moscatelli, G., Bonacini, M., Prandi, M. et al. (2022) Comparative analysis of inbreeding parameters and runs of homozygosity islands in 2 Italian autochthonous cattle breeds mainly raised in the Parmigiano-Reggiano cheese production region. *Journal of Dairy Science*, 105, 2408–2425. Available from: https://doi.org/10.3168/jds.2021-20915

Speiser, J.L., Miller, M.E., Tooze, J. & Ip, E. (2019) A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications*, 134, 93–101. Available from: https://doi.org/10.1016/j.eswa.2019.05.028

Tinarelli, S., Ribani, A., Utzeri, V.J., Taurisano, V., Bovo, C., Dall'Olio, S. et al. (2021) Redefinition of the Mora Romagnola pig breed herd book standard based on DNA markers useful to authenticate its 'mono-breed' products: an example of sustainable conservation of a livestock genetic resource. *Animals*, 11, 526. Available from: https://doi.org/10.3390/ani11020526

Wang, X., Ligang, W., Shi, L., Zhang, P., Li, Y., Li, M. et al. (2022) GWAS of reproductive traits in large white pigs on chip and imputed whole-genome sequencing data. *International Journal of Molecular Sciences*, 23, 13338. Available from: https://doi.org/10.3390/ijms232113338

Wilkinson, S., Archibald, A.L., Haley, C.S., Megens, H.J., Crooijmans, R.P.M.A., Groenen, M.A.M. et al. (2012) Development of a genetic tool for product regulation in the diverse British pig breed market. *BMC Genomics*, 13, 580. Available from: https://doi.org/10.1186/1471-2164-13-580

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

Wilmot, H., Bormann, J., Soyeurt, H., Hubin, X., Glorieux, G., Mayeres, P. et al. (2022) Development of a genomic tool for breed assignment by comparison of different classification models: application to three local cattle breeds. *Journal of Animal Breeding and Genetics*, 139, 40–61. Available from: https://doi.org/10.1111/jbg.12643

Zhang, L., Liang, J., Luo, W., Liu, X., Yan, H., Zhao, K. et al. (2014) Genome-wide scan reveals LEMD3 and WIF1 on SSC5 as the candidates for porcine ear size. *PLoS One*, 9, e102085. Available from: https://doi.org/10.1371/journal.pone.0102085

Zhang, Y., Liang, J., Zhang, L., Ligang, W., Liu, X., Yan, H. et al. (2015) Porcine methionine sulfoxide reductase B3: molecular cloning, tissue-specific expression profiles, and polymorphisms associated with ear size in *Sus scrofa*. *Journal of Animal Science and Biotechnology*, 6, 60. Available from: https://doi.org/10.1186/s40104-015-0060-x

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Schiavo, G., Bertolini, F., Bovo, S., Galimberti, G., Muñoz, M., Bozzi, R. et al. (2024) Identification of population-informative markers from high-density genotyping data through combined feature selection and machine learning algorithms: Application to European autochthonous and cosmopolitan pig breeds. *Animal Genetics*, 00, 1–13. Available from: https://doi.org/10.1111/age.13396