Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

Implementation of the Boruta algorithm for classification and characterization of European pig breeds using High-Density Single Nucleotide Polymorphisms Data

Giuseppina Schiavo[1], Samuele Bovo[1], Matteo Bolner[1], Anisa Ribani[1], Valeria Taurisano[1], Giuliano Galimberti[2], Maurizio Gallo[3], Luca Fontanesi[1]
[1]Dept. of Agricultural and Food Sciences, University of Bologna, Italy; [2]Dept. of Statistical Sciences "Paolo Fortunati", University of Bologna, Italy;[3]Associazione Nazionale Allevatori Suini, Roma, Italy
Correspondence: luca.fontanesi@unibo.it, giuseppina.schiavo2@unibo.it

Animal and Food Genomics Group

Full paper on Animal Genetics

Introduction

Autochthonous pig breeds are a valuable source of genetic diversity, whose conservation at the genomic level is a matter of concern in management programs across Europe. High-density single nucleotide polymorphism (SNP) genotyping is a cost-effective tool for capturing genetic variation, but their processing can be computationally demanding. Feature selection methods often require a compromise between computational feasibility and the retention of biologically relevant markers. This study presents the application of the Boruta algorithm, a wrapper for Random Forests (RF), to address these challenges by identifying informative SNPs.

Dataset

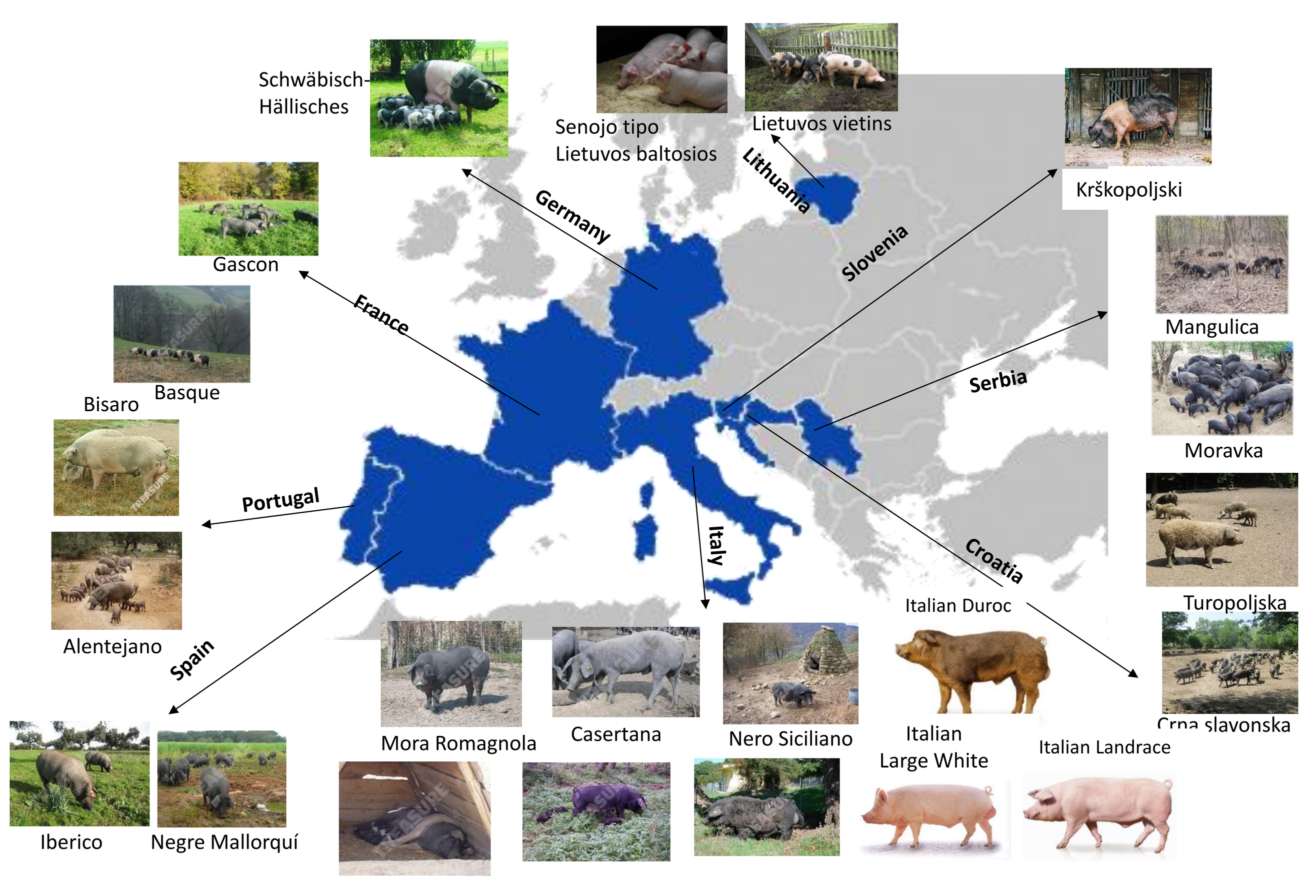

Figure 1. Breeds included in this study and their origin

The work included genomic data from 1,154 pigs representing 23 European autochthonous breeds, genotyped with the GGP 70k Porcine array. Quality control and linkage disequilibrium filtering reduced the dataset to 36,107 SNPs.

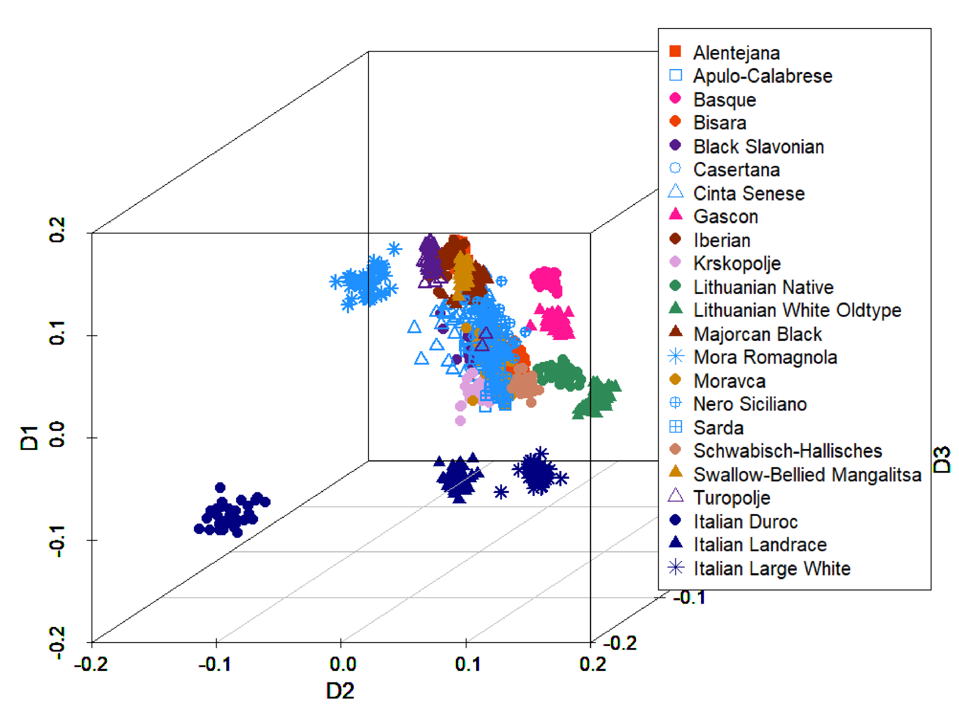

Figure 2. Multi-dimensional Scaling plot of the animals after filtering the dataset for quality. Animals are divided by breed (indicated by different symbols) and by geographical area (indicated by color).

Boruta and Random Forests

Boruta algorithm consists of iterative applications of RF using real features (the SNPs) and shadow features artificially created by randomly permuting the observed ones. Depending on the estimated importance of a feature for classification, three possible output:

real feature importance < shadow feature → Rejected
real feature importance == shadow feature → Tentative
real feature importance > shadow feature → Confirmed

Feature selection steps considered the distribution of SNPs along chromosomes and were designed to keep subsets of SNPs that were representative of all the 18 Sus scrofa autosomes.

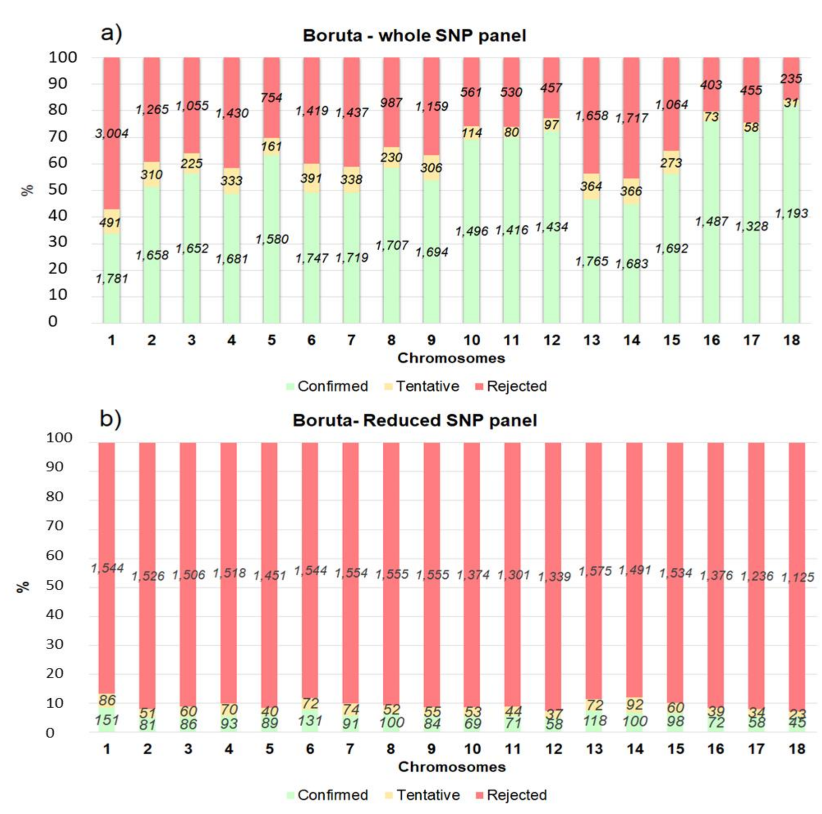

Figure 3. Distribution of «Confirmed» SNPs along the chromosomes

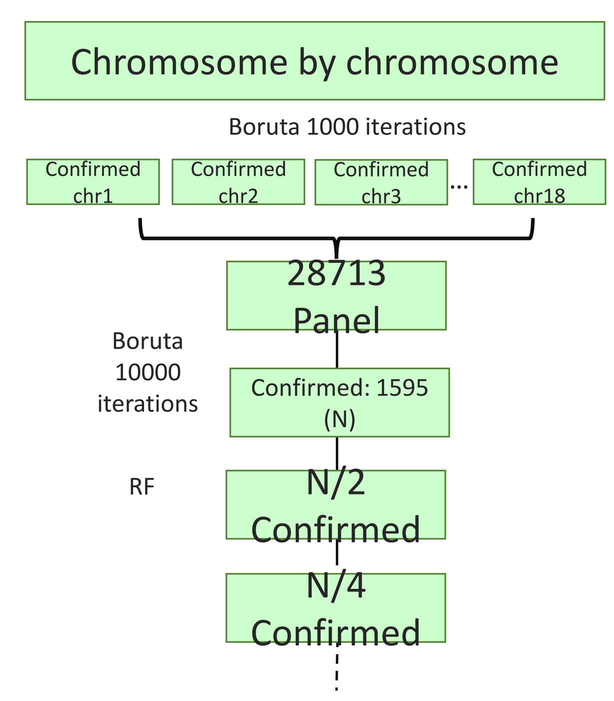

Figure 4. Approach to reduce the dataset after each run of Boruta and simple RF

Main steps of the implemented pipeline:

1. First Boruta run independently on each Chromosome;
2. All "Confirmed" SNPs merged to create a filtered SNP panel (28,713 SNPs);
3. Second Boruta run on the SNP panel to obtain a Reduced panel of Confirmed SNPs (N = 1,595 SNPs);
4. Random Forest on panel "N";
5. Considering RF Mean Decrease Gini (MDG) and Mean Decrease Accuracy (MDA) ranking to select top SNPs in smaller datasets obtained by dividing the initial number N of SNPs (N/2, N/4, N/8, N/16 and N/32);
6. Out of Bag (OOB) error evaluation for each subset.

Results

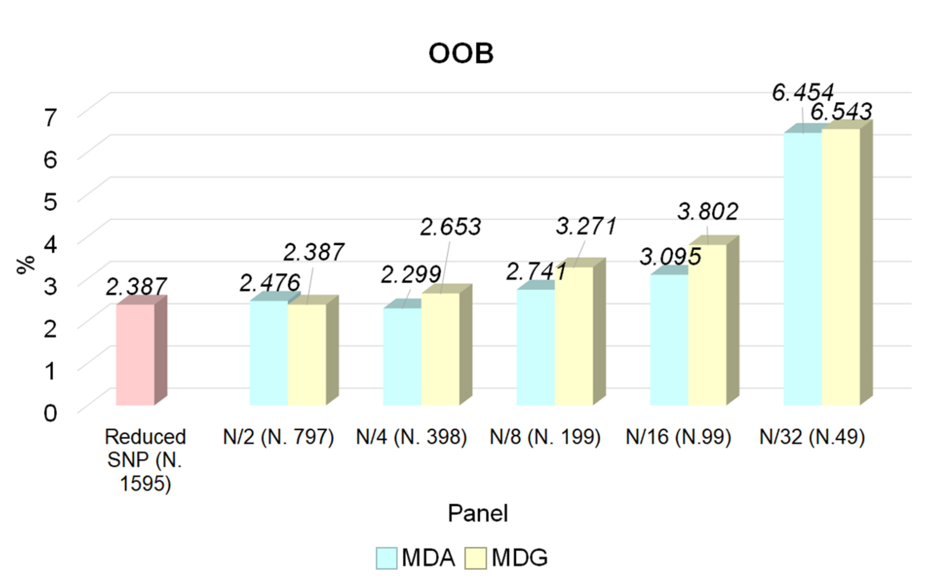

Figure 5. Out Of Bag for the reduced SNP panel and of the the subsets that were defined by subsequently halving the number of SNPs (N) including the SNPs ranked based on their Mean Decrease Accuracy (MDA) and Mean Decrease Gini (MDG). For each SNP panel, the number of SNPs is indicated in parenthesis. At the top of each bar, the OOB (%) error is reported.

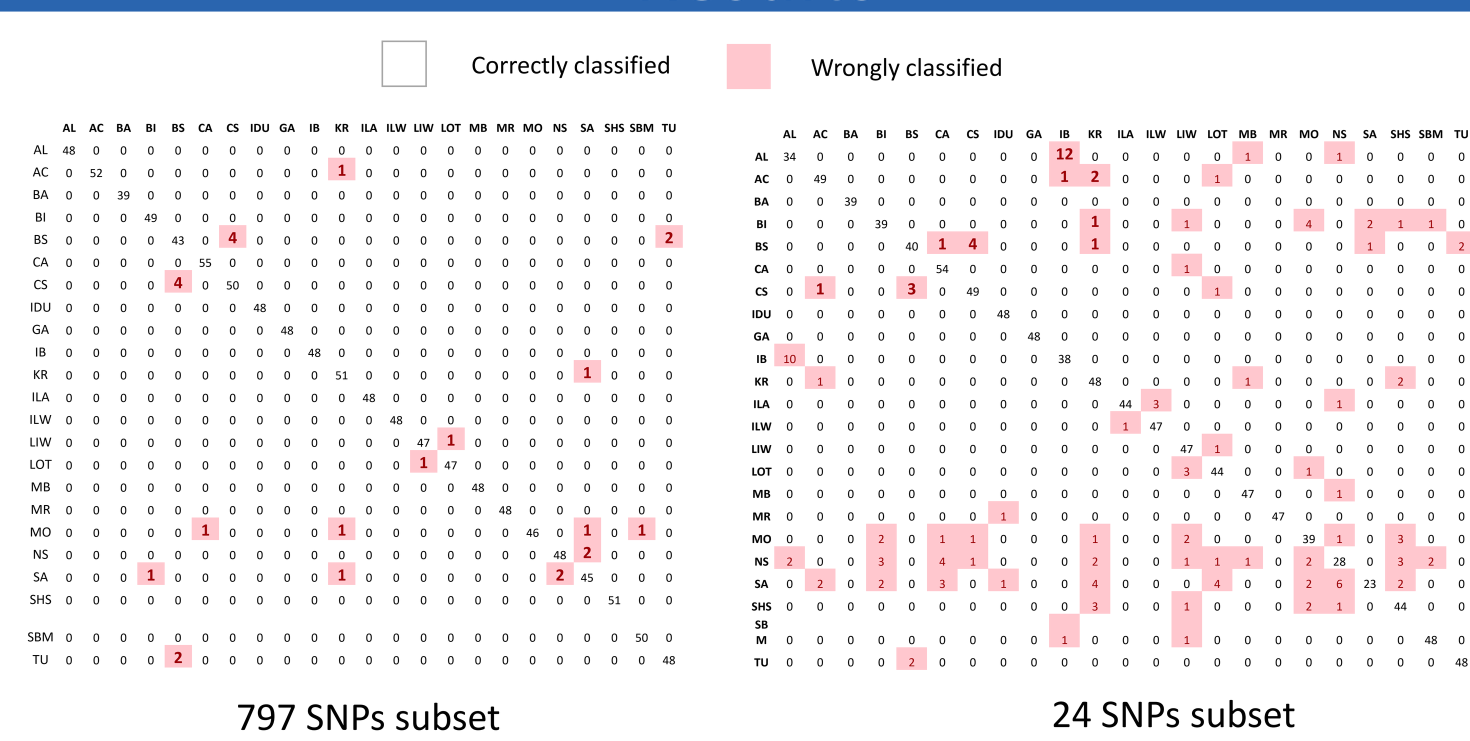

Figure 6. Comparison between the N/2 and N/64 subsets. Red cells represent the number of individuals misclassified from the breed indicated in the row as the breed indicated in the column.

Chrom	Pos	Name	Description
2	138424173	TRPC7	transient receptor potential cation channel subfamily
8	124618687	BMPR1B	bone morphogenetic protein receptor type 1B
8	131373563	DSPP	dentin sialophosphoprotein
16	66468093	SGCD	sarcoglycan delta
8	119588455	PPP3CA	protein phosphatase 3 catalytic subunit alpha
13	80321284	RBP2	retinol binding protein 2
9	136163686	ZPBP	zona pellucida binding protein
6	146470297	PDE4B	phosphodiesterase 4B
8	113988990	HADH	hydroxyacyl-CoA dehydrogenase
17	2770605	SGCZ	sarcoglycan zeta
1	253869470	SLC31A1	high-affinity copper uptake protein
2	87286297	SCAMP1	secretory carrier membrane protein 1
2	5216082	GRK2	G protein-coupled receptor kinase 2
6	4970323	CDH13	cadherin 13%2C H-cadherin (heart)
6	146425347	PDE4B	phosphodiesterase 4B
6	31375428	FTO	fat mass and obesity associated
17	54235613	TSHZ2	teashirt zinc finger homeobox 2
17	2736745	SGCZ	sarcoglycan zeta
5	81804433	IGF1	insulin like growth factor 1

Table 1. Annotation of main genes encompassed by the top SNPs or in the 500kb surrounding regions

The Boruta algorithm identified 2,471 stable SNPs. Annotation of selected SNPs using Ensembl Biomart revealed genomic regions associated with genes that could play significant roles in breed differentiation and adaptation. This work shows the advantages of combining high-density SNP data with machine learning techniques and highlights the potential of AI-driven approaches to identify key genetic markers to enhance conservation and breeding.

Acknowledgments: Funded by the European Union - NextGenerationEU under the National Recovery and Resilience Plan (PNRR) - Mission 4 Education and research - Component 2 From research to business - Investment 1.1 Notice Prin 2022 - DD N. 104 del 2/2/2022, proposal code 202238NP9N – CUP J53D23009570001