ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

AI-driven exploration of molecular data in animal science: applications in genomics and metabolomics to dissect the animal phenome

Samuele Bovo, Giuseppina Schiavo, Matteo Bolner, Francesca Bertolini, and Luca Fontanesi

Animal and Food Genomics Group
Department of Agricultural and Food Sciences
University of Bologna, Italy (IT)

From high-dimensional omics to interpretable feature selection

1 Omics data

High-dimensional: thousands to millions of features, often with correlations and hidden structure.

Metabolomics: 1,000 molecules
Genomics: ~700,000 SNPs.

2 Machine Learning

ML can model many variables jointly and capture nonlinear relationships that univariate scans may miss.

3 The challenge

Using all features:
→ Overfit
→ make the model hard to interpret.

Most features are not biologically relevant.

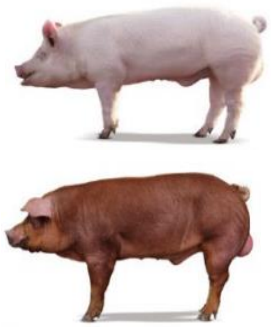
4 Feature selection

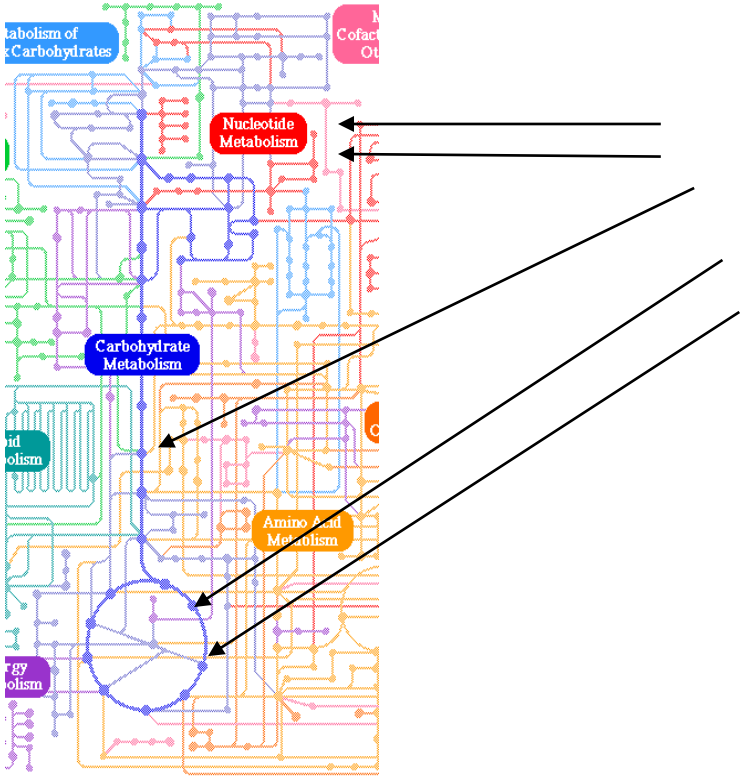
Goal: **retain a compact set of informative features that improves prediction and supports biological interpretation.**

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

The advantage of using BORUTA

- A feature selection algorithm
- Designed as a wrapper around a Random Forest classification algorithm
- **An all-relevant feature selection wrapper**





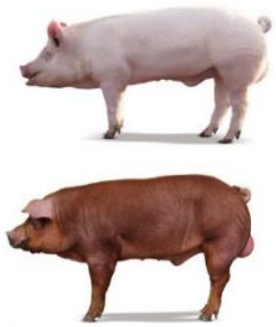
All-relevant feature selection
Several metabolites selected
Global biological insights ☺
→ Example: nucleotides and carbohydrates metabolisms differ between breeds

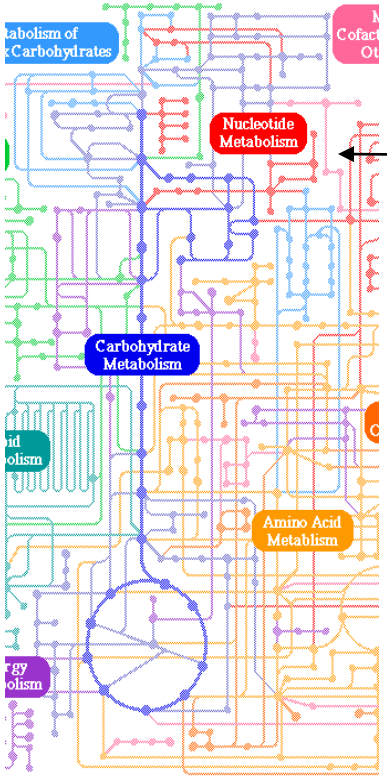
We know where we can play with genetics and/or nutrition

Kursa, M. B., Jankowski, A., & Rudnicki, W. R. (2010). Boruta–a system for feature selection. *Fundamenta informaticae*, *101*(4), 271-285.

The advantage of using BORUTA

- A feature selection algorithm
- Designed as a wrapper around a Random Forest classification algorithm
- **An all-relevant feature selection wrapper**





E.g. sPLS-DA (L1 penalization)

Minimal-optimal model

Example: one metabolite is selected

Sufficient for building a model ☺

→ **No global insights** in the biology of the system ☹

Kursa, M. B., Jankowski, A., & Rudnicki, W. R. (2010). Boruta–a system for feature selection. *Fundamenta informaticae*, *101*(4), 271-285.

The advantage of using BORUTA

The strategy in brief:

1. For each feature we create a corresponding “shadow” feature
   → values are obtained by shuffling the original feature
2. Perform a classification using all attributes of this extended system
   → compute the importance of all attributes.

The importance of a shadow feature

→ can be nonzero only due to random fluctuations.

→reference for deciding which features are truly important.

If Importance of Feature > Importance Shadow feature → reteain/label as important

Kursa, M. B., Jankowski, A., & Rudnicki, W. R. (2010). Boruta–a system for feature selection. *Fundamenta informaticae*, *101*(4), 271-285.

Given a high-dimensional biological dataset, which biological features drive the final phenotype?

Two practical applications of Boruta:

1

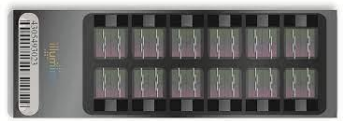
Breed differences

Metabolon
SMALL MOLECULES, BIG INSIGHTS

Metabolomics

2

Breed differences



Genomics

Metabolomics

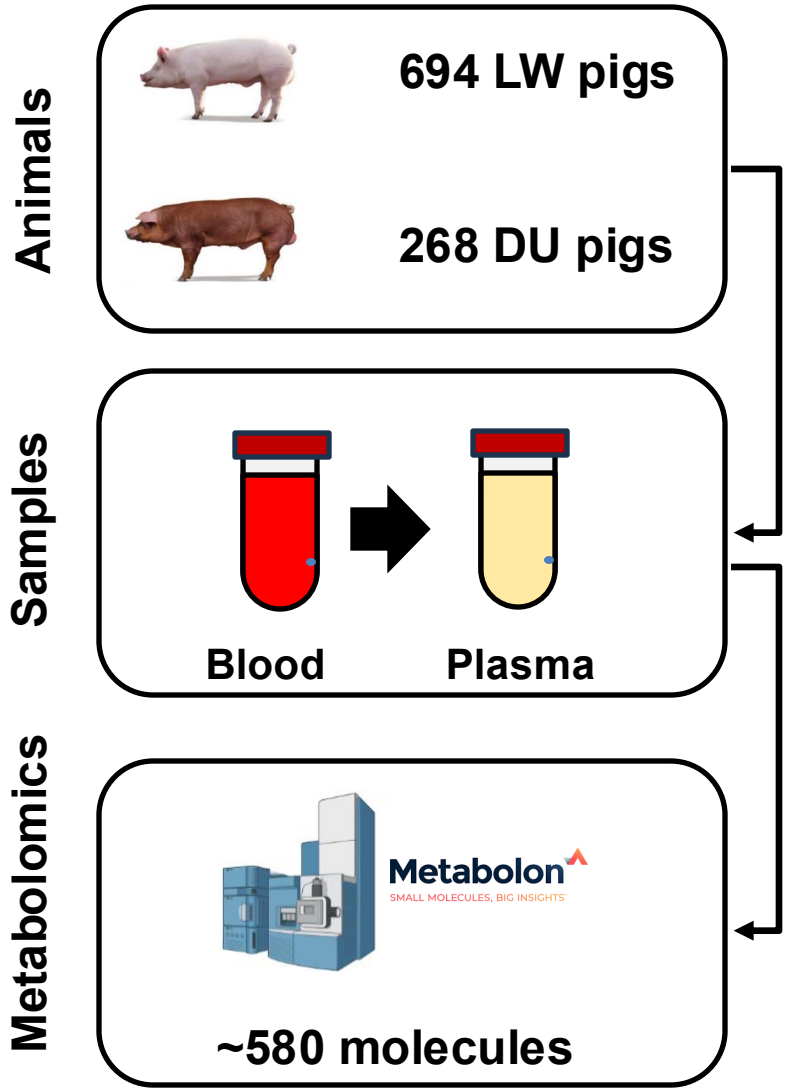
1

Breed differences

Metabolon
SMALL MOLECULES, BIG INSIGHTS

Bovo, S., Bolner, M., Schiavo, G., Galimberti, G., Bertolini, F., Dall'Olio, S., ... & Fontanesi, L. (2025). High-throughput untargeted metabolomics reveals metabolites and metabolic pathways that differentiate two divergent pig breeds. *animal*, *19*(1), 101393.

Dataset and workflow of analysis



Identification differently abundant metabolites

Evaluation of the effect of random seed
- → in data imputation
- → in the feature selection algorithm

1. Imputation:

MICE → it generates 5 datasets

Testing of 5 random seeds → 25 datasets

2. Feature selection (labelling)

Boruta applied to each dataset

→ Testing 5 random seeds → 125 runs of analyses

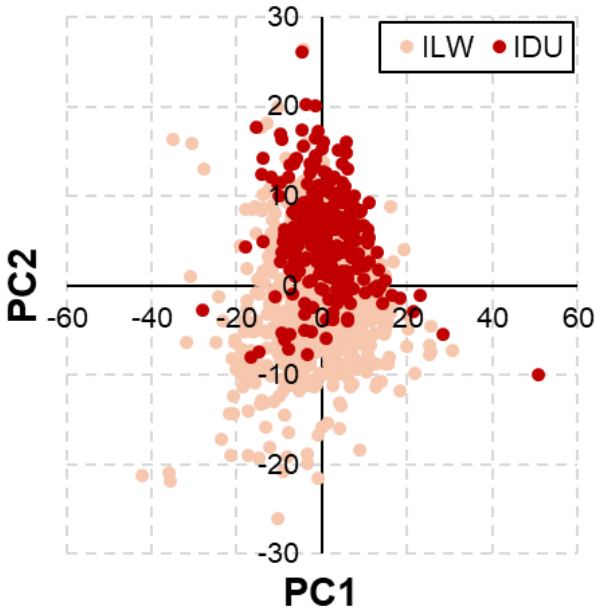
→ 10CV validation → 1,250 runs of analyses

3. Filtering

Keep as discriminant metabolites those always selected across the 1,250 runs (intersection sets)

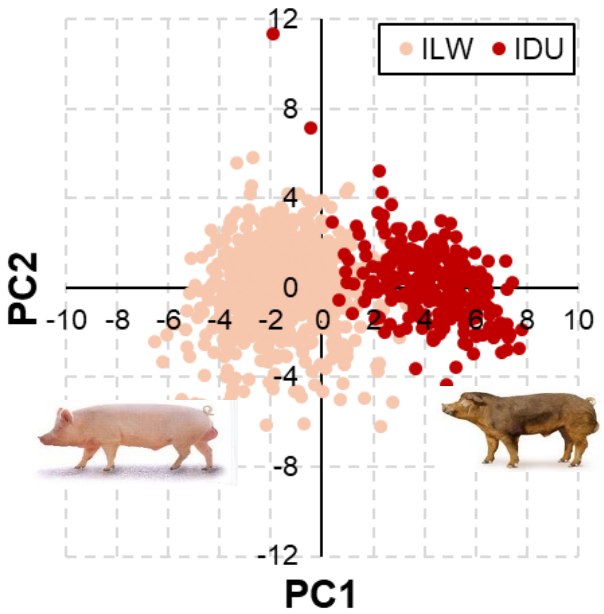
Results: selected discriminant metabolites

PCA whole profile: 580 molecules



Boruta-based pipeline

PCA selected metabolites



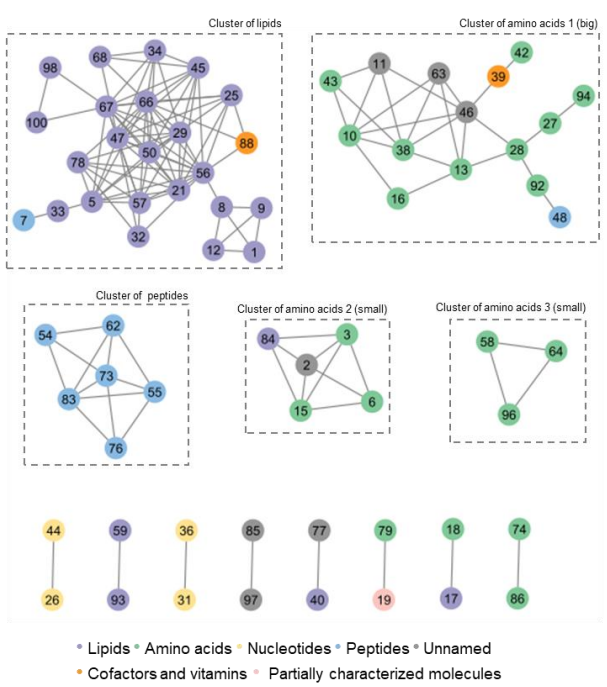
Differences in 100 metabolites

Differences are indirectly due to the different genetic background

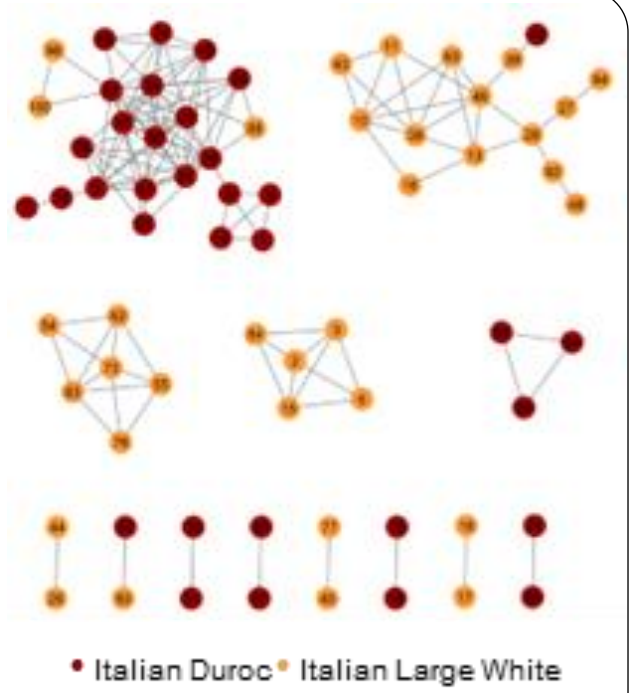
(Genome → Proteome → Metabolome)

Results: biological insights of selected metabolites

Network analysis of selected metabolites

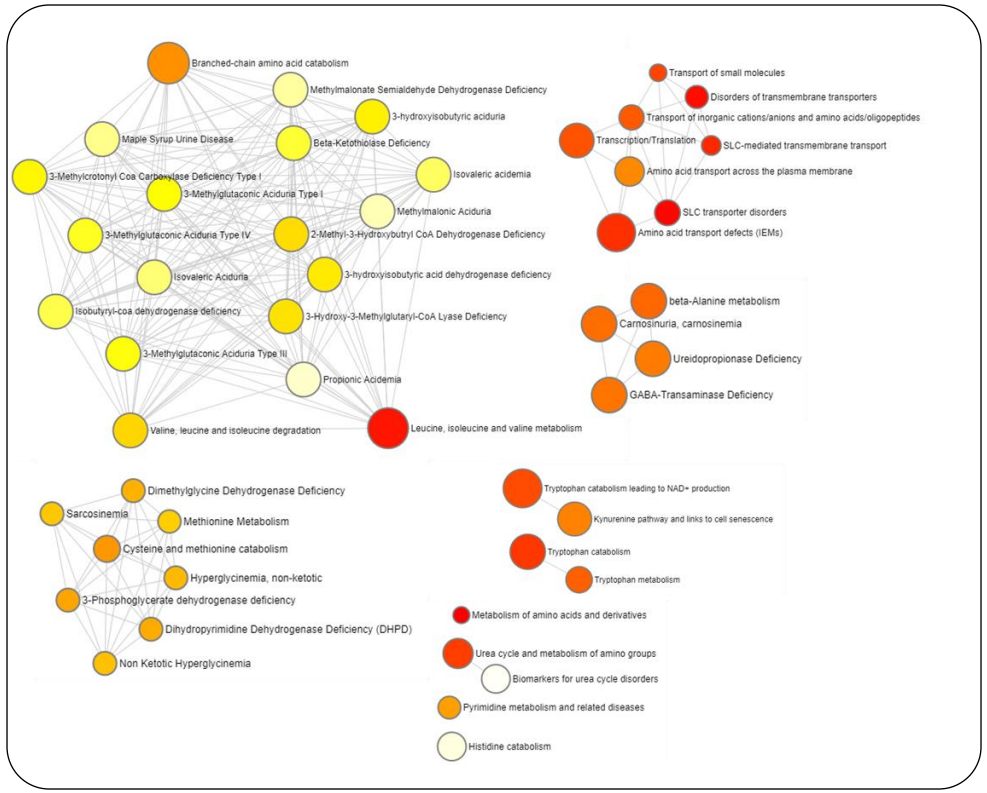


Clusters colored by pathway



Clusters colored by breed

Pathway analysis of selected metabolites



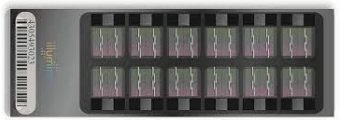
Lipids are generally higher in Duroc
Amino-acids and peptides higher in Large White

Kynurenine pathway:
→modulation of immune, reproductive, and central nervous systems

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

2

Breed differences



Genomics

Schiavo, G., Bertolini, F., Bovo, S., Galimberti, G., Muñoz, M., Bozzi, R., ... & Fontanesi, L. (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*, *55*(2), 193-205.

Breed differences: relevant DNA markers

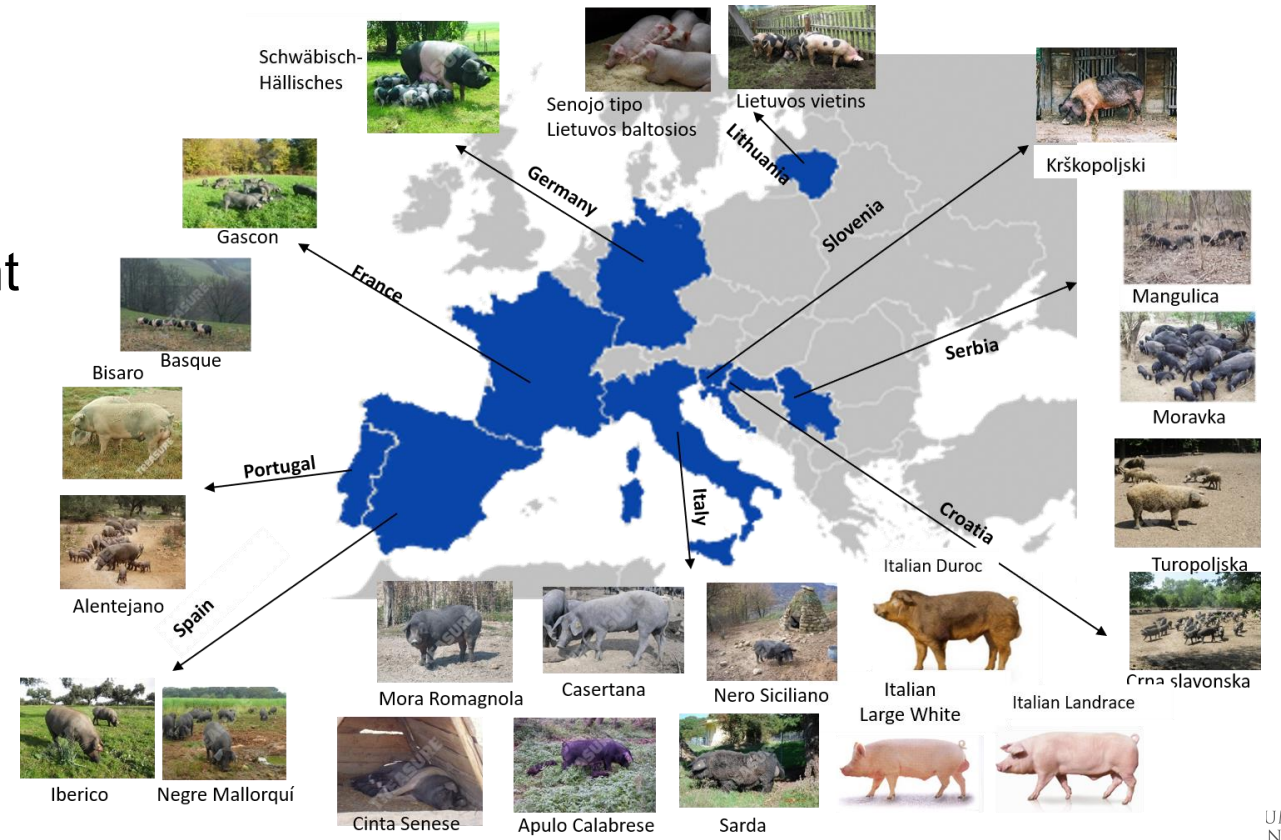
The genome of pig populations has been shaped by natural and artificial selection, resulting in allele frequency differences at loci involved in adaptation and economically important traits, defining breeds.

Autochthonous breeds:

- reservoirs of genetic variability
- adaptation to production environment
- conservation programs strategies

Dataset:

- A total of 1,154 pigs; 23 breeds
- Genomic data: 70,000 SNPs



Workflow of analysis

1. Data quality check and filtering

PLINK → Call rate 0.90 + HWE > 0.001

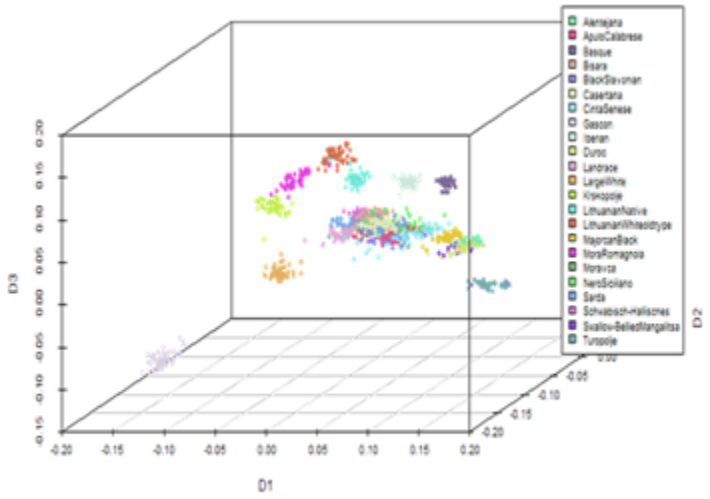
→ **57,877** total SNPs

→ 1,154 pigs; 23 breeds

2. Data coding

Each animal: vector of 0, 1 or 2

→ number of copies of the minor allele of each SNP.



3. Boruta analysis

Identification of relevant markers discriminating across the 23 breeds

→ **Computational problems** ←

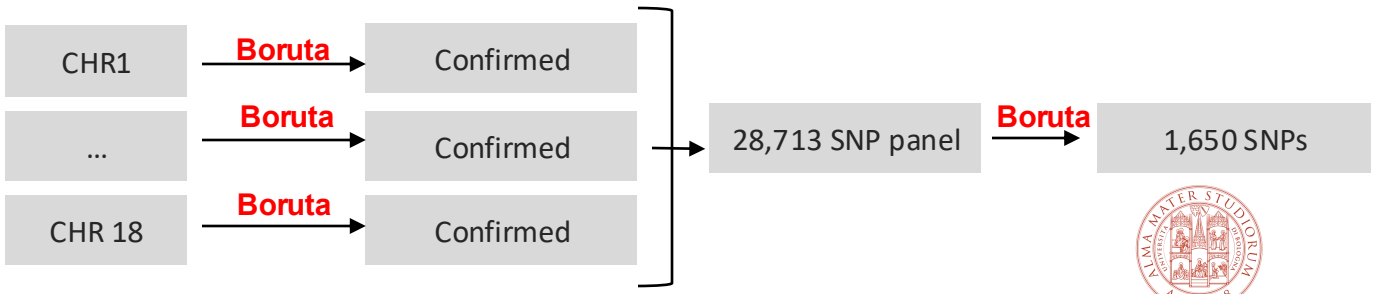
Adopted strategy

1. Boruta for each chromosome → Confirmed SNPs

→ 28,713 SNPs remained.

2. Boruta of the whole set of Confirmed SNPs

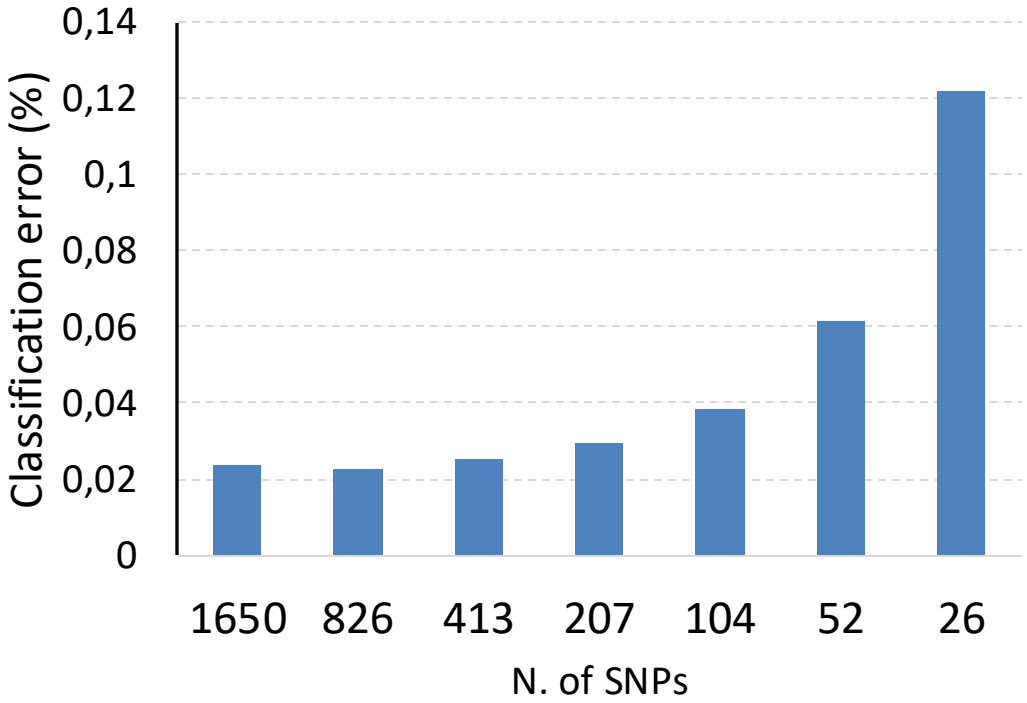
→ 1,650 SNPs remained.



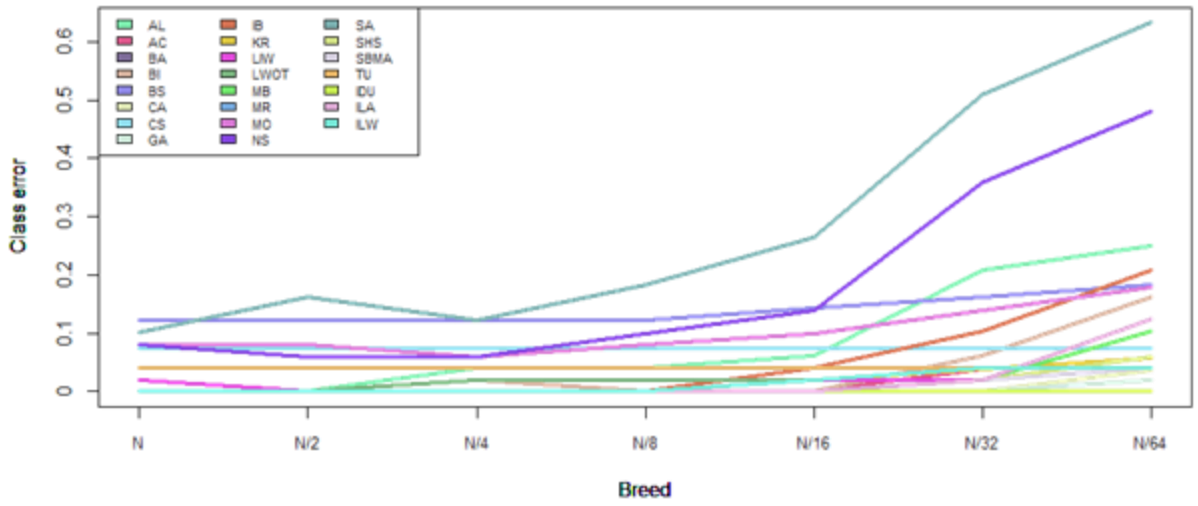
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Workflow of analysis

1. **SNPs (N) used in classic Random Forest**
2. **Testing of reduced set of SNPs**
   → N/2; N/4; N/8; N/16; ..



Low classification error



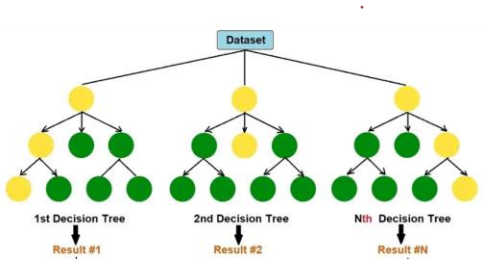
Some breeds "suffer" more of missclassification

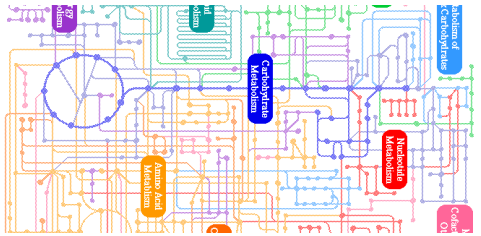
Results: biological insights of selected SNPs

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**

Conclusions

BORUTA results a valid feature-selection algorithms that evaluate variables in context, preserving interactions, correlation structure and nonlinear effects, so selected features (markers) are not only predictive, but also biologically meaningful.





ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

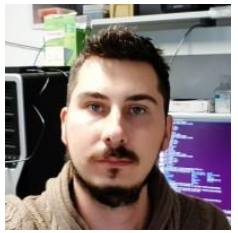
Animal and Food Genomics Group @ DISTAL (UNIBO)

L. Fontanesi
PI, Full Professor

F. Bertolini
Associate Professor

S. Dall'Olio
Associate Professor

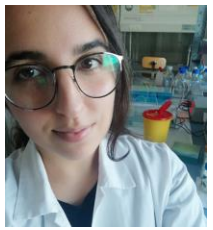
P. Zambonelli
Associate Professor



S. Bovo
Assistant Professor

A. Ribani
Assistant Professor

G. Schiavo
Assistant Professor



V. Taurisano
Research Fellow

M. L. Calabri
PhD Student

S. Agyabeng
PhD Student

G. Zyrianoff
EFSA Fellow

Animal and Food Genomics Group

https://site.unibo.it/animal-and-food-genomics/en

ACKNOWLEDGMENTS

EU-L-PHE EUROPEAN NETWORK ON LIVESTOCK PHENOMICS
Re-Livestock RESILIENT FARMING SYSTEMS
PigPhenomics
ANAS associazione nazionale allevatori suini
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
agritech National Research Center for Technology in Agriculture
FEEDTHEPIG
ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

Thanks for the attention

Questions?

Animal and Food Genomics Group

Animal and Food Genomics Group
Department of Agricultural and Food Sciences
University of Bologna, Italy (IT)

Contacts:
samuele.bovo@unibo.it; luca.fontanesi@unibo.it

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA