Italian Journal of Animal Science

ISSN: 1828-051X (Online) Journal homepage: www.tandfonline.com/journals/tjas20

Comparison of untargeted metabolomic patterns between Italian Landrace and Italian Large White pigs reveals distinct molecular phenotypes

Matteo Bolner, Samuele Bovo, Giuseppina Schiavo, Giuliano Galimberti, Francesca Bertolini, Stefania Dall'Olio, Anisa Ribani, Paolo Zambonelli, Maurizio Gallo & Luca Fontanesi

To cite this article: Matteo Bolner, Samuele Bovo, Giuseppina Schiavo, Giuliano Galimberti, Francesca Bertolini, Stefania Dall'Olio, Anisa Ribani, Paolo Zambonelli, Maurizio Gallo & Luca Fontanesi (2026) Comparison of untargeted metabolomic patterns between Italian Landrace and Italian Large White pigs reveals distinct molecular phenotypes, Italian Journal of Animal Science, 25:1, 190-205, DOI: 10.1080/1828051X.2026.2624165

To link to this article: https://doi.org/10.1080/1828051X.2026.2624165

ITALIAN JOURNAL OF ANIMAL SCIENCE
2026, VOL. 25, NO. 1, 190–205
https://doi.org/10.1080/1828051X.2026.2624165

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Comparison of untargeted metabolomic patterns between Italian Landrace and Italian Large White pigs reveals distinct molecular phenotypes

Matteo Bolner[a]*, Samuele Bovo[a]*, Giuseppina Schiavo[a], Giuliano Galimberti[b], Francesca Bertolini[a], Stefania Dall'Olio[a], Anisa Ribani[a], Paolo Zambonelli[a], Maurizio Gallo[c] and Luca Fontanesi[a]

[a]Dipartimento di Scienze e Tecnologie Agro-alimentari, Gruppo di Genomica Animale e degli Alimenti, Alma Mater Studiorum Università di Bologna, Bologna, Italy; [b]Dipartimento di Scienze Statistiche "Paolo Fortunati", Alma Mater Studiorum Università di Bologna, Bologna, Italy; [c]Associazione Nazionale Allevatori Suini (ANAS), Rome, Italy

ABSTRACT

Breed characteristics are shaped by complex biological mechanisms which, when dissected at the molecular level, can lead to the identification of novel phenotypes useful for the development of phenomics-guided breeding strategies. Building on this, we investigated differences at the metabolome level between the two main maternal lines used in Italian heavy pig crossbreeding schemes: Italian Landrace (ILA) and Italian Large White (ILW), given their relevance in the production of protected designation of origin (PDO) products. We conducted untargeted plasma metabolomics on 205 pigs (53 ILA and 152 ILW), generating high-quality profiles of 576 metabolites. Differentially abundant metabolites were identified using Boruta, a random forest-based feature selection algorithm integrated into a computational pipeline designed to reduce stochastic variability. We also leveraged the unbalanced breed sample sizes to evaluate the influence of population composition and sampling design on metabolite selection. Specifically, ILW animals were either downsampled into three balanced but distinct subsets or included in unbalanced datasets of increasing size. Across all analyses, 59 metabolites differentiated the two breeds. Approximately, 30% were repeatedly identified across all three balanced subsets, and 60–70% appeared in at least two. Inclusion of larger, unbalanced datasets increased the number of selected metabolites by ~25%, indicating improved sensitivity with larger sample sizes while balanced designs enhanced robustness. Biologically, ILW pigs showed elevated gamma-glutamyl amino acids, indicating enhanced glutathione-cycle activity, while ILA pigs exhibited higher hexosylceramides, consistent with greater fat deposition. These findings highlight breed-specific metabolic architectures and support the application of metabolomics in precision breeding for PDO-oriented systems.

HIGHLIGHTS

- Several hundred metabolites were analysed from plasma of Italian Landrace and Italian Large White pigs.
- Distinct metabolic signatures reflect different genetic backgrounds of Italian pig breeds.
- Novel metabolic markers can support improved selection strategies in pig breeding.

ARTICLE HISTORY
Received 15 November 2025
Revised 27 December 2025
Accepted 26 January 2026

KEYWORDS
Metabolic pathway; metabolite; metabolomics; plasma; *Sus scrofa*

Introduction

The global pig production system relies on various divergent breeds and lines for various traits that are utilised in crossbreeding programs to maximise heterotic effects in the production of slaughtered animals (Sellier 1976; Comings and MacMurray 2000; Cassady et al. 2002; Wolf et al. 2006). The Italian pig industry specifically focuses on raising heavy pigs, which are slaughtered at 9 months of age with approximately 170 kg live weight. These pigs are used to produce high-quality dry-cured hams such as Parma and San Daniele protected designation of origin (PDO) hams (Bosi and Russo 2004). This specialised system is based on properly selected pig breeds with specific genetic

CONTACT Samuele Bovo samuele.bovo@unibo.it; Luca Fontanesi luca.fontanesi@unibo.it Dipartimento di Scienze e Tecnologie Agro-alimentari, Gruppo di Genomica Animale e degli Alimenti, Alma Mater Studiorum Università di Bologna, Bologna, Italy
*Both authors contributed equally to this work.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/1828051X.2026.2624165.

programs that meet the required characteristics of this unique industry. The selection objectives include carcases with an adequate fat coverage, high meat quality parameters, with a focus on reducing weight losses of the hams during the curing period and minimising excessive intermuscular fat (Bosi and Russo 2004). Italian Large White (ILW), Italian Landrace (ILA) and Italian Duroc are the primary Italian pig genetic resources specifically selected for this system. These breeds form the foundation of the crossbreeding programs for the Italian heavy pig production system.

Extensive genomic studies have been conducted to characterise these breeds in relation to production and reproductive traits. Their genetic background and selection history are now well documented (Bovo et al. 2020; Bertolini et al. 2024). However, the connection between their genomic diversity and phenotypic differences is only partially understood. The intermediate molecular layers that bridge genotype to phenotype, known as molecular phenotypes or molecular phenome (Houle et al. 2010), remain largely unexplored in these breeds.

One of the most promising proxies for describing the molecular phenome is the metabolome, which refers to the full set of metabolites (small molecules) present in an organism (Fiehn 2002). Metabolites are the final and intermediate products of enzymatic reactions and cellular processes. Their fundamental role in biological pathways makes them valuable indicators to bridge the gap between genome and phenome, and to elucidate the biological context underlying economically relevant traits (Fiehn 2002; Fontanesi 2016). Metabolomics provides a high-throughput characterisation of the metabolome and has gained increasing importance in animal science over the last decade (Goldansaz et al. 2017).

Several studies have utilised metabolomics to investigate various aspects of pig biology, including differences between breeds (Bovo et al. 2016a, 2025a; Hou et al. 2023; Ponsuksili et al. 2025), sex- and fertility-related effects (Bovo et al. 2015, 2025b; Peukert et al. 2021; Zhang et al. 2021; Ponsuksili et al. 2025), feeding and nutritional challenges (Luise et al. 2020; Wang et al. 2021; Metzler-Zebeli et al. 2023), heat stress response (Qu and Ajuwon 2018) and the genetic architecture of the metabolome (Wang and Kadarmideen 2020; Dervishi et al. 2023; Bovo et al. 2025c; Ponsuksili et al. 2025). In terms of Italian breeds, most of the metabolomic studies we have conducted have focused on Italian Duroc and ILW pigs (Bovo et al. 2015, 2016a, 2016b, 2023, 2025a, 2025b, 2025c). These breeds represent the paternal line and part of the maternal lines in three-way crossbreeding schemes, where ILA is typically included to complete the maternal crossbred component. However, ILA pigs remain largely unexplored in terms of characterisation of their metabolome (Bovo et al. 2025d).

In this study, we characterised the plasma metabolome of ILA pigs for the first time using a high-throughput untargeted metabolomic platform that provided information from over 500 plasma metabolites. The metabolomic profile obtained was compared with that of ILW pigs, which make up the other maternal line in the three-way crossbreeding schemes described. This comparison was conducted using a random forest based computational pipeline that we had previously developed, validated and applied to discover discriminant metabolic features in livestock populations (Bovo et al. 2025a, 2025b). This methodology was supported by a statistical procedure that addressed unbalanced datasets to evaluate the effect of sample size and subsampling strategies on metabolite selection. We identified breed-specific metabolic differences that may help explain the varying performances between the two breeds.

Materials and methods

Animals, blood and plasma samples

The study was conducted in accordance with the ARRIVE guidelines. Animals used in the study were kept in compliance with Italian and European legislation for pig production (DLgs 122/2011 and Council Directive 2008/120/EC). Ethical review and approval were not required for this study as the animals were not specifically raised or treated for the study. Blood samples were collected post-mortem during regular slaughter at a commercial slaughterhouse, following guidelines established by Italian and European Union regulations for animal care and slaughter (Council Regulation (EC) No. 1099/2009).

This study involved a total of 205 pigs, including: (i) 53 ILA animals (37 intact gilts and 16 castrated males) slaughtered across six non-consecutive days and (ii) 152 ILW animals (100 intact gilts and 52 castrated males), selected to match the same slaughter days. The ILW pigs were part of a larger population previously studied by Bovo et al. (2025a, 2025b, 2025c). The litters included in the study were not fully independent, as some shared common sires and therefore represented half-sib relationships. In the ILA dataset, the 53 pigs originated from 37 litters generated by crossing 21 sires and 37 dams, with 11 sires shared across multiple litters. In the ILW dataset, the 152 pigs

originated from 115 litters generated by crossing 46 sires and 115 dams, with 22 sires shared across multiple litters. All pigs were part of sib-testing programs consisting of littermate triplets (one castrated male and two females), that were individually performance-tested at the National Pig Breeder Association's Central Station. Testing began at 30–45 days of age and continued until pigs reached approximately 155 ± 5 kg. The animals were uniformly managed and fed, with semi-ad libitum access to a commercial fattening diet from ∼100 kg onwards. After a ∼12 h fast, the pigs were transported to a commercial abattoir and were slaughtered at ∼08:00 am following electrical stunning. Blood was collected post-jugulation from the carotid artery into EDTA tubes, gently inverted 8–10 times, and centrifuged within 2 h at 2420 × *g* for 10 min at +4 °C to obtain plasma, which was stored at −80 °C until metabolomic analysis.

Metabolomic profiling of plasma samples

Untargeted metabolomic profiling of plasma samples was conducted by Metabolon, Inc. (Durham, NC) using the HD4 metabolomics platform, which utilises methanol-based extraction and multiple liquid chromatography–tandem mass spectrometry techniques. Metabolon provided metabolite identification and annotation, including chemical names, database identifiers and associated biological pathways. The final profile included 1029 metabolites grouped in nine metabolic super pathways (i.e. lipids, amino acids, nucleotides, peptides, cofactors and vitamins, carbohydrates, partially characterised molecules, energy and xenobiotics) in addition to unnamed metabolites. More information about the metabolomic platform, surveyed metabolites and pathways can be found in Bovo et al. (2025a).

Data analyses

Considering the variety of statistical analyses carried out, as described in detail in the following sections, Figure 1 provides a concise overview of the main approaches that we applied to extract relevant metabolomic information from the large panel of metabolites analysed. These approaches included: (1) data processing and quality control (QC), (2) construction of five balanced and unbalanced datasets, (3) identification of breed-associated metabolites and (4) assessment of the discriminative power of the selected metabolites along with their functional characterisation.

Quality control, data imputation and cleaning

Quality control and data imputation were conducted as outlined in our previous study (Bovo et al. 2025a), treating each breed separately. First of all, metabolites annotated as xenobiotics by Metabolon were excluded from the analysis. This was done because xenobiotic metabolites are not synthetised by the organism, but rather introduced through food, drugs or pollutants. Therefore, their presence and concentration are not determined by the genetic background of the animal, but by the environment in which it is raised.

Then, outlier values at the metabolite level were removed from the dataset. Outliers were identified as those that deviated by five times the interquartile

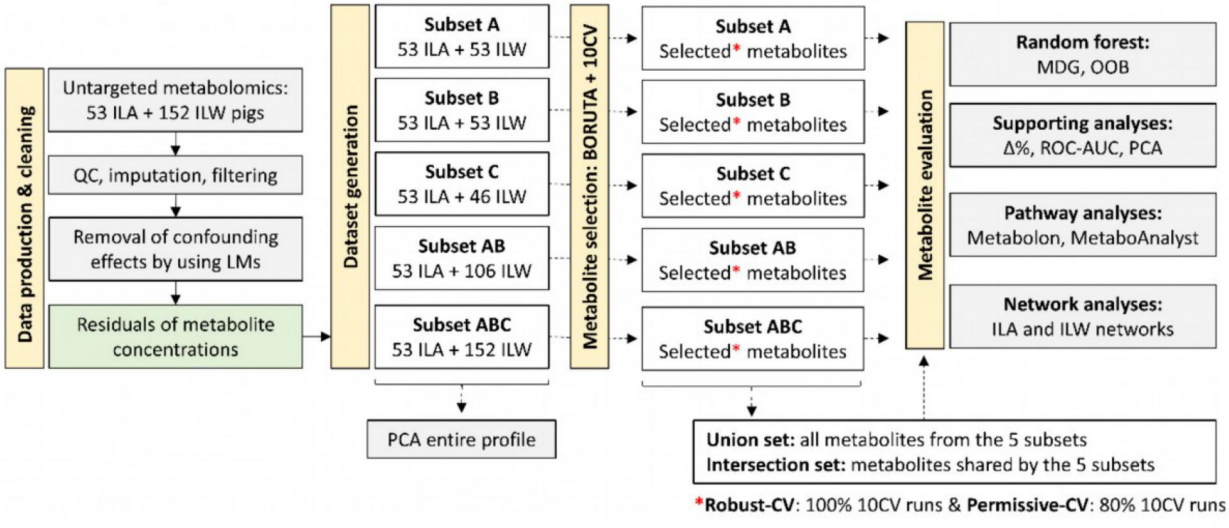

Figure 1. Outline of the data analysis pipeline. CV: cross validation; ILA: Italian Landrace; ILW: Italian Large White; LM: linear model; MDG: mean decrease Gini; OOB: out-of-bag; PCA: principal component analysis; C: quality control; ROC-AUC: receiver operating characteristic curve and area under the curve; Δ%, relative difference in metabolite concentration.

range below or above the median of the metabolite. After marking these outlier values as missing, metabolites with more than 25% missing values were excluded from the dataset. At the sample level, we considered outliers and marked for removal those samples with more than 30% of values missing across the profile.

Missing values still present in the dataset were then imputed using the multivariate imputation by chained equations (MICEs) approach (Azur et al. 2011). Following the recommendation of Faquih et al. (2020), we utilised predictive mean matching as the imputation algorithm generating five distinct datasets, in line with MICE guidelines. Each dataset underwent multiple imputations (with different random state seeds) resulting in 25 imputed datasets for each breed. Subsequently, the ILW and ILA datasets were combined based on the corresponding random state seed and imputation cycle. Confounding factors were then eliminated by regressing each metabolite abundance on sex, carcase weight and sampling day (Bovo et al. 2025a), resulting in residuals used for subsequent statistical analyses. All analyses were performed using Python v3.11.7 and R v.4.2.3 (R Core Team 2024); figures were created using the Python libraries Seaborn (Waskom 2021) and Marsilea (Zheng et al. 2025).

Definition of datasets used for bioinformatics analyses

Analyses were conducted on datasets of varying compositions and sizes to maximise the availability of ILW data, with the sample size being three times greater than that of ILA pigs. This approach aimed to address the limited sample size of ILA pigs and also allowed for the evaluation of the impact of imbalanced group sizes on the results. The ILW dataset was divided into three subsets, each matching the size of the ILA dataset. These subsets were referred to as: (i) 'subset A', consisting of 53 ILA pigs and 53 ILW pigs, (ii) 'subset B', consisting of 53 ILA pigs and 53 ILW pigs and (iii) 'subset C', consisting of 53 ILA pigs and 46 ILW pigs. Each subset was balanced with respect to the slaughter day and sex of the animals and comprised a unique, non-overlapping set of ILW samples. Additionally, the subsets were combined as follows: (i) 'subset AB', including 53 ILA pigs and 106 ILW pigs (combined subsets A and B) and (ii) 'subset ABC', including 53 ILA pigs and the 152 ILW pigs (combined subsets A, B and C). These five datasets were defined to account for the impact of different sample sets (subsets A, B and C) and increasingly imbalanced sample sets (subsets AB and ABC).

Identification of differentially abundant metabolites between pig breeds

The identification of differentially abundant metabolites between pig breeds was performed following the analytical framework detailed by Bovo et al. (2025a). All the analyses described below have been consistently applied to each of the five previously defined subsets.

Briefly, residuals of metabolite abundances were analysed using a combination of multivariate statistical approaches, including principal component analysis (PCA) and Boruta, a machine learning-based feature selection method built on random forests (Kursa et al. 2010). Specifically, Boruta was employed as a supervised algorithm to identify metabolites associated with breed differences. To account for the imbalance of class sample size in datasets AB and ABC, the random forest model was built using the 'balanced' class weight option, adjusting the weight of input data to be inversely proportional to the frequency of classes. In contrast, PCA was adopted as an unsupervised approach to assess the discriminative power of the entire metabolomic profile as well as the subset of metabolites identified by Boruta. PCA was computed using the scikit-learn Python library (Pedregosa et al. 2011), after scaling variables to unit variance. The scikit-learn was used for building the random forest classification models, in combination with the boruta_py package (https://github.com/scikit-learn-contrib/boruta_py) for the feature selection procedure. Metabolites that were labelled as 'confirmed' by Boruta were considered as selected and therefore relevant for classification purposes. Boruta was run five times with five different random seeds on each of the 25 differently imputed datasets. This was done to account for the random component of feature selection. Then, to account for the effect of the samples included in the dataset, we performed an additional 10-fold cross validation (10CV) procedure, by splitting each of the imputed datasets randomly into ten equal parts and running Boruta on 10 rotations, each excluding 1/10th of the dataset. This resulted in a total of 1250 CV runs. Discriminant metabolites were then categorised into two groups based on their reliability of selection: (i) the 'permissive-CV' set, including those metabolites that were confirmed in at least 1000 Boruta runs (80%) and (ii) the 'robust-CV' set, including those metabolites that were confirmed in all 1250 Boruta runs (100%). Results of the five sample subsets

belonging to the 'Permissive-CV' set were then compared by defining two levels of overlap: (i) 'union set', including the whole set of metabolites selected in any of the five sample subsets and (ii) 'intersection set', including metabolites always selected across the five sample subsets.

We then analysed each discriminant metabolite set using a standard random forest analysis (Schiavo et al. 2024), by training a random forest model on the selected metabolites. The mean decrease Gini (MDG) was computed and used to rank the metabolites by feature importance. The out-of-bag (OOB) error was then computed to evaluate the classification power of the metabolite sets. The predictive performance of discriminant metabolites was also assessed with the receiver operating characteristic (ROC) curve and the resulting area under the curve (AUC). The relative difference in concentration between breeds was also assessed for each metabolite, expressed as $\Delta\%_i = \frac{\overline{x}_i^S - \overline{x}_i^P}{\overline{x}_i^S} \times 100$, with $\overline{x}_i^S$ and $\overline{x}_i^P$ being the average metabolite abundance of the ith metabolite in ILW and ILA pigs, respectively.

Comparison of results from subsets A, B and C was based on the relative mean absolute deviation (RMAD) expressed as a percentage, computed for each metabolite (for both $\Delta\%$ and AUC values) as follows: $\text{RMAD}\%_x = \frac{MAD_x}{|\overline{x}|} \times 100$, where $MAD_x = \frac{\sum |x_i - \overline{x}|}{n}$, with x represents a metabolite, i the index of individual observations (corresponding to subsets A, B and C) and $n = 3$, since three subsets were compared.

Pathway analysis

Pathway analysis was conducted using over-representation analysis. Initially, the analysis considered both super- and sub-pathways provided by Metabolon. Subsequently, MetaboAnalyst v6.0 (Pang et al. 2024) was utilised with metabolite identifiers from the Human Metabolome Database (HMDB; Wishart et al. 2022) as input and the RaMP-DB dataset (consisting of 3694 metabolites and lipid pathways) as reference (Braisted et al. 2023). In both instances, terms with a false discovery rate corrected $P_{FDR} < 0.05$ and including at least two of the input metabolites were retained and deemed over-represented.

Network analysis

Selected metabolites were investigated through network and pathway analyses. Networks were computed using Pearson's correlation coefficients (r). For each combination of sample subset and metabolite subset, two correlation networks were investigated: one built using correlations between ILA samples, and one built using correlations between ILW samples. Correlations were considered informative if they presented $r > 0.5$, regardless of their p. Graph-tool v2.97 (Peixoto 2014) was used to build and visualise the networks as well as to extract basic statistics such as node degree and betweenness centrality. The Jaccard similarity index (J) was used to compute the similarity between networks; the edge-based similarity, comparing the overlap of edge sets between two networks, was adopted (Fuxman Bass et al. 2013).

Results

Overview of the metabolomic dataset

A total of 1029 metabolites were measured across 205 samples, distributed over nine super-pathways, in addition to unnamed metabolites. In total, 107 sub-pathways (excluding unnamed metabolites) were covered (Supplementary Figure S1 and Table S1). The largest group was from the lipid super pathway ($n = 463$, 45%), followed by amino acids ($n = 206$, 20%) and unnamed metabolites ($n = 123$, 12%). A total of 77 metabolites (7.5%) were classified as xenobiotics and excluded from the analysis. Metabolites were filtered independently for the two pig breeds. After QC, including the removal of metabolites with excessive missing values or outliers, 576 metabolites were retained. None of the samples had more than 30% of missing values, and therefore all samples were retained for the analysis. The final metabolite dataset covered eight super-pathways and 92 sub-pathways (excluding unnamed metabolites) and broadly maintained the initial distribution across metabolite groups (Supplementary Figure S1).

Identification of metabolites differentiating the two pig breeds

Feature selection with Boruta identified 59 metabolites under the permissive-CV criterion (Table 1; Supplementary Table S2). The number of these metabolites varied across sample subsets, from 33 in subset B to 50 in subset AB (Figure 2(a)). Under the robust-CV criterion, 44 metabolites were retained, ranging from 21 in subset C to 39 in subset ABC (Figure 2(b)). Balanced subsets (A, B, C) yielded comparable results, while larger, unbalanced subsets (AB, ABC) identified more metabolites, suggesting an influence of sample size and composition.

Table 1. Metabolites ($n = 59$) differentially abundant between Italian Landrace and Italian Large White pigs.

Sub-pathways[a]	Metabolites[b]
Amino acid	
Alanine and aspartate metabolism	Asparagine (↓)
Creatine metabolism	Creatinine (↓)
Glutamate metabolism	**Glutamine** (↓)
Glutathione metabolism	cys-gly, oxidised (↓); **cysteinylglycine disulphide** (↓)
Guanidino and acetamido metabolism	1-Methylguanidine (↓)
Histidine metabolism	**1-Methyl-5-imidazolelactate** (↓)
Lactoyl amino acid	N-lactoyl leucine (↑)
Leucine, isoleucine and valine metabolism	3-Hydroxy-2-ethylpropionate (↑); alpha-hydroxyisocaproate (↑); beta-hydroxyisovalerate (↓)
Lysine metabolism	Fructosyllysine (↓)
Methionine, cysteine, SAM and taurine metabolism	S-methylcysteine (↑)
Tryptophan metabolism	Tryptophan (↑)
Urea cycle; arginine and proline metabolism	N-methylproline (↓)
Carbohydrate	
Aminosugar metabolism	N-acetylglucosamine/N-acetylgalactosamine (↑)
Fructose, mannose and galactose metabolism	Fructose (↑)
Glycolysis, gluconeogenesis and pyruvate metabolism	1,5-Anhydroglucitol (1,5-AG) (↑); pyruvate (↓)
Cofactors and vitamins	
Nicotinate and nicotinamide metabolism	Trigonelline (N′-methylnicotinate) (↓)
Riboflavin metabolism	Flavin adenine dinucleotide (FAD) (↓)
Tocopherol metabolism	Gamma-tocopherol/beta-tocopherol (↓)
Lipid	
Corticosteroids	Cortisone (↑)
Fatty acid metabolism (acyl carnitine, dicarboxylate)	**Adipoylcarnitine (C6-DC)** (↓)
Fatty acid, amino	2-Aminooctanoate (↓)
Hexosylceramides (HCERs)	Glycosyl ceramide (d18:1/20:0, d16:1/22:0) (↑); **glycosyl-N-palmitoyl-sphingosine (d18:1/16:0)** (↑); **glycosyl-N-stearoyl-sphingosine (d18:1/18:0)** (↑)
Lysophospholipid	1-Lignoceroyl-GPC (24:0) (↑)
Lysoplasmalogen	1-(1-Enyl-stearoyl)-GPE (P-18:0) (↑)
Metabolite	**Ceramide (d18:1/17:0, d17:1/18:0)** (↑)
Phosphatidylcholine (PC)	1-Linoleoyl-2-arachidonoyl-GPC (18:2/20:4n6) (↑)
Phosphatidylethanolamine (PE)	1-Linoleoyl-2-arachidonoyl-GPE (18:2/20:4) (↑)
Phospholipid metabolism	Choline (↑); trimethylamine N-oxide (↓)
Sphingomyelins	Lignoceroyl sphingomyelin (d18:1/24:0) (↑); sphingomyelin (d17:1/16:0, d18:1/15:0, d16:1/17:0) (↑)
Nucleotide	
Pyrimidine metabolism, uracil containing	5,6-Dihydrouridine (↓)
Peptide	
Dipeptide	Isoleucylglycine (↓); valylglutamine (↓); valylglycine (↓)
Gamma-glutamyl amino acid	**Gamma-glutamylcitrulline** (↓); gamma-glutamylglutamate (↓); **gamma-glutamylglutamine** (↓); gamma-glutamylglycine (↓); **gamma-glutamylhistidine** (↓); **gamma-glutamylleucine** (↓); gamma-glutamylmethionine (↓); **gamma-glutamylphenylalanine** (↓); **gamma-glutamylthreonine** (↓); **gamma-glutamyltyrosine** (↓); **gamma-glutamylvaline** (↑)
Unknown	
Unknown	X-11787 (↓); X-18913 (↓); X-23423 (↑); **X-23593** (↓); X-23997 (↑); X-25265 (↓); **X-25343** (↓)

Detailed statistics are given in Supplementary Table S2.
[a]Classification provided by Metabolon.
[b]The direction of the arrow refers to the average metabolite concentrations of ILA pigs in respect to ILW pigs: higher in ILA (↑) and lower in ILA (↓). In bold are highlighted the 17 metabolites forming a core set confirmed by the separated analysis of the three replicated cohorts (datasets A, B and C).

In subsets A, B and C, 54 metabolites were selected under the permissive-CV criterion, but only 17 were consistently identified across all three (and confirmed also in subsets AB and ABC; Figure 2(a)). An additional 16 metabolites were shared by two subsets, and 21 were unique to a single subset (A: 7, B: 6, C: 8). This highlights a substantial sampling effect, with only 61% of the metabolites being repeatedly selected when altering the composition of the population. Comparable patterns were observed under robust-CV (Figure 3(b)), and despite the strong support in metabolite selection, the low reproducibility across subsets further emphasises the sensitivity of selection to sample composition.

To better assess the impact of samples, overlaps among these five subsets were evaluated using two inclusion criteria. The union set, which includes all metabolites selected in any of the five subsets, comprised the 59 above-mentioned metabolites. Of these, 54 were identified in subsets A, B and C, while the remaining five were only detected in the larger subsets (AB and ABC): four (choline, alpha-hydroxyisocaproate, X-23997, and pyruvate) in subset AB, and one (gamma-tocopherol/beta-tocopherol) in subset

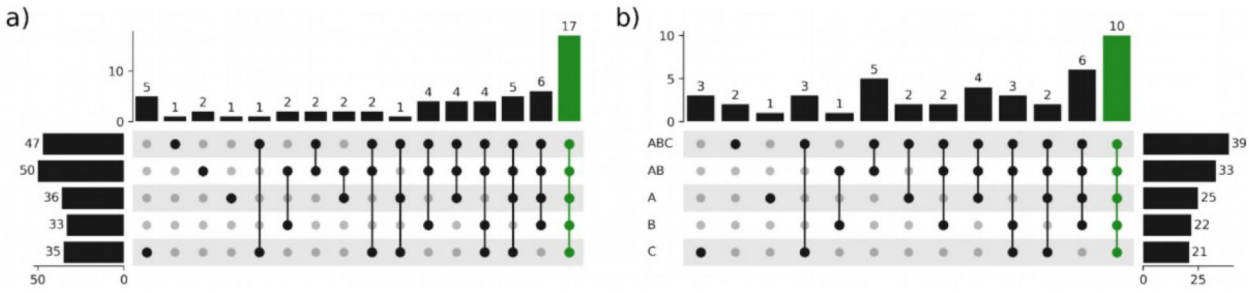

Figure 2. UpSet plot of Boruta-selected metabolites across the five analysed subsets (A, B, C, AB and ABC) under the permissive-CV criterion (a) and the robust-CV criterion (b). In each plot, the upper barplot shows the number of metabolites in each intersection of the five subsets, while the connected dots in the lower panel indicate which subsets are involved in a given intersection. The horizontal bars represent the total number of metabolites selected in each subset. Overall, all metabolites together form the union set, whereas the green bars highlight those included in the intersection set.

ABC. Notably, choline and alpha-hydroxyisocaproate, although present in subset AB, were not confirmed in subset ABC, suggesting limited reproducibility. Under robust-CV, subset AB contributed only one additional metabolite (pyruvate), which was also identified in subset ABC. Subset ABC did not contribute any further metabolite.

Biological properties of selected metabolites

The 59 selected metabolites were assigned to seven of the nine evaluated super-pathways: amino acids (15), lipids (15), peptides (14), carbohydrates (4), cofactors and vitamins (3), nucleotides (1) and seven unnamed compounds (Table 1; Figure 3; Supplementary Table S2). Notably, both the peptide and lipid super-pathways were significantly over-represented ($P_{FDR} < 0.05$; Supplementary Table S3).

The lipid-related metabolites were distributed across 11 sub-pathways. Among these, hexosylceramides (HCERs; three metabolites), sphingomyelins (two metabolites) and phospholipid metabolism (two metabolites) accounted for half of the lipid-related sub-pathways. The amino acid-related metabolites, equally abundant, encompassed 12 sub-pathways, with leucine, isoleucine and valine metabolism (three metabolites), and glutathione (GSH) metabolism (two metabolites) being the most represented. Peptides were classified into two sub-pathways: gamma-glutamyl amino acids (11 metabolites) and dipeptides (three metabolites). Enrichment analysis at the sub-pathway level confirmed significant over-representation ($P_{FDR} < 0.05$; Supplementary Table S4) of (i) gamma-glutamyl amino acids, (ii) HCER and (iii) dipeptides.

Pathway enrichment analysis was also performed using MetaboAnalyst. Of the 59 selected metabolites, 48 could be mapped to valid HMDB identifiers. This analysis identified 27 significantly enriched pathways ($P_{FDR} < 0.05$; Supplementary Table S5), involving only 11 of the 48 mapped metabolites. Among the top enriched pathways were those related to molecular transport and nitrogen metabolism, the latter including L-asparagine, pyruvic acid, glutamine and FAD. A subsequent group of five enriched pathways was consistently driven by the same set of four metabolites (N-acetylgalactosamine, pyruvic acid, glutamine and D-fructose), encompassing amino sugar metabolism and four disease-related pathways (e.g. Sialuria and Tay-Sachs disease). Additionally, though less significant, enrichments included the diet-dependent trimethylamine/trimethylamine N-oxide pathway, aspartate metabolism and hypoacetylaspartia.

Mean difference in concentration of selected metabolites

For the 59 metabolites, the Δ% ranged from −181% (lower in ILA; gamma-glutamylhistidine; subset B) to +54% (higher in ILA; gamma-glutamylvaline; subset C) (Figure 3; Supplementary Table S2). When comparing $|\Delta|$% values (Figure 4(a)), subset-specific metabolites showed lower average differences (46–61%), followed by those in the union set (59 metabolites; 42–46%), whereas the 17 metabolites consistently selected across all five subsets (intersection set) displayed the highest average $|\Delta|$% values (65–74%). This trend suggests that higher selection consistency is associated with both improved discriminative power and larger differences in abundance between breeds. Similar results were obtained for metabolites restricted with the robust-CV criterion (Figure 4(b)).

Balanced subsets (A, B, C) presented some differences, with subset A showing slightly higher values than subsets B and C. Measured by RMAD% (Supplementary Table S2), dispersion was below 35% for all metabolites except for S-methylcysteine (63%)

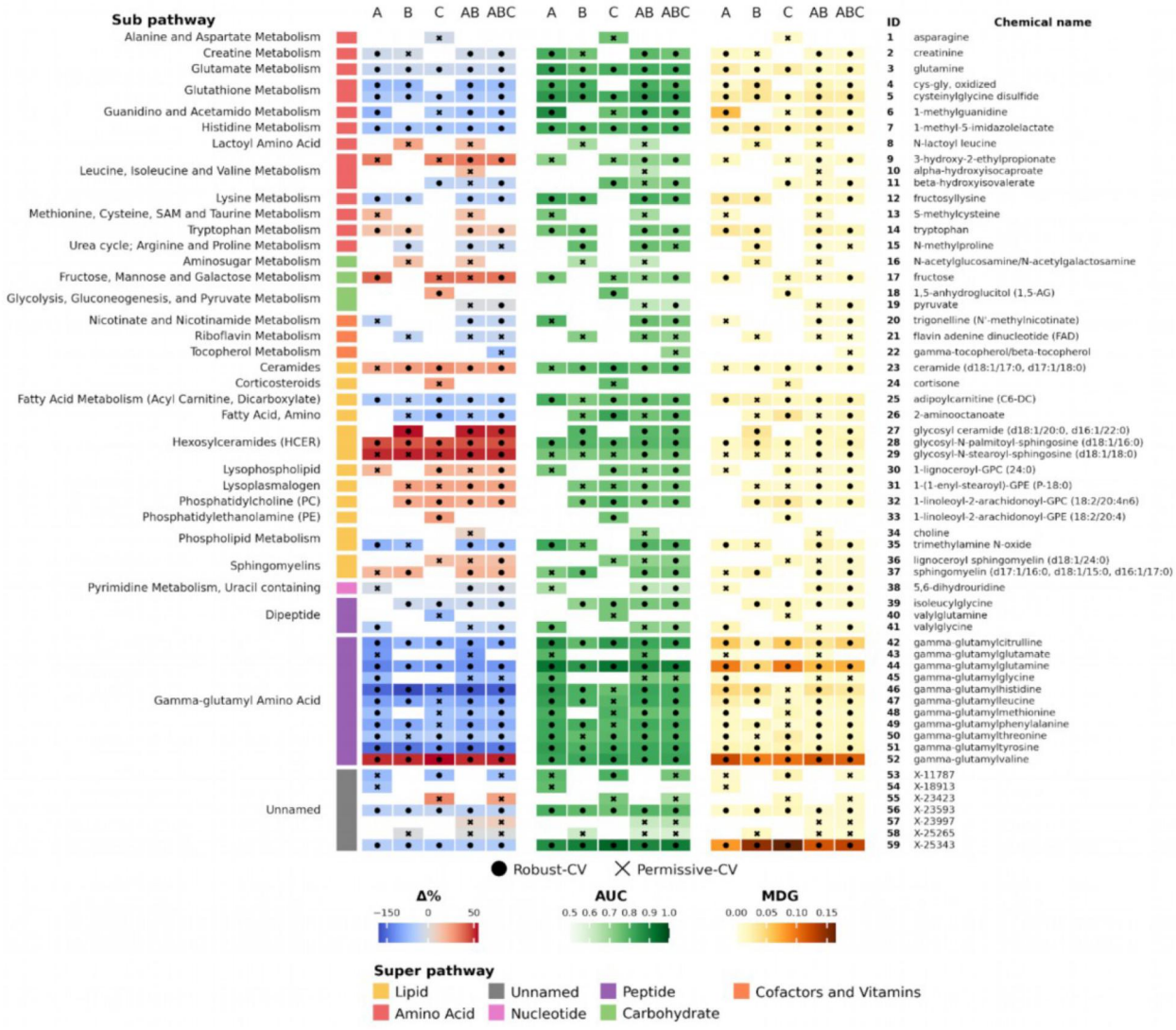

Figure 3. Features of the 59 selected metabolites. On the left side of the main plot are reported the sub-pathway name and super pathway colour of the metabolite; each of the three heatmaps is divided into five columns, one per sample subset (A, B, C, AB and ABC). In each heatmap, cells that are coloured represent metabolites that were selected in that particular sample subset, either with a permissive-CV strategy (represented by a dot) or with a robust-CV strategy (represented by an X symbol). Empty cells indicate that the metabolite was not selected in that sample subset. The leftmost heatmap shows the mean difference between ILA and ILW metabolite abundances (Δ%); the central heatmap shows the ROC AUC level, and the rightmost heatmap shows the mean Gini decrease (MDG). On the right of the heatmaps, metabolites are reported with a progressive identifier (ID) and their chemical name as defined by Metabolon. Full information is given in Supplementary Table S2.

and X-25265 (103%); the latter showed inconsistent Δ% across subsets, being slightly positive in subset A (0.71%) but negative in subset B (−2.55%) and subset C (−2.09%), leading to an RMAD of 103%.

Among the 59 selected metabolites, lipids were mostly more abundant in ILA, with only three out of 15 (20%) showing higher concentrations in ILW. Based on statistics coming from the subset ABC, the most extreme negative Δ% values in ILW (Figure 3; Supplementary Table S2) were observed for 2-aminooctanoate (sub-pathway: fatty acid, Amino), adipoylcarnitine (C6-DC) (sub-pathway: fatty acid metabolism (acyl carnitine, dicarboxylate)) and trimethylamine N-oxide (sub-pathway: phospholipid metabolism). Conversely, ILA exhibited the highest positive Δ% values (Figure 3; Supplementary Table S2) for three lipids from the HCER sub-pathway: glycosyl ceramide (d18:1/20:0, d16:1/22:0), glycosyl-N-stearoyl-sphingosine (d18:1/18:0) and glycosyl-N-palmitoyl-sphingosine (d18:1/16:0).

In contrast to lipids, amino acids were generally more abundant in ILW, with only five out of 15 (33%) being higher in ILA. The most extreme negative Δ% values for ILA (−62% to −71%) (Figure 3; Supplementary Table S2) were observed for cys-gly

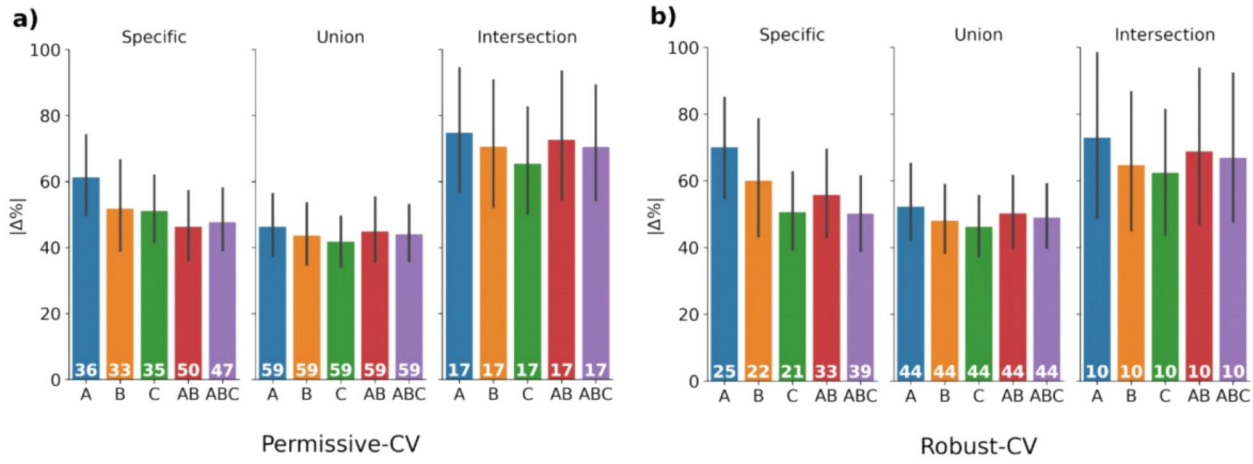

Figure 4. Average $|\Delta|$% (and s.d.) across subset-specific, union and intersection metabolite sets for (a) permissive-CV metabolites and (b) robust-CV metabolites. The number of metabolites considered in the analysis is provided at the bottom of each bar.

oxidised (sub-pathway: GSH metabolism), 1-methyl-5-imidazolelactate (sub-pathway: histidine metabolism) and 1-methylguanidine (sub-pathway: guanidino and acetamido metabolism). On the other hand, a few amino acids showed higher concentrations in ILA, including 3-hydroxy-2-ethylpropionate and alpha-hydroxyisocaproate (both from the leucine, isoleucine and valine metabolism sub-pathway), as well as N-lactoyl leucine (sub-pathway: lactoyl amino acid), with Δ% values ranging from 16 to 34% (Figure 3; Supplementary Table S2).

Selected peptides, including nine molecules from the gamma-glutamyl amino acid sub-pathway and two molecules from the dipeptides sub-pathway, were consistently lower in ILA, with Δ% ranging from −24% to −164%, except for gamma-glutamylvaline, which showed higher concentrations in ILA (+51%) (Figure 3; Supplementary Table S2).

Discriminant power of selected metabolites

Principal component analysis based on the entire metabolite profiles showed minimal separation between pig breeds (Figure 5(a)), unlike the clearer separation based on PC1 observed when analyses were conducted using only the selected metabolites (Figure 5(b–e)).

The discriminative power of individual metabolites was evaluated using AUC values from ROC analyses. Among the 59 selected metabolites, the lowest AUC observed was 0.52 (S-methylcysteine; subset C) while the highest was 0.95 (X-25343; subset C) (Figure 3; Supplementary Table S2). Subset-specific metabolites had average AUC values ranging from 0.75 ± 0.07 to 0.81 ± 0.04 (Figure 6). Generally, AUC values increased progressively from the union set to the intersection set and were higher under the robust-CV criterion compared to permissive-CV (Figure 6), indicating improved discriminative power with greater selection consistency. However, increasing the sample size did not result in higher or more stable AUC estimates at the single-metabolite level.

Figure 6 also highlights significant differences across subsets A, B and C, once again indicating population-specific effects. One notable example was S-methylcysteine, which exhibited the highest RMAD% (12%; Supplementary Table S2) due to highly variable AUCs: 0.74 in subset A, 0.62 in B, and 0.52 in C. A positive correlation was observed between the RMAD% of Δ% and AUC, suggesting that variability in the sampled populations similarly influences both abundance differences and classification power.

Random forest models were constructed using the metabolites identified in each subset to evaluate classification accuracy (OOB) and metabolite importance (MDG scores). OOB errors (Supplementary Table S6) ranged from 2% (permissive-CV, dataset C) to 10% (robust-CV, subset B). Among the balanced subsets A, B and C, subset B consistently displayed the highest classification error (7–10%), while subsets A and C showed lower errors (2–4%), further emphasising the impact of sample composition on model performance. In subsets B, C and ABC, metabolites with the highest AUCs also tended to have the highest MDG scores (Figure 3; Supplementary Table S2). In other subsets, this relationship was less clear. Nonetheless, a moderate correlation between AUC and MDG was observed across all subsets, ranging from $r = 0.52$ (subset AB) to $r = 0.87$ (subset C).

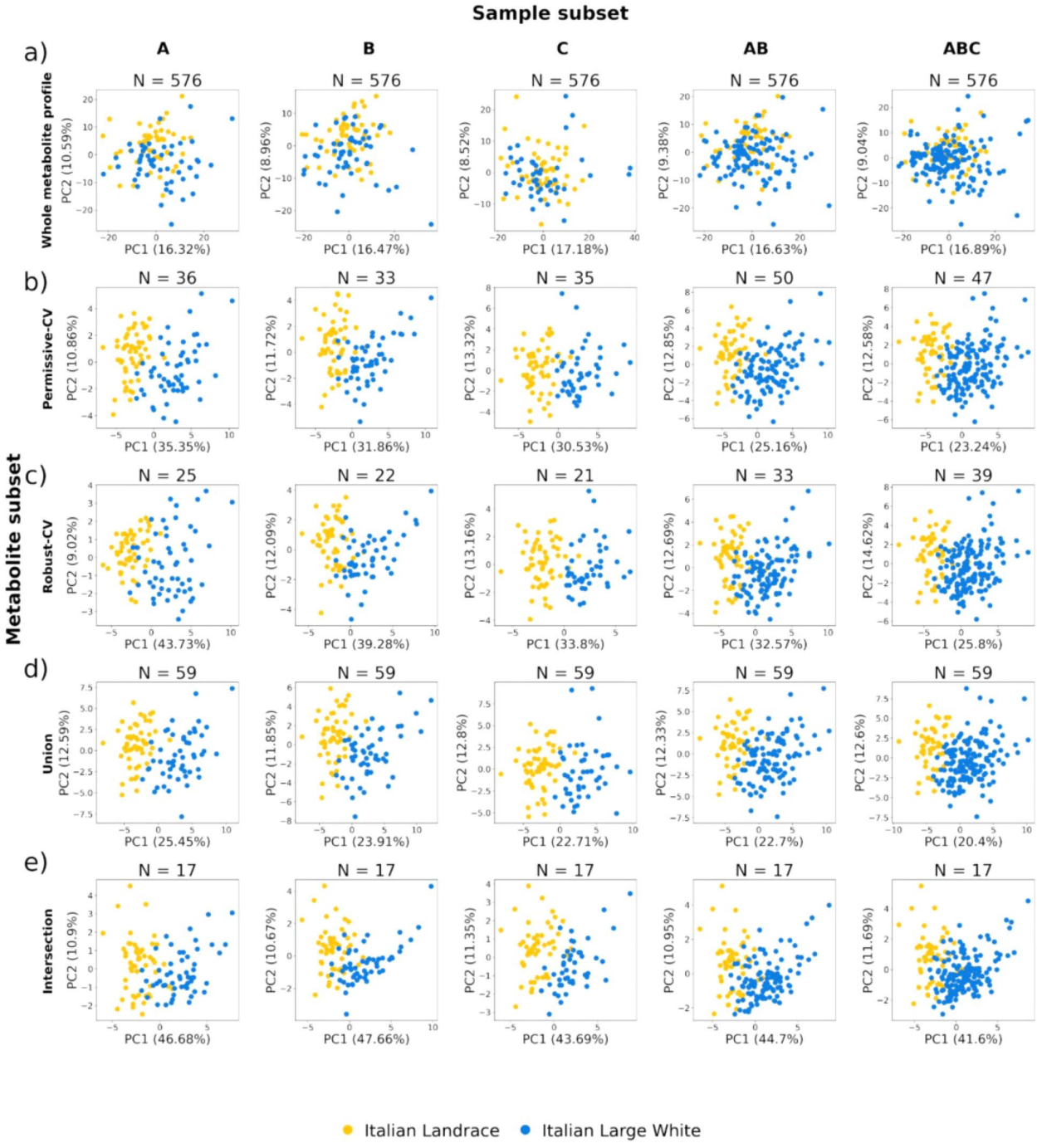

Figure 5. Principal component analysis (PCA) based on: (a) whole-profiles (n. 576 metabolites), (b) subset-specific metabolites based on the permissive-CV criterion and (c) subset-specific metabolites based on the robust-CV criterion, (d) metabolites of the union set and, (e) metabolites of the intersection set. Plots include principal component 1 (PC1) and PC2.

Correlation networks of selected metabolites

Metabolite correlation networks were constructed separately for ILA and ILW using the 59 selected metabolites. Information about nodes and edges is provided in Supplementary Tables S1 and S7, respectively.

In ILA, the resulting network consisted of 38 connected nodes, 87 edges and 21 singletons (Figure 7(a)). Two main modules emerged: one composed exclusively of lipids ($n = 10$) and another containing 28 non-lipid metabolites, including 13 peptides, 10 amino acids and five other compounds. The lipid module contained two sub-clusters: (i) a smaller one including three HCERs and (ii) a larger one composed of the remaining seven lipid-related metabolites. The second, non-lipid module included gamma-glutamyl-threonine, the metabolite with the highest

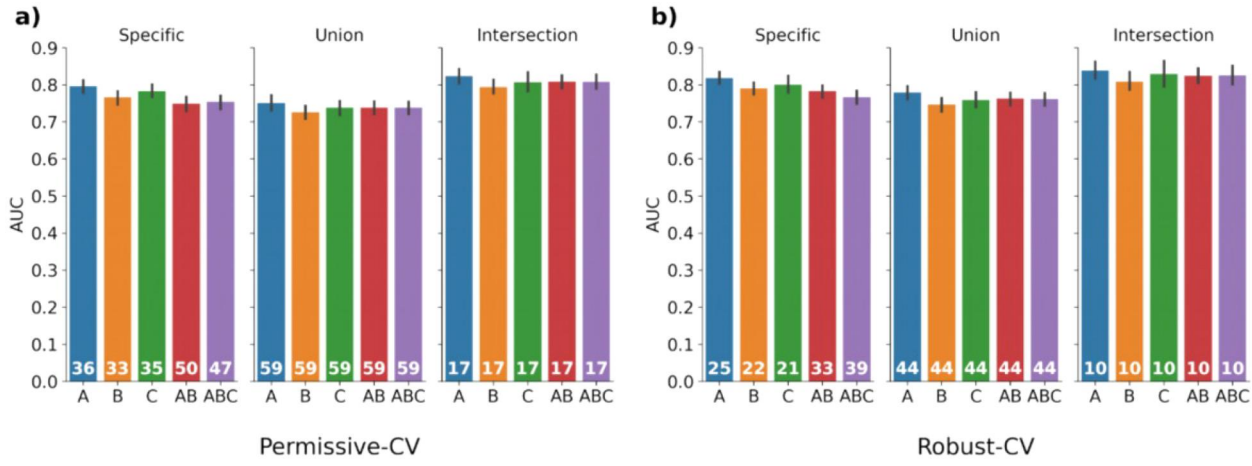

Figure 6. Average area under the curve (AUC) and standard deviation across subset-specific, union and intersection metabolite sets for (a) permissive-CV metabolites and (b) robust-CV metabolites. The number of metabolites considered in the analysis is provided at the bottom of each bar.

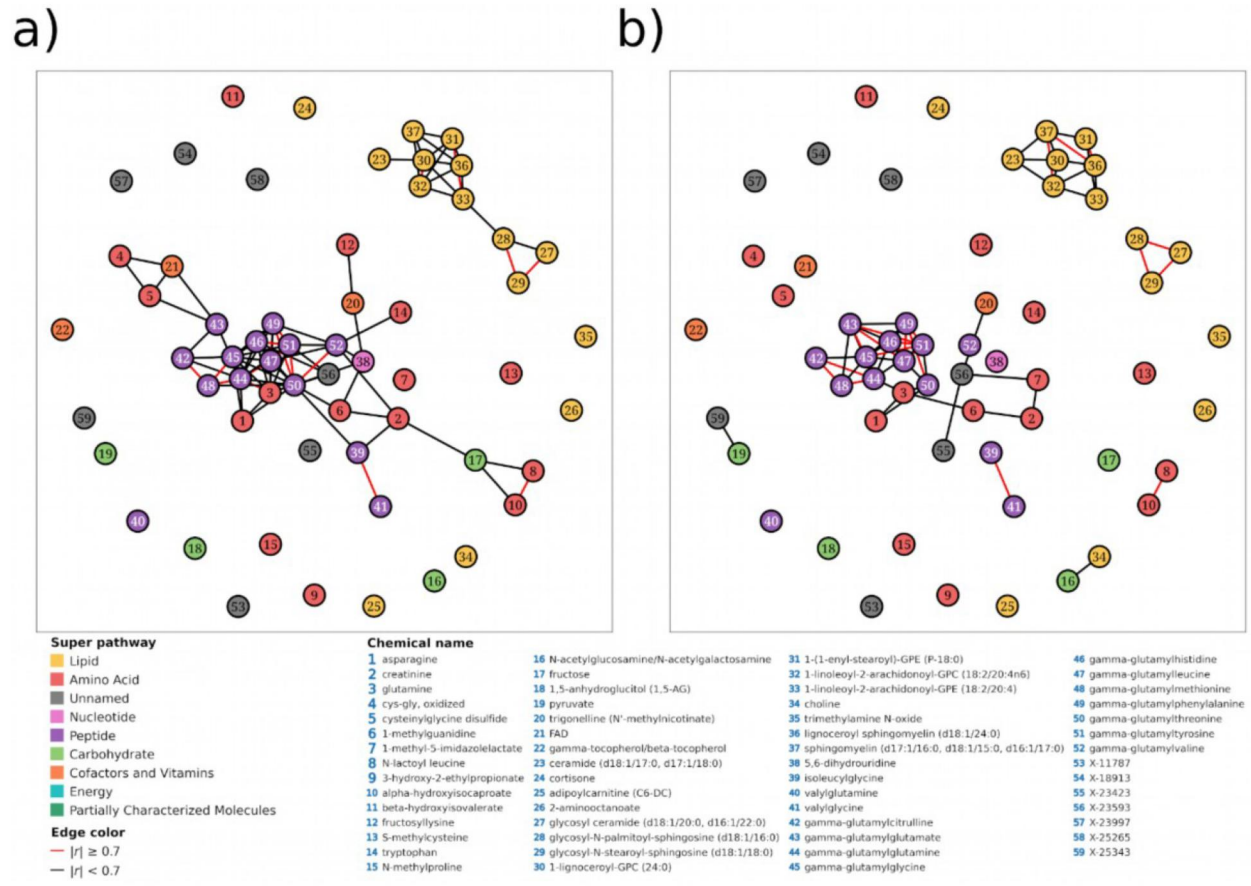

Figure 7. Metabolite networks ($r > 0.5$) based on the 59 selected metabolites for (a) Italian Landrace pigs ($n = 53$ animals) and (b) Italian Large White pigs ($n = 152$ pigs; subset ABC).

betweenness centrality, which was embedded in a highly connected cluster of gamma-glutamyl peptides.

For ILW, five separate networks were evaluated, considering that unlike ILA, ILW subsets varied in size and composition. Network similarity, assessed using the Jaccard index (J), ranged from $J = 0.48$ (between subsets B and C) to $J = 0.91$ (between AB and ABC) (Supplementary Table S8), indicating that network structure was sensitive to both sample size and population composition. Networks A, B and C displayed a comparable number of connected nodes (42, 41 and 41, respectively), with 28 metabolites consistently

shared across all three networks, including those overlapping in subsets AB and ABC. The number of edges was also similar (78–111); however, only 39 edges were consistently retained, again including those found in subsets AB and ABC. These 39 edges connected 28 nodes, 26 of which belonged to the previously identified set of 33 metabolites, thereby defining a stable core (Supplementary Table S7).

In the subset ABC (Figure 7(b)), the ILW network structure diverged from the ILA network, as reflected by a $J = 0.66$. The ILW network was more fragmented, comprising seven modules: two large (18 and 7 nodes) and five small (2 or 3 nodes each) (Figure 7(b)). The largest ILW module (18 nodes) was a subgraph of the ILA main module and included 10 gamma-glutamyl peptides, two unnamed metabolites, five amino acids and one compound from the cofactors and vitamins super-pathway. Notably, gamma-glutamylvaline (node no. 52), the metabolite with the highest concentration difference between breeds, was disconnected in the ILW network but had seven strong correlations in ILA. Furthermore, the lipid-related module in ILW was split into two: a small cluster of three HCER metabolites and a larger cluster of the remaining seven lipids, contrasting with the single lipid module observed in ILA.

Discussion

Dissecting the animal phenome at the molecular level (which describes molecular phenotypes, also known as internal phenotypes) is essential to understand the biological mechanisms underlying external and complex traits, whether related to production, behaviour, reproduction or other features that define a breed (Pérez-Enciso and Steibel 2021). In this context, the identification of metabolic signatures through high-throughput metabolomics provides a biochemical perspective on breed-specific characteristics and descriptors and supports the implementation of precision livestock farming and breeding strategies (Fontanesi 2016; Goldansaz et al. 2017).

In this study, we extended the evaluation of Italian pig breeds used in a three-way crossbreeding system for the production of green legs destined for PDO Parma and San Daniele hams (Bosi and Russo 2004; Bertolini et al. 2024). While our previous works compared IDU and ILW pigs (Bovo et al. 2016a, 2025a), representing a paternal line and a maternal line, respectively, here we focused on comparing ILA and ILW pigs, representing both maternal lines used in this typical crossbreeding scheme. This comparison allowed us to focus on metabolic differences associated with distinct genetic backgrounds (represented by the breed genetic pools), specialised in maximising reproduction performances in this context.

In this metabolomic study, we included a total of 205 pigs from two breeds. Even when considering the breed with the lowest number of pigs (ILA, with 53 pigs), this number is significantly larger than in many other studies comparing metabolomic profiles in farm animals (e.g. Bovo et al. 2016a; Qu and Ajuwon 2018; Luise et al. 2020; Hou et al. 2023; Metzler-Zebeli et al. 2023). This provided sufficient statistical power to detect differences between breeds (Rakusanova and Cajka 2024). For practical reasons, such as the smaller number of ILA pigs compared to ILW pigs in the ANAS breeding programs and the need to maintain a balanced design by pairing animals slaughtered on the same days to reduce environmental confounding effects, the experimental design focused on comparing the two breeds, resulting in an unbalance in the proportion of animals. We could have reduced the number of ILW pigs to match the number of ILA pigs, but we preferred to gather as much informative data as possible in ILW under these experimental conditions and implement a statistical methodology to address the different numbers in the two breed populations included in the study. Therefore, an initial methodological step was taken to address how sampling and population imbalances could affect the identification of metabolic signatures. Because the number of ILW pigs was approximately three times higher than that of ILA pigs, we conducted multiple comparative analyses using both balanced and unbalanced datasets. Three independent subsets of ILW pigs were matched to the ILA group, allowing for balanced replicate comparisons (A, B and C). These subsets were then progressively merged to create imbalanced datasets (AB and ABC), enabling us to assess the stability of metabolite selection under varying sample size conditions. This design is relatively uncommon but innovative in metabolomics, as it allows for a direct evaluation of the effect of population heterogeneity and sampling bias on feature selection. This factor can strongly influence biomarker discovery in omics studies. Unbalanced sample size between groups can represent a major challenge for multivariate modelling and classification performance (Zheng and Jin 2020). Common approaches to mitigate this issue include downsampling, where the larger dataset is randomly reduced to create groups of comparable size, or data augmentation, where synthetic observations are generated to enrich the smaller group (Hasanin et al. 2019). However, both strategies

can introduce artefacts or reduce the representativeness of the original data, potentially biasing model interpretation. To avoid these limitations, we adopted a data reduction approach, focusing on maintaining biological validity and minimising artificial manipulation of the dataset. Additionally, the use of random forest and its feature selection extension, Boruta, was particularly suitable for this dataset, since random forest performs robustly even in the presence of class imbalance and limited training samples, often achieving favourable accuracy despite data skewness (Zheng and Jin 2020).

Across the balanced datasets, a similar number of discriminant metabolites were identified. However, only about 30% of these were consistently selected across all three subsets, regardless of the selection stringency (permissive-CV or robust-CV), indicating a clear influence of the sampled population. When considering metabolites identified in at least two of the three subsets, this overlap increased to approximately 60–70%, highlighting good reproducibility of the selection process. Combining the balanced subsets into progressively larger, unbalanced datasets led to a moderate ($\sim$25%) increase in the number of selected metabolites when the dataset size was doubled (AB), but only a slight additional increase when it was tripled (2–3% from AB to ABC), suggesting that sample imbalance influences metabolite selection up to a certain threshold. Nonetheless, larger sample sizes allowed for the detection of a few additional metabolites not identified in smaller subsets. Overall, the combination of analyses suggests that (i) analysing the entire dataset improves coverage of metabolomic diversity, whereas (ii) down-sampling for subsampling enhances the identification of core, stable metabolites that are potentially less affected by population variability.

The selected metabolites demonstrated strong discriminative power, as evidenced by clear separation between breeds in PCA plots and low classification errors in random forest models (OOB error $\leq$10%). Most metabolites presented AUC values above 0.75, with those identified in all datasets (intersection set) generally showing higher AUC and larger average differences in abundance (Δ%) compared to subset-specific metabolites. A few metabolites, however, showed inconsistent behaviour across subsets, sometimes displaying opposite directionality in abundance between breeds, again reflecting the influence of population composition on metabolite selection. Overall, the robustness of classification performance across subsets supports the reliability of the identified metabolites, though it also underscores the need to interpret metabolomic biomarkers in light of population sampling and model instability due to stochastic components.

From a biological perspective, two metabolite groups that differed between breeds were particularly relevant: the gamma-glutamyl amino acids and the HCERs. HCER metabolites were consistently less abundant in ILW pigs. Similar trends have been reported in cattle, where fattening bulls showed elevated HCER levels due to stimulated ceramide metabolism (Kenéz et al. 2022). The higher abundance of circulating HCER in ILA pigs may therefore reflect their higher fat deposition compared to the leaner ILW breed. In contrast, gamma-glutamyl amino acids were generally more abundant in ILW pigs. These metabolites are intermediates or by-products of the GSH cycle, a key cellular antioxidant and redox regulator (Boysen 2017). Changes in this cycle impair redox homeostasis and can lead to oxidative stress. GSH also plays a pivotal role in several other biological functions, including immune function, detoxification, xenobiotic metabolism and reproduction, potentially influencing resilience and fertility (Wu et al. 2004). Consistent with this, enrichment analyses confirmed the overrepresentation of gamma-glutamyl sub-pathway in the Metabolon dataset. Notably, higher concentrations of gamma-glutamyl amino acids were also reported in ILW compared with IDU pigs in our previous study (Bovo et al. 2025a), suggesting that these metabolites are recurrent features of the ILW pigs and may represent breed-specific molecular markers. Thus, a better understanding of the metabolism of these molecules has implications for reproductive robustness, an important selection goal for maternal lines investigated here. At the network level, gamma-glutamyl amino acids formed a highly connected module. Among the correlated metabolites, one unannotated compound (X-23593) showed strong correlations with three gamma-glutamyl amino acids. The human GWAS Catalog (Cerezo et al. 2025) revealed significant associations of this compound with the gamma-glutamylamine cyclotransferase (*GGACT*) gene, thereby supporting its involvement in the same biochemical context and aiding its putative functional annotation. The identification of putative metabolite-gene associations through network analysis exemplifies again how metabolomics can support functional annotation efforts, which remain challenging for many unknown compounds in untargeted datasets (Krumsiek et al. 2012). Beyond these core pathways, pathway enrichment analysis by the MetaboAnalyst platform

highlighted the perturbation of nitrogen metabolism, with glutamine emerging as a central metabolite. Glutamine plays a key role in ammonia detoxification, acting as a nitrogen carrier and as a precursor for GSH synthesis (Yoo et al. 2020). It also contributes to replenishing the tricarboxylic acid (TCA) cycle through its conversion to pyruvate, linking amino acid and energy metabolism. Consistent with this, pyruvic acid and other metabolites identified in this study (e.g. N-acetylgalactosamine and asparagine) connect glutamine metabolism to carbohydrate and hexosamine biosynthesis pathways (Yoo et al. 2020). Furthermore, the enrichment of solute carrier (SLC) transporters supports the involvement of glutamine transport and utilisation in defining breed-related metabolic profiles. Altogether, these interconnected pathways highlight how amino acid metabolism integrates with energy balance and cellular homeostasis, potentially influencing performance traits that characterise these breeds.

While the methodological and biological findings offer consistent and meaningful insights, it is important to acknowledge several limitations of this study. The moderate sample size, especially for the ILA population, and the reliance solely on plasma metabolomics, restrict the ability to infer tissue-specific metabolic processes. Additionally, while Metabolon's untargeted approach provides broad coverage, there is the possibility that some relevant compounds may go unidentified. Future research should involve targeted quantification of key metabolites, validation in independent populations, and integration with transcriptomic or genomic data to establish causality between metabolic shifts and genetic background.

Overall, these findings reinforce the concept that metabolic variation between ILA and ILW pigs reflects coordinated changes in amino acid and lipid metabolism, particularly involving GSH-related redox balance and ceramide pathways. These molecular differences likely mirror underlying genetic diversity between maternal lines and could serve as functional biomarkers for breed characterisation or selection. Future integration of metabolomics with genomics and digital phenotyping approaches may further advance the development of data-driven, phenomics-informed breeding strategies.

Conclusions

This study provides the first untargeted metabolomic profile of ILA pigs and compares it with ILW animals across multiple balanced and unbalanced datasets. The results reveal consistent breed-associated metabolic differences, particularly within the gamma-glutamyl amino acid and HCER pathways. These metabolites reflect distinct patterns of amino acid and lipid metabolism that may be driven by breed-specific genetic backgrounds. Together, our findings lay the groundwork for future integrative studies combining metabolomic, genomic and phenotypic data to better explain metabolic diversity and productive performance in Italian pig breeds.

Acknowledgements

The authors thank the personnel of the National Pig Breeders Association (ANAS) for their collaboration in this study.

Author contributions

Matteo Bolner: formal analysis, investigation, data curation, writing – original draft, writing – review and editing. Samuele Bovo: methodology, formal analysis, investigation, data curation, writing – original draft, writing – review and editing. Giuseppina Schiavo: investigation, data curation, methodology. Giuliano Galimberti: methodology. Francesca Bertolini: investigation, data curation. Stefania Dall'Olio: supervision. Anisa Ribani: investigation. Paolo Zambonelli: supervision. Maurizio Gallo: writing – review and editing, methodology. Luca Fontanesi: conceptualisation, resources, writing – original draft, writing – review and editing, supervision, funding acquisition.

Ethical approval

Not applicable.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by PRIN2017 PigPhenomics that has received funding from the Italian MUR (for the production metabolomic data); the Horizon Europe Re-Livestock Project that has received funding from the European Union's Horizon Europe Research and Innovation Programme under Grant Agreement No. 101059609 (for the elaboration and analysis of the data); the Italian PRIN2022 Project FEEDTHEPIG, 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 PNRR (DD N. 1409 del 14/09/2022), proposal code P2022FZMJ9 – CUP J53D23018310001 for the analysis of part of the data; the Italian PRIN2022 HamCapture, 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 (for the analysis of part of the data); and was carried out within the Agritech National Research Center (Agritech Spoke 1) and received funding from the European Union – NextGenerationEU (PIANO NAZIONALE DI RI-PRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4 – D.D. 1032 17/06/2022, CN00000022).

ORCID

Matteo Bolner http://orcid.org/0000-0002-4985-0191
Samuele Bovo http://orcid.org/0000-0002-5712-8211
Giuseppina Schiavo http://orcid.org/0000-0002-3497-1337
Giuliano Galimberti http://orcid.org/0000-0002-9161-9671
Francesca Bertolini http://orcid.org/0000-0003-4181-3895
Stefania Dall'Olio http://orcid.org/0000-0003-1384-3771
Anisa Ribani http://orcid.org/0000-0001-6778-1938
Paolo Zambonelli http://orcid.org/0000-0002-2532-5528
Luca Fontanesi http://orcid.org/0000-0001-7050-3760

Data availability statement

The data that support the study findings are publicly available at: https://doi.org/10.5281/zenodo.17536029

References

Azur MJ, Stuart EA, Frangakis C, Leaf PJ. 2011. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res. 20(1):40–49. https://doi.org/10.1002/mpr.329

Bertolini F et al. 2024. Signatures of selection analyses reveal genomic differences among three heavy pig breeds that constitute the genetic backbone of a dry-cured ham production system. Animal. 18(11):101335. https://doi.org/10.1016/j.animal.2024.101335

Bosi P, Russo V. 2004. The production of the heavy pig for high quality processed products. Ital J Anim Sci. 3(4):309–321. https://doi.org/10.4081/ijas.2004.309

Bovo S et al. 2015. Deconstructing the pig sex metabolome: targeted metabolomics in heavy pigs revealed sexual dimorphisms in plasma biomarkers and metabolic pathways. J Anim Sci. 93(12):5681–5693. https://doi.org/10.2527/jas.2015-9528

Bovo S et al. 2016a. Metabolomics evidences plasma and serum biomarkers differentiating two heavy pig breeds. Animal. 10(10):1741–1748. https://doi.org/10.1017/S1751731116000483

Bovo S et al. 2016b. Genome-wide association study for the level of serum electrolytes in Italian Large White pigs. Anim Genet. 47(5):597–602. https://doi.org/10.1111/age.12459

Bovo S 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. Genet Sel Evol. 52(1):33. https://doi.org/10.1186/s12711-020-00553-7

Bovo S et al. 2023. Comparative targeted metabolomic profiles of porcine plasma and serum. Animal. 17(12):101029. https://doi.org/10.1016/j.animal.2023.101029

Bovo S et al. 2025a. High-throughput untargeted metabolomics reveals metabolites and metabolic pathways that differentiate two divergent pig breeds. Animal. 19(1):101393. https://doi.org/10.1016/j.animal.2024.101393

Bovo S et al. 2025b. Description of metabolic differences between castrated males and intact gilts obtained from high-throughput metabolomics of porcine plasma. J Anim Sci. 103:skaf178. https://doi.org/10.1093/jas/skaf178

Bovo S et al. 2025c. Merging metabolomics and genomics provides a catalog of genetic factors that influence molecular phenotypes in pigs linking relevant metabolic pathways. Genet Sel Evol. 57(1):11. https://doi.org/10.1186/s12711-025-00960-8

Bovo S et al. 2025d. Comparative analysis of targeted metabolomic profiles reveals plasma metabolite differences across three Italian heavy pig breeds. Sci Rep. 15(1):39296. https://doi.org/10.1038/s41598-025-23058-z

Boysen G. 2017. The glutathione conundrum: stoichiometric disconnect between its formation and oxidative stress. Chem Res Toxicol. 30(5):1113–1116. https://doi.org/10.1021/acs.chemrestox.7b00018

Braisted J et al. 2023. RaMP-DB 2.0: a renovated knowledgebase for deriving biological and chemical insight from metabolites, proteins, and genes. Bioinformatics. 39(1):btac726. https://doi.org/10.1093/bioinformatics/btac726

Cassady JP, Young LD, Leymaster KA. 2002. Heterosis and recombination effects on pig reproductive traits. J Anim Sci. 80(9):2303–2315. https://doi.org/10.1093/ansci/80.9.230

Cerezo M et al. 2025. The NHGRI-EBI GWAS Catalog: standards for reusability, sustainability and diversity. Nucleic Acids Res. 53(D1):D998–D1005. https://doi.org/10.1093/nar/gkae1070

Comings DE, MacMurray JP. 2000. Molecular heterosis: a review. Mol Genet Metab. 71(1–2):19–31. https://doi.org/10.1006/mgme.2000.3015

Dervishi E et al. 2023. GWAS and genetic and phenotypic correlations of plasma metabolites with complete blood count traits in healthy young pigs reveal implications for pig immune response. Front Mol Biosci. 10:1140375. https://doi.org/10.3389/fmolb.2023.1140375

Faquih T et al. 2020. A workflow for missing values imputation of untargeted metabolomics data. Metabolites. 10(12):486. https://doi.org/10.3390/metabo10120486

Fiehn O. 2002. Metabolomics – the link between genotypes and phenotypes. Plant Mol Biol. 48(1–2):155–171. https://doi.org/10.1007/978-94-010-0448-0_11

Fontanesi L. 2016. Metabolomics and livestock genomics: insights into a phenotyping frontier and its applications in animal breeding. Anim Front. 6(1):73–79. https://doi.org/10.2527/af.2016-0011

Fuxman Bass JI et al. 2013. Using networks to measure similarity between genes: association index selection. Nat Methods. 10(12):1169–1176. https://doi.org/10.1038/nmeth.2728

Goldansaz SA et al. 2017. Livestock metabolomics and the livestock metabolome: a systematic review. PLOS

One. 12(5):e0177675. https://doi.org/10.1371/journal.pone.0177675
Hasanin T et al. 2019. Severely imbalanced big data challenges: investigating data sampling approaches. J Big Data. 6(1):1–25. https://doi.org/10.1186/s40537-019-0274-4
Hou X et al. 2023. Metabolomics and lipidomics profiles related to intramuscular fat content and flavor precursors between Laiwu and Yorkshire pigs. Food Chem. 404(Pt A):134699. https://doi.org/10.1016/j.foodchem.2022.134699
Houle D, Govindaraju DR, Omholt S. 2010. Phenomics: the next challenge. Nat Rev Genet. 11(12):855–866. https://doi.org/10.1038/nrg2897
Kenéz Á, Bäßler SC, Jorge-Smeding E, Huber K. 2022. Ceramide metabolism associated with chronic dietary nutrient surplus and diminished insulin sensitivity in the liver, muscle, and adipose tissue of cattle. Front Physiol. 13:958837. https://doi.org/10.3389/fphys.2022.958837
Krumsiek J et al. 2012. Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information. PLoS Genet. 8(10):e1003005. https://doi.org/10.1371/journal.pgen.1003005
Kursa MB, Jankowski A, Rudnicki WR. 2010. Boruta – a system for feature selection. Fundam Inform. 101(4):271–285. https://doi.org/10.3233/FI-2010-288
Luise D et al. 2020. Targeted metabolomic profiles of piglet plasma reveal physiological changes over the suckling period. Livest Sci. 231:103890. https://doi.org/10.1016/j.livsci.2019.103890
Metzler-Zebeli BU et al. 2023. Creep feeding and weaning influence the postnatal evolution of the plasma metabolome in neonatal piglets. Metabolites. 13(2):214. https://doi.org/10.3390/metabo13020214
Pang Z et al. 2024. MetaboAnalyst 6.0: towards a unified platform for metabolomics data processing, analysis and interpretation. Nucleic Acids Res. 52(W1):W398–W406. https://doi.org/10.1093/nar/gkae253
Pedregosa F et al. 2011. Scikit-learn: machine learning in Python. J Mach Learn Res. 12:2825–2830.
Peixoto TP. 2014. The graph-tool python library. Figshare.
Pérez-Enciso M, Steibel JP. 2021. Phenomes: the current frontier in animal breeding. Genet Sel Evol. 53(1):22. https://doi.org/10.1186/s12711-021-00618-1
Peukert M et al. 2021. Sexual dimorphism of metabolite profiles in pigs depends on the genetic background. Metabolites. 11(5):261. https://doi.org/10.3390/metabo11050261
Ponsuksili S et al. 2025. Genetic regulation and variation of fetal plasma metabolome in the context of sex, paternal breeds and variable fetal weight. Open Biol. 15(3):240285. https://doi.org/10.1098/rsob.240285
Qu H, Ajuwon KM. 2018. Metabolomics of heat stress response in pig adipose tissue reveals alteration of phospholipid and fatty acid composition during heat stress. J Anim Sci. 96(8):3184–3195. https://doi.org/10.1093/jas/sky127
R Core Team. 2024. R: a language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/
Rakusanova S, Cajka T. 2024. Tips and tricks for LC–MS-based metabolomics and lipidomics analysis. TrAC Trends Anal Chem. 180:117940. https://doi.org/10.1016/j.trac.2024.117940
Schiavo G 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. Anim Genet. 55(2):193–205. https://doi.org/10.1111/age.13396
Sellier P. 1976. The basis of crossbreeding in pigs: a review. Livest Prod Sci. 3(3):203–226. https://doi.org/10.1016/0301-6226(76)90016-6
Wang H, Xia P, Lu Z, Su Y, Zhu W. 2021. Metabolome-microbiome responses of growing pigs induced by time-restricted feeding. Front Vet Sci. 8:681202. https://doi.org/10.3389/fvets.2021.681202
Wang X, Kadarmideen HN. 2020. Metabolite genome-wide association study (mGWAS) and gene-metabolite interaction network analysis reveal potential biomarkers for feed efficiency in pigs. Metabolites. 10(5):201. https://doi.org/10.3390/metabo10050201
Waskom ML. 2021. seaborn: statistical data visualization. J Open Source Softw. 6(60):3021. https://doi.org/10.21105/joss.03021
Wishart DS et al. 2022. HMDB 5.0: the Human Metabolome Database for 2022. Nucleic Acids Res. 50(D1):D622–D631. https://doi.org/10.1093/nar/gkab1062
Wolf J, Peškovičová D, Žáková E, Groeneveld E. 2006. Additive and heterotic breed effects in the genetic evaluation of pig sire breeds. Anim Sci. 82(4):455–462. https://doi.org/10.1079/ASC200655
Wu G, Fang Y-Z, Yang S, Lupton JR, Turner ND. 2004. Glutathione metabolism and its implications for health. J Nutr. 134(3):489–492. https://doi.org/10.1093/jn/134.3.489
Yoo HC et al. 2020. Glutamine reliance in cell metabolism. Exp Mol Med. 52(9):1496–1516. https://doi.org/10.1038/s12276-020-00504-8
Zhang Y et al. 2021. Metabolomic analysis and identification of sperm freezability-related metabolites in boar seminal plasma. Animals. 11(7):1939. https://doi.org/10.3390/ani11071939
Zheng W, Jin M. 2020. The effects of class imbalance and training data size on classifier learning: an empirical study. SN Comput Sci. 1(2):71. https://doi.org/10.1007/s42979-020-0074-0
Zheng Y, Zheng Z, Rendeiro AF, Cheung E. 2025. Marsilea: an intuitive generalized paradigm for composable visualizations. Genome Biol. 26(1):5. https://doi.org/10.1186/s13059-024-03469-3