Contents lists available at ScienceDirect

Animal

The international journal of animal biosciences

High-throughput untargeted metabolomics reveals metabolites and metabolic pathways that differentiate two divergent pig breeds

S. Bovo[a,1], M. Bolner[a,1], G. Schiavo[a], G. Galimberti[b], F. Bertolini[a], S. Dall'Olio[a], A. Ribani[a], P. Zambonelli[a], M. Gallo[c], L. Fontanesi[a,*]

[a] *Animal and Food Genomics Group, Division of Animal Sciences, Department of Agricultural and Food Sciences, University of Bologna, 40127 Bologna, Italy*
[b] *Department of Statistical Sciences "Paolo Fortunati", University of Bologna, 40126 Bologna, Italy*
[c] *Associazione Nazionale Allevatori Suini, 00198 Roma, Italy*

ARTICLE INFO

Article history:
Received 14 August 2024
Revised 28 November 2024
Accepted 29 November 2024
Available online 04 December 2024

Keywords:
Machine learning
Metabolite
Molecular phenome
Plasma
Sus scrofa

ABSTRACT

Metabolomics can describe the molecular phenome and may contribute to dissecting the biological processes linked to economically relevant traits in livestock species. Comparative analyses of metabolomic profiles in purebred pigs can provide insights into the basic biological mechanisms that may explain differences in production performances. Following this concept, this study was designed to compare, on a large scale, the plasma metabolomic profiles of two Italian heavy pig breeds (Italian Duroc and Italian Large White) to indirectly evaluate the impact of their different genetic backgrounds on the breed metabolomes. We utilised a high-throughput untargeted metabolomics approach in a total of 962 pigs that allowed us to detect and relatively quantify 722 metabolites from various biological classes. The molecular data were analysed using a bioinformatics pipeline specifically designed for identifying differentially abundant metabolites between the two breeds in a robust and statistically significant manner, including the Boruta algorithm, which is a Random Forest wrapper, and sparse Partial Least Squares Discriminant Analysis (**sPLS-DA**) for feature selection. After thoroughly evaluating the impact of random components on missing value imputation, 100 discriminant metabolites were selected by Boruta and 17 discriminant metabolites (all included within the previous list) were identified with sPLS-DA. About half of the 100 discriminant metabolites had a higher concentration in one or the other breed (48 in Italian Large White pigs, with a prevalence of amino acids and peptides; 52 in Italian Duroc pigs, with a prevalence of lipids). These metabolites were from seven distinct super pathways and had an absolute mean value of percentage difference between the two breeds (|Δ|%) of 39.2 ± 32.4. Six of these metabolites had |Δ|%> 100. A general correlation network analysis based on Boruta−identified metabolites consisted of 31 singletons and 69 metabolites connected by 141 edges, with two large clusters (> 15 nodes), three medium clusters (3–6 nodes) and eight additional pairs, with most metabolites belonging to the same super pathway. The major cluster representing the lipids super-pathway included 24 metabolites, primarily sphingomyelins. Overall, this study identified metabolomic differences between Italian Duroc and Italian Large White pigs explained by the specific genetic background of the two breeds. These biomarkers can explain the biological differences between these two breeds and can have potential practical applications in pig breeding and husbandry.

Implications

This study aimed to compare the metabolomic profiles of two Italian heavy pig breeds (Italian Duroc and Italian Large White) used in crossbreeding programmes. The differences observed may shed light on the complementarities needed to turn on the heterotic effect in the resulting crossbred pigs. The developed metabolomic data analysis approaches provided comprehensive overviews of the metabolome features of the two breeds that aligned with some of their known characteristics. This suggests that overall metabolomic profiles could serve as important descriptors of breed−specific molecular traits.

Introduction

The pig breeding industry has been driving the sustainability of the global pork production system by focusing on the efficiency

[*] Corresponding author.
E-mail address: luca.fontanesi@unibo.it (L. Fontanesi).
[1] Equal contribution.

https://doi.org/10.1016/j.animal.2024.101393
1751-7311/

and performance of boars, sows, and market pigs (Rauw et al., 2020; Kumar et al., 2023). Selective breeding programmes aim to enhance various traits in a variety of specialised breeds. These breeds are then typically used in crossbreeding schemes to maximise the heterotic effect on economically relevant traits (Sellier, 1976; Buchanan and Stalder, 2011). Large White and Duroc are cosmopolitan pig breeds, that, along with a few other breeds, make up the genetic background of many different specialised nuclei (developed at national or breeding industry levels) and synthetic sire and dam lines that serve a variety of crossbreeding programmes (Kim et al., 2020; Mote and Rothschild, 2020; Fabbri et al., 2024). Large White pigs have been traditionally selected to enhance sow productivity, in addition to other complementary traits, while Duroc pigs have been mainly selected for efficiency, growth, meat quality and carcass traits and adaptation (Mote and Rothschild, 2020). Peculiar characteristics of the Duroc pigs are the high levels of inter and intramuscular fat that may lead to excessive fatness covering the muscle fibres of the legs (Armero et al., 1999; Suzuki et al., 2003; Alonso et al., 2015). These two breeds constitute important cosmopolitan breeds, genetically very different, as described by several studies that have analysed their genomes (Wilkinson et al., 2013; Muñoz et al., 2018; Bovo et al., 2020; Zhang et al., 2020). However, what is still missing is an extended phenotyping analysis of their basic physiological and metabolic characteristics that could be captured by more detailed molecular or internal phenotypes that constitute their respective molecular phenomes, which may be at the origin of their differences in terms of performance and production traits.

Recent technological developments have prompted the characterisation of the animal metabolome as part of the molecular phenome aimed to obtain a detailed description of many metabolites (small chemical compounds) present in a biological matrix at a given time (Houle et al., 2010; Fontanesi, 2016; Pérez-Enciso and Steibel, 2021; Hollywood et al., 2006). Metabolites that are considered phenotypes are referred to as metabotypes. These metabotypes represent intermediate internal molecular phenotypes, as opposed to external or end-phenotypes (Fiehn, 2002; Houle et al., 2010; Fontanesi, 2016). The analysis of hundreds of metabolites can produce metabolomic fingerprints that are useful for exploring differences between animals and breeds (Fontanesi, 2016; Goldansaz et al., 2017).

Multivariate statistical methods, particularly machine learning approaches, have been widely used and compared for analysing untargeted metabolomic data (Gromski et al., 2015, Trainor et al., 2017; Mendez et al., 2019; Vu et al., 2019, Galal et al., 2022). Among them, methods of the Partial Least Squares (**PLS**) family (e.g. PLS Regression and PLS Discriminant Analysis) are currently the most widely employed state-of-the-art tools for reducing the dimensionality of noisy and multicollinear data (such as metabolomic datasets). To directly identify a subset of key variables, these approaches were modified and evolved into their sparse versions, which involve feature selection through regularisation (Chung and Keles, 2010). Recently, the Boruta algorithm, a Random Forest wrapper, has emerged as a solution for all-relevant problems of feature selection, allowing the identification of the full set of features carrying information usable for prediction (Kursa et al., 2010). This is of great relevance in metabolomics, where the primary focus is on identifying the complete set of metabolites that are associated with and can explain the observed phenomena rather than just identifying the minimal set of metabolites needed for creating a predictive model (minimal-optimal model). For these features, the Boruta algorithm has been utilised in the analysis of various omics data, including metabolomics (Gromski et al., 2015; Degenhardt et al., 2019; Wenck et al., 2021; Schiavo et al., 2024), resulting in a complementary approach to adopt and further exploit in metabolomic data analysis.

In pigs, metabolomics has been successfully used for various purposes that range from the general characterisation of biofluids to more specific applications, such as intra-breed exploration of meat quality traits, stress and resilience, feed nutrition, sexual dimorphisms and reproductive performance, among others (Bovo et al., 2015, 2023; Carmelo et al., 2020; Goldansaz et al., 2017; Luise et al., 2020; Metzler-Zebeli et al., 2023; Peukert et al., 2021; Wang et al., 2021; Zhang et al., 2021). Moreover, considering that pig breeds are genetically different, comparative metabolomics has been used to explore breed differences in order to evaluate the impact of the genetic background on the metabolome and overall metabolism (D'Alessandro et al., 2011; Straadt et al., 2014; Bovo et al., 2016, Carmelo et al., 2020; Lefort et al., 2020; Xie et al., 2023). As a result, new opportunities are emerging for the development of nutrigenetics and precision feeding strategies tailored to the breed differences. However, metabolomics is still in its early stages, particularly in pigs where only very recently have global untargeted metabolomics analyses begun to take place (e.g. Carmelo et al., 2020; Deng et al., 2023; Xie et al., 2023).

In our previous work (Bovo et al., 2016), we conducted a pilot study to explore metabolomic differences between two pig breeds: Italian Large White and Italian Duroc. We used targeted metabolomics to quantify approximately 180 metabolites. These Italian breeds, derived from the cosmopolitan Large White and Duroc breeds, have been specifically constituted (starting from the 1990 's) to serve as maternal and paternal lines to produce final heavy crossbred pigs (slaughtered at approximately 170 kg of live weight and 9 months of age) that are then used to obtain dry-cured hams.

In this study, we expanded the information on the molecular phenome of pigs by examining the plasma metabolome of Italian Large White and Italian Duroc breeds on a much larger scale than our previous work using an untargeted metabolomics approach. We analysed approximately 800 metabolites from plasma samples collected from over 950 animals. The data were processed using specifically designed bioinformatic pipelines to select metabolites that provide breed–specific metabolomic fingerprints. The pipeline we developed methodologically considered the impact of randomness, including random states in algorithms, subpopulations, and data imputation, on metabolite selection. Additionally, we assessed the metabolites within their biological networks and pathways to provide a comprehensive view of the metabolism of these two divergent pig breeds.

Material and methods

Animals and blood samples

All animals used in this study were kept in accordance with the Italian and European legislations for pig production. All procedures described here were compliant with Italian and European Union regulations for animal care and slaughter. Animals were slaughtered in a commercial abattoir following standard procedures. Fasting of the animals was not specifically conducted for this study, but rather as part of the standard preslaughtering procedure. The pigs were not raised or treated in any way for the purpose of this study; therefore, no additional ethical statement is required.

This study involved a total of 962 healthy pigs: 694 Italian Large White pigs (466 entire gilts and 228 castrated males) and 268 Italian Duroc pigs (191 entire gilts and 77 castrated males) slaughtered over 23 different days. The pigs were part of the sib-testing programmes based on triplets of pigs from the same litter, two females and one castrated male that were individually performance-tested at the Central Station of the National Pig Breeder Association. Performance evaluation started when the pigs were 30–45 days of age and ended when the animals reached about 155 ± 5 kg live weight. During the performance period, all pigs were handled in the same way and fed semi-*ad libitum* with a commercial fattening diet from about 100 kg live weight to the slaughtering weight. When animals were about 155 ± 5 kg live weight, they were subjected to a fasting period of ~12 h, transported to a commercial abattoir and slaughtered, after electrical stunning, in the morning at about 0800 h Animals entered the slaughtering plant within 5 min. Blood samples were collected into an EDTA–containing tube from the draining carotid artery immediately after jugulation. The tubes were inverted eight to ten times and centrifuged within 2 h, at 2 420 × g for 10 min at + 4 °C.

Metabolomic profiling of plasma samples

Untargeted metabolomics profiling of plasma samples was conducted at Metabolon, Inc (Durham, NC, USA). The HD4 metabolomics panel was utilised for this analysis.

Initially, a methanol extraction was performed to eliminate proteins and extract metabolites. The resulting extract underwent various ultra-performance liquid chromatography – tandem mass spectrometry (**UPLC-MS/MS**) steps, including reverse phase UPLC-MS/MS with positive and negative ion mode electrospray ionisation, and Hydrophilic Interaction Liquid Chromatography UPLC-MS/MS with negative ion mode ESI. The samples were analysed in four separate batches, each containing common quality control samples (constituted by a pool of plasma samples). Raw data were extracted, peaks were identified, and quality control was processed using Metabolon's hardware and software. For each metabolite, the raw peak area was divided by the median raw peak area of the quality control samples. Metabolite annotations, including chemical names, database identifiers and biological pathways, were provided by Metabolon.

A total of 722 metabolites were assessed in at least one breed, including 654 named and 68 unnamed metabolites from different metabolite classes belonging to seven metabolic super pathways (lipids, amino acids, nucleotides, peptides, carbohydrates, cofactors and vitamins, energy and partially characterised molecules), were identified in at least on breed. Unnamed metabolites are those that have been detected and measured, but their chemical identity has not yet been elucidated (Krumsiek et al., 2012). Metabolites from the xenobiotic class were not included in the subsequent analyses. This was done because xenobiotic metabolites are not produced directly by the organism but are instead introduced via diet (specific components), drugs, or pollutants. As a result, their presence and concentration in an organism are not solely determined by genetic background but rather by the environment in which the animal is raised.

Data quality control, imputation and cleaning

Quality control and data imputation were conducted on the Italian Large White and Italian Duroc datasets, separately. Following the methodology outlined in our previous works (Bovo et al., 2015), the data were filtered as follows: (i) outlier metabolites were identified as values that deviated by 5 times the interquartile range below or above the median for each metabolite; (ii) these outlier values were then removed from the dataset and marked as missing data; (iii) metabolites with more than 25% missing values were excluded from the dataset; (iv) samples with more than 30% missing values were also removed. Missing values were imputed using the method described by Faquih et al. (2020), which involved the Multivariate Imputation by Chained Equations approach (Azur et al., 2011) using the MICE R package (Van Buuren and Groothuis-Oudshoorn, 2011). The procedure is based on an iterative series of predictive models where each specified variable in the dataset is imputed using the other variables in the dataset. These iterations should be run until it appears that convergence has been met. MICE requires a predictor matrix that defines which columns (metabolites) of the dataset will be used to impute each column with missing values. For each metabolite with missing values, we selected the top ten metabolites with the highest Pearson's correlation coefficient (r) along with the animal sex and sampling date as predictors. Only metabolites of endogenous origin (as defined by Metabolon) were used as predictors. Following the suggestion of Faquih et al. (2020), we used Predictive Mean Matching as the imputation algorithm generating five different datasets, as recommended by MICE authors. Predictive Mean Matching works well for imputing quantitative variables that are not normally distributed and produces imputed values that are much more like real values. The approach works by imputing missing values iteratively using regression models, in which each variable with missing data is modelled conditionally on all other variables in the dataset. Moreover, as part of the MICE strategy, five (or more) imputed datasets are produced, subjected to the intended statistical analysis, and results are then combined (pooled) and reported. To account for the random component of the imputation process, the data were imputed five separate times, each with a different random state seed, resulting in a total of 25 imputed datasets for each breed. Based on the specific random state seed and cycle of imputation, the Italian Large White and Italian Duroc datasets were merged, and confounding factors were removed by regressing each metabolite abundance on covariates (such as sex, animal weight and sampling day) as previously described by Bovo et al. (2015). Residuals were obtained and used in the subsequent statistical analyses. The analyses were conducted in the R v.4.2.3 (R Core Team, 2022) environment and Python v3.11.7.

Selection of differentially abundant metabolites between pig breeds

Differentially abundant metabolites between breeds were identified using unsupervised multivariate statistics and Machine Learning approaches. In the first step of evaluation, we conducted a Principal Component Analysis (**PCA**). The dataset of residuals was analysed in R v.4.2.3 (R Core Team, 2022) through the function prcomp; before PCA, variables were scaled to have unit variance. The first two principal components were extracted and analysed. Subsequently, we utilised two Machine Learning methods for feature selection: Boruta (Kursa et al., 2010), which is a wrapper for a Random Forest classification algorithm, and sparse Partial Least Squares Discriminant Analysis (**sPLS-DA**) (Chung and Keles, 2010).

Boruta analysis consisted of several iterative applications of Random Forest where shadow features were artificially created and compared to the original one to estimate their importance. Features that emerged from the comparison are labelled as 'Confirmed', 'Tentative' and 'Rejected'. Analyses were run in Python v3.11.7, using the BORUTA_py and scikit_learn packages with default parameters except for (i) the max_iter parameter (maximum iterations to perform), which was increased to 1 000 (default: 100), and the alpha parameter (the level at which the corrected P-values will get rejected in the two–step correction adopted in feature selection), which was set to a more stringent value of 0.01 (default: 0.05) to obtain more robust statistics. Metabolites labelled as "confirmed" by Boruta were considered selected and relevant for classification.

sPLS-DA analyses were performed in the R v.4.2.3 environment using the function splsda of the R package spls (Chung and Keles, 2010). They included an internal 10-fold cross-validation (internal 10**CV**) procedure to identify, through a grid search (function cv. splsda), the optimal eta (thresholding parameter; ranging from 0.1 to 0.9) and K (number of hidden components; ranging from 2 to 10) parameters. As a result, metabolites with a non-zero regression coefficient were considered selected and relevant for classification.

Feature selection was performed on each of the 25 datasets with missing data imputed (Supplementary Figure S1). To account for the random component of the feature selection methods, we ran each analysis five times with five different random seeds, for a total of 125 feature selection runs on the original dataset. To evaluate the stability of selected metabolites, each run included an addition 10CV procedure (external 10CV), by randomly splitting each of the 25 imputed datasets into ten equal parts. Thus, each approach tested the discriminative power of metabolites a total of 1 250 times (Supplementary Figure S1), accounting for random components nested in both imputation and selection methods (Supplementary Figure S1). For the two algorithms, we considered informative sets as (i) metabolites classified as confirmed in all 1 250 runs (Boruta) and (ii) metabolites with the sign of the regression coefficient identical in all 1 250 runs (sPLS-DA).

Following the study by Schiavo et al. (2024), the Boruta approach was then coupled with a standard Random Forest analysis of the reduced informative set, meaning that residuals of metabolites selected by Boruta/sPLS-DA were used as starting set for the analysis. The function randomForest included in the R package randomForest (Breiman, 2001) was used, with default parameters. This approach allowed for the computation of the Mean Decrease Gini (**MDG**) parameter, which was used to score the importance of the metabolites and rank them accordingly. The Out-Of-Bag (**OOB**) score and error, two indexes measuring the prediction error of Random Forest (and consequently Boruta), was adopted to assess the ability of the metabolite sets to accurately assign each animal to its breed. This metric provides a computationally convenient approach to evaluate Random Forest without using a testing dataset or cross-validation procedures (Huang and Deng, 2021). In addition to the MDG parameter, discriminant metabolites were assessed for their predictive performance using a Receiver Operating Characteristic curve analysis and the area under the curve (**AUC**) value, a summary metric of the Receiver Operating Characteristic curve, as implemented in the R v.4.2.3 environment with the function auc (default parameters) of the R package pROC (Robin et al., 2011).

We also evaluated the relative difference in concentration between breeds (Δ%; Bovo et al., 2016) for each metabolite (i), expressed as $\Delta\%_i = \frac{\bar{x}_i^S - \bar{x}_i^P}{\bar{x}_i^S} \times 100$, where $\bar{x}_i^S$ and $\bar{x}_i^P$ denote the average metabolite abundance of the ith metabolite in Italian Large White and Italian Duroc pigs, respectively.

Metabolite networks and pathway analysis

After selection, metabolite residuals were subjected to network and pathway analyses. Different networks based on Pearson's correlation coefficients (r) were computed: (i) a network including all pigs of the two breeds (Italian Large White and Italian Duroc), (ii) another network including all animals but with the inclusion of the breed as a fixed effect (in addition to sex, animal weight and sampling day) in the linear model used to clean the data and (iii) two separate networks, one for Italian Large White pigs and one for Italian Duroc pigs. Pearson's correlation coefficients were computed in R v.4.2.3. Edges of the network were considered informative when presenting $r \geq 0.5$ (medium correlation). Networks were drawn using Cytoscape 3.0.1 (Shannon et al., 2003) and basic statistics (e.g. node degree and betweenness centrality) were obtained.

Over-representation analysis was conducted to identify biological features characterising the selected metabolites. The analyses were performed using MetaboAnalyst v.6.0 (Pang et al., 2022). The Human Metabolome Database identifiers were utilised as the input set. The interrogated metabolite sets belonged to the RaMP-DB resource, a comprehensive collection of 3 694 metabolite and lipid pathways from various pathway databases (Zhang et al., 2018). Terms that included at least 2 metabolites from the input set and had a False Discovery Rate corrected $P < 0.05$ were considered statistically over-represented.

Results

Overview of metabolomic profiles

Out of 722 metabolites profiled in both breeds, 594 (82%) were metabolites of endogenous origin, 60 (8%) were metabolites of xenobiotic origin, and 68 (9%) were unnamed metabolites of unknown origin. The distributions of metabolites across super-pathways and sub-pathways are illustrated in Fig. 1A. Endogenous metabolites were categorised into eight super pathways: lipids (49% of the endogenous set), amino acids (30%), nucleotides (7%), peptides (4%), carbohydrates (4%), cofactors and vitamins (4%), energy (1%) and partially characterised molecules (1%). A total of 102 sub-pathways were covered (Supplementary Table S1). Metabolites of xenobiotic origin were excluded from further analyses.

On average, endogenous and unnamed metabolites had very few outlier values across the analysed pigs. For each metabolite the outlier pigs were on average 2.87 ± 6.89 and 2.82 ± 5.85, with median 1 for both endogenous and unnamed metabolites in the Italian Large White breed; 0.92 ± 2.37 and 0.43 ± 0.98 with median 0 for both classes of metabolites in Italian Duroc breed. Outlier values were removed and labelled as missing values. A total of 82 metabolites were discarded because they had a percentage of missing values greater than 25% in at least one of the two breeds. The proportional composition of the different super pathways did not change (Fig. 2A) and 101 sub-pathways (Supplementary Table S1) were still covered. The analysis of the entire metabolomic profile did not identify any outlier samples. After the filtering steps, a high-quality dataset was obtained, including 962 animals and 580 metabolites from the eight super pathways mentioned above. This dataset was used for all subsequent analyses.

Detection of metabolites differentiating the two pig breeds

Differences between the metabolomic profiles of the two breeds were initially examined by conducting a PCA on the entire metabolomic dataset that passed the filtering steps. The first two principal components represented 14.2 and 8.7% of the total variance, respectively. From this first analysis, these initial and global metabolomic profiles, including all metabolites that passed the applied quality filtering steps, did not separate the two breeds into distinct clusters, and the clouds constituted by the two pig groups completely overlapped (Fig. 1C).

Following this initial assessment, we utilised the Boruta Machine Learning approach and the sPLS-DA multivariate approach to identify the most discriminating metabolites within the initial metabolite dataset, which could differentiate the two breeds. Boruta identified a total of 100 metabolites (17.2% of the analysed metabolite dataset; Table 1, Supplementary Table S2) that were stably selected and confirmed across the analysis of the datasets that involved five randoms seeds of analysis (125 runs; 25 datasets × 5 Boruta seeds each) and a 10-fold cross−validation to each Boruta analysis (1 250 runs in total). These 100 stable metabolites derived from the combination (i) of the analyses across the 125 runs, where a total of 163 metabolites were globally selected as having discriminative features (ranging from 134 to

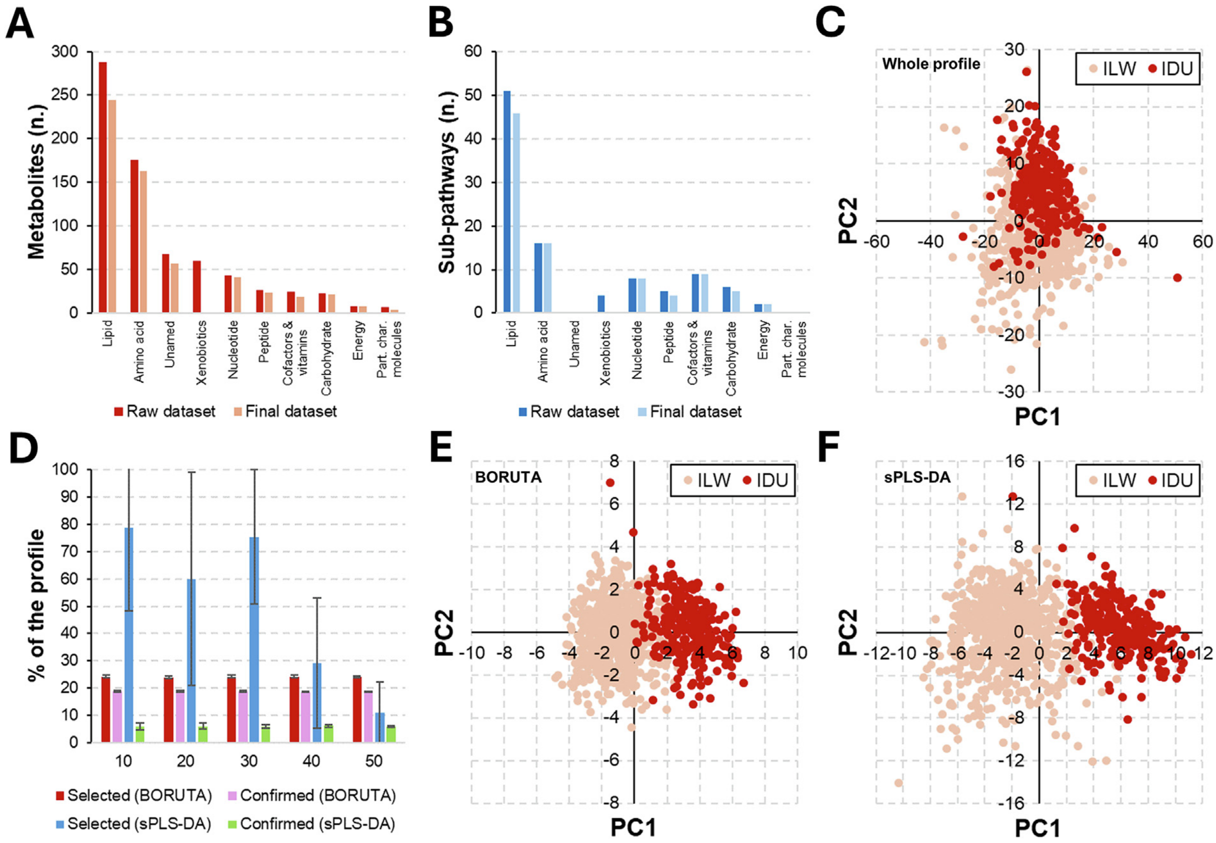

Fig. 1. Metabolomic panel and metabolites differentiating Italian Large White (ILW) and Italian Duroc (IDU) pigs. (A) Distribution of metabolites across super-pathways; (B) Distribution of metabolites across sub-pathways; (C) Principal component analysis (PCA) before feature selection (n. 580 metabolites); (D) Statistics of selection; percentage is given considering the full metabolite set (n. 580 metabolites); (E) PCA based on metabolites selected by Boruta (n. 100); (F) PCA based on metabolites selected only by sparse Partial Least Squares Discriminant Analysis (sPLS-DA) (n. 17). Abbreviations: Part. char. molecules = Partially characterised molecules.

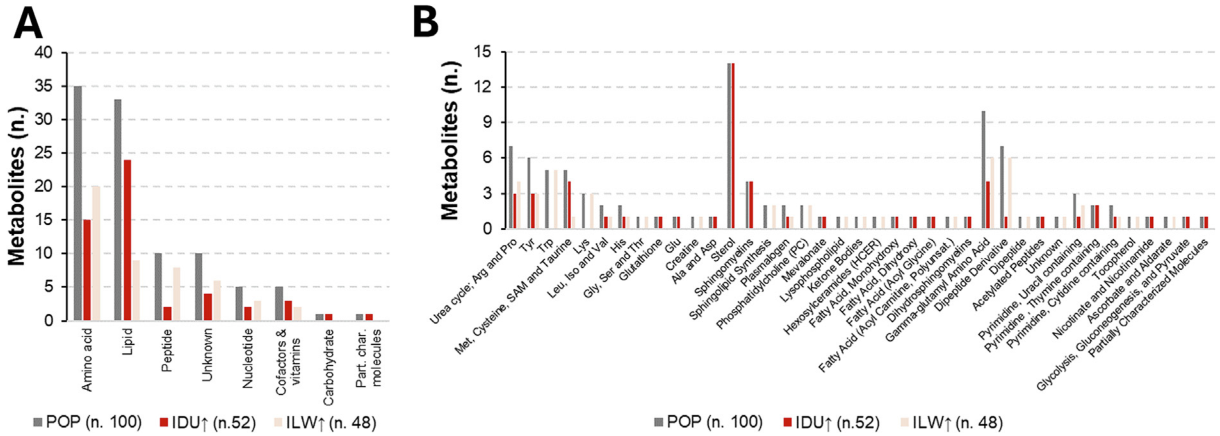

Fig. 2. Biological features of selected metabolites in pigs (n. 100). (A) Distribution of selected metabolites in relation to the super-pathways and the breeds having high concentrations of metabolites; (B) Distribution of selected metabolites in relation to the sub-pathways and the breeds having high concentrations of metabolites. Abbreviations: IDU = Italian Duroc, ILW = Italian Large White, Part. char. molecules = Partially characterised molecules, POP = Whole population.

148 in the single runs, with a mean of 140, representing 24% of the global metabolite dataset; Fig. 1D), and (ii) of the analyses with all 10 rounds of the external 10CV, where a total of 124 metabolites were globally confirmed (ranging from 105 to 113 in the single runs, with a mean of 109, representing 19% of the global metabolite dataset, Fig. 1D). These results indicated that feature selection and confirmation were quite stable despite the random component used for data imputation and analysis and data subsampling via the external 10CV. Approximately, half of the selected metabolites had a higher concentration in Italian Large White pigs (n. 52) while the other half had a higher concentration in Italian Duroc pigs (n. 48).

The sPLS-DA–based analyses led to the identification and confirmation of 17 metabolites as discriminating features (3% of the global metabolite dataset). Similar to the Boruta results, these metabolites were derived from the combination of the analyses (i) across the 125 runs, where all the 722 metabolites were selected at least once (with an average of 295 metabolites selected per run,

Table 1
The list of 100 metabolites showing differences in concentration between the two pig breeds. Metabolites are sorted based on the Area Under the Curve value (from highest to lowest). Additional details are reported in Supplementary Table S2.

First 50 breed-related metabolites				Second 50 breed-related metabolites			
No.[1]	**Metabolite name**	**SP**[2]	**Breed**[3]	**No.**[1]	**Metabolite name**	**SP**[2]	**Breed**[3]
1	Glycosyl ceramide (d18:2/24:1, d18:1/24:2)	Lip	IDU	51	Phenylacetylalanine	Pept	IDU
2	X-15503	U	ILW	52	N-acetyltaurine	AA	IDU
3	Kynurenate	AA	ILW	53	Leucylglycine	Pept	ILW
4	3-methoxytyrosine	AA	IDU	54	Gamma-glutamylcitrulline	Pept	ILW
5	Hydroxypalmitoyl sphingomyelin (d18:1/16:0(OH))	Lip	IDU	55	Gamma-glutamyl-alpha-lysine	Pept	ILW
6	Kynurenine	AA	ILW	56	Palmitoyl sphingomyelin (d18:1/16:0)	Lip	IDU
7	Gamma-glutamylvaline	Pept	IDU	57	Sphingomyelin (d18:1/19:0, d19:1/18:0)	Lip	IDU
8	Glycosyl-N-palmitoyl-sphingosine (d18:1/16:0)	Lip	IDU	58	4-methyl-2-oxopentanoate	AA	IDU
9	Glycosyl ceramide (d18:1/20:0, d16:1/22:0)	Lip	IDU	59	Linoleoylcarnitine (C18:2)	Lip	IDU
10	N-delta-acetylornithine	AA	ILW	60	X-25172	U	IDU
11	X-24736	U	ILW	61	Cystine	AA	ILW
12	Glycosyl-N-stearoyl-sphingosine (d18:1/18:0)	Lip	IDU	62	Gamma-glutamylglutamine	Pept	ILW
13	1-methyl-5-imidazolelactate	AA	ILW	63	X-23423	U	ILW
14	N-acetylmethionine	AA	ILW	64	3-methyl-2-oxobutyrate	AA	IDU
15	N-acetylkynurenine (2)	AA	ILW	65	Anthranilate	AA	ILW
16	3-methylhistidine	AA	ILW	66	Sphingomyelin (d18:1/24:1, d18:2/24:0)	Lip	IDU
17	3-hydroxy-3-methylglutarate	Lip	ILW	67	Sphingomyelin (d17:1/16:0, d18:1/15:0, d16:1/17:0)	Lip	IDU
18	3-methylglutaconate	AA	ILW	68	Sphingomyelin (d18:1/22:2, d18:2/22:1, d16:1/24:2)	Lip	IDU
19	Glycine conjugate of C6H10O2 (2)	PCM	IDU	69	Cystathionine	AA	IDU
20	Sphingomyelin (d18:2/24:2)	Lip	IDU	70	1,5-anhydroglucitol (1,5-AG)	C	IDU
21	Palmitoyl dihydrosphingomyelin (d18:0/16:0)	Lip	IDU	71	6-bromotryptophan	AA	IDU
22	1-methyl-5-imidazoleacetate	AA	ILW	72	Gamma-tocopherol/beta-tocopherol	C&V	IDU
23	N-methylalanine	AA	ILW	73	Gamma-glutamyltyrosine	Pept	ILW
24	S-methylcysteine	AA	IDU	74	Glutamate	AA	IDU
25	Campesterol	Lip	IDU	75	2R,3R-dihydroxybutyrate	Lip	ILW
26	5-methylcytidine	PCM	ILW	76	Gamma-glutamylthreonine	Pept	ILW
27	Carnosine	AA	ILW	77	X-18913	U	ILW
28	Creatinine	AA	ILW	78	Sphingomyelin (d18:1/20:0, d16:1/22:0)	Lip	IDU
29	Sphingomyelin (d18:1/17:0, d17:1/18:0, d19:1/16:0)	Lip	IDU	79	Isobutyrylglycine	AA	IDU
30	Cysteinylglycine disulfide	AA	ILW	80	Ascorbic acid 3-sulfate	C&V	IDU
31	2′-O-methylcytidine	Nucl	IDU	81	3-hydroxybutyrate (BHBA)	Lip	IDU
32	Sphingomyelin (d18:1/20:1, d18:2/20:0)	Lip	IDU	82	2-hydroxyheptanoate	Lip	ILW
33	1-(1-enyl-palmitoyl)-2-linoleoyl-GPC (P-16:0/18:2)	Lip	IDU	83	Gamma-glutamylglycine	Pept	ILW
34	Sphingomyelin (d18:2/23:1)	Lip	IDU	84	Picolinoylglycine	Lip	ILW
35	Hexanoylglycine	Lip	IDU	85	X-11850	U	IDU
36	2′-O-methyluridine	Nucl	IDU	86	Aspartate	AA	IDU
37	X-17354	U	ILW	87	N6-methyllysine	AA	IDU
38	Homocitrulline	AA	ILW	88	Alpha-tocopherol	C&V	ILW
39	trigonelline (N’-methylnicotinate)	C&V	ILW	89	Sarcosine	AA	ILW
40	2-hydroxyoctanoate	Lip	ILW	90	X-22162	U	IDU
41	Thymidine	Nucl	ILW	91	3-indoxyl sulfate	AA	IDU
42	Dimethylglycine	AA	IDU	92	Pro-hydroxy-pro	AA	ILW
43	Methionine sulfone	AA	ILW	93	1-arachidonoyl-GPE (20:4n6)	Lip	ILW
44	Cytidine	Nucl	ILW	94	Anserine	AA	ILW
45	Sphingomyelin (d18:2/24:1, d18:1/24:2)	Lip	IDU	95	Sphinganine-1-phosphate	Lip	ILW
46	X-23593	U	ILW	96	Isoleucine	AA	IDU
47	Sphingomyelin (d18:1/22:1, d18:2/22:0, d16:1/24:1)	Lip	IDU	97	X-11843	U	IDU
48	Leucylhydroxyproline	Pept	ILW	98	1-myristoyl-2-arachidonoyl-GPC (14:0/20:4)	Lip	ILW
49	Tryptophan	AA	IDU	99	2-O-methylascorbic acid	C&V	IDU
50	Sphingomyelin (d18:2/23:0, d18:1/23:1, d17:1/24:1)	Lip	IDU	100	1-palmitoyl-2-palmitoleoyl-GPC (16:0/16:1)	Lip	ILW

[1] Metabolites are ranked by the Area Under the Curve value (see Supplementary Table S2). This identifier is used also for identifying nodes at the network level.

[2] Super-pathway as defined by the Metabolon platform. AA: amino acid; C: carbohydrate; C&V: cofactors and vitamins; Lip: lipid; Nucl: nucleotide; Pept: peptide; PCM: partially characterised molecules; U: unknown.

[3] IDU: higher concentration in Italian Duroc pigs; ILW: higher concentration in Italian Large White pigs.

representing 51% of the global metabolite dataset; ranging from 42 to 578 in the single runs; Fig. 1D) and (ii) with all 10 rounds of the external 10CV, where an average of 35 metabolites were confirmed (6% of the global metabolite dataset; ranging from 27 to 67 in the single runs; Fig. 1D). These results indicate that there was high variability before the application of the CV due to the random component used in selection (rather than the effect of the imputation). However, the instability was resolved when an external 10CV procedure was implemented, as evidenced by the consistent confirmation of a small set of 17 metabolites. Of the selected metabolites, 6 had a higher mean concentration in Italian Large White pigs while the remaining 11 had a lower mean concentration in this breed (and vice versa in the Italian Duroc breed). It is worth mentioning that all 17 metabolites selected with the sPLS-DA analysis are included within the metabolite set chosen with the Boruta analysis (Table 1).

Considering the predictive performance, both selected metabolite sets (one derived from Boruta analyses and one derived from sPLS-DA analyses) were quite effective in separating the pigs of the two breeds: both PCAs based on the two metabolite subsets showed two distinct clusters of pigs representing the two breeds (Fig. 1E and Fig. 1F for Boruta and sPLS-DA, respectively). The separation was primarily driven by Principal Component 1 (as expected), capturing a significant portion of the variance (19.5% for Boruta and 34.2% for sPLS-DA).

As quantitative measures of discrimination power, we calculated the OOB and MDG values as well as the Receiver Operating Characteristic curve for the two sets of metabolites. The metabolites chosen by Boruta had an OOB score of 0.985 (OOB error of 0.015), which was nearly identical to the OOB score value (0.984) obtained with the metabolites selected by sPLS-DA. For the 100 metabolites selected by Boruta, the MDG values ranged from 0.00053 to 0.04878, while the Receiver Operating Characteristic analysis yielded AUC values ranging from 0.53 to 0.92 (mean = 0.74) (Supplementary Figure S2). The correlation between MDG and AUC values was strong (r = 0.83). Out of the 100 metabolites selected with Boruta, the 17 metabolites chosen by sPLS-DA ranked within the top 60, based on MDG (that ranged from 0.003 to 0.04878) and the top 50 based on AUC values (that ranged from 0.74 to 0.93) (Supplementary Figure S2). However, for the remaining 83 metabolites from the Boruta dataset, we also assessed their specific discriminative power. The PCA conducted solely with the 83 metabolites displayed a clear separation of the two breeds (Supplementary Figure S3), with Principal Component 1 explaining 17.2% of the variance and an OOB score of 0.985. The AUC values ranged from 0.53 to 0.89, with approximately 70% of the metabolites having AUC values ≥ 0.7.

Biological features of discriminant metabolites

The metabolites selected by Boruta (100 metabolites) were spread across seven different super-pathways, with amino acids accounting for approximately ~6% of the total profile and lipids also making up around 6%. Within this metabolite dataset, super-pathways were distributed as follows: amino acids (35%), lipids (33%), peptides (10%), nucleotides (5%), cofactors and vitamins (5%), carbohydrates (1%), and partially characterised molecules (1%) (Fig. 2A, grey bars). Unnamed compounds accounted for 10% of the selected metabolites, and no metabolites from the Energy super-pathway were selected. In the case of sPLS-DA (17 metabolites), only four super-pathways were represented, with no nucleotides, cofactors/vitamins and carbohydrates included. Amino acids accounted for ~4% of the total profile, while lipids made up ~3%, similar to Boruta. Within this selection set, super-pathways were distributed as follows: lipids (47%), amino acids (35%), peptides (6%), and partially characterised molecules (6%). Unnamed compounds made up 6% of the selected metabolites.

Stratified by breed, approximately half of the Boruta–identified metabolites showed higher concentrations in Italian Large White pigs (n. 48) while the other half had higher concentrations in Italian Duroc pigs (n. 52). When classified by super-pathways and evaluated in relation to the breed, the relative distributions of metabolites showed some differences. For example, there was a higher concentration of amino acids (n. 20) and peptides (n. 16) in Italian Large White pigs compared to Italian Duroc (n. 15 and n. 2, respectively). Conversely, there was a higher concentration of lipids (n. 24) in Italian Duroc pigs compared to Italian Large White pigs (n. 9) (Fig. 2A). The distribution of selected metabolites in relation to the sub-pathways and breeds is shown in Fig. 2B.

Metabolites had absolute values of Δ% ranging from 6.3 to 193.6 (Table 1), with an average value of 39.2 ± 32.4. Metabolites with a high |Δ|% also showed high MDG and AUC values, with correlations of r = 0.64 and r = 0.58, respectively. A total of 19 metabolites had a |Δ|%>50 and six metabolites had |Δ|% > 100. Specifically, these six metabolites were (i) more abundant in Italian Duroc than Italian Large White, (ii) selected by both Boruta and sPLS-DA [with the exception of glycosyl-N-stearoyl-sphingosine (d18:1/18:0) and glycosyl-N-palmitoyl-sphingosine (d18:1/16:0), which were only selected by Boruta] and belonged to the lipid class except one, the peptide gamma-glutamylvaline. In particular, four lipids belonged to the sub–pathway of hexosylceramides and one (hexanoylglycine) belonged to the Fatty Acid Metabolism (Acyl Glycine). Exosylceramides showed a high correlation with each other but did not correlate with hexanoylglycine. Considering the extreme Δ% values, glycosyl ceramide (d18:2/24:1, d18:1/24:2), a hexosylceramide, had the lowest Δ% value (indicating high concentrations in Italian Duroc pigs), whereas 1-methyl-5-imidazoleacetate, belonging to the histidine metabolism, had the highest Δ% (indicating high concentrations in Italian Large White pigs).

Metabolite networks

The biological characterisation of the selected metabolites initially relied on the construction of a correlation network (based on Pearson's correlation coefficients) annotated by super-pathways. Correlations were calculated initially for each breed, separately, and then at the population level (including all pigs of the two breeds). Within each breed, correlations were normally distributed (Supplementary Figure S4A-B) and similar to each other (Supplementary Figure S4C). At the population level, the distribution was bimodal due to the differential abundance between breeds of metabolites (Supplementary Figure S4D). Correcting for the breed effect removed this bias and allowed true correlations to emerge resulting in a normal distribution (Supplementary Figure S4E). Supplementary Figure S4F shows the relationship between corrected and uncorrected correlations at the population level. Supplementary Figure S5A-B and Supplementary Figure S5C-D show, within each single breed, the relationship with corrected and uncorrected correlations at the population level.

A general correlation network was obtained by retaining only moderate to high correlations (|r|> 0.5) from the three different correlation sets (the two computed within breed and one general metabolite dataset corrected for the breed effects). The network consisted of 31 singletons and 69 metabolites connected by 141 edges (Fig. 3A). The metabolites were grouped into two large clusters (>15 nodes), three medium clusters (3–6 nodes) and eight additional pairs, with most metabolites belonging to the same super pathway (Fig. 3A) or having high abundance in one breed or the other (Fig. 3B). Network statistics can be found in Supplementary Table S2. The first major cluster representing the lipids super-pathway included 24 metabolites, primarily sphingomyelins. This cluster was highly connected and characterised by a high node degree (7.0 ± 4.3). Palmitoyl sphingomyelin (d18:1/16:0) and sphingomyelin (d18:1/24:1, d18:2/24:0) were the two metabolites with the highest node degree, both equal to 14. Two subclusters emerged, connected by the glycosyl-N-palmitoyl-sphingosine (d18:1/16:0) (node n. 8) and palmitoyl sphingomyelin (d18:1/16:0) (node n. 56), which are the two lipids with the highest betweenness centrality. The second major cluster connected several amino acids and urea cycle metabolites. It included 15 metabolites and was less connected (node degree of 3.3 ± 1.9) compared to the cluster of lipids. In this cluster, the un-named compound X-23593 (node n. 46) resulted in the metabolite with the highest node degree (equal to 7) whereas creatinine (node n. 28) resulted in the metabolite with the highest betweenness centrality (0.48) and constituted a central node linking different metabolic routes. Two other un-named metabolites (X-23423 and X-24736) were included in this second–largest cluster. The third major cluster was related to the peptides super pathway and included six metabolites that were highly connected (node degree equal to 4.0 ± 0.8). All of them were classified as gamma-glutamyl amino acids. The other two small clusters included other molecules mainly belonging to the amino acids super pathway. One cluster, that also included an un-named metabolite (X-15503), was constituted by other three metabolites of the kynurenine pathway, related to the tryptophan metabolism and picolinoylglycine. The other small cluster included metabolites related to the leucine, isoleucine and valine metabolism. We did not identify any relationship between node features at the network

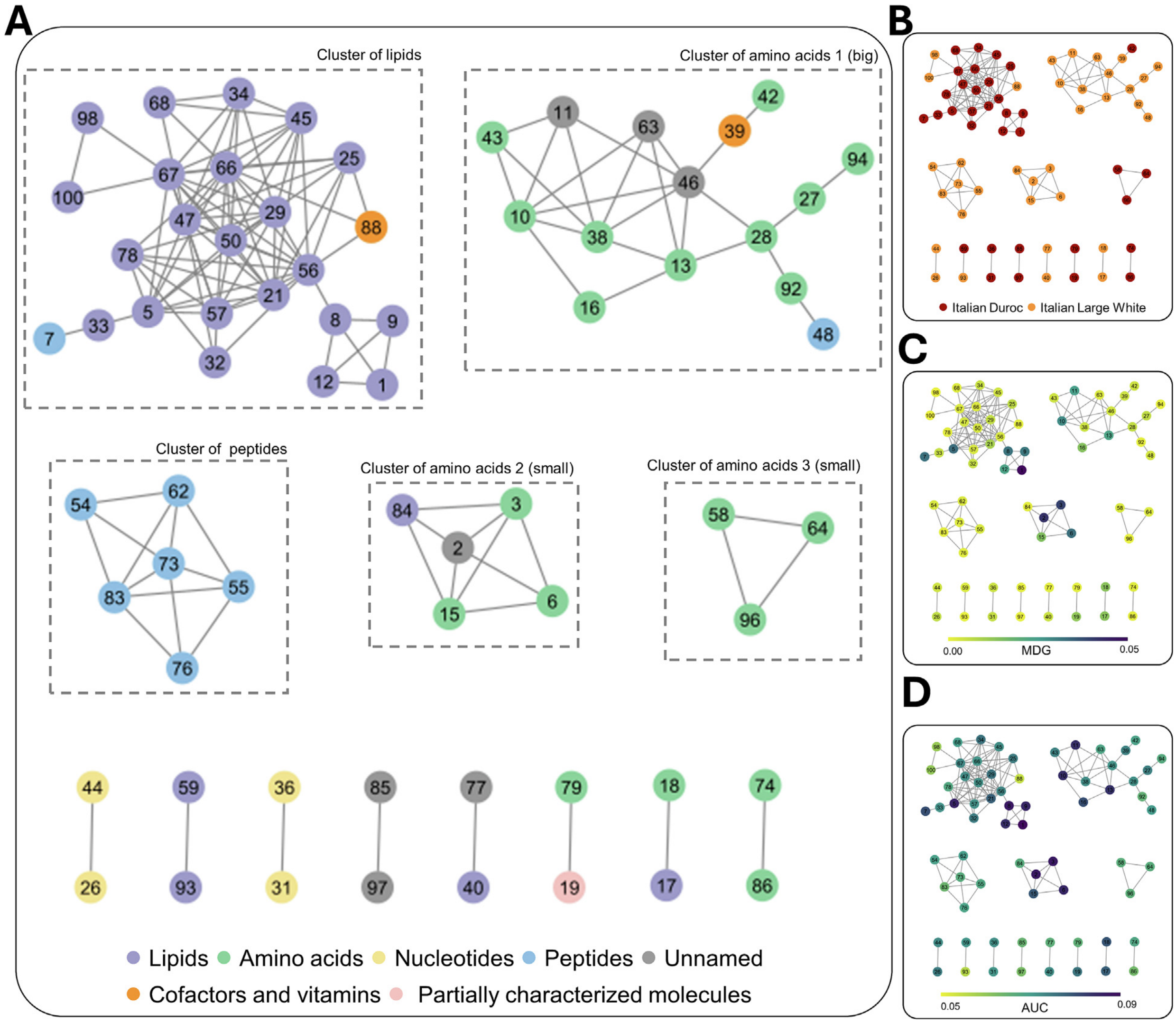
A
Cluster of lipids
Cluster of amino acids 1 (big)
Cluster of peptides
Cluster of amino acids 2 (small)
Cluster of amino acids 3 (small)
● Lipids ● Amino acids ● Nucleotides ● Peptides ● Unnamed
● Cofactors and vitamins ● Partially characterized molecules
B
● Italian Duroc ● Italian Large White
C
0.00
MDG
0.05
D
0.05
AUC
0.09

Fig. 3. Connected components of the correlation network involving selected metabolites (n.100) in pigs. (A) Metabolites are coloured by super-pathway (as provided by Metabolon). Major clusters are evidenced; (B-D) Metabolites are coloured respectively by breed (breed having a higher concentration of the metabolite), MDG (from Random Forest) and AUC (from ROC analysis), respectively. Node identifiers are given in Table 1. Abbreviations: AUC = Area under the curve, MDG = Mean decrease Gini, ROC = Receiver operating characteristic.

level (e.g. node degree, centrality, etc.) and MDG, AUC or |Δ|%. However, we observed that important metabolites tend to be connected with each other as part of subclusters (Fig. 3C–D).

Functional characterisation of discriminant metabolites

The functional characterisation of metabolites was carried out using Metabolon's annotations. The metabolites were grouped into 38 subpathways, six of them containing at least five metabolites each, including Sphingomyelins (n. 14), Gamma-glutamyl Amino Acid (n. 7), Tryptophan Metabolism (n. 7), Methionine, Cysteine, SAM and Taurine Metabolism (n. 6), Histidine Metabolism (n. 5), and Leucine, Isoleucine and Valine Metabolism (n. 5).

In addition to this classification, metabolites were also analysed for over–representation in biological pathways using the MetaboAnalyst webserver. Only 57 metabolites were successfully mapped to the pathway libraries and metabolite sets available in MetaboAnalyst v.5.0; the majority of excluded metabolites were lipids.

A total of 57 pathway libraries and metabolite sets, including 28 out of 57 metabolites, were found to be significantly overrepresented (False discovery rate corrected $P < 0.05$; Supplementary Table S3). It is worth mentioning that metabolite sets in RaMP-DB come from various sources and may contain overlapping information (e.g. the metabolism of amino acids and derivatives from the Reactome database and the amino acid metabolism from WikiPathways). As a result, the ability to link pathways in a network context allowed for the identification of clusters of pathways (Fig. 4). The top enriched metabolite sets (False discovery rate corrected $P < 0.01$; Table 2) included general pathways, that annotated hundreds of molecules each, and featured several of the selected metabolites. Examples include pathways related to the metabolism of amino acids (annotating 109–283 metabolites) and the transport of molecules (annotating 90–208 metabolites). However, more specific sets of metabolites characterised by a lower number (<30) of annotated metabolites emerged. These sets include leucine, isoleucine and valine metabolism, the tryptophan catabolism/metabolism and beta-alanine metabolisms. Considering these smaller sets, a cluster related to tryptophan emerged (Fig. 4), encompassing and linking four metabolite sets (tryptophan metabolism, tryptophan catabolism, tryptophan catabolism leading to NAD + production and the kynurenine pathway with links

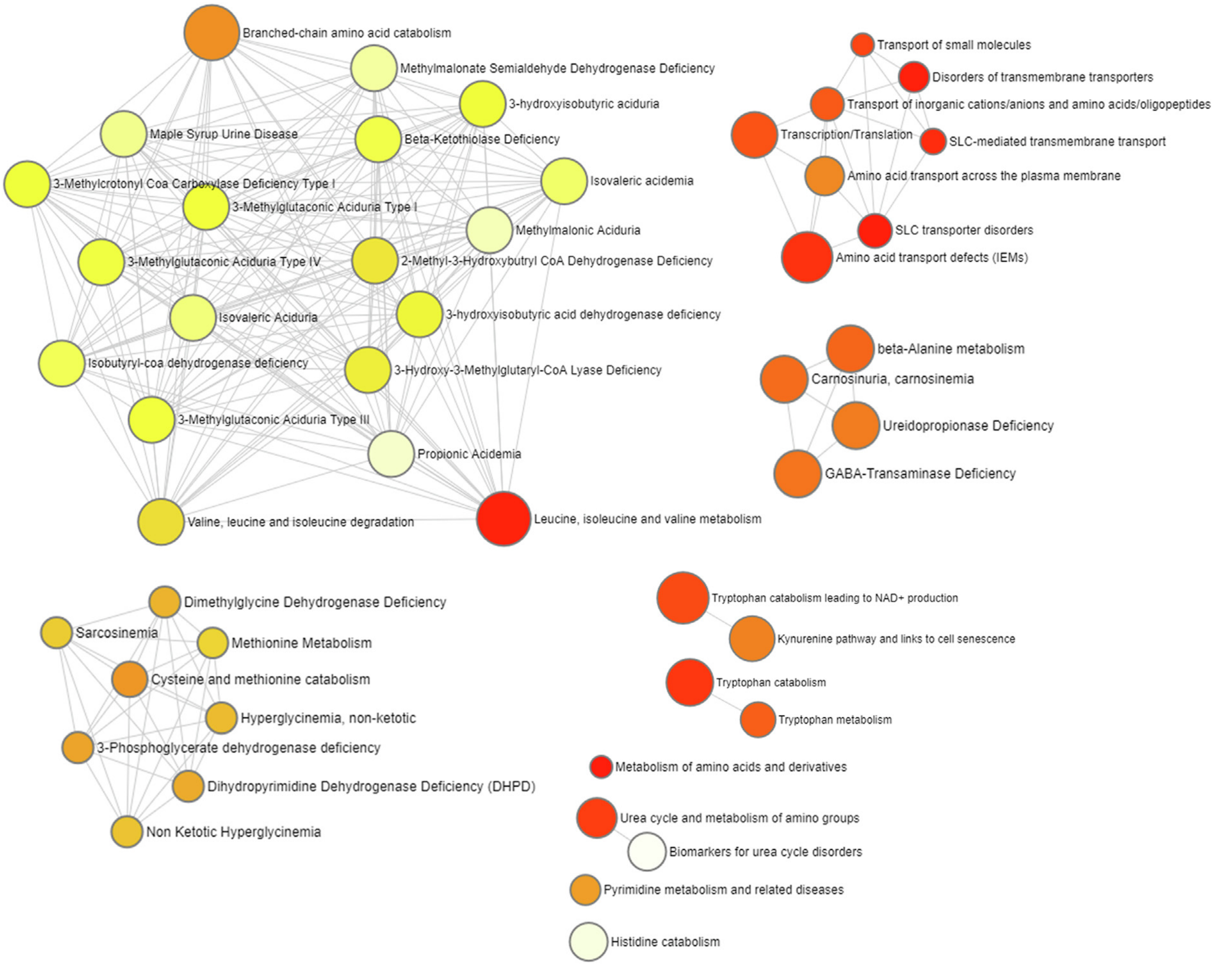
Branched-chain amino acid catabolism
Methylmalonate Semialdehyde Dehydrogenase Deficiency
3-hydroxyisobutyric aciduria
Maple Syrup Urine Disease
Beta-Ketothiolase Deficiency
Isovaleric acidemia
3-Methylcrotonyl Coa Carboxylase Deficiency Type I
3-Methylglutaconic Aciduria Type I
Methylmalonic Aciduria
3-Methylglutaconic Aciduria Type IV
2-Methyl-3-Hydroxybutryl CoA Dehydrogenase Deficiency
Isovaleric Aciduria
3-hydroxyisobutyric acid dehydrogenase deficiency
Isobutyryl-coa dehydrogenase deficiency
3-Hydroxy-3-Methylglutaryl-CoA Lyase Deficiency
3-Methylglutaconic Aciduria Type III
Propionic Acidemia
Valine, leucine and isoleucine degradation
Leucine, isoleucine and valine metabolism
Transport of small molecules
Disorders of transmembrane transporters
Transport of inorganic cations/anions and amino acids/oligopeptides
Transcription/Translation
SLC-mediated transmembrane transport
Amino acid transport across the plasma membrane
SLC transporter disorders
Amino acid transport defects (IEMs)
beta-Alanine metabolism
Carnosinuria, carnosinemia
Ureidopropionase Deficiency
GABA-Transaminase Deficiency
Dimethylglycine Dehydrogenase Deficiency
Sarcosinemia
Methionine Metabolism
Cysteine and methionine catabolism
Hyperglycinemia, non-ketotic
3-Phosphoglycerate dehydrogenase deficiency
Dihydropyrimidine Dehydrogenase Deficiency (DHPD)
Non Ketotic Hyperglycinemia
Tryptophan catabolism leading to NAD+ production
Kynurenine pathway and links to cell senescence
Tryptophan catabolism
Tryptophan metabolism
Metabolism of amino acids and derivatives
Urea cycle and metabolism of amino groups
Biomarkers for urea cycle disorders
Pyrimidine metabolism and related diseases
Histidine catabolism

Fig. 4. Relationships between enriched metabolite sets, based on the metabolites discriminating between the two pig breeds. The colour of nodes represents the *P*-value whereas the size of the nodes represents the enrichment ratio (Supplementary Table S3). Abbreviation: SLC = Solute carrier, GABA = Gamma-aminobutyric acid.

Table 2
Over-represented pathways ($P_{FDR} < 0.01$) on the RaMP-DB database, based on the metabolites discriminating between the two pig breeds. Full results are given in Supplementary Table S3.

Metabolite Set	Total[1]	Hits[2]	Enrichment ratio[3]	*P*	FDR corrected *P*
Metabolism of amino acids and derivatives	283	14	6.83	2.12E-09	7.04E-06
SLC transporter disorders	80	8	13.82	5.81E-08	9.64E-05
Disorders of transmembrane transporters	98	8	11.27	2.89E-07	2.71E-04
Leucine, isoleucine and valine metabolism	67	7	14.43	3.27E-07	2.71E-04
Amino acid metabolism	109	8	10.14	6.62E-07	3.90E-04
Biochemical pathways: part I	445	14	4.35	7.07E-07	3.90E-04
SLC-mediated transmembrane transport	155	9	8.04	8.64E-07	4.09E-04
Amino acid transport defects (IEMs)	27	5	25.51	1.03E-06	4.25E-04
Tryptophan catabolism	33	5	20.92	2.93E-06	0.00108
Urea cycle and metabolism of amino groups	34	5	20.33	3.41E-06	0.00113
Transport of small molecules	208	9	5.96	1.01E-05	0.00303
Tryptophan catabolism leading to NAD+production	23	4	23.95	1.81E-05	0.005
Transcription/Translation	25	4	22.10	2.56E-05	0.00653
Transport of inorganic cations/anions and amino acids/oligopeptides	52	5	13.26	2.91E-05	0.00677
Tryptophan metabolism	53	5	13.02	3.20E-05	0.00677
beta-Alanine metabolism	28	4	19.70	4.08E-05	0.00677
Carnosinuria, carnosinemia	28	4	19.70	4.08E-05	0.00677
GABA-Transaminase Deficiency	28	4	19.70	4.08E-05	0.00677
Ureidopropionase Deficiency	28	4	19.70	4.08E-05	0.00677
Kynurenine pathway and links to cell senescence	28	4	19.70	4.08E-05	0.00677
Amino acid transport across the plasma membrane	32	4	17.24	7.03E-05	0.0111

Abbreviations: FDR = False Discovery Rate, SLC = Solute carrier, GABA = Gamma-aminobutyric acid.

[1] No. of metabolites annotated in the set.

[2] Selected metabolites belonging to the set.

[3] Enrichment Ratio is computed by Hits / Expected, where hits = observed hits; expected = expected hits. Details are given in Supplementary Table S3.

to cell senescence). In these pathways, kynurenate and kynurenine were found to be the metabolites with the highest correlation ($r = 0.76$). Beta-alanine metabolism was part of a different cluster (Fig. 4) also encompassing four disease–related pathways (including GABA-Transaminase deficiency, Ureidopropionase deficiency, carnosinuria and carnosinemia). This cluster also included other four of the selected metabolites: carnosine, anserine, glutamate and aspartate. In particular, anserine and carnosine showed a high correlation ($r = 0.96$) as well as glutamate and aspartate ($r = 0.70$).

Discussion

In recent years, metabolomics has been used to gain valuable insights into the biological complexity of the mammalian metabolism, including the pig, contributing to the dissection of some of the molecular mechanisms underlying economically relevant traits in this livestock species (Bovo et al., 2015, 2016, 2023; Carmelo et al., 2020; Luise et al., 2020; Metzler-Zebeli et al., 2023; Peukert et al., 2021; Wang et al., 2021; Zhang et al., 2021). Metabolomics has opened up the possibility to analyse hundreds of molecular phenotypes that contribute to the dissection of the animal phenome, with a great potential for application in animal breeding and selection (Fontanesi, 2016; Goldansaz et al., 2017; Pérez-Enciso and Steibel, 2021). Different breeds or lines are also typically used in crossbreeding programmes. Therefore, it is crucial to characterise breed-specific metabolomic fingerprints that can offer an overall view of the fundamental metabolic differences between breeds that, when combined, contribute to enhance production performances in crossbred pigs (D'Alessandro et al., 2011; Straadt et al., 2014; Bovo et al., 2016; Carmelo et al., 2020; Lefort et al., 2020; Xie et al., 2023). Because of their relevance for the pig industry, we previously carried out a pilot study where we initially characterised part of the metabolome of the same two heavy pig breeds investigated in this study, i.e. Italian Duroc and Italian Large White. Despite the relatively small number of investigated animals and molecules (12 pigs per breed and ~180 targeted metabolites), we showed that metabolomics can disclose the effect of different genetic backgrounds (represented by the two breeds) on metabolic features (Bovo et al., 2016).

In this study, we scaled up the investigation of the molecular phenome characterisation of the Italian Duroc and Italian Large White breeds by applying MS-based untargeted global metabolomics approach on a total of ~950 pigs. The applied high throughput metabolomic platforms measured ~800 metabolites (722 were then retained) covering seven metabolic super pathways and ~100 specific pathways, analysed with two supervised multivariate machine learning–based approaches. One approach, namely sPLS-DA, derives from PLS discriminant analysis, a popular and ubiquitously applied chemometric method used in the field of metabolomics for dimensionality reduction. Briefly, PLS-DA finds a linear regression model by projecting the predicted variables and the observed variables into a new space and providing several statistics (e.g. loading weights and variable importance on projection scores) for the selection of relevant variables (Brereton and Lloyd, 2014). However, while in PLS discriminant analysis the criteria and threshold for extracting these variables are left to the user, sPLS-DA applies regularisation for the simultaneous dimensionality reduction and variable selection (Chung and Keles, 2010). Thus, this new algorithm was introduced to shift from solely classification and prediction to variable selection (sPLS-DA). Despite being highly performant, one major drawback of these techniques is class separation and classification, even when a dataset does not contain any relevant variables (e.g. a dataset containing n random data points), highlighting the need for careful evaluation of results and the need to involve other steps in (s) PLS-DA (Yi et al., 2016). This includes the implementation of an additional cross-validation procedure (Ruiz-Perez et al., 2020). Moreover, as mentioned at the beginning, sPLS-DA is characterised by the minimal-optimal problem, which leads to selecting only those variables relevant for obtaining a good model rather than all variables related to the analytical problem (all-relevant problem). To better understand the biological problem as well as address algorithmic purposes, features, and popularity, we used a second approach for data analysis. We chose the Random Forest algorithm, specifically Boruta (Kursa et al., 2010), because (i) it is an all-relevant algorithm and (ii) it does not require (in principle) any additional CV procedure (Huang and Deng, 2021). As for other algorithms, despite having performance comparable to the PLS family and being around for a similar length of time (Trainor et al., 2017; Mendez et al., 2019; Vu et al., 2019, Galal et al., 2022), Random Forest has not achieved the same level of popularity as PLS methods, resulting in its underuse. To address the biological problem, we established a bioinformatic pipeline to evaluate and mitigate the randomness introduced by (i) the algorithm (as part of its workflow), (ii) the animals under investigation and (iii) the imputed data.

Application of sPLS-DA let to the identification of a small number of discriminating metabolites (n. 17) whereas Boruta returned a larger number (n. 100) that included all the metabolites coming from sPLS-DA. The first results highlight how the two algorithms provide complementary outputs (Gromski et al., 2015; Mendez et al., 2019). Specifically, 17 metabolites relate to the minimal-optimal problem, while the remaining 83 metabolites relate to the all-relevant problem. Both sets of metabolites demonstrated good and similar classification performance despite differing in size. Secondly, the application of a pipeline evaluating random components at different levels highlighted the instability of sPLS-DA in the selection of relevant features. This result supports the findings of a previous study where the random seed of analysis was shown to impact the selection of optimal tuning parameters by influencing the required tuning parameters of the algorithm (Olson Hunt et al., 2014). Instability is also exacerbated by multicollinearity, due to the presence of correlated variables (which is common in metabolomics data) and affects the stability and accuracy of coefficient estimates (Olson Hunt et al., 2014). This multicollinearity has been shown to degrade the performance of regularisation in sPLS-DA (Fan and Lv, 2008; Lee et al., 2021), which may explain some of the instability observed in our results. A stable core of metabolites was only achieved by sPLS-DA after applying an external CV procedure, underscoring the importance of additional validation steps in (s)PLS-DA (Yi et al., 2016, Ruiz-Perez et al., 2020). In contrast, Boruta remained highly stable both before and after the external CV procedure, highlighting how the Random Forest algorithm typically does not require CV, as each tree is grown from a bootstrapped sample. A direct example that indicates how the larger metabolite set identified by Boruta is stable and weakly affected by randomness and addresses the all-relevant problem derived from metabolites of the kynurenine pathway. Both algorithms (sPLS-DA and Boruta) selected tryptophan and kynurenine, but only Boruta selected other three correlated key molecules of the kynurenine pathway (i.e. kynurenate, N-acetylkynurenine, and anthranilate).

Based on these considerations, we thereafter illustrated differences between breeds and biological pathways defined on all 100 metabolites selected by Boruta, which provide a more comprehensive picture of the biological systems that differ between the two investigated breeds. About half of these 100 metabolites had a higher concentration in one or the other breed, with some peculiar features at the super-pathway level. The level of plasma lipids was typically higher in Italian Duroc pigs whereas peptides and amino acids had higher concentration in Italian Large White pigs.

Specifically, molecules of two classes of sphingolipids, namely hexosylceramides and sphingomyelins (**SM**), were among the lipids with higher level in Italian Duroc than in Italian Large White pigs. Within these classes, glycosyl ceramide (d18:2/24:1, d18:1/24:2) and hydroxypalmitoyl sphingomyelin (d18:1/16:0 (OH)) resulted to be the metabolites with the highest |Δ%| and discriminative power. Among the sphingomyelins, as a proof of concept, the palmitoyl sphingomyelin (d18:1/16:0; also indicated as SM C16:0) confirmed its higher abundance in Italian Duroc pigs reported in our previous study that used another metabolomic platform (Bovo et al., 2016). Sphingolipids are involved in the modulation of several biological processes including immune functions among other roles, and in humans, they have been implicated in pathological states, such as neurodegenerative processes and metabolic disorders (Hannun and Obeid, 2018). The general higher content of circulating lipids in Italian Duroc pigs reflects some characteristics of this breed on other phenotypic traits related to meat and carcass traits. For example, a peculiar characteristic of the Duroc breed is the higher level of intermuscular and intramuscular fat content (Armero et al., 1999; Suzuki et al., 2003; Alonso et al., 2015). The relationship between subcutaneous and intramuscular fat content and composition and blood (serum or plasma) lipid molecule concentration in this breed has been preliminarily investigated (e.g. Alonso et al., 2009, 2015; Muñoz et al., 2012; Tor et al., 2021; Hou et al., 2023). The results obtained in this study enlarge the number of lipids to be considered to provide a better comparative evaluation of lipid composition between different pig tissues (blood vs fat depots).

On the other hand, Italian Large White pigs had a higher plasma content of molecules belonging to the amino acid super-pathways. These molecules include those of the kynurenine pathway, several biogenic amines and gamma-glutamyl amino acids, among several other metabolites, some of which with important roles in modulating the activity of the mammalian immune, reproductive, and central nervous systems (Savitz, 2020), indicating potential peculiar biological characteristics and specialisations of the pigs of this breed. The higher level of creatinine, carnosine and the derived anserine might be indirectly interesting for the pork industry, considering that these molecules can be useful as indicators of muscle metabolism and meat quality attributes, including pH, colour and flavour (Ma et al., 2010; D'Astous-Pagé et al., 2017). Additionally, the higher level of N-delta-acetylornithine and kynurenine in Italian Large White pigs confirms the results reported in our previous exploratory study (Bovo et al., 2016), further supporting the more detailed picture obtained in the current study.

The network and enrichment analyses described all these molecular features that contributed to identifying the metabolic profiles that discriminate between the two breeds. These profiles also showed that not all discriminating lipids or amino acids/peptides within the Boruta list had a higher content exclusively in one breed or the other. This suggests that the different genetic backgrounds of the two breeds may affect, at least in part, the lipid and amino acid metabolisms in different ways.

The constructed network shed light on several unnamed metabolites, particularly those included in amino acid/peptide–enriched networks. For example, the connection of three unnamed metabolites that clustered together (X-23423, X-23593, and X-24736) is indirectly supported in terms of potential function/role. This is evidenced by genome-wide association studies in humans identifying that their levels in urine are all associated with the same chromosome region, including the N-acetyltransferase 8 (putative) (*NAT8*) gene (Schlosser et al., 2020). This gene encodes an enzyme that catalyses the N-acetylation of cysteine conjugates (Veiga-da-Cunha et al., 2010). Genetic variants in the same gene are associated with an increased risk of Chronic Kidney Disease and related molecular markers, such as the circulating level of several N-acetylated amino acids, including N-delta-acetylornithine (Schlosser et al., 2020). N-delta-acetylornithine is connected in the same cluster with all three unnamed metabolites in our pig study. Another example comes from X-15503, which showed high correlations with kynurenine, kynurenate and N-acetylkynurenine. To support its functional link with these metabolites of the kynurenine pathway, genome-wide association studies in humans have reported the association between all these metabolites (including X-15503) and key genes involved in the kynurenine pathway (Yin et al., 2022; Chen et al., 2023). Therefore, by integrating this genetically relevant information with the clustering results we obtained in pigs, as well as other supporting information in humans, it could potentially be possible to assign a biochemical name to some of the metabolites that have not yet been biochemically characterised.

In conclusion, this study has contributed to establishing a comparative metabolomic profile between pig breeds, indicating several significant biochemical differences that are indirectly influenced by the breed–specific genetic background. These differences could be useful to integrate molecular phenotypes in explaining heterosis when these two pigs are crossed. Other studies are currently being conducted to dissect the genetic factors that underlie the metabolic mechanisms shaping the molecular phenome of Italian Duroc and Italian Large White pigs. Once this information becomes available, potential practical applications of these results are also envisioned in designing novel breeding programmes in pigs that include metabotypes.

Supplementary material

Supplementary material to this article can be found online at https://doi.org/10.1016/j.animal.2024.101393.

Ethics approval

Not applicable.

Data and model availability statement

The data that support the study findings are publicly available at: https://doi.org/10.5281/zenodo.14044561.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) did not use any AI and AI-assisted technologies.

Author ORCIDs

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

CRediT authorship contribution statement

S. Bovo: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **M. Bolner:** Writing – review & editing, Writing – original draft, Investigation, Formal analysis, Data curation. **G. Schiavo:** Writing – review & editing, Data curation. **G. Galimberti:** Writing – review & editing, Formal analysis. **F. Bertolini:** Writing – review & editing, Data curation. **S. Dall'Olio:** Writing – review & editing, Data curation. **A. Ribani:** Writing – review & editing, Formal analysis. **P. Zambonelli:** Writing – review & editing, Data curation. **M. Gallo:** Writing – review & editing, Methodology. **L. Fontanesi:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Conceptualization.

Declaration of interest

None.

Acknowledgements

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

Financial support statement

This study was supported by: (1) PRIN2017 PigPhenomics that has received funding from the Italian MUR (for the production metabolomic data); (2) the Horizon Europe Re-Livestock project that has received funding from the European Union's Horizon Europe research and innovation programme under the grant agreement No. 01059609 (for the elaboration and analysis of the data); (3) 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; (4) 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); (5) and was carried out within the Agritech National Research Center (Agritech Spoke 1) and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4 – D.D. 1032 17/06/2022, CN00000022) - This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

References

Alonso, V., Campo, M.D.M., Español, S., Roncalés, P., Beltrán, J.A., 2009. Effect of crossbreeding and gender on meat quality and fatty acid composition in pork. Meat Science 81, 209–217.

Alonso, V., Muela, E., Gutiérrez, B., Calanche, J.B., Roncalés, P., Beltrán, J.A., 2015. The inclusion of Duroc breed in maternal line affects pork quality and fatty acid profile. Meat Science 107, 49–56.

Armero, E., Flores, M., Toldrá, F., Barbosa, J.-A., Olivet, J., Pla, M., Baselga, M., 1999. Effects of pig sire type and sex on carcass traits, meat quality and sensory quality of dry-cured ham. Journal of the Science of Food and Agriculture 79, 1147–1154.

Azur, M.J., Stuart, E.A., Frangakis, C., Leaf, P.J., 2011. Multiple imputation by chained equations: what is it and how does it work? International Journal of Methods in Psychiatric Research 20, 40–49.

Bovo, S., Mazzoni, G., Calò, D.G., Galimberti, G., Fanelli, F., Mezzullo, M., Schiavo, G., Scotti, E., Manisi, A., Samoré, A.B., Bertolini, F., Trevisi, P., Bosi, P., Dall'Olio, S., Pagotto, U., Fontanesi, L., 2015. Deconstructing the pig sex metabolome: targeted metabolomics in heavy pigs revealed sexual dimorphisms in plasma biomarkers and metabolic pathways. Journal of Animal Science 93, 5681–5693.

Bovo, S., Mazzoni, G., Galimberti, G., Calò, D.G., Fanelli, F., Mezzullo, M., Schiavo, G., Manisi, A., Trevisi, P., Bosi, P., Dall'Olio, S., Pagotto, U., Fontanesi, L., 2016. Metabolomics evidences plasma and serum biomarkers differentiating two heavy pig breeds. Animal 10, 1741–1748.

Bovo, S., Ribani, A., Muñoz, M., Alves, E., Araujo, J.P., Bozzi, R., Čandek-Potokar, M., Charneca, R., Di Palma, F., Etherington, G., Fernandez, A.I., García, F., García-Casco, J., Karolyi, D., Gallo, M., Margeta, V., Martins, J.M., Mercat, M.J., Moscatelli, G., Fontanesi, L., 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.

Bovo, S., Schiavo, G., Galimberti, G., Fanelli, F., Bertolini, F., Dall'Olio, S., Pagotto, U., Fontanesi, L., 2023. Comparative targeted metabolomic profiles of porcine plasma and serum. Animal 17, 101029.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Brereton, R.G., Lloyd, G.R., 2014. Partial least squares discriminant analysis: taking the magic away. Journal of Chemometrics 28, 213–225.

Buchanan, D.S., Stalder, K., 2011. Breeds of pigs. In: Rotschild, M., Ruvinsky, A. (Eds.), The Genetics of the Pig. CABI Publishing, Wallingford, UK, pp. 445–472.

Carmelo, V.A.O., Banerjee, P., Da Silva Diniz, W.J., Kadarmideen, H.N., 2020. Metabolomic networks and pathways associated with feed efficiency and related-traits in Duroc and Landrace pigs. Scientific Reports 10, 255.

Chen, Y., Lu, T., Pettersson-Kymmer, U., Stewart, I.D., Butler-Laporte, G., Nakanishi, T., Cerani, A., Liang, K.Y.H., Yoshiji, S., Willett, J.D.S., Su, C.Y., Raina, P., Greenwood, C.M.T., Farjoun, Y., Forgetta, V., Langenberg, C., Zhou, S., Ohlsson, C., Richards, J.B., 2023. Genomic atlas of the plasma metabolome prioritizes metabolites implicated in human diseases. Nature Genetics 55, 44–53.

Chung, D., Keles, S., 2010. Sparse partial least squares classification for high dimensional data. Statistical Applications in Genetics and Molecular Biology 9, 17.

D'Alessandro, A., Marrocco, C., Zolla, V., D´Andrea, M., Zolla, L., 2011. Meat quality of the longissimus lumborum muscle of Casertana and Large White pigs: Metabolomics and proteomics intertwined. Journal of Proteomics 75, 610–627.

D'Astous-Pagé, J., Gariépy, C., Blouin, R., Cliche, S., Sullivan, B., Fortin, F., Palin, M.-F., 2017. Carnosine content in the porcine longissimus thoracis muscle and its association with meat quality attributes and carnosine-related gene expression. Meat Science 124, 84–94.

Degenhardt, F., Seifert, S., Szymczak, S., 2019. Evaluation of variable selection methods for random forests and omics data sets. Briefings in Bioinformatics 20, 492–503.

Deng, L., Li, W., Liu, W., Liu, Y., Xie, B., Groenen, M.A.M., Madsen, O., Yang, X., Tang, Z., 2023. Integrative metabolomic and transcriptomic analysis reveals difference in glucose and lipid metabolism in the longissimus muscle of Luchuan and Duroc pigs. Frontiers in Genetics 14, 1128033.

Fabbri, M.C., Lozada-Soto, E., Tiezzi, F., Čandek-Potokar, M., Bovo, S., Schiavo, G., Fontanesi, L., Muñoz, M., Ovilo, C., Bozzi, R., 2024. Persistence of autozygosity in crossbreds between autochthonous and cosmopolitan breeds of swine: a simulation study. Animal 18, 101070.

Fan, J., Lv, J., 2008. Sure independence screening for Ultrahigh dimensional feature space. Journal of the Royal Statistical Society Series b: Statistical Methodology 70, 849–911.

Faquih, T., Van Smeden, M., Luo, J., Le Cessie, S., Kastenmüller, G., Krumsiek, J., Noordam, R., Van Heemst, D., Rosendaal, F.R., Van Hylckama Vlieg, A., Willems Van Dijk, K., Mook-Kanamori, D.O., 2020. A workflow for missing values imputation of untargeted metabolomics data. Metabolites 10, 486.

Fiehn, O., 2002. Metabolomics – the link between genotypes and phenotypes. Plant Molecular Biology 48, 155–171.

Fontanesi, L., 2016. Metabolomics and livestock genomics: Insights into a phenotyping frontier and its applications in animal breeding. Animal Frontiers 6, 73–79.

Galal, A., Talal, M., Moustafa, A., 2022. Applications of machine learning in metabolomics: disease modeling and classification. Frontiers in Genetics 13, 1017340.

Goldansaz, S.A., Guo, A.C., Sajed, T., Steele, M.A., Plastow, G.S., Wishart, D.S., 2017. Livestock metabolomics and the livestock metabolome: a systematic review. PLoS One 12, e0177675.

Gromski, P.S., Muhamadali, H., Ellis, D.I., Xu, Y., Correa, E., Turner, M.L., Goodacre, R., 2015. A tutorial review: Metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding. Analytica Chimica Acta 879, 10–23.

Hannun, Y.A., Obeid, L.M., 2018. Sphingolipids and their metabolism in physiology and disease. Nature Reviews Molecular Cell Biology 19, 175–191.

Hollywood, K., Brison, D.R., Goodacre, R., 2006. Metabolomics: current technologies and future trends. Proteomics 6, 4716–4723.

Hou, X., Zhang, R., Yang, M., Niu, N., Wu, J., Shu, Z., Zhang, P., Shi, L., Zhao, F., Wang, L., Wang, L., Zhang, L., 2023. Metabolomics and lipidomics profiles related to intramuscular fat content and flavor precursors between Laiwu and Yorkshire pigs. Food Chemistry 404, 134699.

Houle, D., Govindaraju, D.R., Omholt, S., 2010. Phenomics: the next challenge. Nature Reviews Genetics 11, 855–866.

Huang, S., Deng, H., 2021. Data Analytics: A Small Data Approach. Chapman and Hall/CRC, New York, NY, USA.

Kim, J.A., Cho, E.S., Jeong, Y.D., Choi, Y.H., Kim, Y.S., Choi, J.W., Kim, J.S., Jang, A., Hong, J.K., Sa, S.J., 2020. The effects of breed and gender on meat quality of Duroc, Pietrain, and their crossbred. Journal of Animal Science and Technology 62, 409–419.

Krumsiek, J., Suhre, K., Evans, A.M., Mitchell, M.W., Mohney, R.P., Milburn, M.V., Wägele, B., Römisch-Margl, W., Illig, T., Adamski, J., Gieger, C., Theis, F.J., Kastenmüller, G., 2012. Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information. PLoS Genetics 8, e1003005.

Kumar, P., Abubakar, A.A., Verma, A.K., Umaraw, P., Adewale Ahmed, M., Mehta, N., Nizam Hayat, M., Kaka, U., Sazili, A.Q., 2023. New insights in improving sustainability in meat production: opportunities and challenges. Critical Reviews in Food Science and Nutrition 63, 11830–11858.

Kursa, M.B., Jankowski, A., Rudnicki, W.R., 2010. Boruta – a system for feature selection. Fundamenta Informaticae 101, 271–285.

Lee, S.H., Kao, G.D., Feigenberg, S.J., Dorsey, J.F., Frick, M.A., Jean-Baptiste, S., Uche, C. Z., Cengel, K.A., Levin, W.P., Berman, A.T., Aggarwal, C., Fan, Y., Xiao, Y., 2021. Multiblock discriminant analysis of integrative 18F-FDG-PET/CT radiomics for predicting circulating tumor cells in early-stage non-small cell lung cancer treated with stereotactic body radiation therapy. International Journal of Radiation Oncology, Biology, Physics 110, 1451–1465.

Lefort, G., Servien, R., Quesnel, H., Billon, Y., Canario, L., Iannuccelli, N., Canlet, C., Paris, A., Vialaneix, N., Liaubet, L., 2020. The maturity in fetal pigs using a multi-fluid metabolomic approach. Scientific Reports 10, 19912.

Luise, D., Bovo, S., Bosi, P., Fanelli, F., Pagotto, U., Galimberti, G., Mazzoni, G., Dall'Olio, S., Fontanesi, L., 2020. Targeted metabolomic profiles of piglet plasma reveal physiological changes over the suckling period. Livestock Science 231, 103890.

Ma, X.Y., Jiang, Z.Y., Lin, Y.C., Zheng, C.T., Zhou, G.L., 2010. Dietary supplementation with carnosine improves antioxidant capacity and meat quality of finishing pigs: carnosine improve the antioxidant ability and meat quality of pigs. Journal of Animal Physiology and Animal Nutrition 94, e286–e295.

Mendez, K.M., Reinke, S.N., Broadhurst, D.I., 2019. A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. Metabolomics 15, 150.

Metzler-Zebeli, B.U., Lerch, F., Yosi, F., Vötterl, J.C., Koger, S., Aigensberger, M., Rennhofer, P.M., Berthiller, F., Schwartz-Zimmermann, H.E., 2023. Creep feeding and weaning influence the postnatal evolution of the plasma metabolome in neonatal piglets. Metabolites 13, 214.

Mote, B.E., Rothschild, M.F., 2020. Modern genetic and genomic improvement of the pig. In: Bazer, W., Cliff, L.G., Wu, G. (Eds.), Animal Agriculture. Academic Press, New York, NY, USA, pp. 249–262.

Muñoz, M., Bozzi, R., García, F., Núñez, Y., Geraci, C., Crovetti, A., García-Casco, J., Alves, E., Škrlep, M., Charneca, R., Martins, J.M., Quintanilla, R., Tibau, J., Kušec, G., Djurkin-Kušec, I., Mercat, M.J., Riquet, J., Estellé, J., Zimmer, C., Razmaite, V., Araujo, J.P., Radović, Č., Savić, R., Karolyi, D., Gallo, M., Čandek-Potokar, M., Fontanesi, L., Fernández, A.I., Óvilo, C., 2018. Diversity across major and candidate genes in European local pig breeds. PLoS One 13, e0207475.

Muñoz, R., Tor, M., Estany, J., 2012. Relationship between blood lipid indicators and fat content and composition in Duroc pigs. Livestock Science 148, 95–102.

Olson Hunt, M.J., Weissfeld, L., Boudreau, R.M., Aizenstein, H., Newman, A.B., Simonsick, E.M., Van Domelen, D.R., Thomas, F., Yaffe, K., Rosano, C., 2014. A variant of sparse partial least squares for variable selection and data exploration. Frontiers in Neuroinformatics 8, 18.

Pang, Z., Zhou, G., Ewald, J., Chang, L., Hacariz, O., Basu, N., Xia, J., 2022. Using MetaboAnalyst 5.0 for LC–HRMS spectra processing, multi-omics integration and covariate adjustment of global metabolomics data. Nature Protocols 17, 1735–1761.

Pérez-Enciso, M., Steibel, J.P., 2021. Phenomes: the current frontier in animal breeding. Genetics Selection Evolution 53, 22.

Peukert, M., Zimmermann, S., Egert, B., Weinert, C.H., Schwarzmann, T., Brüggemann, D.A., 2021. Sexual dimorphism of metabolite profiles in pigs depends on the genetic background. Metabolites 11, 261.

R Core Team (2022). R: A language and environment for statistical computing. R foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Rauw, W.M., Rydhmer, L., Kyriazakis, I., Øverland, M., Gilbert, H., Dekkers, J.C., Hermesch, S., Bouquet, A., Gómez Izquierdo, E., Louveau, I., Gomez-Raya, L., 2020. Prospects for sustainability of pig production in relation to climate change and novel feed resources. Journal of the Science of Food and Agriculture 100, 3575–3586.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C., Müller, M., 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12, 77.

Ruiz-Perez, D., Guan, H., Madhivanan, P., Mathee, K., Narasimhan, G., 2020. So you think you can PLS-DA? BMC Bioinformatics 21, 2.

Savitz, J., 2020. The kynurenine pathway: a finger in every pie. Molecular Psychiatry 25, 131–147.

Schiavo, G., Bertolini, F., Bovo, S., Galimberti, G., Muñoz, M., Bozzi, R., Čandek-Potokar, M., Óvilo, C., 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, 193–205.

Schlosser, P., Li, Y., Sekula, P., Raffler, J., Grundner-Culemann, F., Pietzner, M., Cheng, Y., Wuttke, M., Steinbrenner, I., Schultheiss, U.T., Kotsis, F., Kacprowski, T., Forer, L., Hausknecht, B., Ekici, A.B., Nauck, M., Völker, U., Walz, G., Oefner, P.J., Kronenberg, F., Mohney, R.P., Köttgen, M., Suhre, K., Eckardt, K.U., Kastenmüller, G., Köttgen, A., 2020. Genetic studies of urinary metabolites illuminate mechanisms of detoxification and excretion in humans. Nature Genetics 52, 167–176.

Sellier, P., 1976. The basis of crossbreeding in pigs; a review. Livestock Production Science 3, 203–226.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research 13, 2498–2504.

Straadt, I.K., Aaslyng, M.D., Bertram, H.C., 2014. An NMR-based metabolomics study of pork from different crossbreeds and relation to sensory perception. Meat Science 96, 719–728.

Suzuki, K., Shibata, T., Kadowaki, H., Abe, H., Toyoshima, T., 2003. Meat quality comparison of Berkshire, Duroc and crossbred pigs sired by Berkshire and Duroc. Meat Science 64, 35–42.

Tor, M., Vilaró, F., Ros-Freixedes, R., Álvarez-Rodríguez, J., Bosch, L., Gol, S., Pena, R. N., Reixach, J., Estany, J., 2021. Circulating non-esterified fatty acids as biomarkers for fat content and composition in pigs. Animals 11, 386.

Trainor, P., DeFilippis, A., Rai, S., 2017. Evaluation of classifier performance for multiclass phenotype discrimination in untargeted metabolomics. Metabolites 7, 30.

Van Buuren, S., Groothuis-Oudshoorn, K., 2011. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software 45, 1–67.

Veiga-da-Cunha, M., Tyteca, D., Stroobant, V., Courtoy, P.J., Opperdoes, F.R., Van Schaftingen, E., 2010. Molecular identification of NAT8 as the enzyme that acetylates cysteine S-conjugates to mercapturic acids. Journal of Biological Chemistry 285, 18888–18898.

Vu, T., Siemek, P., Bhinderwala, F., Xu, Y., Powers, R., 2019. Evaluation of multivariate classification models for analyzing NMR metabolomics data. Journal of Proteome Research 18, 3282–3294.

Wang, H., Xia, P., Lu, Z., Su, Y., Zhu, W., 2021. Metabolome-Microbiome responses of growing pigs induced by time-restricted feeding. Frontiers in Veterinary Science 8, 681202.

Wenck, S., Creydt, M., Hansen, J., Gärber, F., Fischer, M., Seifert, S., 2021. Opening the random forest black box of the metabolome by the application of surrogate minimal depth. Metabolites 12, 5.

Wilkinson, S., Lu, Z.H., Megens, H.-J., Archibald, A.L., Haley, C., Jackson, I.J., Groenen, M.A.M., Crooijmans, R.P.M.A., Ogden, R., Wiener, P., 2013. Signatures of diversifying selection in European pig breeds. PLoS Genetics 9, e1003453.

Xie, Z., Gan, M., Du, J., Du, G., Luo, Y., Liu, B., Zhu, K., Cheng, W., Chen, L., Zhao, Y., Niu, L., Wang, Y., Wang, J., Zhu, L., Shen, L., 2023. Comparison of growth performance and plasma metabolomics between two sire-breeds of pigs in China. Genes 14, 1706.

Yi, L., Dong, N., Yun, Y., Deng, B., Ren, D., Liu, S., Liang, Y., 2016. Chemometric methods in data processing of mass spectrometry-based metabolomics: a review. Analytica Chimica Acta 914, 17–34.

Yin, X., Chan, L.S., Bose, D., Jackson, A.U., VandeHaar, P., Locke, A.E., Fuchsberger, C., Stringham, H.M., Welch, R., Yu, K., Fernandes Silva, L., Service, S.K., Zhang, D., Hector, E.C., Young, E., Ganel, L., Das, I., Abel, H., Erdos, M.R., Bonnycastle, L.L., Kuusisto, J., Stitziel, N.O., Hall, I.M., Wagner, G.R., FinnGen, Kang, J., Morrison, J., Burant, C.F., Collins, F.S., Ripatti, S., Palotie, A., Freimer, N.B., Mohlke, K.L., Scott, L.J., Wen, X., Fauman, E.B., Laakso, M., Boehnke, M., 2022. Genome-wide association studies of metabolites in Finnish men identify disease-relevant loci. Nature Communications 13, 1644.

Zhang, B., Hu, S., Baskin, E., Patt, A., Siddiqui, J., Mathé, E., 2018. RaMP: a comprehensive relational database of metabolomics pathways for pathway enrichment analysis of genes and metabolites. Metabolites 8, 16.

Zhang, Y., Liang, H., Liu, Y., Zhao, M., Xu, Q., Liu, Z., Weng, X., 2021. Metabolomic Analysis and identification of sperm freezability-related metabolites in boar seminal plasma. Animals 11, 1939.

Zhang, S., Zhang, K., Peng, X., Zhan, H., Lu, J., Xie, S., Zhao, S., Li, X., Ma, Y., 2020. Selective sweep analysis reveals extensive parallel selection traits between large white and Duroc pigs. Evolutionary Applications 13, 2807–2820.