Animal Microbiome

RESEARCH

Open Access

Characterization of microbiota signatures in Iberian pig strains using machine learning algorithms

Lamiae Azouggagh[1], Noelia Ibáñez-Escriche[1*], Marina Martínez-Álvaro[1], Luis Varona[2], Joaquim Casellas[3], Sara Negro[4] and Cristina Casto-Rebollo[1]

Abstract

Background There is a growing interest in uncovering the factors that shape microbiome composition due to its association with complex phenotypic traits in livestock. Host genetic variation is increasingly recognized as a major factor influencing the microbiome. The Iberian pig breed, known for its high-quality meat products, includes various strains with recognized genetic and phenotypic variability. However, despite the microbiome's known impact on pigs' productive phenotypes such as meat quality traits, comparative analyses of gut microbial composition across Iberian pig strains are lacking. This study aims to explore the gut microbiota of two Iberian pig strains, Entrepelado ($n=74$) and Retinto ($n=63$), and their reciprocal crosses ($n=100$), using machine learning (ML) models to identify key microbial taxa relevant for distinguishing their genetic backgrounds, which holds potential application in the pig industry. Nine ML algorithms, including tree-based, kernel-based, probabilistic, and linear algorithms, were used.

Results Beta diversity analysis on 16 S rRNA microbiome data revealed compositional divergence among genetic, age and batch groups. ML models exploring maternal, paternal and heterosis effects showed varying levels of classification performance, with the paternal effect scenario being the best, achieving a mean Area Under the ROC curve (AUROC) of 0.74 using the Catboost (CB) algorithm. However, the most genetically distant animals, the purebreds, were more easily discriminated using the ML models. The classification of the two Iberian strains reached the highest mean AUROC of 0.83 using Support Vector Machine (SVM) model. The most relevant genera in this classification performance were *Acetitomaculum*, *Butyricicoccus* and *Limosilactobacillus*. All of which exhibited a relevant differential abundance between purebred animals using a Bayesian linear model.

Conclusions The study confirms variations in gut microbiota among Iberian pig strains and their crosses, influenced by genetic and non-genetic factors. ML models, particularly CB and RF, as well as SVM in certain scenarios, combined with a feature selection process, effectively classified genetic groups based on microbiome data and identified key microbial taxa. These taxa were linked to short-chain fatty acids production and lipid metabolism, suggesting microbial composition differences may contribute to variations in fat-related traits among Iberian genetic groups.

Keywords Microbiome, Iberian pig, 16S rRNA, Machine learning, Classification, Differential abundance

*Correspondence:
Noelia Ibáñez-Escriche
noeibes@dca.upv.es
[1]Institute for Animal Science and Technology, Universitat Politècnica de Valencia, Valencia 46022, Spain
[2]Instituto Agroalimentario de Aragón (IA2), Universidad de Zaragoza, Zaragoza 50013, Spain
[3]Departament de Ciència Animal i dels Aliments, Universitat Autònoma de Barcelona, Bellaterra, Barcelona 08193, Spain
[4]Inga Food, Almendralejo 06200, Spain

Background

In recent years, various studies have shown that the gut microbiome can explain a substantial part of the phenotypic variability in important traits in livestock [1, 2]. The microbiome plays a crucial role in modulating methane emissions in cattle [3], intramuscular fat and resilience in rabbits [4, 5] and growth and carcass composition pigs [6, 7]. Additionally, growing evidence suggests that host genetics can influence gut microbial composition and diversity [8]. Larzul et al. [9] showed a direct response after divergent selection for the abundance of bacterial genera related to feed efficiency in pigs, and a correlated response in this trait. Hence, studying the gut microbiota could help disclose its impact on the variation of key traits in livestock.

An important pig breed that has recently gained considerable attention in both national and international markets is the Iberian pig breed [10]. The increasing demand for its high-quality products is primarily due to its high intramuscular fat deposition and distinctive fatty acid profile, particularly its high oleic acid content. Maltecca et al. [11] found an association between gut microbiota and meat quality traits including fat deposition in pigs. Hence, investigating the gut microbiota differences in Iberian pig strains with different meat quality could help identifying microbial biomarkers with potential application in the Iberian pig industry. Officially, five Iberian strains are recognized: Entrepelado, Retinto, Torbiscal, Lampiño, and Manchado de Jabugo, all of which exhibit substantial genetic diversity [12, 13]. Genomic studies have revealed distinct genomic backgrounds within the Entrepelado, Retinto, and Torbiscal strains, and their reciprocal crosses [14]. This diversity extends to productive and reproductive performance, such as fat deposition and prolificacy, as well as gene expression profiles. In fact, research has highlighted the superior meat quality and prolificacy of the Retinto over the Entrepelado [15, 16]. However, the Entrepelado has demonstrated important maternal effects over the Retinto, positively impacting the offspring growth [16]. Crosses between these strains have shown notable heterosis effects on meat quality and litter size [15–17]. Moreover, Garrido et al. [18] and Villaplana-Velasco et al. [19] found that Retinto animals exhibit higher and healthier fat accumulation and greater expression of key lipogenic genes compared to other strains, such as Torbiscal and Lampiño. Despite their differences in productive traits and the demonstrated influence of the pig microbiome on these traits, the gut microbial composition of different Iberian pig strains has never been compared.

The pig microbiome is not only influenced by host genetics, but also by maternal factors [20], housing environment [21], diet [22], and age [23]. Deciphering the influence of these factors on the microbiome is challenging due to its heterogeneity among individuals. Additionally, microbiome-derived data share complex relationships within each other and with traits of interest and are sparse and compositional [24]. Given these peculiarities, it is crucial to use appropriate models that can effectively extract and utilize all the information from these biological datasets. One effective approach to analyzing these data is the use of Machine Learning (ML) algorithms. These models have shown their ability to capture complex patterns within the microbiota, that traditional analytical methods, such as linear regression and principal component analysis, might overlook [25]. However, the choice of the most effective model depends on the specific use case, so testing multiple ML models is advisable to identify the most suitable one [26].

This study aimed to explore the gut microbiota of Iberian pigs belonging to two different strains (Entrepelado and Retinto) and their reciprocal crosses, and identify the key taxa relevant for distinguishing the genetic background of these Iberian pigs. For this purpose, we evaluated the classification performance of nine widely used ML models, using microbiota abundances derived from 16 S rRNA sequencing as predictors. We also identified the most relevant set of predictive taxa using a feature selection approach and compared them with differentially abundant taxa among different genetic groups. Furthermore, we provided a biological interpretation of the most influential taxa in differentiating the genetic groups, offering insights into their potential impact on host phenotypes.

Methods

Animals and samples

The animals used in this study belonged to two Iberian purebred pig strains (RR; Retinto and EE; Entrepelado) and their reciprocal crosses (ER and RE), where the first letter indicates the sire line and the second the maternal line. These animals originated from the Iberian Testing Center (Almendralejo-Extremadura, Spain) of the company INGA FOOD S.A (Tres Cantos-Madrid, Spain). The pigs were randomly housed in groups of 80, avoiding full sibs, and fattened *ad libitum* by automatic feeders with commercial feedstuffs. In total, 237 castrated males were used, of which 74 pigs belonged to the EE strain, 63 to the RR, 49 to the RE and 51 to the ER. Feces samples were collected at the CTI facilities between October 2021 and November 2022, prior to the animals' transport to the slaughterhouse. The pigs weighted on average 161.6 ± 13.6 kg at the end of the fattening period and were on average 365 ± 35 days of age. The feces samples were homogenized in 50-mL Falcon tubes and aliquoted in 2-mL cryotubes for their immediate freeze in liquid nitrogen. They were then stored at -80 °C at the CTI facility lab until processed.

DNA extraction and 16 S rRNA gene amplicon sequencing

Bacterial DNA for amplicon sequencing was isolated using the HigherPurity™ Soil DNA Isolation Kit (Canvax Reagents SL, Valladolid, Spain), following manufacturer's instructions. DNA concentration and purity were estimated by measuring the 260/280 ratio with a Nanodrop ND-1000 and verifying by a Qubit™ 4 Fluorometer (Invitrogen, Thermo Fisher Scientific, Carlsbad, CA, USA). The hypervariable V3-V4 region of the 16 S ribosomal RNA gene was amplified to identify the microbial community, using previously described primers [27]: forward (5′-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG-3′) and reverse (5′-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC-3′). PCR amplification was carried out in a 25-µl volume per sample, employing a primer concentration of 0.08 µM and NEBNext Q5 Hot Start HiFi PCR Master Mix (New England Biolabs, Ipswich, MA, USA). PCR cycling conditions were as follow: 98 °C for 30 s; 5 cycles at 98 °C for 10 s, 55 °C for 5 min, and 65 °C for 45 s. After this first PCR step, a second PCR was performed to attach Illumina (Illumina, Inc., San Diego, CA, USA) barcodes and full-length Nextera adapters. This second PCR was performed in a total volume of 50-µl, incorporating NEBNext Q5 Hot Start HiFi PCR Master Mix (New England Biolabs) and Nextera XT v2 adaptor primers from Illumina. The cycling conditions for the second PCR were as follows: 98 °C for 30 s, 17 cycles of 98 °C for 10 s, 55 °C for 30 s, and 65 °C for 45 s, and 65 ºC for 5 min. The PCR products were purified using AgenCourt AMPure XP beads (Beckman Coulter, Inc., Brea, CA, USA) to clean up the final library, according to manufacturer's instructions. The amplicons' quality was evaluated using a Fragment Analyzer (Agilent Biosystems, Agilent Technologies, Inc., Santa Clara, CA, USA) and quantified through qPCR using the Kapa library quantification kit designed for Illumina Platforms (Kapa Biosystems) on an ABI 7900HT real-time cycler (Applied Biosystems, Thermo Fisher Scientific, Carlsbad, CA, USA). Amplicons were normalized and pooled in an equimolar concentration of 15 pM and prepared for sequencing on the Illumina MiSeq platform with paired-end reads of 300 bp. A total of 237 samples were randomly distributed across eight plates, including the negative and positive controls, under the same conditions and reagents. Positive controls were set up using the ZymoBIOMICS™ Microbial Community DNA Standard (ref. D6306, Zymo).

Bioinformatic pipeline

Quality control of raw reads was performed using FastQC v0.11.9 [28] and MultiQC v1.18 [29] tools. One sample was discarded due to an abnormal total number of reads. Sequences were processed using nf-core/ampliseq pipeline v2.11.0 [30], which used Cutadapt program [31] for primer trimming, removing the previously described forward and reverse primers. Quality filtering and truncation were handled by DADA2 [32]. Forward reads were trimmed at 280 bases and reverse ones at 220 bases to filter out low-quality base calls. Moreover, reads shorter than 100 bp were discarded. Amplicon sequence variants (ASVs) were inferred for each sample using pooled mode to improve sensitivity in detecting low-abundance variants. ASVs with fewer than 50 reads were filtered to retain only relevant data. Taxonomic classification was conducted with DADA2 using SILVA reference database v.138 [33]. ASVs assigned to Archaea, Eukaryota, Mitochondria and Chloroplasts were excluded, focusing further analyses on the ASVs affiliated with the Bacteria domain. The remaining ASVs were collapsed to the genus level, and those lacking annotation at that taxonomic level were removed from the dataset. The bacterial community from positive controls was adequately characterized, and no contaminants were identified in the negative controls. Principal component analysis (PCA) was performed on the genus-level abundance table to detect samples outliers. Further analysis used a total of 235 samples and 121 genera.

Alpha- and beta-diversity

Alpha- and beta-diversity were computed on different normalized datasets. Alpha-diversity metrics: Shannon's index [34], Pielou evenness [35], and Chao1 estimator [36] were computed on rarefied sequences at the genus level, considering a maximum sampling depth of 6904, that is the minimum read depth registered in the samples after the quality control of sequencing data. Metrics were calculated using the vegan and fossil package in R. We used the Wilcoxon rank-sum test with *p*-values adjusted using the Benjamini-Hochberg (BH) false discovery rate (FDR) correction to compare alpha-diversity measures among different groups. These comparison groups included the four genetic groups (the strains and their reciprocal crosses: EE, ER, RE, RR), age groups (stratified into four quartiles with ranges of 298–344, 345–362, 363–383, and 384–494 days), and four different batches (9, 10, 11, 12). Significant differences in alpha-diversity measures were defined as those with FDR lower than 0.05. Beta-diversity assessment and posterior analysis were performed on a dataset comprising genera present in a minimum of 75% of samples for each genetic group (EE, RR, ER, RE). A total of 121 genera remained in the dataset. Beta-diversity was computed by calculating the Aitchison dissimilarity distance matrix, corresponding to Euclidean distances applied to centered log-ratio (CLR) transformed abundances, to account for their compositional nature [37]. The CLR transformation was

performed with the "compositions" v.2.0–6 R package [38]. The CLR was defined by Aitchison [39] as:

$$clr\left(x_i\right) = \left[ln\frac{x_{11}}{g_m\left(x_i\right)}, ln\frac{x_{12}}{g_m\left(x_i\right)}, \ldots, ln\frac{x_{1k}}{g_m\left(x_i\right)}\right] \quad (1)$$

,

where $g_m\left(x_i\right)$ is the geometric mean of the composition $x_i = [x_{11}, x_{12}, \ldots x_{1k}]$ representing the sample abundance for each genus *k* on the individual *i*. To avoid undefined logarithm, zeros were imputed by the Bayesian-multiplicative replacement approach [40] using the "zCompositions" v.1.4.1 R package [41]. Differences in microbiome beta-diversity were assessed for strains, age and animal batches using a Permutational Multivariable Analysis of Variance (PERMANOVA; p-value ≤ 0.05) after 999 permutations, implemented with the Vegan v.2.6-4 R package [42], with a prior check for homogeneity.

Machine learning analysis

In this study, nine supervised machine learning (ML) algorithms were used to build predictive models to classify the Iberian animals based on their microbiota. For that, the 121 CLR-transformed genera abundances were used after a pre-correction for confounding effects of age (as a continuous variables) and animal batches (4 batches). The ML algorithms employed included tree-based models with various ensemble learning techniques: Decision Tree (DT) [43], Random Forest (RF) [43], XGBoost (XGB) [44], AdaBoost (AB) [45], and CatBoost (CB) [46]. Kernel-based approaches were represented by Support Vector Machine (SVM) [47]. Additionally, probabilistic algorithms including Gaussian Naive Bayes (GNB) [48] and Logistic Regression (LR) [49] were used along with the Partial Least Squares Discriminant Analysis (PLS-DA) [50]. All the algorithms were executed in Python programming language (Python v.3.11.5), using the Scikit-learn module v.1.3.2 [51].

For each algorithm, we explored five different classification scenarios:

1. Four genetic groups scenario: The genetic groups consisted of the two purebred strains and their reciprocal crosses (EE, RR, ER, RE), each treated as a separate class.
2. Purebred scenario: Only purebred individuals were evaluated (EE and RR).
3. Maternal scenario: Individuals were grouped by maternal line (EE/RE and RR/ER).
4. Paternal scenario: Individuals were grouped by paternal line (EE/ER and RR/RE).
5. Heterosis scenario: One class for crossed individuals (ER/RE) and another for purebred individuals (RR/EE).

In each of the five scenarios, the dataset was randomly stratified into training and test sets, with a split ratio of 75/25. The training set was used for hyperparameter tuning and feature selection via 5-fold cross-validation, using RandomizedSearchCV method from Scikit-learn module [51], to identify the best combination of hyperparameters through a search over defined parameter values. Hyperparameters and their range of values used for each ML algorithm are detailed in Additional file 1.

Performance evaluation

The entire dataset was randomly resampled 200 times, creating unique training and test sets each time to simulate a wide range of possible data scenarios. The previously optimized model was then tested against these 200 newly formed test sets, allowing for a comprehensive evaluation of the model's stability and generalization capabilities across varied data subsets. Performance of each of these resampled test sets was assessed using the Area Under the Receiver Operating Characteristic Curve (AUROC), as described by Bradley [52]. Mean and 95% confidence interval (CI) for the AUROC scores were computed by estimating the mean and the 2.5th and 97.5th percentiles of the resulting prediction distribution for the test sets in each scenario. A threshold of 0.60 was considered the minimum AUROC for an acceptable classifier [53].

Feature selection

Once the classification tasks were performed across the five scenarios, a feature selection (FS) process was employed to eliminate noisy variables and reduce dimensionality. For that, the genera were ranked according to their importance scores in every scenario using RF classifier, and these scores were averaged over 20 iterations. In a RF model, the feature importance sums to 1. Therefore, with no differences in contribution among features, the importance of each feature would be 1/121, given that 121 genera were used. Genera with importance scores exceeding this threshold were selected. Besides, we narrowed them down to only those cumulatively explaining 80% of the importance. After FS, the classification was repeated to evaluate the performance of the models with the selected set of genera.

Differential abundance analysis

Differential abundance analysis (DA) was performed on the same dataset used for ML analysis to identify the relevant differentially abundant genera across scenarios. For that, linear models were computed as follows:

$$y_{ikj} = \mu + S_{kj} + e_{ikj},$$

Where y_{ikj} is the CLR-transformed and corrected abundance of the sample *i* for the genus *k* under the scenario-related class *j*; μ is the mean; S_{kj} is the effect of the scenario-related class *j* on the genus *k*; which can be either the four genetic groups effect (4 levels), the maternal effect (2 levels), the paternal effect (2 levels) or the heterosis effect (2 levels); and e_{ikj} is the residual for sample *i* for the genus *k* and class *j*. The models were solved by MCMC with the "brms" R package [54] using four chains with a length of 50,000 iterations, a lag of 10, and a burn-in of 1000 iterations, assuming flat priors. The convergence of the posterior distribution for each model was evaluated by comparing the within-chain variance and between-chain variance known as $\hat{R}$, with $\hat{R}$ close to 1 indicating convergence [55]. Additionally, the convergence plots were visually checked for all cases. The marginal posterior distribution (MPD) of the pairwise differences between classes in each scenario was computed to estimate the posterior mean and the probability of the difference being either positive or negative (P0). The posterior mean of the differences was given in units of standard deviation (SD) of each genus. Genera with a minimum difference mean of 0.50 SD and a P0 higher than 95% were considered differentially abundant between the groups in each scenario.

Results

Microbiota composition of the Iberian pig genetic groups

A total of 16,820,609 raw reads corresponding to 12,158 ASVs were obtained after 16 S rRNA gene sequencing of fecal samples from 237 Iberian pigs, with an average of 70,973 ± 21,461 reads per sample. After taxa and abundance filtering, the number of ASVs dropped to 6986 with a total of 16,396,525 reads (97.47%) and an average of 69,183 ± 20,726 reads per sample. Taxonomic annotation identified a total of 16 phyla, 26 classes, 57 orders, 96 families and 262 genera in the 235 fecal samples, with only 7.1% of the reads not reaching a taxonomic annotation at genus level. The Firmicutes were the most abundant phylum in the fecal microbiota, with an average relative abundance (RA) of 93.9%, followed by Bacteroidota, which represented an average of 4.5% RA. The remaining 1.6% comprised fourteen other phyla which include Actinobacteriota, Spirochaetota, Proteobacteria, Verrucomicrobiota, Planctomycetota, Cyanobacteria, Desulfobacterota, Patescibacteria, Campylobacterota, Fibrobacterota, Elusimicrobiota, Synergistota, Deferribacterota and Fusobacteriota. The Streptococcaceae family, which belongs to the Firmicutes phylum, was the most abundant family in the EE, RE, and ER groups, with relative abundances of 25.8%, 25.5%, and 23.9%, respectively (Fig. 1). It was followed by Clostridiaceae in the EE group and Lactobacillaceae in the RE and ER groups. Meanwhile, the Lactobacillaceae family was the most prevalent in the RR group accounting for 25.1% RA, with the Streptococcaceae family closely behind (Fig. 1). At the genus level, *Streptococcus* was the most dominant genus across the four groups, accounting for an average of 25.7% of the total abundance. *Clostridium sensu stricto 1* was the second most abundant genus in the ER, RE, and EE groups, while *Lactobacillus* held this position in the RR group. A table showing the relative abundance data at the phylum, family, and genus levels for groups with different genetic backgrounds is provided [see Additional file 2].

Differences in the microbiota composition between genetic groups were tested by computing alpha- and beta-diversity at genus level. Alpha-diversity indexes (Chao1 index, Pielou evenness and Shannon index) were computed in a rarified dataset. No significant differences in alpha-diversity metrics were found among the different genetic groups [Fig. 2 and Additional file 3]. However, all alpha-diversity metrics were significantly different between most of age categories as well as animal batches (Fig. 2). On the other hand, beta-diversity was analyzed on the CLR-transformed dataset using Aitchison distances. The analysis indicated a statistically significant effect from genetic group, age, and batch (PERMANOVA; p-value = 0.001). After adjusting the CLR-transformed genera abundances for the confounding effects of age and batch, the results of beta-diversity analysis remained statistically significant for genetic groups (PERMANOVA; p-value = 0.023).

Classification results with machine learning

The analysis of microbial diversity revealed different microbiota compositions between the genetic groups. To test the potential for classifying these animals based on their microbiota, we evaluated the classification performance of nine machine learning (ML) models and computed AUROC scores [see Additional file 4]. In the first scenario, which involved differentiating the four genetic groups, the purebreds EE and RR, and their reciprocal crosses (ER and RE), the best performing ML model was Catboost classifier (CB) (Fig. 3), which achieved a classification performance of 0.64 (95% C.I. [0.63, 0.64]). A detailed analysis of the confusion matrices over 200 data resampling revealed that the purebred individuals (EE and RR) were generally well classified. In contrast, the crossed individuals (ER and RE) were often misclassified as purebred individuals [see Additional file 5]. Indeed, the Purebred scenario showed the best classification performance (Fig. 3) with an AUROC of 0.77 using the same classifier (CB), with a 95% C.I. of [0.76, 0.78]. The Maternal scenario, which aimed to classify individuals into EE-RE and RR-ER groups to explore microbial differentiation based on the maternal effects of these strains, yielded an AUROC of 0.66 (95% C.I. [0.65, 0.66]) achieved by XGB classifier (Fig. 3). Likewise, the Paternal

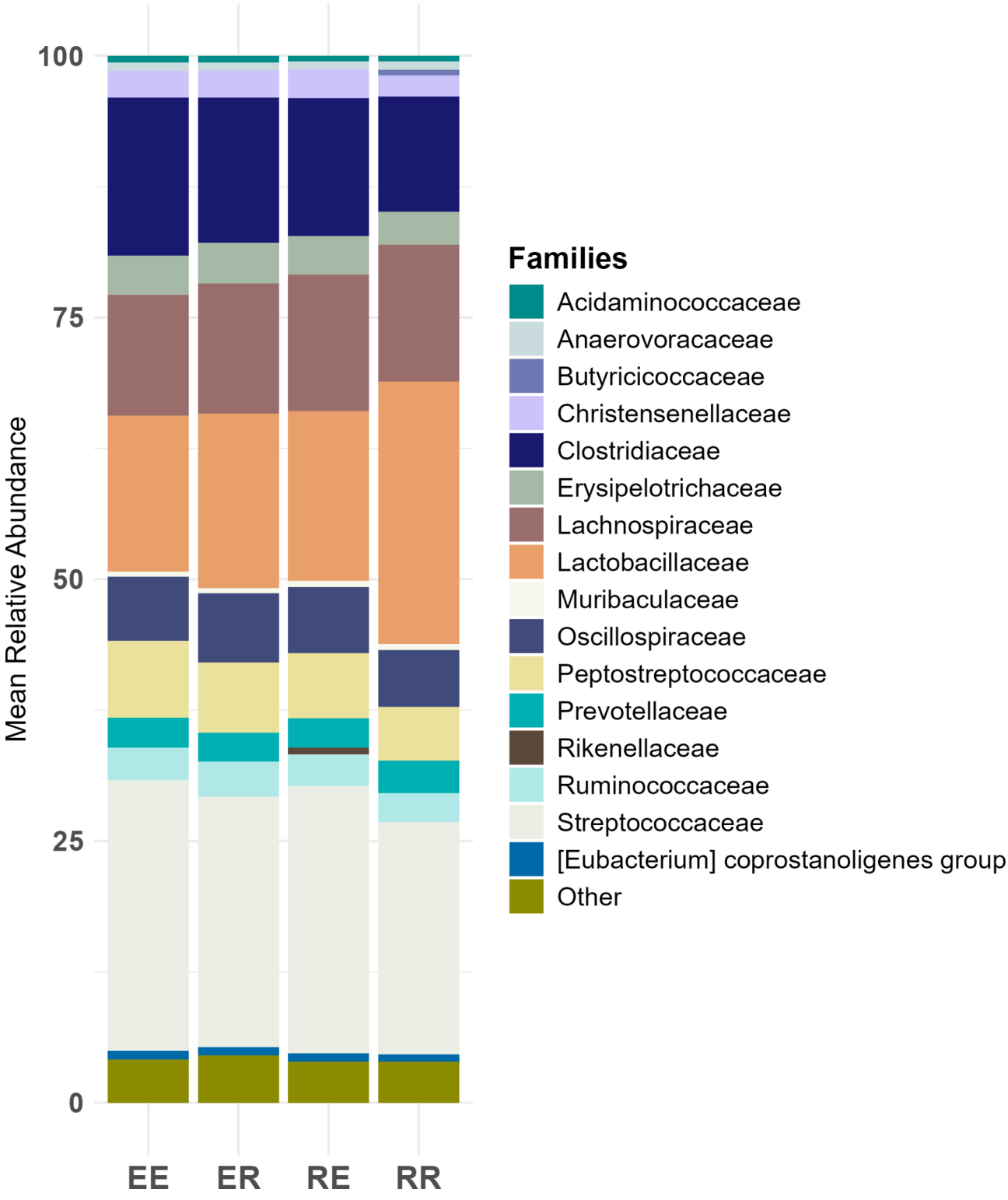

Fig. 1 Relative abundance of the fecal microbiota at the family level. Stacked bar plot of the mean relative abundance (RA) of fecal microbiota in the genetic groups composed of the two Iberian pig Entrepelado (EE) and Retinto (RR) and their reciprocal crosses (ER and RE) at the family level. The 'Other' group includes 80 families, each with an average RA of less than 0.5%

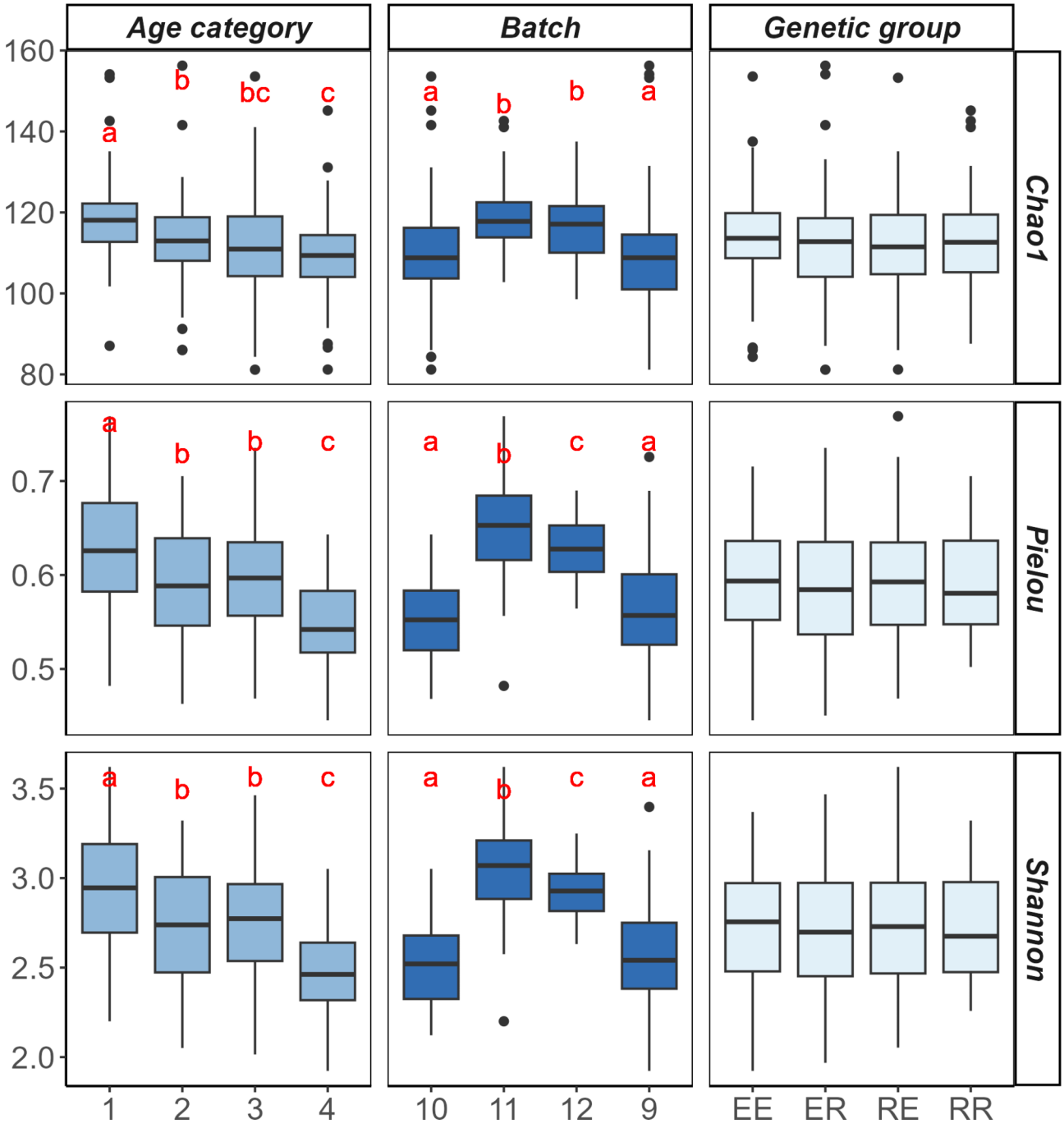

Fig. 2 Comparison of three alpha-diversity indices (Chao1, Pielou evenness and Shannon index) between age category groups, animal batches and genetic groups. Wilcoxon rank-sum test was used for comparisons between the four levels of each factor, with the correction for multiple test comparisons, FDR *p*-value < 0.05. The graph shows values of diversity indices on the Y-axis and factors on the X-axis, including Age category groups (1: 298–344 days; 2:345–362 days; 3:363–383 days and 4:384–494 days), Batch groups (9, 10, 11, 12) and genetic groups (Strains and their reciprocal crosses: EE, ER, RE, RR). No significant differences in alpha-diversity metrics were found between genetic groups. For age category and batch, groups with the same letter or sharing a letter, are not significantly different from each other

scenario aimed to distinguish between the EE-ER and RR-RE groups, to explore the paternal effect of these strains on the microbiota, and the Catboost classifier (LR) showed the best results with a mean AUROC of 0.68 (95% C.I. [0.67, 0.69]) [Additional file 4]. Finally, the classification performance under the heterosis scenario was evaluated, with a focus on distinguishing individuals in the EE-RR group from those in the ER-RE group. The algorithms struggled to differentiate crossed individuals from purebred ones, with the best mean AUROC being

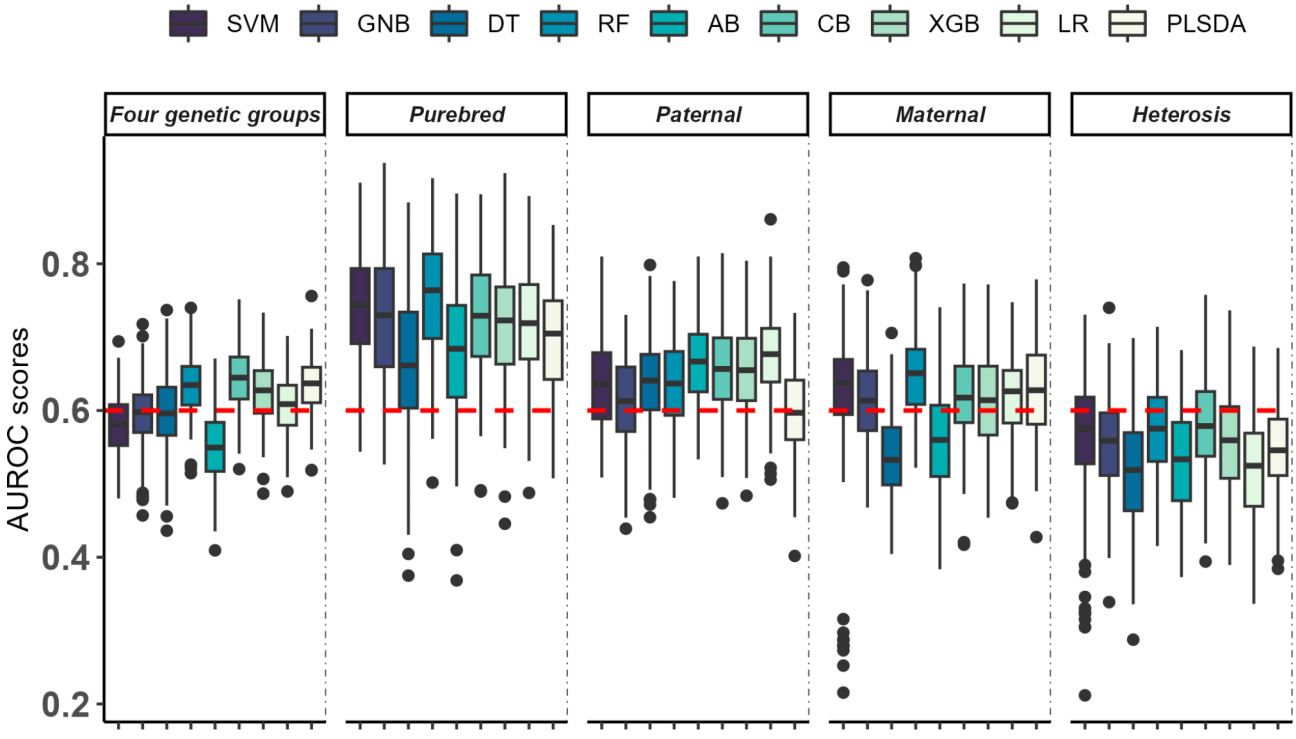

Fig. 3 Boxplots of the AUROC scores of machine learning (ML) models on the test sets across scenarios. ML models include Support Vector Machine (SVM), Gaussian Naïve Bayes (NB), decision tree (DT), random forest (RF), AdaBoost (AB), CatBoost (CB), XGBoost (XGB), Logistic Regression (LR) and Partial Least Squares-Discriminant analysis (PLS-DA). The colors in the boxplots are ordered from right to left as follows: SVM, NB, DT, RF, AB, CB, XGB, LR, and PLS-DA. The dashed red line indicates an AUROC of 0.60, representing the minimum threshold for an acceptable classifier

0.58 (95% CI: [0.57, 0.59]) using once again the Catboost classifier (CB) [Additional file 4].

Feature selection and subsequent classification results

Feature Selection (FS) using RF impurity-based importance scores [see Additional file 6] allowed a dimensionality reduction and focused on identifying the most relevant genera. The identified number of relevant genera ranged between 25 and 35, depending on the scenario. Figure 4 shows the number of relevant genera common to all five scenarios. Acetitomaculum, Escherichia shigella and Lachnospiraceae FCS020 group were the most consistent features, identified as important in all scenarios.

Performances of the nine ML models after FS resulted in moderate to substantial improvements in classification results [see Additional file 7]. The best AUROC scores and the best-performing models before and after FS are reported in Table 1, with the mean percentage increase in performance per scenario. Additionally, comparative results of the algorithms' performances, evaluated through a t-test before and after FS, are detailed in Additional file 8. The best AUROC score in the Purebred scenario improved from 0.77 to 0.83 using the SVM model, with an average performance increase of 6.5% across all models. In the Heterosis scenario, AUROC scores increased from 0.58 to 0.66 with an average performance increase of 11.6%. Per models, GNB registered the highest percentage increase across scenarios after performing FS, followed by AB, SVM, RF, CB, LR, XGB, PLSDA, and DT. This is despite RF being the model used for FS.

Differential abundance analysis

The genera with differential abundance (DA) between genetic groups for each scenario were identified by applying a Bayesian linear model (Table 2). A total of 20 genera demonstrated DA, each characterized by a minimum mean difference of 0.50 SD and a P0 value greater than 0.95. Of these, 16 genera were differentially abundant between purebreds, with *Acetitomaculum*, *Butyricicoccus and Romboutsia* displaying the most relevant DA between EE and RR. All the genera identified in the DA analysis were also selected in the FS process [see Additional file 6], with most achieving the highest importance scores in classification tasks across scenarios. For instance, *Acetitomaculum* displayed the highest importance score in the Four genetic groups, Purebred and Paternal scenarios, with higher abundance in the RR group/paternal group compared to EE one. The EE vs. ER comparison group exhibited the highest number of differentially abundant taxa (seven genera) after the purebred groups, followed by the EE vs. RE comparison group with four taxa exhibiting DA. However, no taxa

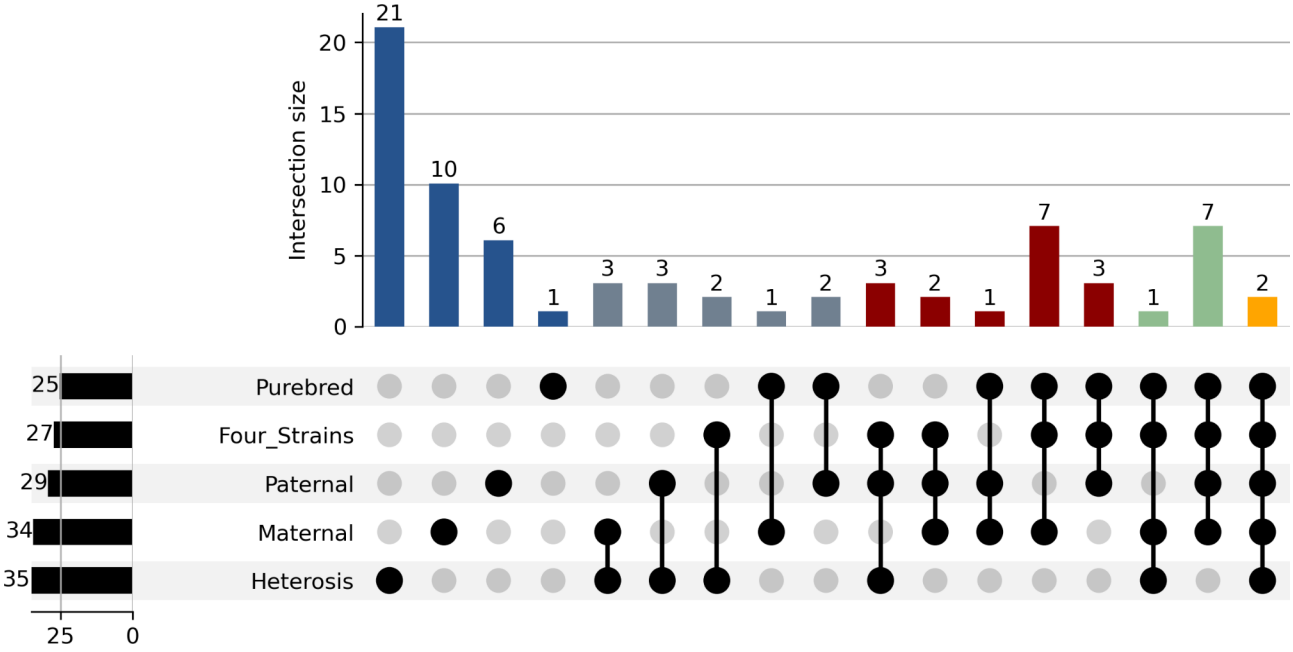

Fig. 4 Upset plot showing intersections among the selected genera of the five scenarios. The blue vertical bars indicate the number of genera selected in only one scenario. Gray and burgundy colors are for genera in two and three scenarios, respectively. The green color is for common genera present in four scenarios. Orange color is for common genera between all scenarios. Connected circles indicate an intersection of genera between scenarios. The circles represent unique genera. The number of genera selected in each one of the five sets is plotted in horizontal bars

Table 1 AUROC scores ± SD of the best model before and after applying feature selection, and the AUROC mean percentage increase per scenario, considering all models

Scenario	No FS		FS		Mean increase
	AUROC	Model	AUROC	Model	
Four genetic groups	0.64 ± 0.04	CB	0.68 ± 0.04	RF	6.9%
Purebred	0.77 ± 0.07	CB	0.83 ± 0.06	SVM	6.5%
Maternal	0.66 ± 0.06	XGB	0.70 ± 0.06	CB	4.3%
Paternal	0.68 ± 0.05	CB	0.74 ± 0.05	CB	8.7%
Heterosis	0.58 ± 0.06	CB	0.66 ± 0.06	CB	11.6%

Abbreviations; FS: Feature selection, CB: Catboost classifier, XGB: XGboost classifier, RF: Random Forest classifier, SVM: Support vector machine classifier

were found to be differentially abundant between maternal groups (Maternal scenario), or between purebred and crossbred animals (Table 2).

Discussion

Microbiome diversity

The microbial composition analysis indicated that the most abundant phylum present in the fecal samples of all groups is Firmicutes (93.9%), significantly ahead Bacteroidota (4.5%) and the rest of phyla (1.3%). Kim et al. [56] andXiao et al. [57] also reported that Firmicutes and Bacteroidota collectively constituted more than 90% of all bacteria present in the pig gut microbiota with a greater abundance in Firmicutes. These two phyla are the primary bacterial divisions of gut microbiota in mammals and are strongly associated with body fat in both humans and mice [58]. However, our results showed that Firmicutes are overwhelmingly dominant compared to Bacteroidota. Ban-Tokuda et al. [59] reported that the fecal level of Firmicutes significantly increased with fattening in pigs, while those of Bacteroidetes significantly decreased. This pattern aligns with the fatty nature of the Iberian pig breed and the fattening stage of the animals used in this study.

At the genus level, *Streptococcus* dominated the microbial composition of animals from all genetic backgrounds, consistent with findings by Heras-Molina et al. [60], who reported that this genus was the most abundant in fecal samples from 210-day-old purebred Iberian pigs. In a study including Torbiscal Iberian pigs, Crespo-Piazuelo et al. [61] found that *Streptococcus* was the most abundant genus in the ileum, however its abundance decreased significantly in the distal colon, a region likely more similar to the fecal microbial composition [62]. In another study also focusing on Torbiscal Iberian pigs using stool samples, López-García et al. [63] revealed

Table 2 Differentially abundant taxa between different genetic groups, with a minimum mean difference of 0.50 SD and a P0 value greater than 0.95. Genera are indicated based on the group in which they are more abundant

Comparison groups	Differentially abundant taxa
EE-RR	*Terrisporobacter (*), Clostridium sensu stricto 1 (*), Butyricicoccus (*), Romboutsia (*), Family XIII UCG 001 (*), Lachnospiraceae FCS020 group (*), UCG 002 (*), Eubacterium brachy group (*), Escherichia Shigella (*),* *Lactobacillus (-), Limosilactobacillus (-), Acetitomaculum (-), UCG 008 (-), Paludicola (-), Ruminococcus gnavus group (-), Eubacterium fissicatena group (*)*
EE-RE	*Colidextribacter (-), Acetitomaculum (-), Phascolarctobacterium (*), Escherichia-Shigella (*)*
EE-ER	*Colidextribacter (-), Acetitomaculum (-), UCG-008 (-), Paludicola (-), Eubacterium fissicatena group (-), Fusicatenibacter (*), Escherichia-Shigella (*)*
ER-RR	*Ruminococcus (*)*
ER-RE	-
RE-RR	-
EE maternal group vs. RR maternal group	-
EE paternal group vs. RR paternal group	*Acetitomaculum (-)*
Purebred vs. crossbred	-

(*): More abundant in the first group of the comparison

(-): More abundant in the second group of the comparison

that Prevotella 9 was the most abundant genus far surpassing *Streptococcus*, with Lactobacillus ranking second, an observation that aligns with our findings in the RR strain. However, it is worth noting that the animals in the studies by Crespo-Piazuelo et al. (2018) and López-García et al. (2021) were 120 and 117 days old, respectively, whereas the animals in our study averaged 365 days of age. Age is a major driver of microbial composition, as demonstrated in various longitudinal studies [64].

The alpha diversity indices did not show relevant differences within the four genetic groups. However, beta diversity analysis, using the Aitchison distance matrix, revealed relevant compositional divergence among the four genetic groups. This suggests that while the diversity of the gut microbial community of the Iberian pig was similar within the four genetic groups, the overall bacterial abundance profile was significantly different among them. On the other hand, both age category and animal batch were a great source of alpha and beta-diversity differences within and between animals respectively. This observation aligns with previous research on pig gut microbiome, highlighting age as a driving factor in microbiota variation [65], in addition to studies showing that pig microbiota may be susceptible to uncontrolled microenvironmental factors such as animal batch [66].

Classification performance

Given the reported phenotypic [15, 67], genetic [12, 14] and transcriptomic [18, 19] differences between different Iberian pig strains, especially in purebreds, our primary hypothesis was that these differences could be also evident at the microbiota level. This hypothesis was initially supported by the beta-diversity analysis and further confirmed by the satisfactory results of the different scenario classification tasks. We showed that the most genetically distant animals, the purebred animals [14], were more easily discriminated based on their microbiota, achieving the highest AUROC among all scenarios: 0.77 before FS and 0.83 after FS. The Four genetic groups scenario failed in the classification of the crossbreeds, while the purebred animals were properly classified, in line with the results obtained in the Purebred scenario. The confusion matrices indicated that crossbred animals were predominantly misclassified as purebred ones rather than being confused within each other. This finding was consistent with the Heterosis scenario (AUROC of 0.58 before FS and 0.66 after FS, 95% C.I. [0.65, 0.67]), where it was difficult to identify patterns that could group crossbreds together. This contrasts with the pattern found by Pena et al. [14], which suggested that crossbred animals (ER and RE) tended to cluster together using genotyping data. Crossbred animals were more likely to be mistaken for their paternal or maternal line. with a slight tendency towards the paternal line. This tendency is reinforced by results of the Paternal (AUROC of 0.74, 95% C.I. [0.72, 0.73]) and Maternal scenarios (AUROC of 0.70, 95% C.I. [0.69, 0.71]), where a slight increase in performance was obtained by the models, indicating that certain microbiota variability may be associated with these effects. These findings support the idea that microbiota may be transmitted to offspring. Camarinha-Silva et al. [1] reported microbiota heritabilities up to 0.57. However, the transmission of the microbiota, either partially or entirely, from one generation to the next is most likely facilitated by physical contact between newborns and their mothers [68]. Piglets come into contact with the dam's microbial communities during and after passing through the birth canal, as well as during nursing, suckling, and maternal care [69, 70]. However, few studies have explored the influence of the paternal microbiota on the phenotypic traits and microbiota of offspring. Animals used were not in physical contact with their sires, this suggests that any paternal effect is likely influenced by genetics. Srihi et al. [71] identified the presence of genomic imprinting, an important epigenetic phenomenon, on reproductive traits in a crossed population between EE and RR Iberian animals. This observation may explain the tendency of

animals to cluster slightly more according to their paternal line than their maternal line, despite the absence of direct contact between sire and offspring. However, further investigations are required to elucidate the influence of the paternal microbiota on their offspring.

Overall, CB, along with RF, proved to be the best-performing machine learning models, both before and after feature selection (FS), except in the Purebred scenario where the SVM model performed well specifically after FS. Various microbial classification studies have demonstrated that CB outperforms other ML methods employed in the present study [72, 73]. Additionally, Roguet et al. [74] emphasized the suitability and relevance of the RF classification approach for fecal source identification using 16 S rRNA gene amplicons. On the other hand, implementation of RF-based feature selection allowed the number of genera to be reduced by three quarters, achieving better performance compared to using the full dataset. With the exception of the maternal scenario, we found that the more effective the initial classification was, the smaller the performance gain observed after FS, since the model tends to identify the correct patterns and generalize properly without the need to reduce the dimensionality of the data. It should be noted that the performance results of each model were obtained using multiple data splits, as there was a risk of obtaining either random or biased estimates of model performance with only a single data split. This approach is particularly important given the limited size and heterogeneous nature of the dataset, as relying on a single random data split to assess performance can lead to misleading conclusions. A systematic study of ML benchmarks by Bouthillier et al. [75], shows that the use of multiple data splits, and therefore multiple test sets, improves the estimation of the general performance of ML algorithms.

Feature importance vs. differential abundance

The use of ML models in our study was not only sought to address the classification aspect of the microbiota data, but also to identify meaningful taxonomic signatures. The biological implications of the five genera with the highest importance scores of the most interesting classification scenarios were evaluated.

Classification tasks in the Purebred scenario were effective, achieving an AUROC of 0.83 after feature selection. This indicates the strength of evidence that genera consistently obtaining the highest importance scores across 20 iterations are unlikely due to random chance. The five more relevant genera of this scenario were *Acetitomaculum*, *Butyricicoccus*, *Limosilactobacillus*, *Erysipelotrichaceae UCG-003 and NK4A214 group*. Interestingly, the most important genus in this scenario -*Acetitomaculum*- was also the most important one in the Four genetic groups and the Paternal scenarios. *Erysipelotrichaceae UCG-003* was found among the top five in the maternal scenario. However, none of the top five important genera in the Purebred scenario appear among the top five in the Heterosis scenario, nor is there any overlap within the top ten. The observed discrepancy is likely attributable to the more accurate classification of purebred animals, which influences the overall performance in the Four Genetic Groups, Paternal, and Maternal scenarios, where purebreds are grouped together.

Acetitomaculum, is a genus part of the Lachnospiraceae family. Members of this family are recognized for their ability to produce short-chain fatty acids (SCFAs) through the fermentation of dietary polysaccharides [76]. SCFAs regulate lipid metabolism by increasing fatty acid oxidation and reducing lipid deposition [77]. Moreover, according to a study by Jiao et al. [78], a SCFA treatment increased the carcass weight and longissimus dorsi area of growing pigs, while also decreasing drip loss, a measure used to evaluate shelf life after slaughter. This indicates that SCFAs can improve carcass traits and meat quality in pigs. Previous research has shown that the RR pig strain outperforms the EE strain in several meat quality traits, including backfat thickness (BFT), total monounsaturated fatty acids (MUFAs), and oleic acid content [15]. *Acetitomaculum* was found to have a relevant DA between the EE and RR strains, with higher abundance in RR animals, which could be related to its superior meat quality.

The genus with the second-highest importance score in the Purebred scenario was *Butyricicoccus*, which showed a relevant higher abundance in EE animals compared to RR ones. It is a butyrate-producing bacterium, a SCFA shown to improve piglet growth performance [79] and to positively influence the gut health and maintenance of intestinal mucosa in pigs [80]. The *Limosilactobacillus* genus, the third in importance and with relevant higher abundance in RR animals compared to EE ones, is a genus of which various species were strongly suggested to have a probiotic potential and to enhance immunological functions [81]. *Erysipelotrichaceae UCG-003*, ranking fourth in importance, is also a significant butyrate producer [82]. Additionally, this genus was found to have a positive correlation with body weight in broilers [83]. Finally, the *NK4A214 group* genus, belonging to the *Oscillospiraceae* family, was identified by Sebastià et al. [84] as having a significant negative correlation with myristic acid content in the *Longissimus dorsi* muscle of a Duroc × Iberian crossed population, and a significant positive correlation with palmitoleic acid content in their backfat. Notably, myristic acid exhibited significant differences in backfat composition between the EE and RR maternal lines [15], while palmitoleic acid showed significant differences in the *Longissimus thoracis* muscle between these lines. Despite these associations, no

relevant differences in the abundance of the *NK4A214 group* genus were observed between the genetic groups, suggesting that its role in fatty acid metabolism may be influenced by other factors.

In the paternal scenario, where classifications achieved interesting results reaching an AUROC of 0.74, *Acetitomaculum and Butyricicoccus* were once again among the top five most important genera. The other three key genera were *Terrisporobacter*, *Family XIII UCG-001* and *Sphaerochaeta*. *Terrisporobacter* was negatively associated with the oleic acid and MUFA (monounsaturated fatty acid) content of the longissimus dorsi muscle in a crossbred population of commercial pigs [85]. Although this genus showed higher abundance in EE animals, which aligns with previous reports indicating higher oleic acid levels in the RR line compared to the EE line [15], these findings did not extend to the EE/RR paternal groups. *Family XIII UCG-001*, second in importance, was positively correlated with liver-related metabolic disturbances in mice [86] and had a relevant higher abundance in EE animals, meanwhile *Sphaerochaeta* genus showed a positive correlation with palmitoleic acid content in the *longissimus dorsi* muscle of a crossbred pig population [84]. While this fatty acid displayed significant differences between the two lines, the abundance of the genus itself did not vary between them.

On the other hands, this study showed a fair classification performance in the Maternal scenario, possibly reflecting the complexity of maternal influences-an effect that may require a larger sample size to elucidate. Maternal effects were observed between the EE and RR strain in some meat quality traits such as palmitoleic acid and backfat thickness [15]. The top five most important genera in this scenario were *Mogibacterium*, *Lachnospiraceae UCG-007*, *Colidextribacter*, *Erysipelotrichaceae UCG-003 and Eubacterium fissicatena group*. All these important genera have been linked to lipid metabolism traits in the literature. *Mogibacterium*, was positively correlated with the concentrations of SCFAs in the feces of piglets [87]. *Lachnospiraceae UCG-007* genus displayed a significant positive correlation with margaroleic acid, and *Colidextribacter* similarly correlated positively with stearic acid, in the Duroc × Iberian population [84]. Interestingly, while the EE maternal line exhibited a higher stearic acid content than the RR line [15], Colidextribacter was less abundant in EE animals compared to RR animals. This inconsistency suggests that additional factors or complex interactions may influence the relationship between *Colidextribacter* and stearic acid. *Erysipelotrichaceae UCG-003* and *Eubacterium fissicatena group* genus, fourth and fifth in importance in this scenario, are significant butyrate producing genera [82, 88]. However, further studies are necessary to properly know if the maternal effect is determinant in Iberian pigs.

Likewise, previous studies have demonstrated a heterosis effect from the cross between EE and RR strains on the IMF content of the Longissimus thoracis [15]. The genus with the highest importance score was *Colidextribacter*. This genus exhibited a relevant DA between the EE strain and the crossbred animals (ER and RE), being more abundant in the group of crossbred animals, and given that a significant positive correlation between *Colidextribacter* and oleic acid was reported [84], this could be due to the increased frequency of RR line, which demonstrated higher levels of this fatty acid compared to EE line [15]. The other important genera include *Lachnospiraceae ND3007 and Eubacterium fissicatena group*, both are SCFAs producing bacteria [88, 89], *Fibrobacter*, *a* genus involved in fiber digestion and energy metabolism [90], and finally *Eubacterium brachy group*, which was found to have a positive correlation with serum HDL cholesterol ("the good cholesterol") in mice [91], suggesting that this taxon may play a role in promoting healthier lipid metabolism. However, none of these genera showed relevant DA between purebred and crossed animals. Moreover, classification performance of the Heterosis scenario reached only barely acceptable levels. Hence, the genera with the highest importance scores, might have been identified by chance. Together these findings indicate that the establishment of a relationship between microbiota and the heterosis effect of these strains would be elusive, and more research is required to disentangle this complex effect in Iberian pigs.

Conclusions

Results of this study suggest that besides factors such as age and housing environment, the genetic background is an important factor influencing the microbiota profile of Iberian pigs. Machine learning algorithms, particularly CatBoost and Random Forest, demonstrated the highest classification performance in the analysis, with SVM achieving notable results in specific cases such as in Purebred and Paternal scenarios. The genera that exhibited DA between different genetic groups were also identified by the RF-based feature selection method as important features and were included as predictors in one or various scenarios. Furthermore, most of the important genera were linked to SCFA production and lipid metabolism, indicating that the differences in the microbial composition between Iberian genetic groups could be contributing to their differences in fat-related traits reported in previous works. Nevertheless, further investigation is needed to determine how these genera might correlate with specific traits that distinguish these genetic groups, particularly those related to meat quality traits.

Abbreviations

AB	AdaBoost
ASVs	Amplicon sequence variants
AUROC	Area Under the ROC curve
CB	CatBoost
CI	Confidence interval
CLR	Centered log-ratio
DA	Differential abundance
DT	Decision Tree
EE	Entrepelado strain
ER	Cross between Entrepelado sire and Retinto dam
FDR	False discovery rate
FS	Feature selection
GNB	Gaussian Naive Bayes
IMF	Intramuscular fat
LR	Logistic Regression
ML	Machine learning
PCA	Principal component analysis
PLS-DA	Partial Least Squares Discriminant Analysis
RA	Relative abundance
RE	Cross between Retinto sire and Entrepelado dam
RF	Random Forest
RR	Retinto strain
SCFA	Short-chain fatty acids
SD	Standard deviation
SVM	Support Vector Machine
XGB	XGBoost

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s42523-025-00378-z.

Supplementary Material 1: Hyperparameters tuned across ML models and their range of search using GridsearchCV

Supplementary Material 2: Relative abundance data (%) at the phylum, family, and genus levels for groups with different genetic backgrounds

Supplementary Material 3: FDR-Adjusted *p*-values from pairwise Wilcoxon tests for alpha-diversity indices (Chao1, Pielou evenness, Shannon index) across genetic groups, animal batches, and age categories

Supplementary Material 4: Mean AUROC scores across different algorithms and scenarios before feature selection

Supplementary Material 5: Average confusion matrix of the Four genetic groups scenario over 200 data splits, with the four components of the confusion matrix (TP, FN, FP and TN)

Supplementary Material 6: Random Forest based importance scores of all genera. Table S1: Genera importance scores in the Purebred scenario. Table S2: Genera importance scores in the Four genetic groups scenario. Table S3: Genera importance scores in the Maternal scenario. Table S4: Genera importance scores in the Paternal scenario. Table S5: Genera importance scores in the Heterosis scenario

Supplementary Material 7: Mean AUROC scores across different algorithms and scenarios after feature selection

Supplementary Material 8: *p*-values of t-test for ML algorithm performances of all scenarios before and after feature selection

Supplementary Material 9: Results of the Bayesian statistical analysis. File includes the name of the relevant genus, posterior mean of the differences among genetic groups (meanDiff), the probability of the difference being higher (if the difference is positive) or lower (if negative) than 0 (P0), the highest posterior interval density of 95% (HPD95), and the comparison group

Acknowledgements

LA gratefully acknowledges the receipt of the GRISOLIA scholarship [CIGRIS/2021/098] awarded by the Generalitat Valenciana.

Author contributions

L.A analyzed the data and wrote the manuscript. L.V and J.C contributed to the study design and the discussion of the results. C.C.R contributed to the data analysis and the discussion of the results and edited the manuscript. M.M.A contributed to the discussion of the results, and edited the manuscript. N.I.E conceived the study, contributed to the discussion of the results, and edited the manuscript. All the authors have read and approved the final manuscript.

Funding

This study was funded by the Spanish Ministry of Science and Innovation in the Project PID2020-114705RB-I00.

Data availability

The data used in this study will be made available on Zenodo upon acceptance of the manuscript, DOI: 10.5281/zenodo.14013900.

Declarations

Ethics approval and consent to participate

All the experimental procedures were approved by the Committee of Ethics and Animal Welfare of the Miguel Hernández University, according to Council Directives 98/58/EC and 2010/63/EU.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 22 August 2024 / Accepted: 20 January 2025
Published online: 03 February 2025

References

1. Camarinha-Silva A, Maushammer M, Wellmann R, Vital M, Preuss S, Bennewitz J. Host genome influence on gut microbial composition and microbial prediction of complex traits in pigs. Genetics. 2017;206:1637–44.
2. Déru V, Tiezzi F, Carillier-Jacquin C, Blanchet B, Cauquil L, Zemb O, et al. Gut microbiota and host genetics contribute to the phenotypic variation of digestive and feed efficiency traits in growing pigs fed a conventional and a high fiber diet. Genet Selection Evol. 2022;54:55.
3. Roehe R, Dewhurst RJ, Duthie C-A, Rooke JA, McKain N, Ross DW, et al. Bovine host genetic variation influences rumen microbial methane production with best selection criterion for low methane emitting and efficiently feed converting hosts based on metagenomic gene abundance. PLoS Genet. 2016;12:e1005846.
4. Martínez-Álvaro M, Zubiri-Gaitán A, Hernández P, Casto-Rebollo C, Ibáñez-Escriche N, Santacreu MA et al. Correlated responses to selection for intramuscular fat on the gut microbiome in rabbits. 2024. https://doi.org/10.3390/ani14142078
5. Casto-Rebollo C, Argente MJ, García ML, Pena RN, Blasco A, Ibáñez-Escriche N. Selection for environmental variance shifted the gut microbiome composition driving animal resilience. Microbiome. 2023;11:147.
6. Bergamaschi M, Maltecca C, Schillebeeckx C, McNulty NP, Schwab C, Shull C, et al. Heritability and genome-wide association of swine gut microbiome features with growth and fatness parameters. Sci Rep. 2020;10:10134.
7. Wu C, Lyu W, Hong Q, Zhang X, Yang H, Xiao Y. Gut microbiota influence lipid metabolism of skeletal muscle in pigs. Front Nutr. 2021;8.
8. Spor A, Koren O, Ley R. Unravelling the effects of the environment and host genotype on the gut microbiome. Nat Rev Microbiol. 2011;9:279–90.
9. Larzul C, Estellé J, Borey M, Blanc F, Lemonnier G, Billon Y, et al. Driving gut microbiota enterotypes through host genetics. Microbiome. 2024;12:116.
10. Horrillo A, Gaspar P, Muñoz Á, Escribano M, González E. Fattening Iberian pigs indoors vs. outdoors: production performance and market value. Animals. 2023;13:506.
11. Maltecca C, Dunn R, He Y, McNulty NP, Schillebeeckx C, Schwab C, et al. Microbial composition differs between production systems and is associated with growth performance and carcass quality in pigs. Anim Microbiome. 2021;3:57.

12. Clemente I, Membrillo A, Azor Ortiz PJ, Polvillo Polo O, Juárez M, Santos E et al. Intra-breed genetic diversity characterization of the Iberian pig. In: XIV Reunión Nacional de Mejora Genética Animal (2008). Sevilla; 2008.
13. Martınez AM, Delgado JV, Rodero A, Vega-Pla JL. Genetic structure of the Iberian pig breed using microsatellites. Anim Genet. 2000;31:295–301.
14. Pena RN, Noguera JL, García-Santana MJ, González E, Tejeda JF, Ros-Freixedes R, et al. Five genomic regions have a major impact on fat composition in Iberian pigs. Sci Rep. 2019;9:2031.
15. Ibáñez-Escriche N, Magallón E, Gonzalez E, Tejeda JF, Noguera JL. Genetic parameters and crossbreeding effects of fat deposition and fatty acid profiles in Iberian pig lines1. J Anim Sci. 2016;94:28–37.
16. Noguera JL, Ibáñez-Escriche N, Casellas J, Rosas JP, Varona L. Genetic parameters and direct, maternal and heterosis effects on litter size in a diallel cross among three commercial varieties of Iberian pig. Animal. 2019;13:2765–72.
17. Varona L, Noguera JL, Casellas J, de Hijas MM, Rosas JP, Ibáñez-Escriche N. A cross-specific multiplicative binomial recursive model for the analysis of perinatal mortality in a diallel cross among three varieties of Iberian pig. Sci Rep. 2020;10:21190.
18. Garrido N, Izquierdo M, Hernández-García FI, Núñez Y, García-Torres S, Benítez R, et al. Differences in muscle lipogenic gene expression, carcass traits and fat deposition among three Iberian pig strains finished in two different feeding systems. Animals. 2023;13:1138.
19. Villaplana-Velasco A, Noguera JL, Pena RN, Ballester M, Muñoz L, González E, et al. Comparative transcriptome profile between Iberian pig varieties provides new insights into their distinct fat deposition and fatty acids content. Animals. 2021;11:627.
20. Mao J, Zhang Y, Liu J, Wang H. Gut microbiota and growth performance of offspring are influenced by wet nurse in pigs using cross-fostering trial. J Sci Food Agric. 2023;103:865–76.
21. Trudeau MP, Mosher W, Tran H, de Rodas B, Karnezos TP, Urriola PE, et al. Experimental facility had a greater effect on growth performance, gut microbiome, and metabolome in weaned pigs than feeding diets containing subtherapeutic levels of antibiotics: a case study. PLoS ONE. 2023;18:e0285266.
22. Verschuren LMG, Calus MPL, Jansman AJM, Bergsma R, Knol EF, Gilbert H, et al. Fecal microbial composition associated with variation in feed efficiency in pigs depends on diet and sex1. J Anim Sci. 2018;96:1405–18.
23. Lim MY, Song E-J, Kang KS, Nam Y-D. Age-related compositional and functional changes in micro-pig gut microbiome. Geroscience. 2019;41:935–44.
24. Ghannam RB, Techtmann SM. Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. Comput Struct Biotechnol J. 2021;19:1092–107.
25. Abavisani M, Khoshrou A, Foroushan SK, Ebadpour N, Sahebkar A. Deciphering the gut microbiome: the revolution of artificial intelligence in microbiota analysis and intervention. Curr Res Biotechnol. 2024;7:100211.
26. Teixeira M, Silva F, Ferreira RM, Pereira T, Figueiredo C, Oliveira HP. A review of machine learning methods for cancer characterization from microbiome data. NPJ Precis Oncol. 2024;8:123.
27. Willis JR, González-Torres P, Pittis AA, Bejarano LA, Cozzuto L, Andreu-Somavilla N, et al. Citizen science charts two major stomatotypes in the oral microbiome of adolescents and reveals links with habits and drinking water composition. Microbiome. 2018;6:218.
28. Andrews S. FastQC: a quality control tool for high throughput sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc. 2010.
29. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32:3047–8.
30. Straub D, Blackwell N, Langarica-Fuentes A, Peltzer A, Nahnsen S, Kleindienst S. Interpretations of environmental microbial community studies are biased by the selected 16S rRNA (gene) amplicon sequencing pipeline. Front Microbiol. 2020;11.
31. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17:10.
32. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: high-resolution sample inference from illumina amplicon data. Nat Methods. 2016;13:581–3.
33. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2012;41:D590–6.
34. Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948;27:379–423.
35. Pielou EC. The measurement of diversity in different types of biological collections. J Theor Biol. 1966;13:131–44.
36. Chao A. Non-parametric estimation of the number of classes in a population. Scand J Stat. 1984;11:265–70.
37. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. Front Microbiol. 2017;8.
38. van den Boogaart KG, Tolosana-Delgado R. Compositions: a unified R package to analyze compositional data. Comput Geosci. 2008;34:320–38.
39. Aitchison J. The statistical analysis of compositional data. Dordrecht: Springer Netherlands; 1986.
40. Martín-Fernández J-A, Hron K, Templ M, Filzmoser P, Palarea-Albaladejo J. Bayesian-multiplicative treatment of count zeros in compositional data sets. Stat Modelling. 2015;15:134–58.
41. Palarea-Albaladejo J, Martín-Fernández JA. zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. Chemometr Intell Lab Syst. 2015;143:85–96.
42. Oksanen J, Simpson G, Blanchet F, Kindt R, Legendre P, Minchin P et al. vegan: Community Ecology Package. 2022.
43. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Routledge; 2017.
44. Chen T, Guestrin C. XGBoost. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2016. pp. 785–94.
45. CAO Y, MIAO Q-G, LIU J-C GAOL. Advance and prospects of AdaBoost algorithm. Acta Automatica Sinica. 2013;39:745–58.
46. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. 2017.
47. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97.
48. Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss. Mach Learn. 1997;29(2/3):103–30.
49. Sperandei S. Understanding logistic regression analysis. Biochem Med (Zagreb). 2014;24(1):12–8.
50. Barker M, Rayens W. Partial least squares for discrimination. J Chemom. 2003;17:166–73.
51. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O et al. Scikit-learn: machine learning in Python. 2012.
52. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. 1997;30:1145–59.
53. Yang S, Berdine G. The receiver operating characteristic (ROC) curve. Southwest Respiratory Crit Care Chronicles. 2017;5:34.
54. Bürkner P-C. Brms: an R package for Bayesian multilevel models using Stan. J Stat Softw. 2017;80.
55. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Stat Sci. 1992;7.
56. Kim HB, Borewicz K, White BA, Singer RS, Sreevatsan S, Tu ZJ, et al. Longitudinal investigation of the age-related bacterial diversity in the feces of commercial pigs. Vet Microbiol. 2011;153:124–33.
57. Xiao Y, Li K, Xiang Y, Zhou W, Gui G, Yang H. The fecal microbiota composition of boar Duroc, Yorkshire, Landrace and Hampshire pigs. Asian-Australas J Anim Sci. 2017;30:1456–63.
58. Ley RE, Bäckhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI. Obesity alters gut microbial ecology. Proceedings of the National Academy of Sciences. 2005;102:11070–5.
59. Ban-Tokuda T, Maekawa S, Miwa T, Ohkawara S, Matsui H. Changes in faecal bacteria during fattening in finishing swine. Anaerobe. 2017;47:188–93.
60. Heras-Molina A, Estellé J, Vázquez-Gómez M, López-García A, Pesantez-Pacheco J-L, Astiz S, et al. The impact of host genetics on porcine gut microbiota composition excluding maternal and postnatal environmental influences. PLoS ONE. 2024;19:e0315199.
61. Crespo-Piazuelo D, Estellé J, Revilla M, Criado-Mesas L, Ramayo-Caldas Y, Óvilo C, et al. Characterization of bacterial microbiota compositions along the intestinal tract in pigs and their interactions and functions. Sci Rep. 2018;8:12727.
62. Flynn KJ, Ruffin MT, Turgeon DK, Schloss PD. Spatial variation of the native colon microbiota in healthy adults. Cancer Prev Res. 2018;11:393–402.
63. López-García A, Benítez R, Núñez Y, Gómez-Izquierdo E, de Mercado E, García-Casco JM, et al. Influence of genetic background and dietary oleic acid on gut microbiota composition in Duroc and Iberian pigs. PLoS ONE. 2021;16:e0251804.
64. Gaire TN, Scott HM, Noyes NR, Ericsson AC, Tokach MD, Menegat MB, et al. Age influences the temporal dynamics of microbiome and antimicrobial resistance genes among fecal bacteria in a cohort of production pigs. Anim Microbiome. 2023;5:2.

65. De Rodas B, Youmans BP, Danzeisen JL, Tran H, Johnson TJ. Microbiome profiling of commercial pigs from farrow to finish. J Anim Sci. 2018;96:1778–94.
66. Le Sciellour M, Zemb O, Hochu I, Riquet J, Gilbert H, Giorgi M, et al. Effect of chronic and acute heat challenges on fecal microbiota composition, production, and thermoregulation traits in growing pigs1,2. J Anim Sci. 2019;97:3845–58.
67. Juárez M, Clemente I, Polvillo O, Molina A. Meat quality of tenderloin from Iberian pigs as affected by breed strain and crossbreeding. Meat Sci. 2009;81:573–9.
68. David I, Canario L, Combes S, Demars J. Intergenerational transmission of characters through genetics, epigenetics, microbiota, and learning in livestock. Front Genet. 2019;10.
69. Lim J-A, Cha J, Choi S, Kim J-H, Kim D. Early colonization of the intestinal microbiome of neonatal piglets is influenced by the maternal microbiome. Animals. 2023;13:3378.
70. Liu S, Zhang Z, Ma L. A review focusing on microbial vertical transmission during sow pregnancy. Vet Sci. 2023;10:123.
71. Srihi H, López-Carbonell D, Ibáñez-Escriche N, Casellas J, Hernández P, Negro S, et al. A Bayesian multivariate gametic model in a reciprocal cross with genomic information: an example with two Iberian varieties. Animals. 2023;13:1648.
72. Duyar C, Senica SO, Kalkan H. Detection of cardiovascular disease using explainable artificial intelligence and gut microbiota data. Intell Based Med. 2024;10:100180.
73. Jing Z, Zheng W, Jianwen S, Hong S, Xiaojian Y, Qiang W, et al. Gut microbes on the risk of advanced adenomas. BMC Microbiol. 2024;24:264.
74. Roguet A, Eren AM, Newton RJ, McLellan SL. Fecal source identification using random forest. Microbiome. 2018;6:185.
75. Bouthillier X, Delaunay P, Bronzi M, Trofimov A, Nichyporuk B, Szeto J et al. Accounting for variance in machine learning benchmarks. 2021.
76. Biddle A, Stewart L, Blanchard J, Leschine S. Untangling the genetic basis of fibrolytic specialization by lachnospiraceae and ruminococcaceae in diverse gut communities. Divers (Basel). 2013;5:627–40.
77. He J, Zhang P, Shen L, Niu L, Tan Y, Chen L, et al. Short-chain fatty acids and their association with signalling pathways in inflammation, glucose and lipid metabolism. Int J Mol Sci. 2020;21:6356.
78. Jiao A, Diao H, Yu B, He J, Yu J, Zheng P, et al. Infusion of short chain fatty acids in the ileum improves the carcass traits, meat quality and lipid metabolism of growing pigs. Anim Nutr. 2021;7:94–100.
79. Hanczakowska E, Niwińska B, Grela ER, Węglarzy K, Okoń K. Effect of dietary glutamine, glucose and/or sodium butyrate on piglet growth, intestinal environment, subsequent fattener performance, and meat quality. Czech J Anim Sci. 2014;59:460–70.
80. Levine UY, Looft T, Allen HK, Stanton TB. Butyrate-producing bacteria, including mucin degraders, from the swine intestinal tract. Appl Environ Microbiol. 2013;79:3879–81.
81. Zhang Q, Vasquez R, Yoo JM, Kim SH, Kang D-K, Kim IH. Dietary supplementation of limosilactobacillus mucosae LM1 enhances immune functions and modulates gut microbiota without affecting the growth performance of growing pigs. Front Vet Sci. 2022;9.
82. Liu S, Li E, Sun Z, Fu D, Duan G, Jiang M, et al. Altered gut microbiota and short chain fatty acids in Chinese children with autism spectrum disorder. Sci Rep. 2019;9:287.
83. Zhi T, Ma A, Liu X, Chen Z, Li S, Jia Y. Dietary supplementation of Brevibacillus laterosporus S62-9 improves broiler growth and immunity by regulating cecal microbiota and metabolites. Probiotics Antimicrob Proteins. 2024;16:949–63.
84. Sebastià C, Folch JM, Ballester M, Estellé J, Passols M, Muñoz M et al. Interrelation between gut microbiota, SCFA, and fatty acid composition in pigs. mSystems. 2024;9.
85. Niu J, Liu X, Xu J, Li F, Wang J, Zhang X et al. Effects of silage diet on meat quality through shaping gut microbiota in finishing pigs. Microbiol Spectr. 2023;11.
86. Song X, Zhong L, Lyu N, Liu F, Li B, Hao Y, et al. Inulin can alleviate metabolism disorders in *ob/ob* mice by partially restoring leptin-related pathways mediated by gut microbiota. Genomics Proteom Bioinf. 2019;17:64–75.
87. Zhu L, Liao R, Tu W, Lu Y, Cai X. Pyrodextrin enhances intestinal function through changing the intestinal microbiota composition and metabolism in early weaned piglets. Appl Microbiol Biotechnol. 2020;104:4141–54.
88. Gan Y, Liu J, Jin M, Zhang Y, Huang S, Ma Q, et al. The role of the gut-joint axis in the care of psoriatic arthritis: a two-sample bidirectional Mendelian randomization study. Dermatol Ther (Heidelb). 2024;14:713–28.
89. Nishiwaki H, Hamaguchi T, Ito M, Ishida T, Maeda T, Kashihara K et al. Short-chain fatty acid-producing gut microbiota is decreased in Parkinson's disease but not in rapid-eye-movement sleep behavior disorder. mSystems. 2020;5.
90. Zhao JB, Liu P, Huang CF, Liu L, Li EK, Zhang G, et al. Effect of wheat bran on apparent total tract digestibility, growth performance, fecal microbiota and their metabolites in growing pigs. Anim Feed Sci Technol. 2018;239:14–26.
91. Kang Y, Oba PM, Gaulke CA, Sánchez-Sánchez L, Swanson KS. Dietary inclusion of yellow mealworms (T. Molitor) and lesser mealworms (A. diaperinus) modifies intestinal microbiota populations of diet-induced obesity mice. J Nutr. 2023;153:3220–36.

Publisher's note