Use of microbiome data and Machine Learning to distinguish Iberian pig strains

L. Azouggagh[1]*, C. Casto-Rebollo[1], L. Varona[2], J. Casellas[3], S. Negro[4], M. Martínez-Álvaro[1] and N. Ibáñez-Escriche[1]

[1] Institute for Animal Science and Technology, Universitat Politècnica de Valencia, 46022 Valencia, Spain, [2] Instituto Agroalimentario de Aragón (IA2), Universidad de Zaragoza, 50013 Zaragoza, Spain, [3] Departament de Ciència Animal i dels Aliments, Universitat Autònoma de Barcelona, Bellaterra, 08193 Barcelona, Spain, [4] Inga Food, 06200, Almendralejo, Spain.

*lazolam@doctor.upv.es

Abstract

There is an increasing interest unraveling the factors that shape microbiome composition due to its association with complex phenotypic traits in livestock. Increasing evidence points to the influence of host genetic variation. This interest has coincided with parallel advances in the field of machine learning (ML), providing valuable insights into microbial communities. This study aimed to explore the gut microbiome of Iberian pigs (n=237) and identify the key taxa relevant for predicting the genetic class of two different strains (EE; Entrepelado and RR; Retinto) and their reciprocal crosses (ER, RE). For this purpose, we evaluated the performance of eight different ML classifiers, based on 16S rRNA sequencing data, and performed traditional differential abundance (DA) analysis between strains. Fecal samples were collected from 237 castrated males and sequenced with 16S Illumina MiSeq platform. The dataset used to perform the classification comprised 37 genera, from which were selected the best features in discriminating samples according to their genetic group. The ML algorithms included tree-based models, kernel-based models and probabilistic models. Classifiers were trained with a subset of the data (train) and their performances were assessed using the area under the ROC curve (AUC) on the remaining set (test). Five scenarios were explored (Genetic groups, Purebreds, Maternal and Paternal line, and Heterosis groups). Our results showed that the most genetically distant animals (EE vs. RR) were more easily discriminated using the trained ML models. The classification of Iberian pigs from EE and RR strains reached a mean AUC of 0.80. However, the crossed animals from ER and RE groups did not exhibit specific patterns and including them in different groups confused the classification process, leading to lower performance results. The best performing ML model was not the same in all the classification tasks, but overall, tree-based models outperformed the other algorithms. Additionally, the most important features partially aligned with the taxa that displayed DA between strains. This study provides a framework that combines ML tools with microbiome-derived data to extract taxa meaningful signatures, that could represent a useful resource to characterize the difference between strains.

Keywords: microbiome, 16S rRNA, Iberian pig, Machine learning, sequencing data

Introduction

The Iberian pig breed is often considered as a single population, but it actually comprises five different strains (Entrepelado, Retinto, Torbiscal, Lampiño and Manchado de Jabugo) recognized in the Iberian Pig Herd Book, each of which has considerable genetic diversity (Clemente et al., 2008; Fabuel et al., 2004). Previous research has highlighted the superior meat quality (Ibáñez-Escriche et al., 2016) and fertility (Noguera et al., 2019) of the Retinto (RR) strain over the Entrepelado (EE) strain. Conversely, the EE strain has shown significant maternal effects compared to RR (Noguera et al., 2019), affecting offspring growth. Furthermore, crosses between these strains have shown heterosis effects on meat quality and litter size traits (Ibáñez-Escriche et al., 2016; Noguera et al., 2019). Beyond genetic factors, research has shown that the host microbiome also plays a role in shaping complex phenotypic traits in pigs, such as growth and carcass composition (Bergamaschi et al., 2020). One approach to explore the interaction between the host genetic variation and the microbiome in Iberian pigs is to use microbiome components as potential biomarkers to classify animals into different genetic groups. The development of ML classifiers based on microbiome-derived features (genera abundances) allowed to uncover meaningful biological patterns between the microbiome and the trait of interest. This study aimed to investigate the gut microbiome of Iberian pigs of two different strains and their reciprocal

crosses, through application of eight ML algorithms on 16S rRNA sequencing data, identify the most suitable set of predictors that could help predict the animals' genetic groups, and then compare the resulting predictors with the results of DA analysis. Finally, we explored explanations for our best predictive models and the set of predictors.

Material and methods

1. Animals and samples

The animals used in this study belonged to two Iberian purebred pig strains (RR and EE) and their reciprocal crosses (ER and RE) from the Iberian Testing Center (CTI) of the company INGA FOOD S.A (Extremadura, Spain). The pigs were randomly housed in groups of 80, avoiding full sibs, and fattened *ad libitum* by automatic feeders with commercial diets. In total, 239 castrated males were used, of which 74 pigs belonged to the EE strain, 63 to the RR strain, 51 to the RE strain, and 51 to the ER strain, where the first letter indicates the paternal line and the second the maternal line. The pigs weighed on average 161.60 ± 13.59 kg at the end of the fattening period and were on average 346.74 ± 45.54 days of age. Feces samples were collected at the CTI facilities before the animals' transport to the slaughterhouse, and stored at -80°C until further analysis. This process involved four separate batches of at least 68 animals each, except the last batch of 26 animals.

2. 16S microbiome profiling

Fecal DNA was extracted and amplified, and amplicons were sequenced in an Illumina MiSeq instrument to generate paired end reads of 2 × 300 bp. Sequences were analyzed using QIIME2-2023.7 (Bolyen et al., 2019) and Amplicon sequence variants (ASV) for each sample were identified using DADA2 algorithm (Callahan et al., 2016). Taxonomic annotation was performed using the SILVA reference database (Quast et al., 2012) and the ASV table collapsed to the genus level was used for further analysis.

3. Diversity and differential abundance analysis

Alpha diversity metrics (Shannon index and Pielou evenness) were calculated based on raw genera counts after rarefaction of the sequences. The Kruskal-Wallis (KW) test was used to assess differences in the alpha diversity metrics between samples, with p-value ≤ 0.05 considered significant for all statistical tests. Posterior analyses were performed using filtered genera abundances, where only genera present in at least 30% of the samples for each strain were included. Genera abundances were centered log-ratio (clr) transformed (Aitchison, 1986) and Beta diversity was evaluated by computing Aitchison dissimilarity distance matrix. Permanova analysis with 999 permutations was performed on the distance matrix to evaluate the differences between strains, age and animal batch. Differential abundances (DA) between strains, maternal lines, paternal lines and heterosis groups were identified by fitting Bayesian linear models. Genera with a minimum difference mean of 0.50 SD and a P0 (probability of the difference being either positive or negative) higher than 90% were considered differentially abundant.

4. Feature selection and Machine learning analysis

We applied the Select K best feature selection method (Tislenko et al., 2022) to find reduced feature sets that maximize model's performance. We used the mutual information as a scoring function. Eight supervised ML classifiers were then implemented to determine the genetic group of the Iberian animals based on the clr-transformed genus abundances. The algorithms include tree-based models (Decision Tree (DT), Random Forest (RF), Adaboost (AB), Catboost (CB), XGboost (XGB)), kernel based models (Support Vector Machine (SVM)) and probabilistic models (Gaussian Naive Bayes (GNB) and Logistic regression (LR)). All the ML algorithms were run using the Sci-kit learn module (Pedregosa et al., 2012) in Python v.3.11.5.

For each algorithm, we explored five different scenarios:

1. Four-strain scenario: Each strain (EE, RR, ER, RE) was treated as a separate class.
2. Purebred scenario: Only purebred individuals were included (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: Grouped by crossed (ER/RE) or purebred individuals (RR/EE).

In each scenario, the dataset was randomly stratified into training and test sets, with a split ratio of 75/25. The training set was used for model building and selection, and hyperparameter tuning via 5-fold cross-validation, while the test set was used for independent evaluation of ML performance using the AUC measure. Distribution of AUC scores were computed using a 200 times bootstrap resampling technique applied to the dataset.

Results and Discussion

1. Results of 16S analysis and Diversity

The 16S amplicon analysis resulted in 683 ASVs with a total of 14 million reads and a mean read count of 59,364 ± 15,413 per sample. Taxonomic annotation identified 3 phyla, 5 classes, 13 orders, 27 families and 55 genera in the 237 fecal samples, from which we retained only genera that appeared in at least 30% of the samples in each strain, reducing the list to 37 genera. Analysis of alpha diversity metrics using KW test did not reveal any significant differences within samples belonging to different strains. Beta diversity revealed a statistically significant, although modest, effect attributed to the strain (p-value = 0.029). Conversely, the effects of the animal batch and the age were more pronounced, showing a highly significant association between microbiome at genus level and batch (p-value = 0.001), and similarly between microbiome and the age (p-value = 0.001).

2. Differential taxonomic composition

After adjusting for batch and age effects, nine genera showed differential abundance (DA) (minimum difference mean of 0.50 SD and a P0 higher than 90%). Table 1 shows taxa found to be differentially abundant between different genetic groups (comparisons with no DA genera are not shown).

Table 1. Differentially abundant taxa between different genetic groups

Comparison groups	Differentially abundant taxa
EE-RR	Lactobacillus*, Clostridium_sensu_stricto_1, Acetitomaculum†, Frisingicoccus°, Romboutsia‡, and Terrisporobacter
EE-RE	Acetitomaculum† and Solobacterium
ER-RR	Lactobacillus* and Romboutsia‡
ER-RE	Intestinibacter
RE-RR	Frisingicoccus° and Prevotella
EE paternal group vs. RR paternal group	Lactobacillus*

*, °, †, ‡: differentially abundant taxa overlapping among different comparison groups

3. Classification results

In the Purebred scenario, with only 136 samples, the highest mean AUC_{test} reached 0.80, which is considered a good classification performance (White et al., 2023) indicating an 80% likelihood of correctly classifying the samples in their correct group. This performance was reached using GNB and LR. Regarding the selected features for this scenario, ten genera were selected using Select K best method. Half of them were found to be differentially abundant, namely *Clostridium sensu stricto 1, Acetitomaculum, Lactobacillus, Frisingicoccus and Terrisporobacter*. *Clostridium sensu stricto 1* had the highest importance score in the Purebred scenario and was linked to increased intramuscular fatness (IMF) and backfat thickness in pigs (Tang et al., 2020). Studies reported that the RR strain surpassed the EE strain in meat quality, including backfat thickness (Ibáñez-Escriche et al., 2016). However, DA analysis revealed a higher abundance of this genus in the EE strain with respect to the RR strain. This inconsistency can be due to the fact that our taxonomic assignation only reached the genus level. Whereas Clostridium sensu stricto 1 genus has demonstrated a functional versatility, and contains certain species that can be opportunistic pathogens (Hu et al., 2021). *Acetitomaculum* and *Lactobacillus* showed the second and third highest importance score respectively, demonstrating a DA between the EE and RR strains, with RR showing higher abundances of these bacteria. *Acetitomaculum* is recognized for its ability to produce SCFAs through the fermentation of dietary polysaccharides (Biddle

et al., 2013), whereas *Lactobacillus* was significantly associated with an increase in fat deposition in commercial and nucleus population of pigs (Maltecca et al., 2021).

In the Maternal scenario, the prediction performance reached an AUC_{test} of 0.64 using CB, considered marginally acceptable. Six predictors were included in the predictive models but no relevant DA was observed in any bacteria between maternal groups. Various studies suggested that the piglet intestinal microbiome is vertically transmitted from the mother (Lim et al., 2023; Liu et al., 2023), as they come into contact with the dam's microbial communities during and after passing through the birth canal, during nursing, or suckling and maternal care. However, there is limited information on how maternal abilities impact the abundance of microbiome bacteria after the partum.

In the Paternal scenario, using only five predictors, the classification performance reached a mean AUC_{test} of 0.71 using GNB, considered a fair performance result. Only one genus, *Lactobacillus*, exhibited a relevant DA between the two paternal groups, being more abundant in the group of RR sire. However, the DA in *Lactobacillu*s between these paternal groups is mainly due to the difference between RR and EE. Some studies suggested the influence of the paternal microbiota on the phenotypic traits and microbiota of offspring, through paternal transgenerational epigenetic mechanisms, though it is still unclear how it regulates offspring microbiota (Li et al., 2022).

In the Heterosis and Four Strains scenarios, the best models' performances, using CB and RF respectively, achieved an AUC_{test} of 0.63 for both cases. In the Heterosis scenario, *Clostridium sensu stricto 1* emerged as the most critical variable impacting the models' prediction, although it was not differentially abundant between heterosis groups, which is in line with its poor performance.

In the Four Strains scenario, the most crucial variable impacting classification was *Phascolarctobacterium*, with its increased abundance correlating with reduced fat content in Large White pigs (Pei et al., 2021).

In general, the genera that exhibited a DA between different genetic groups, were selected as important features and were key for prediction across various classification scenarios. However, 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, with high economic importance in Iberian pigs.

From a performance perspective, CatBoost emerged as the most effective model for classifying Iberian genetic groups, followed by Random Forest and XGboost. Nevertheless, the issue of overfitting in tree-based models must be highlighted. Finally, we showed that the most genetically distant animals (EE vs. RR) were more easily discriminated by their microbiome using the trained ML models.

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Springer Netherlands • Bergamaschi et al. (2020). *Microbiome*, *8*(1), 110 • Biddle et al. (2013). *Diversity*, *5*(3), 627–640 • Bolyen et al. (2019). *Nature Biotechnology*, *37*(8), 852–857 • Callahan et al. (2016). *Nature Methods*, *13*(7), 581–583 • Clemente et al. (2008). *XIV Reunión Nacional de Mejora Genética Animal* • Fabuel et al. (2004). *Heredity*, *93*(1), 104–113 • Hu et al. (2021). *Frontiers in Immunology*, *12* • Ibáñez-Escriche et al. (2016). *Journal of Animal Science*, *94*(1), 28–37 • Li et al. (2022). *Animal Nutrition*, *11*, 142–151 • Lim et al. (2023). *Animals*, *13*(21), 3378 • Liu et al. (2023). *Veterinary Sciences*, *10*(2), 123 • Maltecca et al. (2021). *Animal Microbiome*, *3*(1), 57 • Noguera et al. (2019). *Animal*, *13*(12), 2765–2772 • Pedregosa, et al. (2012). *Scikit-learn: Machine Learning in Python* • Pei et al. (2021). *Frontiers in Microbiology*, *12* • Quast et al. (2012). *Nucleic Acids Research*, *41*(D1), D590–D596 • Tang et al. (2020). *Frontiers in Microbiology*, *11* • Tislenko et al. (2022). *VIII International Conference on Information Technology and Nanotechnology (ITNT)*, 1–5 • Varona et al. (2020). *Scientific Reports*, *10*(1), 21190.