The oral microbiome as a proxy for feed intake in dairy cattle

C. N. Marcos, A. Bach, M. Gutiérrez-Rivas, and O. González-Recio

1Departamento de Producción Agraria, ETSIAAB, Universidad Politécnica de Madrid, Ciudad Universitaria, 28040 Madrid
2Departamento de Mejora Genética Animal, Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria - CSIC, Carretera de la Coruña km 7.5, 28040 Madrid
3ICREA, Passeig de Lluís Companys 23, 08007 Barcelona
4Blanca from the Pyrenees, Hostalets de Tost, 25795 Lleida

**ABSTRACT**

Genetic material from rumen microorganisms can be found within the oral cavity, and hence there is potential in using the oral microbiome as a proxy of the ruminal microbiome. Feed intake (FI) influences the composition of rumen microbiota and might directly influence the oral microbiome in dairy cattle. Ruminal content samples (RS) from 29 cows were collected at the beginning of the study and also 42 d later (RS0 and RS42, respectively). Additionally, 18 oral samples were collected through buccal swabbing at d 42 (OS42) from randomly selected cows. Samples were used to characterize and compare the taxonomy and functionality of the oral microbiome using Nanopore sequencing and to evaluate the feasibility of using the oral microbiome to estimate FI. Up to 186 taxonomical features were found differentially abundant (DA) between RS and OS42. Similar results were observed when comparing OS42 to RS collected at different days. Microorganisms associated with the liquid fraction of the rumen were less abundant in OS42 as these were probably swallowed after regurgitation. Up to 1,102 KEGGs were found to be DA between RS and OS42 and these results differed when comparing time of collection, but differentially abundant KEGGs were mainly associated to metabolism in both situations. Models based on the oral microbiome and rumen microbiome differed in their selection of microbial groups and biological pathways to predict FI. In the rumen, fiber-associated microorganisms are considered suitable indicators of feed intake. On the other hand, biofilm formers like Gammaproteobacteria or Bacteroidia classes are deemed appropriate proxies for predicting feed intake from oral samples. The best lineal model to estimate FI was obtained with the relative abundance of taxonomical feature at genera level, achieving an average $R^2$ equal to 0.88 within the training data, and a root mean square error equal to 3.46 ± 0.83 (standard deviation; SD) kg of DM/ as well as a Pearson correlation coefficient between observed and estimated FI of 0.48 ± 0.30 in the test data. The results from this study suggest that oral microbiome has potential to predict FI in dairy cattle, and it encourages validating this potential in other populations.

**Key words:** dairy cattle, ruminal microbiome, oral microbiome, feed intake

**INTRODUCTION**

Feed efficiency is a moderately ($h^2$ ~0.14 to 0.37) heritable trait (Pryce et al., 2014) that has a large economic impact in dairy production as feeding might be responsible for over 60% of total costs (Connor, 2015; Miglior et al., 2017). Improving feed efficiency through genomic selection could increase the profitability of dairy farms (Pryce et al., 2014). Nevertheless, breeding for feed efficiency is hindered by the difficulty of individually recording feed intake for many animals, which is a requirement in most breeding programs (Seymour et al., 2019). Therefore, methods for determining feed intake at a large scale in the population are needed to expedite genetic progress.

The microbiome found in the rumen of dairy cattle hosts a wide diversity of microorganisms which interact with the environment, the diet and the host genotype (Delgado et al., 2019). Recently, multiple studies have proposed to use the rumen microbiome as a proxy for DMI or feed efficiency (i.e., Wang et al., 2015; Delgado et al., 2019; Monteiro et al., 2022). Unfortunately, current methods to sample rumen content preclude its use in commercial herds, because they require invasive procedures and cannot be conducted at large scale (Young et al., 2020). Rumenocentesis and oral tubing are labor-intensive processes, might have negative effects on the animal welfare and productivity (Nordlund and Garrett, 1994; Henderson et al., 2013; Paz et al., 2016), and the amount of sample collected is small.
and not highly representative (González-Recio et al., 2023). However, because of cattle rumination, a process where the animal regurgitates and comminutes rumen content, ruminal microorganisms and their DNA can be found in the oral cavity. For these reasons, several authors have proposed to use buccal swabs to sample the microbiome within the oral cavity and use it as a proxy of the rumen microbiome. This sampling method is non-invasive, and less expensive and labor intensive (Kittelmann et al., 2015; Young et al., 2020; de Freitas et al., 2022) than stomach tubing. In contrast, resident microorganisms within the oral cavity, including transient facultative anaerobes and feed associated microorganisms, are concomitantly sampled (Young et al., 2020), which may void the use of oral samples to study rumen microbiome. de Freitas et al. (2022) also stated that, although the rumen and the oral cavity share some taxa, saliva samples cannot be used to predict the ruminal microbiome. Similarly, Young et al. (2020) indicated that the oral microbiome sampled through buccal swabs could not be used in classical microbial ecology experiments but might be useful in association studies.

Third generation sequencing technologies offer new opportunities to study the metagenome of complex microbial systems. They may provide both functional and taxonomical information including the abundance of viruses which have a major role in the oral cavity (Thakkar et al., 2021), and offer some advantages over metataxonomic studies (López-García et al., 2022). Herein, we hypothesized that the oral microbiome, assessed using these sequencing technologies, could be used as a proxy to estimate feed intake. Previous studies have demonstrated that the rumen microbiome is moderately heritable ($h^2 \sim 0.10 - 0.30$; Saborío-Montero et al., 2021), and thus the oral microbiome could potentially be utilized in selective breeding programs to improve specific characters. Therefore, the objectives of this study were 1) to characterize the oral microbiome of dairy cattle sampled through buccal swabs in comparison to rumen microbiome using Nanopore sequencing, and 2) to evaluate the potential of using the oral microbiome as a proxy to estimate feed intake.

**MATERIALS AND METHODS**

**Ethical Statement**

All animal experiments were carried out in accordance with EU Directive 2010/63/EU for animal experiments and protocols for animal sampling were approved by the corresponding Ethical Committee (IRTA-257/2021).

**Animals, Feeding and Sampling**

Thirty-two Holstein cows producing 35.3 ± 1.22 kg of milk/d, with 510 ± 10.5 kg of BW, 167 ± 6.4 DIM, and 25.5 ± 0.71 kg of DM/d feed intake (FI), were used in this study. Cows were housed in a single pen at the Blanca from the Pyrenees experimental farm. The pen was equipped with free stalls bedded every 2 d with chopped straw, and cows had free access to fresh water. The pen was equipped with electronic feed bins (MooFeeder, MooSystems, Cortes, Navarra, Spain) that recorded the amount of feed consumed at every visit by every cow throughout the study. During the whole study, cows were fed a TMR twice a day. Ruminal content from each cow were sampled using a stomach tube attached to a mechanical pump at the beginning of the study. Ruminal contents were filtered through 4 layers of cheesecloth and frozen with liquid nitrogen and stored at −80°C until DNA extraction. Ruminal contents from the same 32 cows were sampled again after 42 d, and oral samples from 20 cows randomly selected were collected. Oral samples were obtained about 3 h after the morning feeding using a buccal swab (Puritan HydraFlock, Puritan Diagnostics Llc., Guilford, Maine) inserted laterally through the mouth of the cow coinciding with a regurgitation of a rumen bolus during rumination. Buccal swabs were then stored in vials until further analysis. The average weekly FI of each cow was monitored for 12 weeks, starting 6 weeks before the trial began. The average weekly FI corresponding to the sampling day was used in this study. During this period, an abnormal and sudden decrease in FI was observed in 3 cows, leading to their exclusion from the study. Consequently, a total of 76 samples were used in this study arranged in 58 ruminal samples (RS) collected during 2 sampling moments (0 and 42 d; RS0 and RS42, respectively) and 18 oral samples collected after 42 d (OS42). The phenotypic average of FI for the RS, RS0, RS42 and OS42 subsets were 24.6 ± 0.52, 25.6 ± 0.76, 23.6 ± 0.69, and 24.2 ± 0.87 kg of DM/d, respectively.

**Sample Processing and Bioinformatics**

Genomic DNA was extracted from 250 µL samples of homogenized ruminal content or buccal swabs remnant solution (~750 µL) using a “Dneasy Power Soil PRO” commercial kit (Qiagen, Valencia, CA, USA). Concentration and purity of extracted DNA were assessed using a Qubit fluorometer (ThermoFisher Scientific, Waltham, MA, USA) and Nanodrop ND-1000 UV/Vis spectrophotometer (Nanodrop Technologies Inc., Wilmington, DE, USA). Oxford Nanopore Technologies SQK-LSK109 Ligation Sequencing kit was used.
for sequencing in a Grid-ION automatic sequencer. The 1D Native barcoding Oxford Nanopore Technologies kit (EXP- NBD104 or EXP-NBD114) was used for multiplexing the samples, pooling barcoded DNA from 24 samples for each run. Pooling was done using a 1.5-mL DNA LoBind tube to perform adapter ligation and sequenced using a R9.4.1 flow cell. Basecalling was conducted using high accuracy mode in the MinKNOW built-in basecaller (Guppy v6.4.2). Sequences with quality score <7 and length <150 bp were eliminated. The remaining reads were analyzed with SqueezeMeta v1.6.1 pipeline for long reads that is devised to process long reads from Oxford Nanopore Technologies (Tamames and Puente-Sánchez, 2019; Adrian-López et al., 2022). SqueezeMeta sqm-longreads tool executes Diamond Blastx against GenBank nr taxonomic and KEGG functional databases to map reads through the Last Common Ancestor method for taxonomy, and the fun3 algorithm for functionality. Two RS were withdrawn at this stage as all reads were mapped as unclassified. All reads classified as Arthropoda this stage as all reads were mapped as unclassified. All reads classified as Arthropoda, Streptophyta, Chordata, Mollusca, Nematoda, Chlorophyta or Rotifera within phylum level were discarded. All unclassified reads were discarded as well. Taxonomical features whose relative abundance (RA) was not above 0.01% in at least one sample were also filtered out. Unclassified KEGGs were removed and the remaining functional features were filtered the same way. Taxonomical features included those microorganisms that could not be classified up to their respective level.

**Calculation and Statistical Analyses**

Metagenomic data were normalized using the unweighted centered log ratio (CLR) method within the easyCoda R package (Greenacre, 2018). When needed, zero count values were augmented with the cmultiRepl function of the zComposition R package (Palarea-Albaladejo and Martín-Fernández, 2015). Beta-diversity was assessed through a principal component analysis using the CLR-transformed data at either phylum, class, order family, genus levels, or functionality, and the prcomp function of core R. The $r^2$ between matched samples on the first 2 principal components (PC) was computed as well. A PERMANOVA analysis was conducted using the adonis function of the R package vegan and the matrix of Aitchison distances between samples (Oksanen et al., 2019). Then, a differential abundance analysis was implemented on the CLR-transformed data to determine the differentially abundant (DA) taxonomical features at genus level or KEGGs between RS and OS42 using the Limma R package (Ritchie et al., 2015). The $P$-values were adjusted ($P_{adj}$) by the Benjamin-Hochberg method to control for false positives. A logFC ≥0.5 and $P_{adj}$ < 0.05 were set as threshold values. Spearman correlation coefficients ($\rho$) between FI and the RA of taxonomical features were also estimated using 10,000 bootstrapped (BS) replicates.

**Prediction of Feed Intake**

Linear models to predict FI from all taxonomic microbial features and common productive parameters (milk yield, BW and DIM) were obtained through Bayesian LASSO (BLASSO) regression using the BGLR function from the BGLR package (Pérez and de los Campos, 2014) with 2,000 iterations, a burn-in of 1,000, a thin period of 5, and $\lambda$ parameter as a random variable assigned a Gamma prior with $s = 1.1$. The scaled values of both taxonomical features and productive parameters were used. Models were validated using out-of-bag validation. One thousand BS replicates were withdrawn from the data set and used to train the model. Observations not included in the BS replicates were used to test the model accuracy. The $R^2$ of the model, and the root mean square error (RMSE) and Pearson correlation coefficient ($r$) between the observed and estimated FI within the testing data were calculated.

**RESULTS AND DISCUSSION**

**Taxonomy and Functionality of Microbial Composition**

The final taxonomy composition consisted of 9,516,143 and 3,844,303 reads within RS and OS42, respectively, classified in 229, 170, 120, 74 and 36 taxonomical features up to genus, family, order, class, and phylum level, respectively. Likewise, the final functionality composition consisted of 8,473,563 and 3,654,937 reads from RS and OS42, respectively, classified in 2,172 KEGGs.

Bacteria was the most predominant superkingdom in RS (82.3% of total reads within RS in the final taxonomy composition). The 3 main phyla observed in these samples were Bacteroidetes (36.9%), mainly composed by the Prevotella genus (19.4% of total reads), Firmicutes (24.7%), and Proteobacteria (6.36%) (Figure 1). Other phyla observed in these samples were Fibrobacteres, Spirochaetes, Kiritimatiellaeota, Actinobacteria and Lentisphaerae, which accounted for as much as 5.10% of reads. The RA of Eukaryotes in RS was 16.5% and these were mainly composed by the phyla Ciliophora (3.85%) and Chytridiomycota (0.61%), although a large number of eukaryotes could not be classified (11.6%) due to lack of genome reference in the databases. Archaea representation in RS was 1.17%,
where *Methanobacteria* was the most prevalent class accounting for 0.80% of reads. The relative abundance of viruses in RS was low (0.11%), and both *Uroviricota* and *Nucleocytophoviricota* phyla were identified within these samples. The abundance of viruses was small, but they might be underrepresented as the protocols used in this study were not designed for virus (Koonin et al., 2022). Among KEGGs, up to 29.9% of total reads within RS in the final functionality composition were classified as pathways associated to metabolism (Figure 2), largely carbohydrate metabolism (10.8%), 14.0% to genetic information processing, 8.36% to environmental information processing, 3.30% to cellular processes, and 16.4% of these to other pathways (Figure 2). Further, 27.9% of total reads were classified as BRITe hierarchies. Proteins involved in genetic information processing represented 12.7% of total reads, signaling and cellular processes (9.39%) and metabolism (5.75%). Similar results were reported by López-García et al. (2022) when assessing the ruminal microbiome and functionality using Nanopore sequencing technology, even though the number of unclassified or unassigned reads herein were substantially larger.

The proportion of unclassified reads in OS42 was lower than that observed in RS (48.0 vs 61.4%, respectively). Some differences were observed when comparing the microbiome within RS and OS42. Similarly, bacteria were the most prevalent microorganisms in OS42 (90.3% of total reads within OS in the final taxonomy composition) and the major phyla were *Bacteroidetes* (47.8%) and *Firmicutes* (23.0%). Likewise, *Prevotella* was the largest genus in OS42, in agreement with Young et al. (2020). The RA of *Fibrobacteres* and *Spirillochaetes* (3.78 and 3.16%, respectively) in OS was larger than that of *Proteobacteria* (2.87%). *Chytridiomycota* and *Ciliophora* were the main phyla among eukaryotes, but the RA of eukaryotes in OS42 was considerably lower (8.57%). The RA of archaea and viruses (1.04 and 0.09% respectively) were similar in both RS and OS42 and the same main phyla were identified. Tapio et al. (2016) compared ruminal liquid, ruminal bolus, and buccal swab samples using 16S-18S rRNA and reported that communities were correlated but not identical. Similarly, Young et al. (2020) observed some differences in the ruminal and oral microbiome compositions, although sharing some common taxa. The RA of KEGGs assigned to different pathways and BRITe hierarchies in OS42 were similar to those observed in RS: metabolism (32.9% of total reads within OS42 in the final functionality composition), genetic information processing (13.5%), environmental information processing (6.32%), cellular processes (2.53%), other pathways (16.5%) and BRITe Hierarchies (28.1%); proteins involved in genetic information processing (13.0%), signaling and cellular processes (10.3%), and metabolism (4.82%).

**Comparison Between Ruminal and Oral Microbiomes**

A clear clustering by sample type was highlighted in the Beta-diversity analysis, and significant differences were observed in the PERMANOVA analysis (P < 0.001) at all taxonomical levels or gene function classification (Figure 3). Young et al. (2020) described a clustering similar to current clustering of buccal swabs and rumen samples of cows in a PCoA showing Bray-Curtis dissimilarities and reported significant differences between them. RS0 and RS42 clusters were akin, which suggests that microbiome composition is stable and reproducible at least within a period of a few weeks under similar conditions. These results agree with Monteiro et
al. (2022) who reported that the ruminal microbiome of dairy Holstein cows was stable through lactation given the same environmental conditions. A large proportion of the variance in both the first and second principal components (PC) was shared between matched RS0 and RS42 samples in terms of taxonomy and functionality, accounting for approximately 30% and 31% of the first and second PC, respectively (Table 1). Although a clear distinction between RS and OS42 could be observed in the β-diversity analysis, the shared variance between matched RS and OS samples in the first and second PC was large, and that was greater between RS42 and OS42. These results are expected within oral samples as regurgitated ruminal microorganisms will be concomitantly sampled (Young et al., 2020). In addition, sampling by oral tubing might potentially contaminate ruminal samples with oral microorganisms (Shen et al., 2012). However, a significant proportion of the variability observed in these diversity measures remains associated with the microbiome of the sample itself.

Up to 186 taxonomical features were found to be DA between RS and OS42 (Figure 4), including archaea (10), eukaryotes (24; mainly ciliate protozoa and fungi), viruses (2) and bacteria (150). Predominantly, overabundant microorganisms in RS were comprised of protozoa from the Ciliophora, Apicomplexa, Eouvea and Parabasalia phyla, fungi from the Chytridiomycota, Ascomycota, Basidiomycota, Mucoromycota, and Oomycota phyla, archaea from the Methanobacteria class, and mainly bacteria from the Bacteroidetes, Firmicutes, and Proteobacteria phyla, although bacteria from the Actinobacteria, Cyanobacteria, Fibrobacteres, Kiritaniitellaeota, Lentisphaerae, Spiirochaetes, Synergistetes, and Verrucomicrobia phyla were also overabundant in the rumen. Clostridia and Negativicutes classes were the most observed within the Firmicutes phylum whereas Bacteroidia and Gammaproteobacteria classes were the most observed within the Bacteroidetes and Proteobacteria phyla, respectively. Overabundant Bacteroidetes and Gammaproteobacteria in RS were mainly from the Bacteroidales, and the Aeromonadales, Enterobacteriales and Vibrioanales orders, respectively. On the contrary, 3 taxonomical features from 2 archaeal orders, Methanomassiliicoccales and Methanomicrobiales, and 2 viral taxonomical features from the Cauloviricetes class were overabundant in OS42. Overabundant bacteria in OS42 were within the Actinobacteria (mainly Actinomycetia class), Firmicutes and Proteobacteria phyla. Bacteria from the Bacteroidetes, Chloroflexi, Lentisphaerae, Planctomycetes, Spiirochaetes, Synergistetes, Tenericutes and Verrucomicrobia phyla were also overabundant in OS42. In contrast to RS, Bacilli and Clostridia classes were the most recognized within the Firmicutes phylum, and bacteria from the Alphaproteobacteria, Betaproteobacteria, Gammaproteobacteria (Enterobacteriales, Pseudomonadales and Xanthomonadales orders), Deltaproteobacteria and Epsilonproteobacteria classes were overabundant in OS42. Microorganisms associated within the solid rumen fraction are expected to be more abundant in OS42 rather than those associated with the liquid fraction as the latter is immediately swallowed after regurgitation (Beauchemin, 2018). Klevenhusen et al. (2017) reported higher gene copy numbers from protozoa and fungi in the free ruminal liquid compared with the solid-associated fraction, and reported that the Clostridium, Methanobrevibacter, Pyramidobacter, Ruminobacter, Ruminococcus, Selenomonas and Succinivlasticum genera were within the most abundant identified taxa in the ruminal liquid, which concurs with our results. In agreement with the Beta-diversity analysis, no great differences between RS0 and RS42 were observed, and only 11 features within the Actinobacteria, Firmicutes and Proteobacteria phyla were found to be DA. Moreover, differences between RS0 and RS42 and OS42 were scarce, as both contrasts shared approximately the 80% of DA taxonomical features (Figure 4a), and the rest were features within the same phyla. In general, OS42 microorganisms differentially abundant with respect to RS were also differentially abundant from either RS0 or RS42. Despite there was a time gap of 6 wk between collection of R0 and OS42, results from these samples were similar, which emphasizes that microbiome composition is overall stable and reproducible within a period of 6 wk at least. It must be pointed out that the animals were given the same diet during the trial, and thus no significant changes in the rumen microbiome are expected from diet change (Monteiro et al., 2022). We might hypothesize that changes in the oral microbiome are strongly associated to shifts in the ruminal microbiome or due to dietary changes.

Kittelmann et al. (2015) assumed that bacterial taxa whose maximum RA in buccal swabs was 1% greater than that in any sample collected through oral tubing might be true oral bacteria. With a similar approach, microorganisms found to be overabundant in OS42 in our study could be considered native to the oral cavity whereas overabundant microorganisms in RS could be considered true ruminal microorganisms. Certainly, overabundant microorganisms in OS42 reported herein were in agreement with previous studies where the oral microbiome of cattle was characterized (Borsanelli et al., 2018; McMullen et al., 2020; Young et al., 2020, de Freitas et al., 2022). Many overabundant bacteria in OS42 were also defined in the core oral microbiome of humans (García-Godoy and Hicks, 2008; Kolenbrander et al., 2010). However, observed differences between RS
Figure 3. First 2 principal components for the microbiota composition in the rumen samples (RS) at the beginning of the trial (RS0), RS at 42 d (RS42) and oral samples (OS42) at 42 d. Symbols represent the samples using Euclidean distance of centered log ratio-transformed taxa abundances or functionality colored by sample type. Cows with both OS42 and RS are represented by circles (triangles otherwise). P: significance of sample type effect in the PERMANOVA; R²: variability explained by sample type.
and OS42 might be due to different conditions within the oral cavity compared with those of the rumen (Erdman et al., 1988; Riley et al., 2001). Moreover, contamination by soil microorganisms (de Freitas et al., 2022) or by forage epiphytic bacteria (Young et al., 2020) could also explain differences between RS and OS42. We must also consider that oral microorganisms will be present within ruminal communities (i.e., Moraxella, Neisseria, etc.; Wang et al., 2021) if capable to withstand ruminal environments (de Freitas et al., 2022).

The differential abundance analysis carried out with KEGGs revealed that up to 1,102 KEGGs were DA between RS and OS42 (Figure 4b) (protein families: genetic information processing (136), protein families: metabolism (66), protein families: signaling and cellular processes (90), cellular processes (52), environmental information processing (110), genetic information processing (131), human diseases (12), metabolism (329), not Included in Pathway or BRITE (161) and organismal systems (13)). Among these, 505 were overabundant in RS and 597 overabundant in OS42. No clear distinctive pathways or hierarchies between RS and OS42 could be recognized among differentially abundant KEGGs. However, overabundant KEGGS in RS associated to metabolism were proportionally greater than those observed in OS42 (35.2 vs 25.3%, respectively). Up to 572 and 503 KEGGs were DA between RS0 and RS42, and OS42, respectively, which are considerably smaller numbers than that observed in the RS vs OS42 contrast. Up to 463 differentially abundant KEGGs were common in both contrasts (Figure 4b), and no differentially abundant KEGGs between RS0 and RS42 were observed. Likewise, the proportion of overabundant KEGGs associated to metabolism significantly differed between RS and OS42 in these 2 contrasts. These results suggest that there may be some differences in the metabolic activity between ruminal microorganisms and those found in the oral cavity. Such disparities were expected, given the substantial differences between microorganisms in the rumen and oral cavity observed in the differential abundance analysis.

Furthermore, these results could also suggest that the functionality of microorganisms might be highly variable among sample types.

### Estimation of Feed Intake from Microbiome

The R² in the training data BLASSO regression ranged between 0.71 and 0.92 depending on the taxonomic level used, with mean values of approximately 0.88 using rumen microbial information and 0.81 for oral microbiome. Monteiro et al. (2022) constructed a model to discern how much of the variation in DMI (kg/d) could be explained by the ruminal microbiome, as well as the joint rumen and lower gut microbiome, reporting an adjusted R² equal to 0.35 and 0.82, respectively. Our study showed a much larger proportion of FI variance explained by both the ruminal and oral microbiome, although R² values were not corrected for the number of variables. The larger R² reported here might also suggest overfitting. The mean values for the predictive correlation (r) and RMSE for FI from bootstrapping in the test data ranged between 0.16 – 0.48, and 4.39 – 3.46 kg/d, respectively, depending on the taxonomical level and microbial sample used. Although a large variability was observed for these predictive metrics.

The predictive ability using rumen samples (RS0 and RS42) achieved larger predictive ability using taxa at the phylum level, with a mean predictive correlation of 0.39 (±0.21 standard deviation (SD)) and RMSE of 4.00 (±0.91 kg DM/d SD). The L1-penalized regression (LASSO) can reduce data sparsity and indirectly perform feature selection and is adequate for small sample sizes when \( n \ll p \) (Tibshirani, 1996). However, the LASSO regression requires hyperparameter optimization (i.e., lambda) which is often achieved by cross validation. In BLASSO, lambda may be treated as a random variable, eliminating the need for optimization, which aligns well with our situation. However, it does not strictly perform feature selection, as all variables will receive a coefficient estimate, although it will be

---

### Table 1. Shared variance (r²) between matched samples (rumen samples (RS) at the beginning of the trial (RS0), RS at 42 d (RS42) and oral samples at 42 (OS42)) on the first two principal components (PC) obtained in the β-diversity analysis at both taxonomic and functional levels

<table>
<thead>
<tr>
<th></th>
<th>Phylum</th>
<th>Class</th>
<th>Order</th>
<th>Family</th>
<th>Genera</th>
<th>KEGG</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RS0 vs RS42</td>
<td>0.32</td>
<td>0.29</td>
<td>0.32</td>
<td>0.20</td>
<td>0.40</td>
<td>0.26</td>
</tr>
<tr>
<td>RS0 vs OS42</td>
<td>0.35</td>
<td>0.28</td>
<td>0.19</td>
<td>0.22</td>
<td>0.30</td>
<td>0.18</td>
</tr>
<tr>
<td>RS42 vs OS42</td>
<td>0.40</td>
<td>0.21</td>
<td>0.17</td>
<td>0.36</td>
<td>0.31</td>
<td>0.09</td>
</tr>
<tr>
<td>PC2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RS0 vs RS42</td>
<td>0.33</td>
<td>0.26</td>
<td>0.29</td>
<td>0.31</td>
<td>0.48</td>
<td>0.21</td>
</tr>
<tr>
<td>RS0 vs OS42</td>
<td>0.19</td>
<td>0.15</td>
<td>0.00</td>
<td>0.21</td>
<td>0.50</td>
<td>0.21</td>
</tr>
<tr>
<td>RS42 vs OS42</td>
<td>0.58</td>
<td>0.41</td>
<td>0.04</td>
<td>0.49</td>
<td>0.74</td>
<td>0.71</td>
</tr>
</tbody>
</table>
Table 2. Variables with the ten largest scaled β coefficients sorted by decreasing value, obtained in models to predict feed intake (kg DM/d) using rumen samples (RS), RS at 0 d (RS0), RS at 12 d (RS12) and oral samples at 42d (OS42), counting taxonomical features at phylum (P), class (C), order (O), family (F) or genera (G) levels, and milk yield (MY) and days in milk. The β coefficients are shown between parentheses. U.: unclassified; C. Candidatus

<table>
<thead>
<tr>
<th>Level</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>OS42 (n = 18)</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>Bacteroidetes (−0.33)</td>
</tr>
<tr>
<td>C</td>
<td>U. Bacteroides (−0.22)</td>
</tr>
<tr>
<td>O</td>
<td>U. Bacteroidetes (−0.12)</td>
</tr>
<tr>
<td>F</td>
<td>BW (−0.08)</td>
</tr>
<tr>
<td>G</td>
<td>BW (−0.06)</td>
</tr>
<tr>
<td>RS0 (n = 28)</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>MY (−0.55)</td>
</tr>
<tr>
<td>C</td>
<td>MY (−0.47)</td>
</tr>
<tr>
<td>O</td>
<td>MY (−0.28)</td>
</tr>
<tr>
<td>F</td>
<td>MY (−0.22)</td>
</tr>
<tr>
<td>G</td>
<td>MY (−0.14)</td>
</tr>
<tr>
<td>RS42 (n = 28)</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>BW (−0.70)</td>
</tr>
<tr>
<td>C</td>
<td>BW (−0.39)</td>
</tr>
<tr>
<td>O</td>
<td>BW (−0.28)</td>
</tr>
<tr>
<td>F</td>
<td>BW (−0.21)</td>
</tr>
<tr>
<td>G</td>
<td>BW (−0.16)</td>
</tr>
<tr>
<td>RS (n = 56)</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>MY (−0.75)</td>
</tr>
<tr>
<td>C</td>
<td>U. Bacteroides (−0.99)</td>
</tr>
<tr>
<td>O</td>
<td>U. Bacteroides (−0.62)</td>
</tr>
<tr>
<td>F</td>
<td>MY (−0.55)</td>
</tr>
<tr>
<td>G</td>
<td>MY (−0.43)</td>
</tr>
</tbody>
</table>
Figure 4. Differentially abundant (DA) (a) taxonomical features at genus level and (b) KEGGs between rumen samples (RS), RS obtained at the beginning of the trial (RS0), after 42 d (RS42), and oral samples (OS42) obtained at 42 d. Dots represent DA taxonomical features or KEGGs colored and shaped by either phylum or functionality pathways. Significance thresholds were established at $P_{adj} = 0.05$ and logFC = ± 0.5. Venn diagrams depict the overlap of DA taxonomical features or KEGGs between contrasts.
Figure 5. Estimated root mean square error (RMSE; kg DM/d) and Pearson correlation (r) between the observed and estimated feed intake from out-of-bag validation. The dots are point bootstrapped estimates. The horizontal line displays the confidence interval around the median (first quartile - 1.5 * inter-quartile range (IQR) and third quartile + 1.5 * IQR). The box is drawn from first to third quartiles with the vertical line denoting the median.
close to 0 for unimportant ones. Thus, variables with large scaled β coefficients are deemed important. Productive parameters, such as milk yield and BW were the variables with the largest scaled β coefficients. Other variables with large β coefficients in this model were the RA of bacterial phyla Planctomycetes, Bacteroidetes, C. Melainabacteria, Firmicutes and Fusobacteria, the RA of archaeal phylum Euryarchaeota, as well as the RA of unclassified archaea, and the RA of fungal phylum Oomycota. The scaled RA of taxonomic features and productive parameters values were utilized in model training and testing, potentially increasing collinearity. Some variables may have been selected due to spurious relationships with FI. Nevertheless, variables selected in both ruminal and oral models exhibit a distinct, recognizable pattern, that is, fiber degrading microorganisms were selected in ruminal model whereas biofilm producers were selected in oral models.

Cows were fed a TMR containing more than 50% of fibrous sources and thus, we hypothesize that the relationship between these microorganisms and FI is directly caused by the level of fiber intake and fiber degradation. Planctomycetaceae phyla is a member of the PVC superphyla, alongside the Kiririmatiellaeota, Lentisphaerota and Verrucomicrobia phyla (Rivas-Marin and Devos, 2018). This superphyla is potentially composed of fiber degrading bacteria (McGovern et al., 2018). Similarly, fungi and archaea can be associated to fiber degradation, either because they play a direct role or are symbiotic to microorganisms related to it (Akin and Borneman, 1990; Newbold et al., 2015). Bacteroidetes and Firmicutes phyla encompass bacteria with diverse roles, including fiber degradation. Previous authors reported associations between Bacteroidetes phyla, and some species within the Firmicutes phylum, with feed efficiency in dairy cattle (Delgado et al., 2019; Xue et al., 2022). Microorganisms within the Fusobacteria phylum have been found in the oral microbiome of dairy cattle and goats (Borsanelli et al., 2018; Borsanelli et al., 2022). The association of Fusobacteria with FI could be attributed to potential saliva contamination, considering that rumen samples were collected through oral tubing.

Predictions obtained from RS0 performed worse on the test data (~r = 0.21, ~RMSE = 4.19 kg/d) than the joint rumen data set (RS0 and RS42) at all taxonomical levels. In this case, the model with the greater predictive ability, was observed at class taxonomic level with a mean bootstrapped predictive r of 0.25 (±0.25 SD) and RMSE of 4.11 (±0.80 kg DM/d SD). Milk yield and BW were within the variables with the largest β coefficients, as well as taxonomical features associated to the fibrous fraction of feed, such as Planctomycetes and Kiririmatiellae classes, and archaeal and fungal classes. Microorganisms from the Sphingobacteriia class may also possess the capability to degrade plant cell walls (Zhou et al., 2009).

Models obtained when using only RS42 showed a slight improvement in comparison to RS0 and to the entire set of RS. In this case, the taxonomic level with the largest predictive ability was the order (r = 0.42 ± 0.27 SD, RMSE = 3.51 ± 0.62 kg/d SD) but differences between taxonomical levels were small. Akin the RS and RS0 samples, the best model included BW, and taxonomical features associated to fiber degradation such as the Sphingobacteriia class, U. Planctomycestes, U. Opitutae and U. Verrucomicrobia. This model also included orders from both the Proteobacteria and Tenericutes phyla. In agreement, Delgado et al. (2019) reported an association between Proteobacteria and feed efficiency and Zhang et al. (2021) between Tenericutes and feed intake.

The oral samples showed higher predictive ability than rumen samples at different taxonomical levels (~r = 0.43 and 0.32, ~RMSE = 3.53 and 3.98 kg DM/d). Models using the oral microbiome selected different microbial features compared with RS. The model with the largest predictive accuracy was achieved at the genera level (r = 0.48 ± 0.30 SD, RMSE = 3.46 ± 0.83 kg/d SD). Here, the 95% highest probability density for the predictive ability metrics in the testing set ranged from 1.69 to 4.77 kg DM/d and ~0.21 to 0.95 for RMSE and r, respectively. Results from OS42 also showed large variability in the out-of-bag subsets. Brown et al. (2022) obtained a model to estimate DMI with a root mean squared error of the prediction equal to 2.83 kg of DM/d when using cow descriptors, milk yield and composition, milk fatty acid profile, and production and efficiency transmitting abilities, although sample size used in that study was considerably larger than that used in this manuscript. The National Research Council (NRC) proposed a simple equation to estimate FI using fat-corrected milk, body weight and week of lactation (NRC, 2001). This equation was used to estimate FI, and the predictions were correlated to real observations to estimate the predictive accuracy, which resulted to be r = 0.37 and RMSE = 3.98 kg DM/d on the RS data set, r = 0.34 and 0.40, and RMSE = 4.43 and 3.49 kg/DM/d on the RS0 and RS42 data sets, and a r = 0.14 and RMSE = 3.96 kg DM/d on the OS42 data set. The accuracy of the NRC model was similar to that obtained when using RS, but an improvement was observed when compared with the OS42. It must be pointed out that milk fat was not measured, and hence plain milk yield was used in the NRC equation.

Although β coefficients of productive parameters in oral models were not as large as in the models using RS, oral model included BW as one of the variables with
the largest β coefficients. This model also included the RA of *Anaerovibrio* and *Alloprevotella* genera, from the *Proteobacteria* and *Bacteroidetes* phyla, respectively, as well as viral taxonomical features and other features from the *Proteobacteria* and *Bacteroidetes* phyla. Microorganisms from the *Proteobacteria* and *Bacteroidetes* phyla, specifically microorganisms from the *Gammaproteobacteria* and *Bacteroidales* classes, and viruses were consistently chosen in all models at every level when OS42 were used. The bacterial compositions in the saliva appeared to be associated with the amount of feed consumed by cows. Another reasonable hypothesis is related to the capacity of microorganisms to adhere to surfaces or to form biofilms within the oral cavity which might explain their association with FI. Feed In-take is positively correlated with saliva production and secretion rate, as well as rumination time and chewing activity (Silanikove and Tadmor, 1989; Maekawa et al., 2002a). Bacteria that are unable to adhere to any surface through biofilm formation are washed off the oral cavity by salivary flow (Kolenbrander et al., 2010). Young et al. (2020) indicated that increased salivary flow can modify the abundance of key rumen taxa in oral samples. In agreement, *Gammaproteobacteria* class had the largest proportion of KEGGs associated with biofilm formation (Figure 6). Borsanelli et al. (2022) reported that *Prevotella* genus play an important role in the dental microbiome associated with goat periodontitis, which is a polymicrobial biofilm-induced inflammatory disease. This could help to explain the inclusion of variables such as the *Bacteroidales* order or the *Alloprevotella* genus, as these showed a phenotypic correlation with the RA of *Prevotella*. Much like what was seen with rumen models, the low abundance of virus DNA might be due extraction protocols and redundancy with some bacteria.

Overabundant taxonomical features in rumen samples within classes with large proportions of KEGGs associated with biofilm, showed significant phenotypical correlations between their relative abundance in OS42 and FI (Figure 7). *Prevotella* genus was also overabundant in RS and negatively phenotypically correlated with FI ($r = -0.62$). These microorganisms are probably bacteria within the rumen solids that are regurgitated and can adhere to the oral surfaces (Young et al., 2020), and thus more likely to be sampled through buccal swabs. Ruminal microorganisms can form biofilm and adhere to the oral surfaces depending on the salivary flow, and thus these are washed out at different rates, which might explain the negative association. Oral models included variables related to biofilm formation. Biofilms might be an interesting factor to consider in future studies to elucidate which microorganisms are truly FI predictors. Certain true oral microorganisms (i.e., *Anaerovibrio* genus) showed correlations with FI. Modulation of the oral pH by saliva (Maekawa et al., 2002b) may offer a partial explanation for these associations.

Information regarding the best validation method for extremely small data set with $n \ll p$ is scarce and most is focused on classification tasks. Well-designed cross validation is an almost unbiased validation method. However, it is challenging to implement in small data sets, leading to high variability and underrepresented circumstances. Therefore, bootstrap validation methods have been proposed for small data sets (Efron and Tibshirani, 1997; Braga-Neto and Dougherty, 2004). Bootstrap estimates may incorporate some bias, but some methods have been previously proposed to adjust for this bias. The 0.632 and 0.632+ bootstrap estimates have demonstrated to be satisfactory when extremely small data sets are used (Efron and Tibshirani, 1997) although it still shows some bias for $n < 20$ and might provide anti-conservative estimates (Jiang and Simon, 2007). Moreover, these methods may further under-estimate prediction error when data overfits, which is our case. Conservative estimates for the prediction error might be desirable to avoid false-positive reports in extremely small data sets (Jiang and Simon, 2007). For this reason, we decided to use out-of-bag validation considering the bias-variance trade-off of this method when $n < 20$ (Jiang and Simon, 2007). Despite the small sample size and large variability within out-of-bag estimates, it is worth noting that models achieved remarkable predictive accuracy, which suggests the potential use of both the ruminal and oral microbiome in estimating FI.

In summary, models that incorporate the oral microbiome can predict FI, given our predictions error and the large phenotypical correlations reported between the abundance of biofilm associated microorganisms and FI. These models may be improved in the future using larger data sets. Significant unclassified microorganisms may be uncovered in the future with better reference genomes. The validity of these results across different environments needs to be addressed.

**CONCLUSIONS**

Runen and oral microbiota composition are potential proxies of feed intake in dairy cattle, with predictive correlation as large as 0.48 ± 0.30, and RMSE as low as 3.46 ± 0.83 kg of DM/d depending on the microorganisms included in the model. Despite the large uncertainty from the cross-validation results, the bootstrapped distribution of the predictive accuracy was largely above zero, and more than half of the replicates above 0.50, when oral microbiota was used. The
oral microbiota showed a larger predictive ability than the rumen microbiota, with the advantage that it is easier to collect at large scale. Models using the oral microbiome selected different microbial groups and biological pathways to predict FI, compared with models using rumen microbiome. Fiber-associated microorganisms in the rumen might be acceptable proxies for feed intake, whereas biofilm formers, such as *Gammaproteobacteria* or *Bacteroidia* classes, seem appropriate proxies for feed intake in oral samples. This study proposes oral microbiome as a potential proxy to predict FI in dairy cattle. Although further research is needed as there are yet many unknowns on how to extrapolate these results to a broader population. Validations with larger sample sizes and independent populations are encouraged.

**ACKNOWLEDGMENTS**

C.N. Marcos received a Margarita Salas grant from the Ministerio de Universidades of the Spanish Government (RD 289/2021). We are thankful for the computational support provided by the High-Performance Computing Centre in Galicia (CESGA, Spain).

**REFERENCES**


**Figure 6.** Mean relative abundance (RA; %) of biofilm-associated KEGGs for each classified class sorted by its total RA in decreasing order. C. *Candidatus*.
Figure 7. Density plots of the 10,000 bootstrapped Spearman correlation coefficients (\( \rho \)) between the relative abundance (RA; %) of (a) **Gammaproteobacteria**, (b) **Negativicutes**, (c) **Mollicutes**, (d) **Betaproteobacteria**, and (e) **Spirochaeta** and feed intake (FI; kg of DM/d) in oral samples sampled after 42 d (OS42). Taxonomical features names in green were found overabundant in rumen samples (RS) compared with OS42. Taxonomical features names in blue were found overabundant in OS42 compared with RS. Density plots are sorted by the average \( \rho \) value in OS42 in increasing order. U.: unclassified.


ORCIDS

C. N. Marcos https://orcid.org/0000-0002-1656-9117
A. Bach https://orcid.org/0000-0001-6804-2002
M. Gutiérrez-Rivas https://orcid.org/0000-0001-9752-7498
O. González-Recio https://orcid.org/0000-0002-9106-4063