# Integrating heterogeneous data to predict lamb feed efficiency

Quentin Le Graverand

# THÈSE

**En vue de l'obtention du**

# DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

**Délivré par :**

Institut National Polytechnique de Toulouse (Toulouse INP)

**Discipline ou spécialité :**

Infectiologie, Physipoathologie, Toxicologie, Génétique et Nutrition

---

**Présentée et soutenue par :**

M. QUENTIN LE GRAVERAND

le jeudi 28 septembre 2023

**Titre :**

Intégration de données hétérogènes pour prédire l'efficience alimentaire des agneaux

---

**Ecole doctorale :**

Sciences Ecologiques, Vétérinaires, Agronomiques et Bioingénieries (SEVAB)

**Unité de recherche :**

Génétique, Physiologie et Systèmes d'Elevage (GENPHYSE)

**Directeur(s) de Thèse :**

MME CHRISTEL MARIE-ETANCELIN

**Rapporteurs :**

MME ELLY NAVAJAS, INIA MONTEVIDEO

M. OSCAR GONZALEZ-RECIO, INIA MADRID

**Membre(s) du jury :**

MME CATHERINE LARZUL, INRA TOULOUSE, Présidente

M. MAHENDRA MARIADASSOU, INRAE JOUY-EN-JOSAS, Membre

MME CHRISTEL MARIE-ETANCELIN, INRA TOULOUSE, Membre

MME FLAVIE TORTEREAU, INRA TOULOUSE, Membre

# Integrating heterogeneous data to predict lamb feed efficiency

This work was funded thanks to the partners of the 2018 Joint Call of the Cofund ERA-Nets FACCE ERA-GAS (Grant N° 696356), ICT-AGRI 2 (Grant N° 618123) and SusAn (Grant N° 696231), part of the GrassToGas project. This work also received funding from the European Union's Horizon 2020 research and innovation program (Grant N° 772787), as a part of the SMARTER project.

# Acknowledgements

First, I want to thank Christel Marie-Etancelin and Flavie Tortereau! I know … everyone starts by thanking their supervisors… But, Flavie and Christel deserve my gratitude and so much more. They kick-started the project and guided me throughout the thesis, despite the COVID-19 epidemic and all other hardships. I value all the things I learned from you: from drawing trees to writing a manuscript. Your input, your patience and your support were greatly appreciated. Time went by too fast!

All my work would not have been possible without the help of many other colleagues. Thank you, Annabelle Meynadier and Jean-Louis Weisbecker for all your contributions, discussions and smiles! I want to thank the P3R colleagues, who took care of animals and played an essential role in all experiments. I also thank the technician staff of the ENVT who contributed to analyses.

Thank you, Kim-Anh Lê Cao, for welcoming me in Melbourne, Australia. Discovering and discussing mixOmics with you was a wonderful opportunity! I will also keep fond memories of the lab excursion and the cheese sampling break!

I want to thank all members of the thesis committee: Rachel Rupp, Pauline Martin, Anne Ricard, Donato Andueza, and Sébastien Déjean. Thank you for providing great advice!

I wish to acknowledge all members of the thesis jury: Elly Ana Navajas, Oscar González-Recio, Catherine Larzul and Mahendra Mariadassou. Your input and all discussions were extremely valuable!

I want to thank all the friends who supported me through these crazy years! GenPhySE is one of these wonderful labs: packed with great scientists and friends. Finally, thanks Mom and Dad for pet-sitting my wonderful menagerie, while I was playing with data in Australia.

# Contents

# List of figures

# List of tables

# List of abbreviations

- **ADEI**: Average daily energy intake
- **ADFI**: Average daily feed intake
- **ADG**: Average daily gain (of weight)
- **AUROC**: Area under the receiver operating characteristic curve
- **BER**: Balanced error rate
- **BFT**: Back fat thickness
- **BW**: body weight
- **$BW^{0.75}$**: metabolic weight
- **C-diet**: Concentrate diet (100% concentrates)
- **$CH_4$**: Methane
- **CLR**: Centered logratio
- **$CO_2$**: Carbon dioxide
- **$CO_2$-eq**: Carbon dioxide equivalent
- **DMI**: Dry matter intake
- **DNA**: Deoxyribonucleic acid
- **FCR**: Feed conversion ratio
- **GWAS**: Genome-wide association study
- **$h^2$**: Heritability
- **$h^2_{maternal}$**: Maternal heritability
- **LCFA**: Long-chain fatty acid
- **MD**: Muscle depth (*longissimus dorsi*)
- **M-diet**: Mixed diet (around 67% forages and 33% concentrates)
- **MINT.sPLSDA**: Multivariate integrative sparse partial least squares discriminant analysis
- **MINT.sPLSR**: Multivariate integrative sparse partial least squares regression
- **N-integration**: Integration of different sets of variables
- **NIR**: Near-infrared
- **NIRS**: Near-infrared spectroscopy
- **NMR**: Nuclear magnetic resonance
- **NP-integration:** Integration of different sets of samples and variables
- **OTU**: Operational taxonomic unit
- **PCA**: Principal component analysis
- **PEG**: Partial efficiency of growth
- **pH**: acidy (potential of hydrogen)
- **P-integration**: Integration of different sets of samples
- **$R^2$**: Coefficient of determination
- **REI**: Residual energy intake
- **RFI-**: Efficient line (sheep selected for a low residual feed intake)
- **RFI**: Residual feed intake
- **RFI+**: Inefficient line (sheep selected for a high residual feed intake)
- **RFR**: RandomForest regression
- **RG**: Residual gain (of weight)
- **$r_{genetic}$**: Genetic correlation
- **RNA**: Ribonucleic acid
- **$r_{phenotypic}$**: Phenotypic correlation
- **rRNA**: Ribosomal ribonucleic acid
- **SNP**: single nucleotide polymorphism
- **sPLS**: Sparse partial least squares
- **sPLSDA**: Sparse partial least squares discriminant analysis
- **sPLSR**: Sparse partial least squares regression
- **SVR**: Support vector regression
- **VFA**: Volatile fatty acid

# Abstract

The sheep industry has to face many challenges to become more sustainable. Enhancing feed efficiency may reduce the environmental footprint of the industry, improve the income of farmers and mitigate the feed-food competition. However, feed efficiency is scarcely selected in meat sheep because measuring feed intake individually is expensive. One solution could be to predict feed efficiency or intake. Studying omics may highlight proxies of feed efficiency, which would ease its selection.

The present thesis focused on the prediction of feed efficiency in meat sheep. The aim was to predict feed efficiency and intake from fixed effects, body weight, pedigree and various omics. Data were collected in Romane lambs between 2018 and 2020. Lambs were part of two divergent lines selected for residual feed intake. Production traits and the rumen microbiota were analyzed both under a concentrate diet and a mixed diet including forages, later. Host genomics, the rumen lipidome, faecal near-infrared spectra, rumen and plasma metabolomes were analyzed under a concentrate diet only.

First, principal component analyzes evidenced the main variation patterns of potential predictors of feed efficiency. Pedigree and genomics highlighted the genetic population structure. The rumen microbiota composition, faecal spectra and metabolome were influenced by the environment (*i.e.* the year or period of phenotyping). Second, the rumen microbiota could not reliably predict feed intake when the reference and target populations were raised a different year (correlations between predicted and real phenotypes ≤ 0.35). Third, fixed effects and body weight could predict sheep production traits more accurately or as well as microbiota data. Fourth, studying multiple omics and years highlighted that the best predictors were: the host genetics for residual feed intake (0.54); fixed effects and body weight for feed conversion ratio (0.48) as well as feed intake (0.85). On the opposite, rumen data did not predict well feed efficiency and intake traits. Finally, the integration of heterogeneous datasets (omics and non-omics) was carried out by weight averaging or regressing the predictions of models previously fitted on a single set of predictors. Integrating data significantly increased the prediction accuracy of feed efficiency (≤ 0.59) and intake (≤ 0.88).

The thesis opens up new perspectives to mine omics data and identify feed efficiency predictors. More research is needed to integrate and understand interactions between omics, which may improve our understanding of sheep feed efficiency.

**Keywords**: feed efficiency, omics, data integration, proxies, genetics, sheep.

# Thesis introduction

On the opposite of monogastrics - such as humans, pigs, or poultry - ruminants may digest cellulose thanks to their rumen microbiota (Russell et al., 2009). Thus, ruminants may convert inedible resources for humans into valuable products. Compared to cattle, sheep are often raised on lower quality feed. Thus, sheep waste less human-edible resources for the same meat output, than cattle (Mottet et al., 2017). Many sheep (*Ovis aries*) breeds were created and provide meat, milk, wool and skin products since the sheep domestication around 11,000 years ago (Zeder, 2008; Mazinani and Rude, 2020). Nonetheless, the present thesis will focus on meat sheep production mainly.

The meat sheep industry has to face environmental, economic and societal challenges. With meat sheep raised in Western Europe, the greenhouse gas emission intensity is equal to 109.1 kilograms of equivalent carbon dioxide per kilogram of produced protein (FAO, 2017). Around 55.5% of these greenhouse gas emissions are associated with enteric fermentation and 36.0% to feed production (FAO, 2017). It highlights the importance of animal nutrition to tackle environmental impacts. Feed is also highly important to secure the income of farmers rearing meat sheep: feeding costs may represent around 64% of operational costs (Benoit et al., 2019). Then, ruminant meat production might contribute to the feed-food competition between animals and humans: on average 2.8 kg of human-edible feed are used to produce 1 kg of ruminant meat worldwide (Mottet et al., 2017). These burning issues illustrate why the production of meat sheep has to become more sustainable. To improve the production sustainability, the present thesis will examine one solution: increase feed efficiency.

Selecting feed-efficient sheep would decrease the needs for feed resources without decreasing the production (Koch et al., 1963). Thus, selection for feed efficiency would help tackle sustainability issues. Several studies evidenced that feed efficiency could be selected since they proved that feed efficiency is partially determined by sheep genetics (Mucha et al., 2022). However, sheep are scarcely selected for feed efficiency because it would require expensive measurements of individual feed intakes (Snowder and Vleck, 2003). Predicting feed efficiency instead of measuring it, could ease the selection of meat sheep for efficiency. Therefore, the present thesis does not aim to characterize feed efficiency, it aims to predict it.

Omics, such as metabarcoding and metabolomics, were previously assessed to identify feed efficiency biomarkers in cattle and sheep. For instance, the rumen microbiota was sequenced for the 16S ribosomal RNA gene (Ellison et al., 2019; Marie-Etancelin et al., 2021), or the 18S (Zhang et al., 2020; Clemmons et al., 2021) to identify microorganisms associated with feed efficiency. Blood and ruminal metabolite biomarkers were also cited for feed efficiency (Artegoitia et al., 2017; Goldansaz et al., 2020; Touitou et al., 2022). Faecal near-infrared spectroscopy was proposed to predict voluntary intake under different diets (Andueza et al., 2017). **The predictive ability of omics, pedigree, fixed effects and body weight is assessed in the present thesis. The goal is to identify the best predictors of feed efficiency.**

Improving feed efficiency of meat sheep is under international scrutiny. The SMARTER project aims at increasing the resilience and efficiency of small ruminants through genetic selection, across 13 countries ('SMARTER', 2018). The GrassToGas project aims at decreasing the greenhouse gas emissions of ruminants, across 7 countries ('GrassToGas', 2019). One task of GrasstoGas is to examine how improving feed efficiency could affect greenhouse gas emissions. The thesis work is a part of both SMARTER and GrassToGas projects.

To highlight why predicting feed efficiency is promising, Chapter 1 will first define feed efficiency in meat sheep. Then, discussing drivers of feed efficiency and omics will highlight plausible predictors. Finally, the review will focus on statistical tools to ascertain the predictive ability of predictors.

Chapter 2 will detail the material and methods used throughout the thesis. The chapter will describe which Romane sheep population was studied and how sheep phenotypes were gathered. Protocols used to collect omics data are detailed, as well as the statistical analyzes.

Exploratory analyzes will be featured in Chapter 3. Two questions will be explored: (1) what are the range and variation patterns of lamb production traits and omics data; (2) has divergently selecting for feed efficiency altered traits and omic profiles?

Chapter 4 will assess whether lamb feed intake can be predicted from rumen prokaryotes abundances, when the training and testing populations are raised different years. The

second goal is to assess whether estimated breeding values of predicted intake are close to breeding values of real intake.

Compared to the previous chapter, more traits are predicted from microbiota data in Chapter 5. Prediction accuracies of feed efficiency, intake, growth and body composition traits were assessed under a concentrate diet and then, a mixed diet. Furthermore, prokaryotes and eukaryotes abundances are investigated thanks to 16S and 18S metabarcoding.

Chapter 6 will focus on data integration of multiple omics to predict lamb feed efficiency and intake. The first goal is to identify the best predictors (genomics, metabolomics, lipidomics, phenomics or the microbiota) by sampling different biological matrices (the blood plasma, rumen fluid and faeces). The second goal is to assess if integrating data with different omics collected over 3 years improved the prediction accuracies.

Finally, the thesis results will be discussed in Chapter 7. Consequences of divergent selection for feed efficiency will be pondered. Then, the best and worst proxies of feed efficiency and intake will be discussed. The chapter will also identify perspectives of data integration to predict and select animal traits.

# Chapter 1

## Review: Assessing and predicting feed efficiency in meat sheep

Selecting for feed efficiency is promising when it comes to improving the sustainability of meat sheep production. However, feed efficiency is rarely selected in sheep since recording feed intake is challenging. To understand these challenges, feed efficiency will be defined. Then, omics may reflect biological processes underlying feed efficiency and the review will highlight plausible omic predictors. Finally, this chapter will address statistical tools which may be convenient for the prediction of feed efficiency.

# I – Assessing feed efficiency in growing sheep

The first subsection will define feed efficiency criteria in growing meat sheep. Associations between feed efficiency and other traits will be discussed based on the ruminant literature, if references in meat sheep lack. Finally, feed efficiency drivers will be reviewed similarly.

## A. What is feed efficiency?

As reviewed by Berry and Crowley (2013), a multitude of criteria were proposed to quantify feed efficiency in growing animals, raised for meat production. These criteria differ in their definitions, pros and cons. Nonetheless, all subsequent criteria require several phenotypic measures over one period of time.

### A.1 The diversity of feed efficiency criteria in growing animals

#### Indirect criteria

When feed efficiency criteria are discussed, some indirect traits are regularly mentioned. The relative growth rate can be estimated from the difference between the logarithms of the start and end body weights, divided by the test duration and multiplied by 100 (Fitzhugh and Taylor, 1971). Kleiber's ratio is estimated thanks to the average daily gain (ADG) divided by the metabolic weight - metabolic weight being defined as the body weight at the power of 0.75 (Kleiber, 1961).

These indirect criteria are not rigorously feed efficiency criteria: they rely upon the comparison of growth traits without considering feed intakes (Berry and Crowley, 2013). Therefore, indirect criteria will not be discussed subsequently.

**Direct criteria, ratios**

In animal nutrition, feed requirements are traditionally classified in two categories: maintenance requirements or production requirements. Maintenance requirements encompass needs for thermoregulation, basal metabolism and physical activity (Gonzalez et al., 2018). On the other hand, production requirements mainly correspond to tissue accretion in growing meat lambs. Thus, production requirements can be estimated from weight gain.

Partial efficiency of growth (PEG) denotes feed efficiency as the amount of growth compared to the feed intake while accounting for maintenance requirements (Kellner, 1909):

$$\mathbf{PEG} = \frac{\mathrm{ADG}}{\mathrm{ADFI}_{\mathrm{total}} - \mathrm{ADFI}_{\mathrm{maintenance}}}$$

Where $\mathrm{ADG}$ stands for average daily gain (in body weight). $\mathrm{ADFI}_{\mathrm{total}}$ stands for the total average daily feed intake, while $\mathrm{ADFI}_{\mathrm{maintenance}}$ stands for the amount required to cover maintenance needs. $\mathrm{ADFI}_{\mathrm{maintenance}}$ can be estimated from the metabolic weight, a proxy of maintenance requirements in growing lambs (Garrett et al., 1959). Feed efficient animals have high PEG phenotypes.

Feed conversion ratio (FCR) denotes feed efficiency as the amount of feed needed to gain one unit of weight (Brody, 1945):

$$\mathbf{FCR} = \frac{\mathrm{ADFI}}{\mathrm{ADG}}$$

Feed efficient animals low FCR phenotypes.

**Direct criteria, residuals**

Residual Feed Intake (RFI) denotes feed efficiency as the difference between real feed intake and expected feed intake - based on the animal needs. To estimate expected feed intake (*i.e.* animal needs), one can regress feed intakes over energy sinks as proposed by Byerly (1941) in laying hens, and Koch et al. (1963) in cattle. Few other authors use a pre-established equation to estimate needs, and compute the difference between actual and expected intake. Most authors use regression models including at least the metabolic

weight and weight gain, while some advocate for the inclusion of body composition indicators (Berry and Crowley, 2013). In growing meat sheep, the following equation will be implemented throughout the thesis (adaptation from Tortereau et al., 2020):

$$\mathrm{ADFI} = \mu + \beta_1\,\mathrm{ADG} + \beta_2\,\mathrm{MD} + \beta_3\,\mathrm{BFT} + \beta_4\,\mathrm{final\,BW}^{0.75} + \mathbf{RFI}$$

Where $\mu$ stands for the mean $\mathrm{ADFI}$. $\beta_1$ to $\beta_3$ respectively stand for the effects associated with $\mathrm{ADG}$, muscle depth ($\mathrm{MD}$) and back fat thickness ($\mathrm{BFT}$). These effects account for production requirements in lambs gaining weight, according to their body composition. $\beta_4$ denotes the effect of the metabolic weight ($\mathrm{final\,BW}^{0.75}$) measured at the end of the trial, to account for maintenance requirements. $\mathrm{RFI}$ corresponds to the regression residuals, therefore phenotypes are centered around 0. Moreover, negative values are obtained for feed efficient animals, and positive values for inefficient ones. Computing RFI from tables does not ensure similar distribution properties.

Residual gain (RG) denotes feed efficiency as the difference between real weight gain and expected gain – based on feed intake and metabolic weight (Koch et al., 1963).

$$\mathrm{ADG} = \mu + \beta_1\,\mathrm{ADFI} + \beta_2\mathrm{BW}^{0.75} + \mathbf{RG}$$

$\mathrm{RG}$ corresponds to the regression residuals. Thus, feed efficient animals have positive RG values, against negative values for inefficient ones.

In 2012, Berry and Crowley proposed a new efficiency criterion for growing cattle: the residual intake and body weight gain, which is a linear combination of RFI and RG.

**Note:** Feed efficiency traits are not strictly equivalent even if they are related. For instance, FCR and RFI are phenotypically and genetically correlated in Romane lambs ($r_{phenotypic}$=0.49±0.03, $r_{genetic}$=0.65±0.12) (Tortereau et al., 2020).

**Refining direct criteria of feed efficiency**

Previous direct feed efficiency criteria can be tweaked to account for feed differences. To do so, the animal intake can be expressed according to feed properties rather than the ingested mass. Dry matter intake (DMI) is useful to account for differences in feed moistures. Similarly, energy intakes can be used to account for differences in feed energy densities. Energy density of feeds can be estimated from their chemical composition and

tables. Energy is traditionally partitioned in different levels (Figure 1, Inra et al., 2018) but the thesis will focus on net energy. Net energy is the difference between the ingested amount of gross energy and the energy lost in faeces, urine, respiratory gases and heat.

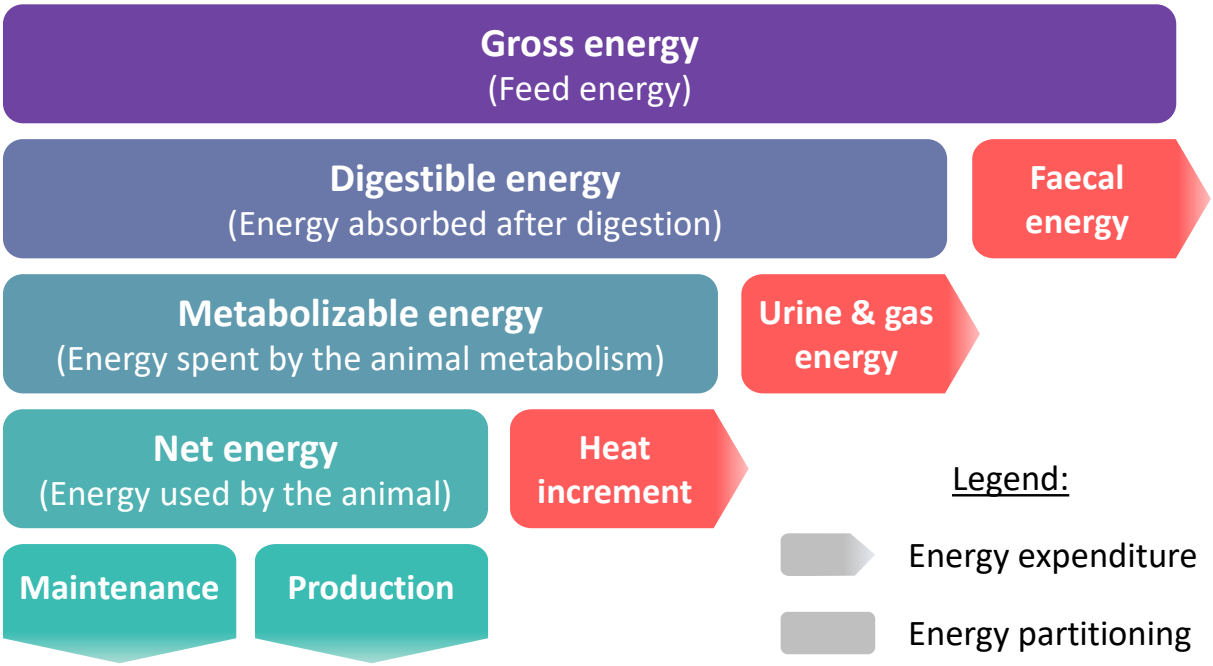



**Figure 1: Energy partioning in animal nutrition**

Feed efficiency criteria can also be tweaked to account for dynamics across time. Most criteria are computed from a few trait values: either averages, or punctual observations. However, feed efficiency criteria could be computed from times series too (Martin et al., 2021b). Leveraging repeated measures would improve the estimation of feed efficiency components: it would better account for the dynamics of efficiency components across time. However, a trade-off must be found between the number of repeated numbers and the trial duration: short disruptions may introduce too much noise if efficiency is estimated over short periods of time.

## A.2 Challenges in recording feed intakes

All direct feed efficiency criteria – ratios or residuals - present the same challenge: they require individual records of feed intake over period of 6-8 weeks on average. In beef, a trial over 42 days was proposed as an adequate compromise between an accurate estimation of the regular intake and costs (Manafiazar et al., 2017). In Romane sheep, trials of six weeks are already implemented in the National Breeding program to index growth and conformation traits in growing male lambs (Tiphine et al., 2011). Indoors, individual intake

recording often relies on expensive automatic feeders. Cost of infrastructures may explain why feed efficiency is seldom selected (Snowder and Vleck, 2003; Hayes et al., 2013).

More investments can be needed when animals are not fed a single standardized feed – such as a pelleted feed. Most ruminants are fed with mixed diets, sooner or later. Then, it would be justified to record individual intakes per feed: ruminants can sort mixed feeds (Cooper et al., 1995; Duncan and Young, 2002). Feed resources may even be more heterogeneous when animals graze because feed resources vary in quality and quantity across time and space (Mialon et al., 2022). Smith et al. (2021) reviewed the methods allowing the study of feed intake by grazing animals: there are plenty of approaches ranging from manual measures, to digital technologies and markers which differ in cost, practicality and accuracy.

## A.3 Pros and cons of feed efficiency criteria

### Direct criteria, ratios

Feed efficiency ratio criteria, such as PEG and FCR, are easily interpretable. For instance, FCR denotes the amount of feed required to produce one kilogram of body weight. Moreover, most ratio criteria can be computed on an animal-by-animal basis, without inferring parameters on a population basis. Thus, most ratio traits can be easily compared between different populations (Berry and Crowley, 2013). FCR is commonly used in research with individual records, or in farms based on pen average intakes.

However, one disadvantage is the strong correlation between ratio traits and the production traits used as numerators or denominators. In fattening Romane sheep, FCR and ADG are phenotypically and genetically correlated ($r_{phenotypic}$=-0.77±0.01; $r_{genetic}$=-0.77±0.09) (Tortereau et al., 2020). FCR and ADFI are also correlated ($r_{phenotypic}$=-0.27±0.03; $r_{genetic}$=-0.10±0.21) (Tortereau et al., 2020). Therefore, in animal breeding, one important disadvantage of ratio traits is the difficulty in anticipating responses to selection: numerator and denominator components can be under different selection pressures (Gunsett, 1984). For instance, FCR can be improved by increasing growth, by decreasing feed intake, or both.

**Direct criteria, residuals**

When feed efficiency criteria are computed via least-squares regressions, regression coefficients are population-specific. Residual phenotypes such as RFI and RG cannot be compared directly between studies.

However, the use of regressions has one virtue: residual efficiency traits are phenotypically independent from the regression explaining variables (Berry and Crowley, 2013). In Romane sheep for example, the estimated phenotypic correlation between RFI and ADG is equal to 0.02±0.04. Thus, using production traits as regressors allows a fair comparison between animals with different production levels (Knott et al., 2008). When ranking animals for feed efficiency, trials are usually set between two calendar dates and not between two ages. So, animals will likely vary in production levels and also in maturity during a trial. In practice, breeding companies set protocols to minimize age differences during the genetic evaluation.

Residual traits are correlated to the regressed variable: for instance, RFI is phenotypically and genetically correlated with feed intake ($r_{phenotypic}$=0.58, $r_{genetic}$=0.59±0.13), while RG is correlated to ADG ($r_{phenotypic}$=0.70, $r_{genetic}$=0.82±0.05) in growing cattle (Berry and Crowley, 2012). Selecting for RFI will likely decrease feed intake while maintaining production, when selecting for RG should increase growth while maintaining feed intake.

Nonetheless, independence between feed efficiency and the regressed production traits does not ensure a genetic independence. For instance, a moderate genetic correlation exists between RFI and muscle depth in Romane fattening lambs (-0.30±0.15) (Tortereau et al., 2020). A regression based on genetic estimates can ensure genetic independence (Kennedy et al., 1993).

RFI has an increasing popularity in the ruminant scientific literature (Berry and Crowley, 2013). However, Berry and Crowley (2012) hypothesized that RFI is not widely adopted by producers because they are afraid that animals with good RFIs may have a low growth rate associated with a low intake. On the other hand, Snowder and Vleck (2003) suggested that breeders may prefer a direct selection for ADG to get a higher economic response to selection, compared to a direct selection for RFI. They also highlighted that it may depend

on the scenario since selection for RFI becomes more interesting when feed costs increase while the lamb prices decrease.

# B. How is feed efficiency associated to other traits?

Selecting for feed efficiency was proposed to mitigate sustainability issues of meat sheep production. However, meat sheep may already be selected for other traits having an economic interest, such as: growth, body composition, body conformation, meat qualities and maternal abilities (Tiphine et al., 2011). Furthermore, new traits might be selected in the future to cope with environmental issues, such as greenhouse gas emissions. Thus, studies assessing associations between feed efficiency and other traits are necessary to anticipate responses to selection for feed efficiency. The review will focus mainly on associations with RFI and FCR.

## B.1 Production and body composition traits

As discussed in the Chapter 1 subsection A.2, FCR is correlated to growth by construction, since its denominator is ADG. Contrastingly, RFI is phenotypically independent from growth and liveweight when feed intake is regressed over metabolic weight and ADG (Zhang et al., 2017; Tortereau et al., 2020).

Results about associations between body composition traits and RFI may be conflicting when RFI is defined by different models. When RFI is computed without regressing intakes over body composition, feed efficiency is phenotypically associated with a decreased eye muscle area and a decreased back fat thickness in Hu lambs (Zhang et al., 2017). When RFI is computed by regressing intakes over the *longissimus dorsi* depth and back fat thickness, feed efficiency is not phenotypically associated with body composition in Romane lambs ($r_{phenotypic}$=0.00±0.04 with muscle depth, $r_{phenotypic}$=0.02±0.04 with back fat thickness) (Tortereau et al., 2020). However, muscle depth might be genetically negatively correlated to RFI in Romane lambs ($r_{genetic}$=-0.30 ± 0.15) (Tortereau et al., 2020). Thus, selection consequences over body composition will depend on the model used to compute RFI: the model may account for body composition or not; it may regress phenotypic values or genetic values.

In Pelibuey lambs, RFI was not phenotypically associated with carcass characteristics, chemical composition and meat quality - except for cooking loss with higher percentages found for efficient animals (Arce-Recinos et al., 2022).

Results are conflicting about visceral organs. Meyer et al. (2015) suggested that diets had a greater effect over organ masses than feed efficiency phenotypes, but efficient animals tended to have bigger spleens and pancreas (p-value=0.09). Another study had discrepant results and highlighted associations between RFI classes and other organs: feed efficient sheep had bigger testis, a lighter intestinal tract, a decreased total stomach weight, a lighter liver and lighter lungs (Zhang et al., 2017). In the Romane breed, the most efficient lambs tended to have larger rumens after three generations of divergent selection for feed efficiency (p-value=0.09) (Conington et al., 2022).

## B.2 Health and reproduction traits

According to the resources allocation theory, improving feed efficiency may affect health, reproduction traits or responses to environmental stressors (Gilbert et al., 2017; Douhard et al., 2021). The assumption is based upon the hypothesis that resources are limited and that trade-offs may exist between different expenditures - such as parasite resistance. In animal breeding contexts, it may be difficult to highlight trade-offs between resources allocations when animals are selected with few nutrient restrictions (Douhard et al., 2021).

To our knowledge, references about associations between sheep feed efficiency and health traits mainly focused on parasite resistance. In Corriedale lambs, RFI phenotypes were neither significantly associated with nematode resistance, nor lines genetically selected for resistance (Ferreira et al., 2021; Navajas et al., 2022). In Romane lambs, no association was found between nematode resistance and lines selected for RFI (Douhard et al., 2022a). However a synergy was found between feed efficiency and resistance in lines genetically selected for parasite resistance (Douhard et al., 2022b). Thus, results are conflicting between studies, but they do not accredit the hypothesis of a trade-off as expected with resources allocation. In meat sheep, there is also a lack of references assessing relationships between feed efficiency, reproductive and maternal traits. Kenny et al. (2018) reviewed the cattle literature and highlighted conflicting results about the relationship between bull fertility and feed efficiency. Studies in cattle, suggested that RFI is not associated with calving difficulty (Basarab et al., 2011; Lawrence et al., 2013). Finally, Kenny et al. (2018)

underlined a lack of consensus about reproductive performances: papers highlighted either the presence or absence of associations between RFI and reproductive traits (rates of pregnancy, calving and weaning).

### B.3 Environmental traits

Feed efficiency was proposed as a trait to indirectly select against greenhouse gas emissions (Boadi et al., 2004). However, when farmers and experts are consulted and have to rank the most efficient and practical measures to mitigate greenhouse gas emissions, selecting for feed efficiency is not seen as one of the preferred mitigation measures (Jones et al., 2013). Furthermore, relationships between efficiency and emissions are inconsistent from one study to another. In lambs from three different breeds, no significant phenotypic correlation was found between RFI and daily methane emissions (Johnson et al., 2022; Navajas et al., 2022; Marques et al., 2022). On the opposite, Paganoni et al. (2017) found significant phenotypic correlations between RFI and daily methane, but correlations were alternatively positive in growing sheep or negative in adults. Finally, two Romane divergent lines were selected for RFI and most efficient lambs emitted more methane daily than inefficient animals (Tortereau et al., 2023).

Differences in emissions might partly be explained by rumen and microbiota differences. Romane lambs from the feed efficient line tended to have bigger rumens (Conington et al., 2022). Bigger rumens and higher retention times were previously associated with higher emissions (Goopy et al., 2014; Conington et al., 2022). One hypothesis is that higher methane emissions are associated with rumen microbial communities producing more hydrogen which is the substrate of methanogens producing methane (Kittelmann et al., 2014). Kamke et al. (2016) hypothesized that higher retention rates could be associated with slow degrading microorganisms producing more hydrogen.

Apart from greenhouse gas emissions, literature is scarce about relationships between feed efficiency and excretion in sheep. In cattle, feed efficient animals were associated with lower manure excretion, partly thanks to lower intakes (Basarab et al., 2013; Berry and Crowley, 2013). Dairy cows fed pasture excreted less nitrogen in faeces if they were feed efficient (Rius et al., 2012). Thus, these studies suggest that improving feed efficiency would reduce environmental impacts associated to manure and its degradation. It remains to check how feed efficiency is associated with urinary excretion. However, feed efficiency and

water ingestion are tied. In steers, RFI was favorably and genetically correlated with water intake ($r_{genetic}$=0.33±0.11) (Ahlberg et al., 2019).

## C. What drives feed efficiency in ruminants?

Feed efficiency traits are multifactorial: many environmental and biological factors may underly feed efficiency (Cantalapiedra-Hijar et al., 2018; Kenny et al., 2018).

Ruminants stand out from many livestock species thanks to the symbiotic relationships between the host and its microbiota in the upper gut (Morgavi et al., 2015). A superorganism including one host and its symbiont can be called a holobiont (Margulis and Fester, 1991). Thus, the host and its ruminal microbiota will be referred as an holobiont in the thesis. Next sections will address how feed efficiency is the result of the interplay between holobiont and environmental determinants (Figure 2).

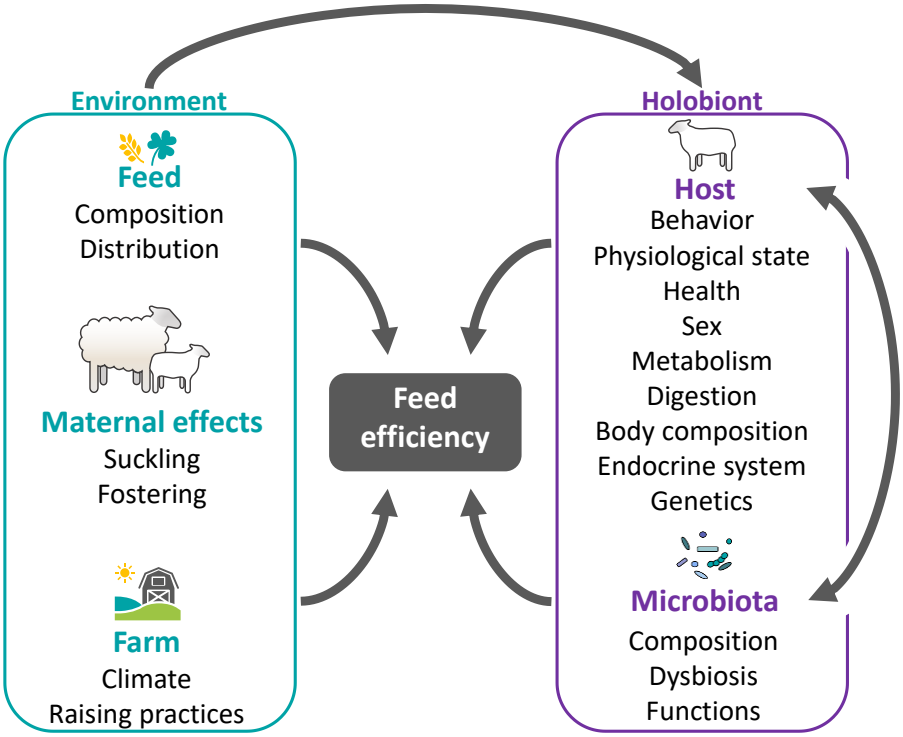



**Figure 2: Main sources of feed efficiency variations (adapted from Cantalapiedra-Hijar et al., 2018; Kenny et al., 2018)**

### C.1 External effects: maternal effects, climate and farming conditions

Feed efficiency may be partially driven by external factors. These external factors may be related to maternal non-genetic effects or environmental effects. Maternal genetic effects will be presented in paragraph C.5.

### Maternal effects

In meat sheep, maternal effects may influence the offspring feed efficiency through pregnancy and fostering. Maternal effects over offspring efficiency may persist after weaning. These effects may be evidenced through the nutrition of dams. For instance, lambs tended to be more feed-efficient when ewes were not feed restricted during gestation (Piaggio et al., 2018). Then, nutrition of beef dams during lactation may alter their offspring feed efficiency after weaning. Allocating dams to low-quality or high-quality pastures during lactation explained 2.5% of FCR variations: dams on high-quality pastures had more feed efficient offspring (Robinson et al., 2013). The interaction between allocation duration and pasture explained 3.6% of offspring RFI variations (Robinson et al., 2013). More studies are mandated to confirm and assess maternal effects, such as nutrition.

### Climate and farming conditions

Ambient conditions such as temperatures, may affect feed efficiency. For instance, heat stress decreased feed intakes and degraded feed conversion ratio in Suffolk lambs (Padua et al., 1997). High temperatures require more efforts from sheep to regulate their body temperature and heat also reduces their appetite (Marai et al., 2007). Furthermore, temperatures may also alter the digestibility of feed by bacteria (Hyder et al., 2017). Sensitivity to temperatures in sheep depends on the fleece length (Armstrong et al., 1959), and by extension shearing practices.

Farming practices can be quick and practical means to improve feed efficiency. For example, feed distribution or availability may influence feed efficiency. After feed restrictions, lambs fed *ad libitum* had an improved feed efficiency than counterparts which did not undergo a restricted diet (Butler-Hogg and Tulloh, 1982; Kamalzadeh et al., 1997). During feed restrictions, growth decreases but it rebounds higher when restrictions end. The phenomenon is known as compensatory growth.

Feed quality could also be considered as a lever. All feed efficiency criteria may not account for differences in dry matter content, energy density, digestibility, or metabolizability. Then, differences in feed dietary characteristics may sway feed efficiency calculations. A good illustration is the forage-to-concentrate ratio in diets: concentrates have a higher energy density and digestibility than forages. Studies observed that lambs fed a higher amount of concentrates were more feed-efficient than lambs receiving more forages, when FCRs were

computed from dry matter intake or digestible matter intake (Haddad and Husein, 2004; Jacques et al., 2011; Papi et al., 2011; Claffey et al., 2018). However, using concentrates to increase feed efficiency might not be sustainable. Concentrates may contain human-edible resources and contribute to the feed-food competition between animals and humans.

When the diet changes and animals age, mechanisms underlying feed efficiency may change too. For example, there were null to moderate phenotypic correlations when sheep RFI was recorded under a diet rich in concentrates, and later under a diet rich in fiber ($0.02 \leq r_{phenotypic} \leq 0.51$) (Redden et al., 2011; Ellison et al., 2022; Marie-Etancelin et al., 2023).

## C.2 Behavior: physical activity and feeding

Behavior may differ between feeding systems (pasture, individual or collective troughs): for example, grazing animal have to explore their environment to feed. Thus, it was hypothesized that physical activity may be a bigger determinant of feed efficiency when ruminants graze (Muir et al., 2020). Most studies assessed the relationship between feed efficiency and behavior indoors, with individual troughs. Two kinds of behaviors are regularly mentioned in the literature: feeding behavior and physical activity. Both behaviors represent energy expenditures and may influence feed efficiency.

Variations in physical activity may alter feed efficiency since activity is not accounted in most cases. For instance, low RFI steers tended to spend more time idle and lying (Gomes et al., 2013). Then, efficient animals would save energy since standing instead of lying would increase energy expenditures by 16-29% (Lobley, 1990). However, other authors reported no significant difference in posture activities between efficient and non-efficient cows, when tied in stalls (De La Torre et al., 2019).

Feeding contributes to physical activity through movements and postural changes. Feeding behavior encompasses several traits. Feed intake and duration can be expressed per day or visit. Both traits were genetically associated to RFI in Romane lambs: intake per visit was correlated negatively to RFI ($r_{genetic} = -0.33 \pm 0.14$) same as feed duration ($r_{genetic} = -0.22 \pm 0.17$) (Marie-Etancelin et al., 2019). These results underlined that feed efficient sheep had larger meals per visit. Besides, several studies demonstrated that feed efficient sheep ate less and visited less often feeders at the genetic and phenotypic levels (Muir et al., 2018; Marie-Etancelin et al., 2019; Sepulveda et al., 2022). Thus, efficient sheep

might save energy by moving less often to visit feeders and eating more per visit. It could decrease the physical activity required by the feeding behavior.

Finally, feeding behavior may vary between sexes and diets (Vandenheede and Bouissou, 1993; Abijaoudé et al., 2000). When sheep are offered several feeds, animals may select feeds to maintain the rumen pH and osmolality (Cooper et al., 1995). Extreme pH and osmolality can lead to metabolic diseases and alter the microbiota.

## C.3 Digestion: an interplay between the microbiota and its host

The holobiont symbiosis is particularly important for digestion. Digestibility is the ability to degrade feed into nutrients. That digestive ability may depend on host and microbiota determinisms.

### Digestibility of the holobiont

Feed efficiency and digestibility are often associated, but it might be difficult to distinguish cause-and-effect relationships. Sauvant and Nozière (2016) carried out a meta-analysis and showed that decreasing DMI results in a higher dry matter digestibility. Thus, Cantalapiedra-Hijar et al. (2018) stressed that improving feed efficiency could cause a higher dry matter digestibility by decreasing feed intake. So, changes in digestibility might be the consequence and not the cause of feed efficiency variations. However, two reviews underlined that in most studies, the dry matter digestibility did not significantly differ between the most and least efficient cattle (Cantalapiedra-Hijar et al., 2018; Kenny et al., 2018). Discrepancies between studies may arise from diet differences and level of feeding.

### The symbiosis underlying the digestion

Generally, the symbiosis is mutually beneficial to the host and its microbiota. The balance between the metabolism and immunity of the host and its symbiont is called eubiosis (Ponziani et al., 2017). We will focus on one symbiont throughout the thesis: the rumen microbiota. The rumen microbiota is a complex ecosystem encompassing: prokaryotes (*e.g.* bacteria and archaea), eukaryotes (*e.g.* protozoa and fungi), and viruses. Bacteria are abundant, diverse and considered as the most active among the rumen microorganisms (Sanjorjo et al., 2023) (Table 1). Archaea are less abundant and are mainly considered for methane production (Sanjorjo et al., 2023). Protozoa constitute a large part of the rumen microbial biomass (up to 60%), but their contribution to the host nutrition is still under

discussion (Newbold et al., 2015). Protozoa may have their own symbionts: associated archaea can be found inside or outside protozoa. Similarly fungi are not the most abundant microorganisms, however they are known for their high ability to degrade complex fibers (Sanjorjo et al., 2023). Viruses are abundant, contribute to interspecific genetic exchanges and they may lyse other microorganisms, but are rarely studied (Sanjorjo et al., 2023).

**Table 1: Abundance of microorganisms in sheep ruminal fluids**

| Microorganisms | Number of DNA copies or particles per mL of sheep rumen fluids | Studies |
|---|---|---|
| Bacteria | $10^8$ to $10^9$ | (Belanche et al., 2019)<br>(Pei et al., 2010)<br>(Stiverson et al., 2011)<br>(Yanagita et al., 2000) |
| Archaea | $10^5$ to $10^7$ | (Belanche et al., 2019)<br>(Stiverson et al., 2011) |
| Protozoa | $10^5$ to $10^8$ | (Belanche et al., 2010)<br>(Belanche et al., 2019)<br>(Lettat et al., 2012)<br>(Skillman et al., 2007) |
| Fungi | $10^4$ to $10^6$ | (Belanche et al., 2019)<br>(Lwin et al., 2011) |
| Viruses | $10^7$ to $10^{10}$ | (Klieve and Bauchop, 1988)<br>(Klieve and Swain, 1993) |

Ruminants have to rely on the microbiota fermentation to degrade the cellulose and hemicellulose -which are abundant fibers in plant cell walls (Selinger et al., 1996). The microbial fermentation produces volatile fatty acids (VFAs) as by-products. These VFAs represent sources of energy for ruminants, and they can fulfill more than 70% of ruminant energy requirements (Bergman, 1990). Rumen microorganisms also contribute to the degradation of lipids, and proteins. Microorganisms produce their own proteins and then, they are digested by the ruminant. Microbial proteins provide 60-85% of the amino acids available in the small intestine (Storm et al., 1983), where amino acids can be assimilated by the host.

On the other hand, ruminants provide an adequate habitat for the microbiota in the foregut. Around 60-70% of the digestive tract volume is associated with a fermentative activity (Parra, 1978), which denotes the importance of microbial digestion. Ruminants are called polygastrics because they have three successive forestomachs and one stomach (Figure 3).

The rumen is the first and largest out of the four components (75% of the "stomachs" volume in sheep), then feed goes through the reticulum (8%), the omasum (4%), and the abomasum (13%) (Membrive, 2016). Studies often focused on the fermentative activity in the rumen due to its large volume, and its physical properties favorable to fermentation – such as its pH, temperature, humidity, and anaerobic conditions.

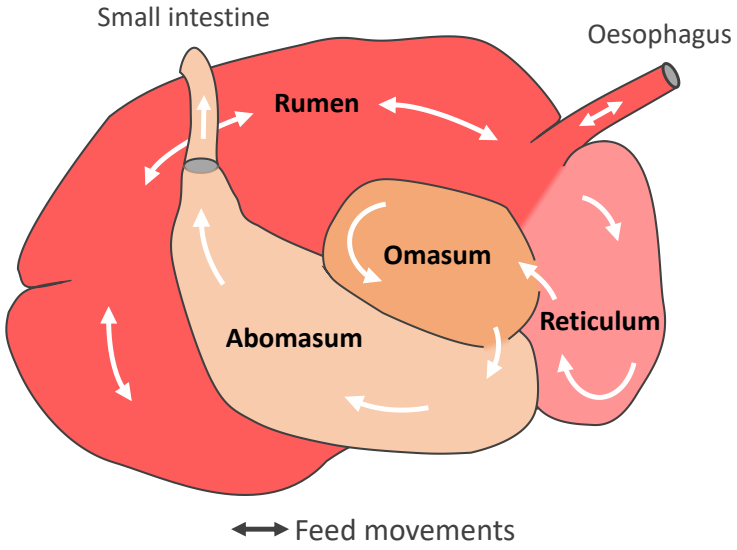



**Figure 3: Anatomy of the ruminant foregut**

In the rumen, abundances of some microorganisms were associated to feed efficiency. However, studies in cattle and sheep are conflicting: the same microbial genus could be positively and negatively associated to feed efficiency (Lovendahl et al., 2018). It suggests that members of a microbial taxa may contribute differently to digestion. On the opposite, different taxa may have the same function and contribute similarly to digestion: such taxa are functionally redundant (Weimer, 2015). In most studies, microorganisms could not be accurately affiliated at the species level, which hinders interpretations. Moreover, many species are not cultivated and their functions may be unclear. Nonetheless, associations were found between feed efficiency and bacteria contributing to the digestion of plant cell walls, protein, and aromatic compounds in the rumen of sheep and cattle (Ellison et al., 2019; McGovern et al., 2020; Marie-Etancelin et al., 2021). Several studies also suggested that microorganisms producing specific volatile fatty acids might result in lower energy losses – through lower methane emissions.

### Production of methane by the symbiont

Hydrogen ions are produced by the microbial fermentation and have to be eliminated: an excess may acidify the rumen and impair its function (Domingues Millen et al., 2016). In the

rumen, methane is produced by archaea which contributes to the elimination of hydrogen ions (Domingues Millen et al., 2016). Methane emissions represent an energy loss: from 2 to 12% of the ingested energy, depending on the diet (Johnson and Johnson, 1995).

Fermentation may result in the production of different VFA profiles. Producing propionate from fiber does not lead to methane production, while producing butyrate and acetate usually leads to methane production (Van Soest, 1994). However, bacteria such as the *Acetitomaculum* genus can also produce acetate from $CO_2$ and $H_2$, which decreases methane emissions (Le Van et al., 1998). Thus, it was hypothesized that feed efficient sheep might lose less energy when the rumen microbiota composition produces less methane, during digestion (Ellison et al., 2017; Marie-Etancelin et al., 2021). However, studies found alternatively positive, negative or no association between feed efficiency and methane emissions (Paganoni et al., 2017; Johnson et al., 2022; Navajas et al., 2022; Marques et al., 2022). Differences in diets may partially explain the discrepancies in greenhouse gas emissions: diets affect both the rumen microbiota composition and the fermentation profile. Ellison et al. (2017) observed that more propionate is produced under concentrate-based diets, while more acetate is produced under forage-based diets. Further investigation is needed to elucidate how the ruminal microbiota, feed efficiency, the diet and greenhouse gas emissions are related together.

**Dysbiosis**

The microbiota functions, composition or distribution may shift and create an imbalance. The imbalance is called a dysbiosis and it can be detrimental for the host (Stecher et al., 2013). It may be caused by feeding practices, such as the distribution of a diet with a high energy density causing an acidosis (Khafipour et al., 2009).

In cattle and sheep, it has been hypothesized that dysbiosis may be associated to feed efficiency (Khiaosa-ard and Zebeli, 2014; Perea et al., 2017). First, dysbiosis may alter the digestion. Then, dysbiosis may also enable the growth of pathogens and the release of toxins, which would induce an immune reaction from the host (Khiaosa-ard and Zebeli, 2014). It would increase the amount of the host energy allocated to non-productive processes. However, experiments did not yet prove a relationship between feed efficiency and dysbiosis.

## C.4 Metabolism: efficiency of maintenance and growth

If we consider growing ruminants, feed efficiency might be explained by differences in maintenance or production requirements. Holobiont physiological processes may influence those requirements such as energy production and protein metabolism.

### Energy production in mitochondria

Maintaining homeostasis and growing require energy. Mitochondria supply energy to cells, by producing adenosine triphosphate thanks to respiration. Results in steers suggested that efficient animals had a higher rate of mitochondrial respiration, thanks to more efficient electron transfers (Kolath et al., 2006). In growing lambs, it was also suggested that a higher feed efficiency is associated to an enhanced mitochondrial respiration (Rajaei Sharifabadi et al., 2012; Giráldez et al., 2021; Touitou et al., 2022).

### Protein metabolism

Proteins have to be continuously renewed to ensure maintenance, since proteins are degraded over time. The process is called protein turnover. Synthesis of proteins represent 25 to 42% of heat production, and around 19% of all energy spent by growing lambs (Millward et al., 1976; Davis et al., 1981; Gill et al., 1989). Most synthesized proteins contribute to the protein turnover: up to 94% of the proteins synthesized by growing bulls, depending on the breed and production level (Lobley, 1990).

Cantalapiedra-Hijar et al. (2018) reviewed the bovine literature and found conflicting results between protein metabolism and feed efficiency. They hypothesized that a lower protein degradation rate could improve feed efficiency. Authors also hypothesized that a higher protein synthesis rate could improve efficiency, if it leads to a higher protein retention. A lower protein degradation or a higher synthesis rate could both explain feed efficiency in growing lambs: concentrations of blood amino acids could be higher in feed efficient lambs (Giráldez et al., 2021), or lower (Touitou et al., 2022).

### Maintenance and growth requirements

In lambs weighing 30 kg, maintenance represent 62% of net energy requirements when lambs grow by 100 g per day, against 42% when they gain 400 g per day (Noziere et al., 2018). Herd and Arthur (2009) postulated that feed efficient cattle might have lower maintenance energy requirements. Literature in cattle and pigs also suggests that feed

efficient animals may have a higher efficiency of growth (Barea et al., 2010; Cantalapiedra-Hijar et al., 2018). More studies are warranted to establish that maintenance requirements and efficiency of growth are related, and how they might affect feed efficiency in ruminants.

Physiological processes, such as protein and energy production, may be associated to variations in requirements for maintenance or growth. Body composition and organ sizes may also affect requirements, depending on the age and maturity. During growth, protein deposition is more energetically efficient than fat deposition (Herd and Arthur, 2009). On the opposite, higher amounts of proteins would be associated to higher maintenance requirements in adults (Herd and Arthur, 2009).

## C.5 Genetics

Feed efficiency traits are complex traits, meaning that they are partially determined by polygenic determinisms - *i.e.* multiple genes (Goddard et al., 2016). The genome determines the potential of an animal for feed efficiency, while genetic expression ultimately affect the phenotype, in conjunction with the environment.

Differences between breeds underline that different genetic backgrounds may result in feed efficiency variations in sheep (Notter et al., 1984) and cattle (Crowley et al., 2010). Depending on the criterion, the most feed efficient breed varies (Notter et al., 1984).

A meta-analysis in meat sheep found low to moderate estimates for feed efficiency heritabilities: 0.12±0.03 for FCR, against 0.32±0.15 for RFI over 7 heritability estimates (Mucha et al., 2022). In Romane lambs, heritabilities were relatively important: 0.30±0.08 for FCR, 0.45±0.08 for RFI (Tortereau et al., 2020). It underlines that feed efficiency could be selected since it is partly inherited. For instance, lambs from efficient sires grew as much but ate 3% less concentrates than lambs from inefficient sires after one generation of divergent selection for RFI (Tortereau et al., 2020).

Dams transmit part of their genetic information to their offspring. The combination of alleles transmitted by dams and sires determine the genetic potential of offspring: the direct additive genetic effect. In meat sheep and beef, dams care for their offspring during the pre-weaning period which might affect feed efficiency too. Maternal caring might be inherited and determined by the dam genetics. Studies are discordant in growing beef: RFI could be either significantly ($h^2_{maternal}$=0.09±0.04; Crowley et al., 2010) or insignificantly associated

to maternal effects (Hoque et al., 2007). However, maternal heritability was not evidenced when RFI was analyzed in New Zealand maternal sheep (Johnson et al., 2022). More studies are warranted, since it can be difficult to disentangle herd effects from maternal effects.

Various genomic mechanisms were associated to feed efficiency, through genome-wide association studies, variant analysis and gene expressions. Cantalapiedra-Hijar et al. (2018) reviewed the beef literature and proposed several categories of biological functions. In sheep, several of these functions may also explain feed efficiency:

- Cell cycle, growth and proliferation (Cockrum et al., 2012);
- Development and morphology (Cockrum et al., 2012);
- Cell transport and signaling (Jonas et al., 2016; Yang et al., 2022);
- Protein, carbohydrate and lipid metabolisms (Giráldez et al., 2021; Zhang et al., 2021);
- Energy production (Giráldez et al., 2021);
- Immunity, inflammation, and oxidative stress (Giráldez et al., 2021; Lin et al., 2023).

Different traits may share similar genetic determinisms. Saatchi et al. (2014) suggested that a pleiotropic quantitative locus may affect feed efficiency but also production and reproduction traits, on the $6^{th}$ chromosome in cattle. It may partially explain why feed efficiency can be genetically correlated to other traits. However, further studies are needed to confirm if only one gene is determinant and which one.

Holobiont genetics may drive feed efficiency through host and microorganisms' genetics. In addition to the above-mentioned effects, host genetics may shape the ruminal microbiota composition. Heritabilities of rumen bacterial abundances varied between 0 and 0.29±0.07, with an average of 0.04±0.03 in dairy sheep (Martinez Boggio et al., 2022). QTLs were identified for a minority of ruminal bacteria in sheep, with candidate genes associated to immune reactions and metabolism (Mani et al., 2022; Martinez Boggio et al., 2022). These experiments demonstrated that a small part of the rumen microbiota was genetically determined, even if further experiments are needed to ascertain the genetic mechanisms. The same conclusion was reached in cattle, where some rumen microorganisms were heritable and associated to feed efficiency (Li et al., 2019b). Thus, feed efficiency might be driven by host genetics directly and indirectly, to a lesser extent, by its influence over the microbiota composition. The combination of host genome and its microbiota genomes is called the hologenome.

Genetic determinisms of feed efficiency may vary across different environments and ages. For instance, the genetic correlation between grower and finisher diets was moderate for RFI in steers ($r_{genetic}$=0.50± 0.48) (Durunna et al., 2011). More studies are needed to confirm results with lower standard errors in cattle and sheep. However, it suggests that one animal will be ranked differently from one environment to another, based on its genetic potential for feed efficiency.

## C.6 Relative importance of feed efficiency determinants

Finally, previous sections underlined the plurality of feed efficiency determinants. It might be difficult to pinpoint some determinants: efficiency variations might be explained by a wide-variety of determinants, but not all at once. Furthermore, feed efficiency is context-dependent: discrepancies might be due to feed efficiency criteria, breeds (Notter et al., 1984), physiological states (Jonas et al., 2016) and genotype by environment interactions (Durunna et al., 2011). Thus, the importance of feed efficiency determinants is likely context-dependent.

Nonetheless, studies attempted to assess the importance of feed efficiency determinants in cattle growing in feedlot (Herd et al., 2004, 2019; Richardson and Herd, 2004).

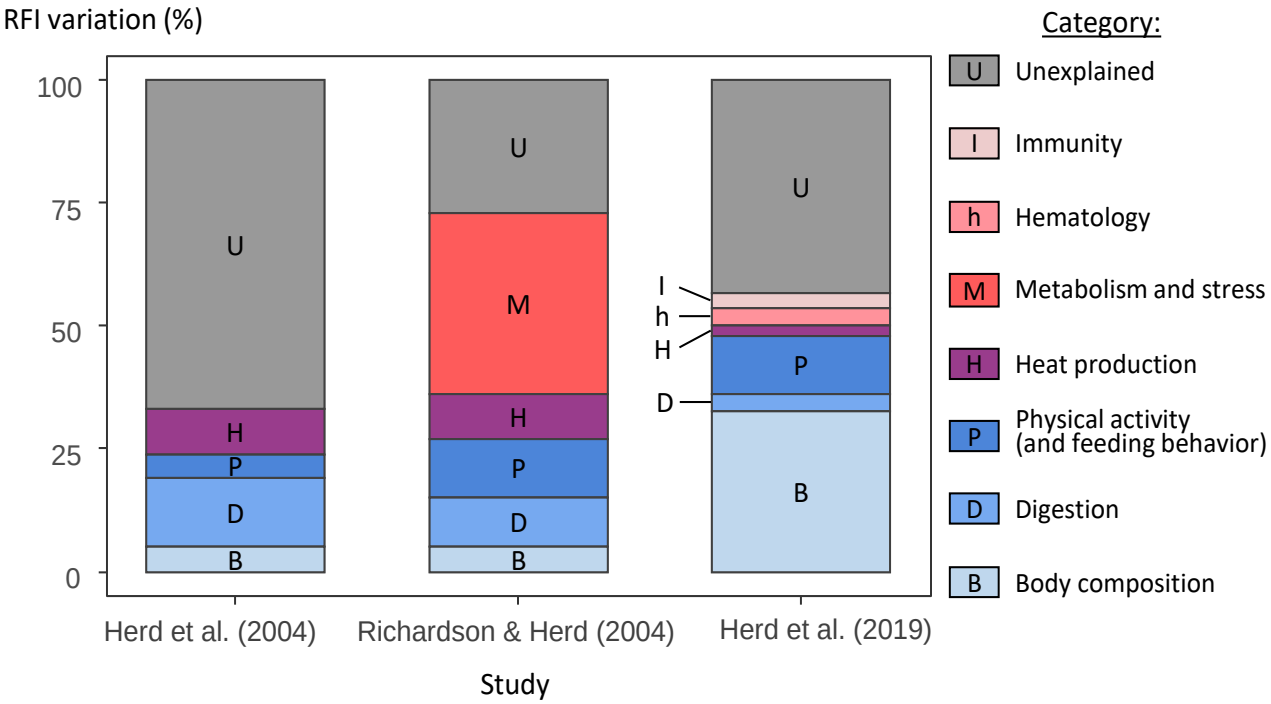



**Figure 4: Estimated contributions of biological mechanisms to residual feed intake, in studies involving beef cattle.**

The biggest sources of variations in RFI were (Figure 4): body composition, digestion, activity, tissue metabolism and stress. However, new determinants of feed efficiency have yet to be discovered since 27 to 67% of RFI variations remained unexplained in feedlot cattle (Figure 4). New technologies might provide deeper or new insights into feed efficiency by studying metabolites, microbiota and genotypes for instance. These new sets of variables might also provide predictors of feed efficiency.

**Take-home messages about feed efficiency:**

- Various feed efficiency traits were defined. All direct criteria require individual feed intake records, which require expensive investments.
- In subsequent analyses of the manuscript, two traits will be retained since they are often reported in the literature: feed conversion ratio (FCR), and residual feed intake (RFI). The two traits present different advantages. FCR can be easily interpreted and values can be compared between different populations. RFI's computation makes RFI phenotypically independent from size and growth phenotypes.
- Few or conflicting results exist about associations between feed efficiency and other traits: such as body composition, organ sizes, health traits or greenhouse gas emissions.
- A large proportion of feed efficiency variations remains unexplained.
- Feed efficiency is complex since it is driven by many factors. There are interplays between the holobiont and its environment. The diversity of possible biological determinisms suggests that studying the genomes and metabolisms of the ruminal microbiota and the host might highlight predictors of feed efficiency.

# II – Predicting feed efficiency from omics

The previous section highlighted that feed efficiency is rarely selected because recording feed intake is expensive. Predicting feed efficiency could help circumvent the record of feed intakes. Thus, the present section will explore options to predict complex traits such as feed efficiency. First, the literature review will highlight omics as predictors of feed efficiency. Then, the review will briefly introduce statistical challenges and integration strategies.

## A. The omics potential as feed efficiency predictors

Terms with the -omics suffix refers to the study of a specific -ome field, *i.e.* a subset of biological information. For instance, genomics refers to the study of the genome. The term "genome" was first coined in the 20s (Lederberg and McCray, 2001), and now stands for the information carried by the genetic make-up of an organism. Various -ome terminologies were coined to name different layers of biological information, from genetics to phenotypes. Thus, studying different omics will highlight different biological questionings (Dettmer et al., 2006).

Previously, the review pointed out that feed efficiency is a complex trait with many determinants. Many omic fields exist and keep emerging: 29 omic fields were already identified in 2010 (Prohaska and Stadler, 2011). Many determinants could be studied through the lens of omics: *e.g.* the microbiota composition, the holobiont metabolism and host genetics. Nonetheless, the review will focus on the most used or promising omics to predict feed efficiency.

Discussing omics will underline why they are promising to predict feed efficiency. Next subsections will underline what information omics represent, how they are obtained and their use to predict sheep feed efficiency.

## A.1 Genomics

*What is the host potential?* – The genome is the set of DNA sequences of an organism (Alberts et al., 2002). That DNA set encodes the biological information which might be expressed by the individual. Thus, genomics can help to predict the potential phenotypes.

In domestic animals, the use of genomics was partly motivated by marker assisted selection and then genomic selection (Blasco and Toro, 2014). Markers are variations of the DNA sequence scattered across the genome. When a marker is close to a gene controlling a trait, both are in linkage disequilibrium: marker and gene will likely segregate together as few genetic recombination between them can be expected. Thus, the marker should remain associated to the gene, and by extension to variations of the trait controlled by the gene.

Nowadays, the most commonly used markers are single-nucleotide polymorphisms (**SNP**s): variations of the DNA sequence by one nucleotide base. DNA microarray chips are a cost-effective solution to collect SNP data and genotype a high number of animals (Meuwissen et al., 2016). Whole-genome sequencing is less cost-effective but allows to discover and detect more SNPs. Overall, many genomic marker effects can be modelled simultaneously to predict traits thanks to technological and statistical progresses.

Meuwissen et al. (2001) helped popularizing genomic selection by including thousands of markers scattered across the genome. Genomic prediction was recommended for traits too expensive or complex to record routinely, such as feed efficiency (Hayes et al., 2013). In growing cattle, the accuracy of RFI genomic predictions varied between -0.01 and 0.67 depending on the population, cross-validation strategy, model and SNP panel (Pryce et al., 2012; Lu et al., 2016; Silva et al., 2016; Brunes et al., 2021). It demonstrates the potential of genomics to predict feed efficiency. Studies are warranted in sheep.

**Note**: Genetic expression partially depends on the genome, on the DNA sequence. However, the epigenome also influences genetic expression. Epigenetics focuses on the changes of DNA expression without any DNA sequence change, such as cytosine methylation. The methylome refers to the methylation profile. The influence of epigenetics over feed efficiency was recently evidenced in lactating cows: efficient cows had less methylated sites (López-Catalina et al., 2022).

**Additional note:** Accuracy of genomic predictions are often expressed as a ratio: the correlation between adjusted phenotypes and genomic breeding values, divided by the square-root of the trait heritability. On the other hand, prediction accuracy of other omics is often expressed directly as the correlation between phenotypes and predictions.

## A.2 Metabarcoding / metagenomics

During the last decade, microbiota data have been proposed to predict animal complex traits (Ross et al., 2013). As the review showed (section I.C.3), the microbiota may contribute to digestion and feed efficiency. Therefore, the microbiota may be relevant to predict feed efficiency.

Moreover the microbiota composition could also help predict efficiency indirectly, by signing for systematic effects (He et al., 2022a). For instance, the microbiota composition is heavily affected by farming and geographical conditions (Henderson et al., 2015; Belanche et al., 2019; Li et al., 2019c; Anderson et al., 2021; Marie-Etancelin et al., 2021; Wei et al., 2021). Thus, when environmental factors affect both efficiency and the microbiota composition, metabarcoding and metagenomics data could be indirectly associated to feed efficiency.

Microbiota data are currently mainly obtained from metabarcoding and metagenomics. Both techniques rely on different sequencing strategies, briefly discussed subsequently.

### Metabarcoding

*Who is part of the microbiota?* – Metabarcoding the microbiota allows to assess the microbiota composition. The goal is to count the number of sequences affiliated to a phylogenetic taxon, which is why metabarcoding is also called metataxonomics. Metabarcoding may help predicting feed efficiency by knowing who may contribute to digestion.

Metabarcoding is carried out by sequencing a targeted genomic sequence - *i.e*. a barcode. Metabarcoding may target: the prokaryote 16S ribosomal RNA (**rRNA**) gene for bacteria and/or archaea abundances, the eukaryote 18S rRNA gene mainly for protozoa and marginally for fungi, while the internal transcribed spacer is preferred for fungi (Breitwieser et al., 2019). Viruses are largely ignored by metabarcoding studies since no universal barcode gene was identified in viruses (Breitwieser et al., 2019). Today, short-read

sequencing is the most commonly used strategy in metabarcoding. Short-reads contribute to the cost-effectiveness of metabarcoding but it also decreases the reliability of taxonomic affiliations (Fuks et al., 2018; Callahan et al., 2019).

In sheep, predicting feed efficiency from rumen metabarcoding looks promising: Ellison et al. (2019) found a strong correlation between actual sheep RFI and predictions from 16S data (r=0.71). In dairy cows, the correlation between REI and 16S predictions was lower (r=0.55) (Tapio et al., 2023). These metabarcoding results still have to be confirmed in larger populations.

Metabarcoding is a fast and cost-effective approach to get an approximate census of microorganisms, but it does not capture information about the microbiota genetic potential (Breitwieser et al., 2019).

**Metagenomics**

*What is the microbiota potential?* – Metagenomics allows to study the "second genome" of the holobiont, by sequencing a wide variety of microbial DNA fragments. Metagenomics may help by depicting how the microbiota can contribute to digestion.

Metagenomics does not target a specific genomic sequence or microbial group. Whole-genome shotgun sequencing can sample all microorganisms' DNAs (bacteria, archaea, fungi and viruses), but also host and feedstuffs DNAs. Thus, metagenomics may help characterizing the functional potential of the microbiota without selecting any particular gene (Scholz et al., 2016). Compared to metabarcoding, metagenomics requires a high coverage, *i.e.* a lot of sequences per sequenced locus. Thus, metagenomics is much more expensive and is rarely used in large animal populations.

Predicting feed efficiency from rumen metagenomics is also promising: Hess et al. (2022, preprint) found moderate correlations between actual sheep RFI and predictions (from 0.19±0.05 to 0.47±0.17). Accuracies varied with the age and diet of sheep. In dairy cows, both metabarcoding and metagenomic studies underlined the potential of the ruminal microbiota to predict sheep feed efficiency.

**Note**: Metagenomics can be cost prohibitive. Thus, inference models were created to infer the abundance of microbial genes or functions from metarcoding data (Picrust, 2013). Results might be less accurate, but more animals can be studied with metabarcoding.

## A.3 Transcriptomics / Metatranscriptomics

### Transcriptomics

*What seems to happen?* – The transcriptome encompasses all RNA transcripts expressed in one host cell or tissue: coding RNA such as messenger RNA, or noncoding RNA such as ribosomal RNA (Vailati-Riboni et al., 2017). RNA is produced based on a DNA sequence during transcription. Thus, studying RNA may indicate which parts of the genome are expressed and how much. RNA can also mediate the genome expression (Romao et al., 2011). Then, transcriptomics could help predict complex traits by shedding light on active genetic mechanisms.

Many technologies exist and keep emerging in transcriptomics. While microarrays allow the study of predetermined sets of RNA, next-generation sequencing allows the discovery of new transcripts (Schneider and Orchard, 2011). RNA-seq is a cost-effective solution, which generally relies on the conversion of RNA into DNA copies prior to sequencing (Schneider and Orchard, 2011).

Past studies suggested that transcriptomics could contribute to feed efficiency prediction. In Hu lambs, the liver transcriptome revealed that coding and non-coding RNA were significantly associated to RFI In Hu lambs (Zhang et al., 2019b, 2022). In pigs, liver transcriptomics predicted RFI breeding values well ($0.63 \leq R^2 \leq 0.65$) (Messad et al., 2019). In dairy sheep, non-invasive samplings identified associations between the milk transcriptome and feed efficiency (Suárez-Vega et al., 2023). In meat sheep, studies are still needed to assess how well feed efficiency can be predicted from transcriptomics, particularly from non-invasive sampling.

### Metatranscriptomics

The host and microbiota transcriptomes -called metatranscriptome- could be relevant to predict feed efficiency too. Metatranscriptomics would unravel which microbiota and host genes are expressed (Aguiar-Pulido et al., 2016). In cattle, feed efficiency groups were associated to different rumen metatranscriptomic profiles (Li and Guan, 2017; Li et al.,

2019a; Xue et al., 2022). However, the prediction accuracy of feed efficiency from metatranscriptomics is not or scarcely documented in ruminants.

## A.4 Proteomics / metaproteomics

### Proteomics

*What makes it happen?* – The proteome is the set of proteins found in one host cell or tissue (Vailati-Riboni et al., 2017). Proteins are synthesized based on RNA sequences, during translation. Proteins are involved in every biological process: recognizing signals, catalyzing reactions, regulating the metabolism, shaping structures, and motion (Vihko and Wagenerb, 1992). Thus, the proteome may help predict complex phenotypes by studying the proteins underlying all the biological processes which eventually lead to the observed trait.

Proteomics may focus on the expression level, structure or function of proteins (Vaz and Tanavde, 2018). While predicting complex traits in animal science, most studies relied on expression-based proteomics: quantitative profiles of proteins are compared between samples. A pre-separation step, such as electrophoresis or chromatography, can be carried out to separate proteins before identification and quantification (Vailati-Riboni et al., 2017). Then, mass spectrometry is often at the core of high-throughput quantitative profiling (Aebersold and Mann, 2003).

Studies showed that proteomics could supply feed efficiency markers. In dairy heifers and Merino lambs, studies highlighted that the hepatic proteome was associated to diet differences (Santos et al., 2018; Zhang et al., 2019a). In both studies, feed efficiencies were contrasted between the diet treatments. But studies still have to assess the proteome predictive ability when ruminants are fed the same diet.

### Metaproteomics

The host and microbiota proteomes -called metaproteome- could also predict feed efficiency. In 12 dairy cows, the rumen proteome accurately predicted the feed efficiency group (error rates as low as 0.17), better than metagenomics and 16S metabarcoding (as low as 0.25) (Sasson et al., 2022). Larger independent studies are needed to confirm results.

## A.5 Metabolomics

*What happens and happened?* – The metabolome corresponds to the set of metabolites found in a cell or tissue: metabolites are small molecules (<1500 daltons) produced by the metabolism (Wishart, 2007). Metabolites can be intermediate molecules or end products. Thus, metabolomics may help predict complex traits by studying what was and what is produced by metabolic reactions.

Metabolomic profiling may be done by coupling separation techniques (chromatography, electrophoresis) and mass spectrometry (Vailati-Riboni et al., 2017). Same as proteomics, mass spectrometry is used to quantify and identify metabolites. The metabolome is also frequently investigated thanks to NMR and infrared spectroscopy (van der Greef et al., 2004). Human expertise or automatic software may identify the molecules attributable to spectral peaks (Lefort et al., 2019). In 'black-box' models, spectral variables may be directly used without attributing metabolites. For instance, NMR spectra may be divided into slices called bins or buckets. Then, the area under the curve of buckets can be used as quantitative variables.

Past studies proved the potential of metabolomics to predict sheep feed efficiency. Plasma metabolites could discriminate the efficient line from the inefficient one: the area under the receiver operating characteristic curve (**AUROC**) ranged from 0.81 to 0.87 (Touitou et al., 2022). Serum metabolites could predict whether a lamb had a high or low RFI phenotype (AUROC=0.80) (Goldansaz et al., 2020). While studies in cattle suggested that the rumen metabolome may also provide feed efficiency biomarkers (Artegoitia et al., 2017; Clemmons et al., 2020), results were less optimistic in the trial involving Romane divergent lines (Touitou et al., 2022). It remains to check the accuracy of quantitative predictions of efficiency phenotypes from metabolomics.

To date, no consensus term is used to specifically refer to the host and microbiota metabolomics (unlike metagenomics, metatranscriptomics or metaproteomics). Host and microbiota metabolites may also pass through different tissues: microorganisms release volatile fatty acids in the ruminal fluid which may pass into the host bloodstream (Membrive, 2016). Thus, plasma, serum and rumen metabolomes will be seen as the result of the holobiont metabolism throughout the thesis.

**Note:** Metabolomics covers a wide range of molecules: vitamins, carbohydrates, nucleic acids, amino acids, fatty acids and more (Vailati-Riboni et al., 2017). Molecular subsets may also be studied: for instance, the lipidome was coined to refer to the study of lipids.

## A.6 Phenomics

*What is observed?* – The definition of phenomics varies depending on authors. Soule (1967) introduced the word phenome to refer to "the phenotype as a whole". The phenome encompasses all phenotypes of an individual across time, organs, tissues and cells (Houle et al., 2010). Thus, Houle et al. (2010) defined phenomics as large-scale phenotyping. One has to prioritize a set of phenotypes, so let's focus on the "spectral phenome". As defined by O'Reilly-Wapstra et al. (2013), the spectral phenome is the set of reflectances, absorbances or transmittances of near-infrared spectra (**NIRS**). NIRS allows to rapidly dress the physico-chemical profile of a biological sample (O'Reilly-Wapstra et al., 2013).

Metabolomics and phenomics technologies and goals are close. However, the thesis makes the distinction between metabolomics and phenomics, since faeces are analyzed with NIRS. The fecal composition does not solely depend on host excreted metabolites: undigested feedstuffs are also excreted. Phenomic variables are quantitative phenotypes, representing physico-chemical properties of samples (Rincent et al., 2018). Thus, spectral phenomics may help predict feed efficiency by indirectly capturing molecular phenotypes contributing to efficiency.

Few studies tried to predict feed efficiency from NIRS data. However, Kneebone and Dryden (2014) used faecal NIRS to predict sheep intake and digestibility under several diets: $R^2$ varied from 0.76 to 0.90 depending on units (dry, organic, digestible matter or crude proteins). Studies are needed under a single diet to see how accurately faecal NIRS may capture inter-individual variations, instead of diet variations. In dairy cows, the prediction accuracy of RFI from milk mid-infrared spectra varied with lactation stages and cross-validation strategies: $R^2$ varied from 0.08 to 0.46 (Shetty et al., 2017). Therefore, infrared spectra may help predict feed efficiency but further investigations are required in meat sheep: a sampling location and prediction strategy have yet to be defined.

## A.7 The cascade of omics

As previous sections hinted, all omics are related and represent a cascade of events (Dettmer et al., 2006, Figure 5). The cascade begins with the genome and metagenome, then continues with the transcriptomes, proteomes, metabolome. Finally, phenotypes are the end result of the cascade.

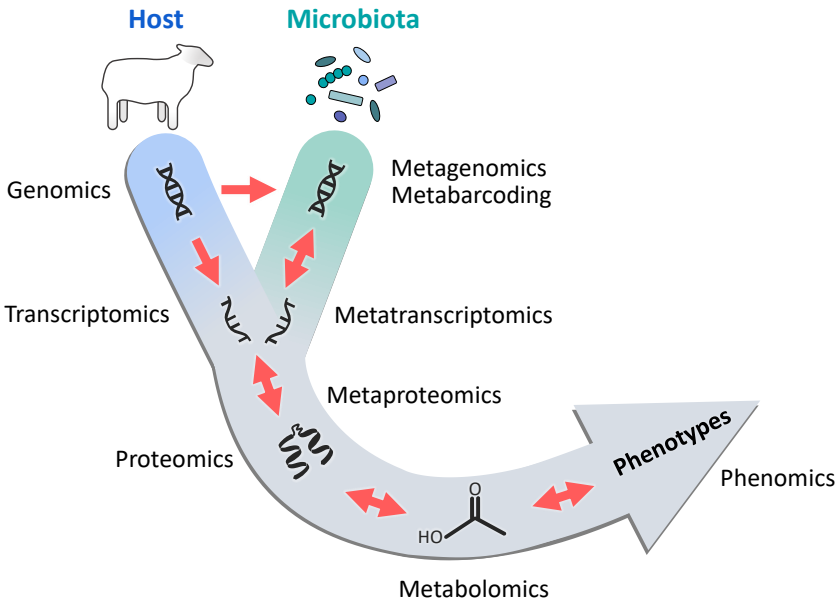



**Figure 5: Omics commonly studied, from the genetic make-up to the phenotype**

Different omic layers are connected through the omic cascade of events, but also because of host and microbiota interactions. For example, the development of the rumen epithelium is the result of a cross-talk between lambs and their microbiota. The rumen epithelium growth is mediated by host gene expression (transcriptome) and volatile fatty acids (rumen metabolome) produced by the microbiota (metabarcoding and metagenomics) (Lin et al., 2019). Moreover, ruminant genetics partially determines the rumen microbiota composition (Li et al., 2019b; Martinez Boggio et al., 2022). On the opposite, the metagenome has virtually no influence over the host genome over short period of times. That assumption does not hold over long time spans: ruminants and their microbiota co-evolved for millions of years (Selinger et al., 1996). Co-evolution occurred thanks to natural selection, genetic drift and horizontal genetic transfers – gene transfers between different the host and different microbial species (Rosenberg, 2021).

# B. Statistical tools to predict feed efficiency from omics

Previous section highlighted that omics provide different layers of information, from the genome to the phenotype. Thus, combining different omics may yield a more comprehensive understanding of the biological system in order to decipher feed efficiency (Widmann et al., 2015). However, different omics may also provide redundant information due to the connectedness of omics. Therefore, several challenges must be addressed by statistical tools when it comes to pinpoint signals out of the omic haystack.

The review of statistical tools will focus on supervised statistical learning. Indeed, the aim of the thesis is to predict feed efficiency. Statistical models will be used to mine omics and identify biological signatures predicting sheep feed efficiency. That is why the review will focus on supervised models, rather than unsupervised approaches used to unravel unknown patterns.

First, statistical challenges of omics data analysis will be reviewed. Then, the review will linger on modelling possibilities to predict complex traits from omic patterns. Finally, data integration will be discussed to understand how multi-omics can be combined to predict traits.

## B.1 The statistical challenges of omics

Analyzing omics offer several challenges: omics are noisy, highly dimensional, and heterogeneous. All these challenges may factor in pre-processing and during statistical learning.

### Noisy data

Omics data are noisy (Picard et al., 2021): meaning that omic variables are the reflect of biological signals and artifacts. Too much noise may even mask biological signals (Ning and Lo, 2010).

First, missing values can be a source of noise in any dataset: there is no record, whether a biological signal exists or not. Missing values may arise at random or arise from technologies (*e.g.* low sequencing coverage, low detection sensitivity of molecules, faulty measurements) (Mirza et al., 2019). When records are incomplete, all individuals with missing records may be discarded or missing values can be imputed. Imputation may be preferred to avoid

wasting data when samples are scarce. Song et al. (2020) reviewed imputation techniques. Briefly, imputation may rely on single omics data: missing genotypes can be imputed from a reference population and genetic algorithms making use of genetic properties such as linkage disequilibrium (Song et al., 2020). In multi-omics settings, missing values can also be inferred from different omics: for instance, missing transcriptomics could be deduced from genomics thanks to machine learning models, transfer learning or matrix factorization (Song et al., 2020). Imputation of missing values may be done prior to supervised learning. Nonetheless, some approaches may handle missing values directly. The single-step framework in genomic selection may include individuals without genotypes, by updating the genomic relationship matrix thanks to pedigree (Legarra et al., 2014). The single-step approach was then applied to different omics: *e.g.* to missing transcriptomics data, thanks to genomics or pedigree data (Westhues et al., 2019).

Batch effects may also be a source of noise. Batch effects may be classed into several categories: biological effects (*e.g.* age or sex), experimental (*e.g.* housing or diet), technical (*e.g.* sampling condition, technician or technology) or computational (*e.g.* software and parameters) (Wang and Lê Cao, 2020). Wang and Lê Cao (2020) reviewed strategies to either account for batch effects, or correct for batch effects – sometimes referred to as normalization. Prior to statistical learning, options to correct batch effects include: correcting the distribution of variables per batch (*e.g.* center variables within batches) or modelling batch effects (*e.g.* Bayesian, linear or partial least-squares models) (Wang and Lê Cao, 2020). All options are not listed, and options must be selected based on their assumptions.

Finally, it may be difficult to accurately pinpoint or select relevant variables when collinearity exists between variables. Collinearity arises from associations, correlations between different variables. One signal may overlap over several variables. Redundancy may exist in single-omics and multi-omics studies, as suggested for transcriptomics and genomics (Westhues et al., 2019).

**The curse of dimensionality**

Omics data are often highly dimensional: datasets include many variables. For instance, genotyping animals with DNA chips may yield dozens of thousands of SNPs per individual. The number of variables often largely exceed the number of samples for practical or

economic reasons (Picard et al., 2021). When machine learning is used to predict an outcome, models may overfit by capturing random noise when the model is trained. Overfitting may be exacerbated with highly dimensional datasets (Domingos, 2012): when there are more variables, it is more likely to find random variable combinations spuriously associated to an outcome observed in few samples. However, capturing these random combinations is not interesting because model predictions will not be generalizable to new independent datasets.

When it comes to omics, "more is better" according to Huang et al. (2017). Multi-omics provide more layers of information. However, it may also aggravate the curse dimensionality by drastically increasing the number of variables. It may bring more noise or redundant information (Picard et al., 2021).

One solution is to decrease dimensionality by extracting features. The aim of feature extraction is to transform the original set of variables into a smaller set of combined variables, without losing too much information (Mirza et al., 2019). Principal component analysis is the most widely used technique to extract features: the analysis is applied to define linear combinations of variables and then principal components coordinates are used as the new variables (Picard et al., 2021). Meng et al. (2016) provide a more thorough review on dimension reduction techniques.

Another solution is to decrease the dimensionality by selecting variables. Only a subset of the most informative variables can be kept. Selection methods are classified into three categories:

- Filter methods applied prior to model building: variables may be filtered based on univariate statistics (*e.g.* correlation analysis, differential analysis) (Mirza et al., 2019);
- Wrapper methods repeatedly used while building successive model versions: a multivariate model is trained several times and the simplest model optimizing accuracy is kept (*e.g.* recursive feature elimination discards the least important variables successively until accuracy decreases too much) (Picard et al., 2021);
- Embedded methods are included in models having a built-in variable selection (*e.g.* multivariate penalized regressions such as Lasso, Bayesian approaches) (de los Campos et al., 2018).

Variable selection and feature extraction can be used jointly to reduce dimensionality. Dimensionality reduction alleviate computational intensity but also improves interpretability (Brouard et al., 2022).

Reducing the dataset dimension can also decrease the complexity of subsequent models too. Let's consider a model whose complexity depends on the number of variables. Training a model from many variables may give it more flexibility. More flexible models capture more information and the variance of predictions increases. Assumptions of flexible models may fit more closely to the data and the prediction bias decreases. The tradeoff bias-variance will affect how well a model can be generalized to new data (Figure 6):

- The model is underfitted when it captures too few signals (the variance is low and the bias high);
- The model is overfitted when it captures too much noise (the variance is high and the bias is low) (Hastie et al., 2021).

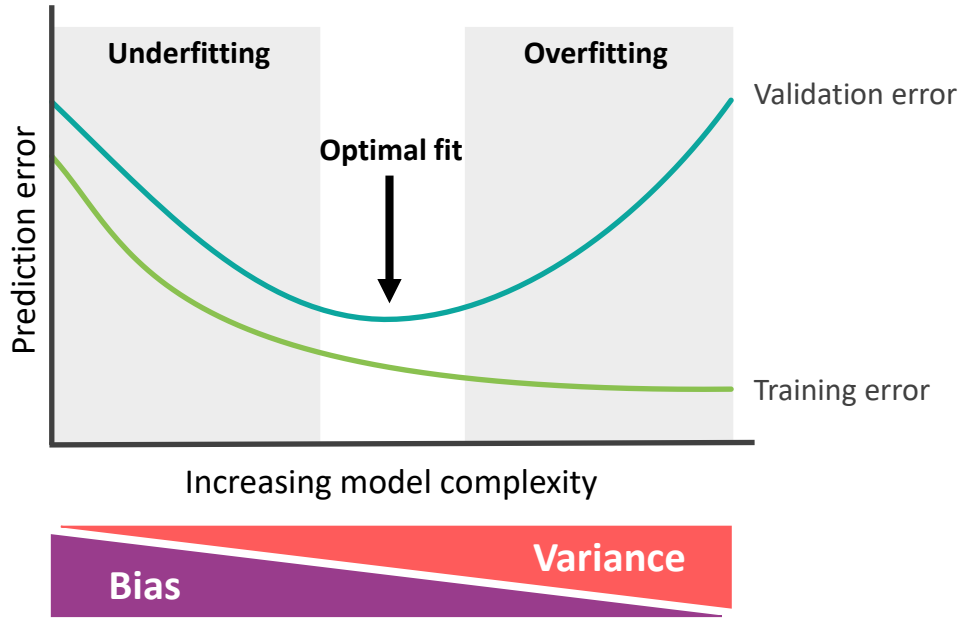



**Figure 6: Model goodness of fit and the underlying trade-off between bias and variance**

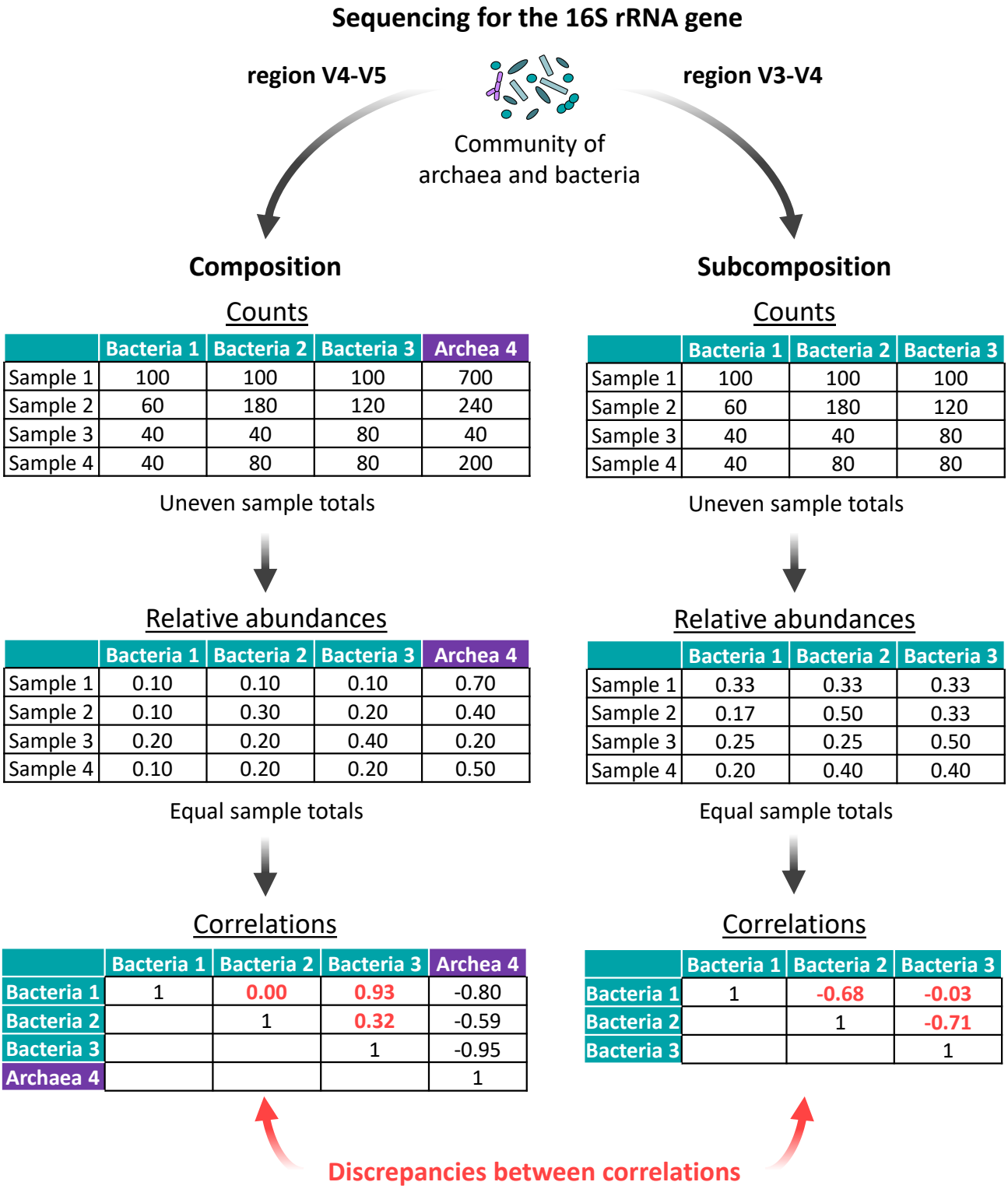
Selecting a dimension reduction strategy may be done thanks to a validation dataset. Validation sets can be study subsets left out during cross-validation, or independent studies. The goal is to train a model (on a training set), which generalizes well to new data (validation set).

**Data heterogeneity and compositionality**

Heterogeneity of data may be challenging, especially in multi-omics experiments. Heterogenous datasets differ based on their distribution (sparsity, normality, …), on their dimension (high or small number of variables per dataset) or data type (continuous, discrete, …). Fusing kernels, projecting to a common latent subspace, using networks or deep learning algorithms are popular approaches to incorporate heterogeneous datasets (Mirza et al., 2019).

Another source of heterogeneity often overlooked by biologists is compositionality. Pondering the nature of variables is not the same as pondering whether variables are compositional. Genotypes can be encoded by allele counts, while transcriptomics and metabarcoding may be encoded by sequence counts. However, genotypes are not considered compositional while transcriptomics and metabarcoding data are. If the relevant information is relative, then data are compositional (Aitchison, 1982). In compositions, the information resides in ratios between parts. For instance, one may ask if having 80 sequences attributed to bacteria 3 is a lot, in samples 2 and 3 (Figure 7)? It depends on how many sequences were attributed to other microorganisms: bacteria 3 is the most abundant in sample 3, while it is not in sample 4. It underlines that information is relative.

A difficulty with compositional data is applying classic statistics which may turn out impractical. For instance, in Figure 7 a mock microbial community is composed of bacteria and archaea. Sequencing the 16S rRNA for the regions V4-V5 would allow to detect bacteria and archaea, while the regions V3-V4 would allow to detect bacteria only. Whatever the sequencing, let's imagine that we would get the same number of sequences per bacteria and sequences (an unrealistic hypothesis). Then, the result of the first sequencing would be a composition (with bacteria and archaea), and the result of the second sequencing would be its subcomposition (with bacteria only). Counts must be converted into relative abundances since sequencing depth varies: the total number of sequences is uneven across samples. Finally, if one applied Pearson correlations, strong discrepancies would be found between bacteria correlations (Figure 7). Correlation analysis is unsound with compositional data because correlations cannot range freely: the constant sum of parts impedes correlations from ranging between -1 and 1 (van den Boogaart and Tolosana-Delgado, 2013).




| | Bacteria 1 | Bacteria 2 | Bacteria 3 | Archea 4 |
|---|---|---|---|---|
| Sample 1 | 100 | 100 | 100 | 700 |
| Sample 2 | 60 | 180 | 120 | 240 |
| Sample 3 | 40 | 40 | 80 | 40 |
| Sample 4 | 40 | 80 | 80 | 200 |

| | Bacteria 1 | Bacteria 2 | Bacteria 3 |
|---|---|---|---|
| Sample 1 | 100 | 100 | 100 |
| Sample 2 | 60 | 180 | 120 |
| Sample 3 | 40 | 40 | 80 |
| Sample 4 | 40 | 80 | 80 |

| | Bacteria 1 | Bacteria 2 | Bacteria 3 | Archea 4 |
|---|---|---|---|---|
| Sample 1 | 0.10 | 0.10 | 0.10 | 0.70 |
| Sample 2 | 0.10 | 0.30 | 0.20 | 0.40 |
| Sample 3 | 0.20 | 0.20 | 0.40 | 0.20 |
| Sample 4 | 0.10 | 0.20 | 0.20 | 0.50 |

| | Bacteria 1 | Bacteria 2 | Bacteria 3 |
|---|---|---|---|
| Sample 1 | 0.33 | 0.33 | 0.33 |
| Sample 2 | 0.17 | 0.50 | 0.33 |
| Sample 3 | 0.25 | 0.25 | 0.50 |
| Sample 4 | 0.20 | 0.40 | 0.40 |

| | Bacteria 1 | Bacteria 2 | Bacteria 3 | Archea 4 |
|---|---|---|---|---|
| Bacteria 1 | 1 | 0.00 | 0.93 | -0.80 |
| Bacteria 2 | | 1 | 0.32 | -0.59 |
| Bacteria 3 | | | 1 | -0.95 |
| Archaea 4 | | | | 1 |

| | Bacteria 1 | Bacteria 2 | Bacteria 3 |
|---|---|---|---|
| Bacteria 1 | 1 | -0.68 | -0.03 |
| Bacteria 2 | | 1 | -0.71 |
| Bacteria 3 | | | 1 |

**Figure 7: Illustration of correlation discrepancies between a composition and one subcomposition**

Compositional data require specific treatments. If a dataset is compositional, each sample may be defined as a vector with N composition parts. The compositional vector $x = (x_1, \ldots, x_D)$ is defined in a geometrical space called simplex $S^D$ (Aitchison, 1982):

$$S^D = \left\{ x = (x_1, \ldots, x_D): \quad x_i \geq 0, \quad \sum_{i=1}^{D} x_i = constant \right\}$$

where all parts $x_1$ to $x_D$ have null or positive values. And the sum of all parts is equal to a constant.

**Note**: A dataset may be compositional even if the sum of raw data is not constant. When sequencing depth varies, the total number of sequences is not constant in metabarcoding and transcriptomics. However, information remains relative: raw counts must be interpreted as relative abundances (percentage, proportion…). And the sum of relative abundances is a constant. It is how data are interpreted that determines whether the dataset is compositional. Not raw data.

Besides transcriptomics and metabarcoding abundances, proteomic and metabolomics concentrations are compositional too. Same as metabolomic buckets which represent parts of the area under the spectrum curve.

Specific transformations are necessary to export compositional data from the simplex space $S^D$ to a real space where Euclidean geometry applies. Logratios consist in logarithms applied to ratios between several parts of a composition. The centered logratio is the most commonly used transformation and is carried out sample-wise (Gloor et al., 2017):

$$clr\,(x) = \left[ \ln\left(\frac{x_1}{g(x)}\right), \ldots, \left(\frac{x_D}{g(x)}\right) \right]$$

Where $x$ is a vector of D composition parts $(x_1, \ldots, x_D)$. The sample geometric mean is denoted by $g(x)$.

After a logratio transformation, compositional data may be used in appropriate standard multivariate statistical analyzes: principal component analysis or partial least-squares, for example.

## B.2 Data integration in statistical learning

Multi-omics experiments may help understand complex traits and predict them more accurately. Indeed, multi-omics would provide several layers of information between genetics and phenotypes. However, integrating omics does not guarantee a better understanding or better predictions. Not all studies managed to improve the prediction accuracy of complex traits from several omics. In ewes and cows, two studies evidenced that omics integration improved the prediction of all examined methane-related traits, compared to the best omic dataset (Ross et al., 2020; Qadri et al., 2022). In lambs and ewes, another study showed the opposite: combining genomics and metagenomics did not improve the prediction of the three respiratory traits assessed (Hess et al., 2022). In pigs, the integration of genomics and microbiota improved the prediction of ADG and FCR (Qadri et al., 2022). In pigs again, predictions of ADFI and digestive coefficients were not improved by omics integration (Qadri et al., 2022; Carillier-Jacquin et al., 2022).

Data integration strategies may be classified into distinct categories. Categories may be defined according to the integration goal: do we integrate different individuals or different variables? On the other hand, categories may be defined according to the strategy: how do we integrate data?

### Difference between P and N integration

P and N integration refer to which data are integrated. Abbreviations comes from standard notations: P is used to denote the number of variables, while N denotes the number of individuals or samples.

P-integration attempts to combine several studies recording the same set of P variables (Rohart et al., 2017b, Figure 8). Integrating data from different studies may increase the number of observations and the reproducibility of the model (Rohart et al., 2017a). Studies may correspond to experiments led by different teams, or correspond to batch effects. The Multivariate INTegrative sparse Partial Least-Squares (MINT-sPLS) is an example of P-integration framework, implemented in mixOmics (Rohart et al., 2017a).

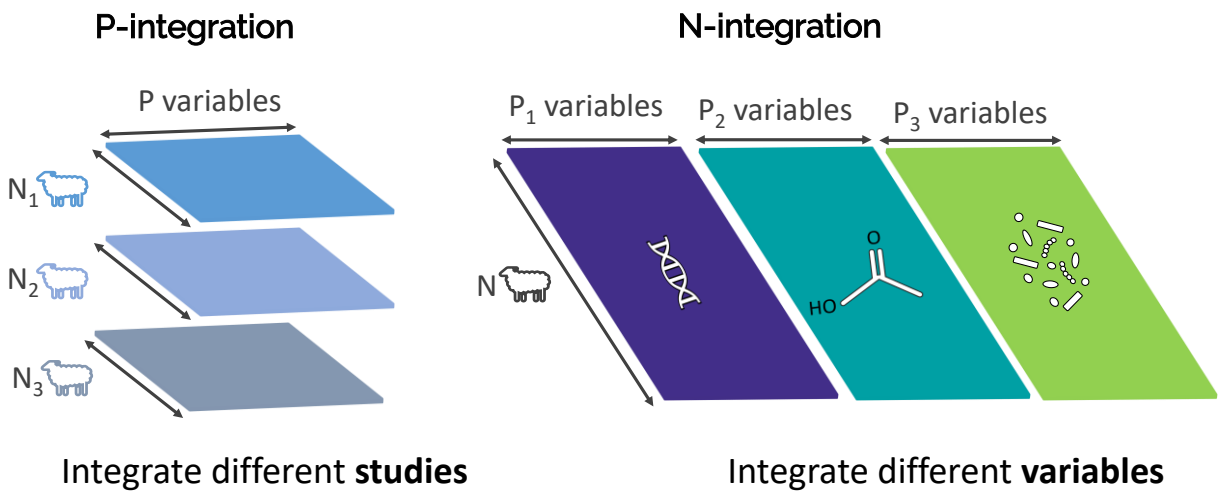



**Figure 8: Difference between N and P integration (adapted from Lê Cao and Welham, 2021)**

N-integration tries to combine different data variables recorded on the same N samples or animals (Rohart et al., 2017b, Figure 8). Combining different data variables may uncover more biological signals, by simultaneously examining different layers of information or uncovering associations between the different variables (Singh et al., 2019). Different data variables (several omics or non-omics data) may be grouped per kind into 'blocks'. The block-sPLS is an instance of N-integration, available in mixOmics (Rohart et al., 2017b).

NP-integration is also feasible: it integrates different studies and different blocks of omics altogether (Rohart et al., 2017b). The MINT.block-sPLS was implemented in mixOmics (Rohart et al., 2017a). Few other integration techniques exist for NP-integration: literature focuses more on N-integration, on multi-omics.

### The diversity of integration strategies

Integration strategies may be distinguished by when and how the different variables are combined. Picard et al. (2021) proposed 5 categories for integration strategies (Figure 9).

1- Early integration

Also called integration by concatenation, early integration is done before modelling. All blocks of variables are concatenated together, to get a single matrix. Any single-omics approach can be applied to the concatenated matrix: from deep learning, Bayesian, tree-based, kernel to linear mixed models and more.

Early integration is the most simple approach (Rappoport and Shamir, 2018). Furthermore, mixed graphical models may be applied to the concatenated matrix to unravel the relationships between all variables (Picard et al., 2021). A disadvantage is that early integration does not account for data heterogeneity (Picard et al., 2021). Another drawback of concatenation is that some omics may be overlooked and unused by the model (Rohart et al., 2017b): a single block of omics may overshadow others if it presents the strongest signals.

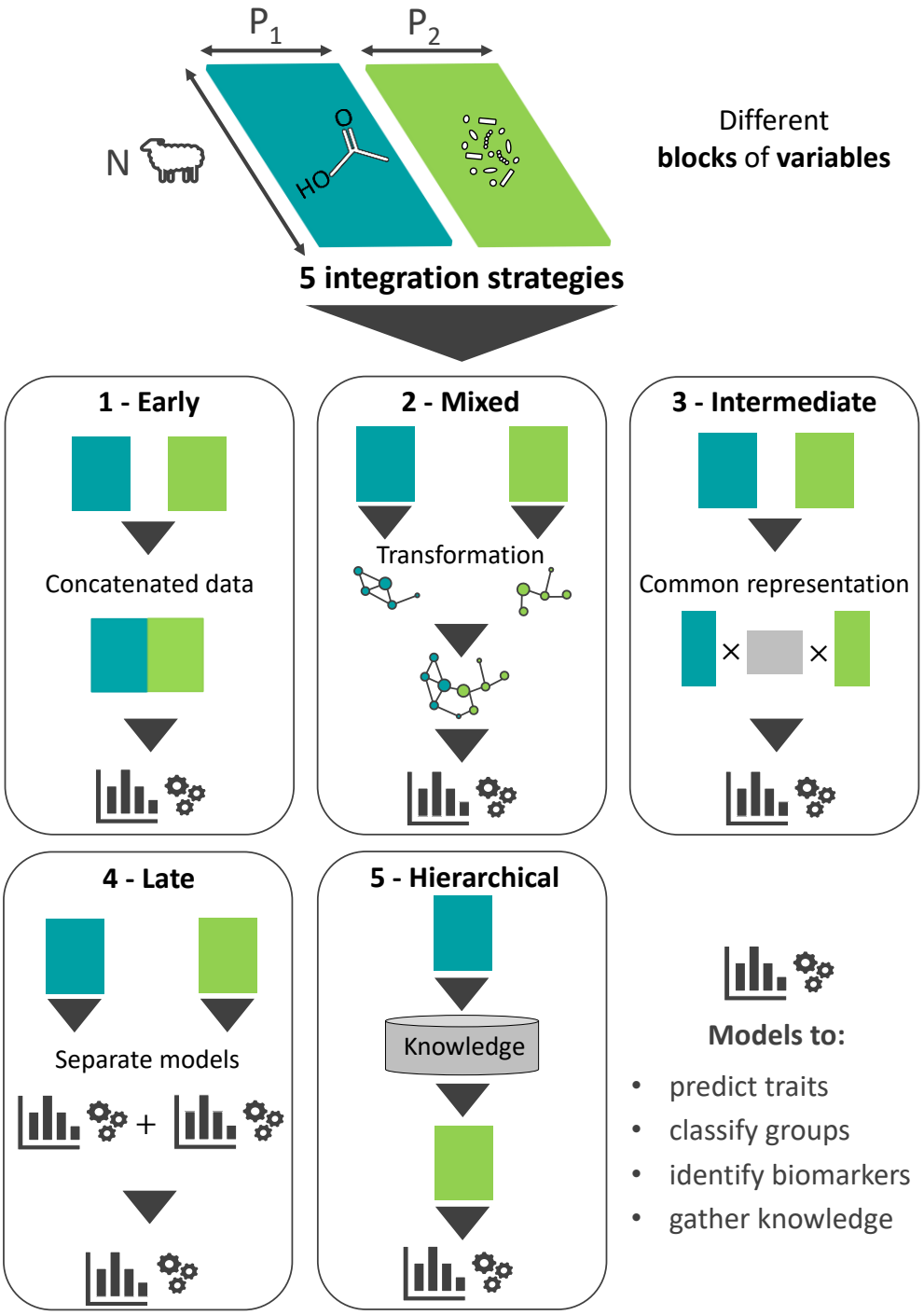



**Figure 9: Integration strategies (adapted from Picard et al., 2021)**

2- Mixed integration

Also called transformation integration, mixed integration can be seen as a two-step procedure. Each dataset may be transformed separately. Kernel, graph and deep learning may be used to transform each set of omics into a simpler representation (Picard et al., 2021). Then, the transformed blocks are modelled together.

Compared to early integration, mixed integration may reduce the heterogeneity existing between different blocks of omics.

3- Intermediate integration

Intermediate integration encompasses methods which do not transform or concatenate variables (Picard et al., 2021). Intermediate methods can handle multiple datasets jointly and directly. Such methods may rely on a common latent space, such as matrix factorization or multi-omics factor analysis (Picard et al., 2021). For instance, MINT-sPLS models integrate different studies by defining global components (also called latent variables in mixOmics), which are common to all integrated studies (Rohart et al., 2017a).

4- Late integration

Also called model integration, late integration can be seen as a two-step procedure. Any single-omics approach is applied to each dataset separately, from deep learning to mixed models (Rappoport and Shamir, 2018). Later, results of all models are aggregated together to get an overall result (by averaging predictions, weighting votes, …).

Late integration allows to choose the most suitable model per dataset, based on its characteristics (dimension, compositionality, distribution…). It may help dealing with heterogeneous data (Picard et al., 2021). Modelling each set of omics separately alleviate the dimensionality curse too: working with all omics together would require to estimate more model parameters simultaneously. However, a major drawback of late integration is that it cannot account for interactions between different blocks of omics (Picard et al., 2021).

5- Hierarchical integration

Briefly, hierarchical models are based on prior knowledge. Knowledge can be incorporated to account for regulatory relationships between omics layers (Picard et al., 2021). Reference databases and literature may provide such information.

An advantage is that hierarchical integration may highlight regulatory networks across several blocks of omics (Picard et al., 2021). However, a drawback is the need for knowledge. Some omics and organisms can be poorly documented. Thus, too much knowledge might lack when many omics are integrated. Hierarchical integration seems compromised to predict feed efficiency from genomics, metabarcoding, metabolomics, lipidomics and phenomics data in sheep.

**Take-home messages about omics and statistical tools:**

- Omics may potentially help predict feed efficiency by representing biological information. Multi-omics may represent intermediate layers, from genetics of the holobiont to its phenotypes.
- The omics cascade illustrates that different omics are connected and may share patterns.
- Predicting from omics present several challenges: the noise which may mask biological signals; the curse of dimensionality arising from many variables and few samples; the data heterogeneity which complicates the selection of a modelling strategy.
- P-integration might be useful to improve reproducibility by integrating data of individuals from different studies or batches. On the other hand, N-integration might prove useful to detect new signals by integrating different blocks of omics.
- Late N-integration might be suitable for experiments with few samples and many different omics. Building models separately per omics may help accounting for data heterogeneity. It may also help alleviate the curse of dimensionality.

# Chapter 2

# Material and methods

Chapter 2 is adapted from the “Material and methods” sections published by Le Graverand et al. (2022, 2023). Reprinted parts are highlighted with quotation marks. Complements were added to fully cover the thesis work.

The present section covers all material and methods employed throughout the manuscript (Chapters 3 to 7): from the population under study, to the phenotyping of sheep production traits, collection of omics data, and statistical methods.

# I-Population

## A. Breed

Data were collected in Romane sheep, a breed formerly known as INRA 401. The Romane is a synthetic breed obtained by crossing two others: the Russian Romanov and the French Berrichon du Cher (Tchamitchian et al., 1986). The crossbreeding goal was to combine the prolificacy and maternal qualities of the Russian breed with the carcass qualities of the French breed (Moreno et al., 2001). The Romane population is considered as fixed and under selection since 1980 (Tchamitchian et al., 1986).

“The present study was conducted from 2018 to 2020 on Romane male lambs, at the INRAE Experimental Unit P3R (UE 332 agreement D18-174-01; Osmoy, France, https://doi.org/10.15454/1.5483259352597417E12)” (Le Graverand et al., 2023).

## B. Divergent lines

“The studied animals were part of a larger design of divergent selection on RFI. Since 2009, feed efficiency was phenotyped in Romane male lambs under a 100% concentrate diet. In 2015, a divergent selection experiment started: two divergent lines were selected for an increased RFI (least feed efficient line, RFI+) or a decreased RFI (most efficient line, RFI-) as described in Tortereau et al. (2020). Briefly, animals were divergently selected for RFI under a 100% concentrate diet. For the genetic evaluation, the RFIs of 1 900 male and female lambs phenotyped since 2009 were calculated according to Tortereau et al. (2020). Estimated breeding values (EBVs) were then computed with PEST software (Groeneveld et al.), considering a heritability of 0.45 (Tortereau et al., 2020) and a pedigree of 6 419 sheep. The mixed model accounted for the pen, suckling method (maternal or artificial), litter size, year of phenotyping and sex as fixed effects.

Every year since 2015, ten to fourteen rams with the lowest and highest RFI breeding values under a concentrate diet were selected for breeding to produce the next generation of selection. Mating was planned within divergent lines: rams with extreme RFI values were selected and mated with ewes chosen to minimize inbreeding. Two years were necessary to complete one generation of selection.

The present study focused only on a subset of Romane male lambs from the divergent lines described above. Part of the second generation was phenotyped in 2018 (103 lambs), while the third generation was phenotyped in 2019 (101 lambs) and 2020 (73 lambs). Part of the lambs phenotyped in 2018 sired lambs studied in 2020. On average, 7.7 male lambs shared the same sire in the study" (Le Graverand et al., 2023).

## II- Phenotyping production traits

Figure 10 illustrates the experimental design. "Animals were reared indoors with wood chip litter and fed successively with two different diets: the first one with concentrates exclusively and the second one with a mixed diet" (Le Graverand et al., 2023).

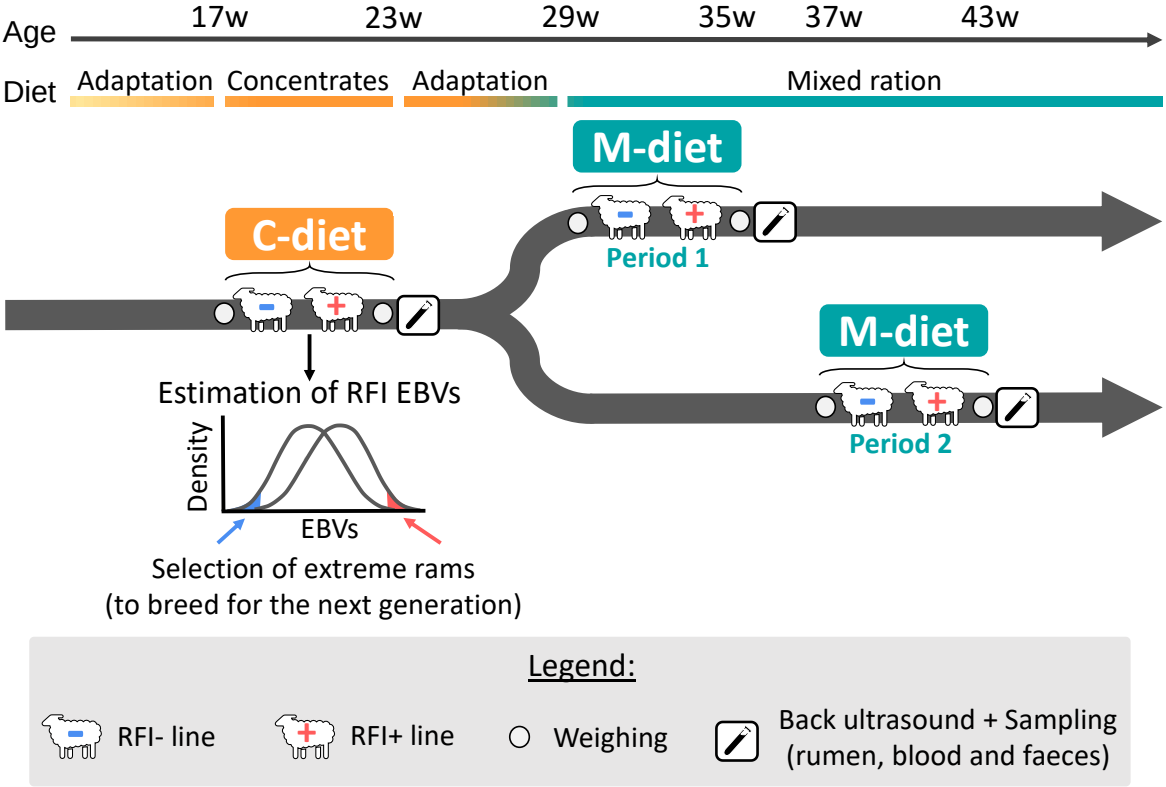



**Figure 10: Experimental design of lamb phenotyping (adapted from Le Graverand et al., 2023)**

## A. Concentrate diet

"Every year, lambs born in the experimental unit were gathered in the experimental barn at approximately 10 weeks of age. Then, lambs were adapted to a 100% concentrate *ad libitum* diet and to its distribution by automatic feeders (for feed nutritional values, see Table 2)" (Le Graverand et al., 2023).

**Table 2: Dietary characteristics of lamb feeds under the concentrate diet (reprinted from Le Graverand et al., 2023)**

| | | Feed characteristics[1] | |
|---|---|---|---|
| **Feed** | **Year** | **DM (g/kg)** | **NE (MJ/kg of DM)** |
| Concentrate | 2018 | 905.05 | 5.78 |
| Concentrate | 2019 | 888.11 | 5.89 |
| Concentrate | 2020 | 892.5 | 5.86 |

[1]: DM: Dry Matter (grams per kilogram of feed); NE: Net Energy (megajoules per kilogram of dry matter)

"Feed intake was recorded using automatic concentrate feeders during six weeks, between 17 and 23 weeks of age on average (Figure 10). Depending on the year, lambs were grouped into four to six pens of approximately 20 lambs with homogeneous body weights to prevent fights. One concentrate feeder was available per pen. From 2018 to 2020, a total of 277 lambs were phenotyped under a concentrate diet.

To account for annual variations of feed compositions, net energy intake was computed from feed intake and feed energy densities. Energy densities of feed were estimated with the INRA 2007 system (Baumont et al., 2007) in megajoules of net energy. The average daily energy intake of concentrate (**ADEI**$_C$) was calculated as the average of daily energy intakes over the six-week period. At the end of the recording period, back ultrasound measurements were carried out to assess the *longissimus dorsi* muscle depth (**MD**$_C$) and back fat thickness (**BFT**$_C$). Starting and final body weights were recorded and used to compute the average daily gain (**ADG**$_C$) over six weeks. Two feed efficiency traits were computed: the feed conversion ratio (**FCR**$_C$) as the ratio of ADEI$_C$ over ADG$_C$ and, REI$_C$ as the residuals of the linear regression of ADEI$_C$ over characterized energy sinks (Equation 1):

$$\mathrm{ADEI_C} = \mu_C + \beta_{1,C}\,\mathrm{ADG_C} + \beta_{2,C}\,\mathrm{final\ BW_C}^{0.75} + \beta_{3,C}\,\mathrm{MD_C} + \beta_{4,C}\,\mathrm{BFT_c} + \mathrm{REI_C} \quad \text{(Eq. 1)}$$

where $\mu_C$ is the mean ADEI$_C$. $\beta_{1,C}$ to $\beta_{4,C}$ respectively stand for the four following covariate effects: ADG$_C$, final metabolic weight ($\mathbf{final\ BW_C^{0.75}}$), MD$_C$ and BFT$_C$. Finally, REIc is the residual energy intake expressed as megajoules of ingested net energy per day" (Le Graverand et al., 2023).

Rumen fluids, blood and faeces were sampled at the end of trials: around 23 weeks of age, under a concentrate diet. "Every year, rumen fluid samples were collected at the end of each feed intake recording trial. Sampling was carried out by trained staff, with a medical gastric tube coupled to a vacuum pump. Ruminal samples were immediately frozen in liquid nitrogen" (Le Graverand et al., 2023). The jugular vein blood was sampled with two different vacutainer tubes. One of the animal's blood samples was kept in EDTA tubes. The second blood sample was centrifuged (2400 × g during 10 minutes) in heparin lithium tubes, before retrieving and freezing the plasma. Faeces were collected directly from the rectum.

## B. Mixed diet

"After the first six-week trial, animals were adapted for 6 weeks to a mixed diet delivered by automatic feeders. In 2018, no restricted concentrate feeder was available; thus, forage feeders delivered an *ad libitum* total mixed ration (33% concentrates, 67% hay). In 2019 and 2020, automatic concentrate feeders ensured restricted access to concentrates (up to 700 g/day), and forage feeders delivered *ad libitum* hay separately (see Table 3 for feed nutritional values)" (Le Graverand et al., 2023).

**Table 3: Dietary characteristics of lamb feeds under mixed diet (1/3 concentrate and 2/3 forage approximately, reprinted from Le Graverand et al., 2023)**

| | | Feed characteristics[1] | |
|---|---|---|---|
| **Feed** | **Year** | **DM (g/kg)** | **NE (MJ/kg of DM)** |
| Concentrate | 2018 | 882.2 | 6.41 |
| Concentrate | 2019 | 874.2 | 6.39 |
| Concentrate | 2020 | 894.1 | 6.33 |
| Orchard hay | 2018 | 914.1 | 3.95 |
| Orchard hay | 2019 (1st period) | 913.5 | 4.22 |
| Orchard hay | 2019 (2nd period) | 915.6 | 4.17 |
| Orchard hay | 2020 | 899.5 | 3.89 |

[1]: DM: Dry Matter (grams per kilogram of feed); NE: Net Energy (megajoules per kilogram of dry matter)

“Due to facility limitations (maximum of 35-40 lambs simultaneously), only lambs having extreme RFI EBVs under a concentrate diet were then phenotyped under a mixed diet. In 2018 and 2019, animals were split during two different periods per year: during summer (period 1, from 29 to 35 weeks of age) or during fall (period 2, from 37 to 43 weeks of age) (Figure 10). In 2020, only one group was phenotyped during fall (period 2). Within one period, animals were allocated by body weight into two pens with 16 animals per pen, on average. Each pen was equipped with two forage feeders and one concentrate feeder in 2019 and 2020. Over the three years, 166 lambs were phenotyped under a mixed diet.

$ADEI_M$, $ADG_M$ and $REI_M$ were estimated under the mixed diet as it was done under the concentrate diet before. However, no $FCR_M$ was calculated under the mixed diet because some individuals had a null or negative $ADG_M$ over this 6-week period” (Le Graverand et al., 2023).

Rumen fluids were sampled at the end of trials: around the age of 35 weeks (Period 1) or 43 weeks (Period 2), under a mixed diet.

## C. Zootechnical data cleaning

“Under one diet, outliers were identified based on feed efficiency (REI, FCR), energy intake (ADEI), growth rate (ADG) and body composition (MD, BFT) traits. The animal was removed from the diet data subset when one of the phenotypes was outside the range [µ - 3 SD; µ + 3 SD], where µ is the mean phenotype and SD its standard deviation. Eight animals were filtered out under a concentrate diet (out of 277 lambs), and three were filtered out under a mixed diet (out of 166 animals)” (Le Graverand et al., 2023).

# III- Omics data acquisition and processing

## A. Genotyping

DNA was extracted from blood samples by LABOGENA (Jouy-en-Josas, France). Genotyping was carried out with the Illumina Ovine SNP50 chip (54 241 SNPs) by LABOGENA.

### A.1 Quality control and imputation

SNP quality control was not stringent (step 1, Figure 11), in order to impute missing genotypes later (step 2, Figure 11). Samples, with an individual genotyping call rate lower than 0.80 were discarded. Out of 570 samples, 567 were kept. SNPs were discarded if they had a call frequency lower than 0.90, a minor allele frequency lower than 0.01 or a p-value lower than $10^{-6}$ for the Hardy-Weinberg equilibrium test. Finally, SNPs were left out if they were not placed on the 26 autosomes, based on the oar_v3.1 reference genome assembly (Jiang et al., 2014). Out of 54 241 SNPs, 42 759 were retained.

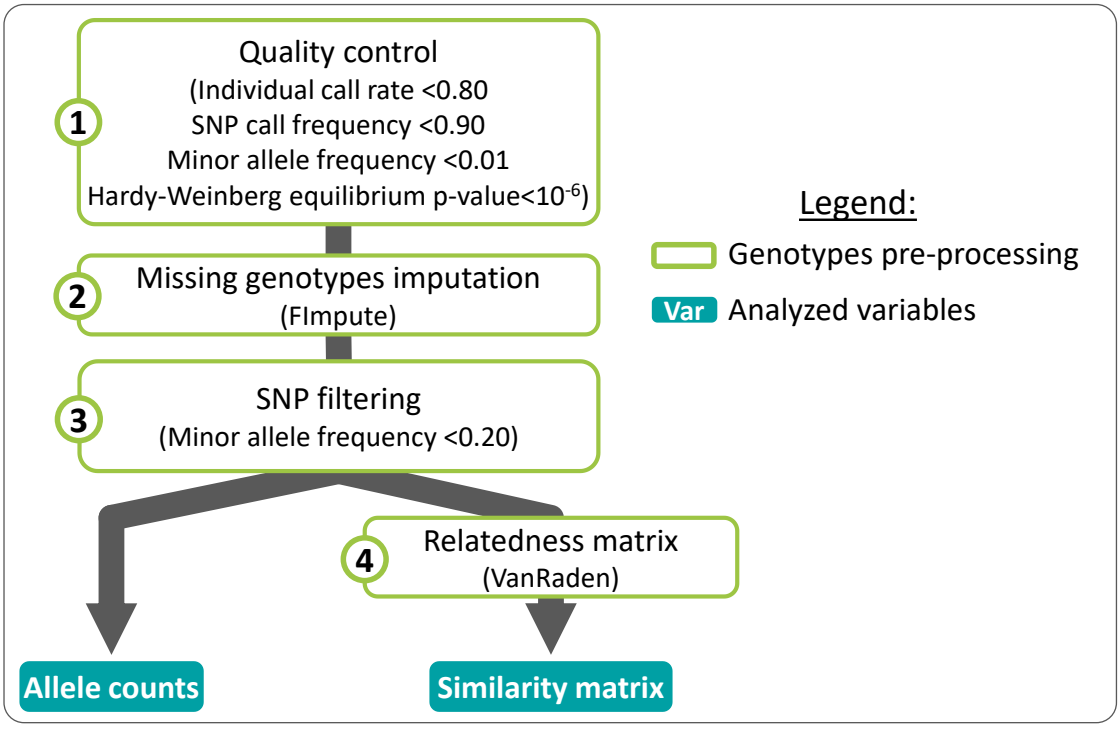



**Figure 11: Workflow of genomics data handling**

No reference population was available to impute missing genotypes in Romane sheep. Missing genotypes were imputed with the FImpute software (step 2, Figure 11) (Sargolzaei et al., 2014). Family imputation was first carried out with pedigree data and 6 428 related

individuals. Then, population imputation relied on the haplotypes defined with 567 genotyped individuals. Mendelian errors -mismatches between progeny and parent genotypes- were checked and corrected with FImpute default parameters (error rates equal to 0.01 and 0.005 to identify mismatches and matches, respectively).

### A.2 Data filtering and transformation

After imputation and prior to statistical learning, SNPs with a minor allele frequency lower than 0.20 were filtered out (step 3, Figure 11). Retaining 29 830 SNPs decreased the computational burden in subsequent statistical learning analyses.

Prior to principal component analysis only, VanRaden's first genomic relationship matrix (VanRaden, 2008) was computed from retained SNPs with the R package AGHmatrix (Amadeu et al., 2016) (step 4, Figure 11).

## B. Rumen microbiota

### B.1 DNA extraction and sequencing

"DNA extraction, amplification and sequencing of microbial fluid samples were carried out in two different batches: one batch with 2018 and 2019 samples and another with 2020 samples. Within each batch, a bead-beating step was carried out with a FastPrep device (MP Biomedicals, Illkirch, France). Then, DNA was extracted with the QIAamp DNA Stool Mini Kit (Qiagen Ltd, West Sussex, UK) from 85 µL of ruminal fluid.

The V4-V5 region of the 16S ribosomal RNA gene was amplified with the forward 515F (5'-CTTTCCCTACACGACGCTCTTCCGATCTGTGYCAGCMGCCGCGGTA-3') and reverse 928R primers (5'-GGAGTTCAGACGTGTGCTCTTCCGATCTCCCCGYCAATTCMTTTRAGT-3') (Wang and Qian, 2009) for 30 PCR cycles. To barcode samples, an index of 6 base pairs was added to 928R primers during a second amplification (12 cycles) with forward (5'-GTGYCAGCMGCC-3') and reverse primers (5'-CCCCGYCAATT-3'), plus adapters. Overlapping paired-end reads of 250 base pairs were produced and aligned to obtain full-length reads with Illumina MiSeq technology (Illumina, San Diego, CA, USA).

The V4 region of the 18S ribosomal RNA gene was amplified with the forward 566F (5'-CTTTCCCTACACGACGCTCTTCCGATCTCAGCAGCCGCGGTAATTCC-3') and reverse 1200R

primers (5'-GGAGTTCAGACGTGTGCTCTTCCGATCTCCCGTGTTGAGTCAAATTAAGC-3') (Hadziavdic et al., 2014) for 30 PCR cycles. To barcode samples, an index of 6 base pairs was added to 1200R primers during a second amplification (12 cycles) with forward (5'-CAGCAGCCGCGGTAATTCC-3') and reverse primers (5'-CCCGTGTTGAGTCAAATTAAGC-3'). Overlapping and non-overlapping paired-end reads of 250 base pairs were produced.

After multiplexing and amplifications, 16S and 18S reads were purified and loaded on an Illumina MiSeq cartridge (Illumina, San Diego, CA, USA) to be sequenced at the Genomic and Transcriptomic Platform (INRAE, Toulouse, France) " (Le Graverand et al., 2023).

## B.2 Bioinformatic processing of microbiota data

### Clustering into operational taxonomic units

"After sequencing, 16S and 18S reads were processed separately with FROGS tools (version 4.0.1) (Escudié et al., 2018). However, within both sets of sequences, reads from different diets and sequencing batches were treated together (step 1, Figure 12). Read processing was performed with the following pipeline: (i) demultiplexing; (ii) reconstruction with 18S amplicon sequences only, (iii) quality control of amplicons based on the presence of primers, ambiguous bases and size (>380 and <500 base pairs for 16S amplicons; >200 and <490 base pairs for 18S amplicons); (iv) clustering into operational taxonomic units (**OTU**s) with the Swarm algorithm using a difference of 1 between sequences in each aggregation step (Mahé et al., 2014); (v) chimaera removal; (vi) pre-filtering, by removing OTUs aggregating less than 0.005% of all sequences; and (vii) taxonomic affiliation with BLAST+ (Camacho et al., 2009) and the Silva 132 16S reference database for bacteria and archaea, or the Silva 138.1 18S database for fungi and protozoa (Quast et al., 2013) " (Le Graverand et al., 2023).

### Functional inference

Functions of 16S OTUs were inferred under a concentrate diet only, by Guibert et al. (Unpublished). Briefly, inference was carried out thanks to FROGSFUNC tools built upon the PICRUSt2 method (Douglas et al., 2020; Darbot et al., 2022). FROGSFUNC affiliates OTUs to a reference phylogenetic tree. Then, the number of 16S gene copies per OTU is determined. Finally, the number of function copies is inferred from the phylogenetic affiliation, the number of 16S gene copies and the abundance of OTUs.

OTUs were excluded from functional inference when their taxonomic affiliation was unreliable. Affiliations were deemed unreliable when: the alignment identity was low (<90% between the OTU seed and the reference sequence), the coverage was low (<90%), or the nearest sequenced taxon index was high (>1.0).

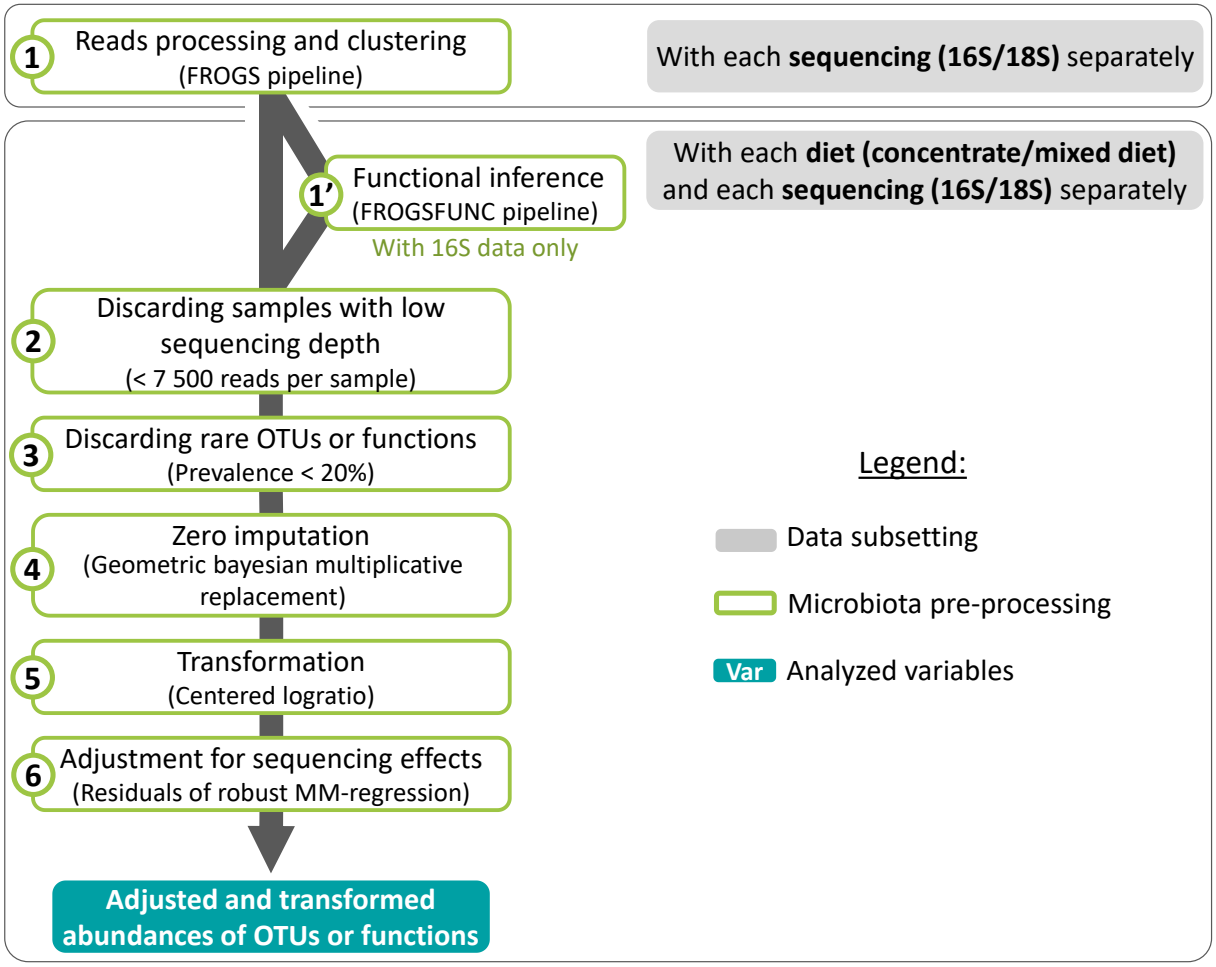



**Figure 12: Workflow of microbiota data handling (adapted from Le Graverand et al., 2023)**

## B.3 Microbiota data cleaning and transformations

"For data cleaning and subsequent analyses (steps 2 to 6, Figure 12), four distinct compositional datasets were considered: one per diet (concentrate or mixed diet) and amplified gene (16S or 18S). Then, sequencing data were filtered at the sample and OTU levels. […] Samples with a [sequencing] depth smaller than 7 500 reads were discarded (step 2, Figure 12). OTU filtering was performed by removing OTUs with a prevalence lower than 20% under the considered diet (step 3, Figure 12) " (Le Graverand et al., 2023). The richness (numbers of OTUs) is reported per sample, before and after data filtering (Table 4).

**Table 4: Number of OTUs and functions detected in the rumen fluid, prior and posterior to data cleaning (adapted from Le Graverand et al., 2023)**

**Before data cleaning**

| | **Diet[1]** | **C-diet** | | **M-diet** | |
|---|---|---|---|---|---|
| | Sequencing (n[2]) | 16S (277) | 18S (275) | 16S (163) | 18S (166) |
| OTUs[3] | Total per dataset | 1 298 | 263 | 1 527 | 269 |
| | Mean per sample | 325 | 60 | 744 | 99 |
| | Min per sample | 185 | 8 | 374 | 34 |
| | Max per sample | 476 | 153 | 993 | 176 |
| Functions[4] | Total per dataset | 301 | - | - | - |
| | Mean per sample | 268 | - | - | - |
| | Min per sample | 236 | - | - | - |
| | Max per sample | 292 | - | - | - |

**After data cleaning**

| | **Diet[1]** | **C-diet** | | **M-diet** | |
|---|---|---|---|---|---|
| | Sequencing (n[2]) | 16S (269) | 18S (205) | 16S (160) | 18S (161) |
| OTUs[3] | Total per dataset | 582 | 124 | 1 148 | 183 |
| | Mean per sample | 296 | 57 | 720 | 91 |
| | Min per sample | 145 | 8 | 363 | 33 |
| | Max per sample | 415 | 95 | 918 | 152 |
| Functions[4] | Total per dataset | 281 | - | - | - |
| | Mean per sample | 266 | - | - | - |
| | Min per sample | 236 | - | - | - |
| | Max per sample | 280 | - | - | - |

[1]: C-diet: animals phenotyped under a 100% concentrate diet; M-diet: phenotyped under a mixed diet.
[2]: Number of samples
[3]: Operational taxonomic units
[4]: Inferred microbial functions (FROGSFUNC). Inference was only carried out with 16S data collected under a concentrate diet

"Several pre-processing steps were carried out to account for compositionality and sequencing effects. Null abundances of OTUs were imputed with the geometric Bayesian multiplicative replacement procedure (step 4, Figure 12) (Martín-Fernández et al., 2015).

Then, a centered logratio (CLR) transformation was carried out with OTU abundances (step 5, Figure 12). CLR coordinates values were adjusted univariately for sequencing effects with a robust MM regression (Maechler et al., 2017) (Equation 4, step 6, Figure 12). Robust MM regression was preferred as it is less sensitive to outlying values than least squares linear regressions.

$$\mathrm{CLR(X)_i} = \beta_{0,\mathrm{i}} + \beta_{1,\mathrm{i}}\,\mathrm{seq_{depth}} + \beta_{2,\mathrm{i}}\,\mathrm{seq_{batch}} + \beta_{3,\mathrm{i}}\,\mathrm{seq_{batch|plate}} + \varepsilon_{\mathrm{i}} \quad (4)$$

where $CLR(X)_i$ stands for the CLR values of the i$^{th}$ OTU. Then, $\beta_{0,i}$ stands for the intercept. For the i$^{th}$ OTU, $\beta_{1,i}$ to $\beta_{3,i}$ represent the effects of sequencing depth, sequencing batches (n=2) and sequencing plates nested in batches (n=5), respectively. Finally, $\varepsilon_i$ stands for residuals" (Le Graverand et al., 2023).

# C. Rumen and plasma metabolomics

Touitou et al. (2022) previously described the plasma and rumen metabolomes in the same sheep population. Metabolite quantifications were assessed under a concentrate diet and mixed diet by Touitou et al. (2022). On the other hand, the present thesis will only focus on metabolite quantifications and buckets, under a concentrate diet only.

## C.1 Nuclear magnetic resonance spectroscopy

Rumen and plasma samples were unfrozen and centrifuged (3000× g at 4°C during 5 minutes). Then, 200 µL of sample supernatant were mixed with 500 µL of phosphate buffer (pH=7). Subsequently, 600 µL of the mix were centrifuged again (4190 × g at 4°C during 15 minutes). The last centrifugation was repeated again for ruminal samples only, to further dilute rumen fluids.

Rumen and plasma samples were analyzed through Nuclear Magnetic Resonance (**NMR**) spectroscopy in two different batches. Sample spectra of 2018 and 2019 were acquired together, while samples of 2020 were processed in a second batch. NMR spectroscopy was carried out at the MetaToul-AXIOM platform (MetaboHUB-ANR-11-INBS-0010, 2011). Spectra were obtained with the Bruker AVANCE III HD 600 MHz spectrometer (Bruker Biospin, Rheinstetten, Germany) and the cpmgpr1D Bruker pulse program. More details are available in Touitou et al. (2022).

## C.2 Bioinformatic processing of NMR spectra

Pre-processing was carried out separately for rumen and plasma samples. Spectra were first processed with Bruker's TopSpin® software (version 4.0.9, Billerica, MA, USA) to correct the zero-order phase and the baseline (step 1, Figure 13). Depending on the biological fluid, chemical shift calibration relied on two different molecules: D-glucose in plasma samples (naturally present), or trimethylsilylpropanoic acid (present in the phosphate buffer) in rumen samples.

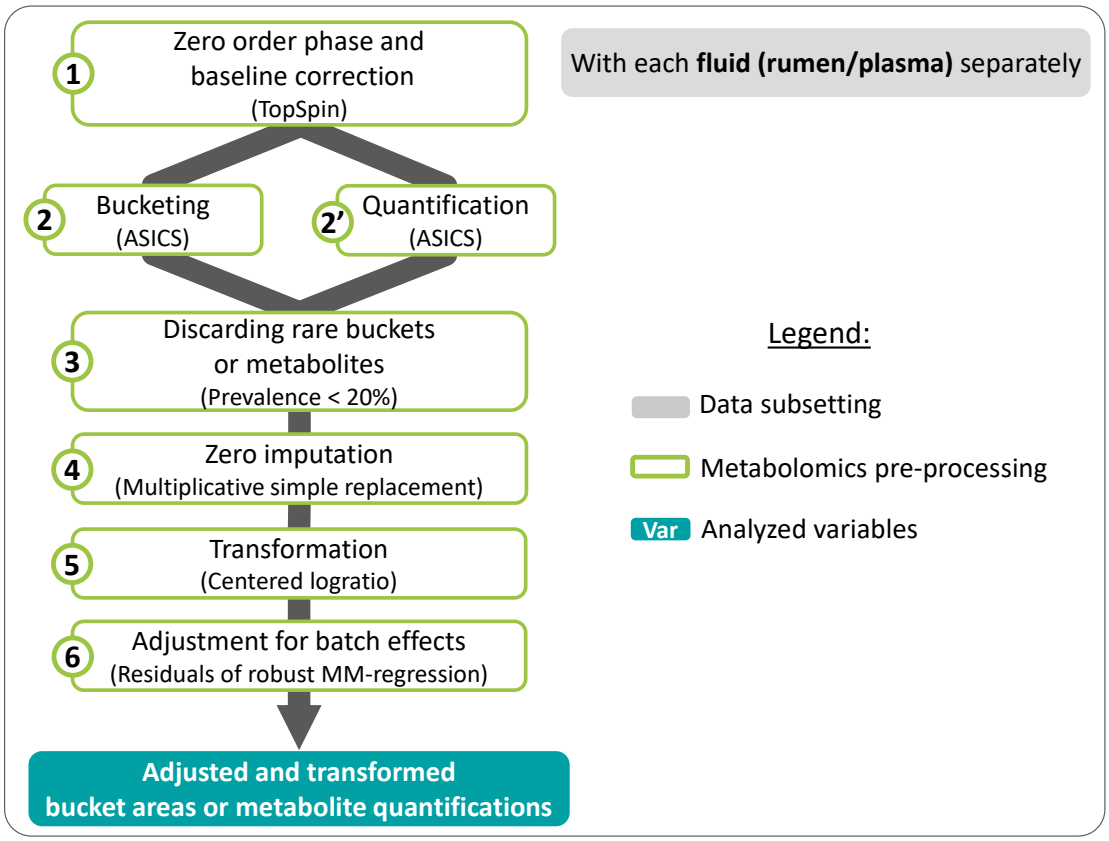



**Figure 13: Workflow of metabolomics data handling**

Then, metabolomic data were processed with the ASICS R package (Lefort et al., 2019). First, spectra areas were normalized to a constant sum, and solvent signals were removed from analysis (water region between 4.5 and 5.1 ppm). ASICS allowed to treat metabolomic data in two different ways (step 2 and 2', Figure 13). First, the "binning" function was used to get buckets from metabolomic data: spectra were divided into parts of equal width (bin option equal to 0.01 ppm). Buckets represent the area under the spectral curve. Second, the "ASICS" function was used to identify metabolites from spectra through deconvolution (maximum chemical shift option equal to 0.01 ppm, noise option equal to 0.02). Identified

metabolites were quantified with ASICS. Metabolite quantifications are expressed as relative concentrations, relatively to the highest concentration (Tardivel et al., 2017).

## C.3 Metabolomic data cleaning and transformations

Under a concentrate diet, four distinct compositional datasets were considered during data cleaning (steps 3 to 6, Figure 13): one per fluid (rumen or plasma) and bioinformatic treatment (buckets or metabolites). Then, metabolomic data were filtered: buckets and metabolites were discarded when they were detected in less than 20% of animals (step 3, Figure 13). The numbers of metabolites and buckets are reported in Table 5, before and after data cleaning.

**Table 5: Number of metabolites and buckets above the detection limit, prior and posterior to data cleaning**

| | | Before data cleaning | | After data cleaning | |
|---|---|---|---|---|---|
| | Diet[1] | C-diet | | C-diet | |
| | Fluid (n[2]) | Rumen (275) | Plasma (275) | Rumen (275) | Plasma (275) |
| Buckets[3] | Total per dataset | 880 | 879 | 871 | 863 |
| | Mean per sample | 851 | 857 | 850 | 856 |
| | Min per sample | 820 | 848 | 818 | 848 |
| | Max per sample | 871 | 864 | 866 | 863 |
| Metabolites[4] | Total per dataset | 72 | 57 | 25 | 25 |
| | Mean per sample | 15 | 21 | 13 | 19 |
| | Min per sample | 2 | 13 | 2 | 13 |
| | Max per sample | 31 | 29 | 22 | 23 |

[1]: C-diet: animals phenotyped under a 100% concentrate diet. Under the mixed diet, metabolomic data were not analyzed during the thesis.
[2]: Number of samples
[3]: Buckets defined with ASICS (fixed width of 0.01 ppm)
[4]: Metabolite quantifications inferred with ASICS.

Like microbiota data, metabolome buckets and metabolite quantifications were deemed compositional. Thus, similar pre-processing steps were carried out: null values were imputed with the Multiplicative simple replacement procedure (Martín-Fernández et al., 2015), and the CLR was applied (steps 4 and 5, Figure 13). CLR values were adjusted with robust MM regressions (Maechler et al., 2017), see Equation 5 for all four metabolomic datasets (steps 6, Figure 13):

$$\mathrm{CLR(X)_i} = \beta_{0,\mathrm{i}} + \beta_{1,\mathrm{i}}\,\mathrm{Batch} + \beta_{2,\mathrm{i}}\,\mathrm{Fasting} + \varepsilon_{\mathrm{i}} \quad (5)$$

where $CLR(X)_i$ stands for the CLR values of the i$^{th}$ metabolite or bucket. Then, $\beta_{0,i}$ stands for the intercept, and $\beta_{1,i}$ for the spectroscopy batch effect (n=2). $\beta_{2,i}$ denotes the fasting effect (n=2): animals were supposed to fast at least 10 hours prior to rumen and metabolome sampling. However, in 2019, feeders kept delivering feed and animals did not fast. Finally, $\varepsilon_i$ stands for residuals.

# D. Rumen lipidomics

Touitou (2023) described the rumen lipidome in the same sheep population. The lipidome will be analyzed under a concentrate diet only, in this thesis.

The rumen lipidome was analyzed through two angles: volatile fatty acids (**VFA**s) and long-chain fatty acids (**LCFA**s). Lipidome profiling protocols are summarized below, whereas details are provided in Touitou (2023).

## D.1 Gas chromatography and mass spectrometry

VFA and LCFA analyses relied on gas chromatography first. LCFAs were additionally analyzed through mass spectrometry.

### Volatile fatty acids

Rumen samples were unfrozen and centrifuged (2 880 x g, during 20 minutes at 4°C). Then, 1mL of supernatant was analyzed according to the protocol of Playne (1985). Gas chromatography was carried out thanks to a chromatograph equipped with a flame ionization detector and a column of 30 meters (Agilent 7890A and column n°125-3232, Agilent Technologies, Santa-Clara, USA).

VFAs concentrations were determined thanks to internal standard solutions (with 4-methylvaleric acid), and the Chromeleon software (version 7.2.10, ThermoFisher Scientific, Waltham, USA).

**Note:** Note that pseudo-quantifications of metabolites were obtained with NMR spectra and ASICs (quantifications were expressed relatively to the highest one). Actual concentrations were obtained for VFAs and LCFAs with chromatography and an internal standard (concentrations expressed in mmol/L).

#### Long-chain fatty acids

Rumen samples were lyophilized, before analyzing 250 mg according to the protocol described in Alves et al. (2013). Gas chromatography was performed with a chromatograph (Shimadzu GC 2010 Plus, Shimadzu, Kyoto, Japan) coupled with a flame ionization detector, a column of 100 meters (Supelco FS CAP SLB-IL111, Sigma-Aldrich, Saint-Louis, USA) and a mass spectrometric detector.

LCFAs were quantified thanks to an internal standard (nonadecanoic acid) and the GCsolution software (Shimadzu, Kyoto, Japan).

### D.2 Fatty acids data cleaning and transformations

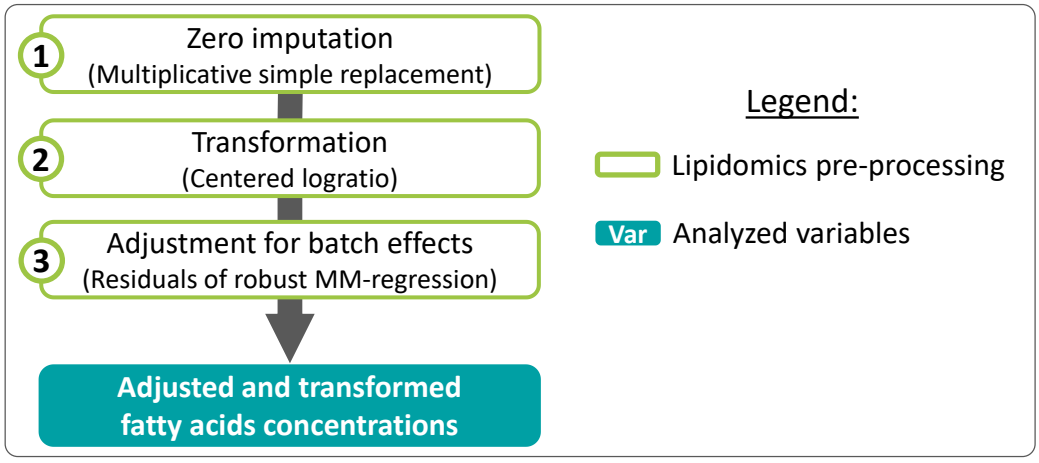



**Figure 14: Workflow of lipidomics data handling**

Fatty acids were not filtered according to their detected prevalence: all VFAs and LCFAs were detected in more than 20% of the samples. Then, fatty acids concentrations were pre-processed exactly like metabolite quantifications (section III.C.3): zeroes were imputed, concentrations were transformed by the CLR and adjusted for batch and fasting effect (steps 1 to 3, Figure 14). VFAs and LCFAs were processed separately. No fatty acid was filtered during data cleaning (all prevalences were higher than 20%), the number of fatty acids is reported in Table 6 per dataset.

**Table 6: Number of fatty acids above the detection limit, prior and posterior to data cleaning**

| | | Before and after data cleaning |
|---|---|---|
| | **Diet[1]** | **C-diet** |
| | Fluid (n[2]) | Rumen (277) |
| VFA[3] | Total per dataset | 6 |
| | Mean per sample | 6 |
| | Min per sample | 6 |
| | Max per sample | 6 |
| LCFA[4] | Total per dataset | 70 |
| | Mean per sample | 68 |
| | Min per sample | 62 |
| | Max per sample | 70 |

[1]: C-diet: animals phenotyped under a 100% concentrate diet. Under the mixed diet, metabolomic data were not analyzed during the thesis.
[2]: Number of samples
[3]: Volatile fatty acids
[4]: Long-chain fatty acids

# E. Faecal phenomics

In the present thesis, the faecal spectral phenome will be analyzed under a concentrate diet only. The faecal phenome was characterized thanks to near-infrared spectroscopy (**NIRS**).

## E.1 Near-infrared spectroscopy

Details about the NIRS protocol can be found in Andueza et al. (2017). Succinctly, faecal samples were homogenized and dried. Then, around 5 g of samples were scanned thanks to the NIRSystems model 6500 spectrometer (Foss NIRSystems, Silver Spring, MD, USA) and the NIRS3 software (Infrasoft International, Port Matilda, PA, USA). Every spectrum was obtained by time-averaging 32 scans. Reflectance was recorded between 400 and 2500 nm, with an interval of 2 nm, and converted into absorbance.

## E.2 NIR spectra data cleaning and transformations

No variable was discarded, since all absorbance values were non-null. Thus, 1050 variables were kept and recorded over 275 faecal samples. NIRS absorbances were either analyzed directly as raw data, or transformed (Figure 15). When transformed, data were first normalized thanks to the standard normal variate transformation and de-trending

correction (Barnes et al., 1989). Then, the first-order derivative was computed (over gaps of 4 points, with 4 points in the first smoothing) (Andueza et al., 2017).

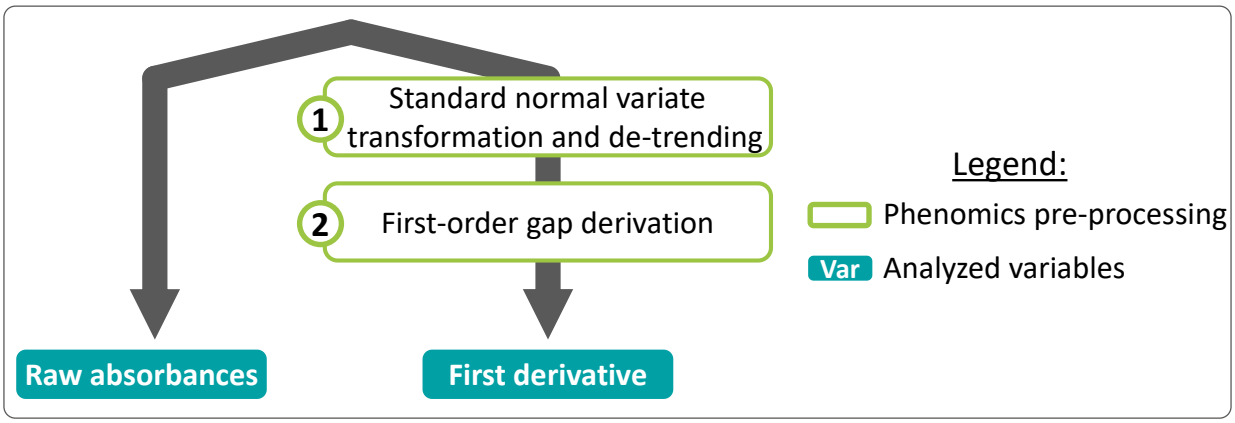



**Figure 15: Workflow of phenomics data handling**

# IV-Statistical analyses

Statistical analyses filled two purposes. The first section, details how the dataset was explored. The second section describes how sheep lines and production traits were predicted.

## A. Data overview

Data overview relied on univariate and multivariate statistical analyses. First, the overview of sheep production traits relied on univariate inferential statistics: the objective is to test if the two divergent feed efficiency lines significantly differed. Then, the overview of omics relied on multivariate unsupervised learning: the aim is to visualize what were the main patterns of omics variations.

### A.1 Sheep production traits

First, least square means were computed per line and host trait to test if lines diverged significantly. The tested host traits were: feed efficiency (REI, FCR), energy intake (ADEI), body weight, growth (ADG) and body composition (MD, BFT). “Linear regressions were fitted with the lm function included in R software (version 4.1.2) (R Core Team, 2021). All traits recorded under a concentrate diet ($\mathrm{Trait_C}$) were regressed as follows (Equation 6):

$$\text{Trait}_C = \mu_{\text{Trait}_C} + \beta_{1,C}\,\text{line} + \beta_{2,C}\,\text{age} + \beta_{3,C}\,\text{suckling} + \beta_{4,C}\,\text{year} + \beta_{5,C}\,\text{pen|year} + \varepsilon_C \quad (6)$$

Where $\mu_{\text{Trait}_C}$ denotes the trait mean. $\beta_{1,C}$ to $\beta_{5,C}$ respectively stand for the effects of the line, age, suckling, year and pen nested in the year. Finally, $\varepsilon_C$ denotes the model residuals.

Traits recorded under a mixed diet ($\text{Trait}_M$) were regressed similarly (Equation 7):

$$\text{Trait}_M = \mu_{\text{Trait}_M} + \beta_{1,M}\,\text{line} + \beta_{2,M}\,\text{age} + \beta_{3,M}\,\text{year} + \beta_{4,M}\,\text{period|year} + \beta_{5,M}\,\text{pen|period|year} + \varepsilon_M \quad (7)$$

Where $\mu_{\text{Trait}_M}$ denotes the trait mean. $\beta_{1,M}$ to $\beta_{5,M}$ respectively stand for the effects of the line, age, year, period nested in the year, and pen nested in the period and the year. $\varepsilon_M$ denotes the model residuals.

Then, regression solutions (Equations 6 and 7) were used to compute least square means per line and host trait with the emmeans R package (version 1.7.3) (Lenth, 2022). Comparison of traits between RFI lines were carried out with the Tukey test. To account for multiple testing, p-values were corrected per diet with Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995) (7 traits under a concentrate diet and 6 under a mixed diet)" (Le Graverand et al., 2023).

## A.2 Omics data

Unsupervised learning was applied to identify the main patterns of omics variations (Analysis A, Figure 16). The analyzed omics were: genomics, rumen metataxonomics, rumen and plasma metabolomics, rumen lipidomics and fecal phenomics. All omics were pre-processed separately prior to analyses (see section III). Principal component analyses (**PCA**s) were carried out per omics block and diet, with the mixOmics R package (version 6.18.1) (Rohart et al., 2017b). Recorded environmental, physiological and technical effects were compared to data projections on the first principal components (**PC**s), to label the main variation patterns.

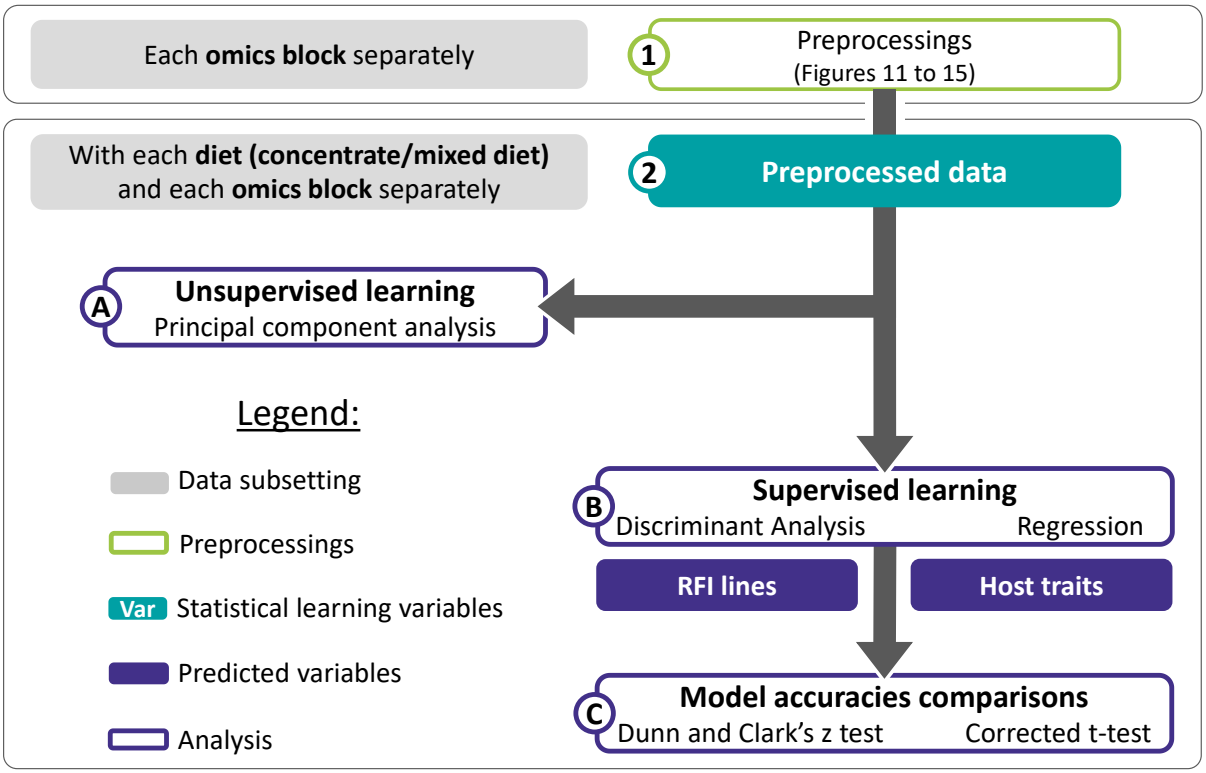



**Figure 16: Workflow of statistical learning with omics data**

# B. Prediction of sheep lines and production traits

Predictions relied on multivariate supervised learning: either discriminant analyses or regressions. The goal of discriminant analyzes was to predict the RFI line from systematic effects (*i.e.* year, pen, suckling, age and body weight effects), pedigree and/or omics data. The goal of regressions was to predict either feed efficiency, energy intake, growth or body composition. This section will first detail how cross-validation was used to select the model hyperparameters and assess the predictive accuracy of models. Then, models will be detailed including the integration of heterogeneous variables and years. Finally, the genetic evaluation of predicted traits will be detailed.

## B.1 Cross-validating and testing the difference between prediction accuracies

Cross-validation strategies varied between studies. Statistical tests were carried out to assess the difference in prediction accuracy, between models trained on different predictors. Choosing a test depended on the cross-validation strategy. With repeated cross-validation, tests had to account for the violation of one assumption: all accuracy estimates are not independent. Indeed, one animal can be part of different testing sets when cross-validation is repeated.

**Repeated k-fold cross-validation (Chapters 3 and 5)**

The first strategy relied on 5-fold cross-validation repeated 50 times. The dataset was split into 5 folds. Four folds -the training set- were used to build the model. The fifth fold -the testing set- was used to select model hyperparameters and evaluate the model accuracy. Thanks to a custom function, cross-validation was stratified per feed efficiency line, year and pen: *i.e.* proportions of each line, year and pen were approximately equal in all training and testing sets. The operation was repeated 50 times, which led to 250 estimates of prediction accuracy per model.

With repeated k-fold cross-validation, accuracy differences between models fitted on different predictors were tested with Bouckaert and Frank's corrected t-test (Bouckaert and Frank, 2004) (Analysis C, Figure 16). Comparisons were made per predicted trait and diet. Benjamini-Hochberg's procedure was applied to adjust p-values (Benjamini and Hochberg, 1995).

**Leave-one-group-out cross-validation (Chapter 4)**

The second strategy relied on leave-one-group-out cross-validation, where groups were defined by the year of phenotyping. Data collected during two different years served as the training set, to build the model. Data of a third independent year was used as the testing set, to select hyperparameters and evaluate the model accuracy. The testing set was either constituted of 2019 or 2020 data. 2018 was not used as a testing set because many animals raised in 2020 had sires phenotyped in 2018. The prediction accuracy was estimated once per year.

With leave-one-group-out cross-validation, differences between accuracies were assessed with Dunn and Clark's z test (Dunn and Clark, 1969). Tests were carried out per testing set. Bonferroni's procedure was used to adjust p-values (Dunn, 1961).

**Repeated random subsampling (Chapter 6)**

The third strategy relied on random subsampling repeated 100 times. The dataset was split into three parts: 60% of the dataset were used as the training set to build the model; 30% constituted the validation set to select the model hyperparameters; 10% were used as the testing set to evaluate the predictive accuracy. The dataset subsampling was stratified per

feed efficiency line, year and pen. The predictive accuracy was assessed 100 times per model.

With repeated random subsampling, differences between accuracies were tested with Nadeau and Bengio's corrected t-test (Nadeau and Bengio, 2003). Models were compared per predicted trait. P-values were adjusted with the Benjamini-Hochberg's procedure.

## B.2 Supervised learning models

Data were preprocessed (see section III) prior to regressions and discriminant analyses. Discriminant analyses were carried out to check whether divergent lines could be predicted from omics data. Regressions were carried out to predict sheep production trait.

### Discriminant analyses

Discriminant analyzes predicted lines from systematic effects and/or omics. Systematic effects included the year, pen, suckling, age and final body weight effects. Sparse partial least-squares discriminant analysis (**sPLSDA**) and multivariate integrative sPLSDA (**MINT-sPLSDA**) were carried out with the mixOmics R package (Rohart et al., 2017b). Discriminant analyses relied on the LASSO algorithm to select predictors in training sets (Tibshirani, 1996). sPLSDA components were built by linearly combining the selected predictors. In MINT-sPLSDA, the construction of components additionally accounted for the year of phenotyping to perform P-integration (Rohart et al., 2017a).

"With sPLSDA [and MINT-sPLSDA], two hyperparameters were tuned thanks to cross-validation: the component number and the number of selected variables per component. The tuning criterion was the Balanced Error Rate (**BER**), calculated as the average of error rates over the RFI lines (Equation 8):

$$\mathrm{BER} = \frac{1}{2} \times \left(\frac{a}{a+b} + \frac{c}{c+d}\right) (8)$$

[with a, b, c and d the numbers of RFI- false predictions, RFI+ true predictions, RFI+ false predictions and RFI- true predictions, respectively].

The following rule of thumb was used to select the final hyperparameters: a more complex model (*i.e.,* with more components and/or more selected variables) was retained if the

averaged BER decreased by one standard error (Kuhn and Johnson, 2013). The average BERs of the selected models are reported" (Le Graverand et al., 2023).

**Regressions**

Regression analyses were performed on raw traits (*i.e.* not adjusted for experimental, physiological or technical effects), given that a genetic evaluation of predicted traits will be carried out later and will account for these effects.

Sparse partial least-squares discriminant regression (**sPLSR**) and multivariate integrative sPLSR (**MINT-sPLSR**) were carried out with mixOmics to predict feed efficiency, energy intake, growth and body composition phenotypes (Rohart et al., 2017b). sPLSR and MINT-sPLSR relied on the LASSO approach to select the best proxies in training sets. The number of components and variables per component were tuned through cross-validations to maximize the prediction accuracy - defined by the coefficient of determination in Chapter 4, the root mean square error in Chapter 5, or the Pearson correlation in Chapter 6.

Support vector regression (**SVR**) was implemented thanks to the e1071 R package (Meyer et al., 2019). In Chapter 4, two hyperparameters were chosen by cross-validating: the kernel (sigmoid, linear, or polynomial of second or third degree) and the regression type (epsilon or nu).

RandomForest regressions (**RFR**) was fitted thanks to the randomForest R package (Liaw and Wiener, 2002). In Chapter 4, one RFR hyperparameter had to be tuned thanks to cross-validation: the number of variables randomly kept as candidates to define decision tree nodes.

## B.3 Blending models trained on different blocks of predictors (Chapter 6)

In Chapter 6, a custom NP-integration strategy proposed to predict sheep feed efficiency (Figure 17). The thesis approach relied on a late integration strategy (Picard et al., 2021): the goal was to blend predictions of several models, instead of variables. The strategy was inspired by the block.sPLSR approach implemented in mixOmics (Singh et al., 2019).

In mixOmics, cross-validation only partitions data into training and testing sets. In the thesis, validation sets were added. Random subsampling was repeated 100 times. Subsampling was stratified per pen, year, and RFI line. First, all omics blocks were modelled

separately: variables were selected and submodels were fitted on training set data (60% of samples). All of these submodels were MINT-sPLSRs.

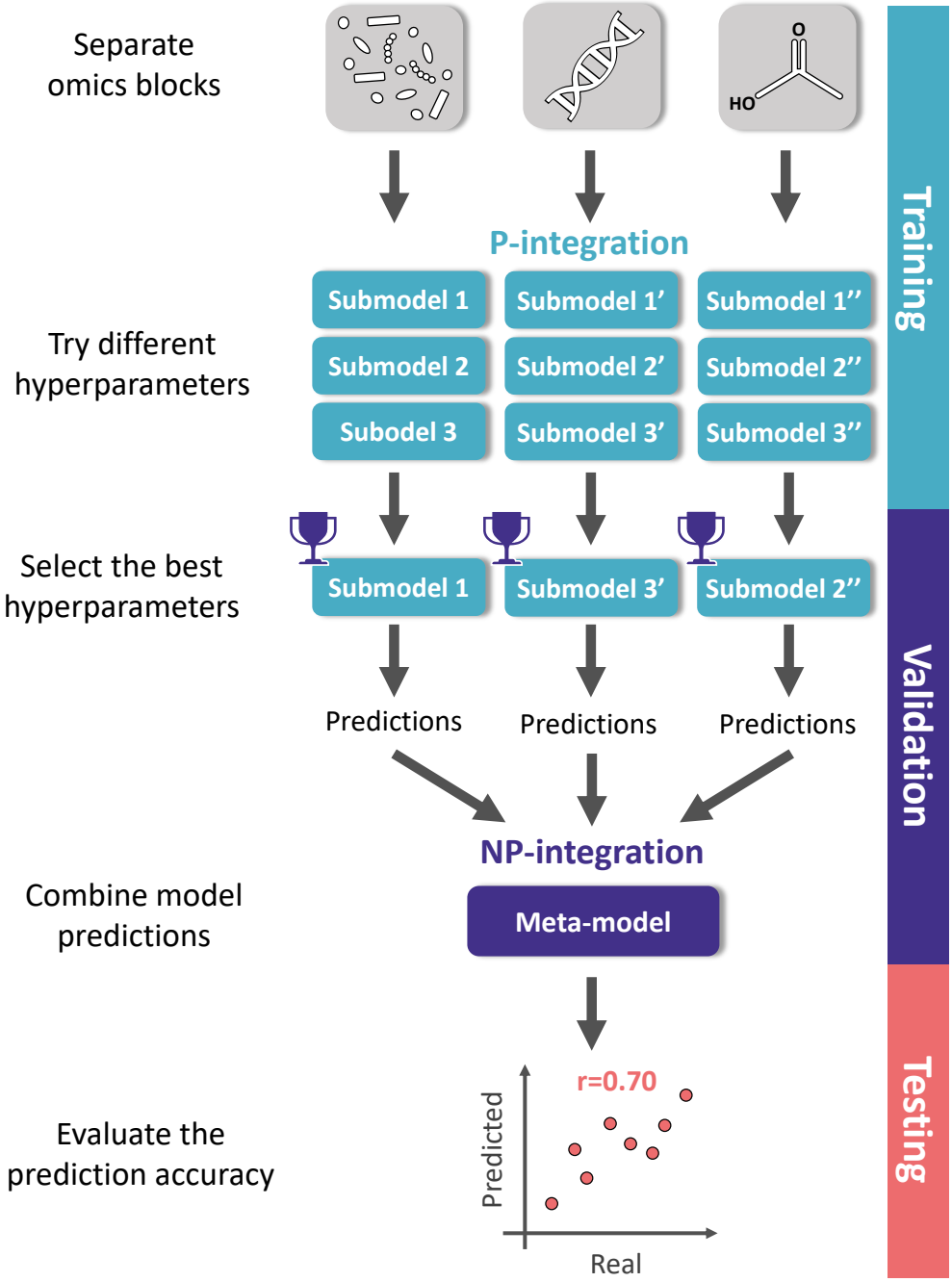



**Figure 17: Blending models built on different variables**

Then, MINT-sPLSRs' hyperparameters were selected to retain submodels maximizing the prediction accuracy in validation sets (30%). Validation sets were also used to fit a meta-model: the meta-model was trained on submodels' predictions. Thus, the meta-model learned which submodels predicted well data unseen during the training. The meta-model was either a weighted mean or a MINT-sPLSR (details below). Finally, the overall performance was assessed by computing the Pearson correlations in testing sets.

**Weighted mean**

Mean weights were computed from the Pearson correlation between validation phenotypes and predictions. Pearson correlations were then set to the exponent k. The exponent k ranged from 1 to 10 and its value was selected to maximize the validation accuracy. Higher exponents increased the gap between weights of blocks having a good predictive ability, and weights of blocks having a low predictive ability.

To assess how much a block contributes to the final prediction, the relative contribution was computed as the relative weight of the block (as a percentage of the total sum). One contribution was computed per block and subsampling repetition.

**MINT-sPLSR**

The number of MINT-sPLSR components could vary from 1 to 5, while the number of selected variables varied from 1 to the number of blocks. The two hyperparameters were selected with nested cross-validations. An inner loop was run in validation sets, to repeat 5-fold cross-validation 10 times per validation set.

To assess how much a block contributes to the final prediction, the value importance in the projection was computed (Tenenhaus, 1998). Then, values were expressed relatively to the highest one. One contribution was computed per block and subsampling repetition.

## B.4 Genetic evaluation of predicted feed intake (Chapter 4)

In Chapter 4, a genetic evaluation was carried out to estimate the breeding value of feed intake predicted from fixed effects, bodyweight and/or the rumen microbiota. “Breeding values were estimated with PEST (Groeneveld et al., 1990), considering a feed intake heritability of 0.28 (Tortereau et al., 2020). Two sets of populations were used to estimate breeding values: an entire Romane population named E (born from 2009 to 2020), with 6,419 animals in the pedigree including 1,900 with ADFI records; one subset population named S (2018 to 2020), with 4,102 animals in the pedigree including 277 with records. The model included the fixed effects of year, pen, early life traits (litter size, suckling method), sex and body weight as a covariate. EBVs of actual ADFI and EBVs of predictions are estimated. EBVs were only estimated for phenotypic ADFI predicted with accurate strategies ($R^2$> 0.7)” (Le Graverand et al., 2022).

**Table 7: Summary of analyzes carried out per thesis chapter**

| | | Chapter 3 | Chapter 4 | Chapter 5 | Chapter 6 |
|---|---|---|---|---|---|
| **Diet** | Concentrate diet | ✓ | ✓ | ✓ | ✓ |
| | Mixed diet | ✓ | | ✓ | |
| **Predicted groups or phenotypes** | RFI lines | ✓ | | ✓ | |
| | Feed efficiency | | | ✓ | ✓ |
| | Feed intake | | ✓ | ✓ | ✓ |
| | Growth | | | ✓ | |
| | Body composition | | | ✓ | |
| **Predictors** | Fixed effects and body weight | ✓ | ✓ | ✓ | ✓ |
| | Genotypes | ✓ | | | ✓ |
| | Plasma NMR buckets | ✓ | | | ✓ |
| | Plasma metabolite quantifications | ✓ | | | ✓ |
| | Rumen eukaryote abundances | ✓ | | ✓ | ✓ |
| | Rumen prokaryote abundances | ✓ | ✓ | ✓ | ✓ |
| | Rumen prokaryote functions | ✓ | | | ✓ |
| | Rumen buckets | ✓ | | | ✓ |
| | Rumen metabolite quantifications | ✓ | | | ✓ |
| | Rumen VFA concentrations | ✓ | | | ✓ |
| | Rumen LCFA concentrations | ✓ | | | ✓ |
| | Spectral absorbances | ✓ | | | ✓ |
| | Spectral first derivative | ✓ | | | ✓ |
| **Models** | sPLSDA | ✓ | | | |
| | MINT-sPLSDA | ✓ | | | |
| | sPLSR | | ✓ | ✓ | |
| | MINT-sPLSR | ✓ | | | ✓ |
| | SVR | | ✓ | | |
| | RFR | | ✓ | | |
| **Cross-validation (statistical test)** | Repeated k-fold (Bouckaert and Franck's t-test) | ✓ | | ✓ | |
| | Leave-one-year-out (Dunn and Clark's z test) | | ✓ | | |
| | Repeated random subsampling (Permutation test) | ✓ | | | ✓ |

Abbreviations: LCFA: Long-chain fatty acid; MINT-sPLSDA: Multivariate integrative sparse partial least squares discriminant analysis; MINT-sPLSR: Multivariate integrative sparse partial least squares regression; NMR: Nuclear magnetic resonance; RFI: Residual feed intake; RFR: Random forest regression; sPLSDA: Sparse partial least squares discriminant analysis; sPLSR: sparse partial least squares regression; SVR: Support vector regression; VFA: Volatile fatty acid.

# Chapter 3

# Study: Divergence between feed efficiency lines and omics overview

Chapter 3 is adapted from the "Results" section published in *animal* by Le Graverand et al. (2023). Reprinted parts are highlighted with quotation marks. Complements were added to cover all omics analyzed during the thesis.

# Rationale

Research is undergoing to identify the consequences of selection for feed efficiency in sheep. Omics may provide new insights into feed efficiency determinants and selection consequences.

Chapter 3 included exploratory analyses of lamb production traits. The first goal was to present the range of production traits, by computing descriptive statistics. The second goal was to assess whether divergent selection for RFI had an effect over production traits, by estimating least-square means per production trait and line.

Chapter 3 also included exploratory analyses of the potential predictors of feed efficiency. Fixed effects, covariates, pedigree and omics were tested during the thesis. Proxies may help predict feed efficiency by signing for biological determinants, the population structure, experimental or environmental effects. The rumen microbiota (prokaryote or eukaryote abundances) was assessed in animals fed a concentrate and then, a mixed diet. Pedigree relatedness, genomics, microbial functions, rumen metabolomics, plasma metabolomics, rumen lipidomics and faecal spectra were analyzed under a concentrate diet only. Variation patterns of potential predictors were explored, by fitting principal component analyses on pre-processed data. Then, discriminant analyses were carried out to predict RFI divergent lines from different predictors. sPLSDAs were fitted on fixed effects, body weight and microbiota data. MINT-sPLSDAs were carried out with pedigree data and all omics. The prediction accuracy of discriminant analyzes was evaluated by calculating the BER and cross-validating. sPLSDA models were evaluated by repeating 5-fold cross-validation 50 times. MINT-sPLSDA models were assessed by repeating random subsampling 100 times.

# Results

The first section details exploratory results of production phenotypes, including: feed efficiency (REI, FCR), energy intake (ADEI), body weight, growth (ADG) and body composition (MD, BFT). Traits are denoted with a C subscript under a concentrate diet; while a M subscript is used under a mixed diet.

The second section focuses on exploratory analyses of potential predictors of feed efficiency, with an emphasis on omics data.

## A. Production phenotypes in lambs of feed efficiency lines

"For feed efficiency traits and their components, descriptive statistics are provided in Table 8, along with least square means computed for the two RFI divergent lines.

Under a concentrate diet (17 to 23 weeks of age), lambs ingested an average of 10.98 MJ of net energy per day and grew by 330.91 g/day. It resulted in an average $FCR_C$ of 0.03 MJ/g. Under a mixed diet (29 to 35, or 37 to 43 weeks of age), sheep consumed an average of 8.06 MJ of net energy per day. No FCR was computed with a mixed diet because of some negative and null $ADG_M$ values. Regardless of the diet, RFI- animals had significantly lower intakes: with decrease of -0.95 MJ of net energy/day under a concentrate diet and -0.46 MJ/day under a mixed diet)

Considering the animals under study, the divergence between the two RFI lines was equal to 1.86 genetic standard deviations of $RFI_C$ least square means. Regardless of the diet, RFI- animals were more feed efficient, with a difference in residual energy intake between the RFI lines equal to 0.69 MJ of net energy/day under a concentrate diet against 0.33 MJ/day under a mixed diet. Under a concentrate diet, the difference in $FCR_c$ was also significant, with RFI+ animals ingesting 0.31 MJ of net energy/day more than RFI- animals to grow by 100 g/day.

Regardless of the diet, no significant difference was found between the two RFI lines for growth ($ADG_C$ or $ADG_M$) and most body composition traits ($BFT_C$, $BFT_M$, or $MD_M$). Only $MD_C$ significantly differed between the two lines: RFI- animals had a lower MD by 0.56 mm. Regarding the final body weight under a concentrate diet, the RFI- line was significantly lighter (-1.8 kg on average, at approximately 23 weeks of age). Later in life, under the mixed diet, the difference between the body weights of both lines was no longer significant (adjusted p-value = 0.196)" (Le Graverand et al., 2023).

**Table 8: Statistical summary of lamb traits and least square means for residual feed intake lines (reprinted from Le Graverand et al., 2023)**

| Diet[1] (n[2]) | Trait[3] | Population statistics[4] | | | | Line least square means[5] | | Adj. p[6] |
|---|---|---|---|---|---|---|---|---|
| | | Mean | SD | Min | Max | RFI- | RFI+ | |
| C-diet (269) | $REI_C$, MJ/day | 0.000 | 0.756 | -1.701 | 2.070 | -0.353 | 0.337 | <0.001 |
| | $FCR_C$, MJ/g | 0.034 | 0.006 | 0.021 | 0.055 | 0.032 | 0.035 | <0.001 |
| | $ADEI_C$, MJ/day | 10.98 | 1.38 | 7.37 | 14.14 | 10.50 | 11.45 | <0.001 |
| | Final $BW_C$, kg | 56.62 | 6.71 | 39.74 | 78.85 | 55.69 | 57.49 | <0.001 |
| | $ADG_C$, g/day | 331 | 58 | 150 | 501 | 333 | 331 | 0.927 |
| | $BFT_C$, mm | 5.7 | 0.8 | 3.9 | 8.3 | 5.8 | 5.8 | 0.927 |
| | $MD_C$, mm | 28.1 | 2.2 | 22.1 | 35.0 | 28.0 | 28.6 | 0.029 |
| M-diet (163) | $REI_M$, MJ/day | 0.000 | 0.899 | -2.740 | 2.222 | -0.203 | 0.125 | 0.013 |
| | $ADEI_M$, MJ/day | 8.06 | 1.22 | 4.31 | 10.50 | 7.77 | 8.23 | 0.008 |
| | Final $BW_M$, kg | 64.49 | 5.80 | 47.80 | 77.90 | 63.74 | 64.72 | 0.196 |
| | $ADG_M$, g/day | 123 | 67 | -58 | 264 | 121 | 128 | 0.386 |
| | $BFT_M$, mm | 4.5 | 0.8 | 2.9 | 6.9 | 4.6 | 4.4 | 0.196 |
| | $MD_M$, mm | 27.3 | 2.4 | 20.8 | 34.0 | 27.0 | 27.5 | 0.196 |

Abbreviations: ADEI: average daily energy intake; ADG: average daily gain; BFT: back fat thickness; C-diet: concentrate diet; FCR: feed conversion ratio; M-diet: mixed diet; MD: muscle depth; REI: residual energy intake; RFI-: efficient line with a low residual feed intake; RFI+: inefficient line with a high residual feed intake.

[1]: C-diet: sheep were fed a 100% concentrate diet; M-diet: sheep were fed a mixed diet including 2/3 of forage and 1/3 of concentrate

[2]: Number of animals

[3]: Subscripts denote the trait diet: C for the C-diet; M for the M-diet.

[4]: Descriptive statistics based on raw phenotypes, without adjusting for confounding effects

[5]: Least squares means were computed for lamb lines divergently selected for residual feed intake. Traits were adjusted for the age, suckling method (under the C-diet only), year, pen and phenotyping period (under the M-diet only).

[6]: Tukey's test adjusted p-values. The Benjamini-Hochberg procedure was applied per diet.

# B. Exploration of potential feed efficiency predictors

## B.1 Main variations patterns

### The genetic background of lamb lines

Lambs were part of two divergent lines, selected during 2-3 generations for RFI. Genetic lines were associated to the first PCs when PCAs were fitted on genomic and pedigree relatedness matrices (11% and 10% of explained variance, respectively) (Figure 18 A and B respectively). Projection on the first two PCs also highlighted the existence of familial clusters, *i.e.* siblings and half-siblings born the same year and sharing the same sire. On the opposite of the familial population structure, generations of selection were not noticeable on the projections defined by the first PCs. The second generation was studied in 2018 and 2019, while the third generation was phenotyped in 2020.

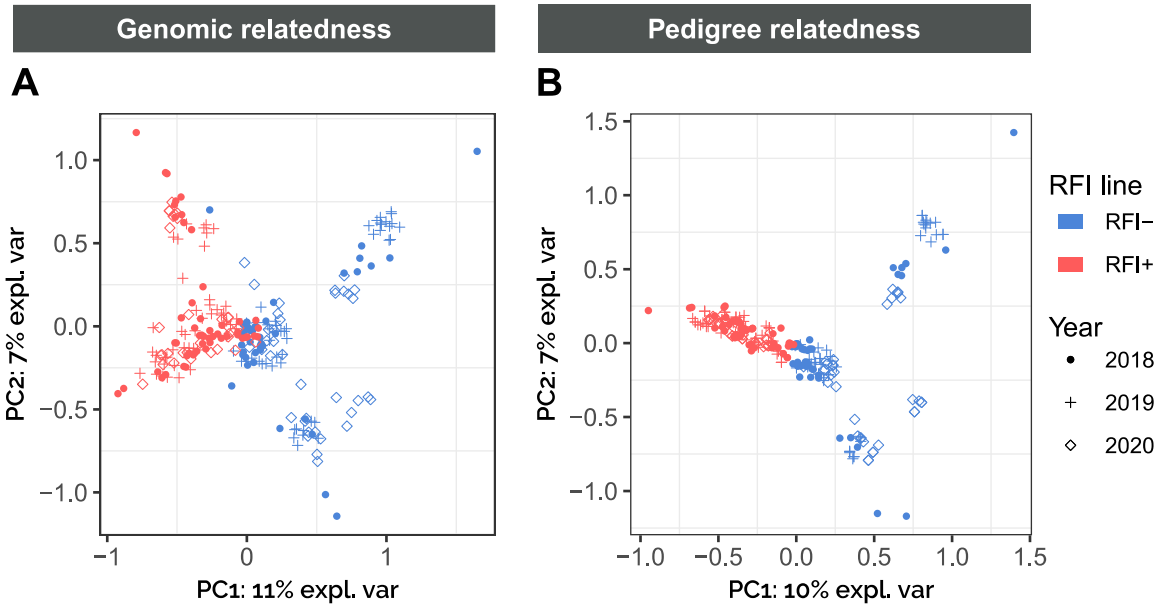



**Figure 18: Principal component analyses of lamb genomic and pedigree relatedness matrices**
Abbreviations: expl. var: explained variance; PC: principal component; RFI-: line selected for a decreased residual feed intake; RFI+: line selected for an increased residual feed intake. Analyses were carried out per relatedness matrix. (A) Genomic relatedness (VanRaden's matrix); (B) Pedigree relatedness (additive genetic relationship). The RFI- and RFI+ lines are denoted by blue and red symbols, respectively. Animals phenotyped in 2018, 2019 and 2020 are represented by dots, crosses and diamonds, respectively.

### The rumen microbiota compositions, in lambs fed a concentrate and a mixed diet later

"After data cleaning, 582 and 1148 OTUs of 16S sequencing remained under the concentrate and mixed diets, respectively, whereas for 18S sequencing, 124 and 183 OTUs were kept under the two diets, respectively )" (Le Graverand et al., 2023).

"OTUs were clustered according to their phylum affiliations, and mean phylum relative abundances are reported in Figure 19. With 16S sequencing, the most abundant phyla were *Bacteroidetes* under a concentrate diet (Figure 19 A) and *Firmicutes* under a mixed diet (Figure 19 C), while *Euryarchaeota* was the third most abundant phylum under both diets. With 18S sequencing and regardless of the diet (Figure 19 B, D), the *Ciliophora* phylum was largely predominant, with relative abundances ranging between 87 and 96%)" (Le Graverand et al., 2023).

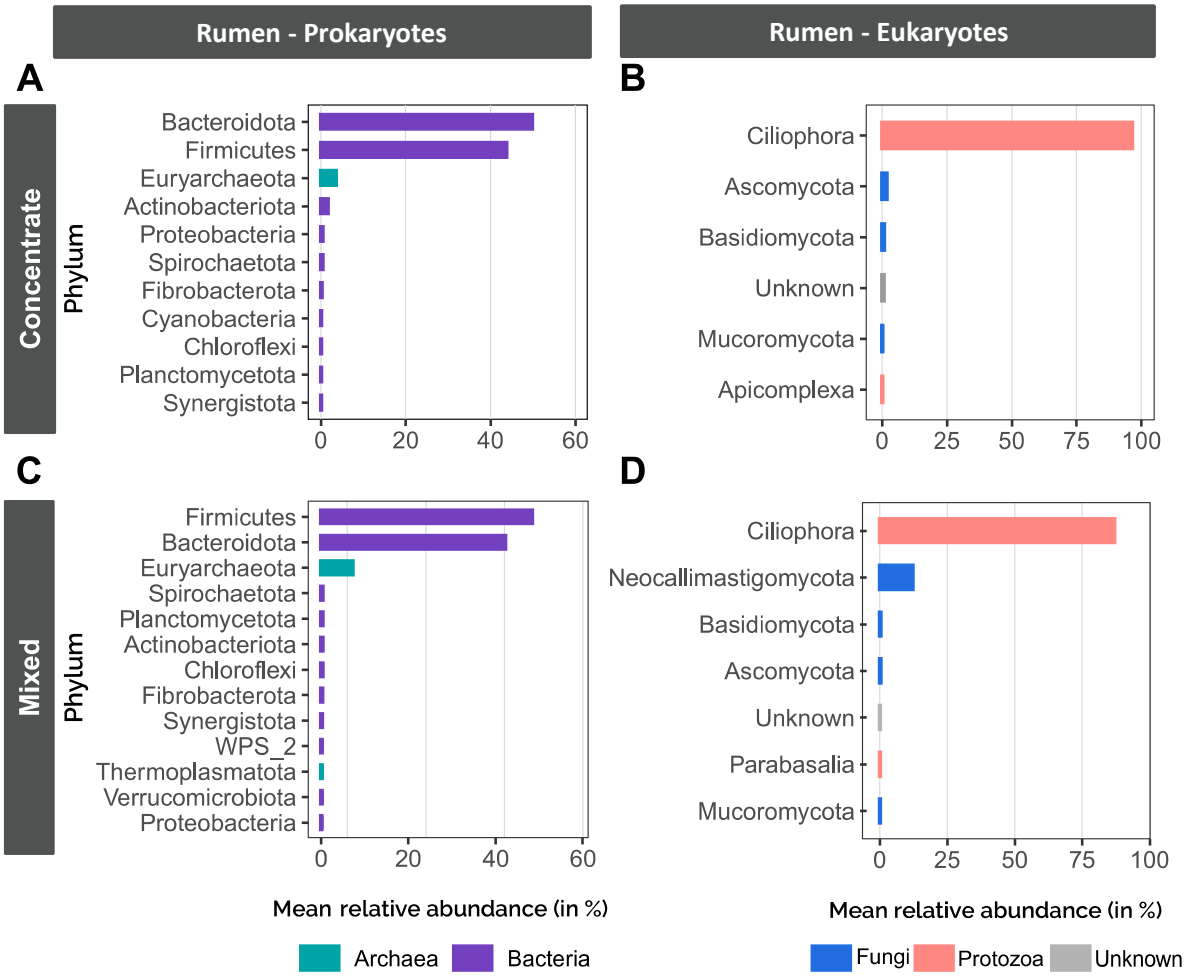



**Figure 19: Mean phylum abundances per sequencing and diet, in lamb ruminal fluids (reprinted from Le Graverand et al., 2023)**

"Relative phylum abundances were computed after data cleaning, per diet and sequencing. (A) 16S sequencing under a concentrate diet; (B) 18S under a concentrate diet; (C) 16S under a mixed diet; (D) 18S under a mixed diet. Taxonomic affiliations were based on the Silva 132 16S and Silva 138.1 18S databases. Taxa affiliated with archaea, bacteria, fungi and protozoa are represented in blue–green, purple, blue, and red, respectively" (Le Graverand et al., 2023).

"PCA allowed the identification of the main variation patterns of microbiota compositions (Figure 20). Under a concentrate diet, the variability of the second PC was associated with the year of lamb phenotyping with 16S data (5% of the variance explained, Figure 20 A).

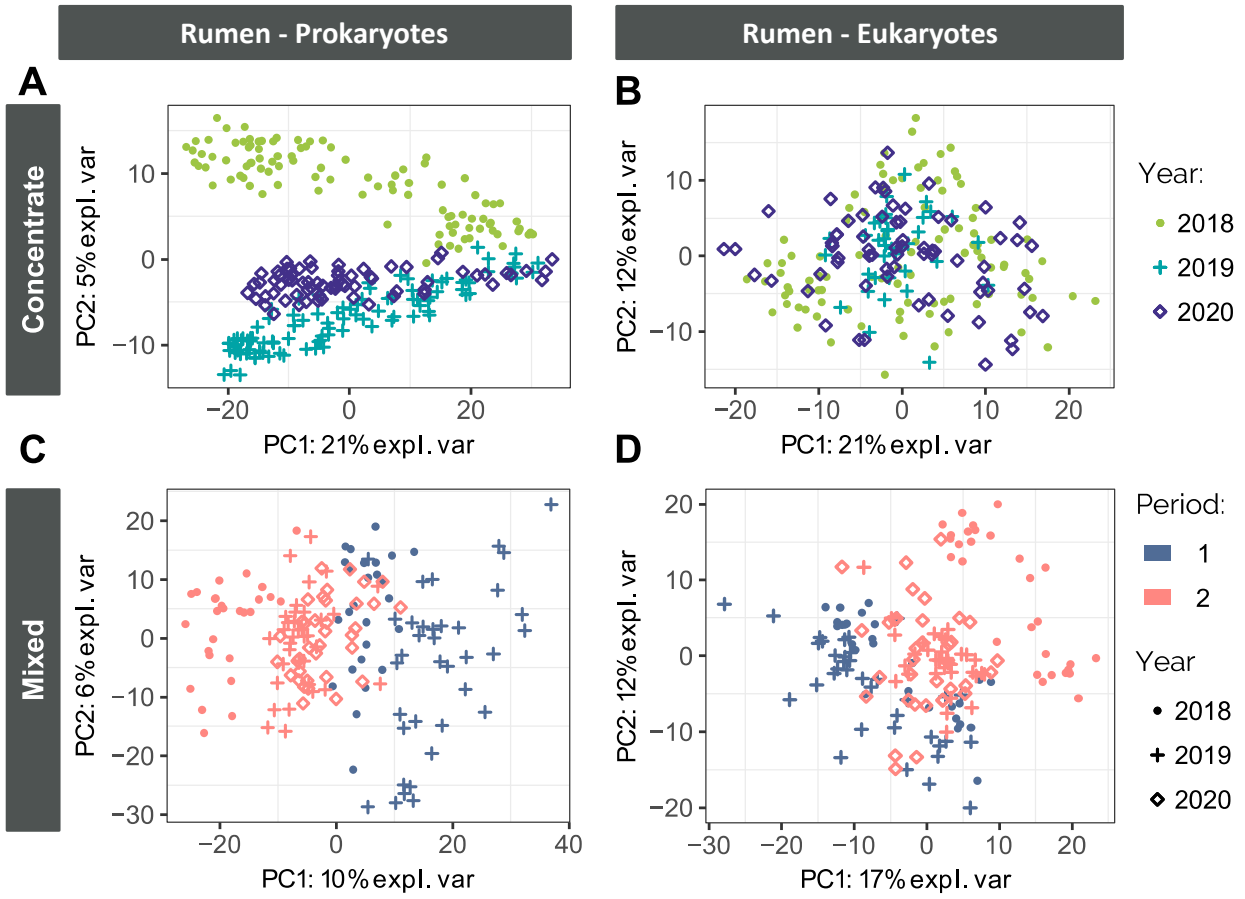



**Figure 20: Principal component analyses of lamb ruminal microbiota, per diet and sequencing (reprinted from Le Graverand et al., 2023)**

"Abbreviations: expl. var: explained variance; OTU: operational taxonomic unit; PC: principal component. Per diet and sequencing, analyses were carried out with OTU centered logratio values, adjusted for sequencing effects. (A) 16S sequencing under a concentrate diet; (B) 18S under a concentrate diet; (C) 16S under a mixed diet; (D) 18S under a mixed diet. Animals phenotyped in 2018, 2019 and 2020 are represented by dots, crosses and diamonds, respectively. Under a mixed diet, animals phenotyped during the first and second periods are represented in dark blue and red, respectively" (Le Graverand et al., 2023).

Under a mixed diet and regardless of the sequencing, the first PC variability seemed to be mostly tied to the phenotyping period and then the year to a lesser extent (Figure 20 C, D). None of the recorded variables (age at sampling, age at weaning, sampling order, pen, suckling method or RFI line) were associated with the other PCs. No main factor of variability could be identified for 18S data under a concentrate diet (Figure 20 B)" (Le Graverand et al., 2023).

**The functional profile of the prokaryotic microbiota, in lambs fed a concentrate diet**

Functions were inferred from 16S data under a concentrate diet, only. After filtering, 281 prokaryote functions were kept. Two groups could be distinguished by the first PC (32% of the explained variance) (Figure 21). Eleven functions were highly correlated with the first

PC coordinates (correlations >90%) and had relatively high loadings (>0.27). These functions were attributed to various pathways: Krebs's cycle, glyoxylate cycle, glycolysis, amino acid biosynthesis, and biodegradation of aromatic compounds. The relative abundance of one bacterial cluster was mainly associated to the eleven functions and the first PC: the OTU 647 affiliated to the *Oribacterium* genus. On average, the OTU 647 was the 261$^{th}$ most abundant OTU out of the 582 retained clusters. The year of study is moderately discriminated by the second PC with transformed abundances of prokaryote functions (25% of variance explained).

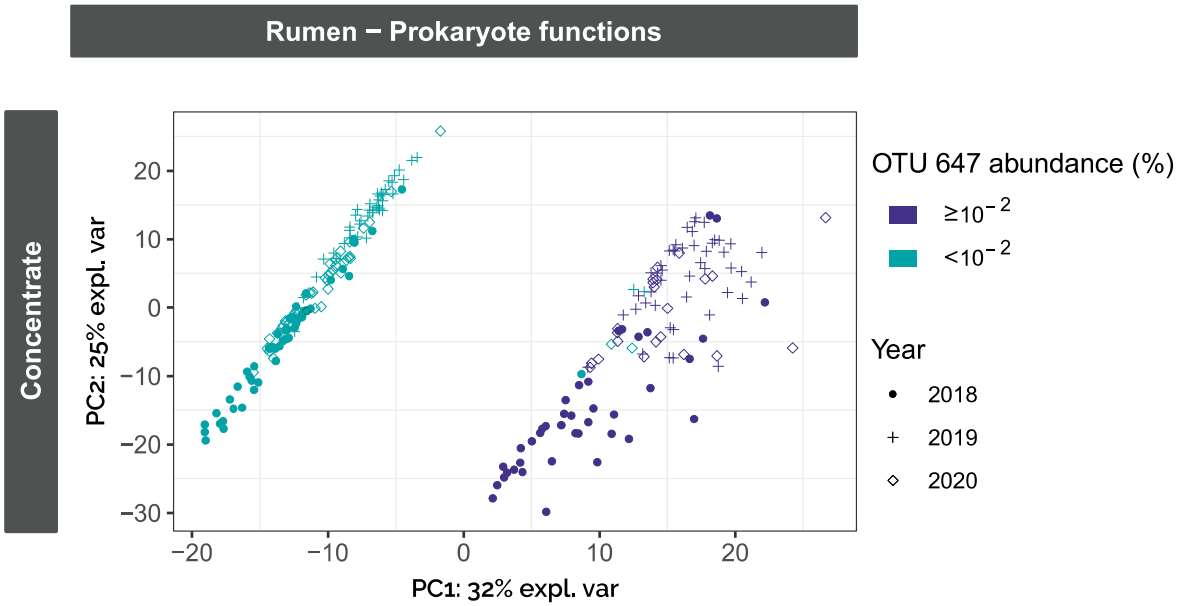



**Figure 21: Principal component analysis of ruminal prokaryote functions, under a concentrate diet**

Abbreviations: expl. var: explained variance; OTU: operational taxonomic unit; PC: principal component. Analyses were carried out under a concentrate diet only, with centered logratio values of function abundances adjusted for sequencing effects. Animals phenotyped in 2018, 2019 and 2020 are represented by dots, crosses and diamonds, respectively. Animals denoted by blue symbols had a relative abundance of OTU 647 superior to 0.01%, while purple symbols denoted abundances inferior to 0.01%.

### Rumen and plasma metabolomics, in lambs fed a concentrate diet

Under a concentrate diet, 871 ruminal buckets and 25 ruminal metabolites concentrations were kept after data cleaning. After logratio transformation and adjusting for fasting and spectroscopy batch effects, PCA projections were compared to recorded factors (Figure 22 A and B). Variations explained by the first PCs could not be associated with the spectroscopy batch, sampling order, fasting effect, pen, suckling method, RFI line, sampling or weaning age.

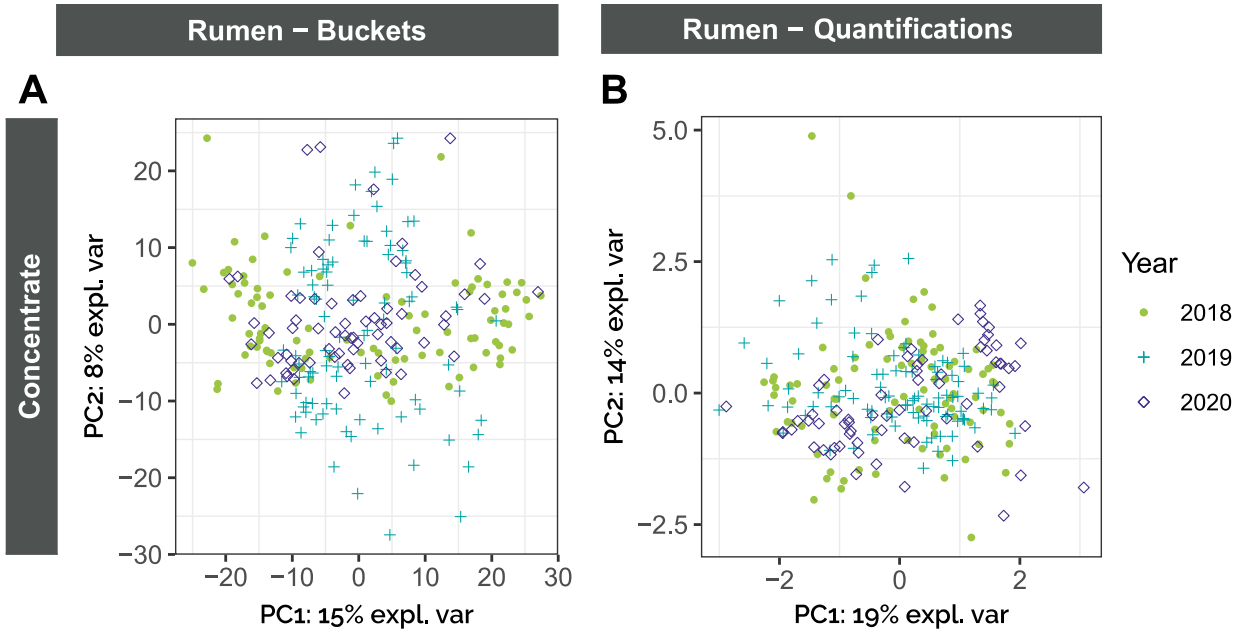



**Figure 22: Principal component analysis of rumen metabolomics, under a concentrate diet**
Abbreviations: expl. var: explained variance; PC: principal component. Analyses were carried out under a concentrate diet only, with centered logratio values of bucket areas or metabolite quantifications adjusted for fasting and spectroscopy batch effects. (A) NMR buckets; (B) Metabolite quantifications. Animals phenotyped in 2018, 2019 and 2020 are represented by green dots, blue crosses and purple diamonds, respectively.

Under the same diet, 863 plasmatic buckets and 25 plasmatic metabolites were retained after filtering. With plasmatic buckets, no known effect could explain the variations represented by the first PCs (Figure 23 A). On the opposite, two groups could be distinguished based on inferred concentrations of plasma metabolites, according to the first PC (23% of explained variance) (Figure 23 B). The two groups were associated with L-threonine concentrations (loading of 0.96 on the first component). On average, L-threonine had the 16$^{th}$ highest concentration out of 25 retained metabolites. Individual having null to low concentrations of L-threonine concentrations were part of the group having negative coordinates on the first PC. The L-threonine concentration of these individuals may be below or close to the detection limit. Intra L-threonine groups, animals phenotyped in 2018 were associated with higher coordinate values on the first PC, compared to animals of 2020. Animals phenotyped in 2019 had intermediate coordinates on the first PC.

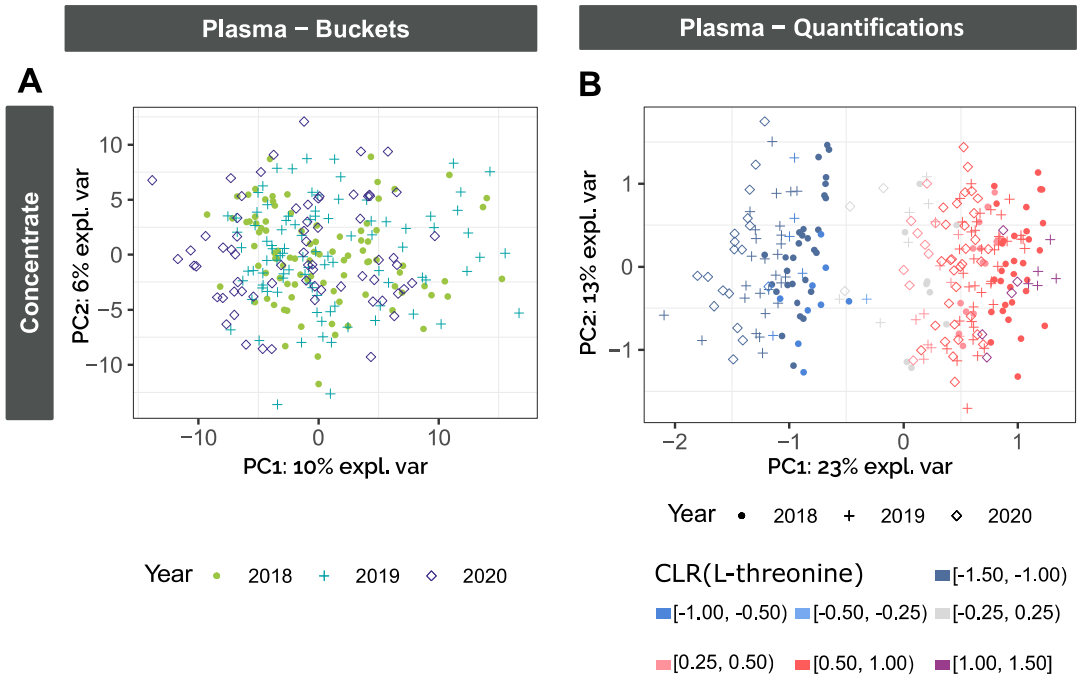



**Figure 23: Principal component analysis of plasma metabolomics, under a concentrate diet**
Abbreviations: expl. var: explained variance; PC: principal component. Analyses were carried out under a concentrate diet only, with centered logratio values of bucket areas or metabolite quantifications adjusted for fasting and spectroscopy batch effects. (A) NMR buckets (fixed width of 0.01 ppm); (B) Metabolite quantifications (inferred from NMR spectra by ASICS). Animals phenotyped in 2018, 2019 and 2020 are represented by dots, crosses and diamonds, respectively. In subplot B, blue shades denote the lowest concentrations of L-threonine, red and purple shades denote the highest concentrations.

### Rumen lipidomics, in lambs fed a concentrate diet

Under a concentrate diet, 6 ruminal VFAs and 70 LCFAs were retained for analysis. After adjusting for analysis batch and fasting effects, no recorded variable could be associated thanks to PCAs fitted on VFA or LCFAs concentrations (Figure 24 A and B).

### Faecal phenomics, in lambs fed a concentrate diet

Under a concentrate diet, 1050 NIRS absorbances were recorded and their first derivative was computed. The year of study was hardly distinguishable when PCA was applied to absorbances: animals phenotyped in 2020 tended to have slightly higher values on the first PC and lower values on the second PC (83% and 11% of the explained variance, respectively) (Figure 25 A). No other recorded variable could be tied to absorbance variations represented by the first PCs. On the opposite, the year of study was easily visualized when PCA was carried with the first derivative: the three years of study were highlighted by the first two PCs (33% and 20% of the explained variance, respectively) (Figure 25 B).

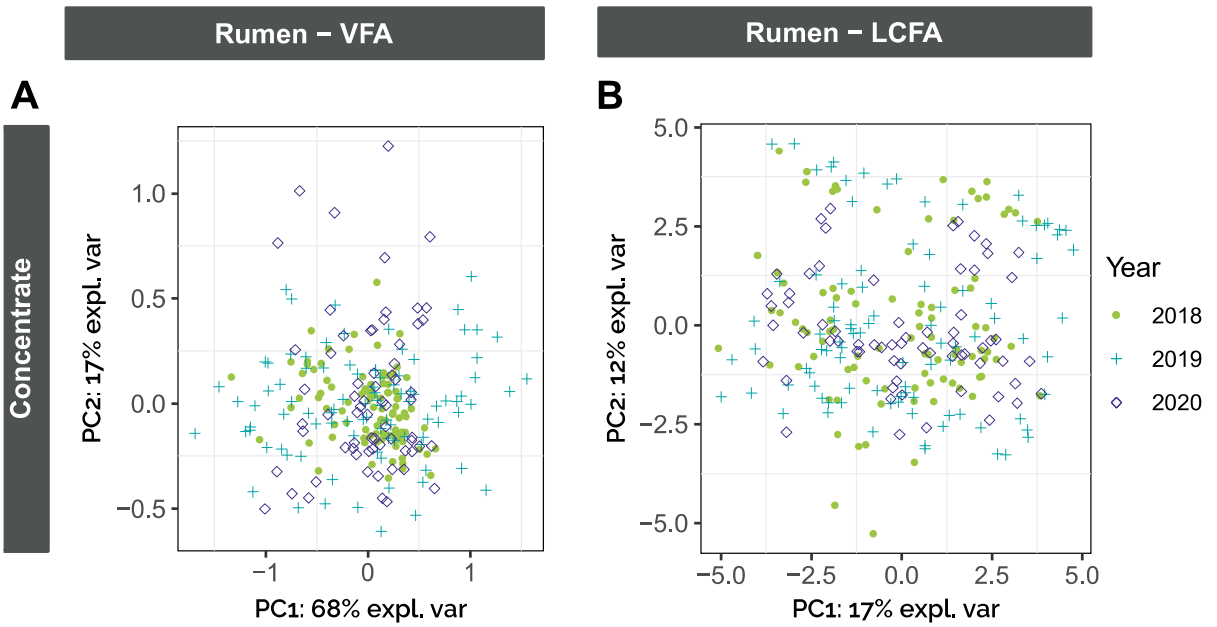



**Figure 24: Principal component analysis of rumen lipidomics, under a concentrate diet**

Abbreviations: CLR: centered logratio; expl. var: explained variance; PC: principal component. Analyses were carried out under a concentrate diet only, with CLR values of fatty acid concentrations adjusted for fasting and analysis batch effects. (A) Volatile fatty acids; (B) Long-chain fatty acids. Animals phenotyped in 2018, 2019 and 2020 are represented by green dots, blue crosses and purple diamonds, respectively.

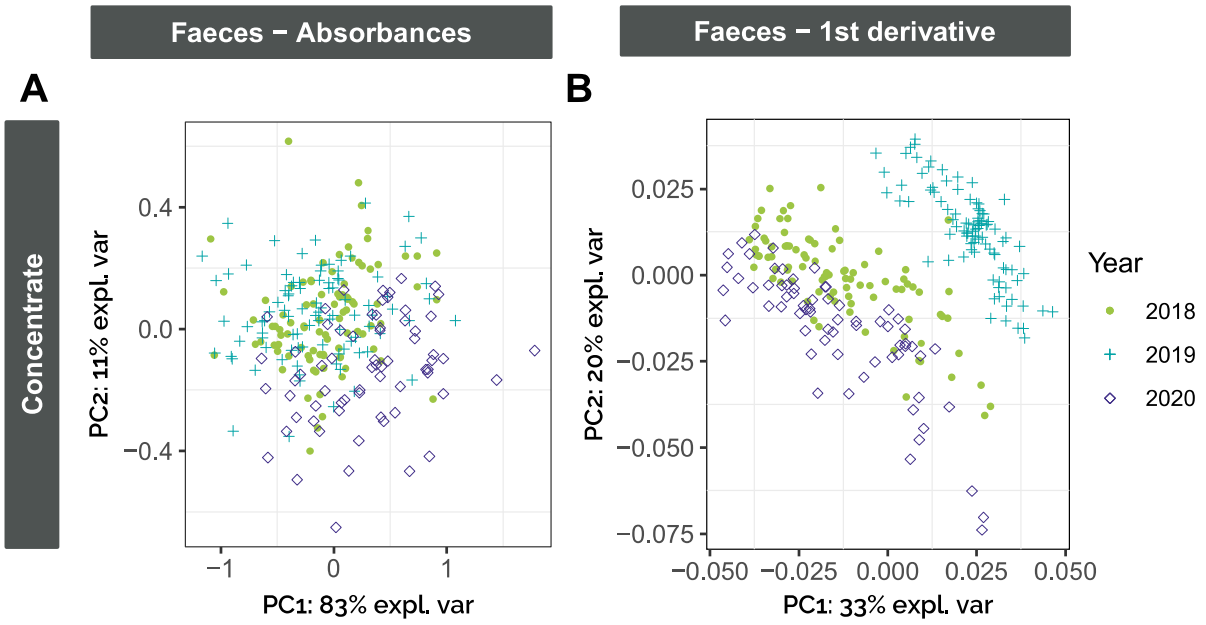



**Figure 25: Principal component analyses of near-infrared spectroscopy of lamb faeces, under a concentrate diet**

Abbreviations: expl. var: explained variance; PC: principal component. Analyses were carried out under a concentrate diet only. (A) Near-infrared spectroscopy absorbances; (B) First derivative of near-infrared spectroscopy absorbances. Animals phenotyped in 2018, 2019 and 2020 are represented by green dots, blue crosses and purple diamonds, respectively.

## B.2 Predicting the feed efficiency lines of lambs from potential predictors

Feed efficiency lines were predicted from different sets of predictors. Depending on omics, the number of phenotyped lambs varied. Thus, lines were also predicted from routine systematic effects – *i.e.* year, pen, suckling, age and final body weight effects – to set a comparable baseline in all population subsets.

### Predicting lamb lines from the microbiota composition and systematic effects, under two diets

sPLSDAs were evaluated by repeating 5-fold cross-validation 50 times. "Regardless of the predictors, diet and sequencing, the BERs of sPLSDA were always high (from 0.35 to 0.55) when predicting RFI lines (Table 9). [Predictions of the RFI line from fixed effects and body weight were more accurate or as accurate as predictions from microbiota data only]. The addition of the microbiota data to the fixed effects and body weight data never decreased BERs significantly" (Le Graverand et al., 2023).

**Table 9: Balanced error rates when lamb feed efficiency lines were predicted from systematic effects and/or microbiota data (reprinted from Le Graverand et al., 2023)**

| | | Set of predictors[3] | | |
|---|---|---|---|---|
| **Diet[1]** | **Sequencing (n[2])** | **M** | **S** | **S+M** |
| C-diet | 16S (269) | 0.46[a] (0.06) | 0.38[a] (0.05) | 0.46[a] (0.06) |
| | 18S (205) | 0.47[b] (0.07) | 0.35[a] (0.06) | 0.43[ab] (0.06) |
| M-diet | 16S (160) | 0.55[a] (0.08) | 0.47[a] (0.07) | 0.53[a] (0.08) |
| | 18S (161) | 0.45[a] (0.08) | 0.49[a] (0.07) | 0.45[a] (0.08) |

"Abbreviations: C-diet: concentrate diet; M-diet: mixed diet; M: microbiota; S: systematic effects; S+M: microbiota plus systematic effects.
Predictions were carried out with sparse partial least squares discriminant analyses. Balanced error rates were averaged over 5-fold cross-validations repeated 50 times. Standard deviations are given in brackets.
1: Diets. C-diet: sheep were fed a 100% concentrate diet; M-diet: sheep were fed a mixed diet (2/3 of forage and 1/3 of concentrate)
2: Number of samples
3: Sets of predictors. M: adjusted centered logratio values of 16S or 18S operational taxonomic units; S: systematic effects, i.e., fixed effects and final body weight; S+M: systematic effects and adjusted centered logratio values of 16S or 18S operational taxonomic units
[a,b]: Balanced error rates with different letters significantly differ (corrected t-test p-value < 0.05 after adjustment). Comparisons and Benjamini-Hochberg adjustments were made per diet and sequencing" (Le Graverand et al., 2023).

**Predicting lamb lines from systematic effects, pedigree and omics, under a concentrate diet**

Other potential predictors of feed efficiency were assessed under a concentrate diet only (Table 10). MINT-sPLSDA models were assessed by repeating random subsampling 100 times.

The prediction accuracy of RFI lines was very low when predictions were made from the functions of rumen prokaryotes (BER = 0.45), same as predictions from rumen metabolomics and lipidomics (0.46 ≤ BERs ≤ 0.51). Predicting from faecal phenomics data (0.43 ≤ BERs ≤ 0.44), plasma metabolomics (0.42 ≤ BERs ≤ 0.43) and systematic effects (BER=0.37) was hardly more accurate.

As expected in a divergent population, RFI lines were accurately predicted from genotypes and pedigree relatedness: averaged BERs were equal to 0.00.

**Table 10: Balanced error rates when lamb feed efficiency lines were predicted from systematic effects, pedigree or omics data under a concentrate diet**

| | Predictors | | | Prediction accuracy |
|---|---|---|---|---|
| Diet[1] (n[2]) | Source | Variables | p[3] | Average BER (SD) |
| C-diet (255) | Farm records | Fixed effects + body weight | 20 | 0.37 (0.09) |
| | | Pedigree relatedness matrix | 255 | 0.00 (0.01) |
| | Blood | Genotypes | 29 830 | 0.00 (0.01) |
| | | Plasma buckets | 863 | 0.42 (0.09) |
| | | Plasma metabolite quantifications | 25 | 0.43 (0.10) |
| | Rumen | Prokaryote abundances | 582 | 0.46 (0.08) |
| | | Prokaryote functions | 281 | 0.45 (0.10) |
| | | Rumen buckets | 871 | 0.50 (0.09) |
| | | Rumen metabolite quantifications | 25 | 0.48 (0.09) |
| | | Rumen VFA concentrations | 6 | 0.46 (0.08) |
| | | Rumen LCFA concentrations | 70 | 0.51 (0.10) |
| | Faeces | Spectral absorbances | 1050 | 0.43 (0.08) |
| | | Spectral first derivative | 1050 | 0.44 (0.09) |

Abbreviations: C-diet: concentrate diet; BER: balanced error rate; SD: standard deviation; LCFA: long-chain fatty acids; VFA: volatile fatty acids.
Predictions were carried out with multivariate integrative sparse partial least squares discriminant analysis. BERs were averaged over testing sets, defined by repeating random subsampling 100 times (training=60% of data, validation=30%, testing=10%). SDs are given in brackets.
[1]: C-diet: sheep were fed a 100% concentrate diet.
[2]: Number of samples
[3]: Number of predictor variables

**Take-home messages about Chapter 3:**

- Feed efficiency traits (REI, FCR) and energy intake (ADEI) significantly differed between RFI divergent lines, when lambs were fed a concentrate or a mixed diet.
- Final $BW_C$ and $MD_C$ were significantly lower in the feed efficient line only when lambs were fed with concentrates, around 23 weeks of age.
- PCAs clearly highlighted that pedigree and genomics data reflect the population structure, in terms of lines and families.
- PCAs underlined that abundances of microorganisms, plasma metabolite quantifications and faecal phenomics can sign temporal effects – i.e. the year or period of phenotyping.
- Inferring microbial functions from 16S metabarcoding, or inferring plasmatic metabolite quantifications from NMR spectra introduced new data structures: animals could be clustered into groups, based on one microorganism or metabolite by PCAs. It may represent inference artefacts.
- Feed efficiency lines could not be predicted accurately from either the rumen microbiota, rumen metabolome, plasma metabolome, faecal phenome or systematic effects.
- Genomics and pedigree can accurately predict the feed efficiency lines.

# Chapter 4

# Study: Predicting feed intake of independent sheep cohorts from their rumen microbiota

Chapter 4 is adapted from the "Results" section published in the proceeding of the 2022 WCGALP congress, by Le Graverand et al. (2022). Reprinted parts are highlighted with quotation marks. Complements were added to introduce the study goal and summarize results.

# Rationale

Feed intake is rarely recorded by sheep breeding companies, on the opposite of fixed effects and body weights which are routinely documented. However, feed intake records are essential to estimate feed efficiency. Thus, predicting intake could help select for feed efficiency.

Chapter 4 focused on the prediction of average daily feed intake under a concentrate diet ($ADFI_C$) from systematic effects and/or microbiota abundance data. Systematic effects included the year, pen, suckling, age and final body weight effects. The matrices were concatenated to integrate microbiota data and systematic effects. The first goal was to train predictive models on a reference population and then, predict $ADFI_C$ in lambs raised a different year. Several machine learning algorithms were used to predict $ADFI_C$ phenotypes: sparse partial least square regression (sPLSR), support vector regression (SVR) and random forest regression (RFR). Prediction accuracy was assessed by carrying out a leave-one-year out cross-validation. Two training sets and testing sets were used: models were trained on 2018 and 2020 to predict 2019 phenotypes; or models were trained on 2018 and 2019 to predict 2020. Since rams phenotyped in 2018 sired lambs in 2020, 2018 was never used as a testing set. The accuracy measure was the Pearson correlation between testing set predictions and real phenotypes.

The second goal was to check whether EBVs of predicted $ADFI_C$ were close to EBVs of observed $ADFI_C$. EBVs of real and predicted $ADFI_C$ were computed with mixed models, before computing the Pearson correlation between EBVs. “Two sets of populations were used to estimate breeding values: an entire Romane population named E (born from 2009 to 2020), with 6 419 animals in the pedigree including 1 900 with ADFI records; one subset population named S (2018 to 2020), with 4,102 animals in the pedigree including 277 with records” (Le Graverand et al., 2022).

# Results

## A. Comparison of different predictors and machine learning approaches for phenotypic ADFIc

"Table 11 details accuracies of ADFIc predictions for the three machine learning models carried out with 16S data, animal traits and a concatenation of both. Considering correlations between actual and predicted ADFIc phenotypes, there was no significant difference between sPLSR, SVR and RFR accuracies whatever the testing set or the predictors.

With all testing sets and machine learning techniques, correlations were significantly lower [and variable] when only 16S data was used as the feature set. Whatever the testing set [(ranging from -0.11 to +0.35)] and the machine learning model, combining animal traits and 16S data together as predictors did not significantly increased correlations [compared to animal traits alone]" (Le Graverand et al., 2022).

**Table 11: Pearson correlations between predicted and actual ADFIc phenotypes of testing set animals (reprinted from Le Graverand et al., 2022)**

| | | Testing sets | |
|---|---|---|---|
| **Models** | **Features** | **2019** | **2020** |
| sPLSR | 16S | -0.12[a] | 0.19[a] |
| | Animal | 0.76[b] | 0.81[b] |
| | Animal+16S | 0.73[b] | 0.82[b] |
| SVR | 16S | -0.04[a] | 0.35[a] |
| | Animal | 0.77[b] | 0.80[b] |
| | Animal+16S | 0.76[b] | 0.82[b] |
| RFR | 16S | 0.03[a] | 0.31[a] |
| | Animal | 0.78[b] | 0.77[b] |
| | Animal+16S | 0.71[b] | 0.74[b] |

[a,b]: Within one testing set, correlations with no common letter significantly differ (adjusted P<0.05, Dunn and Clark's z test).

## B. Relationship between EBVs of predicted ADFIc and EBVs of actual ADFIc

"The quality of predicted ADFIc EBVs is presented in Table 12. Within one machine learning approach and regardless of the testing set, correlations between EBVs are significantly

higher when a full set of records (E) is used during the genetic evaluation compared to a partial set (S).

In 2020, no significant difference could be found between machine learning approaches. However, when only animal traits were used as 2019 $ADFI_C$ predictors, sPLSR performed significantly better than SVR, with RFR having intermediate performances. Finally, the combination of animal traits and 16S data as predictors did not significantly improve correlations between actual and predicted $ADFI_C$ EBVs" (Le Graverand et al., 2022).

**Table 12: Pearson correlations between estimated breeding values of predicted and actual $ADFI_C$, in testing sets (reprinted from Le Graverand et al., 2022)**

| | | Testing sets | | | |
|---|---|---|---|---|---|
| | | 2019 | | 2020 | |
| **Models** | **Features** | **S** | **E** | **S** | **E** |
| sPLSR | Animal | $0.681^{a}$ | $0.868^{c}$ | $0.843^{a}$ | $0.954^{b}$ |
| | Animal+16S | $0.698^{ab}$ | $0.876^{cd}$ | $0.852^{a}$ | $0.956^{b}$ |
| SVR | Animal | $0.624^{b}$ | $0.814^{d}$ | $0.848^{a}$ | $0.962^{b}$ |
| | Animal+16S | $0.650^{ab}$ | $0.817^{cd}$ | $0.839^{a}$ | $0.956^{b}$ |
| RFR | Animal | $0.731^{ab}$ | $0.891^{cd}$ | $0.863^{a}$ | $0.952^{b}$ |
| | Animal+16S | $0.698^{ab}$ | $0.880^{cd}$ | $0.848^{a}$ | $0.959^{b}$ |

[1] Populations differ in the number of records used for EBVs estimation. S: subset population; E: entire population.

$^{a,b,c,d}$: Within one testing set (2019 or 2020), correlations with no common letter significantly differ (adjusted $P<0.05$, Dunn and Clark's z test).

**Take-home messages about Chapter 4:**

- The prediction accuracy of feed intake phenotypes did not vary significantly between the tested machine learning methods (sPLSR, SVR, RFR).
- When $ADFI_C$ phenotypes were predicted from microbiota data only, the prediction accuracy was unstable and varied between years ($-0.12 \leq$ correlations $\leq 0.35$).
- Predictions of $ADFI_C$ phenotypes from fixed effects and body weight were quite accurate ($0.76 \leq$ correlations $\leq 0.81$), significantly more than predictions from the rumen prokaryotic microbiota.
- Concatenating microbiota data with fixed effects and body weight did not improve predictions of $ADFI_C$ phenotypes. It also did not improve the correlation between EBVs of real intake and EBVs of predicted intake.
- The correlation increased between EBVs of predicted $ADFI_C$ and EBVs of real $ADFI_C$, when the number of pedigree individuals (6 419 vs 4 102) and the number of records (1 900 vs 277) both increased. Concomitantly, the correlation increase was also due to the smaller proportion of predicted phenotypes in records (maximum 5% against 36%).

# Chapter 5

# Study: Predicting production traits of contemporaneous sheep from their rumen microbiota

Chapter 5 is adapted from the "Results" section published by Le Graverand et al. (2023). Reprinted parts are highlighted with quotation marks. Complements were added to introduce the study goal and summarize results.

# Rationale

Predictors of feed efficiency are actively researched, including the rumen microbiota. Due to the major environment's influence over the ruminal microbiota and sheep traits, we may have to profile the microbiota of several sheep each year.

Chapter 5 focused on the prediction of production traits from microbiota data. This study differed from the previous one by its ambition: more diets were considered (concentrate and mixed diet), more traits were predicted (feed efficiency, intake, growth, and body composition) and prokaryote and eukaryote abundance proxies were tested (16S and 18S metabarcoding). Chapter 5 also differed by its cross-validation strategy: every year, phenotypes of several sheep are known and used to train predictive models.

The goal was to assess the accuracy of sPLSRs predicting lamb traits from systematic effects and/or microbiota data. Systematic effects included year, pen, suckling, age and body weight effects. Microbiota data included transformed microorganisms' abundances, with 16S and 18S metabarcoding. Prediction accuracy was estimated by calculating the Pearson correlation and repeating 5-fold cross-validation 50 times.

# Results

## A. Predictions from microbiota data only

"Under a concentrate diet, when host feed efficiency was predicted from 16S (Figure 26 A) or 18S (Figure 26 B) data, the average correlations between actual traits and predictions were almost null for $REI_C$ (0.11 for 16S compared to 0.06 for 18S data) or low to moderate for $FCR_C$ (0.35 for 16S, 0.16 for 18S). Under a mixed diet, predicting $REI_M$ led to moderate average correlations (0.35 with 16S, 0.17 with 18S) (Figure 26 C, D)" (Le Graverand et al., 2023).

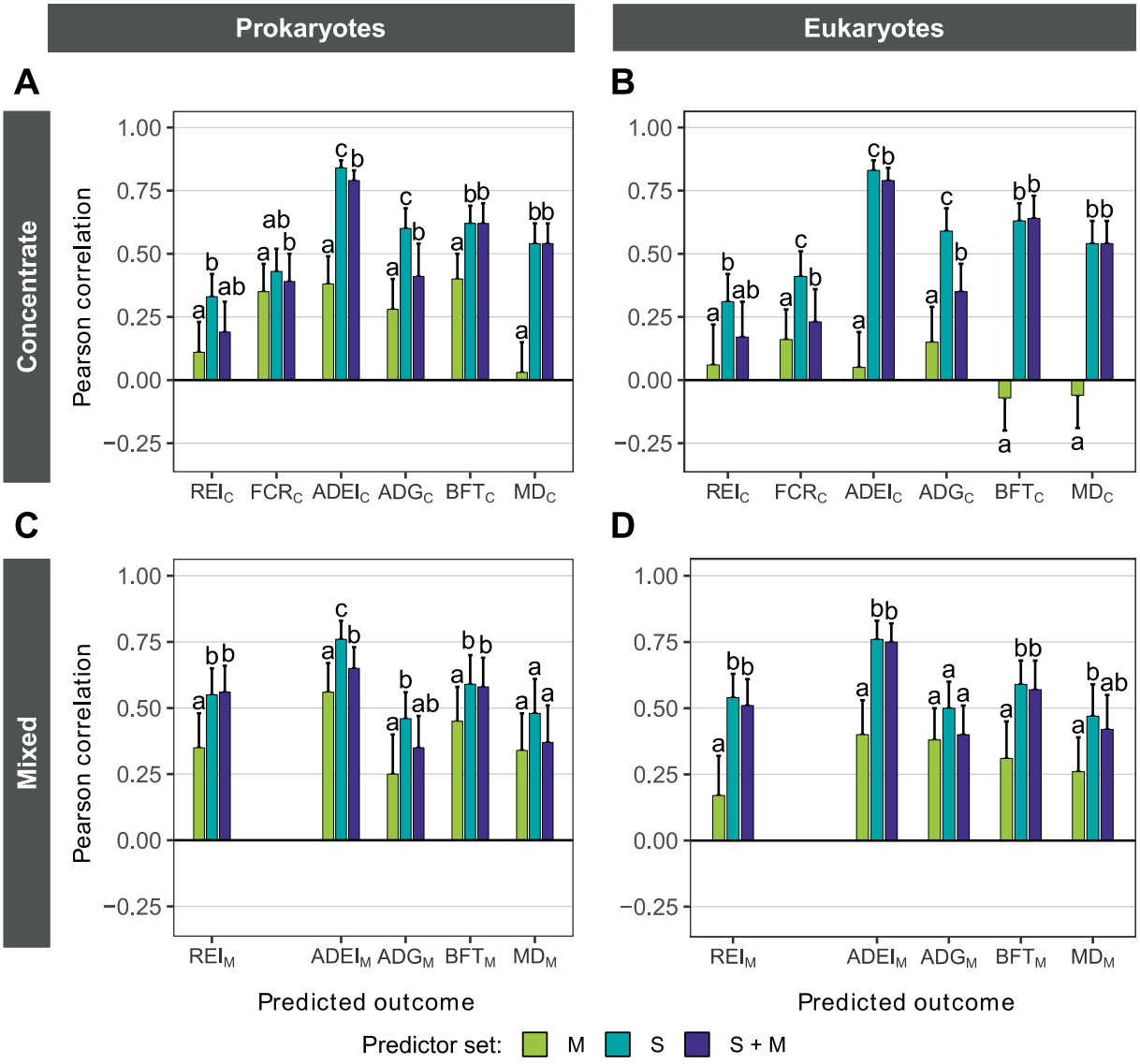



**Figure 26: Average Pearson correlations between lamb feed efficiency, production traits and predictions from systematic effects and/or microbiota data (reprinted from Le Graverand et al., 2023)**

"Abbreviations: ADEI: average daily energy intake; ADG: average daily gain; BFT: back fat thickness; FCR: feed conversion ratio; M: microbiota; MD: muscle depth; REI: residual energy intake; S: systematic effects; S+M: microbiota plus systematic effects. Predictions were carried out with sparse partial least squares regressions. Pearson correlations were averaged over 5-fold cross-validations repeated 50 times. Error bars are equivalent to 1 standard deviation. (A) 16S sequencing under a concentrate diet; (B) 18S under a concentrate diet; (C) 16S under a mixed diet; (D) 18S under a mixed diet. Three sets of predictors were tested, with M: adjusted CLR values of 16S or 18S operational taxonomic units (green); S: systematic effects including fixed effects and final body weight (blue–green); S+M: systematic effects and adjusted CLR of 16S or 18S operational taxonomic units (purple). Traits were recorded under a concentrate diet (C subscript) or a mixed diet (M subscript). $^{a,b,c}$: Correlations with different letters significantly differ (corrected t-test p-value < 0.05 after adjustment). Comparisons and Benjamini–Hochberg adjustments were made per trait, diet and sequencing" (Le Graverand et al., 2023).

Predicting energy intake, growth and body composition "also led to varying average Pearson correlations, depending on the diet and the sequencing dataset. Predicting intakes ($ADEI_C$, $ADEI_M$) with 16S or 18S data led to a large range of correlations: from 0.05 with 18S under a concentrate diet to 0.56 with 16S under a mixed diet. Similar results were obtained for growth traits ($ADG_C$ or $ADG_M$), with correlations ranging from 0.15 with 18S data under a concentrate diet to 0.38 with 18S data under a mixed diet. Finally, average correlations between body composition traits ($BFT_C$, $BFT_M$, $MD_C$ and $MD_M$) and microbiota predictions fluctuated between almost null and moderate values: from -0.07 for $BFT_C$ predicted from 18S data to 0.45 for $BFT_M$ predicted from 16S data" (Le Graverand et al., 2023).

## B. Predictions from systematic effects only

"Regardless of [the diet and trait], average correlations for predictions from systematic effects ranged from 0.31 to 0.84. Furthermore, for almost all recorded traits, correlations were significantly higher than predictions derived from 16S or 18S data only. There were only three exceptions: similar accuracies were reached when FCRC and MDM were predicted from 16S or systematic effects, same as ADGM predicted from 18S or systematic effects" (Le Graverand et al., 2023).

## C. Predictions from microbiota and systematic effects

"Finally, combining microbiota data with fixed effects and final body weight never significantly improved correlations compared to predictions from fixed effects and body weight. Most of the correlations were not significantly different. However, some correlations significantly decreased when 16S data and systematic effects were combined to predict: $ADEI_C$ (by -0.05 units on average), $ADEI_M$ (-0.11) and $ADG_C$ (-0.19). Similar decreases were observed with 18S data when predicting $FCR_C$ (-0.18), $ADEI_C$ (-0.04) and $ADG_C$ (-0.24)" (Le Graverand et al., 2023).

**Take-home messages about Chapter 5:**

- Accounting for the environment is essential to assess the predictive ability of traits by the rumen microbiota. To do so, spatio-temporal variables (pen, year, period) were part of the fixed effects used to predict traits.
- Predictions of production traits – feed efficiency, intake, growth and body composition- from fixed effects and body weight were more accurate or as accurate as predictions from microbiota data only.
- With microbiota data, the highest prediction accuracies were obtained for energy intake (correlations ranging between 0.05 and 0.56). Predictions accuracies were low to moderate for feed efficiency (0.06-0.35), for daily gain (0.15-0.38) and body composition traits (-0.07-0.45). Slightly higher prediction accuracies were obtained under a mixed diet, compared to a concentrate diet.
- Integrating fixed effects, body weight and microbiota data by concatenating matrices never improved prediction accuracies significantly. Sometimes, concatenation integration significantly degraded the prediction accuracy compared to models fitted on systematic effects only.

# Chapter 6

# Study: Predicting feed efficiency of sheep from multi-omics

Chapter 6 presents unpublished results. Part of the work was done thanks to a collaboration with Kim-Anh Lê Cao, at the School of Mathematics and Statistics (internship from October to December 2023, University of Melbourne, Australia).

# Rationale

Omics may provide different insights into variations of feed efficiency. Few studies compared predictions of feed efficiency from different omics, in the same population. Thus, the most promising predictors of feed efficiency still have to be identified. Moreover, few studies integrated data in animals raised during different years. It remains unclear if we can identify suitable proxies when cohorts are raised at different times. Chapter 6 focused on the prediction of feed efficiency or intake under a concentrate diet only.

First, traits were predicted from different blocks of proxies separately: either systematic effects, pedigree or omics. Omics included genomics, rumen metabarcoding, rumen metabolomics, rumen lipidomics, plasma metabolomics, and faecal phenomics. The first goal was to identify the best predictors by fitting one MINT-sPLSR per block (single-block, P integration). Prediction accuracy was computed as the Pearson correlation, by repeating random subsampling 100 times.

Second, predictions from each block of proxies were integrated to predict traits. The goal was to assess how integration influenced the prediction accuracy (multi-block, NP-integration). Feed efficiency and intake were predicted thanks to meta-models trained on the predictions of the single-block regressions above-mentioned. The same strategy was applied to assess the prediction accuracy for single-block and multi-block models, *i.e.* by repeating random subsampling 100 times.

# Results

The first section details prediction accuracies of feed efficiency and intake from single-block models. The second section details accuracies from multi-blocks models.

**Table 13: Prediction accuracy when lamb feed efficiency and energy intake were predicted from single-block and multi-block models**

| Diet[1] (n[2]) | Modelling | Predictors: Source | Variables | p[3] | Prediction accuracy (SD)[4]: $REI_C$ | | $FCR_C$ | | $ADEI_C$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| C-diet (255) | **Single-block** | | | | | | | | | |
| | MINT-sPLSR | Farm records | Fixed effects + body weight | 20 | $0.35^{de}$ | (0.13) | $0.48^{g}$ | (0.13) | $0.85^{f}$ | (0.05) |
| | | | Pedigree relatedness matrix | 255 | $0.54^{g}$ | (0.13) | $0.40^{ef}$ | (0.16) | $0.58^{e}$ | (0.13) |
| | | Blood | Genotypes | 29 830 | $0.54^{g}$ | (0.13) | $0.42^{f}$ | (0.15) | $0.58^{e}$ | (0.12) |
| | | | Plasma buckets | 863 | $0.37^{ef}$ | (0.14) | $0.41^{f}$ | (0.14) | $0.54^{cd}$ | (0.13) |
| | | | Plasma metabolite quantifications | 25 | $0.34^{d}$ | (0.15) | $0.32^{ab}$ | (0.14) | $0.57^{e}$ | (0.10) |
| | | Rumen | Prokaryote abundances | 582 | $0.20^{ab}$ | (0.15) | $0.35^{bcd}$ | (0.17) | $0.51^{bcd}$ | (0.12) |
| | | | Prokaryote functions | 281 | $0.18^{a}$ | (0.17) | $0.35^{bcd}$ | (0.18) | $0.52^{cd}$ | (0.11) |
| | | | Rumen buckets | 871 | $0.23^{b}$ | (0.17) | $0.41^{ef}$ | (0.16) | $0.46^{a}$ | (0.11) |
| | | | Rumen metabolite quantifications | 25 | $0.27^{c}$ | (0.17) | $0.34^{bc}$ | (0.15) | $0.49^{b}$ | (0.11) |
| | | | Rumen VFA concentrations | 6 | $0.24^{b}$ | (0.13) | $0.37^{de}$ | (0.14) | $0.51^{c}$ | (0.10) |
| | | | Rumen LCFA concentrations | 70 | $0.23^{b}$ | (0.15) | $0.31^{a}$ | (0.14) | $0.45^{a}$ | (0.13) |
| | | Faeces | Spectral absorbances | 1050 | $0.39^{f}$ | (0.14) | $0.32^{ab}$ | (0.15) | $0.54^{d}$ | (0.11) |
| | | | Spectral first derivative | 1050 | $0.39^{f}$ | (0.14) | $0.36^{cd}$ | (0.15) | $0.52^{c}$ | (0.12) |
| | **Multi-block** | | | | | | | | | |
| | Weighted mean | All | Single-block predictions | 13 | $0.58^{h}$ | (0.11) | $0.52^{h}$ | (0.12) | $0.86^{g}$ | (0.05) |
| | MINT-sPLSR | | Single-block predictions | 13 | $0.59^{h}$ | (0.12) | $0.49^{g}$ | (0.15) | $0.88^{h}$ | (0.04) |

Abbreviations: $ADEI_C$: average daily energy intake; C-diet: concentrate diet; $FCR_C$: feed conversion ratio; LCFA: long-chain fatty acids; MINT-sPLSR: multivariate integrative sparse partial least-squares regression; $REI_C$: residual energy intake; SD: standard deviation; VFA: volatile fatty acids.

Single-block predictions were carried out with MINT-sPLSRs. Multi-block predictions were carried out by weight averaging the single-block predictions or regressing them with MINT-sPLSR. Pearson correlations between predicted and actual phenotypes were averaged over testing sets. Testing sets were defined by repeating random subsampling 100 times (training=60% of data, validation=30%, testing=10%). Standard deviations are given in brackets.

[1]: C-diet: sheep were fed a 100% concentrate diet.

[2]: Number of samples

[3]: Number of predictor variables

[4]: Average Pearson correlation and standard deviation in brackets

[a,b,c,d,e,f,g,h]: Column-wise, average correlations with different letters significantly differ (p-value < 0.05 of permutation test, after Benjamini–Hochberg adjustment). Comparisons and adjustments were made per trait.

## A. Single-block prediction of lamb feed efficiency and intake

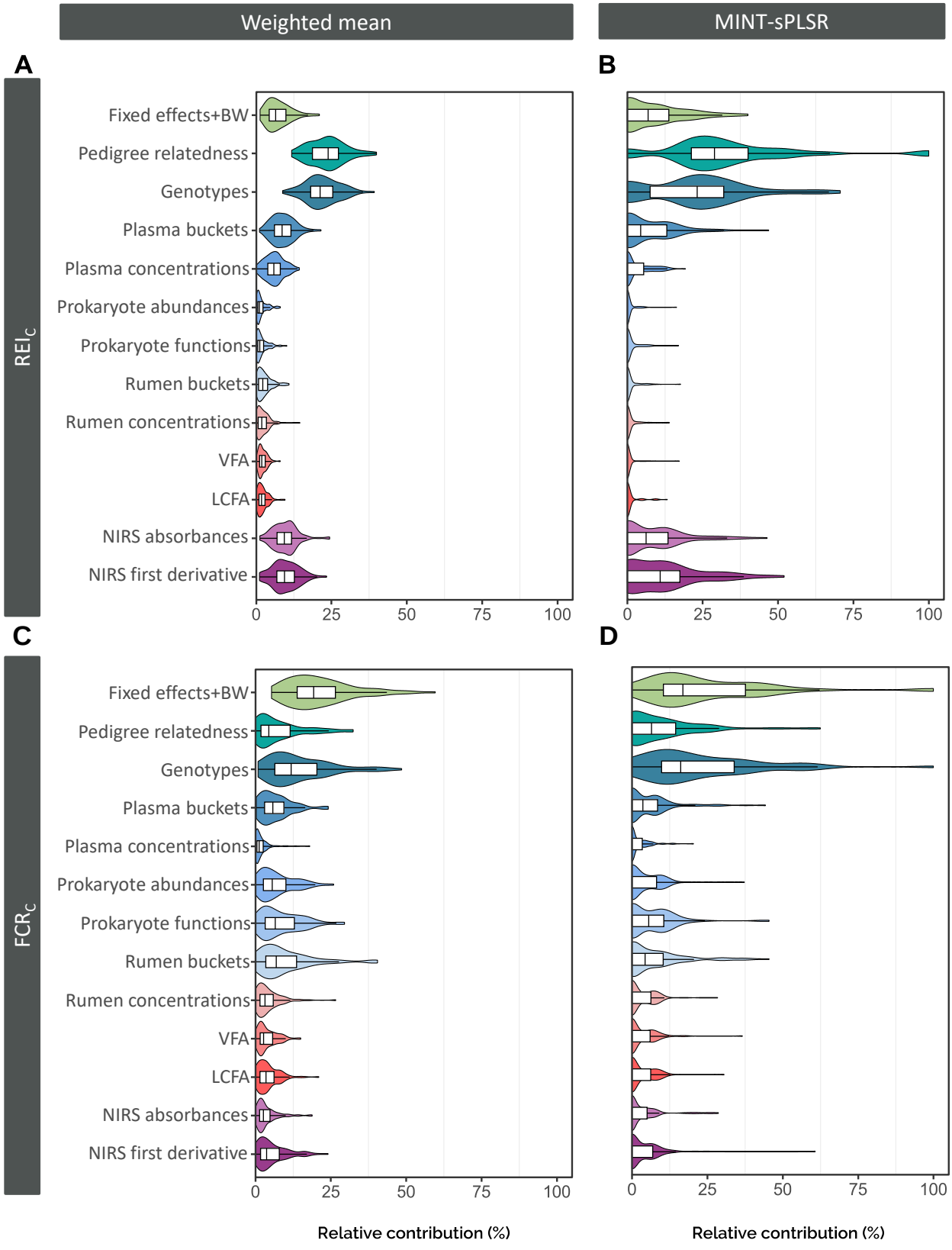
Under a concentrate diet, low to moderate prediction accuracies were obtained for feed efficiency traits ($REI_C$ and $FCR_C$): average Pearson correlations ranged between 0.18 and 0.54. For $REI_C$, the best predictors were either genotypes or pedigree relatedness (average correlations equal to 0.54). On the other hand, the worst predictors of $REI_C$ were ruminal variables: the rumen microbiota, metabolome and lipidome alone did not predict well $REI_C$ (0.20-0.27). For $FCR_C$, the best predictors were fixed effects and body weight (0.48). Then, the next best predictors of $FCR_C$ were the pedigree and genotypes (0.40-0.42), or the rumen and plasma metabolome buckets (0.41).

Whatever the set of predictors, $ADEI_C$ was more or as accurately predicted as feed efficiency traits. Prediction accuracies were moderate to high for $ADEI_C$ (0.45-0.85). Fixed effects and body weight were the best predictors of $ADEI_C$, followed by host genetics (0.58) and concentrations of plasma metabolites (0.57).

Whatever the trait, inferring new variables (*i.e.* prokaryote functions, metabolite quantifications and the first derivative of faecal NIRS) gave mixed results. Sometimes, prediction accuracies marginally increased or decreased compared to predictions drawn from the original variables (*i.e.* prokaryote abundances, NMR buckets, NIRS absorbances).

## B. Multi-block prediction of lamb feed efficiency and intake

When different sets of predictors were integrated, the best prediction accuracies were obtained for $ADEI_C$ (average correlations: 0.86-0.88), followed by $REI_C$ (0.58-0.59) and $FCR_C$ (0.50-0.52) (Table 13). To perform NP integration, using a MINT-sPLSR was the best strategy to predict $ADEI_C$, while weight averaging was the best strategy for $FCR_C$. Similar prediction accuracies were obtained when a weighted mean and a MINT-sPLSR were used to integrate multi-block and predict $REI_C$.




**See the subsequent page for more subplots and information**

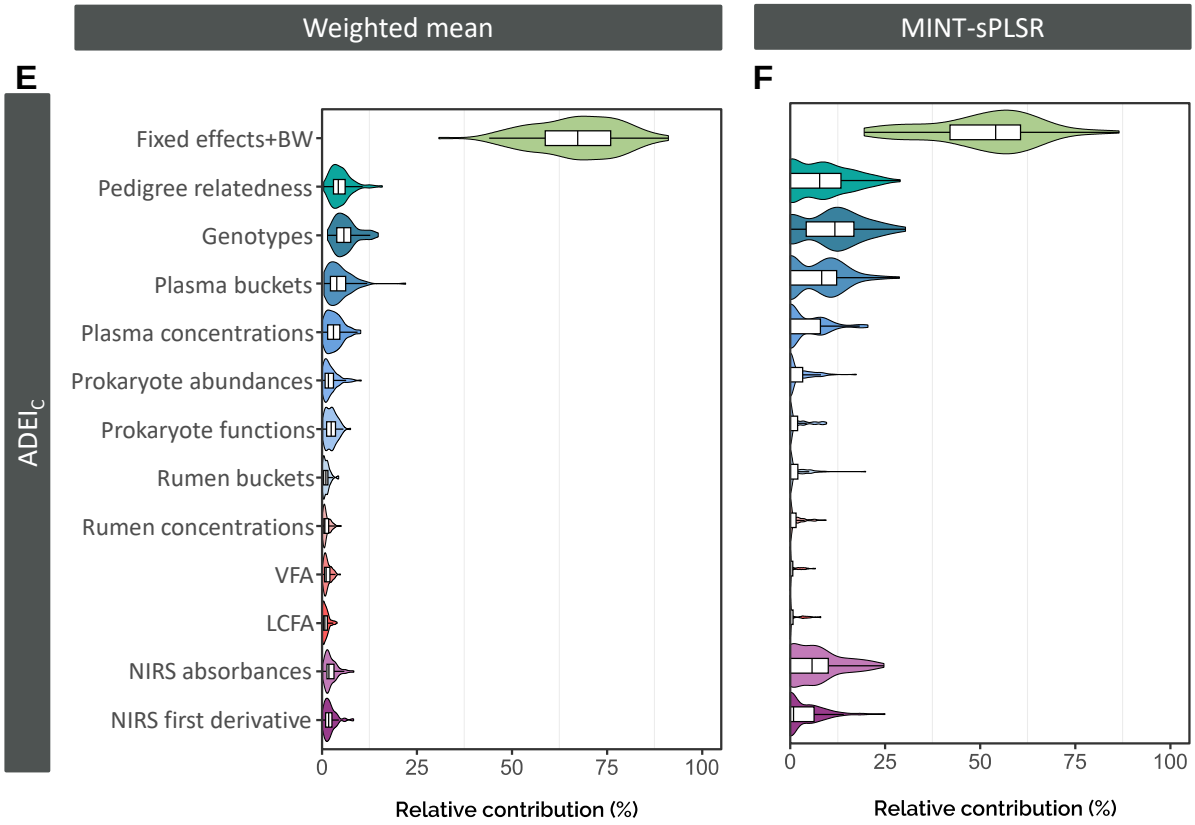



**Figure 27: Distribution of block contributions to the prediction of feed efficiency and intake**
Abbreviations: $ADEI_C$: average daily energy intake; BW: body weight; $FCR_C$: feed conversion ratio; $REI_C$: residual energy intake; LCFA: long-chain fatty acids; VFA: volatile fatty acids; MINT-sPLSR: multivariate integrative sparse partial least-squares regression; NIRS: near-infrared spectroscopy.
Feed efficiency and intake traits were predicted separately, under a concentrate diet. Block integration was carried out with weighted means or MINT-sPLSRs. Relative contributions of blocks were computed over each validation set. Validation sets were defined by repeating random subsampling 100 times (training=60% of data, validation=30%, testing=10%). Violin plots represent the density curves. Boxplots represent the 25th, 50th and 75th percentiles.

However, contributions of predictor blocks varied between predicted traits (Figure 27). Contributions were not compared between different integration strategies, since the computation of contributions differ. Moreover, when the integration strategy relied on a weighted mean, no block selection was made and all 13 blocks contributed to predictions. When the strategy involved a MINT-sPLSR, 6.64 blocks were selected on average to predict $REI_C$, 8.91 for $FCR_C$ and 8.81 for $ADEI_C$. Blocks scarcely selected by the MINT-sPLSR had low average contributions.

**Table 14: Average block contributions to the prediction of feed efficiency and intake**

| Diet[1] (n[2]) | Source | Predictors | p[3] | Average contribution[4] per trait and integration | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | $REI_C$ | | $FCR_C$ | | $ADEI_C$ | |
| | | | | Weighted mean | MINT-sPLSR | Weighted mean | MINT-sPLSR | Weighted mean | MINT-sPLSR |
| C-diet (255) | | Fixed effects + body weight | 20 | 7.5 | 9.4 | 21.5 | 21.3 | 66.9 | 41.8 |
| | | Pedigree relatedness matrix | 255 | 23.7 | 29.4 | 7.9 | 10.8 | 5.0 | 8.7 |
| | Blood | Genotypes | 29 830 | 21.5 | 21.0 | 14.6 | 17.2 | 6.3 | 11.3 |
| | | Plasma buckets | 863 | 9.0 | 8.5 | 6.8 | 5.8 | 4.6 | 8.0 |
| | | Plasma metabolite quantifications | 25 | 6.1 | 4.4 | 2.0 | 2.5 | 3.3 | 5.2 |
| | Rumen | Prokaryote abundances | 582 | 1.7 | 1.0 | 7.0 | 5.8 | 2.3 | 3.4 |
| | | Prokaryote functions | 281 | 1.8 | 1.6 | 8.6 | 7.4 | 2.5 | 2.9 |
| | | Rumen buckets | 871 | 2.7 | 1.3 | 10.0 | 8.4 | 1.2 | 2.5 |
| | | Rumen metabolite quantifications | 25 | 2.2 | 1.1 | 4.2 | 4.8 | 1.2 | 1.8 |
| | | Rumen VFA concentrations | 6 | 2.1 | 0.6 | 3.9 | 3.5 | 1.4 | 1.6 |
| | | Rumen LCFA concentrations | 70 | 2.2 | 1.4 | 4.3 | 3.9 | 1.0 | 1.7 |
| | Faeces | Spectral absorbances | 1050 | 9.6 | 9.1 | 3.7 | 3.8 | 2.3 | 6.5 |
| | | Spectral first derivative | 1050 | 9.8 | 11.2 | 5.5 | 4.9 | 2.1 | 4.6 |

Abbreviations: $ADEI_C$: average daily energy intake; C-diet: concentrate diet; $FCR_C$: feed conversion ratio; LCFA: long-chain fatty acids; MINT-sPLSR: multivariate integrative sparse partial least-squares regression; $REI_C$: residual energy intake; SD: standard deviation; VFA: volatile fatty acids.

Single-block predictions were carried out with MINT-sPLSRs. Multi-block predictions were carried out by weight averaging the single-block predictions or regressing them with MINT-sPLSR. Pearson correlations between predicted and actual phenotypes were averaged over testing sets. Testing sets were defined by repeating random subsampling 100 times (training=60% of data, validation=30%, testing=10%).

[1]: C-diet: sheep were fed a 100% concentrate diet.

[2]: Number of samples

[3]: Number of predictor variables

[4]: Relative contribution averaged over the 100 repetitions (relative weight in weighted means, or value importance in the projection of MINT-sPLSR).

Whatever the integration strategy, blocks were almost ranked in the same order according to their average contribution (Table 14). When $REI_C$ was predicted, the blocks having the highest contribution were: the pedigree, genotypes, NIRS variables and plasma buckets to predict $REI_C$. When $FCR_C$ was predicted, the most important blocks were the fixed effects and body weight, the genotypes, the pedigree and rumen buckets. Depending on the feed efficiency criterion, different buckets had the highest contribution when all blocks were integrated. When $ADEI_C$ was predicted, the block including fixed effects and body weight had by far the greatest contribution.

**Take-home messages about Chapter 6:**

- Feed efficiency and intake were predicted from single-block models (either systematic effects, pedigree or one set of omics). For feed efficiency, average prediction accuracies varied between 0.18 and 0.54. For energy intake, accuracies were higher: from 0.46 to 0.85.
- For $REI_C$, the best predictors were the host genotypes or pedigree, prior to NP-integration. After, the best model was a MINT-sPLSR fitted 6.64 predictor blocks on average.
- For $FCR_C$, the best predictors were the fixed effects and body weight, without NP-integration. Then, the best model was the weighted mean fitted on all blocks.
- For $ADEI_C$, the best proxies were fixed effects and body weight, prior to NP-integration. After, the best model was a MINT-sPLSR fitted on 8.81 blocks on average. However, one block had the highest contribution by far: the block with fixed effects and body weight.
- The rumen microbiota and lipidome did not predict well feed efficiency and intake. The plasma metabolome and faecal phenome contributed moderately to predictions.
- The prediction accuracy of multi-block models was marginally higher than the best single-block models (+0.05 maximum for REIc; +0.04 for FCRc; +0.03 for ADEIc).

# Chapter 7

# Thesis discussion

In Chapter 7, a few elements were adapted from the “Discussion” section published in *animal* (Le Graverand et al., 2023). Reprinted parts are highlighted with quotation marks. Most parts were added to cover all the thesis work.

In sheep, finding proxies of feed efficiency would widespread its selection. The literature suggests that proxies might be identified by profiling the host and its microbiota thanks to omics (Karisa et al., 2014; Shetty et al., 2017; Zhang et al., 2020; Marie-Etancelin et al., 2021). The thesis assessed the potential of various proxies: fixed effects, body weight, pedigree, genomics, rumen metabarcoding, rumen and plasma metabolomics, rumen lipidomics or faecal phenomics. **The main goal was to identify the best predictors of feed efficiency and intake, in meat sheep.**

The thesis relied on one original experimental design: no study has ever collected as many omics data in the same sheep population (8 techniques, 34 927 retained variables). Furthermore, the sheep number is relatively high compared to studies investigating several omics (around 250 lambs). Moreover, the rumen microbiota was studied under two successive diets: a concentrate diet and later a mixed diet.

Compared to the literature, the experimental design also stands out thanks to two sheep lines genetically selected for a higher or a lower feed efficiency. Divergent selection can help identify feed efficiency determinants by exacerbating the genetic differences between the most and least efficient sheep. Divergently selecting also helps explore how correlated traits will respond to selection (Gilbert et al., 2017; Tortereau et al., 2020).

First, the discussion will ponder the consequences of the divergent selection for feed efficiency over production traits and omics profiles. Then, the most promising predictors of feed efficiency will be reviewed. Finally, this chapter will underline the potential of data integration to predict animal traits.

# I- The consequences of feed efficiency selection

The present section will assess the consequences of selection for feed efficiency over production traits first and omics profiles later.

## A. Selection altered several lamb traits

### A.1 Feed efficiency and intake

Selecting for a lower residual feed intake should increase efficiency by reducing feed intake and maintaining the production (Koch et al., 1963). Under a concentrate diet, we did observe a significant difference between the feed efficiency and intake least-square means of the two Romane divergent lines. As expected, efficient lambs had lower $REI_C$, $FCR_C$ and $ADEI_C$ (adjusted p-values <0.001, Chapter 3 Table 8). Similar results were observed in cattle and pigs divergently selected for RFI, as well as rabbits selected for a decreased RFI (Arthur et al., 2001; Gilbert et al., 2017; Garreau et al., 2019).

In Romane male lambs, we observed an indirect and favorable response to selection when animals switched to a mixed diet. The feed efficient line still had a significantly higher feed efficiency and a lower intake under the mixed diet (adjusted p-values < 0.05, Chapter 3 Table 8). In whether lambs, Ellison et al. (2022) observed a positive correlation between RFI under a concentrate diet and a pelleted forage-based diet. Nowadays, most breeding companies select rams under a concentrate diet, during the growing period. Results suggests that if these companies selected for feed efficiency, it would also improve the efficiency in growing animals eating concentrates and forages. However, more lambs may graze or consume low-quality forages in the future.

Only growing male lambs were studied during the thesis. The diet and physiological determinants of feed efficiency likely differ between lambs and mature ewes. Mature ewes do not grow but they may mate, gestate, suckle and mobilize their body reserves. In dairy cattle, Macdonald et al. (2014) selected the 10% least efficient growing Holstein calves and the 10% most efficient. While heifer calves RFIs differed by 17.0% during growth, later, the difference was reduced to 2.4% during the first lactation (Macdonald et al., 2014). It suggests that feed efficiency may remain weakly and favorably correlated at different

physiological stages. More research is needed to assess the association between feed efficiency in growing lambs and mature ewes. Ewes constitute most of the mature sheep population.

## A.2 Final body weight and muscle depth under a concentrate diet

Final body weight and muscle depth significantly differed between the two lines, under a concentrate diet only (adjusted p-value < 0.05, Chapter 3 Table 8). Selection for RFI could directly result in a body weight decrease: a similar selection experiment in rabbits showed that selecting for a decreased RFI during 10 generations significantly decreased body weights of young animals too (Garreau et al., 2019). Even if we regress feed intake phenotypes over the metabolic weight and muscle depth to compute RFI, it does not ensure that regressors and RFI are genetically independent (Kennedy et al., 1993). However, the genetic correlation was almost null between $RFI_C$ and final $BW_C$ in Romane sheep ($r_{genetic}$= −0.03±0.19), while the correlation was moderate between $RFI_C$ and $MD_C$ ($r_{genetic}$= −0.30±0.15) (Tortereau et al., 2020). Results still have to be checked with more data and more generations.

Changes in body weight and $MD_C$ might also be indirect consequences of selection. Only 10-12 sires mated per year which may result in a bottleneck effect. It takes two years to phenotype a complete generation of selection, thus only 1.5 generations were phenotyped between 2018 and 2020. The difference between body weight and muscle depth might not remain significant when different sires mate and produce the next generations.

Under a mixed diet, final body weight and muscle depth differences decreased and were not significant anymore. Differences may decrease because male lambs are getting close to the end of growth, around 43 weeks age. Compensatory growth phenomena might also affect differently the two lines' growth. Differences might also be less pronounced under a mixed diet, since selection was carried out under a concentrate diet.

## B. Selection structured the study population

In our experiment, the divergent selection clearly structured the lamb population. PCA evidenced that lines and families could be easily distinguished by their genetic (pedigree relatedness) and genomic backgrounds (genomic relatedness). The result was expected: the goal of divergent selection is to exacerbate the difference between extreme animals' genetics. In Romane lambs, the divergence was equal to 1.86 genetic deviations after selecting for $RFI_C$ during almost three generations. In pigs, the divergence reached 3.84 deviations after selecting for RFI during nine generations (Gilbert et al., 2017).

In our experiment, the predictive ability of genotypes and pedigree data is likely high (see section II B.2 for the prediction of phenotypes). We studied a population strongly structured by family, with two divergent lines part of one flock. Discriminant analyzes easily predicted the Romane efficiency line from genotypes and pedigree data (average BERs=0.00, Chapter 3 Table 10). Such results are unlikely if we predict extreme phenotypic groups, in a commercial population where animals may be less related. For instance, Shirzadifar et al. (2022) picked beef cattle to retain the top 15% and the worst 15% RFI phenotypes, from commercial and research populations. They observed a moderate error rate (0.39) when the RFI group was predicted from beef genotypes with a k-nearest neighbor classifier.

## C. Rumen, plasma and faecal omic profiles scarcely diverged between selected lines

The thesis assessed whether the rumen microbiota, metabolome and lipidome / the plasma metabolome / the faecal NIR spectra diverged between the two selected lines.

First, discriminant analyzes highlighted that the examined rumen, plasma and faecal omics could not predict accurately the Romane divergent line. When we looked for rumen, plasma and faecal signatures of selection, the BERs of discriminant analyzes were very high: from 0.42 to 0.51 (Chapter 3 Table 10). Thus, divergently selecting for feed efficiency had little effect over rumen, plasma and faecal omics. However, it does not mean that selection had no effect at all. It only suggests that divergence between the rumen, plasma and faecal profiles are not sufficient to discriminate lines.

Second, fitting PCAs on rumen microbiota abundances, plasma metabolite concentrations and the first derivative of faecal NIR spectra variations showed that the year or period of phenotyping were associated with the first components. Thus, it may be difficult to identify microbiota, metabolomic and phenomic variables discriminating efficiency groups when animals are raised at different times or under different diets.

## C.1 The spectral phenome of faeces

Faecal NIRS may not discriminate feed efficiency categories well when all animals are fed the same diet. In lambs, the faecal NIRS poorly discriminated feed efficiency lines, under a concentrate diet (0.43 ≤ BER ≤ 0.44, Chapter 3 Table 10). In beef steers, faecal NIRS (error rate = 0.46) also poorly discriminated extreme RFI groups, under a mixed diet (Meale et al., 2017). Previously, faecal NIR spectra were used to estimate the feed quality, digestibility or intake in sheep eating different diets (Andueza et al., 2017). However, breeding companies select rams under the same diet.

## C.2 The rumen and plasma metabolomics

Under a concentrate diet, Romane divergent lines were better discriminated by plasma metabolomics (0.42 ≤ BERs ≤ 0.43, Chapter 3 Table 10) than rumen metabolomics (0.48 ≤ BERs ≤ 0.50). In the same population, Touitou et al. (2022) evidenced it before by computing the AUROC: 0.81 for plasma, against 0.70 for rumen metabolomics. Very similar results were obtained when extreme RFI groups were discriminated from steers' plasma NIRS (error rate= 0.46) (Meale et al., 2017), or lambs' serum metabolome (AUROC=0.80) (Goldansaz et al., 2020).

If the objective is to discriminate extreme feed efficiency groups, it suggests that the plasma or serum metabolomes may discriminate just as well. If the objective is to understand how the metabolism influences feed efficiency, plasma and serum may provide different insights. No common metabolite was pinpointed in both studies, but similar biological pathways were proposed: Touitou et al. (2022) highlighted metabolites involved either in the energy production by mitochondria or the protein turnover; while Goldansaz et al. (2020) identified proxies involved either in the energy regulation or protein synthesis. Discrepancies may also be explained by differences in diet, analysis, and group partition: one study attempted to

discriminate genetic lines (Touitou et al., 2022), while the other examined phenotypic groups (Goldansaz et al., 2020).

### C.3 The rumen microbiota

The rumen microbiota hardly discriminated Romane feed efficiency lines. It would suggest that selecting for feed efficiency did not impair the microbiota. In dairy sheep, divergently selecting for either milk persistency or somatic cell count scarcely altered the rumen microbiota too: BERs of discriminant analyzes ranged from 0.50 to 0.71, on average (Martinez Boggio et al., 2021). In the future, longer-term selection could further alter the rumen microbiota composition of sheep. “In monogastrics, divergent selection of pigs for RFI during 9-10 generations resulted in significant differences in the abundances of 52 faecal bacterial genera between the RFI lines (Aliakbari et al., 2021). With rabbits, discriminant analysis of principal components was able to discriminate a line selected for a decreased RFI during 9 generations and the ancestral line based on their caecal bacterial phylotypes (Drouilhet et al., 2016). The discrepancy between these monogastrics results and ours may be due to much more advanced divergence” (Le Graverand et al., 2023). Discrepancies may also be explained by which microbiota was studied: ruminal in sheep, faecal in pigs, caecal in rabbits. Furthermore, the microbiota composition was assessed by cross-validating discriminant analyzes in sheep. Therefore, slight microbiota differences were overlooked if they were not sufficient to discriminate lines accurately. In rabbits and pigs, slight microbiota differences could be picked up by differential analyzes and unvalidated discriminant analyzes.

Furthermore, environmental effects might exceed the host genetics influence. Past studies underlined the dramatic effect of the diet over the microbiota composition. Four times more rumen OTUs differed between diets than lamb RFI categories (Ellison et al., 2017). Similarly, the abundance of 91 genera out of 114 differed between two diets, while no genera was associated with Romane lamb RFI (Marie-Etancelin et al., 2021).

### C.4 The rumen lipidome

The rumen lipidome was partially studied in Romane divergent lines: VFAs and LCFAs hardly discriminated the two lines (0.46 ≤ BER ≤ 0.51, Chapter 3 Table 10). Thus, it suggests that lipidome profiles diverged little between the two efficiency lines.

Fregulia et al. (2021) reviewed 14 studies in cattle and highlighted discrepancies. Depending on experimental conditions, the same VFA could be alternatively associated with the most efficient cattle, the least efficient animals or none at all. Thus, it may be difficult to pinpoint discriminant variables when lambs are raised in different conditions.

**Take-home messages about the consequences of divergent selection:**

- Production traits, rumen, plasma and fecal omics were investigated in lambs divergently selected for feed efficiency.
- Divergent selection altered both efficiency and intake under a concentrate diet and a mixed diet later. Muscle depth and final body weight decreased in the feed efficient line, under a concentrate diet only. Such decreases could result from indirect responses dictated by genetic correlations, or a bottleneck effect.
- Rumen microbiota, metabolomics and lipidomics hardly discriminated feed efficiency lines. Plasma metabolomics and faecal NIR spectra could discriminate lines slightly better.
- On the other hand, divergent lines could be easily discriminated from genotypes and pedigree data. It highlights that their prediction accuracy is high in divergent experiments, where the population has a strong familial structure. Lower accuracies could be expected when animals are less closely related, such as in commercial populations.

# II- Proxies of feed efficiency and intake

This section will focus on proxies of feed efficiency and intake phenotypes. The thesis did not assess whether EBVs could be predicted from omics. While phenotypes are directly measured, EBVs are estimations. We studied the last divergent lines generations, which had few or no descendants. Thus, we deemed that the reliability of EBVs was not high enough to assess models predicting EBVs from omics. For instance, coefficients of determination of feed intake breeding values were low: 0.41 on average (range: 0.34 to 0.66). Nevertheless, EBVs from predicted phenotypes are discussed later (section III B.2).

Next sections will discuss how we assessed potential proxies of feed efficiency and intake: the fixed effects, body weight, pedigree and omics data. Then, best potential proxies will be discussed based on the thesis results and the literature. Proxies will be considered separately, without N-integration.

## A. Assessing the worth of proxies

Feed efficiency and intake phenotypes are partially determined by environment and experiment conditions, such as the diet and climate (Butler-Hogg and Tulloh, 1982; Kamalzadeh et al., 1997; Padua et al., 1997). These conditions may have an incidence on potential proxies too, as highlighted before (Chapter 7, section I. C). Thus, studying lambs raised during different years (*e.g.* in different environments) may help assess the worth of proxies.

Most papers assess the generalizability of proxies: they are promising if predictive models generalize well. A model is generalizable when it is fitted on one dataset (training) and the model may predict well the same outcome on a new dataset (testing) (Kuhn and Johnson, 2013). The testing set must be unseen during the training. Computing the prediction accuracy with testing sets is necessary to check whether proxies could predict new data, or results only overfits.

Most papers do not assess if proxies are practical: new proxies are not useful if they do not bring new insights into traits. New proxies may be redundant with variables which are easier

to collect. Integrating redundant variables is just counterproductive: multiplying the number of proxies is a waste of resources if it does not improve the prediction accuracy.

In breeding companies, several variables are already recorded: the experimental conditions such as the age; and production traits such as the live weight. It is less expensive to record these variables routinely, than to collect omics data. Per sample, analysis costs can range between 24€ for genotypes to 211€ for LCFAs. To justify expenses, models including omics must predict better than regressions fitted on cheaper variables only. For instance, Maltecca et al. (2019) and Velasco-Galilea et al. (2021) assessed the predictive ability of the microbiota with and without spatiotemporal fixed effects: such as the pen or batch of animals. In rabbits, authors showed that adding microbiota data and systematic effects improved the prediction of RFI (Velasco-Galilea et al., 2021). In swine, a similar observation was reached for growth and carcass traits (Maltecca et al., 2019).

During the thesis, we considered the following cheap and routine variables: the pen, the age, the body weight, the year and/or period of sampling. All these variables may already help predict feed efficiency and intake. Indeed, male lambs could not be grouped randomly: the body weight had to be homogeneous per pen to prevent excessive fighting. Then, conditions slightly varied between years despite all attempts: lambs were slightly older in 2018; feeders did not stop in 2019 the day before sampling; the diet chemical composition varied; and weather conditions changed from year to year.

So, lambs' weights and ages were spuriously correlated with the spatiotemporal environment (Figure 28). Microbiota, metabolomics, lipidomics, phenomics and lamb traits are indirectly correlated because of environmental effects. The correlation may exist even if there are no causal relationship between omics and traits. Thus, conclusions about omics might be overoptimistic when correlations are ignored with experiment conditions. Intermediate omics were suggested to replace unregistered environmental effects. He et al. (2022b) proposed the microbiota metabarcoding as an indicator of animals' raising conditions (more perspectives are discussed in the section III. B).

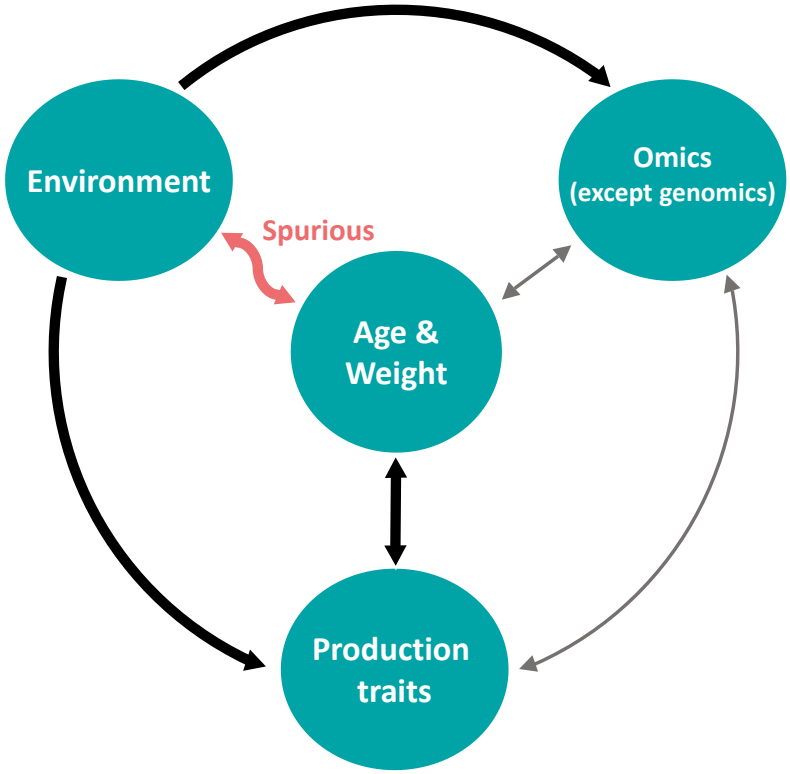



**Figure 28: Associations between the environment, intermediate omics and production traits**

## B. Most and least promising proxies

Since determinisms differed between efficiency and intake traits, the best proxies varied accordingly. Here, the predictive abilities of proxies examined during the thesis will be discussed separately - without data integration.

### B.1 Fixed effects and body weight were the best proxies for weight-dependent traits

In Romane lambs, fixed effects and final body weight were the best proxies of feed intake, whatever the diet: Pearson correlations varied between 0.76 and 0.85 (Chapter 5 Figure 26 and Chapter 6 Table 13). Fixed effects and body weight predicted $FCR_C$ less well, but they remained its best predictors: Pearson correlations ranged between 0.40 and 0.48. On the opposite, fixed effects and body weight did not predict $REI_C$ as accurately as feed intake and $FCR_C$: correlations ranged from 0.31 to 0.35. Similar conclusions were reached in rabbits for feed intake and efficiency (Velasco-Galilea et al., 2021).

Such results were expected: feed intake and $FCR_C$ vary with the animal weight. Phenotypic correlations evidenced this relationship in Romane lambs: 0.78±0.01 between $ADEI_C$ and final $BW_C$; -0.77±0.01 between $FCR_C$ and final $BW_C$ (Tortereau et al., 2020).

On the opposite, $REI_C$ phenotypes are independent of animal weight differences (Koch et al., 1963). Indeed, $REI_C$ is a composite trait computed by regressing intake over the metabolic weight, growth and body composition effects. It explains why models fitted on fixed effects and body weight do not generalize well to predict $REI_C$. It stresses the need for new proxies to predict residual efficiency traits, such as pedigree and omics data.

## B.2 Genetics and genomics were the best proxies of residual feed intake in divergent lines

In Romane divergent lines, the best $REI_C$ proxies were genotypes and pedigree relatedness: average correlations were both equal to 0.54 (Chapter 6 Table 13). Compared to our experiment population, animals may be less closely related in commercial populations. Thus, genomic prediction accuracies might be lower in commercial populations. In cattle for example, past studies observed lower genomic prediction accuracies: average correlations between RFI phenotypes and genomic EBVs reached a maximum of 0.19, in growing Australian Holsteins or Nellores (Pryce et al., 2012; Silva et al., 2016). However, RFI heritabilities were low in these cattle populations: $h^2$ =0.22±0.07 in Holstein, and 0.17±0.07 in Nellore (Pryce et al., 2012; Silva et al., 2016). On the other hand, heritability was higher in one Romane flock: $h^2$=0.45±0.08, before divergent selection (Tortereau et al., 2020).

Genotypes and pedigree relatedness were also the second-best proxies for $FCR_C$ and $ADEI_C$: with average correlations ranging between 0.40 and 0.42 for $FCR_C$, against 0.58 for $ADEI_C$ (Chapter 6 Table 13).

Given that heritabilities are significant, genomic and genetic selections seem feasible in Romane sheep. However, accuracies and reliabilities have to be checked more thoroughly in a commercial population. Furthermore, the thesis experimental design and results cannot demonstrate whether genomic selection should be preferred over genetic selection: MINT-sPLSRs performed similarly when regressions were fitted on pedigree relatedness or genotypes allelic counts. Finally, only SNPs with a minor allele frequency higher than 0.20 could be retained to carry out sPLS models. Otherwise, sPLS models did not always

converge. Mixed models and Bayesian models may include more SNPs and help compare genetic and genomic accuracies.

### B.3 Rumen omics were the least promising proxies

Before, the rumen microbiota, metabolome and lipidome seemed promising since the digestion of plants by rumen microorganisms is essential in ruminants (Domingues Millen et al., 2016).

In lambs, however, the rumen microbiota did not predict well phenotypes. Predictions from the rumen microbiota were less or as accurate as predictions from fixed effects and/or production traits, whatever the sequencing (16S or 18S rRNA gene), the regression model and the cross-validation strategy (Chapter 4 Table 11 and Chapter 6 Table 13). For REI, correlations were null to low between actual phenotypes and predictions from microbiota data (0.06-0.35). Prediction accuracies seemed a bit higher under a mixed diet. Differences between diets might be explained by fiber content differences, or the experimental design: lambs were phenotyped during summer or fall under a mixed diet, and the rumen microbiota composition differed between periods (Chapter 5 Figure 26). Our accuracy results contradicts Ellison et al. (2019): they observed a correlation of 0.71 between lamb RFI and predictions from 16S data (against 0.35 for $REI_M$). Ellison et al. (2019) picked a small subset of animals to get three groups: one with average RFI phenotypes and two with extreme RFI values (16 lambs to train the model, 20 to test it). Thus, the retained phenotypes do not follow the same distribution as the whole original population. Having contrasting groups may inflate the Pearson correlation between phenotypes and predictions (Aggarwal and Ranganathan, 2016).

We do not advocate for rumen metabarcoding to predict feed efficiency and intake. However, untargeted analyses could help identify better proxies, by censing all kinds of microorganisms. For instance, shotgun sequencing or restriction enzyme-reduced representation sequencing can detect bacteria, archaea, fungi, protozoa and viruses (Hess et al., 2020). Even if the rumen microbiota did not predict well feed efficiency, rumen metagenomics and metabarcoding data could help predict methane emissions. Methane is a direct by-product of rumen microorganisms which degrade fiber by fermenting (Hill et al., 2016). Furthermore, additive models assessed how much phenotypic variance was

associated with genetics and the rumen microbiota. In dairy cattle, microbiabilities varied between 0.13 and 0.31 for methane production (Difford et al., 2018; Saborío-Montero et al., 2021). However, high microbiabilities do not guarantee that microbiota data can predict traits accurately. Models may overfit. In sheep, microbiabilities ranged between 0.78 and 0.96 for methane emissions, but prediction accuracies ranged from 0.29 to 0.38 (Hess et al., 2022).

Rumen metabolome and fatty acids did not predict well $REI_C$: correlations between actual and predicted phenotypes ranged from 0.23 to 0.27 (Chapter 6 Table 13). Thus, all models fitted on rumen data did not generalize well to predict $REI_C$.

The low repeatability of rumen proxies might explain why ruminal traits cannot be considered as accurate proxies of host traits. Fresco et al. (2022) studied several rumen omics in dairy ewes sampled two weeks apart. Most microbiota abundances, metabolites and fatty acids were unstable from one week to another. The median repeatabilities were low: 0.15 for OTUs (range of OTU repeatabilities : 0 - 0.93); 0.44 for VFAs (0.21 to 0.57); 0.21 for LCFAs (0.02-0.86) (Fresco et al., 2022). Correlations between variables were also unstable. Twenty OTUs were correlated with rumen LCFAs the first week, but only 3 correlations remained significant the second week (Fresco et al., 2022). Low repeatabilities and correlation instabilities suggest that rumen omics are sensitive to random, environmental, technical and/or analysis parameters. For instance, saliva and different rumen fractions may be found in rumen fluid samples collected with gastric tubes (Henderson et al., 2013; Terré et al., 2013).

Thus, the thesis suggests that the rumen is not the best gut location to sample, if the goal is to predict feed intake and efficiency in growing lambs. Other gut locations, such as the rectum, might provide better proxies.

## B.4 Faecal and plasma omics may provide non-invasive proxies

Sampling faeces and blood is less invasive than sampling the rumen. Non-invasive sampling is primordial to collect data routinely.

Faeces can be collected directly from the rectum. In pigs, such sampling was performed to predict feed efficiency, growth and carcass traits from the faecal microbiota (Maltecca et al., 2019; Aliakbari et al., 2022).

Besides the faecal microbiota, faecal NIRS might provide feed efficiency proxies too. In Romane divergent lines, average correlations between real $REI_C$ and predictions were equal to 0.39 (Chapter 6 Table 13). First derivative or absorbances were the second-best proxies of $REI_C$, right after pedigree and genotypes. Faecal NIRS may predict traits strongly related with the physico-chemical profile of samples. In pigs for example, faecal NIRS could accurately predict digestibility coefficients: correlations varied between 0.82 for crude fiber digestibility, and 0.95 for the nitrogen digestibility (Labussière et al., 2019). In ruminants, faecal NIRS is mainly used to predict diet characteristics. Lamb FCR could not be predicted under a single diet accurately ($R^2 \leq 0.15$), however the accuracy increased when lambs were fed two different diets ($R^2=0.69$) (SMARTER Deliverable 1.1, unpublished) In dairy cows however, milk mid-infrared spectroscopy was proposed to predict RFI: coefficients of determination were as high as 0.46 at the beginning of lactation (Shetty et al., 2017).

In Romane lambs, the plasma metabolome moderately predicted $REI_C$: correlations ranging from 0.34 to 0.37 (Chapter 6 Table 13). Accuracies were intermediate between faecal NIRS and rumen variables. In beef, 32% of RFI phenotypic variation could be explained by regressing RFI on three plasma metabolites (creatine, carnitine and hippurate) (Karisa et al., 2014). In Romane lambs however, the plasmatic concentrations of creatine and carnitine were quantified with ASICS but they were not associated with feed efficiency (VIP<1) (Touitou et al., 2022). Hippurate is part of the ASICS reference library, but it could not be detected and quantified from NMR spectra in Romane lambs.

## B.5 Data inference and transformation

In machine learning, feature engineering consists in selecting and transforming an original set of variables into a new set of variables (Kuhn and Johnson, 2013). Inference is an instance of feature engineering based on knowledge. Inference may be useful to get more interpretable predictors.

Inference was used to estimate the abundance of prokaryote functions from the abundance of prokaryote microorganisms. Original variables (transformed prokaryote abundances) and inferred data (function abundances) had similar prediction accuracies, when feed efficiency and intake were predicted. Correlations did not significantly differ (Chapter 6). We expected that inferred data would be more interesting for predictions, since different bacteria may have redundant functions (Weimer, 2015). Furthermore, in cows metagenomics discriminated feed efficiency groups better than metataxonomics (AUROC>0.9, AUROC>0.8 respectively) (Kruger Ben Shabat et al., 2016). However, inference models are only as good as the available knowledge. Picrust2 relies on a phylogenetic tree to infer the abundance of genes from the abundance of microorganisms (Douglas et al., 2020). Picrust2's tree is less exhaustive than the latest SILVA databases releases (Darbot et al., 2022). Therefore, while the OTU 647 belonged to the *Christensenellaceae R-7 group* based on SILVA's classification, the closest match in Picrust2's tree belonged to the genus *Alicyclobacillus*. The two genera are not closely related since they do not belong to the same class taxa. It may have introduced an inference artefact, which explains why two lamb groups were evidenced after the functional inference of prokaryotes (Chapter 3). Shotgun metagenomics would likely provide a more accurate census of microbial functions than inference could.

Inference was used to infer the concentration of metabolites from NMR spectra. No trend was clear: original variables (transformed bucket areas) could predict feed efficiency or intake traits better than, worse than or as well as inferred data (metabolite quantifications). The accuracy of metabolite concentrations relies on the exhaustiveness of the reference library. For instance, the ASICS library does not include urea (Lefort et al., 2019). Thus, urea quantifications could not be inferred from NMR spectra, even if the metabolite could help assess nitrogen cycles.

Transformation is another instance of feature engineering. The first derivative of NIRS was computed from spectra absorbances. NIRS absorbances and the first derivative could predict $REI_C$ with the same accuracy, but absorbances predicted $ADEI_C$ better and the first derivative predicted $FCR_C$ better.

Thus, the thesis does not allow to conclude whether data inference and data transformation are preferable to predict feed efficiency and intake from microbiota data, metabolomics and phenomics.

## B.6 Using other phenotypes to predict feed efficiency

While the thesis scrutinized omics a lot, they might not be the easiest and cheapest proxies of feed efficiency and intake. Some other trait phenotypes could be easier to collect and analyze. Many traits were associated with feed efficiency and could represent valuable proxies: such as heat production, behavior, and respiratory traits. However, caution is advised: few papers carried out cross-validation to check how accurately these traits could predict feed efficiency.

Heat production was proposed as a proxy for feed efficiency. When bull RFI is computed by regressing feed intake over weight gain and metabolic weight, 24% of efficiency variations may be explained by infrared thermography - with temperatures measured at 6 body locations (Montanholi et al., 2010). Feed efficient bulls had a significantly lower cheek, eye and feet temperatures. It was hypothesized that feed efficient ruminants may produce less heat thanks to a higher efficiency of mitochondrial respiration, or a lower protein turnover (Cantalapiedra-Hijar et al., 2018). Lines et al. (2014) found no association between cattle lines divergently selected for RFI and $CO_2$ entry rate - used as a proxy of heat production. However, Lines et al. (2014) underlined that heat production varies with the amount of ingested energy. Thus, feed efficient ruminants may produce less heat when they ingest less energy, not because they have a more efficient metabolism.

Behavioral traits may also provide valuable proxies of feed efficiency. Lamb feeding behavior was associated with feed efficiency under a concentrate diet: efficient Romane lambs visited less often feeders, spent less time eating per meal and ate when less animals were accessing feeders (Marie-Etancelin et al., 2023). It suggests that efficient animals might save energy by moving less to feed. On the other hand, inefficient animals might spend more energy moving or dominating other animals (Marie-Etancelin et al., 2023). Sepulveda et al. (2022) observed a strong correlation in ewes between RFI and the eating duration per visit ($correlation_{genetic}$= -0.50 ± 0.19). Such results and modelling responses to selection suggested that selecting ewes for longer meal duration would improve RFI (Sepulveda et al., 2022). However, in Romane lambs the genetic correlation seemed lower and less interesting: -0.22 ± 0.17 (Marie-Etancelin et al., 2019). In both cases, standard errors were relatively high. Bigger populations may be used to estimate more accurately feed efficiency gains if longer eating durations are selected per visit. Water intake behavior might also

provide proxies: RFI and water intake were genetically correlated in steers ($r_{genetic}$=0.33±0.11) (Ahlberg et al., 2019). Radio frequency identification and accelerometers could help measure behavioral traits, without investing as much as in automatic feeders. Furthermore, such systems may be used to phenotype new traits and putative proxies: such as study social traits and physical activity.

Respiratory gases were also considered as feed efficiency proxies: they may vary with the ingested feed, the physical activity and metabolism phenomena. In dairy cows, methane and carbon dioxide concentrations were examined as proxies of RFI respiratory gases (Difford et al., 2020). However, correlations between RFI and respiratory gases could be alternatively null, positive or negative: values differed between populations (Dutch or Danish cows) and RFI computations (phenotypic, genetic, single-step RFI) (Difford et al., 2020). In lambs and ewes, daily methane emissions could also be either insignificantly, positively or negatively correlated with RFI (Paganoni et al., 2017; Muir et al., 2020; Johnson et al., 2022; Navajas et al., 2022; Marques et al., 2022; Tortereau et al., 2023). These results clearly demonstrate that proxies of feed efficiency may be context-dependent.

Still, combining respiratory gases and production traits is promising: a PCA was fitted on body weight, ADG, $CO_2$, $O_2$ and $CH_4$ emissions in Uruguayan Merinos, then the first component could predict feed intake in Corriedale and Dohne flocks (correlations=0.73; SMARTER Deliverable 1.1, unpublished).

In Romane lambs fed a mixed diet, FCR was correlated with the difference between plasma and diet abundances of the $^{15}N$ isotope (correlations=-0.67; SMARTER Deliverable 1.1, unpublished). However, correlations were negligible under a concentrate diet. Correlations were also negligible with RFI.

**Take-home messages about the proxies of feed efficiency and intake:**

- During the thesis, the best proxies of REI were genotypes and pedigree relatedness. However, the family structure is strong in our divergent selection experiment. Thus, the potential of genetic and genomic selection should be assessed in a commercial population, where animals may be less closely related. The best proxies of $ADEI_C$ and $FCR_C$ were fixed effects and body weight during the thesis.
- Plasma metabolomics and fecal NIR spectra showed intermediate prediction accuracies for feed efficiency and intake.
- Collecting and analyzing omics require much resources, while the literature showed that several traits could be used as feed efficiency and intake proxies: such as heat production, behavior, respiratory traits. The predictive ability of many traits still has to be checked with cross-validations. Otherwise, claims might be overoptimistic.

## III- Integration perspectives to predict and select animal traits

As discussed before, past studies evidenced that the potential of proxies to predict animal traits could vary from one context to another. P-integration could help identify proxies which are generalizable in different contexts, by integrating different experimental conditions. Furthermore, the previous section lingered on the best potential proxies of feed efficiency and intake, but each kind of proxy was considered separately. N-integration could help predict animal traits more accurately by integrating different proxies.

Now the discussion will focus on the perspectives of N and P integration, to predict animal traits. First, the discussion will highlight the contributions and flaws of strategies used during the thesis. Then, the discussion will focus on how data integration could help select animals for complex traits.

# A. Integration strategies

Several integration approaches were tried throughout the thesis. The first subsection will discuss the prediction accuracies, opportunities and drawbacks of the thesis integration attempts. The second subsection will underline how the attempted integration strategies could be improved.

## A.1 Prediction accuracies, opportunities and drawbacks of strategies implemented in the thesis

### Integration of data collected during different years

The effect of P-integration over prediction accuracies may only be assessed for 16S data. In Chapter 5, sPLSR were carried out and did not account for the year of lambs' rearing. On the opposite, in Chapter 6, MINT-sPLSR accounted for the three different years. Prediction accuracies were close when sPLSR or MINT-sPLSR predicted traits from 16S data. However, the difference between models was not be tested, because cross-validation strategies differed.

The aim was not necessarily to increase the prediction accuracy. Compared to classic sPLS discriminant analyzes, past studies showed that MINT discriminant analyzes could either improve the prediction accuracy (Rohart et al., 2017a) or decrease it (Poirier et al., 2020). The goal of MINT-sPLSR was to account for batch effects, in order to identify generalizable proxies. For instance, feeders in 2019 kept delivering feed when animals were supposed to fast for the night, prior to blood and rumen sampling in the morning. Fasting duration has a major incidence on metabolomic profiles. Thus, P-integration provides the opportunity to try to account for a batch effect, instead of discarding a faulty batch dataset.

### Integration of fixed effects, covariates and microbiota data

In the first thesis studies (Chapter 4 and Chapter 5), early integration was attempted by concatenating the matrices: the incidence matrix of fixed effects and covariates and the transformed microbiota abundance matrix.

Few authors performed an early integration to predict host traits simultaneously from fixed effects, covariates and prokaryote abundances. Chapter 5 (Figure 26) prediction accuracies were consistent with past studies. Indeed, correlations between phenotypes and

predictions were equal to: 0.11 in sheep against 0.17 in rabbits for $REI_C$; 0.79 in sheep against 0.73 in rabbits for $ADEI_C$; 0.79 in sheep against 0.73 in rabbits for $ADG_C$, 0.62 in sheep against 0.50 in pigs for $BFT_C$ (Maltecca et al., 2019; Velasco-Galilea et al., 2021).

“Therefore, low to high prediction accuracies can be reached when systematic effects and microbiota are combined as predictors of host traits. However, it does not demonstrate that predictions benefit from the inclusion of microbiota variables. Prediction models may rely more on systematic effects, such as the live weight, than microbiota variables. Indeed, the present study illustrates that models integrating microbial predictors and systematic effects never significantly outperformed predictions from systematic effects only. Therefore, microbiota predictors did not provide any added value to improve prediction accuracies. Prediction accuracies did not benefit from any [complementarity] between microbial, fixed effects and body weight predictors. In addition, systematic effects (*i.e.,* fixed effects and body weight here) represented less-expensive predictors than metabarcoding variables. In contrast, Velasco-Galilea et al. (2021) demonstrated that sPLSR models predicted rabbit RFI more accurately when microbiota variables were included than models with systematic effects only. Similarly, Maltecca et al. (2019) showed that machine learning accuracy was improved by including microbiota data to predict swine growth and carcass traits. This was not the case in our study, regardless of the trait, microbiota sequencing or diet.

The sheep rumen microbiota may not have a negligible predictive ability for feed efficiency or its components. However, the present study suggests that recording systematic effects, such as animal weight and environmental effects, might be more effective than sampling and metabarcoding the rumen fluid. Most studies did not consider the predictive ability of systematic effects to establish a baseline and prove the utility of the microbiota in host trait predictions. We conclude that host trait predictions did not benefit from the inclusion of microbiota predictors because these were compared to cheaper predictors that proved to be more or as effective” (Le Graverand et al., 2023).

The advantage of early integration by concatenation is its simplicity. But concatenating fixed effects, covariates and microbiota data, might be flawed. Singh et al. (2019) observed that training models on concatenated data may favor the most predictive variable blocks, and overlook other integrated blocks. Moreover, concatenating data increases the data

dimensionality. A higher number of variables may increase the noise and exacerbate the dimensionality curse, especially when there are few samples (Picard et al., 2021). Thus, if the goal is to integrate many blocks, other integration strategies could be more appropriate.

**Integration of fixed effects, covariates, pedigree and omics collected during different years**

Fixed effects, covariates, pedigree and omics were integrated via a late integration (Chapter 6). First, one MINT-sPLSR submodel was trained per block of predictors. Second, a metamodel was trained to get the final prediction estimates: either a weighted mean or a MINT-sPLSR. MINT-sPLSRs are convenient, as they can account for the experiment design (the three different years) and handle correlated variables. Indeed, estimates from different blocks may be correlated: predictions from genotypes and pedigree relatedness coefficients are expected to be close, for example.

Then, the thesis integration strategy voluntarily differed from the block sPLSR implemented in mixOmics (Singh et al., 2019). In mixOmics, estimates are either averaged or weight averaged. Block weights are defined by the correlation between latent variables and the outcome to predict. However, mixOmics computes the correlation with latent variables defined during the training. Thus, values of latent variables could be sensitive to overfitting. Thus, we postulated that the block sPLSR implemented in mixOmics may give too much importance to overfitting blocks. That is why, a cross-validation design with training, validation and testing sets was developed in this thesis. Furthermore, cross-validation stratification per year, pen and RFI lines was introduced in the thesis. The weighted mean and MINT-sPLSR metamodels were trained on validation sets to give more importance to blocks which predict well data unseen during the training of submodels. The integration strategy and data partitioning are directly inspired by examples set in competitive machine learning (Töscher and Jahrer, 2009).

Models integrating different omics may predict traits slightly better. In Romane lambs, fixed effects, covariates, pedigree and omics predicted feed efficiency and intake better than models fitted on a single block of predictors. The most challenging trait to predict was REI, since fixed effects and body weight could predict ADEI and FCR well. Integration, showed that pedigree, genotypes, faecal NIRS data and plasma buckets were the most promising

predictors – *i.e.* with the highest average contributions. Compared to the best single-block model, multi-block models had slightly but significantly higher prediction accuracies: the average correlation increased by 0.03 units for $ADEI_C$ (from 0.85 to 0.88), by 0.04 for $FCR_C$ (0.48 to 0.52), by 0.05 for $REI_C$ (0.54 to 0.59). In Romane divergent lines, the prediction accuracy of $REI_C$ from pedigree or genomics was already high (correlations = 0.54). In a commercial population, lower prediction accuracies could be expected before and after integration. Hess et al. (2022) integrated genomics and metagenomics to predict sheep RFI with mixed models: compared to genomic predictions, integration improved accuracies by 0.03 units maximum (from 0.28 to 0.31) (Hess et al., 2022). Similarly, integrating genomics with microbiota or metabolomic data increased the prediction accuracy of sheep residual methane emissions by 0.07 units maximum (from 0.20 to 0.27) (Ross et al., 2020).

As Huang et al. (2017) suggested, having more omics is better. However, no breeding company might actually consider collecting more omics data to predict and select traits: the prediction accuracy improvement is too marginal. Collecting more omics could be useful to identify novel trait determinisms by pinpointing proxies. The thesis focused on the prediction accuracy but it would be interesting to explore the feed efficiency determinisms underlying the selected predictors. The explaining variables of MINT-sPLSR could be assessed thanks to their value importance in the projection (Tenenhaus, 1998). Variables having the highest values may highlight feed efficiency determinisms. As in any late integration, each block was analyzed separately. A random forest mixed graphical model could be used to identify associations between the best predictors of different blocks (Fellinghauer et al., 2013).

Late integration might not be the optimal approach when it comes to understand trait determinisms. Indeed, late integration may not evidence interactions between different omics (Picard et al., 2021). However, late integration is an easy solution to pre-process and model each set of omics separately.

## A.2 Possible improvements

The thesis focused on NP integration to rank best proxies of feed efficiency and intake. Several leads could be pursued to improve the integration of different blocks of predictors. Other models or cross-validation strategies might be used.

### Refining the blending strategy

The multi-block integration developed in the thesis consisted in building one MINT-sPLSR per block, before using a meta-model to aggregate block predictions. This strategy is called "blending" because the meta-model is trained on validation data (Nti et al., 2020). Another strategy exists: the "stacking" strategy where the metamodel is fitted on training or testing data (Singh et al., 2019; Nti et al., 2020). We preferred to implement the blending strategy, since we expected that meta-models fitted on training data may favor overfitting submodels. On the other hand, it is not possible to assess if meta-models overfit when they are trained on testing data. Thus, blending was preferred over stacking.

While a MINT-sPLSR was fitted on all blocks of predictors, a different submodel could be applied per omics type. This change would account for data heterogeneity between different blocks. During the thesis, different kinds of models were tried only with microbiota data. Feed intake was predicted from microbiota data with random forest, sparse partial least squares and support vector regressions. None of these three models performed significantly better than the other, when phenotypes were predicted (Chapter 4 Table 11). However, R packages selbal (Rivera-Pinto et al., 2018) and predomics (Prifti et al., 2020) were introduced to predict outcomes from compositional microbiota data. In both packages, models may define balances: ratios between microbial groups.

Fitting MINT-sPLSR on microbiota data did not pose any computational issue, compared to genotypes. During the thesis, MINT-sPLSR models did not converge well when genotypes had a low minor allele frequency: genotypes with frequencies below 0.20 were filtered out. However, rare variants could help predict phenotypes and diseases. Thus, other models could be more suited for genomics: mixed models, Bayesian regressions and other machine learning techniques – such as gradient boosting and random forest (González-Recio and Forni, 2011).

The thesis blending strategy could be adapted to fit several models per block. In competitive machine learning, Töscher and Jahrer (2009) predicted multimedia ratings by building linear and non-linear submodels on the same predictors. Authors observed that aggregating predictions of different models improved the overall prediction accuracy. Thus, we could try to combine multiple omics and multiple models per omics. Furthermore, Töscher and Jahrer

(2009) also used a training set to determine the submodel hyperparameters, and a validation set to define the metamodel hyperparameters. After selecting all hyperparameters, Töscher and Jahrer (2009) proposed to retrain submodels and metamodel on the concatenated training and testing sets.

Finally, only linear and additive models were used as metamodels during the thesis: a weighted mean or a MINT-sPLSR. Non-linear models accounting for interactions might help when predictions made from different blocks of predictors are collinear. For instance, predictions from inferred data (*i.e.* prokaryote functions) and the original data (*i.e.* prokaryote abundances) are likely correlated.

**Switching from late integration to another strategy**

Stacking is an example of late integration strategy. Sometimes stacking does not improve the prediction accuracy, compared to models fitted only on the best block of predictors. In growing pigs, stacking SNPs with genomic EBVs of feed intake, ADG and body composition traits did not improve the prediction of RFI (Mora et al., 2023). In dairy cows, stacking behavior, metabolomic and production traits also did not improve the prediction of RFI (Martin et al., 2021a). However, two early or mixed integrations performed better with behavior, metabolomic and production traits. Indeed, the prediction accuracy of RFI improved slightly when an artificial neural network was used (early or mixed integration): $R^2$ increased from 0.01 to 0.08 (Martin et al., 2021a). Accuracy also increased when a multiple linear regression was carried out (early integration): from 0.02 to 0.13 (Martin et al., 2021a). Thus, late integration might not always be the best integration to optimize the prediction accuracy.

Early and mixed integration can be carried out with various models: with deep learning, tree-based, kernel and graph models for example (Picard et al., 2021). Such models may account for interactions between omics. However, it would considerably increase the complexity of models, which may be problematic with few samples. No study has ever attempted to fit models accounting for interactions, between as many omics as the thesis examined. The thesis focused on late integration because 250 phenotyped animals would not be enough to account for interactions between all kinds of omics.

# B. Perspectives in animal breeding

Recently, studies assessed whether omics could help select animals or plants. The inclusion of omics provides several opportunities, it also requires additional considerations to collect data.

## B.1 Why include omics data?

### Using omics as phenotype proxies

Omics were already proposed to phenotype larger populations for traits difficult and expensive to record. In dairy cows, milk mid-infrared spectroscopy was proposed to predict milk composition or cheese-making traits (Grelet et al., 2021; Sanchez et al., 2022). In pigs, coefficients of digestibility could be estimated from near-infrared spectra (Labussière et al., 2019). As long as prediction accuracies are high enough, selecting predicted traits may be advantageous if larger populations can be selected. However, the thesis compared and ranked several omics based on their predictive ability. Alone, no omic block could predict REI accurately (0.28 ≤ correlations ≤ 0.54, Table 13). Thus, no model was proposed during the thesis to predict feed efficiency from omics. Similarly, no model was proposed to predict feed intake from omics: fixed effects and production traits were better predictors. However, if feed intake is predicted from production traits, predicted intake cannot be used to estimate feed efficiency.

Even if we cannot use omics to completely replace feed efficiency and intake measurements, the inclusion of omics could present other advantages.

### Increasing the genomic prediction accuracy

In plants and animals, combining genotypes and other omics was proposed to increase the genomic prediction accuracy. In wheat, multi-trait genomic predictions were carried out with bivariate genomic models: the two explained variables were the measured trait and the predicted trait from either NIR or NMR spectroscopy (Hayes et al., 2017). Multi-trait predictions were most often more accurate than single-trait predictions. Higher prediction accuracies could help select faster. The genetic correlation between the measured trait and the predicted trait can help increase the prediction accuracy. After genomics, the best omics to predict REI were the fecal NIR spectra and plasma metabolomic buckets, in Romane

divergent lines. Thus, fecal phenomics and plasma metabolomics might be used to try to increase the genomic prediction accuracy. Phenotypic correlations were rather low between real REI phenotypes and predictions from fecal NIR spectra (correlation = 0.39) and plasma metabolomic buckets (correlation = 0.37) (Chapter 6 Table 13). It remains to check whether genetic correlations are high enough between real REI and predicted REI. If genetic correlations are high enough, the inclusion of predictions from omics may improve the genomic prediction accuracy in bivariate models.

**Coping with missing phenotypes**

Including omics may help when some phenotypes are missing, when they are not recorded in a fraction of the population. Christensen et al. (2021) developed models to perform a genetic evaluation and handle incomplete input records: records of phenotypes, genotypes and/or omics. With simulated data, authors observed that including omics improved the genetic evaluation accuracy for candidates without recorded phenotypes. Christensen et al. (2021) hypothesized that the improvement is possible thanks to the correlation between the omics used and the evaluated trait. Again, thesis results advocate for testing the inclusion of fecal NIR spectra or plasma buckets, when REI is evaluated in lambs fed a concentrate diet. However, it remains to check if the improvement is worth the cost of collecting phenomics or metabolomics.

**Studying the hologenome**

Including omics can also help dissect a trait, by studying the holobiont: the trait expression is the result of host and microbiota determinisms. The goal was to study the holobiont microorganisms, metabolites and fatty acids. The more the microbiota influences the trait, the more important it is to have insights into the holobiont. When it comes to the methane production of dairy cows, holobiont effects (0.38 ≤ proportion of phenotypic variance ≤ 0.59) accounted for more variance than host additive genetic effects (0.15 to 0.22) and rumen microbiota (0.15 to 0.31) (Saborío-Montero et al., 2021). Variance proportions varied with the linear model formula and the relatedness matrix calculation. More research is needed to choose how the relatedness matrix can be computed from different omics.

Authors advocated for selection indices including methane and microbiome traits to reduce the environmental footprint (Gonzalez-Recio et al., 2023). The microbiota composition

could be partially inherited vertically, thanks to genetics and physical contacts between parents and offspring (David et al., 2019). In pigs, selection for microbiota enterotypes during three generations increased the prevalence of the most abundant genera per enterotype (Larzul et al., 2023). It proved that the pig fecal microbiota could be partially inherited and partially selected. Finally, high microbiabilities do not prove that the inherited microbiota can explain a large part of the phenotypic variance in traits.

However, when the RFI of sheep is predicted, combining genomics and rumen metagenomics did not always perform better than genomic models (Hess et al., 2022). Authors showed that holobiont effects predicted RFI better (correlations up to 0.49 ± 0.06) than genotypes alone (0.29 ±0.08) only in lambs fed with alfalfa. However, there was no significant improvement in lambs and adults fed grass (Hess et al., 2022). In pigs, RFI variance was assessed to decompose the genetic variance into two components: the direct effect of host additive genetics over RFI and, the indirect effect via the inherited fecal microbiota affecting RFI (Weishaar et al., 2020). Authors found a moderate ratio between the microbiota additive variance and the host additive variance: 0.31. In pigs, it evidenced that the fecal microbiota can be partially inherited and, that it can also determine feed efficiency.

Holobiont effects over residual feed efficiency traits likely depend on which gut microbiota is studied, under which diet and physiological stage. In sheep fed concentrates, the rumen microbiota very poorly predicted REI (Chapter 5 and Chapter 6). Thus, we do not advocate for rumen metabarcoding. As suggested above, the faecal microbiota could bring new insights into ruminants' feed efficiency. Moreover, faeces sampling could be carried out on a larger scale than rumen sampling. Large-scale sampling and phenotyping are important to constitute reference populations to enable genomic selection programs.

## B.2 The collection of omics to predict feed efficiency

The goal of the thesis was to identify feed efficiency predictors to ease the genetic selection of feed efficiency. Thus, the husbandry protocol was close to protocols implemented in the national breeding program: each year 75-100 growing lambs were raised and phenotyped together for feed efficiency, under a concentrate diet.

The previous section concluded that large-scale sampling and phenotyping are essential in animal breeding. Genotypes, growth, weight, body conformation and composition phenotypes are already collected routinely in breeding companies. On the other hand, feed efficiency phenotypes and many omics are rarely collected. Thus, the discussion will focus on feed efficiency and omics data. The data collection strategy must be well-thought to get exploitable results. The thesis and the literature highlighted several points to be taken into consideration: how data are collected and, which animals are phenotyped.

**How to sample and phenotype**

Where? – This might be the first question which comes to mind when omics data are collected, except for genomics. Invasive sampling is not suited for animal breeding, especially if it requires sacrificing animals… Even if Zhang et al. (2019b) found associations between sheep efficiency and liver RNAs, it would be difficult to profile the liver transcriptome in large animal cohorts. Similarly, rumen sampling is challenging: even if tubing is less invasive than cannulating (da Cunha et al., 2023), both sampling methods would not be convenient in breeding companies. Tubing and cannulating may not remain authorized by European ethical regulation too. Furthermore, the rumen metabolome and microbiota vary between the different rumen fractions: solid, epithelial, and liquid (Su et al., 2022; da Cunha et al., 2023). When animals are tubed, it may be difficult to ensure how solid and liquid fractions are sampled. Anyway, rumen omics poorly predicted REI in Romane lambs (0.18 ≤ correlations ≤ 0.27, in Chapter 6 Table 13). Sampling other gut locations might be more encouraging. "Monteiro et al. (2022) suggested that the dairy cow faecal microbiota was less correlated with feed intake than the rumen microbiota and that the faecal microbiome was more correlated with production efficiency. They hypothesized that more efficient animals could be associated with a different fermentation profile in the rumen. Then, the host uptake of nutrients might lead to differences in nutrient availability for microorganisms in the lower gut (Monteiro et al., 2022). In addition to the rumen, it might be the accumulation throughout the whole digestive tract of differences in digestion or assimilation that influences the host efficiency." (Le Graverand et al., 2023). Faecal sampling would be much less invasive than rumen sampling. Similarly, blood sampling is less invasive, plus it is already used for genotyping. Blood samples could also be used to quantify feed proxies by carrying out NMR or NIR spectroscopy (Meale et al., 2017; Touitou et al., 2022).

When? – Timing may also influence how well feed efficiency can be predicted from omics. First, feed efficiency determinisms likely vary between different ages and physiological stages, as discussed in section II B.6. Second, the predictive ability of omics may vary depending on when samples are collected, relatively to when phenotypes are recorded. Ross et al. (2020) predicted more accurately sheep methane emissions when metabolomic data were collected during the same trial, compared to trials carried out 7- 8 months sooner or later. Then, Maltecca et al. (2019) predicted more accurately pig growth and carcass traits when microbiota faecal samples were collected in the middle of the growth trial, rather than the beginning or end. Thus, sampling time may affect the microbiota or metabolome composition over large time spans. Sampling time could also affect omics profiles over short time spans: sampling time may affect how long the animal fast since its last meal. When dairy Lacaune sheep were sampled over a day, the sampling hour significantly affected more than 10% of the rumen bacterial abundances (Martinez Boggio et al., 2021). When Romane lambs were all sampled in the morning, the hour and sampling order effects were negligible over the rumen microbiota (Chapter 5). Nonetheless, microbiota variations over time might be worth considering, especially if traits have a diurnal pattern. For instance, greenhouse gas emissions varied according to sheep grazing activities: emissions reached a peak around sunset (Lockyer, 1997).

How often? – No or few papers looked into how often omics data can be collected to predict feed efficiency. During the thesis, no biological replicate was used to collect omics data. Collecting omics data once downsizes costs and allows to phenotype more animals. However, sampling only once could limit the prediction accuracy of feed efficiency which is estimated over 6 weeks. Sampling more often might help cope with the low repeatability of some omics data. Bacteria abundances and fatty acids concentrations had low repeatabilities (medians ranged from 0.15-0.44), in Lacaune dairy ewes sampled two weeks apart (Fresco et al., 2022). Similarly, the median of 6 blood metabolites repeatabilities was equal to 0.27, in heifers sampled during the growing and finishing periods (Kelly et al., 2010). Low repeatabilities may be explained by biological, random and technical factors.

Technical replicates consist in analyzing several times the same sample, while biological replicates correspond to several samples. Technical replicates would be the easiest solution, as they require less animal handling and potentially less invasive procedures. In RNA-seq,

Liu et al. (2014) showed that focusing the efforts on the number of biological replicates may be more interesting than privileging technical settings: increasing the sequencing depth improved the power of differential analysis until a limit was reached, while increasing the number of biological replicates kept increasing the power. Thus, biological replicates should not be dismissed. Furthermore, if biological samples were taken at different times, longitudinal analysis might help understand relationships between the phenotype and omics data (Maltecca et al., 2019). It could be particularly interesting for traits recorded over a large period of times, such as feed efficiency. Martin et al. (2021b) advocated for longitudinal analysis of feed efficiency, since it is estimated from several traits which vary dynamically across time.

**Who to sample and phenotype**

Partitioning datasets into training (or reference) and testing (target) populations is essential when traits are predicted. In dairy cows, the correlation between cows' dry matter intake and predictions from rumen metagenomics was equal to 0.39, when the training and testing populations were raised in different countries (Delgado et al., 2019). During the thesis, feed intake was predicted from microbiota data by splitting data in two different ways. In Romane lambs, the correlation between feed intake and predictions from 16S data varied from -0.12 to 0.35, when training and testing populations were raised during different years (Chapter 4, Table 11). Average correlations between energy intake and 16S predictions ranged between 0.38 and 0.52, when training and testing populations included contemporaneous lambs (Chapter 5 Figure 26 and Chapter 6 Table 13). Thus, connections between training and testing populations may affect the predictive ability of omics data.

Having contemporaneous animals might enable models to learn which omic profile is associated to which environment: omics such as metataxonomics can sign for environmental effects (He et al., 2022b, Chapter 5). The microbiota composition is susceptible to environmental parameters such as the climate, the country location, the animal pen, the diet and the farming system (Thompson and Holmes, 2009; Henderson et al., 2015; Belanche et al., 2019; Marie-Etancelin et al., 2021). When predicted traits and omics are all associated to environmental variations, it seems ideal to mix contemporaneous animals in training and testing sets. However, it requires more resources to phenotype animals in each environment.

Nonetheless, continuous phenotyping is already advised for genomic selection. The association between markers and quantitative trait loci decays over generations (Blasco and Toro, 2014). However, it remains unclear how many feed efficiency phenotypes should be collected to constitute a proper training set. For instance, the correlation between EBVs of predicted $ADFI_C$ and EBVs of actual intake varied with the number of animals having real records, predicted records and registered in the pedigree (Chapter 4 Table 12). Research is needed to optimize the ratio between measured and predicted phenotypes, in terms of genetic evaluation accuracy.

**Take-home messages about the perspectives:**

- P-integration may be useful to account for batch effects and identify proxies generalizable across the different batches.
- N-integration of fixed effects, body weight and microbiota data in sheep, illustrated that microbiota data did not bring new information to predict feed efficiency or feed intake. Microbiota proxies were associated to spatiotemporal variations (across years, period and pens). However, fixed effects could help predict the environmental variations of feed efficiency and intake at a lower cost.
- NP-integration showed that integrating systematic effects, pedigree, as well as rumen, plasma and faecal omics improved significantly the prediction accuracy of feed efficiency and intake. However, improvements were marginal. Thus, multi-omic integration might not be interesting in breeding companies, but it may help understand trait variations better.
- During the thesis, the NP-integration was customized to blend predictions from submodels built on each block of predictors. A MINT-sPLSR submodel was fitted on all blocks. However, the submodel could be chosen per block, to better account for omics heterogeneity. A different integration strategy could be tried to model inter-omics interactions. However, the number of samples could hinder the use of complex models accounting for interactions.
- In the future, omics such as plasma metabolomics and faecal NIRS could potentially help increase the prediction accuracy of genomic models and, cope with missing phenotypes. Omics could also help understand the molecular dialogue between the host and its microbiota, which partially shapes feed efficiency and intake phenotypes.

# Conclusion

Improving feed efficiency is one key towards a more sustainable sheep industry. Feed efficiency has many known and unknown determinisms. Thus, the thesis assessed and ranked different potential proxies of feed efficiency and intake. The best proxies of feed intake were fixed effects and animal production traits. However, predicting feed intake from such proxies would not allow to infer feed efficiency from the predicted intake. The best examined proxies of residual feed intake were the host genetics and genomics (correlations between real and predicted phenotypes = 0.54). However, the thesis relied on two Romane lines genetically selected for feed efficiency. Thus, genomic and genetic prediction accuracies were likely exaggerated in divergent lines. Then, the next best proxies of feed efficiency were the plasma metabolomics and faecal near-infrared spectra. Rumen fatty acids, metabolites and microbiota poorly predicted feed efficiency, despite the rumen importance for the ruminant's nutrition.

The fecal microbiota might be more convenient to sample and study feed efficiency. In addition, the literature showed that omics are not the only promising efficiency proxies: respiratory gases, heat production, and behavior could be used as predictors.

The thesis showed that integrating pedigree, systematic effects and multiple omics, significantly improved the prediction accuracy of feed efficiency phenotypes. However, the accuracy gain was marginal. Thus, integrating as many omics as the thesis (8 techniques) cannot be justified in breeding programs. New studies could focus on a few omics: such as plasma metabolomics and faecal NIRS, which were the most promising proxies after genomics. Studies could test if including plasma and faecal omics would improve the genomic evaluation of feed efficiency. Including these omics could increase the accuracy of genomic predictions, or help cope with missing phenotypes.

Multi-omic integration remains promising to study and understand the biological determinants of feed efficiency. Omics remain useful to disentangle holobiont determinisms: the molecular dialogue between the host and its microbiota partially determines feed efficiency.

# References

Abijaoudé JA, Morand-Fehr P, Tessier J, Schmidely P and Sauvant D 2000. Diet effect on the daily feeding behaviour, frequency and characteristics of meals in dairy goats. Livestock Production Science 64, 29–37.

Aebersold R and Mann M 2003. Mass spectrometry-based proteomics. Nature 422, 198–207.

Aggarwal R and Ranganathan P 2016. Common pitfalls in statistical analysis: The use of correlation techniques. Perspectives in Clinical Research 7, 187.

Aguiar-Pulido V, Huang W, Suarez-Ulloa V, Cickovski T, Mathee K and Narasimhan G 2016. Approaches for Microbiome Analysis. Evolutionary Bioinformatics 12, 5–16.

Ahlberg CM, Allwardt K, Broocks A, Bruno K, Taylor A, McPhillips L, Krehbiel CR, Calvo-Lorenzo M, Richards CJ, Place SE, DeSilva U, VanOverbeke DL, Mateescu RG, Kuehn LA, Weaber R, Bormann J and Rolf MM 2019. Characterization of water intake and water efficiency in beef cattle. Journal of Animal Science 97, 4770–4782.

Aitchison J 1982. The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B (Methodological) 44, 139–160.

Alberts B, Johnson A, Lewis J, Raff M, Roberts K and Walter P 2002. The Structure and Function of DNA. In Molecular Biology of the Cell. Garland Science, New York, NJ, USA.

Aliakbari A, Zemb O, Billon Y, Barilly C, Ahn I, Riquet J and Gilbert H 2021. Genetic relationships between feed efficiency and gut microbiome in pig lines selected for residual feed intake. Journal of Animal Breeding and Genetics, 491–507.

Alves SP, Santos-Silva J, Cabrita ARJ, Fonseca AJM and Bessa RJB 2013. Detailed Dimethylacetal and Fatty Acid Composition of Rumen Content from Lambs Fed Lucerne or Concentrate Supplemented with Soybean Oil. PLOS ONE 8, e58386.

Amadeu RR, Cellon C, Olmstead JW, Garcia AAF, Resende MFR and Muñoz PR 2016. AGHmatrix: R Package to Construct Relationship Matrices for Autotetraploid and Diploid Species: A Blueberry Example. The Plant Genome 9, 1–10.

Anderson CJ, Koester LR and Schmitz-Esser S 2021. Rumen Epithelial Communities Share a Core Bacterial Microbiota: A Meta-Analysis of 16S rRNA Gene Illumina MiSeq Sequencing Datasets. Frontiers in Microbiology 12, 1–14.

Andueza D, Picard F, Dozias D and Aufrère J 2017. Fecal Near-Infrared Reflectance Spectroscopy Prediction of the Feed Value of Temperate Forages for Ruminants and Some Parameters of the Chemical Composition of Feces : Efficiency of Four Calibration Strategies. Applied Spectroscopy 71, 2164–2176.

Arce-Recinos C, Ramos-Juárez JA, Hernández-Cázares AS, Crosby-Galván MM, Alarcón-Zúñiga B, Miranda-Romero LA, Zaldívar-Cruz JM, Vargas-Villamil L, Aranda-Ibáñez EM, Vargas-Bello-Pérez E and Chay-Canul AJ 2022. Interplay between feed efficiency indices, performance, rumen fermentation parameters, carcass characteristics and meat quality in Pelibuey lambs. Meat Science 183, 108670.

Armstrong DG, Blaxter KL, Graham NM and Wainman FW 1959. The effect of environmental conditions on food utilisation by sheep. Animal Production 1, 1–12.

Artegoitia VM, Foote AP, Lewis RM and Freetly HC 2017. Rumen fluid metabolomics analysis associated with feed efficiency on crossbred steers. Scientific Reports 7, 1–14.

Arthur PF, Archer JA, Herd RM and Melville GJ 2001. Response to selection for net feed intake in beef cattle. In Proceedings of the Association for the Advancement of Animal Breeding and

Genetics, pp. 135–138.

Barea R, Dubois S, Gilbert H, Sellier P, van Milgen J and Noblet J 2010. Energy utilization in pigs selected for high and low residual feed intake. Journal of Animal Science 88, 2062–2072.

Barnes RJ, Dhanoa MS and Lister SJ 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Applied spectroscopy 43, 772–777.

Basarab JA, Beauchemin KA, Baron VS, Ominski KH, Guan LL, Miller SP and Crowley JJ 2013. Reducing GHG emissions through genetic improvement for feed efficiency : effects on economically important traits and enteric methane production. Animal 7, 303–315.

Basarab JA, Colazo MG, Ambrose DJ, Novak S, McCartney D and Baron VS 2011. Residual feed intake adjusted for backfat thickness and feeding frequency is independent of fertility in beef heifers. Canadian Journal of Animal Science 91, 573–584.

Baumont R, Dulphy JP, Sauvant D, Meschy F, Aufrere J and Peyraud J-L 2007. Valeur alimentaire des fourrages et des matières premières: tables et prévision. In Alimentation des ruminants. Besoins des animaux et valeur des aliments (ed. Editions Quae). Editions Quae, Versailles, France.

Belanche A, Balcells J, De La Fuente G, Yañez-Ruíz DR, Fondevila M and Calleja L 2010. Description of development of rumen ecosystem by PCR assay in milk-fed, weaned and finished lambs in an intensive fattening system. Journal of Animal Physiology and Animal Nutrition 94, 648–658.

Belanche A, Kingston-Smith AH, Griffith GW and Newbold CJ 2019. A multi-kingdom study reveals the plasticity of the rumen microbiota in response to a shift from non-grazing to grazing diets in sheep. Frontiers in Microbiology 10, 122.

Benjamini Y and Hochberg Y 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) 57, 289–300.

Benoit M, Sabatier R, Lasseur J, Creighton P and Dumont B 2019. Optimising economic and environmental performances of sheep-meat farms does not fully fit with the meat industry demands. Agronomy for Sustainable Development 39.

Bergman EN 1990. Energy contributions of volatile fatty acids from the gastrointestinal tract in various species. Physiological reviews 70, 567–590.

Berry DP and Crowley JJ 2012. Residual intake and body weight gain : A new measure of efficiency in growing cattle. Journal of Animal Science 90, 109–115.

Berry DP and Crowley JJ 2013. CELL BIOLOGY SYMPOSIUM : Genetics of feed efficiency in dairy and beef cattle. Journal of Animal Science 91, 1594–1613.

Blasco A and Toro MA 2014. A short critical history of the application of genomics to animal breeding. Livestock Science 166, 4–9.

Boadi D, Benchaar C, Chiquette J and Massé D 2004. Mitigation strategies to reduce enteric methane emissions from dairy cows : Update review. Canadian Journal of Animal Science 84, 319–335.

van den Boogaart KG and Tolosana-Delgado R 2013. Analyzing compositional data with R. Springer Berlin, Heidelberg.

Bouckaert RR and Frank E 2004. Evaluating the replicability of significance tests for comparing learning algorithms. In Advances in Knowledge Discovery and Data Mining (eds. D. Honghua, S. Ramakrishnan and Z. Chengqi), pp. 3–12. Springer, Sydney, Australia.

Breitwieser FP, Lu J and Salzberg SL 2019. A review of methods and databases for metagenomic classification and assembly. Briefings in Bioinformatics 20, 1125–1139.

Brody S 1945. Bioenergetics and growth, with special reference to the efficiency complex in

domestic animals. Reinhold Publishing Corp.

Brouard C, Mariette J, Flamary R and Vialaneix N 2022. Feature selection for kernel methods in systems biology. NAR Genomics and Bioinformatics 4, 1–17.

Brunes LC, Baldi F, Lopes FB, Narciso MG, Lobo RB, Espigolan R, Costa MFO and Magnabosco CU 2021. Genomic prediction ability for feed efficiency traits using different models and pseudo-phenotypes under several validation strategies in Nelore cattle. Animal 15.

Butler-Hogg B V and Tulloh NM 1982. Growth patterns in sheep: the effects of weight losses on compensatory growth and feed intake in Corriedale sheep. The Journal of Agricultural Science 99, 641–649.

Byerly TC 1941. Feed and other costs of producing market eggs.

Callahan BJ, Wong J, Heiner C, Oh S, Theriot CM, Gulati AS, McGill SK and Dougherty MK 2019. High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. Nucleic acids research 47, e103.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K and Madden TL 2009. BLAST+: Architecture and applications. BMC Bioinformatics 10, 1–9.

Cantalapiedra-Hijar G, Abo-Ismail M, Carstens GE, Guan LL, Hegarty R, Kenny DA, McGee M, Plastow G, Relling A and Ortigues-Marty I 2018. Review: Biological determinants of between-animal variation in feed efficiency of growing beef cattle. Animal 12, s321–s335.

Carillier-Jacquin C, Deru V, Tusell L, Bouquet A, Jacquin L and Gilbert H 2022. 403. Predicting pig digestibility coefficients with microbial and genomic data using machine learning prediction algorithms. In Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP), pp. 403–1680. Wageningen Academic Publishers.

Christensen OF, Börner V, Varona L and Legarra A 2021. Genetic evaluation including intermediate omics features. Genetics 219.

Claffey NA, Fahey AG, Gkarane V, Moloney AP, Monahan FJ and Diskin MG 2018. Effect of forage to concentrate ratio and duration of feeding on growth and feed conversion efficiency of male lambs. Translational Animal Science 2, 419–427.

Clemmons BA, Powers JB, Campagna SR, Seay TB, Embree MM and Myer PR 2020. Rumen fluid metabolomics of beef steers differing in feed efficiency. Metabolomics 16, 1–9.

Clemmons BA, Shin SB, Smith TPL, Embree MM, Voy BH, Schneider LG, Donohoe DR, McLean KJ and Myer PR 2021. Ruminal protozoal populations of angus steers differing in feed efficiency. Animals 11, 1–10.

Cockrum RR, Pickering NK, Anderson RM, Hyndman DL, Bixley MJ, Dodds KG, Stobart RH, McEwan JC and Cammack KM 2012. Identification of single nucleotide polymorphisms associated with feed efficiency in rams. In Proceedings of the Western Section of the American Society of Animal Science, pp. 79–83.

Conington J, Lambe N, Tortereau F, McGovern F, Navajas E, De Barbieri I, Ciappesoni G, Jakobsen J, Smith E, Yates J, Le Graverand Q, McDermott K, Steinheim G, Aspeholen Aby B, D?nnem I, McHugh N, Farrell L, Marie-Etancelin C, Johnson P and Rowe S 2022. Strategies to mitigate greenhouse gas emissions from pasturebased sheep systems ? an EU project consortium view. In Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP), pp. 107–110. Wageningen Academic Publishers, Wageningen.

Cooper SDB, Kyrizakis I and Nolan J V. 1995. Diet selection in sheep: The role of the rumen environment in the selection of a diet from two feeds that differ in their energy density. British Journal of Nutrition 74, 39–54.

Crowley JJ, McGee M, Kenny DA, Crews DH, Evans RD and Berry DP 2010. Phenotypic and genetic parameters for different measures of feed efficiency in different breeds of Irish performance-tested beef bulls. Journal of Animal Science 88, 885–894.

da Cunha LL, Monteiro HF, Figueiredo CC, Canisso IF, Bicalho RC, Cardoso FC, Weimer BC and Lima FS 2023. Characterization of rumen microbiome and metabolome from oro-esophageal tubing and rumen cannula in Holstein dairy cows. Scientific Reports 13, 1–14.

Darbot V, Samb M, Bernard M, Rué O and Pascal G 2022. FROGSFUNC: Smart integration of PICRUSt2 software into FROGS pipeline. In JOBIM. Rennes, France.

David I, Canario L, Combes S and Demars J 2019. Intergenerational transmission of characters through genetics, epigenetics, microbiota, and learning in livestock. Frontiers in Genetics 10, 1–14.

Davis SR, Barry TN and Hughson GA 1981. Protein synthesis in tissues of growing lambs. British Journal of Nutrition 46, 409–419.

Delgado B, Bach A, Guasch I, González C, Elcoso G, Pryce JE and Gonzalez-Recio O 2019. Whole rumen metagenome sequencing allows classifying and predicting feed efficiency and intake levels in cattle. Scientific Reports 9, 1–13.

Dettmer K, Aronov PA and Hammock BD 2006. Mass spectrometry-based metabolomics. Mass Spectrometry Reviews 26, 51–78.

Difford GF, Løvendahl P, Veerkamp RF, Bovenhuis H, Visker MHPW, Lassen J and de Haas Y 2020. Can greenhouse gases in breath be used to genetically improve feed efficiency of dairy cows? Journal of Dairy Science 103, 2442–2459.

Difford GF, Plichta DR, Løvendahl P, Lassen J, Noel SJ, Højberg O, Wright A-DG, Zhu Z, Kristensen L, Nielsen HB, Guldbrandtsen B and Sahana G 2018. Host genetics and the rumen microbiome jointly associate with methane emissions in dairy cows. PLOS Genetics 14, e1007580.

Domingos P 2012. A Few Useful Things to Know about Machine Learning. Commun. ACM 55, 78–87.

Domingues Millen D, De Beni Arrigoni M and Dias lauritano Pacheco R 2016. Rumenology ( PDFDrive ).

Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, Huttenhower C and Langille MGI 2020. PICRUSt2 for prediction of metagenome functions. Nature Biotechnology 38, 685–688.

Douhard F, Douhard M, Gilbert H, Monget P, Gaillard J-M and Lemaître J-F 2021. How much energetic trade-offs limit selection? Insights from livestock and related laboratory model species. Evolutionary Applications 14, 2726–2749.

Douhard F, Rupp R and Gilbert H 2022a. 54 . Feed efficiency and resource allocation trade-offs : theory , evidence and prospects. 264–267.

Douhard F, Rupp R and Gilbert H 2022b. Feed efficiency and resource allocation trade-offs: theory, evidence and prospects. In Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP), pp. 264–267. Wageningen Academic Publishers, Rotterdam.

Drouilhet L, Achard CS, Zemb O, Molette C, Gidenne T, Larzul C, Ruesche J, Tircazes A, Segura M, Bouchez T, Theau-Clément M, Joly T, Balmisse E, Garreau H and Gilbert H 2016. Direct and correlated responses to selection in two lines of rabbits selected for feed efficiency under ad libitum and restricted feeding: I. Production traits and gut microbiota characteristics1. Journal of Animal Science 94, 38–48.

Duncan AJ and Young SA 2002. Can goats learn about foods through conditioned food aversions and preferences when multiple food options are simultaneously available? Journal of Animal Science 80, 2091–2098.

Dunn OJ 1961. Multiple Comparisons among Means. Journal of the American Statistical Association

56, 52–64.

Dunn OJ and Clark V 1969. Correlation Coefficients Measured on the Same Individuals Author ( s ): Olive Jean Dunn and Virginia Clark Source : Journal of the American Statistical Association , Vol . 64 , No . 325 ( Mar ., 1969 ), pp . 366- Published by : American Statistical Associ. Jasa 64, 366–377.

Durunna ON, Plastow G, Mujibi FDN, Grant J, Mah J, Basarab JA, Okine EK, Moore SS and Wang Z 2011. Genetic parameters and genotype × environment interaction for feed efficiency traits in steers fed grower and finisher diets 1. Journal of Animal Science 89, 3394–3400.

Ellison MJ, Cockrum RR, Means WJ, Meyer AM, Ritten J, Austin KJ and Cammack M 2022. Effects of feed efficiency and diet on performance and carcass characteristics in growing wether lambs. Small Ruminant Research 207, 106611.

Ellison MJ, Conant GC, Lamberson WR, Austin KJ, Van Kirk E, Cunningham HC, Rule DC and Cammack KM 2019. Predicting residual feed intake status using rumen microbial profiles in ewe lambs. Journal of Animal Science 97, 2878–2888.

Ellison MJ, Conant GC, Lamberson WR, Cockrum RR, Austin KJ, Rule DC and Cammack KM 2017. Diet and feed efficiency status affect rumen microbial profiles of sheep. Small Ruminant Research 156, 12–19.

Escudié F, Auer L, Bernard M, Mariadassou M, Cauquil L, Vidal K, Maman S, Hernandez-Raquet G, Combes S and Pascal G 2018. FROGS: Find, Rapidly, OTUs with Galaxy Solution. Bioinformatics 34, 1287–1294.

FAO 2017. Global Livestock Environmental Assessment Model (GLEAM) [online]. Retrieved on 8 March 2023, from *www.fao.org/gleam/en/*.

Fellinghauer B, Bühlmann P, Ryffel M, Von Rhein M and Reinhardt JD 2013. Stable graphical model estimation with Random Forests for discrete, continuous, and mixed variables. Computational Statistics and Data Analysis 64, 132–152.

Ferreira GF, Ciappesoni G, Castells D, Amarilho-Silveira F, Navajas EA, Giorello D, Banchero G and De Barbieri I 2021. Feed conversion efficiency in sheep genetically selected for resistance to gastrointestinal nematodes. Animal Production Science 61, 754–760.

Fitzhugh HA and Taylor CS 1971. Genetic Analysis of Degree of Maturity. Journal of Animal Science 33, 717–725.

Fregulia P, Neves ALA, Dias RJP and Campos MM 2021. A review of rumen parameters in bovines with divergent feed efficiencies: What do these parameters tell us about improving animal productivity and sustainability? Livestock Science 254.

Fresco S, Marie-Etancelin C, Meynadier A and Martinez Boggio G 2022. Variation in Rumen Bacteria of Lacaune Dairy Ewes From One Week to the Next. Frontiers in Microbiology 13.

Fuks G, Elgart M, Amir A, Zeisel A, Turnbaugh PJ, Soen Y and Shental N 2018. Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling. Microbiome 6, 1–13.

Garreau H, Ruesche J, Gilbert H, Balmisse E, Benitez F, Richard F, David I, Drouilhet L and Zemb O 2019. Estimating direct genetic and maternal effects affecting rabbit growth and feed efficiency with a factorial design. Journal of Animal Breeding and Genetics 136, 168–173.

Garrett WN, Meyer JH and Lofgreen GP 1959. The Comparative Energy Requirements of Sheep and Cattle for Maintenance and Gain1. Journal of Animal Science 18, 528–547.

Gilbert H, Billon Y, Brossard L, Faure J, Gatellier P, Gondret F, Labussière E, Lebret B, Lefaucheur L, Floch N Le, Louveau I, Merlot E, Montagne L, Mormede P, Renaudeau D, Riquet J, Milgen J Van, Vincent A and Noblet J 2017. Review : divergent selection for residual feed intake in the growing pig. animal 11, 1427–1439.

Gill M, France J, Summers M, McBride BW and Milligan LP 1989. Simulation of the Energy Costs Associated with Protein Turnover and Na+,K+-Transport in Growing Lambs. The Journal of Nutrition 119, 1287–1299.

Giráldez FJ, Santos N, Santos A, Valdés C, López S and Andrés S 2021. Fattening lambs with divergent residual feed intakes and weight gains: Unravelling mechanisms driving feed efficiency. Animal Feed Science and Technology 273.

Gloor GB, Macklaim JM, Pawlowsky-Glahn V and Egozcue JJ 2017. Microbiome datasets are compositional: And this is not optional. Frontiers in Microbiology 8, 1–6.

Goddard ME, Kemper KE, Macleod IM, Chamberlain AJ and Hayes BJ 2016. Genetics of complex traits : prediction of phenotype , identification of causal polymorphisms and genetic architecture. Proceedings of the Royal Society B: Biological Sciences 283, 20160569.

Goldansaz SA, Markus S, Berjanskii M, Rout M, Guo AC, Wang Z, Plastow G and Wishart DS 2020. Candidate serum metabolite biomarkers of residual feed intake and carcass merit in sheep. Journal of Animal Science 98, 1–15.

Gomes R da C, Sainz RD and Leme PR 2013. Protein metabolism, feed energy partitioning, behavior patterns and plasma cortisol in Nellore steers with high and low residual feed intake. Revista Brasileira de Zootecnia 42, 44–50.

González-Recio O and Forni S 2011. Genome-wide prediction of discrete traits using bayesian regressions and machine learning. Genetics Selection Evolution 43, 1–12.

Gonzalez-Recio O, Scrobota N, López-Paredes J, Saborío-Montero A, Fernández A, López de Maturana E, Villanueva B, Goiri I, Atxaerandio R and García-Rodríguez A 2023. Review: Diving into the cow hologenome to reduce methane emissions and increase sustainability. Animal.

Gonzalez LA, Kyriazakis I and Tedeschi LO 2018. Review: Precision nutrition of ruminants: Approaches, challenges and potential gains. Animal 12, S246–S261.

Goopy JP, Donaldson A, Hegarty R, Vercoe PE, Haynes F, Barnett M and Oddy VH 2014. Low-methane yield sheep have smaller rumens and shorter rumen retention time. British Journal of Nutrition 111, 578–585.

GrassToGas 2019. Retrieved on 24 April 2023, from *https://www.eragas.eu/en/eragas/research-projects/grasstogas.htm*.

Le Graverand Q, Marie-Etancelin C, Meynadier A, Weisbecker J-L, Marcon D and Tortereau F 2023. Predicting feed efficiency traits in growing lambs from their ruminal microbiota. animal, 100824.

Le Graverand Q, Marie-Etancelin C, Weisbecker JL, Meynadier A, Marcon D and Tortereau F 2022. Using machine learning to predict feed intakes of meat sheep from animal traits and ruminal microbiota. In Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP), pp. 142–618. Wageningen Academic Publishers.

van der Greef J, Stroobant P and van der Heijden R 2004. The role of analytical sciences in medical systems biology. Current Opinion in Chemical Biology 8, 559–565.

Grelet C, Dardenne P, Soyeurt H, Fernandez JA, Vanlierde A, Stevens F, Gengler N and Dehareng F 2021. Large-scale phenotyping in dairy sector using milk MIR spectra: Key factors affecting the quality of predictions. Methods 186, 97–111.

Gunsett FC 1984. Linear Index Selection to Improve Traits Defined as Ratios. Journal of Animal Science 59, 1185–1193.

Haddad SG and Husein MQ 2004. Effect of dietary energy density on growth performance and slaughtering characteristics of fattening Awassi lambs. Livestock Production Science 87, 171–177.

Hadziavdic K, Lekang K, Lanzen A, Jonassen I, Thompson EM and Troedsson C 2014. Characterization

of the 18s rRNA gene for designing universal eukaryote specific primers. PLoS ONE 9, 2.

Hastie T, Tibshirani R, James G and Witten D 2021. An introduction to statistical learning (2nd ed.). Springer texts 102, 618.

Hayes BJ, Lewin HA and Goddard ME 2013. The future of livestock breeding : genomic selection for efficiency , reduced emissions intensity , and adaptation. Trends in Genetics 29, 206–214.

Hayes BJ, Panozzo J, Walker CK, Choy AL, Kant S, Wong D, Tibbits J, Daetwyler HD, Rochfort S, Hayden MJ and Spangenberg GC 2017. Accelerating wheat breeding for end-use quality with multi-trait genomic predictions incorporating near infrared and nuclear magnetic resonance-derived phenotypes. Theoretical and Applied Genetics 130, 2505–2519.

He Y, Maltecca C, Howard J, Huang Y, Gray K and Tiezzi F 2022a. 501. Comparing methods to summarize gut microbiota composition in estimating microbiability of host phenotypes in swine. In Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP), pp. 501–2081. Wageningen Academic Publishers.

He Y, Tiezzi F, Jiang J, Howard JT, Huang Y, Gray K, Choi JW and Maltecca C 2022b. Use of Host Feeding Behavior and Gut Microbiome Data in Estimating Variance Components and Predicting Growth and Body Composition Traits in Swine. Genes 13.

Henderson G, Cox F, Ganesh S, Jonker A, Young W, Janssen PH, Abecia L, Angarita E, Aravena P, Arenas GN, Ariza C, Attwood GT, Avila JM, Avila-Stagno J, Bannink A, Barahona R, Batistotti M, Bertelsen MF, Brown-Kav A, Carvajal AM, Cersosimo L, Chaves AV, Church J, Clipson N, Cobos-Peralta MA, Cookson AL, Cravero S, Carballo OC, Crosley K, Cruz G, Cucchi MC, De La Barra R, De Menezes AB, Detmann E, Dieho K, Dijkstra J, Dos Reis WLS, Dugan MER, Ebrahimi SH, Eythórsdóttir E, Fon FN, Fraga M, Franco F, Friedeman C, Fukuma N, Gagić D, Gangnat I, Grilli DJ, Guan LL, Miri VH, Hernandez-Sanabria E, Gomez AXI, Isah OA, Ishaq S, Jami E, Jelincic J, Kantanen J, Kelly WJ, Kim SH, Klieve A, Kobayashi Y, Koike S, Kopecny J, Kristensen TN, Krizsan SJ, LaChance H, Lachman M, Lamberson WR, Lambie S, Lassen J, Leahy SC, Lee SS, Leiber F, Lewis E, Lin B, Lira R, Lund P, Macipe E, Mamuad LL, Mantovani HC, Marcoppido GA, Márquez C, Martin C, Martinez G, Martinez ME, Mayorga OL, McAllister TA, McSweeney C, Mestre L, Minnee E, Mitsumori M, Mizrahi I, Molina I, Muenger A, Munoz C, Murovec B, Newbold J, Nsereko V, O'Donovan M, Okunade S, O'Neill B, Ospina S, Ouwerkerk D, Parra D, Pereira LGR, Pinares-Patino C, Pope PB, Poulsen M, Rodehutscord M, Rodriguez T, Saito K, Sales F, Sauer C, Shingfield K, Shoji N, Simunek J, Stojanović-Radić Z, Stres B, Sun X, Swartz J, Tan ZL, Tapio I, Taxis TM, Tomkins N, Ungerfeld E, Valizadeh R, Van Adrichem P, Van Hamme J, Van Hoven W, Waghorn G, Wallace RJ, Wang M, Waters SM, Keogh K, Witzig M, Wright ADG, Yamano H, Yan T, Yanez-Ruiz DR, Yeoman CJ, Zambrano R, Zeitz J, Zhou M, Zhou HW, Zou CX and Zunino P 2015. Rumen microbial community composition varies with diet and host, but a core microbiome is found across a wide geographical range. Scientific Reports 5, 14567.

Henderson G, Cox F, Kittelmann S, Miri VH, Zethof M, Noel SJ, Waghorn GC and Janssen PH 2013. Effect of DNA extraction methods and sampling techniques on the apparent structure of cow and sheep rumen microbial communities. PloS one 8, 1–14.

Herd RM and Arthur PF 2009. Physiological basis for residual feed intake. Journal of Animal Science 87, 64–71.

Herd RMA, Oddy VHA and Richardson ECB 2004. Biological basis for variation in residual feed intake in beef cattle . 1 . Review of potential mechanisms. Australian Journal of Experimental Agriculture 44, 423–430.

Herd RM, Velazco JI, Smith H, Arthur PF, Hine B, Oddy H, Dobos RC and Hegarty RS 2019. Genetic variation in residual feed intake is associated with body composition, behavior, rumen, heat production, hematology, and immune competence traits in Angus cattle. Journal of Animal Science 97, 2202–2219.

Hess MK, Rowe SJ, Van Stijn TC, Henry HM, Hickey SM, Brauning R, McCulloch AF, Hess AS, Kirk MR,

Kumar S, Pinares-Patiño C, Kittelmann S, Wood GR, Janssen PH and McEwan JC 2020. A restriction enzyme reduced representation sequencing approach for low-cost, high-throughput metagenome profiling. PLoS ONE 15, 1–18.

Hess M, Zetouni L, Hess A, Budel J, Dodds K, Henry H, Brauning R, McCulloch A, Hickey S, Johnson P, Elmes S, Wing J, Bryson B, Knowler K, Hyndman D, Baird H, McRae K, Jonker A, Janssen P and Rowe S 2022. Combining host and rumen metagenome profiling for selection in sheep: prediction of methane, feed efficiency, production, and health traits (Preprint). Research Square. Retrieved on 2 May 2023, from *https://www.researchsquare.com/article/rs-2290213/v1*.

Hoque MA, Arthur PF, Hiramoto K, Gilmour AR and Oikawa T 2007. Variance components due to direct genetic, maternal genetic and permanent environmental effect for growth and feed-efficiency traits in young male Japanese Black cattle. Journal of Animal Breeding and Genetics 124, 102–107.

Houle D, Govindaraju DR and Omholt S 2010. Phenomics: The next challenge. Nature Reviews Genetics 11, 855–866.

Huang S, Chaudhary K and Garmire LX 2017. More is better: Recent progress in multi-omics data integration methods. Frontiers in Genetics 8, 1–12.

Hyder I, Ravi Kanth Reddy P, Raju J, Manjari P, Srinivasa Prasad C, Aswani Kumar K and Sejian V 2017. Alteration in Rumen Functions and Diet Digestibility During Heat Stress in Sheep BT - Sheep Production Adapting to Climate Change. In (eds. V. Sejian, R. Bhatta, J. Gaughan, P.K. Malik, S.M.K. Naqvi and R. Lal), pp. 235–265. Springer Singapore, Singapore.

Jacques J, Berthiaume R and Cinq-Mars D 2011. Growth performance and carcass characteristics of Dorset lambs fed different concentrates: Forage ratios or fresh grass. Small Ruminant Research 95, 113–119.

Jiang Y, Xie M, Chen W, Talbot R, Maddox JF, Faraut T, Wu C, Muzny DM, Li Y, Zhang W, Stanton J-A, Brauning R, Barris WC, Hourlier T, Aken BL, Searle SMJ, Adelson DL, Bian C, Cam GR, Chen Y, Cheng S, DeSilva U, Dixen K, Dong Y, Fan G, Franklin IR, Fu S, Fuentes-Utrilla P, Guan R, Highland MA, Holder ME, Huang G, Ingham AB, Jhangiani SN, Kalra D, Kovar CL, Lee SL, Liu W, Liu X, Lu C, Lv T, Mathew T, McWilliam S, Menzies M, Pan S, Robelin D, Servin B, Townley D, Wang W, Wei B, White SN, Yang X, Ye C, Yue Y, Zeng P, Zhou Q, Hansen JB, Kristiansen K, Gibbs RA, Flicek P, Warkup CC, Jones HE, Oddy VH, Nicholas FW, McEwan JC, Kijas JW, Wang J, Worley KC, Archibald AL, Cockett N, Xu X, Wang W and Dalrymple BP 2014. The sheep genome illuminates biology of the rumen and lipid metabolism. Science 344, 1168–1173.

Johnson PL, Hickey S, Knowler K, Wing J, Bryson B, Hall M, Jonker A, Janssen PH, Dodds KG, Mcewan JC and Rowe SJ 2022. Genetic parameters for residual feed intake , methane emissions , and body composition in New Zealand maternal sheep. Frontiers in Genetics 13.

Johnson KA and Johnson DE 1995. Methane emissions from cattle. Journal of Animal Science 73, 2483–2492.

Jonas E, Martin GB, Celi P, Li L, Soattin M, Thomson PC and Raadsma HW 2016. Association of polymorphisms in leptin and leptin receptor genes with circulating leptin concentrations , production and efficiency traits in sheep. Small Ruminant Research 136, 78–86.

Jones AK, Jones DL and Cross P 2013. Informing decision making in agricultural greenhouse gas mitigation policy : A Best – Worst Scaling survey of expert and farmer opinion in the sheep industry. Environmental Science & Policy 29, 46–56.

Kamalzadeh A, Van Bruchem J, Koops WJ, Tamminga S and Zwart D 1997. Feed quality restriction and compensatory growth in growing sheep: Feed intake, digestion, nitrogen balance and modelling changes in feed efficiency. Livestock Production Science 52, 209–217.

Kamke J, Kittelmann S, Soni P, Li Y, Tavendale M, Ganesh S, Janssen PH, Shi W, Froula J, Rubin EM

and Attwood GT 2016. Rumen metagenome and metatranscriptome analyses of low methane yield sheep reveals a Sharpea-enriched microbiome characterised by lactic acid formation and utilisation. Microbiome 4, 1–16.

Karisa BK, Thomson J, Wang Z, Li C, Montanholi YR, Miller SP, Moore SS and Plastow GS 2014. Plasma metabolites associated with residual feed intake and other productivity performance traits in beef cattle. Livestock Science 165, 200–211.

Kellner O 1909. The Scientific Feeding of Animals. McMillan Company, New York.

Kelly AK, McGee M, Crews DH, Sweeney T, Boland TM and Kenny DA 2010. Repeatability of feed efficiency, carcass ultrasound, feeding behavior, and blood metabolic variables in finishing heifers divergently selected for residual feed intake. Journal of Animal Science 88, 3214–3225.

Kennedy BW, van der Werf JH, Meuwissen TH, Properties S and Intake RF 1993. Genetic and statistical properties of residual feed intake. Journal of animal science 71, 3239–3250.

Kenny DA, Fitzsimons C, Waters SM and Mcgee M 2018. Invited review : Improving feed ef fi ciency of beef cattle – the current state of the art and future challenges. animal 12, 1815–1826.

Khafipour E, Li S, Plaizier JC and Krause DO 2009. Rumen microbiome composition determined using two nutritional models of subacute ruminal acidosis. Applied and Environmental Microbiology 75, 7115–7124.

Khiaosa-ard R and Zebeli Q 2014. Cattle's variation in rumen ecology and metabolism and its contributions to feed efficiency. Livestock Science 162, 66–75.

Kittelmann S, Pinares-Patiño CS, Seedorf H, Kirk MR, Ganesh S, McEwan JC and Janssen PH 2014. Two different bacterial community types are linked with the low-methane emission trait in sheep. PLoS ONE 9, 1–9.

Kleiber M 1961. The fire of life. An introduction to animal energetics. John Wiley & Sons, Inc., New York.

Klieve A V and Bauchop TOM 1988. Morphological Diversity of Ruminal Bacteriophages from Sheep and Cattle. Applied and Environmental Microbiology 54, 1637–1641.

Klieve A V and Swain RA 1993. Estimation of Ruminal Bacteriophage Numbers by Pulsed- Field Gel Etectrophoresis and Laser Densitometry. Applied and Environmental Microbiology 59, 2299–2303.

Kneebone DG and Dryden GM 2014. Prediction of diet quality for sheep from faecal characteristics: comparison of near-infrared spectroscopy and conventional chemistry predictive models. Animal Production Science 55, 1–10.

Knott SA, Cummins LJ, Dunshea FR and Leury BJ 2008. The use of different models for the estimation of residual feed intake (RFI) as a measure of feed efficiency in meat sheep. Animal Feed Science and Technology 143, 242–255.

Koch RM, Swiger LA, Chambers D and Gregory KE 1963. Efficiency of feed use in beef cattle. Journal of animal science 22, 486–494.

Kolath WH, Kerley MS, Golden JW and Keisler DH 2006. The relationship between mitochondrial function and residual feed intake in Angus steers. Journal of Animal Science 84, 861–865.

Kruger Ben Shabat S, Sasson G, Doron-Faigenboim A, Durman T, Yaacoby S, Berg Miller ME, White BA, Shterzer N and Mizrahi I 2016. Specific microbiome-dependent mechanisms underlie the energy harvest efficiency of ruminants. ISME Journal 10, 2958–2972.

Kuhn M and Johnson K 2013. Applied predictive modeling. Springer, New-York, NY, USA.

De La Torre A, Andueza D, Renand G, Baumont R, Cantalapiedra-Hijar G and Nozière P 2019. Digestibility contributes to between-animal variation in feed efficiency in beef cows. Animal 13,

2821–2829.

Labussière E, Ganier P, Condé A, Janvier E and Van Milgen J 2019. Development of a NIRS method to assess the digestive ability in growing pigs. In Book of abstracts of the 70th Annual Meeting of the European Federation of Animal Science, p. 604. Wageningen Academic Publishers, Ghent, Belgium.

Larzul C, Estellé J, Borey M, Blanc F, Lemonnier G, Billon Y, Thiam M-G, Quinquis B, Galleron N and Jardet D 2023. Driving gut microbiota enterotypes through host genetics. Preprint.

Lawrence P, Kenny DA, Earley B and Mcgee M 2013. Intake of conserved and grazed grass and performance traits in beef suckler cows differing in phenotypic residual feed intake. 152, 154–166.

Lê Cao K-A and Welham ZM 2021. Multivariate Data Integration Using R: Methods and Applications with the mixOmics Package. Chapman and Hall/CRC, Boca Raton.

Lederberg BJ and McCray AT 2001. ' Ome Sweet ' Omics-- A Genealogical Treasury of Words. The Scientist 15, 8.

Lefort G, Liaubet L, Paris A, Canlet C, Tardivel P, Père M-C, Quesnel H, Iannuccelli N, Vialaneix N and Servien R 2019. Systems biology ASICS : an R package for a whole analysis workflow of 1D 1 H NMR spectra. Bioinformatics 35, 4356–4363.

Legarra A, Christensen OF, Aguilar I and Misztal I 2014. Single Step, a general approach for genomic selection. Livestock Science 166, 54–65.

Lenth R V. 2022. The Comprehensive R Archive Network. emmeans: Estimated Marginal Means, aka Least-Squares Means. Retrieved on 24 August 2022, from *https://cran.r-project.org/web/packages/emmeans/emmeans.pdf*.

Lettat A, Nozière P, Silberberg M, Morgavi DP, Berger C and Martin C 2012. Rumen microbial and fermentation characteristics are affected differently by bacterial probiotic supplementation during induced lactic and subacute acidosis in sheep. BMC Microbiology 12.

Li F and Guan LL 2017. Metatranscriptomic Profiling Reveals Linkages between the Active Rumen Microbiome and Feed Efficiency in Beef Cattle. Applied and Environmental Microbiology 83.

Li F, Hitch TCA, Chen Y, Creevey CJ and Guan LL 2019a. Comparative metagenomic and metatranscriptomic analyses reveal the breed effect on the rumen microbiome and its associations with feed efficiency in beef cattle. Microbiome 7.

Li F, Li C, Chen Y, Liu J, Zhang C, Irving B and Fitzsimmons C 2019b. Host genetics influence the rumen microbiota and heritable rumen microbial features associate with feed efficiency in cattle. Microbiome 7, 1–17.

Li R, Teng Z, Lang C, Zhou H, Zhong W, Ban Z, Yan X, Yang H, Farouk MH and Lou Y 2019c. Effect of different forage-to-concentrate ratios on ruminal bacterial structure and real-time methane production in sheep. PLOS ONE 14, e0214777.

Liaw A and Wiener M 2002. Classification and Regression by randomForest. R News 2, 18–22.

Lin C, Wang W, Zhang D, Huang K, Li X, Zhang Y, Zhao Y, Wang J, Zhou B, Cheng J, Xu D and Li W 2023. Polymorphisms in SHISA and RFC genes and their association with feed conversion ratio in Hu sheep. Frontiers in Veterinary Science 9.

Lin L, Xie F, Sun D, Liu J, Zhu W and Mao S 2019. Ruminal microbiome-host crosstalk stimulates the development of the ruminal epithelium in a lamb model. Microbiome 7, 1–16.

Lines DS, Pitchford WS, Bottema CDK, Herd RM and Oddy VH 2014. Selection for residual feed intake affects appetite and body composition rather than energetic efficiency. Animal Production Science 58, 175–184.

Liu Y, Zhou J and White KP 2014. RNA-seq differential expression studies: More sequence or more replication? Bioinformatics 30, 301–304.

Lobley GE 1990. Energy metabolism reactions in ruminant muscle: responses to age, nutrition and hormonal status. Reproduction, nutrition, development 30, 13–34.

Lockyer DR 1997. Methane emissions from grazing sheep and calves. Agriculture, Ecosystems and Environment 66, 11–18.

López-Catalina A, Bach A, Gutiérrez-Rivas M, Peiró-Pastor R and González-Recio O 2022. 544. An approach to study the association between the blood cell methylome with feed efficiency traits. In Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP), pp. 544–2257. Wageningen Academic Publishers, Rotterdam, Netherlands.

de los Campos G, Vazquez AI, Hsu S and Lello L 2018. Complex-Trait Prediction in the Era of Big Data. Trends in Genetics 34, 746–754.

Lovendahl P, Difford GF, Li B, Chagunda MGG, Huhtanen P, Lidauer MH, Lassen J and Lund P 2018. Review: Selecting for improved feed efficiency and reduced methane emissions in dairy cattle. Animal 12, S336–S349.

Lu D, Akanno EC, Crowley JJ, Schenkel F, Li H, De Pauw M, Moore SS, Wang Z, Li C, Stothard P, Plastow G, Miller SP and Basarab JA 2016. Accuracy of genomic predictions for feed efficiency traits of beef cattle using 50K and imputed HD genotypes. Journal of Animal Science 94, 1342–1353.

Lwin KO, Hayakawa M, Ban-Tokuda T and Matsui H 2011. Real-Time PCR Assays for Monitoring Anaerobic Fungal Biomass and Population Size in the Rumen. Current Microbiology 62, 1147–1151.

Macdonald KA, Pryce JE, Spelman RJ, Davis SR, Wales WJ, Waghorn GC, Williams YJ, Marett LC and Hayes BJ 2014. Holstein-Friesian calves selected for divergence in residual feed intake during growth exhibited significant but reduced residual feed intake divergence in their first lactation. Journal of Dairy Science 97, 1427–1435.

Maechler M, Rousseeuw P, Croux C, Todorov V, Ruckstuhl A, Salibian-Barrera M, Verbeke T, Koller M, Conceicao EL and di Palma MA 2017. The Comprehensive R Archive Network. Robustbase: basic robust statistics R package. Retrieved on 7 December 2021, from *https://cran.r-project.org/web/packages/robustbase/robustbase.pdf*.

Mahé F, Rognes T, Quince C, de Vargas C and Dunthorn M 2014. Swarm: Robust and fast clustering method for amplicon-based studies. PeerJ 2014, 1–13.

Maltecca C, Lu D, Schillebeeckx C, McNulty NP, Schwab C, Shull C and Tiezzi F 2019. Predicting Growth and Carcass Traits in Swine Using Microbiome Data and Machine Learning Algorithms. Scientific Reports 9, 1–15.

Manafiazar G, Basarab JA, McKeown L, Stewart-Smith J, Baron V, Macneil MD and Plastow G 2017. Optimizing feed intake recording and feed efficiency estimation to increase the rate of genetic gain for feed efficiency in beef cattle. Canadian Journal of Animal Science 97, 456–465.

Mani S, Aiyegoro OA and Adeleke MA 2022. Association between host genetics of sheep and the rumen microbial composition. Tropical Animal Health and Production 54, 1–16.

Marai IFM, El-Darawany AA, Fadiel A and Abdel-Hafez MAM 2007. Physiological traits as affected by heat stress in sheep-A review. Small Ruminant Research 71, 1–12.

Margulis L and Fester R 1991. Symbiosis as a source of evolutionary innovation: speciation and morphogenesis. MIT press.

Marie-Etancelin C, Francois D, Weisbecker J-L, Marcon D, Moreno-Romieux C, Bouvier F and Tortereau F 2019. Detailed genetic analysis of feeding behaviour in Romane lambs and links with residual feed intake. Journal of Animal Breeding and Genetics 136, 174–182.

Marie-Etancelin C, Tortereau F, Gabinaud B, Martinez Boggio G, Le Graverand Q, Marcon D, De Almeida ML, Pascal G, Weisbecker JL and Meynadier A 2021. Apart From the Diet, the Ruminal Microbiota of Lambs Is Modified in Relation to Their Genetic Potential for Feed Efficiency or Feeding Behavior. Frontiers in Microbiology 12, 1–15.

Marie-Etancelin C, Weisbecker J-L, Marcon D, Estivalet L, Le Graverand Q and Tortereau F 2023. Selecting for feed efficiency with concentrates in meat sheep also improved feed efficiency with dry forage, but what about feeding behavior? (in press). In 74th Annual Meeting of the European Federation for Animal Science. Wageningen Academic Publishers, Lyon, France.

Marques CB, De Barbieri I, Velazco J, Navajas EA and Ciappesoni G 2022. 28. Genetic parameters for feed efficiency, gas emissions, oxygen consumption and wool traits in Australian Merino. In Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP), pp. 28–160. Wageningen Academic Publishers.

Marquisseau A, Tortereau F, Marty-Gasset N, Marie-Etancelin C and Le Graverand Q 2023. Prediction of feed efficiency related traits from plasma NMR spectra. In 74th Annual Meeting of the European Federation for Animal Science. Wageningen Academic Publishers, Lyon, France.

Martín-Fernández JA, Hron K, Templ M, Filzmoser P and Palarea-Albaladejo J 2015. Bayesian-multiplicative treatment of count zeros in compositional data sets. Statistical Modelling 15, 134–158.

Martin MJ, Dórea JRR, Borchers MR, Wallace RL, Bertics SJ, DeNise SK, Weigel KA and White HM 2021a. Comparison of methods to predict feed intake and residual feed intake using behavioral and metabolite data in addition to classical performance variables. Journal of Dairy Science 104, 8765–8782.

Martin P, Ducrocq V, Gordo DGM and Friggens NC 2021b. A new method to estimate residual feed intake in dairy cattle using time series data. Animal 15, 100101.

Martinez Boggio G, Meynadier A, Buitenhuis AJ and Etancelin CM 2022. Host genetic control on rumen microbiota and its impact on dairy traits in sheep. Genetics Selection Evolution 54, 1–19.

Martinez Boggio G, Meynadier A, Daunis-I-Estadella P and Marie-Etancelin C 2021. Compositional analysis of ruminal bacteria from ewes selected for somatic cell score and milk persistency. PLoS ONE 16, 1–16.

Mazinani M and Rude B 2020. Population, world production and quality of sheep and goat products. American Journal of Animal and Veterinary Sciences 15, 291–299.

McGovern E, McGee M, Byrne CJ, Kenny DA, Kelly AK and Waters SM 2020. Investigation into the effect of divergent feed efficiency phenotype on the bovine rumen microbiota across diet and breed. Nature Publishing Group UK.

Meale SJ, Morgavi DP, Cassar-Malek I, Andueza D, Ortigues-Marty I, Robins RJ, Schiphorst AM, Laverroux S, Graulet B, Boudra H and Cantalapiedra-Hijar G 2017. Exploration of Biological Markers of Feed Efficiency in Young Bulls. Journal of Agricultural and Food Chemistry 65, 9817–9827.

Membrive CMB 2016. Anatomy and Physiology of the Rumen. In Rumenology (eds. D.D. Millen, M. De Beni Arrigoni and R.D. Lauritano Pacheco), pp. 1–38. Springer International Publishing, Cham.

Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM and Culhane AC 2016. Dimension reduction techniques for the integrative analysis of multi-omics data. Briefings in Bioinformatics 17, 628–641.

Messad F, Louveau I, Koffi B, Gilbert H and Gondret F 2019. Investigation of muscle transcriptomes using gradient boosting machine learning identifies molecular predictors of feed efficiency in growing pigs. BMC Genomics 20.

Meuwissen THE, Hayes BJ and Goddard ME 2001. Prediction of total genetic value using genome-

wide dense marker maps. Genetics 157, 1819–1829.

Meuwissen T, Hayes B and Goddard M 2016. Genomic selection: A paradigm shift in animal breeding. Animal Frontiers 6, 6–14.

Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F, Chang C-C, Lin C-C and Meyer MD 2019. Package 'e1071'. The R Journal.

Meyer AM, Vraspir RA, Ellison MJ and Cammack KM 2015. The relationship of residual feed intake and visceral organ size in growing lambs fed a concentrate- or forage-based diet. Livestock Science 176, 85–90.

Mialon M-M, Aigueperse N, Parisot S, Durand C, Guittard A, Ginane C, Bonnafe G, Douls S, Tortereau F and Boivin X 2022. Early experience and genetic predispositions: what impact on behavioural adaptation and performance of ewe lambs on pasture? In 26. Rencontres autour des Recherches sur les Ruminants (3R 2022), 26ème Rencontres Recherches Ruminants, pp. 573–576. Paris, France.

Millward DJ, Garlick PJ and Reeds PJ 1976. The energy cost of growth. Proceedings of the Nutrition Society 35, 339–349.

Mirza B, Wang W, Wang J, Choi H, Chung NC and Ping P 2019. Machine learning and integrative analysis of biomedical big data. Genes 10.

Montanholi YR, Swanson KC, Palme R, Schenkel FS, McBride BW, Lu D and Miller SP 2010. Assessing feed efficiency in beef steers through feeding behavior, infrared thermography and glucocorticoids. Animal 4, 692–701.

Mora M, González P, Quevedo JR, Montañés E, Tusell L, Bergsma R and Piles M 2023. Impact of multi-output and stacking methods on feed efficiency prediction from genotype using machine learning algorithms. Journal of Animal Breeding and Genetics, 1–15.

Moreno C, Bouix J, Brunel JC, Weisbecker JL, François D, Lantier F and Elsen JM 2001. Genetic parameter estimates for carcass traits in the INRA401 composite sheep strain. Livestock Production Science 69, 227–232.

Morgavi DP, Rathahao-paris E, Popova M, Boccard J, Nielsen KF and Boudra H 2015. Rumen microbial communities influence metabolic phenotypes in lambs. Frontiers in Microbiology 6.

Mottet A, de Haan C, Falcucci A, Tempio G, Opio C and Gerber P 2017. Livestock: On our plates or eating at our table? A new analysis of the feed/food debate. Global Food Security 14, 1–8.

Mucha S, Tortereau F, Doeschl-Wilson A, Rupp R and Conington J 2022. Animal Board Invited Review: Meta-analysis of genetic parameters for resilience and efficiency traits in goats and sheep. Animal 16, 100456.

Muir SK, Linden N, Kennedy A, Knight MI, Paganoni B, Kearney G, Thompson AN and Behrendt R 2020. Correlations between feed intake, residual feed intake and methane emissions in Maternal Composite ewes at post weaning, hogget and adult ages. Small Ruminant Research 192, 1–7.

Muir SK, Linden N, Knight M, Behrendt R and Kearney G 2018. Sheep residual feed intake and feeding behaviour: Are 'nibblers' or 'binge eaters' more efficient? Animal Production Science 58, 1459–1464.

Nadeau C and Bengio Y 2003. Inference for the Generalization Error. Machine Learning 52, 239–281.

Navajas EA, Ciappesoni G, Gimeno D, Velazco JI and De Barbieri I 2022. 37. Association of genetic resistance to internal nematodes and production traits on feed efficiency and methane emissions in Corriedale lambs. In Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP), pp. 195–198. Wageningen Academic Publishers, Rotterdam.

Newbold CJ, de la Fuente G, Belanche A, Ramos-Morales E and McEwan NR 2015. The Role of Ciliate Protozoa in the Rumen . Frontiers in Microbiology 6.

Ning MM and Lo EH 2010. Opportunities and challenges in omics. Translational Stroke Research 1, 233–237.

Notter DR, Ferrell CL and Field RA 1984. Effects of breed and intake level on growth and feed efficiency in ram lambs. Journal of Animal Science 58, 560–576.

Noziere P, Sauvant D and Delaby L 2018. Inra, 2018. Alimentation des ruminants. Editions Quae, Versaill.

Nti IK, Adekoya AF and Weyori BA 2020. A comprehensive evaluation of ensemble learning for stock-market prediction. Journal of Big Data 7.

O'Reilly-Wapstra JM, Freeman JS, Barbour R, Vaillancourt RE and Potts BM 2013. Genetic analysis of the near-infrared spectral phenome of a global Eucalyptus species. Tree Genetics & Genomes 9, 943–959.

Padua JT, Dasilva RG, Bottcher RW and Hoff SJ 1997. Effect of high environmental temperature on weight gain and food intake of Suffolk lambs reared in a tropical environment. In Proceedings of 5th international symposium, pp. 809–815. Bloomington, Minnesota, USA.

Paganoni B, Rose G, Macleay C, Jones C, Brown DJ, Kearney G, Ferguson M and Thompson AN 2017. More feed efficient sheep produce less methane and carbon dioxide when eating high-quality pellets. Journal of Animal Science 95, 3839.

Papi N, Mostafa-Tehrani A, Amanlou H and Memarian M 2011. Effects of dietary forage-to-concentrate ratios on performance and carcass characteristics of growing fat-tailed lambs. Animal Feed Science and Technology 163, 93–98.

Parra R 1978. Comparison of foregut and hindgut fermentation in herbivores. In The ecology of arboreal folivores (ed. G. Montgomery), pp. 205–230. Smithsonian Institution Press, Washington DC.

Pei C, Liu Q, Dong C, Li H, Jiang J and Gao W 2010. Anaerobe Diversity and abundance of the bacterial 16S rRNA gene sequences in forestomach of alpacas ( Lama pacos ) and sheep ( Ovis aries ). Anaerobe 16, 426–432.

Perea K, Perz K, Olivo SK, Williams A, Lachman M, Ishaq SL, Thomson J and Yeoman CJ 2017. Feed efficiency phenotypes in lambs involve changes in ruminal, colonic, and small-intestine-located microbiota. Journal of Animal Science 95, 2585–2592.

Piaggio L, Quintans G, Julián RS, Ferreira G, Ithurralde J, Fierro S, Pereira ASC, Baldi F and Banchero GE 2018. Growth , meat and feed ef fi ciency traits of lambs born to ewes submitted to energy restriction during mid-gestation. Animal 12, 256–264.

Picard M, Scott-Boyer MP, Bodein A, Périn O and Droit A 2021. Integration strategies of multi-omics data for machine learning analysis. Computational and Structural Biotechnology Journal 19, 3735–3746.

Picrust S 2013. PICRUSt2 for prediction of metagenome functions.

Playne MJ 1985. Determination of Ethanol , Volatile Fatty Acids , Lactic and Succinic Acids in Fermentation Liquids by Gas Chromatography. Journal of the Science of Food and Agriculture 36, 638–644.

Poirier S, Déjean S, Midoux C, Lê Cao KA and Chapleur O 2020. Integrating independent microbial studies to build predictive models of anaerobic digestion inhibition by ammonia and phenol. Bioresource Technology 316, 123952.

Ponziani FR, Zocco MA, D'Aversa F, Pompili M and Gasbarrini A 2017. Eubiotic properties of rifaximin: Disruption of the traditional concepts in gut microbiota modulation. World Journal of Gastroenterology 23, 4491–4499.

Prifti E, Chevaleyre Y, Hanczar B, Belda E, Zucker J, Danchin A and Cl K 2020. Interpretable and accurate prediction models for metagenomics data. 1–11.

Prohaska SJ and Stadler PF 2011. The Use and Abuse of -Omes. In Bioinformatics for Omics Data: Methods and Protocols (ed. B. Mayer), pp. 173–196. Humana Press, Totowa, NJ, USA.

Pryce JE, Arias J, Bowman PJ, Davis SR, Macdonald KA, Waghorn GC, Wales WJ, Williams YJ, Spelman RJ and Hayes BJ 2012. Accuracy of genomic predictions of residual feed intake and 250-day body weight in growing heifers using 625,000 single nucleotide polymorphism markers. Journal of Dairy Science 95, 2108–2119.

Qadri QR, Zhao Q, Lai X, Zhang Z, Zhao W, Pan Y and Wang Q 2022. Estimation of Complex-Trait Prediction Accuracy from the Different Holo-Omics Interaction Models. Genes 13.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J and Glöckner FO 2013. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Research 41, 590–596.

R Core Team 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rajaei Sharifabadi H, Zamiri MJ, Rowghani E and Bottje WG 2012. Relationship between the activity of mitochondrial respiratory chain complexes and feed efficiency in fat-tailed Ghezel lambs. Journal of Animal Science 90, 1807–1815.

Rappoport N and Shamir R 2018. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Research 46, 10546–10562.

Redden RR, Surber LMM, Roeder BL, Nichols BM, Paterson JA and Kott RW 2011. Residual feed efficiency established in a post-weaning growth test may not result in more efficient ewes on the range. Small Ruminant Research 96, 155–159.

Richardson EC and Herd RM 2004. Biological basis for variation in residual feed intake in beef cattle. 2. Synthesis of results following divergent selection. Australian Journal of Experimental Agriculture 44, 431–440.

Rincent R, Charpentier J, Faivre-rampant P, Paux E, Le Gouis J, Bastien C and Segura V 2018. Phenomic Selection Is a Low-Cost and High-Throughput Method Based on Indirect Predictions : Proof of Concept on Wheat and Poplar. G3 Genes|Genomes|Genetics 8, 3961–3972.

Rius AG, Kittelmann S, Macdonald KA, Waghorn GC, Janssen PH and Sikkema E 2012. Nitrogen metabolism and rumen microbial enumeration in lactating cows with divergent residual feed intake fed high-digestibility pasture. Journal of Dairy Science 95, 5024–5034.

Rivera-Pinto J, Egozcue JJ, Pawlowsky-Glahn V, Paredes R, Noguera-Julian M and Calle ML 2018. Balances: a New Perspective for Microbiome Analysis. mSystems 3, 10.1128/msystems.00053-18.

Robinson DL, Cafe LM and Greenwood PL 2013. Meat Science And Muscle Biology Symposium: Developmental programming in cattle: Consequences for growth, efficiency, carcass, muscle, and beef quality characteristics1,2. Journal of Animal Science 91, 1428–1442.

Rohart F, Eslami A, Matigian N, Bougeard S and Lê Cao KA 2017a. MINT: A multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. BMC Bioinformatics 18, 1–13.

Rohart F, Gautier B, Singh A and Lê Cao KA 2017b. mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS Computational Biology 13, 1–19.

Romao JM, Jin W, Dodson M V, Hausman GJ, Moore SS and Guan LL 2011. MicroRNA regulation in mammalian adipogenesis. Experimental Biology and Medicine 236, 997–1004.

Rosenberg E 2021. Evolution of Holobionts: The Hologenome Concept. In Microbiomes - Current

Knowledge and Unanswered Questions, pp. 317–352. Springer International Publishing, Cham, Switzerland.

Ross EM, Hayes BJ, Tucker D, Bond J, Denman SE and Oddy VH 2020. Animal Genetics and Genomics Genomic predictions for enteric methane production are improved by metabolome and microbiome data in sheep ( Ovis aries ). Journal of Animal Science 98.

Ross EM, Moate PJ, Marett LC, Cocks BG and Hayes BJ 2013. Metagenomic Predictions : From Microbiome to Complex Health and Environmental Phenotypes in Humans and Cattle. PLOS ONE 8, e73056.

Russell JB, Muck RE and Weimer PJ 2009. Quantitative analysis of cellulose degradation and growth of cellulolytic bacteria in the rumen. FEMS Microbiology Ecology 67, 183–197.

Saatchi M, Schnabel RD, Taylor JF and Garrick DJ 2014. Large-effect pleiotropic or closely linked QTL segregate within and across ten US cattle breeds. BMC Genomics 15.

Saborío-Montero A, Gutiérrez-Rivas M, López-García A, García-Rodríguez A, Atxaerandio R, Goiri I, Jiménez-Montero JA and González-Recio O 2021. Holobiont effect accounts for more methane emission variance than the additive and microbiome effects on dairy cattle. Livestock Science 250.

Sanchez MP, Tribout T, Fritz S, Wolf V, Laithier C, Brochard M and Boichard D 2022. Opportunities for genomic selection of cheese-making traits in Montbéliarde cows. Journal of Dairy Science 105, 5206–5220.

Sanjorjo RA, Tseten T, Kang M and Kwon M 2023. In Pursuit of Understanding the Rumen Microbiome. Fermentation 9.

Santos A, Valdés C, Giráldez FJ, López S, France J, Frutos J, Fernández M and Andrés S 2018. Feed efficiency and the liver proteome of fattening lambs are modified by feed restriction during the suckling period. Animal 12, 1838–1846.

Sargolzaei M, Chesnais JP and Schenkel FS 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15.

Sasson G, Moraïs S, Kokou F, Plate K, Trautwein-schult A and Jami E 2022. Metaproteome plasticity sheds light on the ecology of the rumen microbiome and its connection to host traits. ISME 16, 2610–2621.

Sauvant D and Nozière P 2016. Quantification of the main digestive processes in ruminants : the equations involved in the renewed energy and protein feed evaluation systems. Animal 10, 755–770.

Schneider M V and Orchard S 2011. Omics Technologies, Data and Bioinformatics Principles. In Bioinformatics for Omics Data: Methods and Protocols (ed. B. Mayer), pp. 3–30. Humana Press, Totowa, NJ, USA.

Scholz M, Ward D V, Pasolli E, Tolio T, Zolfo M, Asnicar F, Truong DT, Tett A, Morrow AL and Segata N 2016. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nature Methods 13, 435–438.

Selinger LB, Forsberg CW and Cheng KJ 1996. The rumen: A unique source of enzymes for enhancing livestock production. Anaerobe 2, 263–284.

Sepulveda BJ, Muir SK, Bolormaa S, Knight MI, Behrendt R, MacLeod IM, Pryce JE and Daetwyler HD 2022. Eating Time as a Genetic Indicator of Methane Emissions and Feed Efficiency in Australian Maternal Composite Sheep. Frontiers in Genetics 13, 1–14.

Shetty N, Løvendahl P, Lund MS and Buitenhuis AJ 2017. Prediction and validation of residual feed intake and dry matter intake in Danish lactating dairy cows using mid-infrared spectroscopy of milk. Journal of Dairy Science 100, 253–264.

Shirzadifar A, Plastow G, Basarab J, Miar Y, Li C, Fitzsimmons C, Riazi M and Manafiazar G 2022. A machine learning approach for predicting the most and the least feed-efficient groups in beef cattle. In Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP) (eds. R.F. Veerkamp and Y. De Haas), pp. 1656–1659. Wageningen Academic Publishers, Rotterdam, Netherlands.

Silva RMO, Fragomeni BO, Lourenco DAL, Magalhães AFB, Irano N, Carvalheiro R, Canesin RC, Mercadante MEZ, Boligon AA, Baldi FS, Misztal I and Albuquerque LG 2016. Accuracies of genomic prediction of feed efficiency traits using different prediction and validation methods in an experimental Nelore cattle population. Journal of Animal Science 94, 3613–3623.

Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ and Cao KAL 2019. DIABLO: An integrative approach for identifying key molecular drivers from multi-omics assays. Bioinformatics 35, 3055–3062.

Skillman LC, Toovey AF, Williams AJ and Wright G 2007. Development and Validation of a Real-Time PCR Method To Quantify Rumen Protozoa and Examination of Variability between Entodinium Populations in Sheep Offered a Hay-Based Diet. Applied and Environmental Microbiology 72, 200–206.

SMARTER 2018. Retrieved on 28 April 2023, from *https://www.smarterproject.eu/*.

Smith WB, Galyean ML, Kallenbach RL, Greenwood PL and Scholljegerdes EJ 2021. Understanding intake on pastures: how, why, and a way forward. Journal of Animal Science 99, 1–17.

Snowder GD and Vleck LD Van 2003. Estimates of genetic parameters and selection strategies to improve the economic efficiency of postweaning growth in lambs 1. Journal of Animal Science 81, 2704–2713.

Van Soest PJ 1994. Nutritional Ecology of the Ruminant. Cornell University Press.

Song M, Greenbaum J, Luttrell J, Zhou W, Wu C, Shen H, Gong P, Zhang C and Deng HW 2020. A Review of Integrative Imputation for Multi-Omics Datasets. Frontiers in Genetics 11, 1–15.

Soule M 1967. Phenetics of Natural Populations I. Phenetic Relationships of Insular Populations of the Side-Blotched Lizard. Evolution 21, 858.

Stecher B, Maier L and Hardt WD 2013. 'Blooming' in the gut: How dysbiosis might contribute to pathogen evolution. Nature Reviews Microbiology 11, 277–284.

Stiverson J, Morrison M and Yu Z 2011. Populations of Select Cultured and Uncultured Bacteria in the Rumen of Sheep and the Effect of Diets and Ruminal Fractions. International Journal of Microbiology 2011, 750613.

Storm E, Øskov ER and Smart R 1983. The nutritive value of rumen micro-organisms in ruminants: 2. The apparent digestibility and net utilization of microbial N for growing lambs. British Journal of Nutrition 50, 471–478.

Su M, Hao Z, Shi H, Li T, Wang H, Li Q, Zhang Y and Ma Y 2022. Metagenomic Analysis Revealed Differences in Composition and Function Between Liquid-Associated and Solid-Associated Microorganisms of Sheep Rumen. Frontiers in Microbiology 13, 1–14.

Suárez-Vega A, Frutos P, Gutiérrez-gil B, Esteban-blanco C, Toral PG, Arranz J and Hervás G 2023. Feed efficiency in dairy sheep : An insight from the milk transcriptome. Frontiers in Veterinary Science 10.

Tapio M, Fischer D, Mäntysaari P and Tapio I 2023. Rumen Microbiota Predicts Feed Efficiency of Primiparous Nordic Red Dairy Cows. Microorganisms 11.

Tardivel PJC, Canlet C, Lefort G, Tremblay-Franco M, Debrauwer L, Concordet D and Servien R 2017. ASICS: an automatic method for identification and quantification of metabolites in complex 1D 1H

NMR spectra. Metabolomics 13, 109.

Tchamitchian L, Lefevre C, Brunel C, Bibe B and Ricordeau G 1986. Development of a New Synthetic Prolific Line of Sheep (INRA 401). In 3rd World Congress on Genetics Applied to Livestock Production, pp. 535–540. Lincoln.

Tenenhaus M 1998. La régression PLS: théorie et pratique. Editions technip.

Terré M, Castells L, Fàbregas F and Bach A 2013. Short communication: Comparison of pH, volatile fatty acids, and microbiome of rumen samples from preweaned calves obtained via cannula or stomach tube. Journal of Dairy Science 96, 5290–5294.

Thompson CL and Holmes AJ 2009. A window of environmental dependence is evident in multiple phylogenetically distinct subgroups in the faecal community of piglets. FEMS Microbiology Letters 290, 91–97.

Tibshirani R 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58, 267–288.

Tiphine L, David I, Raoul J, Guerrier J, Praud JP, Bodin L, François D, Jullien E and Poivey JP 2011. Estimation of breeding values for meat sheep in France. In 62nd Annual Meeting of the European Federation of Animal Science (EAAP), p. 344. Wageningen, NLD: Wageningen Academic Publishers, Stavanger.

Tortereau F, Marie-Etancelin C, Weisbecker JL, Marcon D, Bouvier F, Moreno-Romieux C and François D 2020. Genetic parameters for feed efficiency in Romane rams and responses to single-generation selection. Animal 14, 681–687.

Tortereau F, Weisbecker J-L, Marcon D, François D, Le Graverand Q and Marie-Etancelin C 2023. Improving feed efficiency in meat sheep increases CH4 emissions measured indoors with dry forage or on pasture (in press). In 74th Annual Meeting of the European Federation for Animal Science. Wageningen Academic Publishers, Lyon, France.

Töscher A and Jahrer M 2009. The BigChaos Solution to the Netflix Prize 2009. Netflix prize documentation, 1–52.

Touitou F 2023. Comprendre le lien entre l'activité des micro-organismes présents dans le rumen, le métabolisme de l'hôte et l'efficience alimentaire des agneaux. Université Paul Sabatier - Toulouse III, 316pp.

Touitou F, Tortereau F, Bret L, Marty-Gasset N, Marcon D and Meynadier A 2022. Evaluation of the Links between Lamb Feed Efficiency and Rumen and Plasma Metabolomic Data. Metabolites 12, 304.

Vailati-Riboni M, Palombo V and Loor JJ 2017. What Are Omics Sciences? In Periparturient Diseases of Dairy Cows: A Systems Biology Approach (ed. B.N. Ametaj), pp. 1–7. Springer International Publishing, Cham, Switzerland.

Le Van TD, Robinson JA, Ralph J, Greening RC, Smolenski WJ, Leedle JAZ and Schaefer DM 1998. Assessment of reductive acetogenesis with indigenous ruminal bacterium populations and Acetitomaculum ruminis. Applied and Environmental Microbiology 64, 3429–3436.

Vandenheede M and Bouissou MF 1993. Sex differences in fear reactions in sheep. Applied Animal Behaviour Science 37, 39–55.

VanRaden PM 2008. Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science 91, 4414–4423.

Vaz C and Tanavde V 2018. Proteomics. In Omics Approaches, Technologies And Applications (eds. P. Arivaradarajan and G. Misra), pp. 57–73. Springer Singapore, Singapore.

Velasco-Galilea M, Piles M, Ramayo-Caldas Y and Sánchez JP 2021. The value of gut microbiota to

predict feed efficiency and growth of rabbits under different feeding regimes. Scientific Reports 11, 1–18.

Vihko P and Wagenerb C 1992. Structure and genetic engineering of antigens and antibodies : applications in immunoassays. Clinica Chimica Acta 207, S5–S11.

Wang Y and Lê Cao KA 2020. Managing batch effects in microbiome data. Briefings in Bioinformatics 21, 1954–1970.

Wang Y and Qian PY 2009. Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. PLoS ONE 4, 10.

Wei X, Dong Z, Cheng F, Shi H, Zhou X, Li B, Wang L, Wang W and Zhang J 2021. Seasonal diets supersede host species in shaping the distal gut microbiota of Yaks and Tibetan sheep. Scientific Reports 11, 1–11.

Weimer PJ 2015. Redundancy , resilience , and host specificity of the ruminal microbiota : implications for engineering improved ruminal fermentations. Frontiers in Microbiology 6.

Weishaar R, Wellmann R, Camarinha-Silva A, Rodehutscord M and Bennewitz J 2020. Selecting the hologenome to breed for an improved feed efficiency in pigs—A novel selection index. Journal of Animal Breeding and Genetics 137, 14–22.

Westhues M, Heuer C, Thaller G, Fernando R and Melchinger AE 2019. Efficient genetic value prediction using incomplete omics data. Theoretical and Applied Genetics 132, 1211–1222.

Widmann P, Reverter A, Weikard R, Suhre K, Hammon HM, Albrecht E and Kuehn C 2015. Systems biology analysis merging phenotype, metabolomic and genomic data identifies Non-SMC Condensin I Complex, Subunit G (NCAPG) and cellular maintenance processes as major contributors to genetic variability in Bovine feed efficiency. PLoS ONE 10, 1–22.

Wishart DS 2007. Current Progress in computational metabolomics. Briefings in Bioinformatics 8, 279–293.

Xue MY, Xie YY, Zhong Y, Ma XJ, Sun HZ and Liu JX 2022. Integrated meta-omics reveals new ruminal microbial features associated with feed efficiency in dairy cattle. Microbiome 10.

Yanagita K, Kamagata Y, Kawaharasaki M, Suzuki T, Nakamura Y and Minato H 2000. Phylogenetic Analysis of Methanogens in Sheep Rumen Ecosystem and Detection of Methanomicrobium mobile by Fluorescence In Situ Hybridization. Bioscience, Biotechnology, and Biochemistry 64, 1737–1742.

Yang X, Wang W, Wang X, Zhang D, Li X, Zhang Y, Zhao Y, Zhao L, Wang J, Xu D, Li W, Zhou B, Lin C, Zeng X and Zhai R 2022. Polymorphism in ovine ADCY8 gene and its association with residual feed intake in Hu sheep. Animal Biotechnology, 1–8.

Zeder MA 2008. Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. Proceedings of the National Academy of Sciences of the United States of America 105, 11597–11604.

Zhang Y, Li F, Chen Y, Wu H, Meng Q and Guan LL 2020. Metatranscriptomic Profiling Reveals the Effect of Breed on Active Rumen Eukaryotic Composition in Beef Cattle With Varied Feed Efficiency. Frontiers in Microbiology 11, 1–12.

Zhang J, Shi H, Li S, Cao Z, Yang H and Wang Y 2019a. Integrative hepatic metabolomics and proteomics reveal insights into the mechanism of different feed efficiency with high or low dietary forage levels in Holstein heifers. Journal of Proteomics 194, 1–13.

Zhang X, Wang W, Mo F, La Y, Li C and Li F 2017. Association of residual feed intake with growth and slaughtering performance, blood metabolism, and body composition in growing lambs. Scientific Reports 7, 1–11.

Zhang D, Zhang X, Li F, Li C, La Y, Mo F, Li G, Zhang Y, Li X, Song Q, Zhao Y and Wang W 2019b.

Transcriptome Analysis Identifies Candidate Genes and Pathways Associated With Feed Efficiency in Hu Sheep . Frontiers in Genetics 10.

Zhang D, Zhang X, Li F, Li X, Zhao Y, Zhang Y, Zhao L, Xu D, Wang J, Yang X, Cui P and Wang W 2022. Identification and characterization of circular RNAs in association with the feed efficiency in Hu lambs. BMC Genomics 23.

Zhang D, Zhang X, Li F, Yuan L and Zhang Y 2021. Polymorphisms in ovine ME1 and CA1 genes and their association with feed efficiency in Hu sheep. Journal of Animal Breeding and Genetics 138, 589–599.

# Appendix 1: Congress abstract

Appendix 1 is based upon the analyzes of Anaïs Marquisseau, a master 2 trainee supervised by Flavie Tortereau, Christel Marie-Etancelin, and me during 6 months. The following abstract was accepted at the EAAP 2023 congress.

"**Prediction of feed efficiency related traits from plasma NMR spectra**

Anaïs Marquisseau[1], Flavie Tortereau[1], Nathalie Marty-Gasset[1], Christel Marie-Etancelin[1], Quentin Le Graverand[1]

[1] GenPhySE, Université de Toulouse, INRAE, ENVT, 31326, Castanet Tolosan, France

Feed efficiency is a key trait to integrate in breeding programs, particularly in order to limit the feed-food competition and the environmental impact of livestock production. The calculation of feed efficiency criteria requires individual feed intakes to be recorded, which is too expensive in small ruminants to be reasonably proposed. An option to make this trait more affordable is to predict feed intake or efficiency from a variety of predictors that can be easily recorded. As it has been evidenced that the animal metabolism is one of the main biological function underlying feed efficiency, we propose to examine the predictability of feed efficiency related traits from plasma metabolome. Plasma samples from 265 Romane male lambs fed a 100% concentrate diet were analyzed with NMR. NMR spectra were divided into 877 buckets of 0.01 ppm, and the considered values were the area under the curve of each bucket. These variables were CLR transformed before multivariate analyses (sparse Partial least squares, sPLS) used for prediction. Prediction performances of feed intake and RFI were assessed through 5-fold nested cross-validation repeated 50 times, i.e. over 250 models. Accuracies of prediction from NMR buckets were compared to the accuracy obtained from body weights, growth, body composition (called zootechnical traits hereafter). As a result, we highlighted that buckets did not improve the prediction of feed intake from zootechnical traits: an average $R^2$ of 0.7 was obtained from zootechnical traits with or without buckets, against 0.2 from buckets only. For RFI, $R^2$ were below 0.1 whatever the set of predictors. Considering whole spectra did not help predict feed efficiency nor feed intake. However, the main buckets involved in RFI prediction were consistent with metabolites previously associated to feed efficiency: such as beta-hydroxyisovaleric acid or L-tyrosine" (Marquisseau et al., 2023).

# Appendix 2: Congress abstract

The following abstract was accepted at the EAAP 2023 congress, for an oral presentation.

"**Blending multivariate models to predict feed efficiency and explore multiple omics in meat sheep**

Le Graverand, Q.[1], Tortereau, F.[1], Marie-Etancelin, C.[1], Meynadier, A.[1], Weisbecker, J.L.[1], Lê Cao, K.A.[2]

[1] GenPhySE, Université de Toulouse, INRAE, ENVT, 31326, Castanet-Tolosan, France;

[2] School of Mathematics and Statistics, University of Melbourne, VIC 3010, Australia

Selecting sheep for feed efficiency would improve the sustainability of sheep farming by decreasing feeding needs. However, due to the costs of recording feed intake, feed efficiency is rarely selected in sheep. Identifying feed efficiency biomarkers could help resolve this issue. A total of 258 Romane male lambs were phenotyped in the growing period for Residual Feed Intake (RFI)-in three different batches. Rumen fluid and blood were sampled as potential sources of biomarkers for feed efficiency. Multivariate analyses were performed with six distinct 'blocks' of predictors: fixed effects and covariates (FC), genotypes (SNPs), plasma NMR spectra (NMR), ruminal volatile fatty acids (VFAs), long-chain fatty acids (LFAs), bacteria and archaea abundances (16S amplicon sequencing). We modified a Partial Least Square regression approach (PLS) to account for the three batches while selecting biomarkers of feed efficiency (Rohart et al., 2017). Cross-validation was repeated to fit one model per block on our training data (60% of the samples). Then, predictions for the validation set (30% of the samples) were obtained by using a weighted aggregation-based on the performance on each validation set. Testing data (10%) were independently used to assess the overall prediction accuracy based on Pearson correlations. When RFI was predicted from separate blocks, the average accuracy was low to moderate: 0.08 (standard deviation: 0.17) from VFAs to 0.44 (0.13) from SNPs. When RFI was predicted with our approach combining different omics, accuracy increased and reached an average of 0.55 (0.11). Based on weights attributed to blocks of predictors, we were able to rank the most predictive blocks to explain RFI: SNPs, FC, NMR, 16S, LFA and VFA. Furthermore, within each block we identified variables that were highly associated with feed efficiency RFI, including β-hydroxyisovaleric acid and a SNP located on the chromosome 3. To conclude, blending models is useful to integrate heterogeneous omics data: from predicting efficiency, to identifying associations between multi-omics predictors."

# Appendix 3: Training

Several courses were undertaken, to develop skills during the thesis.

**Personal development**

- EDEN21 – Workshop for PhD students and supervisors (October 12, 2021 – October 15, 2021, Satillieu, EDEN) – 28 hours

**Ethics and scientific integrity**

- Ethics and scientific integrity (February 10, 2021 - February 10, 2021, Toulouse, Université Fédérale de Toulouse Midi Pyrénées) - 6 hours

**Literature review and publication methodology**

- Writing and Presenting Scientific Papers (August 29, 2021 - August 29, 2021, Davos, EAAP 2021) - 8 hours

**Career development**

- PhDOOC - MOOC PhD and Career development (January 12, 2022 - February 28, 2022, online, PhDOOC) - 12 hours

**Pedagogy for the University**

- Design: how to integrate digital technology into teaching (March 28, 2023, Toulouse, Université Fédérale Toulouse Midi Pyrénées) - 3.5 hours

- Evaluate: The functions and objectives of evaluation (May 25, 2023, Toulouse, Université Fédérale Toulouse Midi Pyrénées) - 3 hours

- MOOC Dyslexic students in my lecture hall: understanding and helping (January 17, 2023, online, FUN MOOC) - 12 hours

- Training to teach in higher education (February 01, 2021, online, FUN MOOC) - 24 hours

**Scientific**

- Compositional Data Analysis with CoDaPack and R (October 05, 2020 - September 16, 2021, online, Coda Association) - 25 hours

- First steps in AWK programming (March 25, 2021, online, Génotoul) - 6 hours

- FROGS: tools for bioinformatics and statistics analyses with amplicon metagenomics data (March 15, 2021, online, Génotoul) - 26 hours

- Data integration with mixOmics and mixKernel (June 08, 2021, online, Génotoul) - 12 hours

- mixOmics R Essentials for Biological Data Integration (October 11, 2021, online, The University of Melbourne) - 40 hours

- MOOC Bioinformatics: algorithms and genomes (February 01, 2021, online, FUN MOOC) - 13 hours

- Practicals on machine learning and deep learning (October 18, 2021, Castanet-Tolosan, Génotoul) - 9 hours

- RNAseq alignment, quantification and transcript discovery with statistics (November 02, 2020, online, Génotoul) - 24 hours

- Using sed and AWK to modify large text files (March 24, 2021, online, Génotoul) - 6 hours

- Care and use of Laboratory animals (Function B) (March 01, 2021, Toulouse, ENVT) - 56 hours

# Appendix 4: Teaching and supervising

## Teaching

As a PhD student, I taught at « Prépa des INP » (undergraduate preparatory classes) and at ENSAT (agriculture engineering school) in Toulouse.

**Animal breeding (56 hours)**

- Practical courses: Population effective size and inbreeding (ENSAT – Licentiate 3rd year); Relatedness, genetic evaluation and genetic progress (ENSAT – Master 1st year).

**Population genetics (8 hours)**

- Practical courses: Allele and genotype frequencies under the Hardy-Weinberg equilibrium, or migration and genetic drift scenarios (ENSAT – Licentiate 3rd year).

**Microbiology (10 hours)**

- Practical course: Preparation of Petri dish cultures from wine samples, Gram staining (ENSAT – Licentiate 3rd year).

**Molecular biology (48 hours)**

- Practical courses: Enzymatic digestion of bacterial plasmids and electrophoresis (Prépa des INP Toulouse – Licentiate 1st year); Electrophoresis to check the presence of genes in genome-edited plants (ENSAT - Licentiate 3rd year).

**Scientific projects (6 hours)**

- Tutoring: Developing a scientific problematic and experimental design, inspired by IgNobels (ENSAT – Master 1st year).

**Agronomy (4 hours 30)**

- Assessing internship presentations: Technical and economic analysis in dairy cattle farms (ENSAT - Licentiate 3rd year).

- Assessing internship reports: Analyzing and advising dairy goat and dairy sheep farms ENSAT - Licentiate 3rd year).

**Diverse (16 hours)**

- Exam surveillance

# Supervising

**Master 2 – Internship**

- Anaïs Marquisseau (Master in bioinformatics at the University of Clermont Auvergne) was co-supervised by Flavie Tortereau, Christel Marie-Etancelin and me during 6 months. Topic: feed efficiency prediction from animal traits and plasma NMR spectra.

# Appendix 5: Scientific communications

## Published papers

- **Le Graverand Q**, Marie-Etancelin C, Meynadier A, Weisbecker J-L, Marcon D and Tortereau F 2023. Predicting feed efficiency traits in growing lambs from their ruminal microbiota. Animal, 100824.

- Marie-Etancelin C, Tortereau F, Gabinaud B, Martinez Boggio G, **Le Graverand Q**, Marcon D, De Almeida ML, Pascal G, Weisbecker JL and Meynadier A 2021. Apart from the diet, the ruminal microbiota of lambs is modified in relation to their genetic potential for feed efficiency or feeding behavior. Frontiers in Microbiology 12, 1–15.

## Congress theatre presentations

- **Le Graverand Q**, Tortereau F, Marie-Etancelin C, Meynadier A and Lê Cao KA 2023. Blending multivariate models to predict feed efficiency and explore multiple omics in meat sheep. August 2023, Lyon, France. In proceedings of the 74th Annual Meeting of the European Federation for Animal Science (pp. 569). Wageningen Academic Publishers.

- **Le Graverand Q**, Marie-Etancelin C, Weisbecker JL, Meynadier A, Marcon D and Tortereau F 2022. Using machine learning to predict feed intakes of meat sheep from animal traits and ruminal microbiota. July 2022, Rotterdam, Netherlands. In Proceedings of the 12th World Congress on Genetics Applied to Livestock Production (WCGALP), pp. 142–618. Wageningen Academic Publishers.

- **Le Graverand Q**, Tortereau F, Meynadier A, Marcon D and Marie-Etancelin C 2021. The rumen microbiota is modified in lambs divergently selected for residual feed intake. August 2021, Davos, Switzerland. In proceedings of the 72nd Annual Meeting of the European Federation for Animal Science. Wageningen Academic Publishers.

## Congress abstracts

- Conington J, Lambe N, Tortereau F, McGovern F, Navajas E, De Barbieri I, Ciappesoni G, Jakobsen J, Smith EM, Yates J, **Le Graverand Q**, Mcdermott K, Steinheim G, Aspeholen Åby B, Dønnem I, Mc Hugh N, Farrell L, Marie-Etancelin C, Johnson P and Rowe S 2022. Strategies to mitigate greenhouse gas emissions from pasture-based sheep systems–an EU project consortium view. July 2022, Rotterdam, Netherlands. In Proceedings of 12th World Congress on Genetics Applied to Livestock Production (pp. 107-110). Wageningen Academic Publishers.

- Marquisseau A, Tortereau F, Marty-Gasset N, Marie-Etancelin C and **Le Graverand Q** 2023. Prediction of feed efficiency related traits from plasma NMR spectra. August 2023, Lyon,

France. In proceedings of the 74th Annual Meeting of the European Federation for Animal Science. Wageningen Academic Publishers, Lyon, France.

- Tortereau F, Weisbecker JL, Coffre-Thomain C, Legoff Y, François D, **Le Graverand Q**, Marie-Etancelin C 2023. Improving feed efficiency in meat sheep increases CH4 emissions measured indoor or at pasture. August 2023, Lyon, France. In proceedings of the 74th Annual Meeting of the European Federation for Animal Science. Wageningen Academic Publishers, Lyon, France.

- Marie-Etancelin C, Weisbecker JL, Marcon D, Estivalet L, **Le Graverand Q**, Tortereau F 2023. August 2023, Lyon, France. Selecting feed-efficient sheep with concentrates alters their efficiency with forages and behaviour. In proceedings of the 74th Annual Meeting of the European Federation for Animal Science. Wageningen Academic Publishers, Lyon, France.

## Collaboration

- Collaborated with Kim-Anh Lê Cao, as a visiting student in the School of Mathematics and Statistics (October to December 2023, University of Melbourne, Australia) – Integration of multiple studies and omics.

# Résumé

L’élevage ovin doit relever de nombreux défis pour améliorer sa durabilité. Améliorer l'efficience alimentaire pourrait réduire l'empreinte environnementale de l’élevage, améliorer les revenus des éleveurs et atténuer la concurrence entre l’alimentation humaine et animale. Cependant, l'efficience alimentaire est rarement sélectionnée chez l’ovin allaitant car mesurer l'ingestion d'aliments individuellement coûte cher. Prédire l'efficience alimentaire serait une solution. L'étude des omiques pourrait identifier des prédicteurs qui faciliteraient la sélection des ovins pour l'efficience alimentaire.

La thèse s'est concentrée sur la prédiction de l'efficience alimentaire et de l'ingestion chez l’ovin allaitant. Les prédicteurs potentiels comprenaient des effets fixes, le poids vif de l’animal, le pedigree et des données omiques collectées chez des agneaux Romane entre 2018 et 2020. Les agneaux faisaient partie de lignées divergentes sélectionnées pour la consommation résiduelle. Les caractères de production et le microbiote du rumen ont été analysés avec des animaux nourris avec des concentrés, puis avec des fourrages. Les génotypes des animaux, le lipidome ruminal, le spectre proche-infrarouge fécal, les métabolomes ruminal et plasmatique ont été analysés seulement avec des concentrés.

Tout d'abord, les analyses en composantes principales ont mis en évidence les principaux facteurs de variation des potentiels prédicteurs de l'efficience. La généalogie et la génomique ont mis en évidence la structure de population. D'autre part, la composition du microbiote ruminal, les spectres fécaux et le métabolome sont influencés par l'environnement (c'est-à-dire l'année ou la période de phénotypage). Deuxièmement, le microbiote ruminal ne pouvait pas prédire l’ingestion de façon fiable si les populations de référence et cible ont été élevées lors d’années distinctes (corrélations entre les phénotypes prédits et réels ≤ 0,35). Par rapport au microbiote ruminal, les effets fixes et le poids corporel pouvaient prédire les caractères de production aussi précisément si ce n’est plus. L'étude de données omiques et non-omiques collectées sur plusieurs années a mis en évidence que les meilleurs prédicteurs étaient : la génétique de l'hôte pour la consommation résiduelle (0,54) ; les effets fixes et le poids pour l'indice de consommation (0,48) ainsi que la consommation (0,85). En revanche, les données ruminales prédisaient mal l'efficience alimentaire et la consommation. Enfin, l'intégration de données hétérogènes (omiques ou non) a été réalisée en pondérant la moyenne des prédictions de modèles entrainé sur un seul jeu de prédicteurs, ou en régressant les prédictions. L'intégration a significativement augmenté la précision de prédiction de l'efficience (≤ 0,59) et de l'ingestion (≤ 0,88).

La thèse ouvre de nouvelles perspectives pour l'exploitation des données omiques et l’identification de prédicteurs de l'efficience alimentaire. Des recherches sont nécessaires pour intégrer et comprendre les interactions entre les différentes omiques, ce qui pourrait améliorer notre compréhension de l'efficience alimentaire des moutons.

**Mots-clés** : efficience alimentaire, omiques, intégration de données, prédicteurs, génétique, ovin.