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. GUILLERMO MARTINEZ BOGGIO

le mardi 6 décembre 2022

Titre :

Understanding the role of ruminal microbiota in phenotype expression in Lacaune dairy sheep

---

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

MME ANNABELLE MEYNADIER

Rapporteurs :

M. CHRISTIAN MALTECCA, NC STATE UNIVERSITY

M. GABRIEL CIAPPESONI, INIA LAS BRUJAS

Membre(s) du jury :

MME CATHERINE LARZUL, INRA TOULOUSE, Président

M. JORDI ESTELLE, INRA JOUY EN JOSAS, Membre

M. JUAN PABLO SANCHEZ, IRTA BARCELONA, Membre

MME ANNABELLE MEYNADIER, ECOLE NATIONALE VETERINAIRE DE TOULOUSE, Membre

MME CHRISTEL MARIE-ETANCELIN, INRA TOULOUSE, Membre

Al fin y al cabo,

somos lo que hacemos para cambiar lo que somos.

Eduardo Galeano

Acknowledgments

First, I would like to thank Christel Marie-Etancelin and Annabelle Meynadier for their guidance throughout this thesis. They strongly contributed to my intellectual upbringing in genetics, microbiology and animal nutrition, inspired and encouraging to write this manuscript. This work would not have been possible without the sharing of their vast knowledge. Their patience and understanding are inestimable.

Second, I would like to thank Andres Legarra, Ole Christensen, Albert J. Buitenhuis and Pepus Daunis-i-Estadella for enlightening discussions and lots of ideas. Their advices and contributions to my work have been very important.

Third, I would like to thank all the staff of the INRAE Experimental Unit of La Fage and the National Veterinary of Toulouse for their work to be possible to have the data used in this work.

I have also fortunate to be surrounded by incredible people, who are now friends and without them my stay would have been harder. Especially with much less barbacues.

Some people come into our lives and quickly go. Some stay for a while and leave footprints… Thus, I would like to thank all the others I have met and who influenced me because without them I might not be here.

Last but not least, I would like to thank my family and Giuliana for their love, support and understanding. They encouraged me to take the road which led to where I am now.

Finally, this thesis would not have been possible without the financial support of INRAE and La Region Occitanie.

Table of Contents

Acknowledgments I

Table of Contents III

List of Tables VII

List of Figures VII

List of abbreviations IX

Chapter 1 - Introduction and objectives 1

Chapter 2 - Scientific background 5

2.1 Holobionts 5

2.2 Integrating the holobiont in quantitative genetics 6

2.2.1 Host genome on phenotype (G → P) 10

2.2.2 Microbiota on phenotype (M → P) 10

2.2.3 Host genome on microbiota (G → M) 11

Host genetic control over individual microbial abundance 11

Host genetic control over the microbial community 12

2.3 Host genome (G) 12

2.4 Microbiome (M) 14

2.4.1 Metagenome 14

2.4.2 Digestive ecosystem of ruminants 16

Forestomach biotope 17

Ruminal microbiota 18

Bacteria 19

2.4.3 Sequencing technologies for microbiome study 20

Short-read sequencing 21

Long-read sequencing 23

Characterization of microbial communities 24

2.5 Phenotype (P) 25

2.5.1 Milk quantity (1965-) ... 27
2.5.2 Fat and protein quantity and rate (1985-) ... 27
2.5.3 Somatic cell count, mammary morphology and scrapie disease (2005-) ... 27
2.5.4 Fine milk composition ... 28
Fatty acids ... 28
Proteins ... 30
2.6 Conclusion ... 30
Chapter 3 - Materials and methods ... 33
3.1 Experimental Unit of La Fage ... 33
3.2 Experimental design ... 35
3.2.1 Animals ... 35
3.2.2 Feeding practices ... 35
3.2.3 Genetic structure of the dairy flock ... 36
3.3 Animal's genotyping ... 37
3.4 Milk traits ... 38
3.4.1 Fine milk composition predicted from MIR spectra ... 40
3.4.2 Milk fatty acids quantified by gas chromatography ... 40
3.5 Sampling and analysis of rumen samples ... 41
3.5.1 Rumen fatty acids ... 42
3.5.2 Rumen microbial composition ... 42
DNA extraction and targeting 16S rRNA gene sequencing ... 43
Bioinformatics process ... 44
Analysis of microbial abundances ... 45
Compositional data analysis ... 47
3.6 Statistical models ... 49
3.7 Data availability ... 50
3.8 Conclusion ... 50
Chapter 4 - Working with microbial abundances ... 53
4.1 Introduction ... 53

4.2 Compositional analysis of ruminal bacteria from ewes selected for somatic cell score and milk persistency 54

4.3 Variation in rumen bacteria of Lacaune dairy ewes from one week to the next 71

4.4 Ruminal bacteria heritabilities: which is the impact of data processing? 84

4.5 Conclusion 87

Chapter 5 - Host genetic control on ruminal microbiota 89

5.1 Introduction 89

5.2 Host genetic control on rumen microbiota and its impact on dairy traits in sheep 90

5.2.1 Supplementary results 110

Distribution of OTUs heritabilities 110

Alpha-diversity heritability 110

5.3 Evidence of genetic links between rumen bacteria and milk somatic cell score in Lacaune dairy sheep 112

5.4 Conclusion 118

Chapter 6 - Impact of rumen microbiota on phenotypes 119

6.1 Introduction 119

6.2 Microbiability of milk composition and genetic control of microbiota effects in sheep 120

6.2.1 Supplementary results 136

6.3 Conclusion 137

Chapter 7 - General discussion 139

7.1 What impact does the rumen microbiota have on phenotypic variability? 139

7.1.1 Links between OTU abundances and phenotypes 140

7.1.2 Microbial community effect on phenotypes 142

7.2 What degree of control do host genetics have over microbial composition? 143

7.2.1 Host genetic control on OTU abundance 143

7.2.2 Host genetic control over the microbial community 145

7.3 How could the holobiont concept be implemented in animal breeding? 147

7.3.1 Genetic selection of microbiota 147

7.3.2 Short-term changes in microbiota composition 147

7.3.3 Selection of microbiota as a nongenetic effect 148

Chapter 8 - Conclusions and perspectives 151

References 153

Training and education during the PhD thesis 162

Publications during the PhD thesis 163

Résumé 165

Summary 166

List of Tables

Table 1. Description of short- and long-read next generation sequencing technologies. 23
Table 2. Total number of animals with records grouped by year and genetic lines (PERS, SCS or Socs2 genotype). 37
Table 3. Description of milk traits and fine milk composition traits. 39
Table 4. Number of DNA sequences and OTUs along the bioinformatics process for datasets. 45
Table 5. Number of taxon and frequency of unknown affiliation per taxonomic level. 46
Table 6. Alpha-diversity indices and their heritabilities estimates. 111
Table 7. Estimates of variance components ($\sigma ar2$ residual genetic, $\sigma u2$ microbial, and $\sigma e2$ residual variances), heritability ($h2$), microbiability ($cm2$), statistics of the relationship matrices (average of the diagonal $diag(\boldsymbol{O})$ and average relationship $\boldsymbol{O}$) for beta-lactoglobulin, for four different types of the relationships ($\boldsymbol{O_1}$, $\boldsymbol{O_2}$, $\boldsymbol{O_3}$, $\boldsymbol{O_4}$). 137

List of Figures

Figure 1. Multi-omics system from the genome to phenome (through intermediate omics: transcriptome, proteome and metabolome) and metagenome. 8
Figure 2. Models including gut microbial data. 9
Figure 3. Family of sheep illustrating how offspring receive their genes from each of their parents. 13
Figure 4. Main vectors of transmission for genetic (red arrow) and microbial (blue arrow) inheritance within generation (n) and between generations (n and n+1). 16
Figure 5. Rumen as a continuous fermentation chamber. 18
Figure 6. Illumina sequencing process. 22
Figure 7. Traditional dairy sheep production regions in France with their local breeds. 26
Figure 8. Timeline with the most relevant scientific programs developed at the INRAE Experimental Unit of La Fage. 34
Figure 9. Distribution of ewes based on selection index of somatic cell score for SCS line (SCS+ and SCS-), and selection index of coefficient of variation of milk production for PERS line (PERS+ and PERS-). 36

Figure 10. Mid-infrared (MIR) spectra technique (adapted from Sanchez et al., 2019). 40
Figure 11. Rumen sampling process. 41
Figure 12. Experimental design to obtain and process rumen samples. 43
Figure 13. Bioinformatics process of DNA sequences using FROGS 3.0 pipeline. 44
Figure 14. Frequency distribution of OTU heritabilities. 110
Figure 15. Main results obtained for individual and microbiota effects on phenotypes. 140
Figure 16. Main results obtained for host genetic effect on individual and community microbial composition 143

List of abbreviations

CLR: Centered log-ratio
DIM: Days in milk
EBV: Estimated breeding value
EU: Experimental unit
FA: Fatty acid
FAME: Fatty acid methyl esters
FC: Fat content
GBM: Geometric Bayesian method
GBS: Genotyping by sequencing
GWAS: Genome-wide association study
h^2: Heritability
INRAE: Institut national de la recherche pour l'agriculture, l'alimentation et l'environnement
ISOL: Index synthetique ovin lait (Synthetic index for milk sheep)
LCFA: Long-chain fatty acid
LD: Linkage disequilibrium
LSCS: Lactation somatic cell score
MIR: Mid-infrared
MY: Milk yield
c_m^2: Microbiability
MUFA: Monounsaturated fatty acid
MWAS: Microbiome-wide association study
NGS: Next generation sequencing
OTU: Operational taxonomic unit
PC: Protein content
PCR: Polymerase chain reaction
PUFA: Polyunsaturated fatty acid
QTL: Quantitative trait loci
rRNA: Ribosomal ribonucleic acid
SBL: Sequencing by ligation
SBS: Sequencing by synthesis
SCC: Somatic cell count
SCS: Somatic cell score
SFA: Saturated fatty acid
SNP: Single-nucleotide polymorphism
Socs2: Suppressor of cytokine signaling 2
VFA: Volatile fatty acid

Chapter 1
Introduction and objectives

The population of commensal microorganisms in and on living things was largely unknown for years. So much so that the microorganisms as a whole were considered a black box. However, this has changed and it is now well known as a community called microbiome that has a crucial impact on the host's life. These advances have been strengthened by the advent of new DNA sequencing technologies and bioinformatics tools allow for us to unveil a microscopic world of which little was known until a few years ago. This new look at the microbiome as a crucial element of the host, no longer as a black box, is well represented by what Blaser (2014) calls the microbiome revolution. And this is where the concept proposed by Margulis (1993) of holobiont fits in, as the term that allows for assemble the host with its microbiome community, to consider it as a single unit.

Animal breeding is not left out of this microbiome revolution, and in the last ten years has moved in the direction of evaluating the potential use of digestive microbiome information for system-level decision making. The gut microbiome has been shown to influence many disease-associated pathways in humans and animals, and studies of association to zootechnical traits are increasing in livestock species such as pigs, poultry and ruminants. Microbiota would serve as a tool to increase knowledge and improve prediction of complex traits, which are difficult and costly to measure, as well as those traits most closely associated with the microbial composition. In ruminants, the presence of symbiotic microorganisms in the rumen is closely related to productive traits (Difford *et al.*, 2018; Xue *et al.*, 2018; Matthews *et al.*, 2019). Therefore, this thesis will focus on dairy traits that are directly associated with rumen microbiota through volatile fatty acids and proteins produced by it. Milk is the basis of the world's food supply, and is a sector for which an annual growth rate of 1.7% (to 981 megatons by 2028) is faster than most other main agricultural commodities (OECD/FAO, 2022). The main marketable product is fluid milk, followed by cheese, where sheep systems, mainly in Asia and Europe aim to produce high quality cheese.

Based on this context, the hypotheses of this project are the rumen microbiota influence milk composition and udder health traits, and there is a host genetic control over its rumen microbiota. To address these hypotheses, this thesis aims to integrate the relationship between the host genome and ruminal metagenome in the modeling of the variability of dairy production and udder health traits in ewes. In that way, this thesis will contribute to a precise physiological characterization of animals in a population, through the use of high-dimensional data, which will improve the robustness of animals and the efficiency of breeding in the context of the sustainability of agricultural systems. In particular, through the ruminal microbiota it will allow for select animals according to their bacterial composition, and/or to improve the phenotype (health, welfare or performances) by means of a control based on ruminal microbiota via feed.

First of all, **Chapter 2** presents the scientific background that supports the objective of this thesis, as well as the missing information in terms of the holobiont and the links between the three key elements that compose it: the host genome, microbiome and phenotype. **Chapter 3** describes the scientific project, the experimentation and the data processing used to address the hypotheses. The results are presented from Chapter 4 to Chapter 6, through four original scientific publications and two congress communications.

The first step before answering any biological question, is a technical one: how to work properly with microbial abundances? In **Chapter 4**, a workflow to be applied on microbiome data is proposed and tested on a real dataset with the objective of discriminating genetic lines and estimating associations between rumen bacterial abundance and dairy traits. In addition, to assess the consistency of microbial composition data and the results of associations between microbiota and phenotypes over a short period of time, the stability of rumen bacterial abundance from one week to the next and the impact of this on the phenotypic association with milk and rumen fatty acids was determined. Finally, the proposed compositional workflow was compared with different data processing methods through the estimation of genetic parameters for rumen bacterial abundances.

The results in **Chapter 5** contribute to answer the biological question: what is the host genetic control over its rumen microbial composition? Then, the objectives were first to quantify the host genetic effect on rumen bacterial abundance, then to estimate genetic correlations with milk and udder health traits and to identify possible genomic regions in common between microbial abundances and dairy traits. In **Chapter 6**, the objectives were to disentangle the genetic effects on dairy traits mediated and non-mediated by rumen microbial abundances, to

answer the question: what genetic control does the host have over the phenotype through the microbiota?

Chapter 7 integrates the main results to understand the impact of the rumen microbiota on the phenotype, the degree of host genetic control over the microbial composition, and how to implement the holobiont concept in the animal breeding, while objectifying methodological limitations of this work. A general conclusion and perspectives are developed in **Chapter 8**.

Chapter 2
Scientific background

This chapter presents a synthesis of the available literature on the holobiont concept and its potential integration into quantitative genetics theory. For this purpose, the three key elements that make up the genome-microbiome-phenotype system are presented. First, the host genome and microbiome are described, with special emphasis on ruminants and the rumen ecosystem, composition and measurement, followed by the phenotype represented by milk traits.

Before starting, it is necessary to define some terms based on Marchesi and Ravel (2015) that are used throughout this text:

- The word **microbiome** is of ancient Greek origin, from micro (**μικρός**, small) and biome, derived from the Greek word bios (**βΐος**, life). The microbiome refers to a microbial community occupying a reasonable, well-defined habitat that has distinct physiochemical properties and includes the microbes and their theater of activity (Whipps *et al.*, 1988).
- The word **microbiota** is of ancient Greek origin, from micro (**μικρός**, small) and biota (**βιοτα**, the living organisms of an ecosystem). The microbiota consists of all the living members that make up the microbiome and belong to different kingdoms.
- The **metagenome** is all the genetic information contained in the microbes that are members of the microbiota.

2.1 Holobionts

All animals establish symbiotic relationships with microorganisms. Thus, the animal is no longer conceived as an autonomous entity but as a holobiont that includes the host plus all its numerous associated microorganisms (Zilber-Rosenberg and Rosenberg, 2021). The relevance of symbiotic microbes in the host's life is well known, as they affect almost every aspect of the host, such as reproduction, health, fitness, and production traits in livestock species. These impacts could occur through influencing individual performance by enhancing fitness or expanding the range of environments in which the host can adapt and compete successfully (Henry *et al.*, 2021). However, a holobiont can also be considered a polygenomic entity that

includes the host genome and metagenome, *i.e.,* the hologenome as defined by Zilber-Rosenberg and Rosenberg (2008). The holobiont concept is especially relevant because phenotypic variation in a holobiont can be subject to selection operating through the transmission of the best combination of genes and/or microbes to the next generation.

However, how is this holobiont be transmitted to the next generation? The mechanism of transmission of the genome and metagenome will be detailed later; briefly, the host genome is inherited following Mendelian laws, and the metagenome is inherited mainly by physical transmission from parents to offspring and from the environment. Microbiota transmission is a relevant aspect because, with a high transmission fidelity, the beneficial associations between the host and its microbiome are maintained through generations by increasing the production (increasing the phenotypic mean) but with a low fidelity, the loss of these associations could be helpful under changing environmental conditions (increasing phenotypic variance) (Bruijning *et al.*, 2022). Thus, animal selection and evolution can no longer be considered separately from the microbes, necessitating analysis of the holobiont that includes the host, microorganisms and their interactions. For this purpose, quantitative genetics theory can be adapted to integrate the holobiont into genetic models.

2.2 Integrating the holobiont in quantitative genetics

In quantitative genetics, the most common genetic models are designed to be used with metric characters that show a continuous expression (*e.g.,* birth weight, milk yield, litter size, *etc.*). In animal production, this phenotypic expression is supposed to be explained by the simultaneous effect of many genes with a small effect as well as variation arising from nongenetic causes, *i.e.,* environmental effects (Falconer, 1986). However, for certain traits, major genes have been identified that explain a large part of the phenotypic variance, such as the suppressor of cytokine signaling 2 (Socs2) gene that is mainly associated with somatic cell count (SCC), which explains 12% of the phenotypic variance, but also with milk yield, size and weight in dairy sheep (Rupp *et al.*, 2015).

According to the classic genetic theory proposed by Fisher in 1919, the measure of performance of a trait in an individual, *i.e.,* phenotypic value (y_i) is explained by an overall mean (μ), its genotypic value (g_i) and an error term (e_i):

$$y_i = \mu + g_i + e_i \quad (1)$$

The genotypic value of an individual is equal to the sum of the additive (u_i), dominance (d_i) and epistatic effects (ε_i), as defined Falconer (1986):

$$g_i = u_i + d_i + \varepsilon_i \quad (2)$$

However, in general, the genotypic value is equal to the additive effect due to the small magnitude of the dominance and epistatic effects. Then, Equation 1 becomes:

$$y_i = \mu + u_i + e_i \quad (3)$$

where u_i is the breeding value of the ith individual. The breeding value represents only the part of the genotypic value of an individual that can be transmitted to its progeny. However, since it cannot be observed, it is estimated by genetic evaluation using the phenotype, genealogy and, more recently, genomics from polymorphisms identified in the genome.

The advent of next-generation sequencing (NGS) technologies after the 2000s made panels of thousands of single-nucleotide polymorphisms (SNPs) available that, together with the study by Meuwissen *et al.* (2001), encouraged the use of genomic information for genetic evaluations. Given the large number of SNPs in the panels and their known positions in the animal genome, some are expected to be tightly and nonrandomly associated, *i.e.,* in linkage disequilibrium (LD) with quantitative trait loci (QTL) that largely affect the phenotypic variance of a trait. Then, genomic evaluation consists of predicting the breeding values of individuals based on the LD between SNPs and casual loci and the genomic relatedness between observed and unobserved individuals (Meuwissen *et al.*, 2001).

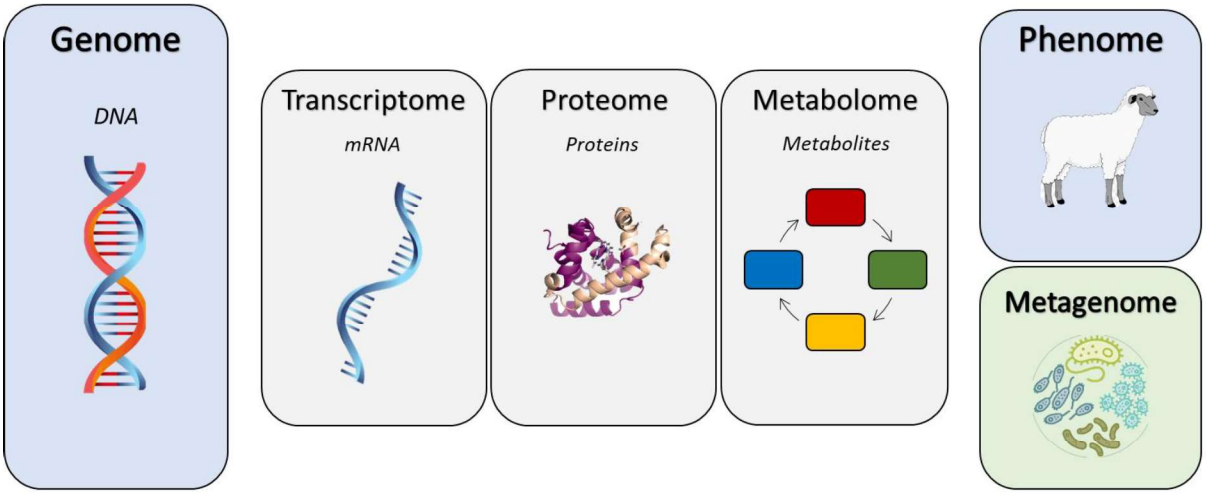


Figure 1. Multi-omics system from the genome to phenome (through intermediate omics: transcriptome, proteome and metabolome) and metagenome.

More recently, NGS technologies have made it possible to combine high-dimensional multiomics data, *e.g.,* transcriptomics, proteomics, and metabolomics, as predictor variables for a more comprehensive modeling of phenotypes. In addition, modeling the complexity of and relationships between multiomics data improves the understanding of the mechanism or causal relationship of complex trait architecture (Ritchie *et al.*, 2015). All of the abovementioned omics measurements are considered intermediates between deoxyribonucleic acid (DNA) action and phenotype expression, as shown in Figure 1. However, metagenomics, *i.e.,* microbial genome information, is not a strictly intermediate omics measurement because, from a biological point of view, it has complex interactions with the phenotype, host genome (Figure 1) and environment. The relevance of the metagenome is given by its capacity to expand the host genetic repertoire and influence the proportion of phenotypic variance explained by genetic effects, classically referred to as heritability. Thus for example, the gut microbiome is considered to contribute to the total heritability of a trait (Sandoval-Motta *et al.*, 2017; Henry *et al.*, 2021).

The first implementation of metagenomics for phenotype prediction in livestock species was described by Ross *et al.* (2013) to predict methane production in cows. Since then, the use of multiomics data has been expanded to predict complex phenotypes in animals and plants (Riedelsheimer *et al.*, 2012; Guo *et al.*, 2016; Z. Li *et al.*, 2019; Morgante *et al.*, 2020) as well as improve the accuracy of the estimated breeding values (EBVs) (Hayes *et al.*, 2017; Weishaar *et al.*, 2020; Christensen *et al.*, 2021; Zhao *et al.*, 2022). These models for genomic prediction

must consider the genetic and microbiota effects on phenotype in addition to the genetic control on the microbial composition. Therefore, more complex models that include the interactions between each element of the genome (G) – microbiome (M) – phenotype (P) system should be considered.

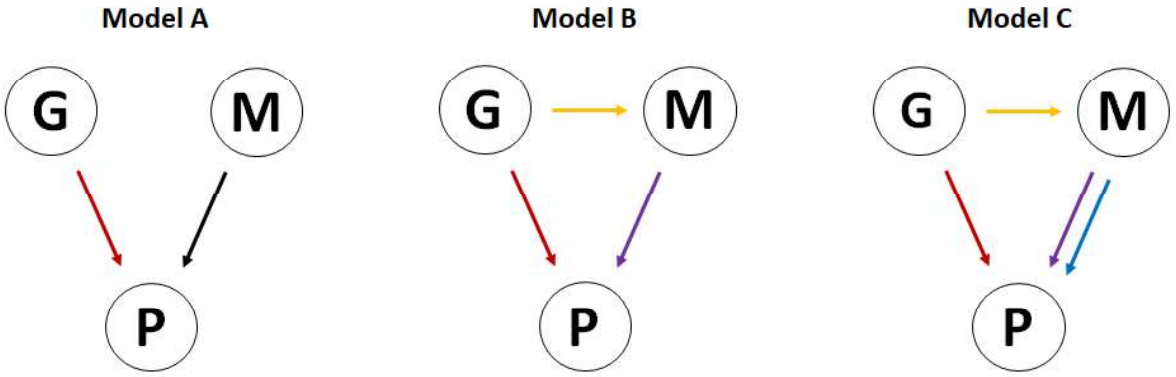


Figure 2. Models including gut microbial data. Host genome (G), microbiome (M), and phenotype (P). A red arrow from G to P indicates that there is a subset of SNPs that impact on P. A black arrow from M to P indicates there is a subset of microbial abundances that influence P, when this arrow is violet it is because the subset of microbial abundances is controlled by host genetics, if it is not controlled by host genetics but it is transmissible it is a blue arrow from M to P. A yellow arrow from G to M indicates that there is a subset of G that influences a subset of abundances in M.

Three models incorporating gut microbial data are presented in Figure 2. Model A considers that the host genome and microbiota impact the phenotype with no interaction between them. This is the most widely used model for phenotype predictions, typically including gut microbiome data (Camarinha-Silva *et al.*, 2017; Difford *et al.*, 2018; Rothschild *et al.*, 2018). Model B considers that the host genome and microbiota influence the phenotype and the host genome has a control over a subset of microbial abundances. This model was recently proposed to estimate the genetic effect on phenotypes mediated by the gut microbiome (Weishaar *et al.*, 2020; Tiezzi *et al.*, 2021) and intermediate omics measurements (Christensen *et al.*, 2021). Finally, Model C is similar to Model B but considers that the impact of the microbiota on the phenotype is explained by subsets of microbial abundances with and without host genetic control.

2.2.1 Host genome on phenotype (G → P)

The phenotype is classically explained by the genotype and environmental effects. The expression of a phenotype is determined by the action of many genes with small effects as well as by the intermediate steps shown in Figure 1, such as gene expression, protein expression and posttranslational modifications, which are between the gene action and the phenotype (Ritchie *et al.*, 2015). However, until a few years ago, animal models included genetic effects only without the intermediate factors.

In quantitative genetics, the contribution of the host genome to the phenotype is estimated as the heritability. Heritability in the broad sense is the proportion of phenotypic variance of a given trait in a given population that is due to genetic differences among individuals and, in a narrow sense, it expresses the extent to which phenotypes are determined by the genes passed on by the parents (Falconer, 1986). Then, as heritability is an estimation of the strength of the relationship between a phenotype and breeding values, when the heritability is high (*e.g.,* in sheep: carcass and growth traits, milk fat and protein contents), there is a strong relationship between the phenotype and breeding values; however, when the heritability is low (*e.g.,* in sheep: calving interval and fertility), the phenotype and breeding values have little bearing on each other (Bourdon, 2000).

2.2.2 Microbiota on phenotype (M → P)

Microorganisms in symbiosis with their host will contribute to the phenotype through microbial functions that are crucial for phenotype expression; the presence and/or abundance of microbes in the ecosystem will provide information about its relevance to the host (Sandoval-Motta *et al.*, 2017). In the literature, the host-associated microbiome is known to influence phenotypes, from health traits in humans and animals (Gebreyesus *et al.*, 2020; Rothschild *et al.*, 2022) to production traits in livestock species (Difford *et al.*, 2018; Buitenhuis *et al.*, 2019; Hess *et al.*, 2020; Vollmar *et al.*, 2020; Khanal *et al.*, 2021; He *et al.*, 2022). However, for most traits, the biological mechanism through which the microbes affect the phenotype is not known and sometimes does not have a direct effect on the expression of the phenotype.

In quantitative genetics, the microbiota is considered not only as a trait for which we could estimate host genetic variability but also complementary information contributing to the phenotype. The latter led evolutionary biologists to propose the concept of general heritability (Danchin *et al.*, 2011). In this regard, to make a parallelism with heritability, Difford *et al.* (2018) defined the contribution of microbiota composition to the phenotype variability as

microbiability, which is the proportion of phenotypic variance due to the gut microbial composition. In addition, to identify the microbes that contribute the most to the phenotype, as in a genome-wide association study (GWAS) using the host genome, the use of the microbial ecosystem through a microbiome-wide association study (MWAS) was proposed (Wang and Jia, 2016).

2.2.3 Host genome on microbiota (G → M)

The host genetic control over the microbiota could be estimated by each individual microbial abundance or by the microbial community as a whole.

Host genetic control over individual microbial abundance

Following the classic heritability estimated for each phenotype (G → P), it is possible to do the same with the microbial abundance as the phenotype determining the proportion of microorganisms whose abundances are genetically controlled by the host (G → M) and then evaluate whether these microbes have an effect on the phenotype of interest (M → P). In livestock species, the literature showed a small group of microbes under host genetic control (<16% total microbial community), with low to moderate heritability depending on the species and sequencing technologies used (Camarinha-Silva *et al.*, 2017; Difford *et al.*, 2018; Wen *et al.*, 2019; Grieneisen *et al.*, 2021). However, there are no reported results for the microbial effect on phenotype using separately microbiota genetically controlled and nongenetically controlled by the host.

Model C considers that the impact of the microbiota on the phenotype (M → P) is explained by subsets of microbial abundances that are genetically and not genetically controlled by the host. As mentioned, a low proportion of microbes are commonly under host genetic control due to imperfect transmission measured by the transmission fidelity, colonization from the environment and the number of microbial generations within one host generation (Bruijning *et al.*, 2022). In that scenario, Model C, which includes the different transmission modes of the microbiota, could improve the understanding of the genome-microbiome-phenotype system. The microbiota is characterized by vertical transmission from parents to offspring, which in livestock species is mainly transmitted from the mother, and by horizontal transmission from the group of comates sharing the same environment. The proposed Model C is inspired by the transmissibility model presented by David and Ricard (2019), which estimates the transmissible value accounting for all inheritance sources, such as additive genetic effects and nongenetic effects (*e.g.,* epigenetics, gut microbiota, cultural effects). In that way, genetic selection

classically based on EBVs could include genetic and nongenetic effects and become a selection on transmissible values, which allows for us to obtain maximal progress on selection.

Host genetic control over the microbial community

Another way to estimate the host genetic control over the microbiota is to use all microbial abundance data. Recently, models such as Model B have been proposed to quantify the impact of the indirect genetic effect on the phenotype mediated by the gut microbiome (Weishaar *et al.*, 2020) or intermediate omics (Christensen *et al.*, 2021). Using these models, unlike what is presented for each individual abundance, we first estimate the effect of the microbiota on the phenotype (M → P) and then the proportion of the microbial community explained by host genetics (G → M). A recent study estimated microbiabilities for milk yield and volatile fatty acids (VFAs) in cows and calculated the contribution to the phenotype of rumen bacteria controlled and not controlled by host genetics (Zang *et al.*, 2022). The authors determined a greater contribution from genetically controlled bacteria than nongenetically controlled bacteria, but this could be associated with the calculation of an average microbial effect based on the number of bacteria in each subset. To date, there are no results reported in livestock species using the methodology proposed by Christensen *et al.* (2021).

Having described the general concepts, the following three sections present each element of the genome-microbiome-phenotype system in detail.

2.3 Host genome (G)

Most of the references used in these paragraphs are taken from Bourdon (2000) and Falconer (1986).

The host genome is all the genetic information of an organism contained in long strands of DNA called chromosomes. In livestock, the hosts are commonly higher organisms, which are diploids with two copies of each chromosome, one chromosome inherited from the father and the other from the mother. In each chromosome, there are genes, and segments of DNA that are transcribed into ribonucleic acids (RNAs) to encode proteins, located at specific chromosomal sites called loci. At each locus is a pair of genes, one gene on the paternal chromosome and one on the maternal chromosome. The genes could have alternative forms called alleles. The combination of alleles at a single locus or at a number of loci is called genotype. Therefore, a genotype that has identical alleles is called a homozygote, while it is heterozygote if the alleles are different.

In sexual populations, individuals do not necessarily produce offspring whose genotypes match their own. Following Mendel's laws of inheritance, prior to reproduction, sexual individuals produce haploid gametes with a single copy of the chromosome by cell division called meiosis, which is known as the first Mendelian's law of segregation. Then, the second Mendelian law refers to genes that assort independently during meiosis if all possible gametes are formed in equal proportions. However, there is an exception to Mendel's second law that it is caused by the linkage. Two loci that are closed in the same chromosome are linked and thus have a greater chance of segregating together during meiosis unless recombination occurs. The diploid state is restored when gametes from two parents fuse to form a zygote. Therefore, each animal receives half of its genes from its father and half from its mother. However, as shown in Figure 3, it is not possible to predict which half of the genes are passed on to the offspring. This is a chance factor, also called Mendelian sampling, which refers to the random sampling of parental genes.

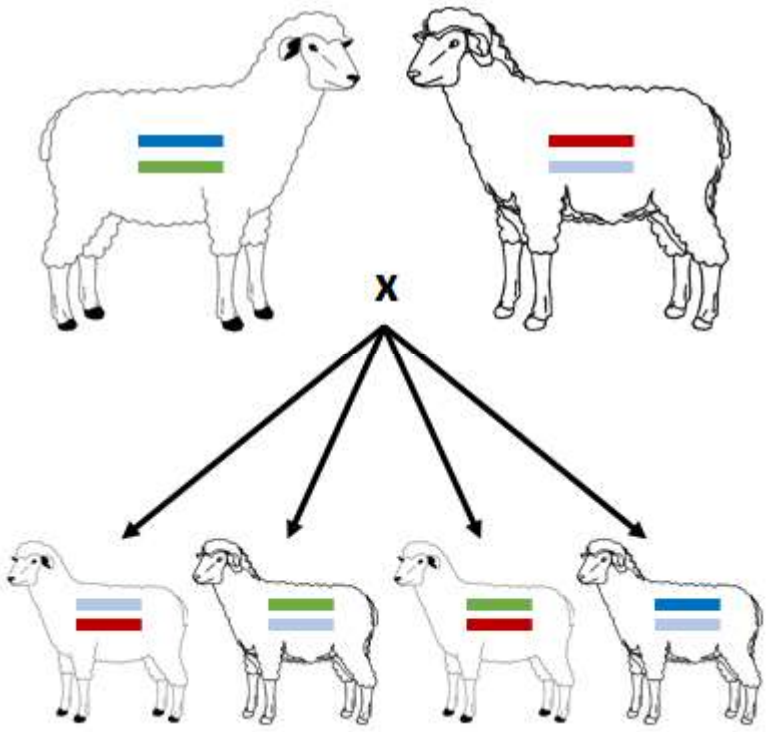


Figure 3. Family of sheep illustrating how offspring receive their genes from each of their parents.

Mendel's principles are explained at the individual level, but in animal breeding, the focus is on populations. How can the genetic content of the genome in a population be described? We

can do this through allele frequencies, which are the relative frequency of a particular allele in a population, and genotypic frequencies, which are the relative frequency of a particular genotype at a locus. Allele frequencies can change in a predictable direction and amount through systematic processes, such as selection, mutation and migration; they can also change in a predictable amount through a dispersive process, such as random drift arising in small populations due to the effect of sampling. However, it is generally well known that there is a predictable relationship between allele and genotypic frequencies over generations based on the Hardy-Weinberg equilibrium that occurs in a large population in which selection, mutation and migration do not act.

Selection is a process that determines which individuals become parents, how many offspring they produce, and how long they remain in the population. The objective of selection is to increase the frequency of favorable alleles of genes associated with the phenotype of interest. Migration is the introduction of individuals from another population with different genetic compositions, which produces a change in the allele frequency. Mutation is a process that alters DNA and generates new alleles. Mutation is rare, but if persistent during successive generations, it can produce important effects on the population. Then, the genetic diversity of a population is a balance between selection and random drift, which eliminates variation and renews through mutation and migration.

In summary, the genome is all the genetic information contained in chromosomes and transmitted to the next generation during sexual reproduction, based on Mendel's laws. Is the genetic content of microorganisms in symbiotic communities determined by the same processes as in the host genome? In what way is the metagenome transmitted?

2.4 Microbiome (M)

2.4.1 Metagenome

In general, microbial genetic information is contained in a single circular chromosome as a haploid organism, and microbes have generational intervals shorter than that of their host (Bruijning *et al.*, 2022). The enormous diversity of microbial genes, ~1000 times greater than in the human genome, enriches the host genome, giving the holobiont enormous genetic potential to adapt to changing environmental conditions. However, the microorganisms that compose the host microbial communities represent only a very low proportion (less than 30%) of the global species diversity in different ecosystems (Zilber-Rosenberg and Rosenberg, 2021). The considerable individual variation in gut microbiota composition is caused by the great

diversity of existing species as well as numerous external factors (*e.g.,* diet, age, host genetics). The large variation in microbial composition imposes relevance to the identified group of microbes present in most individuals, which could be associated with animal fitness. The core microbiome is commonly defined as a group of microbial taxa or genes that are shared by the majority of individuals in a population (Shapira, 2016). In addition, the core microorganisms have genes that could be involved in common keystone functions to maintain the symbiotic relationship with their host (Zilber-Rosenberg and Rosenberg, 2021). However, noncore microbes are also relevant to host possibilities for variation and adaptation to new environments.

Genetic diversity in host microbial communities is based on the presence/abundance of a multitude of microbial species with or without differences in their gene content. This microbial genetic diversity varies independently of its host through different processes, such as recombination and mutation, particularly by, as proposed Zilber-Rosenberg and Rosenberg (2021) by:

1) Acquisition of novel microbes from the surrounding environment.
2) Increase or decrease in the number of specific microbial groups. This is based on the dynamic nature of the microbiota and their multiplication or decrease as a function of local conditions.
3) The transmission of genes between microorganisms, *i.e.,* horizontal gene transfer, mediated by transposons, plasmids, genomic islands and bacteriophages.

To make use of this metagenomic diversity and transmit it to subsequent generations, it is necessary to know the mechanisms by which the microbiota are transmitted. In general, microbial communities have complex inheritance processes that can be vertical and/or horizontal (Danchin *et al.*, 2011; Sandoval-Motta *et al.*, 2017). Figure 4 illustrates the most common methods of transmission in ruminants. As is well known, vertical transmission consists of the passage of microorganisms from one generation to the next (*e.g.,* parents to offspring). Vertical transmission was shown in ruminants during parturition and/or the suckling period (Sandoval-Motta *et al.*, 2017), as the mother has direct physical contact with the newborn through the genital tract or by licking, and through contact with the udder skin during the lactation period. However, horizontal transmission is also possible between animals of the same generation through physical contact and/or contamination of the shared environment through feces and sputum (Dehority, 2004).

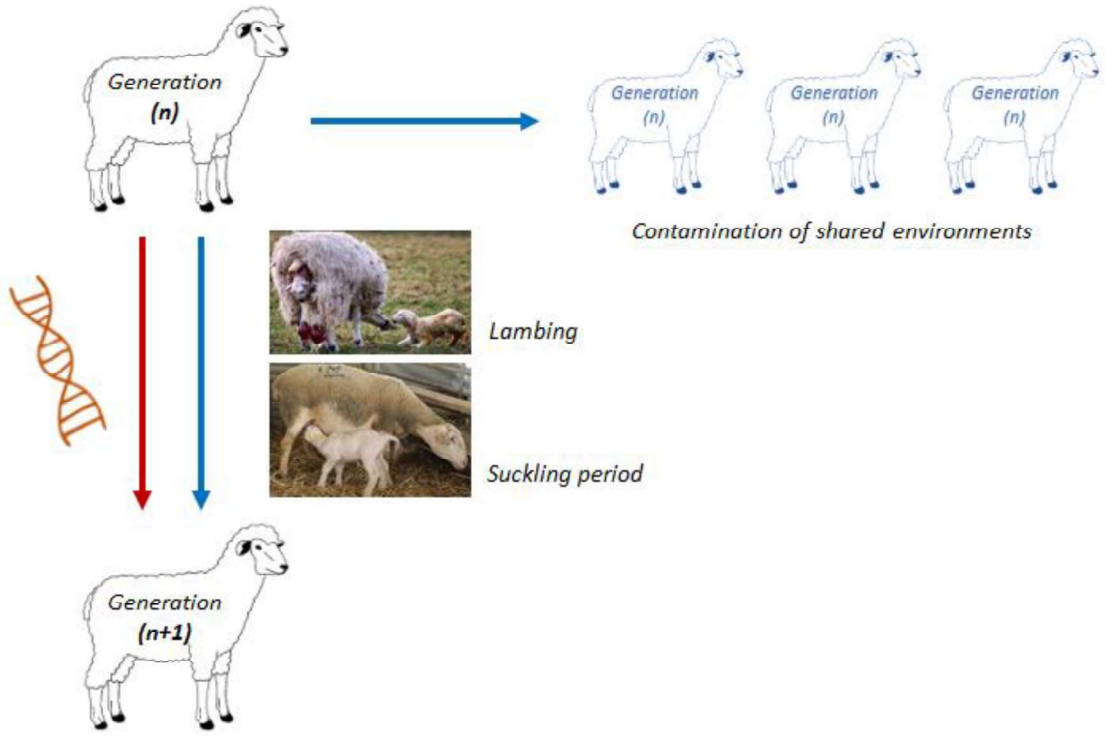


Figure 4. Main vectors of transmission for genetic (red arrow) and microbial (blue arrow) inheritance within generation (n) and between generations (n and n+1).

These concepts apply to all symbiotic communities of different hosts but are exemplified in ruminants here since this work focuses on them and their symbiotic community located in the digestive tract, more precisely in the rumen. The following sections are devoted to presenting the particularities of the rumen microbial community and how microbial genetic information can be obtained with sequencing technologies.

2.4.2 Digestive ecosystem of ruminants

Ruminants are holobionts because of the symbiosis between the host and its ruminal microbiota, where both parts are very beneficial. The host provides necessary living conditions for microbial growth, and the microbiota is responsible for the digestion of food, especially fibers, and nutrient synthesis through enzymes not produced by any higher animal species (Mazza Rodrigues, 2016).

Ruminants are polygastric herbivores with three aglandular forestomaches and a glandular stomach. Thus, in ruminants, the feed first goes into the forestomaches to be available for microbes, which digest most of the dietary components, and is then subjected to the action of enzymes produced by the digestive tube and attached glands. The main feature that allows for

fiber fermentation is a voluminous fermentative chamber represented by the rumen, inhabited by a wide microorganism population that ferments organic matter, mainly parietal carbohydrates.

Forestomach biotope

The ruminant digestive system consists of a long muscular tube that extends from the mouth to the anus and a group of glands attached to this digestive tube. In particular, the feed passes successively through three forestomaches: the rumen, reticulum and omasum, and finally through a glandular stomach called the abomasum. The proportions of digestive compartments in terms of volume in large ruminants, such as bovines, are 80% of rumen (120-300 liters), 5% of reticulum, 8% of omasum, and 7% abomasum; and in small ruminants, such as sheep, they are 75% of rumen (15 liters), 8% of reticulum, 4% omasum and 13% abomasum (Bertan Membrive, 2016).

The digestion process starts when the animal ingests feed, the ruminants chew it very roughly, the food is soaked with saliva, and this poorly chewed feed is transported through the esophagus to the rumen where it mixes with the preexisting rumen content. Then, rumination allows for regurgitation and a second strong chewing of the large food particles, reducing the size of feed particles, producing saliva rich in buffer substances, and accelerating the feed passage through the reticulo-omasal orifice. Once returned to the rumen, the fragmented feed is digested by the microbes in the rumen with subsequent gas production (mainly methane and dioxide carbon). The contents of the rumen are heterogeneous, composed of a complex mass of digesta, which may float or sediment at the bottom depending on the density of the feed, a liquid fraction with microbial cells and fine feed particles, and a free gas cap at the top. In the rumen, two elements are crucial to fulfill its digestive functions: 1) its motor activity consisting of synchronized contractions of propulsive and mixing movements to perform functions of mixing, regurgitation, eructation of fermentative gases as well as moving the feed bowl to the next forestomach; and 2) its epithelial structure consisting of highly irrigated mucosal tissue that allows for the absorption of fermentative products, such as VFAs and ammonia, through the wall. In addition, the epithelial lining of the reticulum is raised into folds forming a honeycomb structure, while the rumen itself is lined with papillae of varying size and shape (Dehority, 2004).

The rumen is a large chamber in which ingested feed is subjected to microbial digestion. The rumen is an ideal microbial habitat because the conditions allow for the survival and growth of microorganisms. The normal conditions in the rumen based on Dehority (2004) are 1) a

temperature of 38 to 40 °C; 2) a normal pH of 6.4 to 6.6, which is fairly well buffered between 5.5 and 7.0; 3) osmotic pressure, hypotonic to blood plasma except shortly after feeding; 4) dry matter content that is relatively constant under most feeding conditions, generally in the range of 10 to 13%; and 6) reducing redox potential in connection with anaerobiosis, ranging from -150 to -350mV.

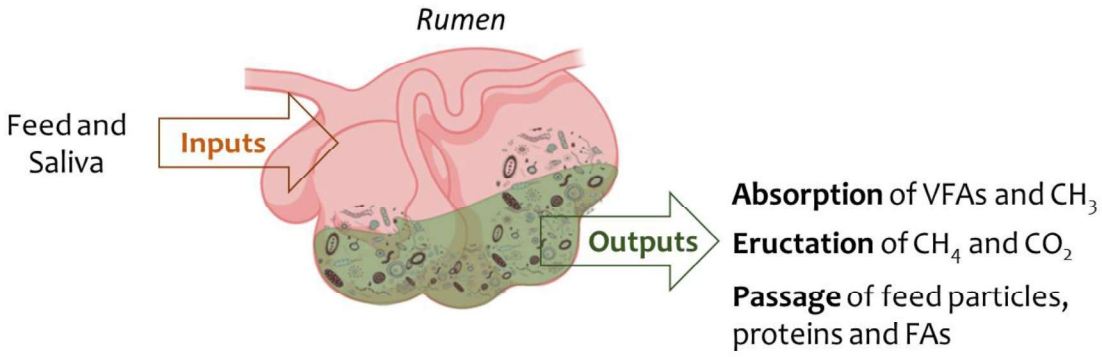


Figure 5. Rumen as a continuous fermentation chamber. VFAs Volatile fatty acids, CH_3 ammonia, CH_4 methane, and CO_2 carbon dioxide.

The rumen ecosystem is similar to a continuous culture system that functions as a biological fermentation unit. As shown in Figure 5, the fermentation chamber has inputs such as the more or less continuous availability of feed and outputs that are the elimination of end products by absorption, the eructation of gases and the passage to the small intestine of degraded particles, microbial proteins and fatty acids (FAs) (Nagaraja, 2016).

Ruminal microbiota

The rumen is inhabited by a multitude of symbiotic microorganisms that colonize, grow, and live in harmony with the host. That microbial population in the rumen includes members of all three domains, Eubacteria or Bacteria, Archaea represented by Methanogens, and Eukarya represented by Protozoa and Fungi, and viruses as bacteriophages. The most abundant are bacteria (10^{10} to 10^{11}/ml), representing 40-90% of the total microbial mass, followed by methanogenic archaea (10^{6} to 10^{8}/ml), anaerobic fungi (10^{5}/ml) and ciliated protozoa (10^{5}/ml) (Nagaraja, 2016).

Bacteria are not only the most abundant but also contribute the most to energy production from the fermentation of plant carbohydrates to VFAs, such as acetic acid (C2:0), propionic acid (C3:0) and butyric acid (C4:0). They also contribute to microbial protein synthesis and fat lipolysis followed by unsaturated fatty acid biohydrogenation, which influence the long-chain

fatty acid (LCFA) composition in the rumen (Hungate, 1966; Nagaraja, 2016). Methanogens, as members of Archaea, constitute approximately 2 to 4% of the total microbial biomass in the rumen, and most species can reduce carbon dioxide to form methane. Protozoa are anaerobic microbes with a low contribution to rumen fermentation but play a key role in maintaining the anaerobic condition in the rumen and in regulating bacterial populations. Finally, the fungal contribution in the rumen is not well known, and bacteriophages are viruses that infect bacteria; thus, they can control the bacterial population in the rumen (Nagaraja, 2016).

Bacteria

The main members of the rumen ecosystem are bacteria, which are unicellular prokaryotes, most of which are strict anaerobes. Bacteria that participate actively in the digestion of feedstuff in the rumen can be found free-floating in the rumen fluid (~30%) or attached to feed particles (~70%). This is relevant for sampling, since free-floating bacteria are part of the liquid fraction and bacteria attached to feed particles are part of the solid fraction of rumen samples. However, in practice, liquid rumen samples are composed of a liquid phase and small solid particles. Therefore, it represents the composition of the microbiota well (Ramos-Morales *et al.*, 2014), even though sampling of the liquid fraction obtains ~30% of the total microbial biomass in the rumen, and some differences in proportions of the *Prevotellaceae* and *Lachnospiraceae* families were reported by Henderson *et al.* (2013). A very small fraction (~1%) of the bacteria is not reached by rumen sampling as it is attached to epithelial cells, and it is speculated that their function is to maintain anaerobic conditions in the rumen.

Ruminal bacteria vary in their enzymatic specificity; however, most of them utilize monomers or oligomers, which are liberated from plant material by microbial enzymatic hydrolysis of polymers, including starch, pectin, cellulose, hemicellulose and lipids (Hobson and Stewart, 1997). In the rumen, there are keystone functional groups of microorganisms, such as those based on the ability to hydrolyze polysaccharides: fibrolytic and amylolytic bacteria. Fibrolytic bacteria include 1) cellulolytic species that are able to hydrolyze cellulose, the most effective being *Ruminococcus flavefaciens* and *R. albus*, *Fibrobacter succinogenes*, and to a lesser extent *Butyrivibrio fibrisolvens* and *Eubacterium celullosolvens*; 2) hemicellulolytic species that are able to hydrolyze hemicellulose, including *Butyrivibrio fibrisolvens* but also *Ruminococcus flavefaciens* and *R. albus*; and 3) pectin-fermenting species, especially *Butyrivibrio fibrisolvens*. Amylolytic bacteria are starch digesters, such as *Ruminobacter amylophilus* or *Succinomonas amylolytica*. However, this distinction in terms of substrate utilization is relative because some bacteria are capable of using many polymers, such as *Prevotella ruminicola* which uses almost

all of them except cellulose and is generally classified as amylolytic. This redundancy of the rumen microbial community is based on the smaller number of functions than the microbial species involved. Weimer (2015) defined redundancy as the overlapping physiological capabilities of multiple microbial taxa, where a large number of species can contribute to the degradation of the same substrate and production of the same nutrient.

Historically, much of the information on the composition and taxonomy of the ruminal bacterial community has been obtained by cultivation-based methods. However, less than 1% of bacteria are cultivable under laboratory conditions due to their anaerobic nature. This stimulated the development of high-throughput sequencing technologies, increasing the identification of bacterial species in ecosystems such as the rumen (Santos *et al.*, 2020).

2.4.3 Sequencing technologies for microbiome study

Advances in sequencing technologies enable decoding of the DNA composition of biological life and the uncultured organisms of host-associated microbial communities. However, is has only been possible since the DNA structure was discovered by Watson and Crick in 1953. The sequencing methods consist of determining the order of the nucleotides adenine, thymine, cytosine, and guanine on a DNA strand from a sample.

The first sequencing technology developed by Sanger *et al.* in 1977 was based on the addition of chain-terminating and radioactively or fluorescently labeled dideoxynucleotides to perform the sequencing of a DNA strand complementary to the interrogated template strand. However, given that only one sequencing reaction could be performed at a time, this methodology had limited throughput. To increase sequencing capabilities, new technologies known as NGS were introduced between 2004 and 2006. The major feature of NGS is the combination of high throughput and single-stranded DNA sequencing. These sequencing technologies involve massively parallel short-read sequencing, in which millions of individual sequencing reactions occur in parallel. However, due to the nature short-read technologies, the reassembly process of long stretches of DNA is challenging. Third-generation sequencing overcomes the limitations of short-read technologies, achieving read lengths upward of 10 kilobases (Hu *et al.*, 2021).

The next sections provide an overview of the NGS classified into short- and long-reads, details about the benefits and limitations of each method, and descriptions of how these technologies are used in microbial community sequencing.

Short-read sequencing

Short-read NGS, or second-generation sequencing, has the common feature of massive sequencing of short DNA fragments (75-700 base pairs (bp)), which are clonally amplified and sequenced in parallel. These technologies include two consecutive steps, clonal amplification and the sequencing. Clonal amplification consists of solid-phase amplification of DNA fragments to produce enough detectable signals during sequencing. Depending on the technology used, the amplification could be by polymerase chain reaction (PCR) emulsion (Ion Torrent and 454 pyrosequencing) or by bridging PCR (Illumina), generating millions of DNA fragments. The sequencing process consists of the detection of each of the nucleotides of the DNA chain, which can occur by ligation or synthesis. Short-read sequencing technologies are classified based on the sequencing process, such as sequencing by ligation (SBL) or sequencing by synthesis (SBS).

SBL consists of DNA fragments being ligated to a fluorescent label, and the color emitted by the label is recorded. Then, from the color output, the DNA nucleotides are encoded. NGS that implements this approach includes sequencing by oligonucleotide ligation and detection (SOLiD) technology, with high accuracy of reads but relatively short reads and long run times (Table 1).

However, SBS involves DNA polymerase-dependent nucleotide incorporation on the extended DNA chain, which produces a detected signal based on a fluorophore or an ionic concentration change. Three technologies implement the SBS approach: Ion Torrent, 454 pyrosequencing and Illumina. Ion Torrent and 454 pyrosequencing are based on single-nucleotide addition, which consists of a single signal to detect the incorporation of deoxynucleotides into the elongated DNA strand. However, the main difference between these two technologies is the way in which they detect nucleotide incorporation because 454 pyrosequencing is based on pyrophosphate release, and Ion Torrent is based on hydrogen ion release, which generates a change in pH in solution. Illumina technology is based on cyclic reversible termination using a terminator molecule to prevent elongation (Goodwin *et al.*, 2016).

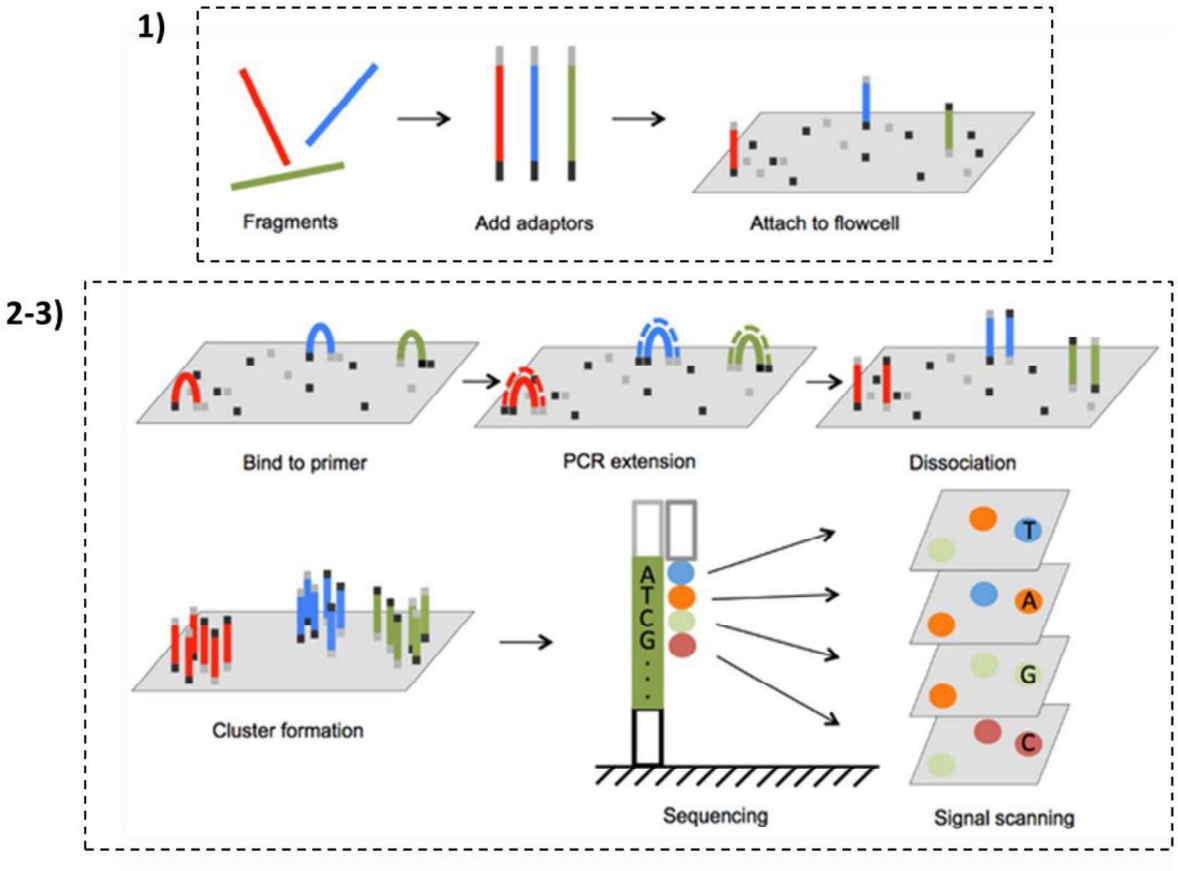


Figure 6. Illumina sequencing process. 1) Sample preparation, and 2-3) cluster generation and sequencing (Adapted from http://en.biomarker.com.cn/platforms/illumina).

Illumina sequencing consists of three steps, as shown in Figure 6: 1) sample preparation, 2) cluster generation, and 3) sequencing and signal scanning. First, adapters are ligated to both ends of extracted single-stranded DNA fragments. Then, these adapters are fixed to the complementary adapters placed on a solid plate. Second, the single-stranded DNA fragments attached to the solid plate are amplified by PCR bridge amplification, which creates several identical copies of each DNA fragment by several cycles of strand synthesis and double-stranded denaturation, generating a group of DNA fragments called clusters. Third, SBS uses four labeled reversible terminators, primers and DNA polymerases to progressively reconstitute the second DNA strand of each amplified single strand within a cluster. The nature and number of nucleotides used in each step are obtained by capturing the emitted fluorescence of the cluster excited by the laser. Finally, computer programs translate these signals into nucleotide sequences.

NGS resolved most problems encountered in Sanger sequencing (Sanger *et al.*, 1977), with high-throughput and clonal sequencing dominating the market. Illumina has become the

standard sequencing technology and is most applied in microbial community studies (Santos *et al.*, 2020) due to its simplicity of use and the combination of read length with high accuracy (Goodwin *et al.*, 2016), as shown in Table 1. However, short reads have limitations, such as the run time of analysis and assembly difficulties, that prevent the identification of specific structural variants (Hu *et al.*, 2021).

Table 1. Description of short- and long-read next generation sequencing technologies.

	Read length (bp)	Chemistry	Error rate
Short-read sequencing			
SOLiD	75	SBL	~0.1%
Ion Torrent	400	SBL – SNA	1%
454 pyrosequencing	700	SBS – SNA	1%
Illumina	75-300	SBS – CRT	0.1 – 1%
Single-molecule real-time long reads sequencing			
Pacific Bioscience	10 – 15 kb		15%
Oxford Nanopore	>200,000		12%

SBL: sequencing by ligation, SBS: sequencing by synthesis, SNA: single nucleotide addition, CRT: cyclic reversible termination, bp: base pairs, kb: kilobase.

Long-read sequencing

Long-read NGS, or third-generation sequencing overcomes the limitations of second-generation sequencing. Genomes are extremely complex, with many long repetitive elements, alterations and structural variations, which cannot be well covered by short-read NGS. The long-read sequencing of several kilobases, such as by Pacific Bioscience and Oxford Nanopore (Table 1), is able to obtain a better resolution of complex genomes (Goodwin *et al.*, 2016). Currently, the most common long-read technologies are based on single-molecule real-time sequencing and do not require prior clonal amplification, which is an advantage in runtime compared with the short-read NGS. However, long-read technologies have some drawbacks,

such as higher error rates (~15%) and cost (750-1,000US$ per gigabase) than short-read technologies such as Illumina (45-300US$ per gigabase).

Characterization of microbial communities

Short- and long-read NGS are used to characterize microbial communities through three approaches: 1) amplicon sequencing, 2) shotgun sequencing, and 3) genotyping by sequencing (GBS).

1) Amplicon sequencing: example of 16S rRNA gene

The advances in sequencing technologies presented above were accompanied by the development of phylogenetic markers, such as the gene coding for the 16S rRNA of prokaryote ribosomes introduced by Woese and Fox in 1977.

16S rRNA is part of the small ribosomal subunit (30S) present in all prokaryotic cells. The gene encoding this molecule has distinctive features that make it suitable for taxonomic profiling: it is relatively short, approximately 1,500 bp, and it is composed of ten conserved regions and nine hypervariable regions, which are good properties to design primers (conserved regions), while hypervariable regions allow for phylogenetic assignment. However, some limitations to bacterial identification were detected, such as the variable number of copies in bacterial genomes, the taxonomic assignment discrepancies based on different hypervariable regions used, and the low taxonomic resolution beyond the genus level (Poretsky *et al.*, 2014). This limited taxonomic resolution is due to the analysis of short regions of 16S rRNA but can be improved by using long-read sequencing technologies that cover all regions of the gene (V1 to V9 region) (Santos *et al.*, 2020; Callahan *et al.*, 2021).

2) Shotgun metagenomic sequencing

Shotgun metagenomics is the untargeted (shotgun) sequencing of all microbial genomes present in a sample. This method is used to profile taxonomic composition, assess the functional potential of microbial communities and recover whole genome sequences. Shotgun sequencing consists of DNA extraction from all cells in a community, and all DNA is subsequently shared into tiny fragments that are independently sequenced with NGS. This results in DNA sequences that align to various genomic locations, including nonmicrobes, where some of the reads are sampled from taxonomically informative genomic regions (*e.g.,* 16S) and others are sampled from coding sequences that provide insight into the biological functions (Quince *et al.*, 2017). However, the application of this method has some limitations: it is expensive to sequence and

analyze a large number of samples, and the set of microbial genomes available as catalog references is biased toward pathogens and easily cultivable bacteria (Quince *et al.*, 2017).

3) Genotyping by sequencing

GBS, or restriction enzyme-reduced representation sequencing, is a sequencing technique that reduces the genome complexity by digesting genomic DNA with restriction enzymes, followed by NGS of the resulting fragments. The principle of this method is that sequencing only a specific fraction of 0.5 to 1% of any microbial genome is sufficient to capture a large part of the information necessary to determine the composition and diversity of the microbial community. Compared with previous metagenomic sequencing, GBS is not limited to detecting microbes with a particular gene, as with 16S rRNA gene targeted sequencing; thus, this method can be applied to study a wider variety of organisms, such as hosts, viruses, and fungi (Hess *et al.*, 2020).

To summarize this section on the microbiome, the metagenome is all the genetic information contained in the symbiotic microorganism population, which is of special interest in ruminants because of its impact on productive traits. The rumen microbial composition is measured by the use of NGS technologies, which provide a good representation of the whole community.

2.5 Phenotype (P)

Previously, it was expressed that the genome and metagenome have an impact on the phenotypes of interest; however, this interest is defined by the production system under study. As this work is focuses on dairy sheep, the phenotype of interest are the milk composition traits.

Milk is a complex biological fluid due to its multifaceted physical nature and the diversity of its molecular components, such as carbohydrates, lipids, proteins, vitamins and minerals. For this reason, milk of animal origin plays an important role in the diets of the world's population, especially children. World milk production by species corresponds to cattle (81%), buffaloes (15%), goats (2%), sheep (1%) and camels (0.4%). Sheep milk is mainly produced in Asia (46%), followed by Europe (29%), Africa (24%) and North and South America (1%). The countries bordering the Mediterranean and Black Sea have a quarter of the ewes and produce half of the world's sheep's milk (FAOSTAT, 2020). Regarding the ewe's performance, the milk yield of dairy ewes in Europe is more than double the world's average milk quantity.

In France, dairy sheep production is considered a relevant agricultural activity developed in three traditional regions exploiting local breeds (Figure 7): south of the Central Massif with the Lacaune dairy breed, the western Pyrenees with the Basco-Béarnaise, Manech Tête Noire and

Manech Tête Rousse breeds, and Corsica Island with the Corse breed. Currently, the dairy sheep sector in France has 1.6 million dairy ewes (FAOSTAT, 2020), and the production systems are intended for the production of 'Appellation d'Origine Protégée'-certified cheese, which gives it a higher market value (*e.g.,* the world famous Roquefort cheese produced with Lacaune milk) and, to a lesser extent, milk to produce a variety of other cheeses (Lagriffoul *et al.*, 2016).



Figure 7. Traditional dairy sheep production regions in France with their local breeds.

Milk production in France increased significantly from 1970 onward, rising by 3.6% per year, from 57 to 275 million liters, explained by the increase in flock size and individual milk yield. However, the crucial points that produced this development were the implementation of mechanical milking, the improvement of animal nutrition, and the adoption of artificial insemination and selection schemes with genetic evaluations (Lagriffoul *et al.*, 2016). In this section, the focus is on the last point: the genetic selection schemes.

The selection scheme is based on a pyramidal structure. A nucleus of approximately 20% of the ewes is under official performance recording, and every year, a group of rams are evaluated through progeny testing. After the selection of rams for artificial insemination and natural service, they are distributed in the population outside the selection nucleus. The selection objectives for each breed have evolved depending on the production context of each traditional

region and the payment of milk to the farmers by the industry. The five local breeds presented are currently undergoing a genetic evaluation program, but the Lacaune breed includes the largest number of dairy traits in the selection index called ISOL (of the French 'Index Synthétique Ovin Lait'). The selection criteria included in the genetic evaluation of the Lacaune breed based on Lagriffoul *et al.* (2016) and Barillet *et al.* (2016) are detailed below.

2.5.1 Milk quantity (1965-)

Milk quantity was included as a selection criterion in genetic evaluations from 1965 to 1985. This initiated an easy-to-measure trait during the official performance control, which allowed for inclusion of as many animals as possible and fast genetic progress. Milk quantity is mainly determined by the quantity of lactose, which is the most abundant carbohydrate in milk and which induces mammary uptake of water and dilutes the solid components. This is because lactose is synthesized by the mammary gland from blood glucose and is responsible for maintaining the osmotic equilibrium between blood and the mammary gland. However, it is now well known that selection based only on milk quantity for a long time is not a good practice because it produces a decay in milk solid components such as fat and proteins. Therefore, the Lacaune scheme modified its selection objective by selecting for milk quantity based on the milk solid components.

2.5.2 Fat and protein quantity and rate (1985-)

In France, ewe's milk is used entirely for cheese production, hence the importance of evaluating and including in the selection criteria the solid content composition of the milk, which is crucial for cheese yield and texture. However, it was not until 1985 that changes began to be implemented in the criteria with the progressive inclusion of the milk solids content. In the first step, through the use of a selection index, *i.e.,* combining traits with different economic weights, the quantity of fat and protein was included to improve the cheese quantity. In a second step, the index was adapted to include the fat and protein rates, which allowed for simultaneous improvement of the cheese quantity and yield. That change in the ISOL index allowed for countering of the effect of selection for milk quantity alone and valuation of the milk in terms of cheese production and payment to the farmers.

2.5.3 Somatic cell count, mammary morphology and scrapie disease (2005-)

After years of effectively selecting for milk quantity and later for milk composition, in 1990 there was a growing interest in including functional traits associated with udder health and morphology that were impaired by genetic selection. However, the health crisis due to scrapie

disease in sheep between 1990 and 2001 imposed a complete adaptation of selection objectives to rapidly eradicate this fatal disease from flocks in France. A major step in eradication was taken after identifying a prion protein or PrP gene associated with resistance to this disease. The selection schemes eliminated ewes with susceptible genotypes and use of only homozygous resistant rams. The most affected breed was Manech Tête Rousse in the western Pyrenees region, with an initial frequency of the resistant allele (ARR) of 16%, compared to 60% for the dairy breed Lacaune (Barillet *et al.*, 2016). Once the eradication of the sensible allele (VRQ) of scrapie disease has been achieved, all rams of the five breeds commercialized in France should be homozygous resistant (ARR/ARR) for the PrP gene.

Mastitis is an inflammation of the mammary gland, which in dairy sheep manifests itself mainly in a subclinical form without clinical signs, for which the only indicator is the increase in the somatic cell count in the milk. Thus, mastitis is not only a disease in the ewe but also impairs milk coagulation during cheese making. The incidence of mastitis in the flock can be decreased by reducing the number of infected animals (*e.g.,* culling infected ewes), limiting new infections in the flock (*e.g.,* hygiene in the milking process), and by genetic selection of mastitis-resistant animals. To select ewes for mastitis resistance, the criterion used is the milk SCC (transformed into the somatic cell score (SCS)) through lactation represented by the lactation somatic cell score (LSCS), for which there is evidence of efficacy after the use of divergent genetic lines selected for this trait in the INRAE Experimental Unit of La Fage (Rupp *et al.*, 2009). In 2005, mastitis resistance and udder morphology were incorporated into the ISOL index of the Lacaune dairy breed.

2.5.4 Fine milk composition

The impact of the inclusion of fat and protein contents in selection led to further research in this direction for the identification of novel phenotypes that better represent the fine composition of milk. Currently, the ISOL index does not include the fine milk composition because the industry does not pay for differential milk composition, and more information is needed on the impact on other milk traits in order to include them.

The following is a description of the traits that determine milk quality, such as milk FAs and proteins, based on Park *et al.* (2007), Léonil *et al.* (2013) and Raynal-Ljutovac *et al.* (2008).

Fatty acids

Lipids are one of the most characteristic components of ewe's milk, being twice as high as in cow's milk, and highly variable according to the stage of lactation. Lipids are present in a

globular form in an emulsion state, called fat globules, composed of triacylglycerols and a membrane of lipoproteins and phospholipids. The fat content is made up of 98% triacylglycerols and less than 2% phospholipids, sterols and fat-soluble vitamins. Triacylglycerols are composed of three FAs and one molecule of glycerol. A fatty acid (FA) is a hydrocarbon chain, comprising single or double carbon-carbon bonds, and with two groups at its ends, one methyl (-CH3) and one carboxyl (-COOH). FAs can be classified by different criteria:

- Based on the length of the carbon chain:
  - Short-chain FA (2 to 10 carbons)
  - Medium-chain FA (12 to 17 carbons)
  - Long-chain FA (18 or more carbons)

- Based on the degree of saturation of the carbon chain:
  - Saturated FA (SFA) with no double bond
  - Monounsaturated FA (MUFA) with one double bond
  - Polyunsaturated FA (PUFA) with at least two double bonds

- Based on the position of the first double bond on the carbon chain counting from the methyl end (-CH3):
  - Omega-3 (*n-3*) with a double bond at the 3rd carbon
  - Omega-6 (*n-6*) with a double bond at the 6th carbon

- Based on the configuration of the double bond: the *cis* configuration characterizes the isomers whose priority groups are on the same side with respect to the plane formed by the double bond (most FAs), *trans* otherwise.

In ruminants, milk FAs are derived from de novo synthesis in the mammary gland and absorption from the blood. New FA synthesis uses acetate (85%), β-hydroxybutyrate (10-15%) and propionate (<1%) as precursors and produces short- and medium-chain FAs. From the blood, it mainly takes up long-chain FAs, which can originate from diet or adipose tissue, part of which could be desaturated by the udder. In small ruminants, milk fat has a high SFA content (60%) represented mainly by short- and medium-chain FAs, while MUFAs (mainly oleic acids C18:1*n-9*) account for 20%, and PUFAs (mainly C18:2*n-6* to C18:3*n-3*) account for 5% (Park *et al.*, 2007). This fatty acid profile in milk is of great relevance to cheese quality because it is little affected by the cheese-making process. In addition, two characteristics, namely, the small

size of the fat globules and the high levels of short- and medium-chain FAs, provide sheep and goat cheese with beneficial effects on fat assimilation and energy supply for human consumption (Raynal-Ljutovac *et al.*, 2008).

Proteins

Proteins in milk are of high quality with well-balanced amino acids. Eighty percent are caseins, which are phosphoproteins of four types: alpha-S1, alpha-S2, beta, and kappa, and the remaining 20% are soluble proteins or whey proteins, which are mainly beta-lactoglobulin, alpha-lactalbumin, immunoglobulin, serum albumin and lactoferrin. All caseins, beta-lactoglobulin, and alpha-lactalbumin are synthesized by the mammary gland, and the other serum proteins are taken up from blood.

Caseins have different properties that allow for them to be organized into a stable complex called micelles. The casein micelle is composed of 92% caseins (alpha-S1, alpha-S2, beta and kappa in relative proportions of 3:1:3:1) and 8% minerals (90% calcium phosphate and 10% magnesium and citrate ions). The structure of the micelle according to the model proposed by Ono and Obata (1989) is an aggregation of two types of submicelles of variable composition. One is composed of alpha-S (alpha-S1 and/or alpha-S2) and beta caseins, and the other is composed of alpha-S (alpha-S1 and/or alpha-S2) and kappa caseins. The submicelles with more kappa casein are preferably located on the surface of the micelle, playing an important role in the colloidal stability of the micelle. The destabilization of this colloidal structure, which can be induced by the acidification of the milk or the action of proteolytic enzymes, is the basic process by which milk is transformed into cheese. Soluble proteins may impair cheesemaking because they do not coagulate under the effect of proteolytic enzymes, but their amino acid profiles are of interest with a high level of essential amino acids.

Sheep's milk is unique for its nutritional properties and has recently gained interest in human diets due to the presence of omega-3 and omega-6 essential FAs, as well as conjugated linoleic acid (CLA) (Collomb *et al.*, 2008). In addition, it has a higher concentration of butyric acid than that of other ruminants and fully covers the requirement for all essential amino acids (Mohapatra *et al.*, 2019).

2.6 Conclusion

In ruminants, the symbiotic microorganisms in the rumen play an important role in the life of the animals. However, to date, the holobiont concept has not been integrated into quantitative genetics. One reason is because the microbiome represents a type of high-dimensional data for

which there is no standard methodology of processing and analysis. On the other hand, there is little published information on the genetic control of the host on the rumen microbial composition and even less on the genetic associations between the rumen microbiota and phenotypes or the metabolic processes that explain the associations between them. Most of the work used cows and focused on complex traits such as methane emissions or feed efficiency. However, in sheep, there are no studies to date linking rumen microbiota and dairy traits.

In this thesis, we proposed to investigate a methodological aspect, such as whether a single rumen sample obtained through a gastric tube is representative of the microbial composition in the rumen of adult animals. Based on this, two relevant biological hypotheses are proposed: 1) the rumen microbiota influences milk composition and udder health traits, and 2) there is host genetic control over rumen microbiota.

Chapter 3
Materials and methods

This thesis is based on the MicroGenOL project of which two scientific groups of INRAE-GenPhySE, GeSPR (Genetics and Selection of Small Ruminants) and NED (Nutrition and Digestive Ecosystem), are part, which allow for work on the integration of the genome and metagenome. To fulfill the objectives proposed in this thesis and obtain the appropriate materials, the experimentation was carried out at the INRAE Experimental Unit (EU) of La Fage (UE 321 agreement A312031, Roquefort, France).

3.1 Experimental Unit of La Fage

The INRAE EU of La Fage is located in the Central Massif of France, in the Roquefort cheese production area. The EU of La Fage has two experimental flocks bred for different purposes. One flock of Romane sheep bred for meat production, under a purely open-air grazing system, and the other flock, on which this thesis is based, composed of milk-producing Lacaune ewes, under a mixed system of open-air grazing during the spring-summer months and stabling during the winter months. The Lacaune flock has been under selection for several years after its incorporation to the EU in 1990. The particularity of the La Fage flock is that ewes evaluated belonged to the two Lacaune populations selected by the genetic selection companies in France: Confédération Générale de Roquefort and Coopérative OVI-TEST.

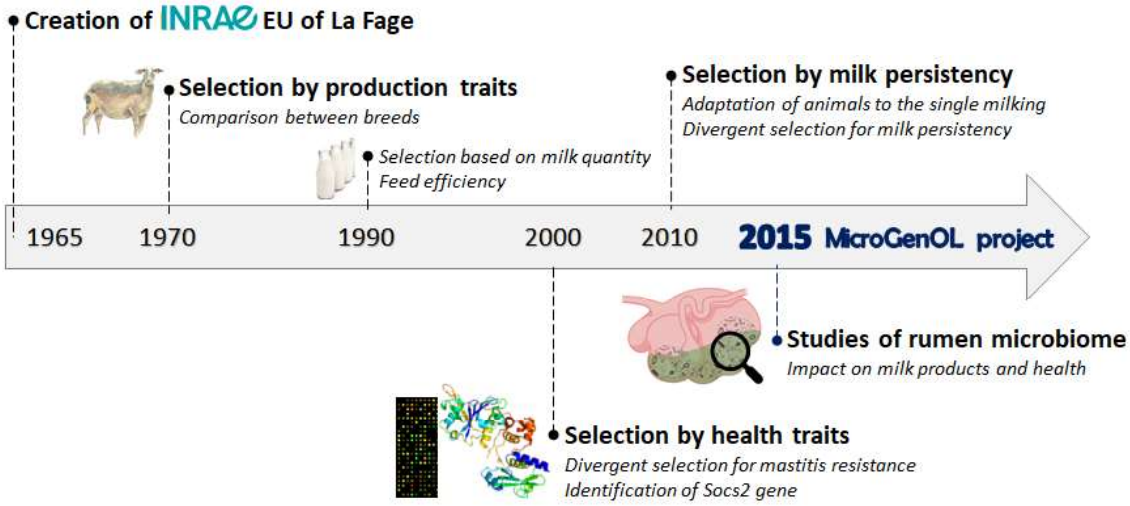


Figure 8. Timeline with the most relevant scientific programs developed at the INRAE Experimental Unit of La Fage.

Since the creation of the EU of La Fage in 1965, the objectives for which the dairy flock has been evaluated have evolved over time, adapting to new production requirements (Figure 8). The EU of La Fage was established in the Central Massif region not by chance, but at the request and effort of local dairy farmers who visualized the need for INRAE to work on local problems and improve milk production. Therefore, from 1970 to 1990, the main objectives were to produce animals that were adapted to indoors conditions and produce as much milk as possible. To this end, different dairy breeds and crossbreeds were compared, and the Lacaune breed selected as a purebred on the basis of milk quantity was chosen. However, after years of selection for milk quantity alone, the subclinical mastitis became evident through an increase in SCC in milk that impact cheese production. Therefore, starting in 2000, at the EU began to study and select for mastitis in sheep based on milk somatic cell score (SCS), but years later, the Socs2 gene was identified (Rupp *et al.*, 2015) changing the way of genetic selection within the EU. The Socs2 is an important gene involved in the expression of SCS values, which resulted in starting to genotype animals and selecting based on the absence of the mutant allele in double copy.

On the other hand, and considering the high feed costs involved in keeping animals indoors, after 1990, experimental works began to explore the genetic basis of feed efficiency of dairy ewes. But also, the high costs for farmers arise to cover the high energy needs when ewes have a very pronounced lactation peak, led the UE of La Fage since 2010 to start evaluating the

possible selection of ewes according to their lactation curve. That was the beginning of the divergently selection on milk persistency (PERS). Since this selection allows for obtain ewes that have flat lactation curves, thus, the same annual milk yield would be better distributed during lactation and feeding costs at the peak of lactation would be reduced. In turn, based on the farmers' preference to have only one milking per day, it was tested which of the two divergent lines in PERS is more adapted to it. Finally, due to the growing worldwide scientific interest in the role of rumen microbiota in production animals and the impact on the traits under study in the EU of La Fage, the project producing the data for this thesis was initiated in 2015.

3.2 Experimental design

3.2.1 Animals

The experimental design of the MicroGenOL project consisted on 806 Lacaune dairy ewes from INRAE EU of La Fage, which were rumen sampled and measured from 2015 to 2019. The final number of animals analyzed were 795 ewes, after to remove those ewes without any phenotypic record. These 795 dairy ewes were part of two datasets, **Dataset 1** composed by 795 ewes with one rumen sampling, and **Dataset 2** composed by 118 ewes with repeated rumen sampling one week apart.

The experiment included adult multiparous ewes with two or more lactations, where 83% of ewes (N= 664) were between 100 and 133 days in milk (DIM). The decision to include these categories of animals was made to reduce the experimental variability, given the differences in milk yield and rumen microbiota composition in the first lactation (Pitta *et al.*, 2014), as well as the variation in milk composition of ewes during the period close to lambing.

3.2.2 Feeding practices

All ewes weighing mean of 77 ± 9 kg, were raised indoors and fed a mixed ration of on average 90% meadow hay and silage plus 10% barley (on a gross matter basis) supplemented with approximately 150g of a commercial protein concentrate (38% of crude protein on dry matter basis) distributed in the milking parlor. The adjustment of the percentage of concentrates and forages was done each year according to the feeding value of the forages to cover the needs of the ewes, which therefore received the same amounts of nutrients over the five years. On average during this period, the ewes ingested 3.27kg of dry matter in total, comprising 16% of crude protein and 30% of crude fiber.

3.2.3 Genetic structure of the dairy flock

The genetic structure of the dairy flock includes four independent divergent genetic lines, two selected for milk persistency (PERS: PERS+ and PERS-), and two selected for milk somatic cell score (SCS: SCS+ and SCS-), which will be replaced after 2018 by the selection based on the Socs2 genotype (Socs2: CC, CT and TT). The genetic lines SCS+ and SCS-, and PERS+ and PERS- showed differences greater than two genetic standard deviations. In addition, we also obtained clear differences between lines using the EBVs for each of the genetic lines (Figure 9).

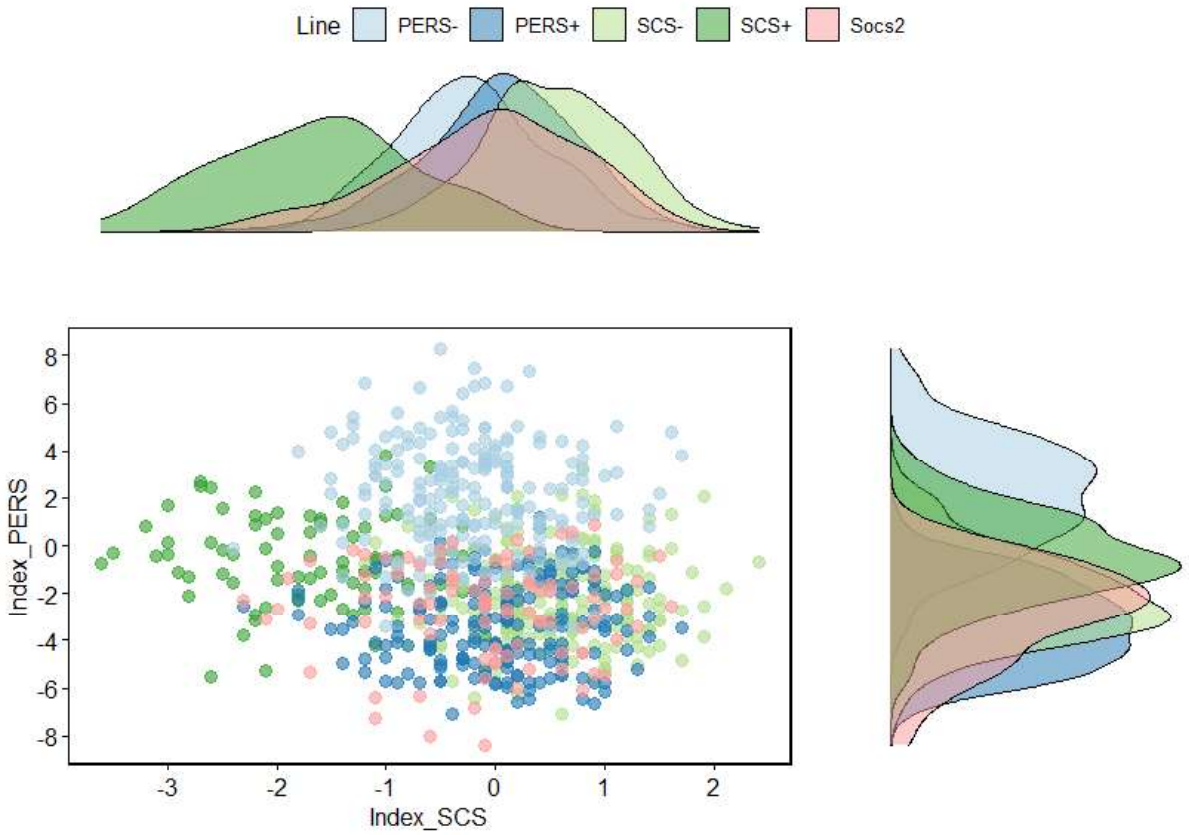


Figure 9. Distribution of ewes based on selection index of somatic cell score for SCS line (SCS+ and SCS-), and selection index of coefficient of variation of milk production for PERS line (PERS+ and PERS-).

Genetic selection of SCS lines was based on the EBVs for milk SCS of sires of the whole Lacaune population and dams within the La Fage flock, and was initiated in 2003 (Rupp *et al.*, 2003). Two groups of ewes with extreme EBVs were created according to the log-transformed SCC, a high-SCS line (SCS+) and a low-SCS line (SCS-). This selection was demonstrated by Rupp *et al.* (2009) to produce ewes with susceptibility/resistance to natural clinical and sub-clinical mastitis. But as was presented before, the selection for mastitis resistance changed after

the identification of the Socs2 gene. Then in the experiment, there are ewes selected for SCS index, but also genotyped for Socs2 gene (Rupp *et al.*, 2015): the new selection objective is to increase the mutant allele frequency (T) in the experimental Lacaune population to investigate possible associations with other traits.

Genetic selection of PERS lines was based on EBVs of Lacaune sires estimated relative to the whole Lacaune population based on the coefficient of variation in milk production on the testing day. The selection started in 2010, with extreme sires mated to produce the PERS divergent lines in the La Fage flock. Two extreme groups of ewes were created, one with high persistence (PERS+) and one with low persistence (PERS-) in milk production.

In Table 2 is presented the total number of ewes in the experiment grouped by year, genetic line and if they have repeated rumen samples.

Table 2. Total number of animals with records grouped by year and genetic lines (PERS, SCS or Socs2 genotype).

	Lactation	PERS+	PERS-	SCS+	SCS-	Total
2015	Adults (L2+)	89/**55** [30]	92/**60** [29]	63/**34** [29]	120/**88** [30]	364/**237** [118]
2016	L2	24/**22**	25/**23**	**18**	39/**33**	106/**96**
2017	L2	**31**	**28**	**13**	45/**44**	117/**116**
2018	L2	26/**25**	**34**	53†		113/**59**
2019	L2	30/**29**	**23**	42†		95/**52**
Total		200	202	94	204	795/**560**

L2+: second lactation or more, L2: second lactation, SCS+: high-SCS lines, SCS-: low-SCS line, PERS+: high-PERS line, PERS-: low-PERS line, †Socs2 genotype.
In bold the number of ewes with fine milk composition records from MIR spectra, and between brackets the number of ewes with repeated rumen samples one week apart.

3.3 Animal's genotyping

For all 795 ewes, blood samples were collected for DNA extraction and genotyping. Among 795 ewes, 743 were genotyped using Illumina Ovine SNP50 BeadChip (54,241 SNPs), 314 at

LaboGenA (Jouy en Josas, France) and 429 at Aveyron-Labo (Rodez, France). The remaining 52 ewes were genotyped using Illumina Ovine SNP15 (16,681 SNPs) at Neogen (Lansing, USA), followed by imputation to a medium-density SNP chip as part of the Lacaune dairy sheep genomic selection program (Larroque *et al.*, 2017). Genotypes were subjected to quality control, including minimum call rates of 90% for SNPs and 95% for individuals and exclusion of SNPs with a minor allele frequency lower than 5%. The final dataset included 773 genotyped individuals and 35,492 autosomal SNPs, corresponding to the SNP used in genomic selection. Markers were positioned on 26 autosomal *Ovis aries* chromosomes and mapped to the *Ovis aries* genome assembly Oar_v3.1 (https://www.ncbi.nlm.nih.gov/assembly/GCF_000298735.1/). The SNP associated with the Socs2 gene mutation was included in the map on autosomal chromosome 3 (129,722,200 bp).

3.4 Milk traits

Every three weeks, ewes milk yield (MY) was quantified and sampled at the morning and afternoon milking. The milk fat content (FC), protein content (PC) and SCC were measured by Interprofessional Milk Analysis Laboratory (Agrolabs, Aurillac, France), which also produced the mid-infrared (MIR) spectra of milk. We analyzed only the milk records from the milk control closest to the rumen sampling and the associated MIR spectra. The SCC was transformed to SCS, with the following formula [SCS = 3 + log2(SCC/100,000)]. A description of the daily milk traits analyzed is presented in Table 3. For the ewes with repeated samples one week apart (Dataset 2), milk samples were additionally analyzed with gas chromatography.

Table 3. Description of milk traits and fine milk composition traits.

Trait	Units	Mean	SD	CV (%)
Milk yield	ml	1946	589	30.3
Fat content	g per 100ml	7.37	1.14	15.5
Protein content	g per 100ml	5.71	0.52	9.1
Somatic cell score		3.11	1.91	61.4
Saturated fatty acids				
Butyric acid (C4:0)	g per 100ml	0.25	0.04	12.0
Caproic acid (C6:0)	g per 100ml	0.21	0.03	14.3
Caprylic acid (C8:0)	g per 100ml	0.20	0.03	15.0
Capric acid (C10:0)	g per 100ml	0.73	0.12	16.4
Lauric acid (C12:0)	g per 100ml	0.49	0.08	16.3
Palmitic acid (C16:0)	g per 100ml	1.96	0.37	18.9
Unsaturated fatty acids				
Oleic acid (*cis-9* C18:1)	g per 100ml	0.80	0.31	38.7
Rumenic acid (*cis-9 trans-11* C18:2)	g per 100ml	0.04	0.02	50.0
Alpha-linolenic acid (C18:3*n-3*)	g per 100ml	0.04	0.01	25.0
Caseins				
Alpha-S1-casein	g per 100ml	1.38	0.15	10.9
Alpha-S2-casein	g per 100ml	0.66	0.26	39.4
Beta-casein	g per 100ml	2.10	0.23	10.9
Kappa-casein	g per 100ml	0.45	0.05	8.9
Soluble proteins				
Alpha-lactalbumin	g per 100ml	0.14	0.01	7.7
Beta-lactoglobulin	g per 100ml	0.46	0.05	108.7

SD: standard deviation, CV (%): coefficient of variation.

3.4.1 Fine milk composition predicted from MIR spectra

We predicted by pre-established equations the fine profile of milk proteins and FAs from MIR spectra obtained with Milko-ScanTM FT6000 (Foss, Nanterre, France) (Figure 10). The prediction equations used in this study were characterized in Ferrand *et al.* (2012) for milk proteins, four caseins and two soluble proteins, and in Ferrand-Calmels *et al.* (2014) for FAs concentration, saturated and unsaturated FAs.

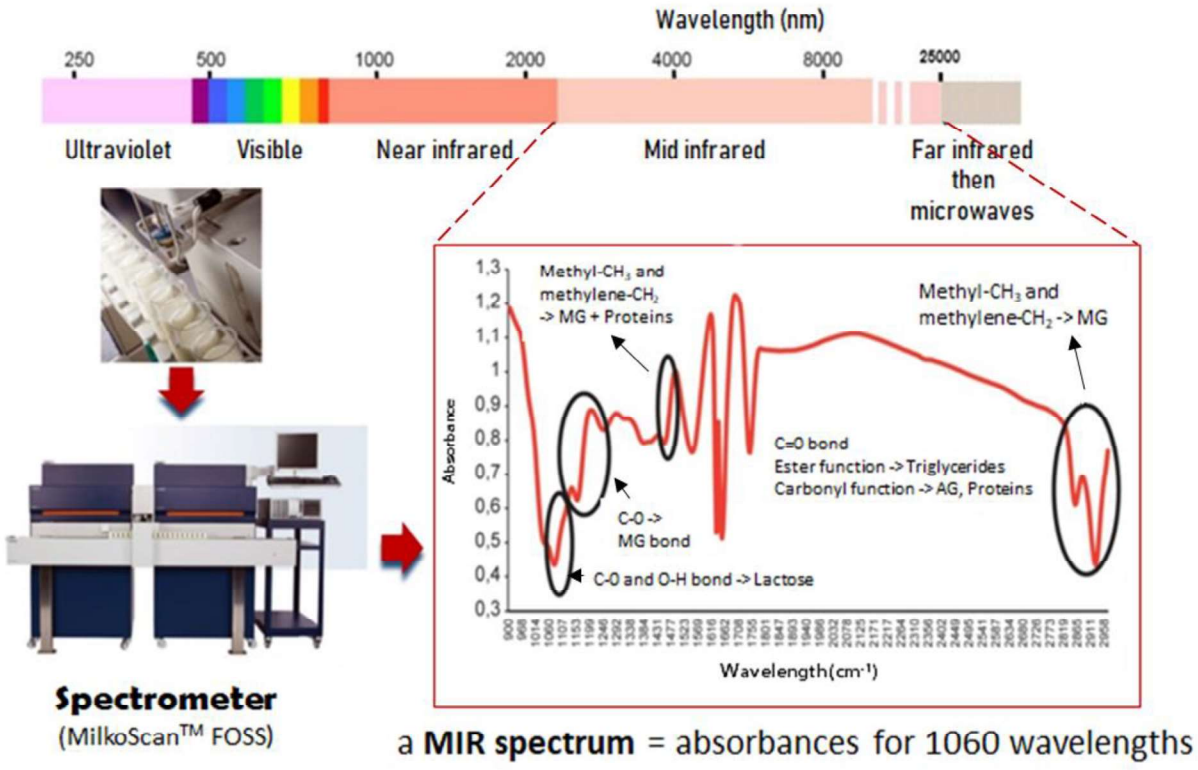


Figure 10. Mid-infrared (MIR) spectra technique (adapted from Sanchez et al., 2019).

3.4.2 Milk fatty acids quantified by gas chromatography

For the morning milk samples of the 118 ewes with repeated samples one week apart, we measured long-chain fatty acid (LCFA) percentages in total FA methyl esters (FAME) by gas chromatography at the National Veterinary of Toulouse (Toulouse, France), following the same method as for the rumen samples. In total, we quantified 38 FAs with eight FAs measured only in milk samples (C6:0, C7:0, C8:0, C9:0, C10:0, C11:0, C14:1, C16:1) and 30 FAs measured in both milk and rumen samples (C4:0, C12:0, C13:0, *anteiso*-C13:0, *iso*-C13:0, C14:0, *iso*-C14:0, C15:0, *anteiso*-C15:0, *iso*-C15:0, C16:0, C17:0, *anteiso*-C17:0, *iso*-C17:0, C18:0, *cis-9* C18:1, a mix of *cis-11* C18:1 and *trans-15* C18:1, *cis-12* C18:1, *cis-15* C18:1, a mix of *trans-6*, *trans-7*, *trans-8* C18:1, *trans-9* C18:1, *trans-10* C18:1, *trans-11* C18:1, *trans-12* C18:1, *trans-16* C18:1, C18:2, *cis-9 trans-11* C18:2, *trans-11 cis-15* C18:2, C18:3*n-3*, C20:1).

3.5 Sampling and analysis of rumen samples

Rumen sampling was performed according to a detailed protocol that received the approval of the Ministère de l'Enseignement Supérieur de la Recherche et de l'Innovation (APAFIS#6292–2016080214271984 v8). The most relevant points of rumen sampling are shown in Figure 11 and are detailed below.



Figure 11. Rumen sampling process. 1- Equipment used for animal's immobilization and rumen sampling, 2- Rumen sampling with a gastric tube, 3- On-site laboratory process of the rumen samples, 4- Milking parlor.

To avoid dilution of samples by feed or water, the animals did not have access to feed and water 10 hours and two hours prior to sampling, respectively. The date and time of each sample were recorded. The animal is immobilized in a special cage adapted for ewes (Figure 11-1), and sampling is performed by competent staff (Figure 11-2). Rumen sampling is performed using a medical gastric tube that is introduce through the mouth until it reaches the rumen (Figure 11-2), and a vacuum pump that allows for the sample to be extracted from the rumen through the gastric tube. On average, we collected rumen samples of 30ml with a liquid fraction and small solid particles. This rumen sampling method was chosen because of the ability to obtain representative samples of the rumen microbial community in a large number of animals (Henderson *et al.*, 2013; Ramos-Morales *et al.*, 2014).

Immediately after sampling, the rumen samples are kept on ice (Figure 11-3) for less than five minutes until aliquots are taken directly, one sample for FAs composition and one sample for microbiota composition, both frozen and stored in liquid nitrogen at -80°C (Figure 11-3). The ewes, after rumen sampling, go directly to the milking parlor for milking (Figure 11-4), but milk samples analyzed were obtained at the closest milking control within three days of rumen sampling.

3.5.1 Rumen fatty acids

The VFAs and LCFAs compositions were measured from rumen samples by two gas chromatography analyses at the National Veterinary of Toulouse (Toulouse, France). Six VFAs, C2:0, C3:0, C4:0, valeric acid (C5:0), isobutyric acid (*iso*-C4:0) and isovaleric acid (*iso*-C5:0), were quantified using automated gas separation, according to the method of Playne (1985) and modified as follows. The rumen samples were first centrifuged at 2,880 × g for 20min to separate the liquid phase. For protein removal, 1ml of supernatant was mixed with 200µl of (25% v/v) metaphosphoric acid and further centrifuged at 20,000 × g for 15min. One hundred microliters of supernatant were added to 75µl (0.2% v/v) of 4-methylvaleric acid as internal standard and 900µl of ultrapure water. One microliter of the mixture was then injected into a gas chromatograph (Hewlett Packard, Model 7890A) equipped with a DB-FFAP column (30m × 0.53mm *i.d.*, 1µm film thickness, Agilent Technologies, Palo Alto, USA) and an FID detector (Avondale, USA). Chromatograms were integrated using the Chromeleon software (Thermofisher Scientific, version 6.8, Whaltham, USA). The sum of the six VFAs concentration was defined as the total concentration and was used to obtain the molar proportions of each VFA.

The LCFAs of rumen content were extracted and methylated *in situ* using the procedure of Park and Goins (1994), except that the solution of 14% of boron trifluoride in methanol was replaced by a solution of methanol-acetylchloride. Nonadecanoic acid was used as the internal standard at a dose of 0.8mg. The FAME were then quantified by gas chromatography (Agilent 6890N, Network GC System, equipped with a model 7,683 auto injector, Agilent Technologies, Palo Alto, USA) using a fused silica capillary column (100m × 0.25mm *i.d.*, 0.20µm film thickness, CPSil88, Varian, Middelburg, the Netherlands) as described in Zened *et al*. (2011). Peaks were identified and quantified by comparison with commercial standards (Sigma Co., St Louis, USA), except for *cis*-9 C18:1, which were identified by the order of elution. Chromatograms were integrated using the Chromeleon software (Thermofisher Scientific, version 6.8, Whaltham, USA). Results were expressed as % of FA in total FAME.

3.5.2 Rumen microbial composition

The experimental design used to obtain and process the microbial composition from the sheep rumen samples is shown in Figure 12.

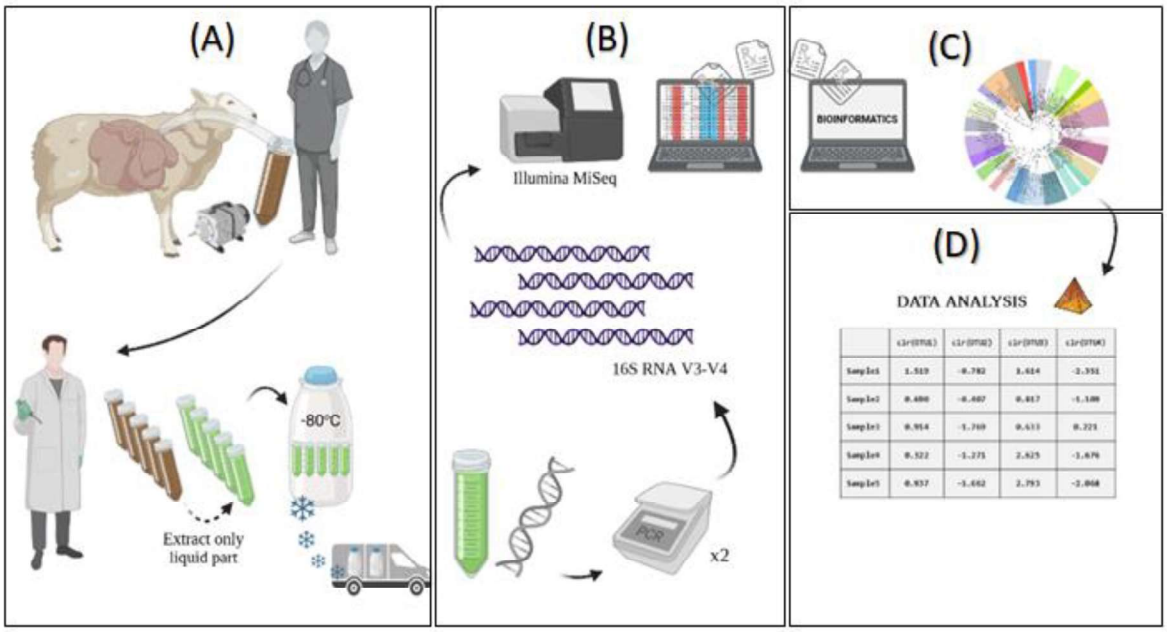


Figure 12. Experimental design to obtain and process rumen samples. (A) Rumen sampling, (B) DNA extraction and sequencing process, (C) bioinformatics process, (D) data analysis of rumen sample composition.

DNA extraction and targeting 16S rRNA gene sequencing

Total DNA from 80µl of ruminal sample was extracted and purified using the QIAamp DNA Stool Mini Kit (Qiagen Ltd, West Sussex, UK) according to the manufacturer's instructions, with a previous bead-beating step in a FastPrep instrument (MP Biomedicals, Illkirch, France).

As this project aims to work with the composition of microbiota in the rumen, the decision to use 16S rRNA gene sequencing was based on obtaining an accurate taxonomic affiliation, as well as lower sequencing costs for a large number of animals. The V3-V4 regions of 16S rRNA gene were amplified (first PCR: 30 cycles) from purified genomic DNA with the primers F343 (forward) and R784 (reverse). As Illumina MiSeq technology enables 250 bp paired-end reads, we obtained overlapped reads that generated extremely high-quality, full-length reads of the entire V3 and V4 regions in a single run. Single multiplexing was performed using a 6-bp index, which was added to R784 during a second round of PCR with 12 cycles using forward primer (AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGAC) and reverse primer (CAAGCAGAAGACGGCATACGAGATGTGACTGGAGTTCAGACGTGT). The resulting PCR products were purified and loaded onto an Illumina MiSeq cartridge (Illumina, San Diego, USA) at the Genomic and Transcriptomic Platform (INRAE, Toulouse, France)

according to the manufacturer's instructions. This process was repeated each year between 2015 and 2019, but in the first three years, the sequencing process was carried out at different times, so the samples were not sequenced in the same batch.

Bioinformatics process

The bioinformatics process of the DNA sequences was done using FROGS 3.0 pipeline (Escudié *et al.*, 2018) with the procedure presented in Figure 13.

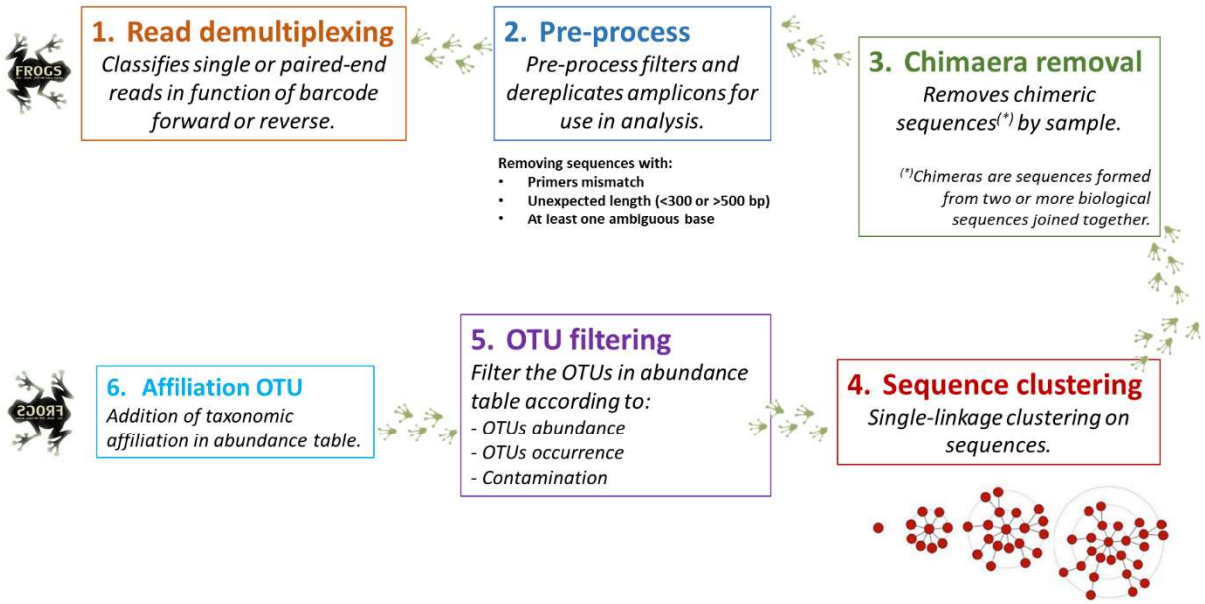


Figure 13. Bioinformatics process of DNA sequences using FROGS 3.0 pipeline.

The read demultiplexing is the first step of the bioinformatics process, and consisted of assigning each paired-end read to its sample on the basis of the previously integrated index. The second step consisted of read pre-processing, where we defined some parameters to remove sequences, such as the presence of primer mismatch, sequence length less than 300 bp or greater than 500 bp, or the presence of at least one ambiguous base. The third step is the chimaera removal defined in Figure 13, followed by the sequence clustering. In FROGS 3.0, de novo clustering is performed where sequences grouped according to pairwise distance measures. We used one sequence of difference between each of the three-aggregation step of clustering. The fifth step is the operational taxonomic unit (OTU) filtering. The Bokulich filter was the only one used in this work, which consists of remove OTUs with an abundance lower than 0.005% of the total sequences (Bokulich *et al.*, 2013). The decision was based on the fact that a standardized mock community was not included in each sample for filtrate calibration, so the recommendation to apply a conservative filter at 0.005% was followed. Finally the last step is

the OTU affiliation using the SILVA database (version 138) (Quast *et al.*, 2012). It is an ideal reference due to their extensively curated taxonomy and the new non-redundant datasets. All the parameters, filters and values used for the bioinformatics process were based on support of the FROGS pipeline technical group of INRAE (Toulouse, France).

As result of the independent bioinformatics process applied for the two datasets in this work, we retained around 47% of the initial DNA sequences and we obtained very similar final number of OTUs in abundance tables (Table 4).

Table 4. Number of DNA sequences and OTUs along the bioinformatics process for datasets.

	Dataset 1 N= 795	Dataset 2 N= 236
Initial sequences	20,853,771	5,258,562
Sequences with 5' primer	18,484,436	4,863,246
Sequences with 3' primer	18,383,617	4,836,379
Sequences with expected length	18,372,188	4,830,627
Sequences without aberrant nucleotides	18,322,421	4,830,596
Sequences without chimera	14,989,216	3,853,044
Final sequences	9,536,442	2,500,763
Number of OTUs	2,059	2,079

N: number of samples, OTUs: operational taxonomic units.

Analysis of microbial abundances

For Dataset 1, the OTU abundance table had 795 samples (as rows) and 2,059 affiliated OTUs (as columns), represented by 750 to 168,617 sequences (median of 1,755 DNA sequences). Rumen samples contained in average 1,322 OTUs per sample. Rare OTUs were identified in the dataset represented by those with lesser than 2,034 sequences across all samples.

For Dataset 2, the OTU abundance table had 236 samples and 2,079 affiliated OTUs, represented by 193 to 113,090 sequences (median of 453 DNA sequences). Rumen samples contained in average 1,406 OTUs per sample. In Table 5 is present the frequency of unknown taxonomic affiliation for OTUs in both datasets.

Table 5. Number of taxon and frequency of unknown affiliation per taxonomic level.

Taxonomic level	Dataset 1 (2,059 OTUs)		Dataset 2 (2,079 OTUs)	
	N	Unknown affiliation	N	Unknown affiliation
Phylum	11	0%	10	0%
Class	17	0%	15	0%
Order	39	1%	34	0%
Family	56	3%	56	4%
Genus	112	20%	117	23%
Species	70	94%	67	93%

N: number of identified taxonomic affiliation, OTUs: operational taxonomic units.

Using the abundance table of the Dataset 1, for each rumen sample, we calculated some microbial diversity indices which summarize the structure of a community: first of all, a basic measure of microbial richness as the number of OTUs observed in each sample, as well as three alpha-diversity indices that are measures of richness and evenness such as Shannon, inverse Simpson and Simpson. The Shannon index is a measure of the evenness of the species abundance distribution, Inverse Simpson index is the inverse probability that two sequences sampled at random come from the same species, and Simpson index is 1- probability that two sequences sampled at random come from the same species. Another common measure but of beta-diversity is the Bray-Curtis distance between samples, which gives information about similarities between communities. However, to calculate Bray-Curtis it is necessary to apply the rarefaction process for which there is clear evidence against it (McMurdie and Holmes, 2014), so in this work we do not estimate this distance measure.

Microbial abundance data is a random sample of the relative abundance of species in the sampled ecosystem. This is because the abundances are generated by high-throughput sequencing technologies, which have a fixed size, imposing a sampling constraint. The microbial abundance table is composed of counts (representing relative abundances) with always positive values that sum to a constant, so the relative abundances represent compositional data.

Compositional data analysis

In compositional data the relevant information is not in the parts, but in the ratios between the parts of the composition (Aitchison, 1986). A composition is a vector of N parts $x = [x_1, x_2, \dots, x_N]$, where each part is assumed positive. For example, in a microbial abundance table, each microbial sample is a row compositional vector and each part is the abundance of a specific OTU. Then, the group of real positive vectors closed to a constant is called the simplex of N parts (Eq. 4),

$$S^N = \left\{ \boldsymbol{x} = [x_1, x_2, \dots, x_N] \middle| x_j > 0, j = 1, 2, \dots, N; \sum_{j=1}^{N} x_j = \Psi \right\} \quad (4)$$

where, S^N is the simplex space of N parts defined by a set of vectors $\boldsymbol{x}$ with N parts, each part x_j is strictly a positive real number, and the sum of all components is equal to a constant Ψ, which can be 1 (representing proportions), 100 (percentages), 10^6 (parts per million), and 10^9 (parts per billion). This constraint in S^N to a constant Ψ produces an interdependence between parts of a composition, where if one of the parts increases, some of the other parts must decrease.

The compositional data, with thousands of variables (*e.g.,* OTUs), should be analyze using multivariate methodologies which are based on the Euclidean space (a finite-dimensional vector space over the reals $\mathbb{R}$, with an inner product). However, as compositional data is on the simplex space, Aitchison (1986) proposed to project the simplex space to the real space, using functions that must be:

1. Scale invariant, which means that no matter the value of the part, the results must be the same.
2. Permutation invariant, which means that if the parts of the vector changes their order, the result will be equivalent (*e.g.,* [OTU_1, OTU_2, OTU_3] and [OTU_2, OTU_1, OTU_3] have equivalent result).
3. Sub-composition consistent, which means that if a sub-composition of the data is taken, the ratios between the parts will be the same as in the complete data.

The logratio is the most common function which follows these three principles. A logratio transformation provides a one-to-one mapping on to the real space led to the use of a variety of these transformations, such as additive, centered and isometric logratio. In turn, these allowed for the use of unrestricted multivariate methods applied to the transformed data, with translatable inferences back into compositional statements. In this thesis, we used the centered

log-ratio (CLR) transformation, which is the most commonly used in animal microbiota data (Zhang *et al.*, 2020; Pérez-Enciso *et al.*, 2021; López-García *et al.*, 2022).

Since microbiota data are characterized by a high proportion of zeros, which poses a problem for using logarithms, zero values must be imputed before applying the CLR transformation. In compositional data, a zero value could be:

a) Essential zeros, a value that was not measured, *i.e.,* true zero in the data.

b) Counting zeros, which appear in sampling studies involving counts (*e.g.,* microbial data).

c) Rounded zeros are data not be observed because their true value is below the maximum round-off error or detection limit (*e.g.,* milk fatty acid concentrations).

The most common zeros in microbial data are count zeros that refer to the non-presence in the sample of a specific OTU. The assumption about this zero is that this OTU may not be observed due to limited size of the sample or limited capacity of the sequencing instrument. That is, if the sample size increases, some OTUs will probably be observed, *i.e.,* non-zero probability. Consequently, it seems logical to impute count zeros with non-zero values. In the literature, the most common methods used are two: the addition of a constant (*e.g.,* 1) for all values, or the addition of a small value (*e.g.,* 0.001) only for count zeros in the dataset. However, Martín-Fernández *et al*. (2015) proposed an adapted replacement method for compositional data, called geometric Bayesian multiplicative (GBM). The GBM consists in imputing zero values using an informative prior distribution, and in a multiplicative modification of the non-zero values to preserve the ratios between parts.

Let $\boldsymbol{X}$ denote a dataset formed by compositional vectors $\boldsymbol{x_i}$ with N parts ($i = 1, \dots, N$). When one wants to replace the count zeros in a vector $\boldsymbol{x_i}$, the parameters considered are:

$$\alpha_{ij} = \sum_{\substack{k=1 \\ k \neq i}}^{N} x_{kj} \tag{5}$$

$$t_{ij} = \frac{\alpha_{ij}}{\sum_{k=1}^{D} \alpha_{ik}} = \widehat{m}_{ij} \tag{6}$$

where, α_{ij} is the prior information in j for x_{ij} considering the other vectors $\boldsymbol{x_i}$ leaving out their own value ($k \neq i$), t_{ij} is the prior information α_{ij} divided by the total sum of each row k in $\boldsymbol{\alpha}$. Then, we should consider a new parameter s that regulates the importance of the prior estimate,

which in the GBM method corresponds to $1/g_i$. The Equation 7 replaces the vector x_i by the vector $\boldsymbol{r_i} = (r_{i1}, \ldots, r_{iN})$:

$$r_{ij} = \begin{cases} \dfrac{\widehat{\boldsymbol{m}}_{\boldsymbol{ij}}}{g_i * n_i + 1} & if\ x_{ij} = 0, \\ x_{ij} * \left(1 - \dfrac{\sum_{k|x_{ik}=0} \widehat{\boldsymbol{m}}_{\boldsymbol{ik}}}{g_i * n_i + 1}\right) & if\ x_{ij} > 0, \end{cases} \quad (7)$$

where, $\widehat{\boldsymbol{m}}_{\boldsymbol{ij}}$ was defined above, $\widehat{\boldsymbol{m}}_{\boldsymbol{ik}}$ is the prior information use to replace the zero values, g_i is geometric mean of $\widehat{m}_i$ and n_i is the number of observations.

Once all zeros have been imputed, the logarithms could be used. The CLR transformation is symmetric in their parts, keeping the same number of parts (or OTUs). Thus, if a sample have N parts there are N logratios in the CLR set, using as denominator the geometric mean of the sample (Eq. 8).

$$\boldsymbol{CLR}(\boldsymbol{r}) = \left[ln\frac{r_1}{g(\boldsymbol{r})}, ln\frac{r_2}{g(\boldsymbol{r})}, \ldots, ln\frac{x_N}{g(\boldsymbol{r})}\right] \quad (8)$$

where $\boldsymbol{r}$ is the vector with replaced values of $\boldsymbol{x_i}$ of N parts, ln is the neperian logarithm, and $g(r) = (r_1 \cdot \ldots \cdot r_D)^{1/N}$ is the geometric mean of the parts of r. As result of CLR transformation, microbial data are mapped in real space with all OTUs, which is relevant for community analysis. However, a limitation of CLR is that, although it expresses data in real space, it is not isometric. Although recently Greenacre *et al.* (2021) mention that it is so close to being isometric that, for practical purposes, the CLR is isometric.

In this thesis, we used compositional data analysis not only for rumen bacterial abundance, following the workflow that will be proposed in Chapter 4, but also for the fine milk composition and rumen fatty acids detailed above. This was a decision considering the nature of microbial abundances (Gloor *et al.*, 2017), and proteins and fatty acids (Graeve and Greenacre, 2020) whose values are constrained by the methodology used for their sampling.

3.6 Statistical models

The remaining sources of variation, then those considered in the experimental design of the project, were evaluated by ANOVA at a significance level of 5%. For the microbial abundances, in both datasets, the environmental factors included in the models were those significant for more than the 10% of the total number of OTUs in the abundance table. The threshold of 10% was defined on the basis that the most relevant effects for OTUs were generally above this

threshold, and those mainly biological effects of lesser effect on abundances presented percentages below 5% (approx. 103 OTUs). The mean R-squared of models for microbial abundances was 0.20, with a range from 0.04 to 0.77.

The models used in this thesis were linear models for the correction of fixed effects of OTUs and milk traits, and mixed models for the genetic parameter estimations. In the first analysis, we used multiple traits models with a pedigree matrix to estimate the heritability of microbial abundance of Dataset 1 (2,059 OTUs) and genetic correlations between microbial abundances and dairy traits (Chapter 5). The inclusion of a pedigree matrix was due to convergence problems using the genomic matrix in models for 2,059 OTUs and 18 dairy traits. And, the use of multiple traits model was used to account the genetic structure of the population (Zhou and Stephens, 2014) with two genetic lines, SCS and PERS, for which selection traits such as LSCS and coefficient of variation of milk production were included as traits in the models.

For the last analysis, we calculated the total heritability of dairy traits (Chapter 6) based on the method proposed by Christensen *et al.* (2021) with a joint model for omics and phenotypes. As an omics, we included microbiota through a microbial similarity matrix $\boldsymbol{O}$, but the authors called $\boldsymbol{M}$ (Difford *et al.*, 2018; Weishaar *et al.*, 2020). We constructed the matrix $\boldsymbol{O}$ in different ways that are described in the supplementary results of Chapter 6.

3.7 Data availability

The sequence read archive are publicly available on the National Library of Medicine repository under two NCBI BioProjects (https://www.ncbi.nlm.nih.gov/bioproject/): the first called **"Rumen microbiota of Lacaune dairy sheep from INRAE La Fage"** (accession number PRJNA723543) with the 795 samples sequenced once, and the second called **"MicroGenOL week repeatability"** (accession number PRJNA765197) with the 236 samples corresponding to 118 ewes sequenced twice. In addition, a Dataset called **"Lacaune dairy ewes"** is available in Data INRAE repository (https://doi.org/10.15454/IOET0P) including the 795 samples with environmental effects, and MY and milk composition traits.

3.8 Conclusion

The experimental data for this thesis consisted of 795 Lacaune dairy ewes from the Experimental Unit of La Fage. The adult multiparous ewes were reared under the same housing and feeding conditions during the five years of the experiment, allow for reduce the environmental variability between animals. The genetic structure of the population was

composed of two divergent genetic lines, SCS and PERS, and more recently the replacement of the SCS line by the selection by Socs2 genotype. All ewes have pedigree and genomic information, milk records and rumen microbial abundance from 16S rRNA gene sequencing analyzed using the compositional data workflow that is defined in Chapter 4. The compositional workflow was compared with other published methodologies to benchmark against other works, and to quantify the impact of the method used in genetic parameter estimations. In addition, a subgroup of animals had rumen samples repeated one week apart, which allowed for us to measure the stability of the microbial composition and to evaluate the conservation of associations between OTUs and phenotypes from one week to the next, which is also presented in Chapter 4.

As a second part of this work, in the experimental design, rumen sampling was planned to be performed within no more than three days of milk recording and sampling, which enabled the correct linkage between rumen microbiota composition and daily milk traits. Thus, it was possible to estimate phenotypic and genetic correlations between rumen microbiota and milk traits, as presented in Chapters 4 and 5. Genetic determinism of microbiota composition was also investigated in Chapter 5. Finally, due to the availability for all animals of phenotype records, genomic and microbial information, it was possible to implement mixed models including two random effects (animal and microbiome) to explain the phenotype. Based on this, estimates of the direct genetic effect, and of the indirect genetic effect on the phenotype mediated by the rumen microbiota are presented in Chapter 6.

Chapter 4
Working with microbial abundances

4.1 Introduction

Microbial abundances are multivariate high dimensional data and constrained by a constant sum, then they should not be analyzed directly with the classical methodologies, as they could produce unreliable results, but should be preprocessed for analysis. However, there is no standard methodology for metagenomics, which, coupled with the enormous source of variation from sampling to data processing (*e.g.,* biological, technical and computational factors) (Wang and LêCao, 2019), makes it difficult to compare the results obtained by different studies. (Gloor *et al.*, 2016) defined this as the reproducibility problem of microbiome studies.

In this thesis, in order to obtain reliable and reproducible results, we defined a workflow considering the compositional nature of the microbial data proposed by Gloor *et al.* (2017). This proposed compositional workflow was applied on a real dataset of dairy ewes, with the objectives of discriminating two divergent genetic lines using the rumen microbiota composition, and estimating associations between rumen microbial abundance and fine milk composition traits. However, given the identified sources of variation affecting the microbiota, could the bacterial composition and estimated associations with phenotypes be maintained over a short period of time? To answer this question, we evaluated the individual and community stability of the bacterial composition from one week to the next and the conservation of the associations between bacterial composition and milk and rumen FAs, using repeated measurements of a subset of one hundred dairy ewes. Finally, due to the difficulties encountered in comparing the results with the literature, we compared the proposed compositional workflow to different data processing methods used to estimate genetic variability of microbial abundances.

The first results obtained for the compositional data workflow, discriminant analysis between genetic lines SCS and PERS, and phenotypic association between rumen microbiota and dairy

traits were presented as Oral presentation at the 71nd Annual Meeting of the European Federation of Animal Science (EAAP) in December 2020, and later published under the title **"Compositional analysis of ruminal bacteria from ewes selected for somatic cell score and milk persistency"** in PLOS ONE Journal in July 2021. In addition, the methodological pipeline developed in this study was applied on another dataset of rumen microbiota from meat sheep, resulting in a publication entitled "Apart from the diet, the ruminal microbiota of lambs is modified in relation to their genetic potential for feed efficiency or feeding behavior" published in Frontiers in Microbiology Journal in October 2021.

In 2021 within the framework of a European Master's degree, Solène Fresco, supervised by Christel Marie-Etancelin and myself, carried out the analyses and processed the results corresponding to the stability over a short-period of time of the ruminal bacteria composition and the associations between ruminal bacteria and milk and rumen FAs. One of the main objectives was to evaluate the possibility of minimizing rumen sampling, which is invasive for the animals and time-consuming. These results were published under the title **"Variation in rumen bacteria of Lacaune dairy ewes from one week to the next"** in Frontiers in Microbiology Journal in June 2022.

Finally, the results obtained from the evaluation and comparison of data processing methods through the impact on the heritability estimates were presented under the title **"Ruminal bacteria heritabilities: which is the impact of data processing?"** in Poster format at the 72nd Annual Meeting of the European Federation of Animal Science (EAAP) in September 2021 in Davos (Switzerland).

4.2 Compositional analysis of ruminal bacteria from ewes selected for somatic cell score and milk persistency

PLOS ONE

RESEARCH ARTICLE

Compositional analysis of ruminal bacteria from ewes selected for somatic cell score and milk persistency

Guillermo Martinez Boggio[1]*, **Annabelle Meynadier**[1☯], **Pepus Daunis-i-Estadella**[2], **Christel Marie-Etancelin**[1☯]

1 GenPhySE, INRAE, INPT, ENVT, Université de Toulouse, Castanet-Tolosan, France, **2** Department of Computer Science, Applied Mathematics and Statistics, University of Girona, Girona, Spain

☯ These authors contributed equally to this work.
* guillermo.martinez-boggio@inrae.fr

Check for updates

OPEN ACCESS

Citation: Martinez Boggio G, Meynadier A, Daunis-i-Estadella P, Marie-Etancelin C (2021) Compositional analysis of ruminal bacteria from ewes selected for somatic cell score and milk persistency. PLoS ONE 16(7): e0254874. https://doi.org/10.1371/journal.pone.0254874

Editor: Arda Yildirim, Tokat Gaziosmanpasa Universitesi, TURKEY

Received: April 24, 2021

Accepted: June 25, 2021

Published: July 26, 2021

Copyright: © 2021 Martinez Boggio et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All sequences and data files are available from NCBI (accesion number:PRJNA723543) and data INRAE (DOI: https://doi.org/10.15454/IOET0P).

Funding: GMB Funders: INRAE and La Région Occitanie https://www.inrae.fr/ https://www.laregion.fr/ The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Ruminants are dependent on their rumen microbiota to obtain energy from plants. The composition of the microbiome was well-known to be associated with health status, and production traits, but published results are difficult to reproduce due to large sources of variation. The objectives of this study were to evaluate the associations of ruminal microbiota and its association with genetic lines selected by somatic cell score (SCS) or milk persistency (PERS), as well as milk production, somatic cell score, fat and protein contents, and fatty acids and proteins of milk, using the principles of compositional data. A large sample of 700 Lacaune dairy ewes from INRAE La Fage feeding the same diet and belonging to two divergent genetic lines selected for SCS or PERS was used. The ruminal bacterial metagenome was sequenced using the 16S rRNA gene, resulting in 2,059 operational taxonomic units affiliated with 112 genera. The abundance data were centred log-transformed after the replacement of zeros with the geometric Bayesian method. Discriminant analysis of the SCS showed differences between SCS+ and SCS- ewes, while for PERS no difference was obtained. Milk traits as fat content, protein content, saturated fatty acids and caseins of milk were negatively associated with *Prevotella* (R = [-0.08;-0.16]), *Suttonella* (R = [-0.09;-0.16]) and *Ruminococcus* (R = [-0.08;-0.16]), and positively associated with *Lachnospiraceae* (R = [0.09;0.16]) and *Christensenellaceae* (R = [0.09;0.16]). Our findings provide an understanding of the application of compositional data to microbiome analysis, and the potential association of *Prevotella*, *Suttonella*, *Ruminococcaceae* and *Lachnospiraceae* with milk production traits such as milk fatty acids and proteins in dairy sheep.

Introduction

Ruminants are able to obtain energy from plant fibre to produce foods for human consumption. This is achieved through rumen symbiosis with colonizing microorganisms, such as bacteria, protozoa and fungi. Bacteria are the most abundant microorganisms in the rumen and

make the greatest contribution to the digestion and conversion of feeds to short-chain fatty acids, microbial proteins and vitamins [1]. Associations of the ruminal microbiota with sire breed [2] and with different traits, such as feed efficiency [3], methane yield [4–6], and milk composition [7–9], have been reported, mainly in cows. However, in sheep only a few studies reported changes in the rumen bacteria with different diets [10–12], but no associations with milk production traits in dairy ewes. Research on cows considered a few animals with a maximum sample size of 16 [7–9] and used phenotypic differences, not genetic selection.

The main problem in published studies concerning the association between the microbiome and production traits is reproducibility. In the general workflow of microbiome analysis, the sources of variation, from sampling to statistical analysis, are almost infinite [13]. High-throughput sequencing technologies have made an important contribution to the knowledge of ruminal microbiome diversity. However, technologies with a limited number of sequencing reads obtained per sample, such as metabarcoding of the 16S rRNA gene, place a constraint on microbial data. Thus, the observed read counts is a fixed-size random sample of the relative abundance of the operational taxonomic units (OTUs) in the ecosystem. Moreover, the counts obtained are not related to the absolute value of the OTU, but to the probability of counting the OTU [14].

This kind of data is referred to as compositional, and a statistical approach adapted to this data must be applied. The term compositional data [15, 16] is used to describe a data set in which the parts of each sample have an arbitrary or noninformative sum, such as 100 for percentages. As result, the data contain only information about the relationships between the different parts of the composition. Three principles should be fulfilled in any statistical analysis of compositions: scale invariance, permutation invariance and sub-compositional coherence [16]. To meet one of the most important principles, scale invariance, it was proposed to work with the log-ratio, whose invariant form is called the log-contrast [15]. Compositional data are represented in a non-Euclidean space called a simplex. The log-ratio transformations proposed by Aitchison [15] and Egozcue et al. [17] allow observations to be represented in Euclidean space, on which most association analyses are statistically based. Centred log-ratio (CLR) and isometric log-ratio (ILR) transformations are the most widely used types of log-ratio transformations; both are isometric and allow correct operation in Euclidean space. However, only the ILR is orthonormal, generating a complete set of independent transformed variables on an orthonormal basis (as a coordinate system). Thus, the ILR works with balances [17], while the CLR works with OTU abundance, which allows a simple interpretation of the results.

Zero values are slightly more problematic in compositional data analysis than in standard multivariate statistical analysis because it is not possible to work with log-ratios if we have zero values in the data set. Microbiome metabarcoding data represent the probabilities of counts per OTU through a random sampling process [18], so some values in the data are true zero values due to true absence in the ecological environment, while others could arise randomly because of the constraint generated by high-throughput sequencing technologies. In the literature, different ways of correcting these zero values are applied, from the use of arbitrary corrections such as adding an offset of 1 to all values in the data set to the use of Bayesian models [18, 19].

Another procedure that must be considered when working with microbiome data, which contributes to the reproducibility of the results, is adjusting the data according to the different sources of experimental variation mentioned above. In the literature, these effects are known as batch effects, and they can include technical factors such as sample collection and storage, sample processing, and DNA sequencing; biological factors such as animal breed, health status and environmental effects; and computational factors such as bioinformatic pipelines and the statistical analysis used [20].

Thus, to obtain robust and reproducible results when working with microbiome data, it is crucial to use a compositional data approach, as stated by Gloor et al. [14] (*"Microbiome datasets are compositional*: *and this is not optional"*), and to adjust the data according to the principal sources of experimental variation.

The main purposes of this study are to present a conceptual framework for the compositional data approach applied to metabarcoding data in a discriminant analysis of divergent genetic lines of sheep selected on the basis of either somatic cell score (SCS) or milk persistency (PERS) and to link ruminal bacteria with milk production and milk quality traits.

Materials and methods

Animal handling and sampling

Data were obtained from the INRAE Experimental Unit of La Fage (UE 321 agreement A312031, Roquefort, France) between 2015 and 2019. The animals under study were adult Lacaune dairy ewes (weighing 77 kg on average) raised indoors and fed 93% meadow hay and silage plus 7% of concentrates (on dry matter basis). The genetic structure of the INRAE La Fage flock includes independent divergent genetic lines of Lacaune dairy ewes: two selected for milk SCS and the other two for PERS. Divergent selection based on estimated breeding values (EBVs) for milk SCS of sires of the whole Lacaune population and dams within the La Fage flock was initiated in 2003 [21]. Two groups of ewes with extreme EBVs were created according to the log-transformed somatic cell count (SCC): a high-SCS line, represented as SCS+, and a low-SCS line, represented as SCS-. This selection was demonstrated to produce ewes with susceptibility/resistance to natural clinical and sub-clinical mastitis [22]. Estimated breeding values of Lacaune sires were estimated relative to the whole Lacaune population based on the coefficient of variation of milk production on the testing day. Starting in 2009, extreme sires were mated to produce the PERS divergent lines in the La Fage flock. Two extreme groups of ewes were generated, one with high persistence (PERS+) and one with low persistence (PERS-) in the milk production curve.

Ruminal contents were sampled from each ewe using a vacuum pump and a medical gastric tube, that allows a qualitative representation of the rumen microbial community in a large number of animals [23]. To avoid dilution of samples by feed or water, the animals did not have access to feed and water 10 hours and 2 hours prior to sampling, respectively. Immobilization was performed with a special cage adapted for ewes, sampling was performed by competent staff, and the gastric tube was thoroughly rinsed with clean water between animal sampling to minimize cross-contamination. The rumen samples were directly aliquoted, frozen and stored at -80˚C. This protocol received approval from the Ministere de l'Enseignement Superieur de la Recherche et de l'Innovation–Animal ethics committee with the following approval number: APAFIS#6292–2016080214271984 v8.

The experimental data consisted of 700 ruminal samples, including 94 from SCS+ ewes, 204 from SCS- ewes, 200 from PERS+ ewes and 202 from PERS- ewes. The genetic difference within the SCS and PERS lines was obtained by estimating index differences between SCS +/SCS- and PERS+/PERS- expressed in standard deviations of the indexes estimated for the whole Lacaune dairy population.

16S rRNA gene amplicon sequencing

Total DNA from 80 μL of ruminal sample was extracted and purified using the QIAamp DNA Stool Mini Kit (Qiagen Ltd, West Sussex, UK) according to the manufacturer's instructions, with a previous bead-beating step in a FastPrep instrument (MP Biomedicals, Illkirch, France). The 16S rRNA V3-V4 regions of the extracted DNA strands were amplified (first PCR: 30

cycles) from purified genomic DNA with the primers F343 (5′–CTTTCCCTACACGACGCTCTTCCGATCTACGGRAGGCAGCAG–3′; [24]) and reverse R784 (5′–GGAGTTCAGACGTGTGCTCTTCCGATCTTACCAGGGTATCTAATCCT–3′; [25]). As Illumina MiSeq technology enables 250-bp reads, the ends of each read are overlapped and can be stitched together to generate full-length reads of the entire V3 and V4 regions in a single run. Single multiplexing was performed using a 6-bp index, which was added to R784 during a second PCR with 12 cycles using forward primer (`AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGAC`) and reverse primer (`CAAGCAGAAGACGGCATACGAGATGTGACTGGAGTTCAGACGTGT`). The PCR products were purified and loaded onto an Illumina MiSeq cartridge (Illumina, San Diego, CA, USA) at the Genomic and Transcriptomic Platform (INRAE, Toulouse, France) according to the manufacturer's instructions. This process was repeated each year between 2015 and 2019, but in the first three years, the sequencing process was carried out at different times, so the samples were not sequenced in the same batch.

The sequences of the 700 samples were processed using the FROGS pipeline [26] by following the FROGS workflow operational procedure: (i) read demultiplexing, i.e., assigning each paired-end read to its sample on the basis of the previously integrated index; (ii) read pre-processing, i.e., removing sequences presenting a primer mismatch, displaying an unexpected length (>300 or <500 bp) or with at least one ambiguous base; (iii) chimaera removal; (iv) sequence clustering with denoising and a defined difference of d = 1 between sequences in each aggregation step of clustering; (v) cluster filtering, i.e., removing clusters with abundances <0.005% of the total sequences [27]; and (vi) taxonomy assignment to OTUs using SILVA database (version 138). OTU number refers to the identification of the OTU in the abundance table.

Abundance data

The abundance table and taxonomy files were imported into R (v4.0.2). Zeros were imputed under the assumption that the probability of occurrence of the OTUs in the multinomial experiment was not zero. Therefore, the geometric Bayesian-multiplicative (GBM) method [19] was used, where the zero values were replaced in each sample by the posterior probability obtained from the Bayesian model, which considered all the available data, and weighted by the geometric mean. To maintain the ratios between all the abundance values, the non-zero values were multiplied by a value generated as a function of the posterior probability and the geometric mean. The GBM method was performed with the following formula, through the cmultRepl function of the zCompositions package [28] in R (v4.0.2):

$$r_{ij} = \begin{cases} \dfrac{\hat{m}_{ij}}{g_i \cdot n_i + 1}, & if\ x_{ij} = 0, \\ x_{ij} \cdot \left(1 - \dfrac{\Sigma_{k|x_{ik}=0}\, \hat{m}_{ik}}{g_i \cdot n_i + 1}\right), & if\ x_{ij} > 0. \end{cases} \quad (1)$$

where r_{ij} is a vector of replacement abundance values defined as $r_{ij} = (r_{i1},\ldots,r_{i2059})$, $\hat{m}_{ij}$ is the prior probability estimate for each category j, g_i is the geometric mean of the i-th row, and n_i is the number of samples.

The abundance table with no zero values was CLR transformed with the following formula through the compositions package [29] in R (v4.0.2):

$$clr(x) = \left[\log\left(\frac{x_1}{g(x)}\right), \log\left(\frac{x_2}{g(x)}\right), \ldots, \log\left(\frac{x_D}{g(x)}\right)\right], \quad (2)$$

where x is a row vector with abundances for the OTUs in the sample (x_1 = OTU 1, x_2 = OTU 2, to x_D = OTU 2059), $g(x) = \sqrt[D]{x_1 \cdot x_2 \cdot \ldots \cdot x_D}$, is the geometric mean of x, and D is equal to 2,059.

The CLR-transformed bacterial abundances were adjusted with a unique general linear model, performed with the sasLM package in R (v4.0.2), and the fixed effects that were significant ($P<0.05$) for more than 10% of the OTUs were retained.

Finally, the model was:

$$y = \mu + \alpha * DIM + Nseq + Year + Run(Year) + Lact(Year) + Hour(Year) + e \quad (3)$$

where y is the CLR-transformed traits of the OTUs, μ is the overall mean, *DIM* is the lactation stage (from 28 to 133 days in milk) included as a covariable, *Nseq* is the number of sequences per sample as a fixed effect (7 levels from <5,000 to >30,000 sequences), *Year* is the year fixed effect (6 levels), *Run*(*Year*) is the fixed effect of run within year (5 levels), *Lact*(*Year*) is the fixed effect of lactation number within year (3 levels), *Hour*(*Year*) is the fixed effect of the hour of sampling in the morning/afternoon within year (8 levels), and e is the residual random effect.

The genetic line effect represented by differences among SCS+, SCS-, PERS+ and PERS- was not considered in the model since it was used as a discriminating factor in the multivariate discriminant analysis.

Phenotypic data

Daily recordings of milk production, milk somatic cell count (SCC) quantified with a Fossomatic cell counter (Foss, Nanterre, France), and milk fat and protein contents (FC and PC) were performed as part of the official milk recording of the flock. Ruminal samplings were performed between 0 and 3 days after the milking recordings were made in the morning and afternoon milkings. Two samples per animal were sent for analysis at the Interprofessional Milk Analysis Laboratory (Agrolab's Aurillac, France). Milk FC and PC were analysed with mid-infrared (MIR) techniques with a Milko-Scan™ FT6000 instrument (Foss, Nanterre, France). The daily milk production traits studied were daily milk yield (MP), daily FC and PC (as weighted averages), and daily SCS [$SCS = 3 + \log_2(SCC/100{,}000)$].

Moreover, for these official milk recordings (with the exception of those made in 2016), the MIR spectra were recovered in order to predict the fine profile of milk proteins and fatty acids. Fresh milk samples were analysed using MIR spectrometry [30]. The spectral data of the individual milk samples were obtained on a Milko-Scan™ FT6000 instrument (Foss, Nanterre, France). The proteins included in the analysis were the 4 caseins (CNs) α_{s1}-CN, α_{s2}-CN, β-CN and κ-CN and the 2 soluble proteins α-lactalbumin and β-lactoglobulin [31]. The saturated fatty acids (SFAs), unsaturated fatty acids (UFAs) and polyunsaturated fatty acids (PUFAs) included in the analysis were only the FAs used in ewe milk predictions [30], such as butyric acid (C4:0), caproic acid (C6:0), caprylic acid (C8:0), capric acid (C10:0), lauric acid (C12:0), palmitic acid (C16:0), oleic acid (*cis-9* C18:1), conjugated linoleic acid (*cis-9 trans-11* C18:2) and α-linoleic acid (C18:3*n-3*). Milk proteins and fatty acids are expressed in g per 100 ml.

The daily FC and PC, milk proteins and milk FAs were CLR transformed to account for their compositional nature, and all traits were adjusted using the sasLM package in R (v4.0.2) according to:

$$y = \mu + \alpha * DIM + Year + Lact(Year) + e \quad (4)$$

where y is the milk production traits; μ is the overall mean; *DIM* is the lactation stage (from 28 to 133 days in milk) included as a covariable; *Year* is the fixed effect of year (6 levels); *Lact*

(*Year*) is the fixed effect of lactation number within year (3 levels); and *e* is the residual random effect.

Multivariate analysis

The multivariate analysis was performed with the residuals obtained from Eq (3) for bacterial abundances and Eq (4) for milk traits.

Two discriminant analyses were performed on OTU abundances to discriminate the divergent lines (for SCS and PERS separately), using sparse partial least-squares discriminant analysis (sPLS-DA). The number of components selected was based on principal component analysis, from which the sum of components explained at least 60% of the variation. The number of variables was selected using the CLR-lasso penalty method considering the optimal number as a function of the lambda value after 25-fold cross-validation. The loading values indicate the weight of a subset of OTUs whose linear combination maximizes the differences between genetic lines.

Regression analyses of the relationships of ruminal bacteria with milk production traits and MIR-predicted traits performed on all divergent lines together, using sparse partial least-squares (sPLS) analysis. A single sPLS analysis was carried out for milk production traits and fine milk FA and protein compositions predicted with MIR. The analysis included 561 ewes with information for all traits. As previously described, principal components analysis and the CLR-lasso penalty method were used to define the numbers of components and variables for the sPLS model. The multivariate analysis were implemented using mixOmics package [32] in R (V4.0.2). A Pearson correlation matrix was calculated with only the OTUs selected according to the first principal component (PC1) and second principal component (PC2) of the corresponding sPLS analysis. Statistical significance was declared at a P value <0.05. Then, clustering of OTUs and traits was performed with the heatmaply function in R (v4.0.2).

The classification reliability corresponding to the discriminant analysis model was assessed as a function of the maximum prediction distance between the overall misclassification error rate and balanced error rate (BER) after fivefold cross-validation repeated 10 times. BER was calculated as 1 –balanced accuracy.

Results

As a result of the bioinformatics analysis, 9,536,442 sequences were retained (63% of the initial total DNA sequences). The abundance table included 2,059 affiliated OTUs, represented by 751 to 168,617 sequences (mean of 1,761 DNA sequences). Rare OTUs represented by fewer than 2,034 sequences across all samples were excluded from the analysis. Genera were defined as the finest taxonomic level due to an unknown species frequency of 94%.

Overall, the 2,059 OTUs from the 700 samples were attributed to 11 phyla and 112 genera. Expressed as a percentage of total sequences for all samples, the most representative phyla were *Bacteroidetes* (50.8%), *Firmicutes* (43.3%) and *Proteobacteria* (2.7%), and the most abundant genera were *Prevotella* (34%), *Lachnospiraceae NK3A20 group* (6.4%), *Ruminococcus* (5.8%), *Christensenellaceae R-7 group* (5.3%), *Oscillospiraceae NK4A214 group* (3.8%) and *Rikenellaceae RC9 gut* (3.6%). The percentage of zero values in whole abundance table is shown in Fig 1.

Discriminant analysis of SCS and PERS lines

Divergent selection created large differences between the lines: 2.19 units of SCS EBVs (i.e., 3.6 genetic SD) created between the 94 SCS+ and 204 SCS- ewes and 5.52 units of milk CV EBVs (i.e., 2.1 genetic SD) created between the 200 PERS+ and 202 PERS- ewes.

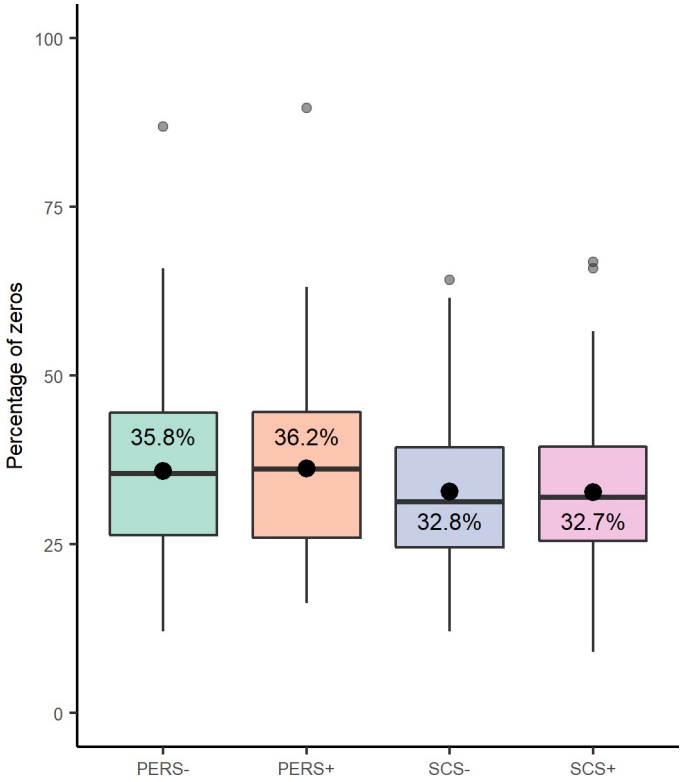


Fig 1. Percentage of zero values in data by genetic line. SCS+ and SCS- as somatic cell score lines susceptibility/resistance, and PERS+ and PERS- as milk persistency line high/low persistence.

https://doi.org/10.1371/journal.pone.0254874.g001

The discriminant model defined for SCS lines included 100 principal components (63% of variance explained) and 17 variables. The SCS+ and SCS- ewes could be discriminated on the basis of their ruminal bacteria (Fig 2). Table 1 includes the 34 OTUs most associated with the SCS lines in each of the first 2 principal components. Only two OTUs were removed from the abundance table because of zero values for all samples. The BER obtained from the model was 0.50, and the first two principal components explained 4% of the variance.

The *Prevotella* genus was well represented, with 11 OTUs associated with either the SCS + or the SCS- ewes, through components 1 and 2 (Table 1). Only OTU1145 was associated with SCS- ewes for the two main components. The *Christensenellaceae R-7 group* genus appeared to be associated with SCS- ewes in PC1, but in PC2, OTU285 and OTU382 belonging to this same genus were associated with SCS+ ewes. The family *Lachnospiraceae* was well represented by *Lachnospiraceae NK3A20 group*, *Lachnospiraceae NK4A136 group* and *Lachnospiraceae AC2044 group*, which were associated with either SCS- or SCS+ ewes.

The discriminant model for PERS lines included 120 principal components (62% of variance explained) and 5 variables. The PERS+ and PERS- lines could not be discriminated according to their ruminal bacteria (Fig 3). Table 2 includes the 10 OTUs most associated with the PERS lines in each of the first 2 principal components. The BER obtained from the model was 0.71, and the first two principal components explained 2% of the variance. The *Prevotella* genus, represented by OTU1482 (PC1) and OTU1395 (PC2), was positively associated with PERS- ewes. In addition, the PERS- line was associated with *Oscillospiraceae NK4A214*, *Blautia* and an unknown genus (order *Clostridia UCG-014*) through component 1 and with *Streptococcus* through component 2. Thus, the genera *Ruminococcus* and *Oscillospiraceae NK4A214* were associated with PERS+ ewes.

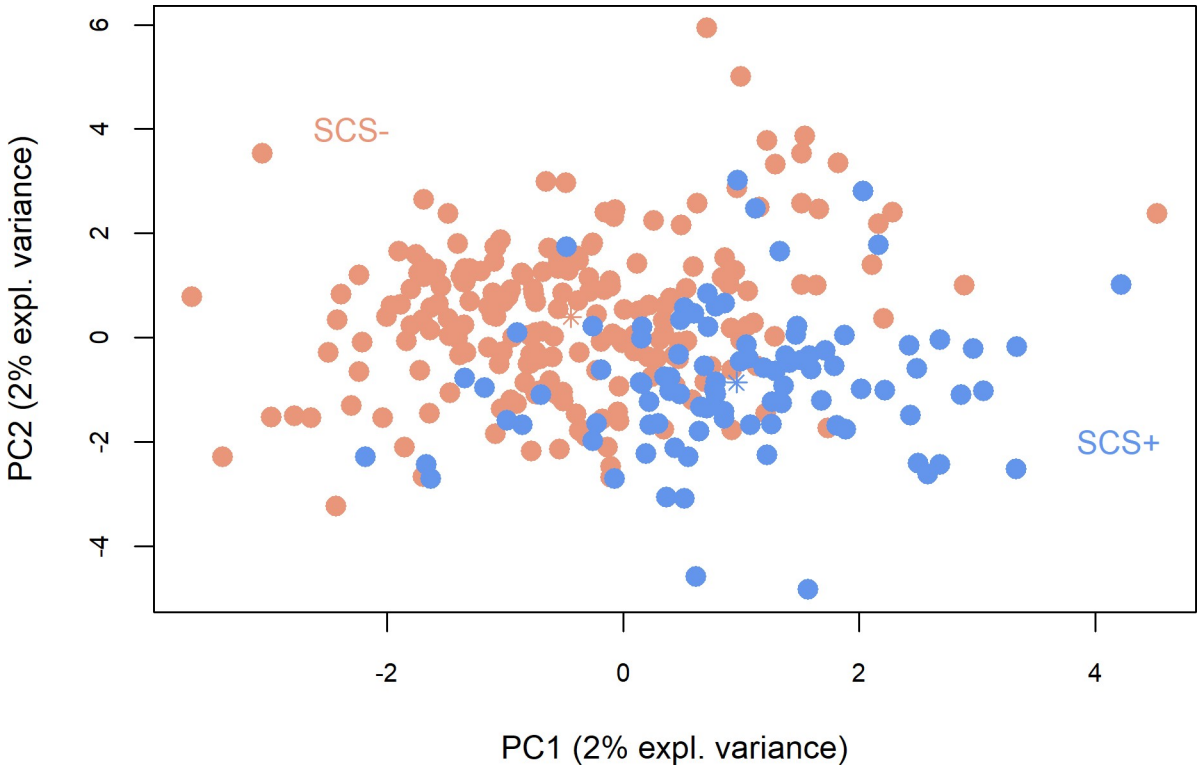


Fig 2. Sparse partial least squares discriminant analysis between divergent somatic cell score (SCS) lines. SCS+: ewes selected for high somatic cell score i.e. susceptible to mastitis; SCS-: ewes selected for low somatic cell score i.e. resistant to mastitis.

https://doi.org/10.1371/journal.pone.0254874.g002

Regression analysis between ruminal bacterial abundance and milk traits

The sPLS regression model included 150 components and 9 variables. Fig 4 shows only the 17 most representative OTUs from PC1 and PC2 (OTU1593 was representative for both components and all traits).

Daily milk production and SCS were each correlated with one OTU of the genus *Prevotella* (Fig 4). Milk FC and PC had similar correlations with 5 common OTUs: they were negatively correlated with 2 *Prevotella* OTUs (R = [-0.11;-0.13], P< 0.01) and with *Suttonella* (R = [-0.09;-0.12], P< [0.05;0.01]) and positively correlated with *Lachnospiraceae NK4A136 group* (R = [0.10;0.15], P< [0.05;0.01]) and *Christensenellaceae R-7 group* (R = [0.09;0.11], P< [0.05;0.01]). Moreover, PC was specifically correlated with *Endomicrobium*, while FC had numerous correlations, such as positive correlations with *Lachnospiraceae probable genus 10* and *Christensenellaceae R-7 group*, negative correlations with 2 *Ruminococcus* OTUs and variable correlations with 2 OTUs of the *Muribaculaceae* family (R = [0.12;-0.14], P< 0.01).

α_{s1}-CN, κ-CN and β-lactoglobulin were positively correlated with *Lachnospiraceae NK4A136 group* (R = [0.13;0.16], P< 0.01) and negatively correlated with *Prevotella* and *Suttonella* (R = [-0.14;-0.16], P< 0.01). To a lesser extent, *Christensenellaceae R-7 group* and the family *Muribaculaceae* showed positive and negative correlations with α_{s1}-CN and κ-CN, respectively (Fig 4). α_{s2}-CN and β-CN exhibited the same trend as the other caseins but with weaker correlations: negative correlations with *Suttonella* and *Prevotella* and positive correlations with *Lachnospiraceae NK4A136 group*. α-Lactalbumin was clearly different from the other protein since it was positively correlated with *Prevotella* and with an unknown genus of the family *Muribaculaceae* (R = [0.15;0.18], P< 0.001), while the families *Lachnospiraceae* and p-251-o5 showed negative correlations with this protein (R = -0.11, P< 0.01).

Table 1. Loading values per OTU with genus affiliation, associated genetic line and percentage of abundance, for the two first components from the somatic cell score (SCS) line sparse partial least squares discriminant analysis.

Affiliated *genus*	OTU	PC1	Line	Abundance (%)	Affiliated *genus*	OTU	PC2	Line	Abundance (%)
Prevotella	14	0.40	SCS+	0.80	*Prevotella*	1064	0.35	SCS-	0.01
Ruminococcaceae/ unknown genus	501	0.34	SCS+	0.04	*Prevotella*	1092	0.34	SCS-	0.05
Prevotella	819	0.27	SCS+	0.01	*Ruminococcus*	657	0.32	SCS-	0.02
Prevotellaceae UCG-001	467	0.25	SCS+	0.08	*Prevotella*	112	0.23	SCS+	0.13
F082/unknown genus	3286	0.13	SCS+	0.01	*Lachnospiraceae NK4A136 group*	2186	0.22	SCS-	0.01
Lachnospiraceae AC2044 group	230	0.08	SCS+	0.08	*Prevotellaceae YAB2003 group*	517	0.17	SCS-	0.10
Prevotella	1486	0.08	SCS+	0.01	*Anaeroplasma*	220	0.16	SCS-	0.06
Lachnospiraceae NK3A20 group	625	-0.06	SCS-	0.02	*Prevotella*	1145	0.11	SCS-	0.03
Clostridia UCG-014/ unknown genus	530	-0.09	SCS-	0.05	*Christensenellaceae R-7 group*	382	-0.02	SCS+	0.19
Prevotella	1145	-0.12	SCS-	0.03	*F082/unknown genus*	2208	-0.03	SCS+	0.01
Moryella	328	-0.15	SCS-	0.04	*Pseudoramibacter*	226	-0.08	SCS+	0.04
Syntrophococcus	1739	-0.19	SCS-	0.01	*Christensenellaceae R-7 group*	285	-0.19	SCS+	0.17
Fibrobacter	1640	-0.23	SCS-	0.02	*Family XIII UCG-001*	450	-0.24	SCS+	0.04
F082/ unknown genus	175	-0.26	SCS-	0.08	*Mogibacterium*	79	-0.26	SCS-	0.17
Prevotella	367	-0.33	SCS-	0.05	*Prevotella*	443	-0.29	SCS+	0.05
Prevotella	1842	-0.34	SCS-	0.01	*Lachnospiraceae NK3A20 group*	39	-0.31	SCS+	0.27
Christensenellaceae R-7 group	2009	-0.36	SCS-	0.01	*Oscillospiraceae/ unknown genus*	538	-0.36	SCS+	0.04

When the genus is unknown, the family affiliation is included before the backslash (/).

OTU, identification number of the operational taxonomic unit.

PC1, first principal component; PC2, second principal component.

Abundance (%), expressed as percentage of total sequences.

https://doi.org/10.1371/journal.pone.0254874.t001

The strongest correlations were observed with SFAs, which were negatively correlated with all 4 *Prevotella* OTUs selected by sPLS analysis and with *Suttonella*, particularly for C10:0 and C12:0 (R = -0.16, P< 0.001). Some genera of the phylum *Firmicutes* were correlated with SFAs. For example, 2 OTUs belonging to *Christensenellaceae R-7 group* were positively correlated with SFAs, and 2 OTUs belonging to *Ruminococcus* were negatively correlated with SFAs. An unknown genus of the *Muribaculaceae* and the p-251-o5 family showed the maximum correlations of 0.20 (P< 0.001) with C10:0.

Compared to SFAs, MUFAs had fewer significant correlations with OTUs. As presented in Fig 4, the MUFA *cis-9* C18:1 was negatively associated with *Endomicrobium* and *Prevotella*, while the PUFA C18:3*n-3* was positively associated with *Christensenellaceae R-7 group* and *Probable genus 10* and negatively associated with *Prevotella* and an unknown genus of *Muribaculaceae*. Finally, *cis-9 trans-11 C18:2* was not correlated with any of the 17 OTUs selected by sPLS analysis.

Discussion

Bacteroidetes, *Firmicutes* and *Proteobacteria* were the most dominant phyla in the rumen of dairy ewes. The same phyla were reported by other authors studying sheep [11, 33] and dairy cows [7–9, 34–37] with different rumen sampling methods and statistical analysis.

The analysis of microbiome abundance data with the commonly applied methodology [7–9, 34, 37], i.e., data treatment with a normalization process, such as rarefaction, and using non-metric distances (i.e., Bray Curtis), provides results that seem satisfactory, irrespective of the compositional nature of the data. However, statistical knowledge since Pearson [38] has shown that processing such data without considering them as compositional could lead to spurious

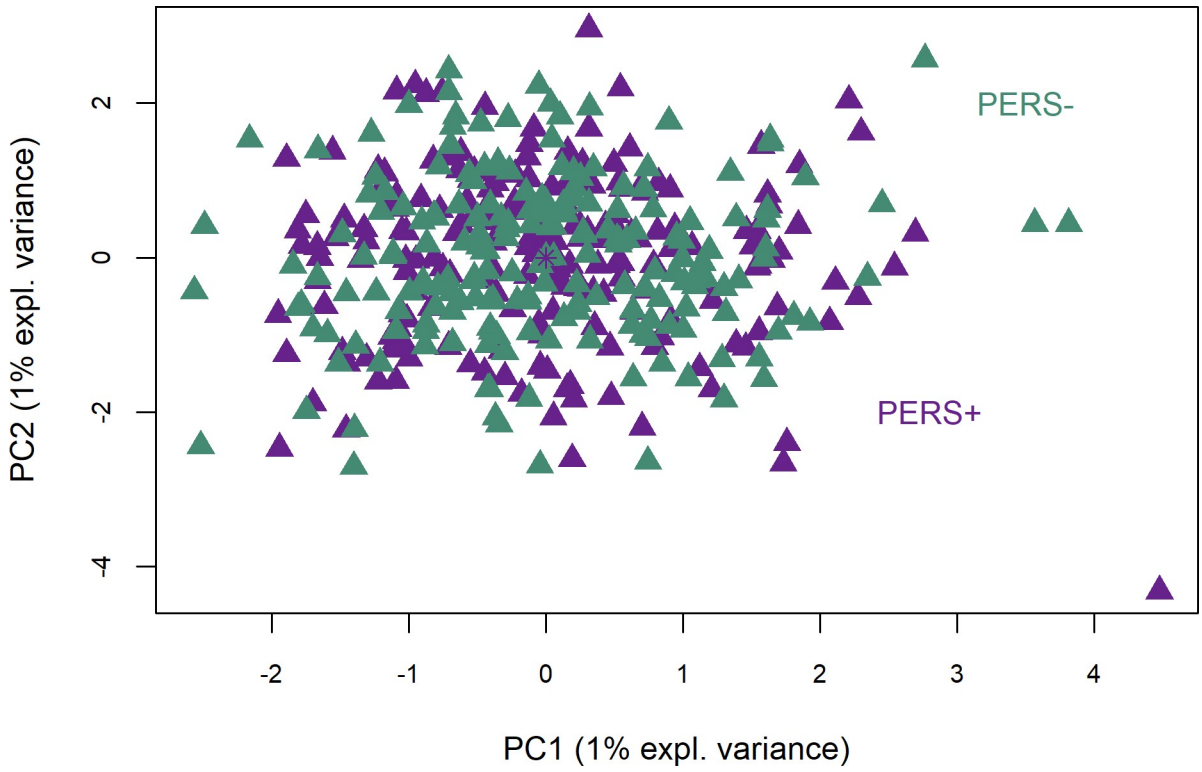


Fig 3. Sparse partial least squares discriminant analysis between divergent milk persistency (PERS) lines. PERS+: ewes selected for a high milk persistence; PERS-: ewes selected for a low milk persistence.

https://doi.org/10.1371/journal.pone.0254874.g003

correlations. More recently, Gloor et al. [14] pointed out that the use of traditional methods to analyse data without considering their compositional nature can lead to "misleading and unpredictable" results [13, 14, 16].

Thus, this work aimed to apply compositional data analysis to the rumen bacterial metagenome obtained by metabarcoding and to correct for technical and zootechnical effects in order to obtain robust and reproducible results. The compositional workflow of the study consisted of the following steps:

Table 2. Loading values per OTU with genus affiliation, associated genetic line and percentage of abundance, for the two first components from the milk persistency (PERS) line sparse partial least squares discriminant analysis.

Affiliated *genus*	OTU	PC1	Line	Abundance (%)	Affiliated *genus*	OTU	PC2	Line	Abundance (%)
Prevotella	1428	0.79	PERS-	0.01	*[Eubacterium] coprostanoligenes group/ unknown genus*	1501	0.39	PERS+	0.01
Clostridia UCG-014/ unknown genus	411	-0.22	PERS-	0.04	*Prevotella*	1395	0.11	PERS-	0.01
Blautia	216	-0.29	PERS-	0.07	*Streptococcus*	634	-0.15	PERS-	0.04
Oscillospiraceae NK4A214	292	-0.34	PERS-	0.07	*Anaerovoracaceae/ unknown genus*	1131	-0.58	PERS+	0.02
Ruminococcus	537	-0.36	PERS +	0.05	*Oscillospiraceae NK4A214*	823	-0.69	PERS+	0.02

When the genus is unknown, the family affiliation is included before the backslash (/).

OTU, identification number of the operational taxonomic unit.

PC1, first principal component; PC2, second principal component.

Abundance (%), expressed as percentage of total sequence.

https://doi.org/10.1371/journal.pone.0254874.t002

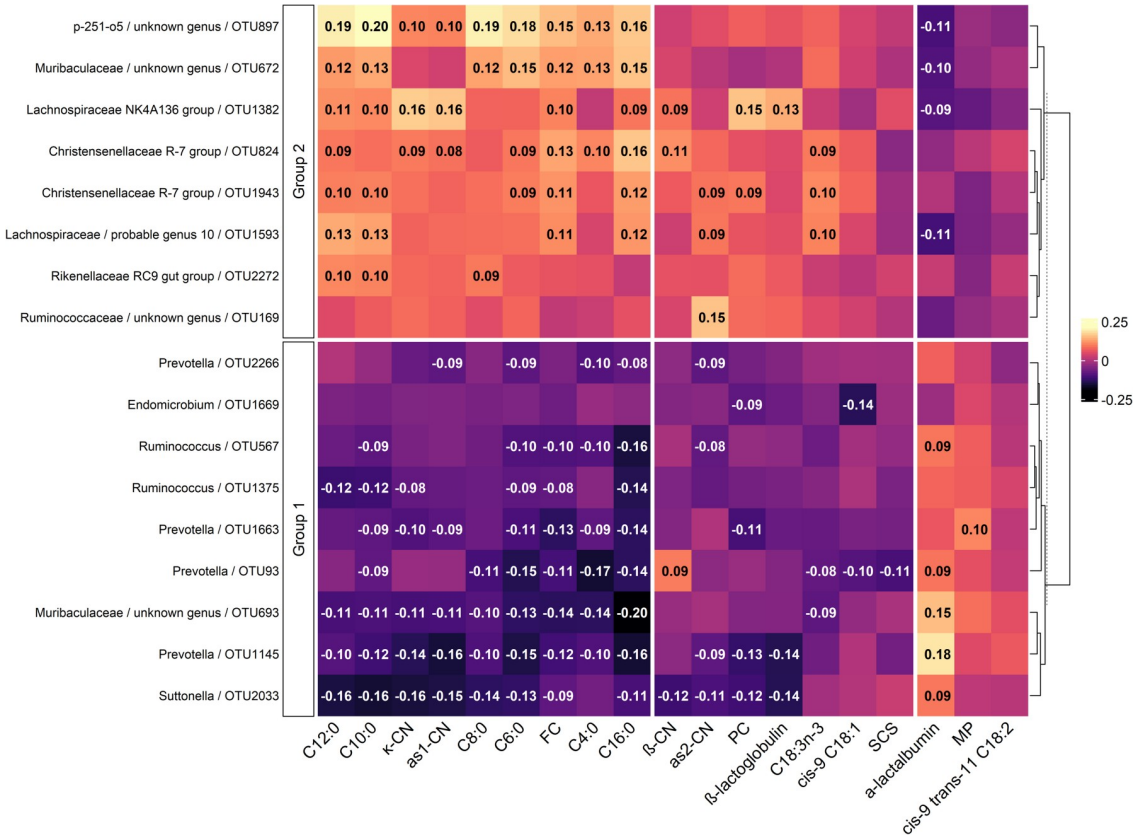


Fig 4. A correlation matrix heatmap between bacterial taxa and milk traits. OTUs selected by the 2 first components of the sparse least squares analysis; daily milk production (MP), somatic cell score (SCS), daily milk protein contents (PC), daily milk fat content (FC), milk fatty acids (butyric acid (C4:0), caproic acid (C6:0), caprylic acid (C8:0), capric acid (C10:0), lauric acid (C12:0), palmitic acid (C16:0), oleic acid (*cis-9* C18:1), conjugated linoleic acid (*cis-9 trans-11* C18:2) and α-linoleic acid (C18:3*n-3*), expressed as % of total fatty acids) and milk proteins, as casein (CN) (α_{s1}-CN, α_{s2}-CN, β-CN, κ-CN, expressed as % of total proteins), and soluble proteins (α-lactalbumin and β-lactoglobulin, expressed as % of total proteins).

https://doi.org/10.1371/journal.pone.0254874.g004

1. Zero values were corrected with the GBM method [19]. Theoretically, this method is appropriate since it generates a minor distortion in the ratios between OTU abundances, based on the correction of zero values and the multiplication of non-zero values. In addition, the GBM method considers the multivariate nature of microbiome data through a Bayesian model, where new values are generated on the basis of the posterior probabilities of zero values in the raw data.

2. OTU abundance was CLR transformed. This transformation allows a simple interpretation of the biological results, since each OTU in each sample is compared with the geometric mean of the sample. The limit of CLR transformation is that OTUs remain dependent because of the use of the geometric mean. Therefore, CLR transformation partially solves the problem identified by Pearson in 1897 [38]. However, the statistically correct alternative to CLR transformation proposed by Egozcue et al. [17], i.e., the ILR transformation, does not allow easy interpretation of the results [39]. Indeed, ILR transformation works with balances (linear combinations of OTUs) to achieve total independence among the OTUs, and it is currently not possible to back-transform the results after multivariate analyses. Further work is needed in this sense.

3. The microbiome and phenotypic data were adjusted through linear models. These models must include in their definition batch effects [20] which are any unwanted source of

variation representing biological and technical effects. When the effects are balanced in the experiment, linear models are an interesting method to correct for batch effects [20]. From these models, the residual values (variation not explained by the included effects), which are ultimately the input values for multivariate analyses, are obtained. However, the consequence of using residual values for sPLS-DA and sPLS analyses is that the remaining variation in the residuals exploited by these regression models is reduced, as shown below.

In this way, we considered not only the nature of the available microbiome data to work in the appropriate geometric space (Euclidean) but also the residuals to allow a more correct analysis of the effect under study, i.e., the genetic lines based on SCS and PERS.

Discriminant analysis

Discriminant analysis for both the SCS and PERS lines showed low explained variance (Figs 2 and 3) for the first two principal components. Using residuals leads to a smaller variance of the values and therefore affects the variance explained by the discriminant effect for the first components. As a result, it is necessary to include a large number of components in the analysis. Variable selection was performed through the CLR-lasso method and allowed some OTUs with low abundance that carried irrelevant information to be excluded.

Since all ewes were Lacaune breed receiving the same diet and batch effects (except that of line) were corrected for by the linear model, the remaining variation in the rumen bacteria may be explained by the genetic lines. In spite of this, the variance among the genetic lines was not explained by the composition of the host animal's microbiome for PERS, and only slightly for SCS. This is demonstrated by the BER obtained for the sPLS-DA analyses of 0.50 for SCS and 0.71 for PERS. Nevertheless, Fig 2 shows a slight difference between the SCS+ and SCS- lines, despite only 4% of the total variance being explained (for both components). Some OTUs assigned to *Prevotella*, *Christensenellaceae R-7 group* and unknown genus of the family *Ruminococcaceae* were the main discriminants for the first component (Table 1). Zhong et al. [36] did not report differences in these three genera between the rumens of cows with phenotypically high and low SCCs; in this comparison, the authors noticed only enrichment of *Proteobacteria* (especially an unclassified *Succinivibrionaceae*) in the ruminal microbiota of cows with low SCCs. In our study, these OTUs were not significantly different according to SCS. Therefore, the hypothesis of a link between selection on SCS and modifications in the rumen microbial population was not rejected, but its validity remains unclear in terms of the bacteria involved. The PERS line analysis revealed a complete absence of differences between PERS + and PERS- ewes, as shown in Fig 3. However, three OTUs presented loading values greater than 0.5 (Table 2) along PC1 and PC2, and they belonged to *Prevotella*, *Oscillospiraceae NK4A214* and an unknown genus of *Anaerovoracaceae*. Nevertheless, there was no hypothesis of a correlated response of ruminal microbiome abundance to PERS selection.

The results for both divergent lines suggest that genetic selection for zootechnical traits, such as udder health and milk production curves, did not modify the abundance of rumen bacteria and therefore the animals' ability to digest their feed.

Links between ruminal bacteria and milk traits

Daily milk production was positively associated with a *Prevotella* OTU, similar to the results reported by Huang et al. [40]. This genus is known to have major metabolic activity in the production of propionate [41], which is the main precursor for gluconeogenesis in the liver [1] leading to lactose production. Some authors [9, 35] reported that some genera of the *Lachnospiraceae* family were positively correlated with daily milk yield, while we found that two OTUs

affiliated with this family were weakly but negatively associated with daily milk production, as reported by Huang et al. [40]. The results obtained in dairy cows can be considered as references for dairy sheep, since as shown [42] the differences in terms of rumen microbiota among species are smaller when the diet is based on a mixture of forage and concentrates.

The SCS was correlated with a *Prevotella* OTU, but a possible association between this genus and the SCC in milk has not been reported, and these results are in line with the difficulty of differentiating the genetic lines selected for SCS. As expressed by Zhong et al. [36], the bacterial communities in the rumen are stable in animals with different SCCs, and this is probably true of ewes, where mastitis is overwhelmingly sub-clinical. However, the main hypothesis is a link between the intestinal microbiota and intramammary infection (i.e., clinical mastitis) [43].

Concerning milk composition, we identified two groups of OTUs (Fig 4): group 1, with negatively linked OTUs belonging to *Prevotella*, *Suttonella*, *Ruminococcus* and *Endomicrobium*, and group 2, with positively linked OTUs belonging *Lachnospiraceae NKA136*, *probable genus 10*, *Rikenellaceae RC9*, *Ruminococcaceae*, *Christensenellaceae* and *p-251-o5*. Muribaculaceae was represented by one OTU in the two groups, and for α-lactalbumin and daily milk production, the relation was reversed. Group 1 was represented mostly by propionic acid and proteolytic bacteria such as *Prevotella* [41], *Suttonella* [44] and some *Ruminococcus* [45], characterized by increasing milk production with a possible dilution of milk components. In contrast, group 2, with mostly butyric and acetic acid-producing bacteria such as *Lachnospiraceae* [46], had less proteolytic activity [47, 48], leading to the opposite effect for the concentration of milk components. These results are in accordance with other studies in dairy cows that also found *Prevotella*ceae family negatively correlated with milk fat, and *Lachnospiraceae* positively correlated with milk fat and protein contents [8, 37, 49].

In sheep [33] as well as in cows [50] a close relationship between the rumen microbiota composition and short-chain fatty acids in rumen was reported that could influence the synthesis of milk components. Therefore, the most likely hypothesis is that bacteria of group 2, through butyric and acetic acid, promote the production of short- and medium-chain SFAs.

In conclusion, this study applying the compositional data approach to a significant sample size of Lacaune dairy ewes revealed that rumen bacteria belonging to *Prevotella*, *Suttonella*, *Ruminococcaceae* and *Lachnospiraceae* are associated with milk production traits such as milk fatty acids and proteins. However, despite the large genetic differences between lines, ruminal bacteria were able to only weakly discriminate between SCS lines and unable to discriminate between PERS lines. Although dilution of the ruminal samples by saliva could be expected, correction of the rumen microbiota for the number of sequences per sample could have reduced this effect.

Since some abundant OTUs were correlated with milk composition traits, it would be interesting to further investigate the mechanism by which rumen bacterial metabolites affect milk composition traits in order to understand the relationships detected in this work.

Acknowledgments

The authors would like to thank Rachel Rupp and Hélène Larroque for access to the genetic resources produced, Beatrice Gabinaud and Marie-Luce Chemit for DNA extraction, and La Fage's technicians for animal care and particularly rumen sampling.

Author Contributions

Conceptualization: Annabelle Meynadier, Christel Marie-Etancelin.

Data curation: Guillermo Martinez Boggio, Annabelle Meynadier, Christel Marie-Etancelin.

Formal analysis: Guillermo Martinez Boggio.

Funding acquisition: Annabelle Meynadier, Christel Marie-Etancelin.

Investigation: Guillermo Martinez Boggio.

Methodology: Guillermo Martinez Boggio, Annabelle Meynadier, Pepus Daunis-i-Estadella, Christel Marie-Etancelin.

Supervision: Annabelle Meynadier, Christel Marie-Etancelin.

Writing – original draft: Guillermo Martinez Boggio.

Writing – review & editing: Guillermo Martinez Boggio, Annabelle Meynadier, Pepus Daunis-i-Estadella, Christel Marie-Etancelin.

References

1. Stewart CS, Flint HJ, Bryant MP. The rumen bacteria. In: Hobson PN, Stewart CS (eds). The Rumen Microbial Ecosystem. Blackie academic & professional, London, UK, 1997, pp 10–55.
2. Hernandez-Sanabria E, Goonewardene LA, Wang Z, Zhou M, Moore SS, Guan LL. Influence of Sire Breed on the Interplay among Rumen Microbial Populations Inhabiting the Rumen Liquid of the Progeny in Beef Cattle. PLoS ONE. 2013; 8(3):e58461. https://doi.org/10.1371/journal.pone.0058461 PMID: 23520513
3. Li F, Guan LL. Metatranscriptomic Profiling Reveals Linkages between the Active Rumen Microbiome and Feed Efficiency in Beef Cattle. Appl Environ Microbiol. 2017; 83(9):e00061–17. https://doi.org/10.1128/AEM.00061-17 PMID: 28235871
4. Difford GF, Plichta DR, Løvendahl P, Lassen J, Noel SJ, Højberg O, et al. Host genetics and the rumen microbiome jointly associate with methane emissions in dairy cows. PLoS Genet. 2018; 14(10): e1007580. https://doi.org/10.1371/journal.pgen.1007580 PMID: 30312316
5. Kittelmann S, Pinares-Patiño CS, Seedorf H, Kirk MR, Ganesh S, McEwan JC, et al. Two Different Bacterial Community Types Are Linked with the Low-Methane Emission Trait in Sheep. PLoS ONE. 2014; 9 (7):e103171. https://doi.org/10.1371/journal.pone.0103171 PMID: 25078564
6. Wallace RJ, Sasson G, Garnsworthy PC, Tapio I, Gregson E, Bani P, et al. A heritable subset of the core rumen microbiome dictates dairy cow productivity and emissions. Sci Adv. 2019; 5(7):eaav8391. https://doi.org/10.1126/sciadv.aav8391 PMID: 31281883
7. Bainbridge ML, Cersosimo LM, Kraft J. Rumen bacterial communities shift across a lactation in Holstein, Jersey and Holstein × Jersey dairy cows and correlate to rumen function, bacterial fatty acid composition and production parameters. FEMS Microbiology Ecology. 2016; 92:fiw059.
8. Jami E, White BA, Mizrahi I. Potential Role of the Bovine Rumen Microbiome in Modulating Milk Composition and Feed Efficiency. PLoS ONE. 2014; 9(1):e85423. https://doi.org/10.1371/journal.pone.0085423 PMID: 24465556
9. Tong J, Zhang H, Yang D, Zhang Y, Xiong B, Jiang L. Illumina sequencing analysis of the ruminal microbiota in high-yield and low-yield lactating dairy cows. PLoS ONE. 2018; 13(11):e0198225. https://doi.org/10.1371/journal.pone.0198225 PMID: 30423588
10. Toral PG, Belenguer A, Shingfield KJ, Hervás G, Toivonen V, Frutos P. Fatty acid composition and bacterial community changes in the rumen fluid of lactating sheep fed sunflower oil plus incremental levels of marine algae. Journal of Dairy Science. 2012; 95:794–806. https://doi.org/10.3168/jds.2011-4561 PMID: 22281344
11. Castro-Carrera T, Toral PG, Frutos P, McEwan NR, Hervás G, Abecia L, et al. Rumen bacterial community evaluated by 454 pyrosequencing and terminal restriction fragment length polymorphism analyses in dairy sheep fed marine algae. Journal of Dairy Science. 2014; 97:1661–1669. https://doi.org/10.3168/jds.2013-7243 PMID: 24440247
12. Belenguer A, Toral PG, Frutos P, Hervás G. Changes in the rumen bacterial community in response to sunflower oil and fish oil supplements in the diet of dairy sheep. Journal of Dairy Science. 2010; 93:3275–3286. https://doi.org/10.3168/jds.2010-3101 PMID: 20630243
13. Gloor GB, Reid G. Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. Can J Microbiol. 2016; 62(8):692–703. https://doi.org/10.1139/cjm-2015-0821 PMID: 27314511

14. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome Datasets Are Compositional: And This Is Not Optional. Front Microbiol. 2017; 8:2224. https://doi.org/10.3389/fmicb.2017.02224 PMID: 29187837

15. Aitchison JA. The Statistical Analysis of Compositional Data. Chapman & Hall Ltd.: New York, USA, 1986.

16. Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R. Compositional data and their sample space. In: Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R. (eds). Modelling and analysis of compositional data. Wiley, Chichester, West Sussex, UK, 2015, pp 8–21.

17. Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barcelo-Vidal C. Isometric Logratio Transformations for Compositional Data Analysis. Mathematical Geology. 2003; 35(3):279–300.

18. Fernandes AD, Macklaim JM, Linn TG, Reid G, Gloor GB. ANOVA-Like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq. PLoS ONE. 2013; 8(7):e67019. https://doi.org/10.1371/journal.pone.0067019 PMID: 23843979

19. Martín-Fernández J-A, Hron K, Templ M, Filzmoser P, Palarea-Albaladejo J. Bayesian-multiplicative treatment of count zeros in compositional data sets. Statistical Modelling. 2015; 15(2):134–158.

20. Wang Y, LêCao K-A. Managing batch effects in microbiome data. Briefings in Bioinformatics. 2019; 0:1–17.

21. Rupp R, Lagriffoul G, Astruc JM, Barillet F. Genetic Parameters for Milk Somatic Cell Scores and Relationships with Production Traits in French Lacaune Dairy Sheep. Journal of Dairy Science. 2003; 86 (4):1476–1481. https://doi.org/10.3168/jds.S0022-0302(03)73732-1 PMID: 12741573

22. Rupp R, Bergonier D, Dion S, Hygonenq MC, Aurel MR, Robert-Granié C, et al. Response to somatic cell count-based selection for mastitis resistance in a divergent selection experiment in sheep. Journal of Dairy Science. 2009; 92(3):1203–1219. https://doi.org/10.3168/jds.2008-1435 PMID: 19233814

23. Henderson G, Cox F, Kittelmann S, Miri VH, Zethof M, Noel SJ, et al. Effect of DNA Extraction Methods and Sampling Techniques on the Apparent Structure of Cow and Sheep Rumen Microbial Communities. PLoS ONE. 2013; 8:e74787. https://doi.org/10.1371/journal.pone.0074787 PMID: 24040342

24. Liu Z, Lozupone C, Hamady M, Bushman FD, Knight R. Short pyrosequencing reads suffice for accurate microbial community analysis. Nucleic Acids Research. 2007; 35:e120–e120. https://doi.org/10.1093/nar/gkm541 PMID: 17881377

25. Andersson AF, Lindberg M, Jakobsson H, Bäckhed F, Nyrén P, Engstrand L. Comparative Analysis of Human Gut Microbiota by Barcoded Pyrosequencing. PLoS ONE. 2008; 3:e2836. https://doi.org/10.1371/journal.pone.0002836 PMID: 18665274

26. Escudié F, Auer L, Bernard M, Mariadassou M, Cauquil L, Vidal K, et al. FROGS: Find, Rapidly, OTUs with Galaxy Solution. Bioinformatics. 2018; 34(8):1287–1294. https://doi.org/10.1093/bioinformatics/btx791 PMID: 29228191

27. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, et al. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nat Methods. 2013; 10(1):57–59. https://doi.org/10.1038/nmeth.2276 PMID: 23202435

28. Palarea-Albaladejo J, Martín-Fernández JA. zCompositions—R package for multivariate imputation of left-censored data under a compositional approach. Chemometrics and Intelligent Laboratory Systems. 2015; 143:85–96.

29. van den Boogaart KG, Tolosana-Delgado R. "compositions": A unified R package to analyze compositional data. Computers & Geosciences. 2008; 34(4):320–338.

30. Ferrand-Calmels M, Palhière I, Brochard M, Leray O, Astruc JM, Aurel MR, et al. Prediction of fatty acid profiles in cow, ewe, and goat milk by mid-infrared spectrometry. Journal of Dairy Science. 2014; 97 (1):17–35. https://doi.org/10.3168/jds.2013-6648 PMID: 24268398

31. Ferrand M, Miranda G, Larroque H, Leray O, Guisnel S, Lahalle F, et al. Determination of protein composition in milk by mid-infrared spectrometry. In ICAR 38. Annual Meeting 2012. p. 5.

32. Rohart F, Gautier B, Singh A, Lê Cao K-A. mixOmics: an R package for 'omics feature selection and multiple data integration. PLoS Comput Biol. 2017; 13(11):e1005752. https://doi.org/10.1371/journal.pcbi.1005752 PMID: 29099853

33. Li H, Yu Q, Li T, Shao L, Su M, Zhou H, et al. Rumen Microbiome and Metabolome of Tibetan Sheep (Ovis aries) Reflect Animal Age and Nutritional Requirement. Front Vet Sci. 2020; 7:609. https://doi.org/10.3389/fvets.2020.00609 PMID: 32984417

34. Jami E, Mizrahi I. Composition and Similarity of Bovine Rumen Microbiota across Individual Animals. PLoS ONE. 2012; 7(3):e33306. https://doi.org/10.1371/journal.pone.0033306 PMID: 22432013

35. Xue M, Sun H, Wu X, Guan LL, Liu J. Assessment of Rumen Microbiota from a Large Dairy Cattle Cohort Reveals the Pan and Core Bacteriomes Contributing to Varied Phenotypes. Appl Environ Microbiol. 2018; 84(19):e00970–18. https://doi.org/10.1128/AEM.00970-18 PMID: 30054362

36. Zhong Y, Xue M, Liu J. Composition of Rumen Bacterial Community in Dairy Cows With Different Levels of Somatic Cell Counts. Front Microbiol. 2018; 9:3217. https://doi.org/10.3389/fmicb.2018.03217 PMID: 30619238

37. Lima FS, Oikonomou G, Lima SF, Bicalho MLS, Ganda EK, de Oliveira Filho JC, et al. Prepartum and Postpartum Rumen Fluid Microbiomes: Characterization and Correlation with Production Traits in Dairy Cows. Appl Environ Microbiol. 2015; 81(4):1327–1337. https://doi.org/10.1128/AEM.03138-14 PMID: 25501481

38. Pearson K. Mathematical contributions to the theory of evolution.—On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proc R Soc Lond. 1897; 60:489–498.

39. Greenacre M, Grunsky E. The isometric logratio transformation in compositional data analysis: a practical evaluation. Barcelona: Universitat Pompeu Fabra; 2019 p. 44. Report No.: 1627.

40. Huang S, Ji S, Suen G, Wang F, Li S. The Rumen Bacterial Community in Dairy Cows Is Correlated to Production Traits During Freshening Period. Front Microbiol. 2021; 12:630605. https://doi.org/10.3389/fmicb.2021.630605 PMID: 33746924

41. Bickhart DM, Weimer PJ. Symposium review: Host–rumen microbe interactions may be leveraged to improve the productivity of dairy cows. Journal of Dairy Science. 2018; 101(8):7680–7689. https://doi.org/10.3168/jds.2017-13328 PMID: 29102146

42. Henderson G, Cox F, Ganesh S, Jonker A, Young W, et al. Rumen microbial community composition varies with diet and host, but a core microbiome is found across a wide geographical range. Sci Rep. 2015; 5:14567. https://doi.org/10.1038/srep14567 PMID: 26449758

43. Ma C, Sun Z, Zeng B, Huang S, Zhao J, Zhang Y, et al. Cow-to-mouse fecal transplantations suggest intestinal microbiome as one cause of mastitis. Microbiome. 2018; 6(1):200. https://doi.org/10.1186/s40168-018-0578-1 PMID: 30409169

44. Dewhirst F, Paster B, La Fontaine S, Rood J. Transfer of Kingella indologenes (Snell and Lapage 1976) to the Genus Suttonella gen. nov. as Suttonella indologenes comb. nov.; Transfer of Bacteroides nodosus (Beveridge 1941) to the Genus Dichelobacter gen. nov. as Dichelobacter nodosus comb. nov.; and Assignment of the Genera Cardiobacterium, Dichelobacter, and Suttonella to Cardiobacteriaceae fam. nov. in the Gamma Division of Proteobacteria on the Basis of 16s rRNA Sequence Comparisons. International Journal of Systematic Bacteriology. 1990; 40(4):426–433. https://doi.org/10.1099/00207713-40-4-426 PMID: 2275858

45. Reichardt N, Duncan SH, Young P, Belenguer A, McWilliam Leitch C, Scott KP, et al. Phylogenetic distribution of three pathways for propionate production within the human gut microbiota. ISME J. 2014; 8(6):1323–1335. https://doi.org/10.1038/ismej.2014.14 PMID: 24553467

46. Vacca M, Celano G, Calabrese FM, Portincasa P, Gobbetti M, De Angelis M. The Controversial Role of Human Gut Lachnospiraceae. Microorganisms. 2020; 8(4):573. https://doi.org/10.3390/microorganisms8040573 PMID: 32326636

47. Amaretti A, Gozzoli C, Simone M, Raimondi S, Righini L, Pérez-Brocal V, et al. Profiling of Protein Degraders in Cultures of Human Gut Microbiota. Front Microbiol. 2019; 10:2614. https://doi.org/10.3389/fmicb.2019.02614 PMID: 31803157

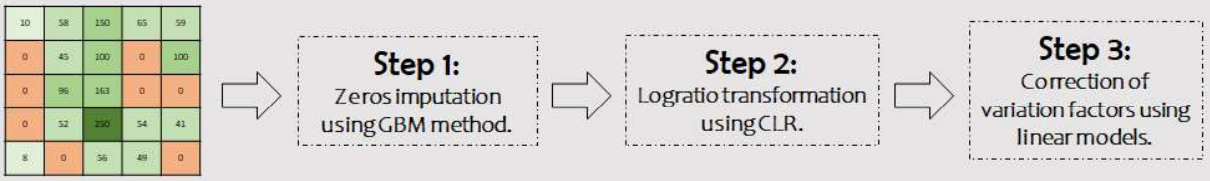
48. Bach A, López-García A, González-Recio O, Elcoso G, Fàbregas F, Chaucheyras-Durand F, et al. Changes in the rumen and colon microbiota and effects of live yeast dietary supplementation during the transition from the dry period to lactation of dairy cows. Journal of Dairy Science. 2019; 102(7):6180–6198. https://doi.org/10.3168/jds.2018-16105 PMID: 31056321

49. Indugu N, Vecchiarelli B, Baker LD, Ferguson JD, Vanamala JKP, Pitta DW. Comparison of rumen bacterial communities in dairy herds of different production. BMC Microbiol. 2017; 17:190. https://doi.org/10.1186/s12866-017-1098-z PMID: 28854878

50. Schären M, Frahm J, Kersten S, Meyer U, Hummel J, Breves G, et al. Interrelations between the rumen microbiota and production, behavioral, rumen fermentation, metabolic, and immunological attributes of dairy cows. Journal of Dairy Science. 2018; 101:4615–4637. https://doi.org/10.3168/jds.2017-13736 PMID: 29454699

Conclusion of section 4.2

The idea that microbiota data is compositional is not new, as far back as 2017 Gloor *et al.* stated that *"Microbiome dataset are compositional: and this is not optional"*. However, in this work, a theoretical framework of compositional data theory is given, and a workflow is proposed which is tested with a real dataset. The compositional data workflow consists of:



To conclude, microbiota data is compositional and methodologies must be adapted accordingly. The rumen microbiota abundances were not able to discriminate divergent genetic lines for mastitis resistance or for milk persistency. A small group of OTUs which contribute the most to the microbial community variance showed low phenotypic correlations with fine milk composition traits.

4.3 Variation in rumen bacteria of Lacaune dairy ewes from one week to the next

frontiers | Frontiers in Microbiology

ORIGINAL RESEARCH
published: 23 June 2022
doi: 10.3389/fmicb.2022.848518

Variation in Rumen Bacteria of Lacaune Dairy Ewes From One Week to the Next

Solène Fresco, Christel Marie-Etancelin, Annabelle Meynadier and Guillermo Martinez Boggio*

GenPhySE, INRAE, INPT, ENVT, Université de Toulouse, Castanet-Tolosan, France

Bacteria are the most abundant microorganisms in the rumen microbiota and play essential roles, mainly fermenting plant compounds that yield fatty acids. In this study, we aimed at assessing stability of both bacterial composition and of its associations with rumen and milk fatty acids phenotypes over a 1-week period. The study was performed using 118 Lacaune dairy ewes from the INRAE Experimental Unit of La Fage. Rumen and milk samples were obtained from the ewes twice, 1 week apart, and microbiota composition, volatile and long-chain fatty acid concentrations were analyzed. Bacterial composition was assessed using 16S rRNA gene sequencing, and microbiota and fatty acids were analyzed as compositional data. As we worked with relative abundances expressed in a constrained space, the centered log-ratio transformation enabled to transform data to work with multivariate analyses in the Euclidian space. Bacterial composition differed between the 2 weeks of sampling, characterized by different proportions of the two main phyla, *Bacteroidetes* and *Firmicutes*. The repeatability of the operational taxonomic units (OTUs) was low, although it varied significantly. However, 66 of them presented a repeatability of over 0.50 and were particularly associated with fatty acid phenotypes. Even though the OTUs from the same bacterial families presented similar correlations to fatty acids in both weeks, only a few OTUs were conserved over the 2 weeks. We proved with the help of sequencing data that there is significant change in microbial composition over a week in terms of abundance of different families of bacteria. Further studies are required to determine the impact of bacterial composition alterations over 1 week, and the specificities of the highly repeatable OTUs.

Keywords: rumen bacteria, dairy sheep, fatty acids, compositional data, stability, repeatability

OPEN ACCESS

Edited by:
Franck Carbonero, Washington State University Health Sciences Spokane, United States

Reviewed by:
Emma Hernandez-Sanabria, Flanders Institute for Biotechnology, Belgium
Brian B. Oakley, Western University of Health Sciences, United States

****Correspondence:***
Solène Fresco
solene.fresco@inrae.fr

Specialty section:
This article was submitted to Microorganisms in Vertebrate Digestive Systems, a section of the journal Frontiers in Microbiology

Received: *04 January 2022*
Accepted: *06 June 2022*
Published: *23 June 2022*

Citation:
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. Front. Microbiol. 13:848518. doi: 10.3389/fmicb.2022.848518

INTRODUCTION

Ruminant evolution resulted in adaptations to digest plant fiber efficiently through the development of multi-chamber stomachs hosting a microbiota composed of bacteria, archaea, protozoa, and fungi (Dehority, 2003). Among these organisms, bacteria are the most abundant and contribute the most to energy production, mainly volatile fatty acids (VFAs), from the fermentation of plant carbohydrates (Hungate, 1966). They are also involved in lipolysis and biohydrogenation (Hungate, 1966) and thus influence the long-chain fatty acid (LCFA) composition in the rumen,

which in turn greatly determines milk LCFA composition (Lourenço et al., 2010; Jami et al., 2014; Buitenhuis et al., 2019).

The rumen bacterial composition of adults can be influenced by biological factors, such as diet (Fernando et al., 2010; Henderson et al., 2015), parity (Pitta et al., 2014), and genetics (Sasson et al., 2017; Difford et al., 2018). Technical and computational factors may induce bias in the observed bacterial composition (Wang and LêCao, 2020). Some of the main factors include sampling techniques (Geishauser and Gitzel, 1996; Lodge-Ivey et al., 2009; Henderson et al., 2013), the bioinformatics pipeline (Schloss, 2010), and the statistical approach, including normalization method and tests applied to the data (McMurdie and Holmes, 2014). A specific methodology must be applied to microbiota data which is considered as compositional data (Gloor et al., 2017), as the only relevant information is contained in the ratios between the variables, and not in their numerical values.

Microbiota stability can be defined as the conservation of bacterial proportions at different taxonomic levels over a period of time, or high repeatability of bacterial abundance. Studies have shown that during adulthood of ruminants, bacterial composition appears to vary over long periods (3–4 months; Bainbridge et al., 2016; Zhu et al., 2021), but to be stable over shorter ones (3 days to 2 weeks; Skarlupka et al., 2019; Huang et al., 2020; Mamun et al., 2020). However, few studies focused on the short period rumen bacterial stability, and small numbers of animals were used, varying from 5 to 12 for above cited studies.

Assessing whether bacterial composition is stable over short periods of time is essential to confirm that phenotypic correlations between rumen microbiota and rumen and milk fatty acids are independent from sampling time and the influence of environmental factors (i.e., diet). The objective of this study was to answer the following two questions: Is the rumen bacterial composition of dairy ewes stable over a short period? Does it affect the correlations of bacteria with rumen and milk fatty acids?

MATERIALS AND METHODS

Animals and Experimental Design

Data were obtained from 118 Lacaune dairy ewes reared on the Experimental Unit La Fage (INRAE UE 321 agreement A312031, Roquefort, France). The genetic structure of the INRAE La Fage flock includes independent divergent genetic lines of Lacaune dairy ewes: two selected for milk SCS, based on estimated breeding values (EBVs) for milk SCS, and the other two for PERS, based on EBVs for the coefficient of variation of milk production on the testing day. Two groups of ewes with extreme EBVs were created according to the log-transformed somatic cell count (SCC): a high-SCS line (SCS+) and a low-SCS line (SCS−). And also, two extreme groups of ewes were generated, one with high persistence (PERS+) and one with low persistence (PERS-) in the milk production curve (**Table 1**). Ewes were in second ($n = 67$) or third lactation ($n = 51$), ranging from 119 to 133 days in milk, and were milked twice a day (8 a.m. and 5 p.m.) with an average daily production of 1.53 ± 0.30 kg. Ewes were fed with a total of 7 kg of a mixture composed of 72% grass silage, 21% hay, and 7% barley (on a DM basis) after morning and evening milkings. They were also supplied with 100 g of barley and 100 g of a commercial protein-rich concentrate (Brebitanne®, RAGT, Albi) in the milking parlor. Ewes had free access to water and stayed indoors with no access to grazing.

Samples and Data Collection

Samples were collected on 2 days that were 1 week apart, following the same protocol. The animal was immobilized standing in a restraint cage, and a medical gastric tube introduced into the esophagus until reaching the rumen, the introduction depth being standardized by a graduation on the tube. The vacuum pump was turned on once the gastric tube in the rumen and turned off before removing it. The sample was placed in a cold storage box for transport to the lab, where three aliquots were collected: (1) the first aliquot of 2 ml for microbiota analysis; (2) the second aliquot of 5 ml of rumen fluid was mixed with 0.2 ml of sulfuric acid (25% v/v) for VFA analysis; and (3) the last aliquot of 40 ml for LCFA analysis. All three aliquots were frozen and stored at −20°C until analysis, except for those used for microbiota analysis, which were stored at −80°C. The gastric tube was cleaned with hot water after each collection.

Rumen samples were collected before milking from each animal. For both days of sampling, a first group of 60 animals were collected from 8:30 to 11:30 a.m. and a second group of 60 animals were collected from 1:30 to 4:30 p.m. The same ewes were always sampled in the morning or the evening. To avoid dilution of samples by water, the ewes were denied access to water 2 h before sampling. Animals did not have access to feed from the previous evening (10:00 p.m.) for the group sampled in the morning, and from the early morning (7:00 a.m.) for the group sampled in the afternoon. Consequently, the number of hours of fasting varied from 6.5 to 13.5 depending on the order and moment of sampling.

For each of the 2 weeks, milk yields were recorded on the previous evening and the morning of the sampling day, and milk samples were collected and preserved with bronopol (Agrolab, Aurillac, France) for milk composition analysis or by immediate freezing for LCFA analysis. Somatic cell count and fat and protein contents analyses were performed on both

TABLE 1 | Number of animals per genetic line and genotype.

	Genetic lines[1]			
Genotypes[2]	**PERS+**	**PERS−**	**SCS+**	**SCS−**
CC	21	25	18	23
CT	9	4	10	7
TT	0	0	1	0

[1]*Genetic lines selected for high (PERS+) and low (PERS−) milk persistency or high (SCS+) and low (SCS−) somatic cell score.*
[2]*C being the wild allele and T the mutant allele of the SOCS2 gene.*

the evening and the morning milk samples. LCFA analysis was performed on the morning milk samples only.

Bacterial DNA Extraction, PCR Amplification, and Sequencing

Total DNA from 80 µl of rumen samples was extracted and purified using the QIAamp DNA Stool Mini Kit (Qiagen Ltd., West Sussex, United Kingdom) according to the manufacturer's instructions, with a bead-beating step in a FastPrep instrument (MP Biomedicals, Illkirch, France). The 16S rRNA V3–V4 regions gene of the extracted DNA strands were amplified (first PCR: 30 cycles) with the primers F343 (5′-CTTT CCCTACACGACGCTCTTCCGATCTACGGRAGGCAGCAG-3′; Liu et al., 2007) and R784 (5′-GGAGTTCAGACGTGTG CTCTTCCGATC TTACCAGGGTATCTAATCCT-3; Andersson et al., 2008). As Illumina MiSeq technology enables 250 bp reads, the ends of each read were overlapped and stitched together to generate full-length reads of the entire V3 and V4 regions in a single run. Single multiplexing was performed using a 6 bp index, which was added to R784 during a second round of PCR with 12 cycles using the forward primer (AATGATACGGCGACCACCGAGATCTACACTCT TTCCCT ACACGAC) and reverse primer (CAAGCAGAAGACGGCA TACGAGATGTGACT GGAGTTCAGACGTGT). The PCR products were purified and loaded onto an Illumina MiSeq cartridge (Illumina, San Diego, CA, United States) at the Genomic and Transcriptomic Platform (INRAE, Toulouse, France) according to the manufacturer's instructions.

Sequence reads were demultiplexed, and each paired-end read was assigned to its sample based on the previously integrated index, and processed with the FROGS 3.0 pipeline (Escudié et al., 2018). The procedure consisted of the following steps: (1) read pre-processing, removing sequences with primer mismatch, displaying unexpected length (<300 or >500 bp), or with ambiguous bases; (2) sequence clustering with denoising and one sequence difference between each of the three aggregation steps of clustering; (3) chimera removal; (4) cluster filtering with Bokulich filter (removing clusters with abundances <0.005%; Bokulich et al., 2013); and (5) taxonomy assignment to operational taxonomic units (OTUs) using the SILVA 138.16S pintail 100 database. From this process, an abundance table containing the number of sequences per OTU and rumen sample was obtained.

Rumen Fatty Acids Composition Analyses

Two gas chromatography analyses were performed at the National Veterinary School of Toulouse (Toulouse, France) with rumen samples: one for VFAs and the other for LCFAs.

Six VFAs, acetic acid (C2:0), propionic acid (C3:0), butyric acid (C4:0), valeric acid (C5:0), isobutyric acid (*iso*-C4:0), and isovaleric acid (*iso*-C5:0), were quantified using automated gas separation, according to the method of Playne (1985) and modified as follows. The rumen samples were first centrifuged at 2,880 × *g* for 20 min to separate the liquid phase. For protein removal, 1 ml of supernatant was mixed with 200 µl of (25% v/v) metaphosphoric acid and further centrifuged at 20,000 × *g* for 15 min. Then, 100 µl of the supernatant was added to 75 µl (0.2% v/v) of 4-methylvaleric acid as an internal standard and 900 µl of ultrapure water. From this mixture, 1 µl was then injected into a gas chromatograph (Hewlett Packard, Model 7890A) equipped with a DB-FFAP column (30 m × 0.53 mm i.d., 1-µm film thickness, Agilent Technologies, Palo Alto, CA, United States) and an FID detector (Avondale, PA, United States). Chromatograms were integrated using Chromeleon software (Thermo Fisher Scientific, version 6.8, Waltham, MA, United States). The sum of the six VFA concentrations was defined as the total concentration and was used to obtain the molar proportions of each VFA.

The LCFAs of rumen content were extracted and methylated *in situ* using the procedure described by Park and Goins (1994), except that the solution of 14% boron trifluoride in methanol was replaced by a solution of methanol–acetylchloride (10:1). Nonadecanoic acid (C19:0) was used as the internal standard at a dose of 0.8 mg. The fatty acid methyl esters (FAMEs) were then quantified by gas chromatography (Agilent 6890N, Network GC System, equipped with a model 7,683 auto injector, Agilent Technologies, Palo Alto, CA, United States) using a fused silica capillary column (100 m × 0.25 mm i.d., 0.20 µm film thickness, CPSil 88, Varian, Middelburg, the Netherlands) as described by Zened et al. (2011). Peaks were identified and quantified by comparison with commercial standards (Sigma Co., St Louis, MO, United States), except for C18:1, C18:1 *trans*-9, C18:1 *trans*-11, and C18:1 *cis*-9, which were identified by the order of elution. Chromatograms were integrated using the Peak Simple software (Peak Simple Data System, version 2.83, SRI, Torrance, CA, United States). Results are expressed as the percentage of total FAME. The 29 measured fatty acids (FAs) were: C12:0, C13:0, *anteiso*-C13:0, *iso*-C13:0, C14:0, *iso*-C14:0, C15:0, *anteiso*-C15:0, *iso*-C15:0, C16:0, C17:0, *anteiso*-C17:0, *iso*-C17:0, C18:0, C18:1 *cis-9*, a mix of C18:1 *cis-11* and C18:1 *trans-15*, C18:1 *cis-12*, C18:1 *cis-15*, a mix of C18:1 *trans-6*, *trans-7*, *trans-8*, C18:1 *trans-9*, C18:1 *trans-10*, C18:1 *trans-11*, C18:1 *trans-12*, C18:1 *trans-16*, C18:2, C18:2 *cis-9*, *trans-11*, C18:2 *trans-11*, *cis-15*, C18:3, and C20:1.

Milk Composition Analyses

The four milk samples from each ewe were analyzed at Agrolab (Aurillac, France). Fat and protein contents were obtained using mid-infrared spectrometry (Milk-Scan™ FT 6000 instrument, Foss, Nanterre, France) and somatic cell count was quantified using a Fossomatic cell counter (Foss, Nanterre, France), to which a log-transformation was applied to obtain the SCS (Ali and Shook, 1980). The data were translated into daily variables. Daily milk yield was obtained by summing the morning and evening milk yields. Daily protein content, fat content, and SCS were computed as the average of the morning and evening values weighted by the corresponding milk yields.

Long-chain fatty acid percentages in total FAME were measured in the morning milk for the 2 weeks by gas chromatography at the National Veterinary School of Toulouse (Toulouse, France), following the same method as for the rumen samples. The 38 measured FAs included the 29 measured in

TABLE 2 | Models for OTUs, VFAs, long-chain fatty acids (LCFAs) in the rumen, milk yield, fat and protein contents, somatic cell score, and LCFAs in the milk.

Variables	Line	SOCS2[1]	Parity	Line *x* Parity[2]	Week	Line *x* Week[2]	Time after feeding	Number of sequences	Number of ewes
Rumen									
OTUs	■		■	■	■		■	■	111
VFAs	◆				◆		◆		116
LCFAs	■	■			■	■			116
Milk									
LCFAs	◆	◆			◆	◆			117
Milk yield	◆	◆			◆				118
Fat content		◆	◆		◆				118
Protein content	◆				◆				118
Somatic cell score	◆	◆	◆						118

[1]*Genotype for the SOCS2 gene.*

[2]*x Represents the interaction between effects.*

■ *Fixed effects selected while significant for at least 10% of the variable.*

◆ *Fixed effects selected when significant for each variable.*

the rumen in addition to C4:0, C6:0, C7:0, C8:0, C9:0, C10:0, C11:0, C14:1, and C16:1.

Statistical Analysis

All analyses were performed using R software (R Core Team, 2021)[1]. Statistical significance was set at $p < 0.05$.

Microbiota and fatty acids data are compositional data (Gloor et al., 2017), meaning that the information is contained in the ratios between variables, not in the values themselves, due to the restrictions imposed by either the sequencing technology or the measurement unit (percentages). Thus, we applied the compositional data approach proposed by Aitchison (1986) for the composition of OTUs and FAs. It consisted of imputing zero values in the data set to transform the counts into log-ratios using the centered log-ratio (CLR). Then, the multivariate analysis described in a previous study by Martinez Boggio et al. (2021) was applied. Two approaches were applied to impute the zero values, considering their nature. In the bacterial abundance table, as zeros referred to a probability of count, geometric Bayesian multiplicative replacement was applied (cmultRepl function from the zCompositions package); in the LCFA datasets, the zero values reflected the detection limit of the chromatograph; thus, the expectation–maximization procedure was applied (lrEM function from the zCompositions package). Then, the CLR transformation was applied to OTUs, VFA, and LCFA data (clr function from the compositions package).

Table 2 presents the models used to correct the data for fixed effects and repeated measures (random animal effect) and the number of animals considered. Ewes were removed from OTUs models when one of their samples had a low sequencing quality (<500 OTUs). Ewes missing one of the two samples were removed from rumen and milk LCFA models. Ewes presenting values of VFAs under the detection limit of the gas chromatograph were removed. The fixed effects included in the models were: genetic lines (four levels: PERS−, PERS+, SCS-, and SCS+), week of sampling (two levels: "week 1" and "week 2"), parity (two levels), SOCS2 genotype (two levels: "TT/CT" or "CC"), time after feeding (six levels of equal size), and number of sequences per sample (four levels of equal size).

The significance of the fixed effects was assessed using the ANOVA function from the sasLM package. For milk yield, fat and protein contents, SCS, and the six VFAs, the fixed effects were included in the model when they were significant. For OTUs, rumen, and milk LCFAs, fixed effects were retained in the models when they were significant for at least 10% of the variables. Linear mixed models were defined for each trait (one OTU being one trait), using the lmer function from the lmerTest package, and the corresponding variances of random animal and residual effects were obtained (VarCorr function from the lme4 package), allowing to compute the repeatability, defined as the animal variance divided by the total variance. Then, Spearman correlations were performed between the OTU repeatability values and their percentage of zeros or average abundance.

To perform multivariate analyses assessing the effect of the week, residuals of OTUs and FAs obtained from ANOVA were used without correcting for the week effect (**Table 2**). Sparse partial least square discriminant analyses (sPLS-DA) and sparse partial least square analyses (sPLS) were performed, to assess the influence of the week on bacterial composition and identify relationships between the OTUs and the FAs, respectively (spls and splsda functions from the mixOmics package). The sparse procedure allowed to reduce the number of components, i.e., the dimensionality of the analyses, and select only the most relevant variables, i.e., the OTUs, for each component. The number of components was previously chosen to explain 90% of the variance, using a principal component analysis. The number of variables retained by component was determined using the CLR lasso penalty method considering the penalization obtained with a 10-fold cross-validation (cv.glmnet and glmnet functions from the glmnet package).

The predictive ability of the sPLS-DA model was evaluated through the overall misclassification error rate after a 5-fold cross-validation repeated 10 times (perf function). For each

[1]https://www.r-project.org/

sPLS, the 20 OTUs presenting the highest association with the FAs were selected and Pearson correlations were computed. Only the FAs common in both weeks and having at least one significant correlation with the 20 selected OTUs were presented. Fisher exact tests were performed to estimate over- or under-representation of specific phyla or highly repeatable OTUs among those selected for sPLS-DA and sPLS compared to all OTUs.

RESULTS

Rumen Bacterial Composition

A total of 2,500,763 sequences were obtained, with an average of 10,736 ± 3,745 reads per sample, grouped into 2,079 OTUs, with an average of 1,406 ± 211 OTUs per sample. Ten phyla were identified; the three major phyla, based on the percentage of sequences assigned to each phylum out of the total number of DNA sequences, were *Bacteroidetes* (53.5%), *Firmicutes* (35.4%), and *Proteobacteria* (6.2%). The other phyla were *Fibrobacterota* (2.7%), *Spirochaetota* (1.2%), *Actinobacteria* (0.5%), *Patescibacteria* (0.3%), and *Desulfabacterota*, *Elusmicrobiota*, and *Campylobacterota*, each lower than 0.2%. We found that 81.2% of the DNA sequences enabled genus-level classification, resulting in the identification of 117 genera. The most important ones were *Prevotella* (30.78%), *Lachnospiraceae_NK3A20_group* (6.3%), *Ruminococcus* (5.3%), *Christensenellaceae_R_7_group* (4.7%), *Sphingomonas* (4.5%), and *Rikenellaceae_RC9_gut_group* (3.9%). The lowest taxonomic rank considered in this study was genus, as 92.7% of the reads were assigned to unknown species or were multi-affiliated species as expected with short 16S reads.

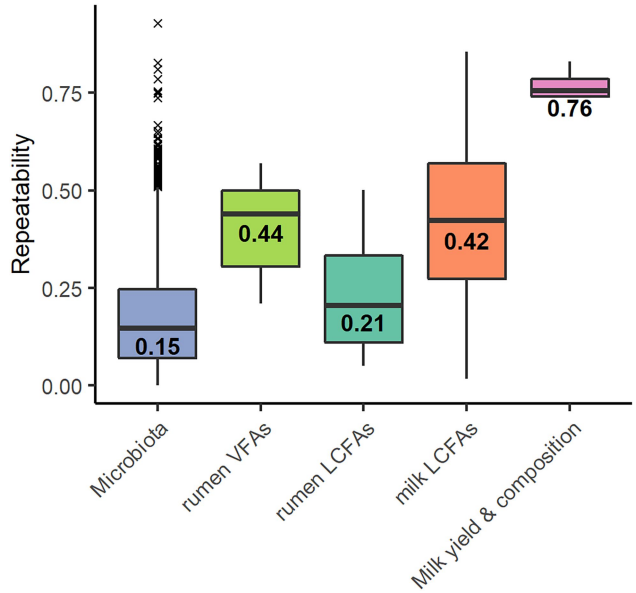


FIGURE 1 | Repeatability values of the 1,599 operational taxonomic units (OTUs)* of the microbiota, the six volatile fatty acids (rumen VFAs), the 27 rumen long-chain fatty acids* (rumen LCFAs), the 37 milk long-chain fatty acids* (milk LCFAs) and the traits milk yield, and fat content and protein content and somatic cell score (Milk yield & composition). Values indicate the median. *Repeatability values were estimated for the OTU and trait models that converged.

Fixed Effects and Repeatability

Operational taxonomic unit data were corrected by six effects that were significant for at least 10% of the OTUs: genetic lines (25.9%), week of sampling (16.2%), time after feeding (13.9%), number of sequences per sample (12.8%), parity (11.4%), and the interaction between the genetic line and parity (11.9%), with percentages in brackets indicating the percentage of OTUs for which the given factor was significant. The explained variance was estimated for all OTUs, resulting in an average of 0.12 ± 0.04, ranging from 0.03 to 0.50 according to OTUs.

Out of the 2,079 OTUs, 1,665 showed an estimated repeatability ranging from 0 to 0.93 with a median of 0.15 ± 0.14 (**Figure 1**). The remaining 414 OTUs had no estimate; linear models did not converge because of the small number of samples. Of the 1,665 OTUs, 66 presented a repeatability of over 0.50, corresponding to a frequency of 4%. Spearman correlations were performed between the repeatability values of the OTUs and their percentage of zeros ($R = -0.01$, $p = 0.50$) and between the repeatability of the OTUs and their average abundance ($R = 0.24$, $p < 0.001$).

The repeatabilities ranged from 0.21 to 0.57 for VFAs, from 0.05 to 0.50 for rumen LCFAs, and from 0.02 to 0.86 for milk LCFAs (**Figure 1**). The repeatability was 0.83 for milk yield, 0.74 for fat content, 0.77 for protein content, and 0.74 for SCS. The models did not converge for the LCFAs C18:1 *trans*-12 and C18:2 *cis*-9, *trans*-11 in the rumen and C20:1 in the milk.

Bacterial Composition Over One Week

The model of the sPLS-DA on bacterial composition over a week included 160 components and 57 variables per component. Based on OTU residual abundances, rumen samples were discriminated by the week of collection (**Figure 2**). The overall error rate of the model was 0.21. In the first component (**Table 3**), week 2 was characterized by a lower number of OTUs belonging to the phylum *Firmicutes* ($p < 0.001$) and a higher number of OTUs belonging to the phylum *Bacteroidetes* ($p < 0.01$) than when considering all OTUs. In the second component (**Table 3**), week 1 had a higher presence of OTUs from the phyla *Firmicutes* ($p < 0.01$) and a lower presence of OTUs from the phyla *Bacteroidetes* ($p = 0.01$) than when considering all OTUs.

In the first component, 49 OTUs among the 57 selected to discriminate week 1 and week 2 had a repeatability of less than 0.50 (**Supplementary Table S1**), three had a repeatability of over 0.50, and the remaining five OTUs did not have an estimated repeatability. In the second component (**Supplementary Table S2**), 44 OTUs of 57 had a repeatability of less than 0.50, one had a repeatability of over 0.50, and the remaining 12 OTUs had no estimated repeatability. The proportion of OTUs presenting repeatability values of less than 0.50 among the 57 selected for each component of the sPLS-DA was not

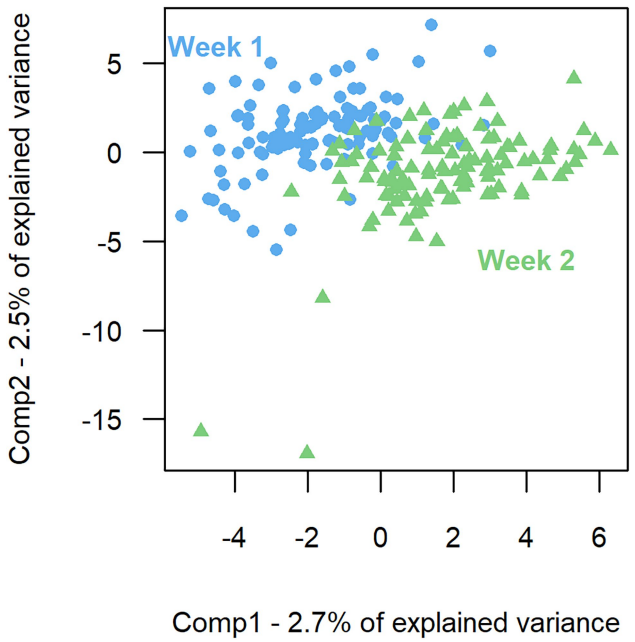


FIGURE 2 | Sparse partial least square discriminant analysis of the rumen bacterial composition between the 2 weeks of sampling. First component (Comp1) plotted against second component (Comp2) presenting the largest explained variance between weeks.

TABLE 3 | Number of OTUs belonging to the phyla *Firmicutes* and *Bacteroidetes* considering the 2,079 OTUs or the OTUs selected on the two main components for the sPLS-DA associated with each of the 2 weeks.

	Total number of OTUs	Number of OTUs from the phyla *Firmicutes*	Number of OTUs from the phyla *Bacteroidetes*
All OTUs	2,079	880	1,026
Component 1			
week 1	25	12	8
week 2	32	4	24
Component 2			
week 1	32	6	23
week 2	25	8	15

significantly different from that considering all OTUs ($p = 0.46$ and $p = 0.47$ for components 1 and 2, respectively).

Correlations Between Rumen Bacteria and Fatty Acid Phenotypes

Correlations Between Rumen Bacteria and VFAs

In both weeks 1 and 2, sPLS models for VFAs included 77 components, with 152 variables retained per component in week 1 and 187 in week 2. For the two main components, the explained variance was 7 and 9% for weeks 1 and 2, respectively. The significant Pearson correlations between the 20 most associated OTUs and *iso*-C4:0 and *iso*-C5:0 ranged from −0.49 to −0.19 and from 0.21 to 0.46 (**Figure 3**). Three OTUs were conserved between week 1 and week 2, and the corresponding correlations were of the same sign. For *Lachnospiraceae* and *Rikenellaceae* families, the sign of the correlation with a given VFA was the same for all the OTUs of the family on both weeks, while bacteria from *Prevotellaceae* family had both positive and negative correlations with the VFAs. Highly repeatable OTUs (> 0.50) were particularly selected by the sPLS on both weeks ($p = 0.04$ and $p < 0.001$ for weeks 1 and 2, respectively).

Correlations Between Rumen Bacteria and Rumen LCFAs

For rumen LCFAs, sPLS models in both week 1 and week 2 included 77 components, with 19 variables retained per component on week 1 and 15 on week 2. For the two main components, the explained variance was 7 and 8% for weeks 1 and 2, respectively. The significant Pearson correlations ranged from −0.51 to −0.19 and from 0.19 to 0.48 for the 20 OTUs most associated with the five rumen LCFAs (**Figure 4**). Three OTUs appeared on both weeks, and the sign of their correlation was conserved. For *Lachnospiraceae* and *Rikenellaceae* families, the sign of the correlation with a given rumen LCFA was the same for all the OTUs of the family on both weeks, while bacteria from *Prevotellaceae* family had both positive and negative correlations with the rumen LCFAs. Highly repeatable OTUs (>0.50) were particularly selected by the sPLS on both weeks ($p < 0.001$).

Correlations Between Rumen Bacteria and Milk LCFAs

In both weeks 1 and 2, sPLS models for milk LCFAs included 77 components, with 41 variables retained per component in week 1 and 54 in week 2. For the two main components, the explained variance was 6 and 9% for weeks 1 and 2, respectively. The significant Pearson correlations between the 20 most associated OTUs and C18:1 *trans*-11 and C18:2 *trans*-11, *cis*-15 ranged from −0.39 to −0.22 and from 0.19 to 0.45 (**Figure 5**). Four OTUs were conserved between week 1 and week 2, two of which were also conserved in the VFA analysis (OTU_408 and OTU_459), and the corresponding correlations were of the same sign. For *Rikenellaceae* and *Ruminococcaceae* families, the sign of the correlation with a given milk LCFA was the same for all the OTUs of the family on both weeks, while bacteria from *Prevotellaceae* family had both positive and negative correlations with the milk LCFAs. Highly repeatable OTUs (>0.50) were particularly selected by the sPLS on both weeks ($p = 0.04$ and $p < 0.001$ for weeks 1 and 2, respectively).

DISCUSSION

Biological, Technical, and Computational Effects

Bacteroidetes, *Firmicutes*, and *Proteobacteria* were the three main phyla observed in the rumen of adult dairy ewes fed a mixed diet of forages and concentrates, in accordance with that observed in sheep (Belanche et al., 2019; Liu et al., 2020) and cows (Bainbridge et al., 2016; Plaizier et al., 2017). However, the

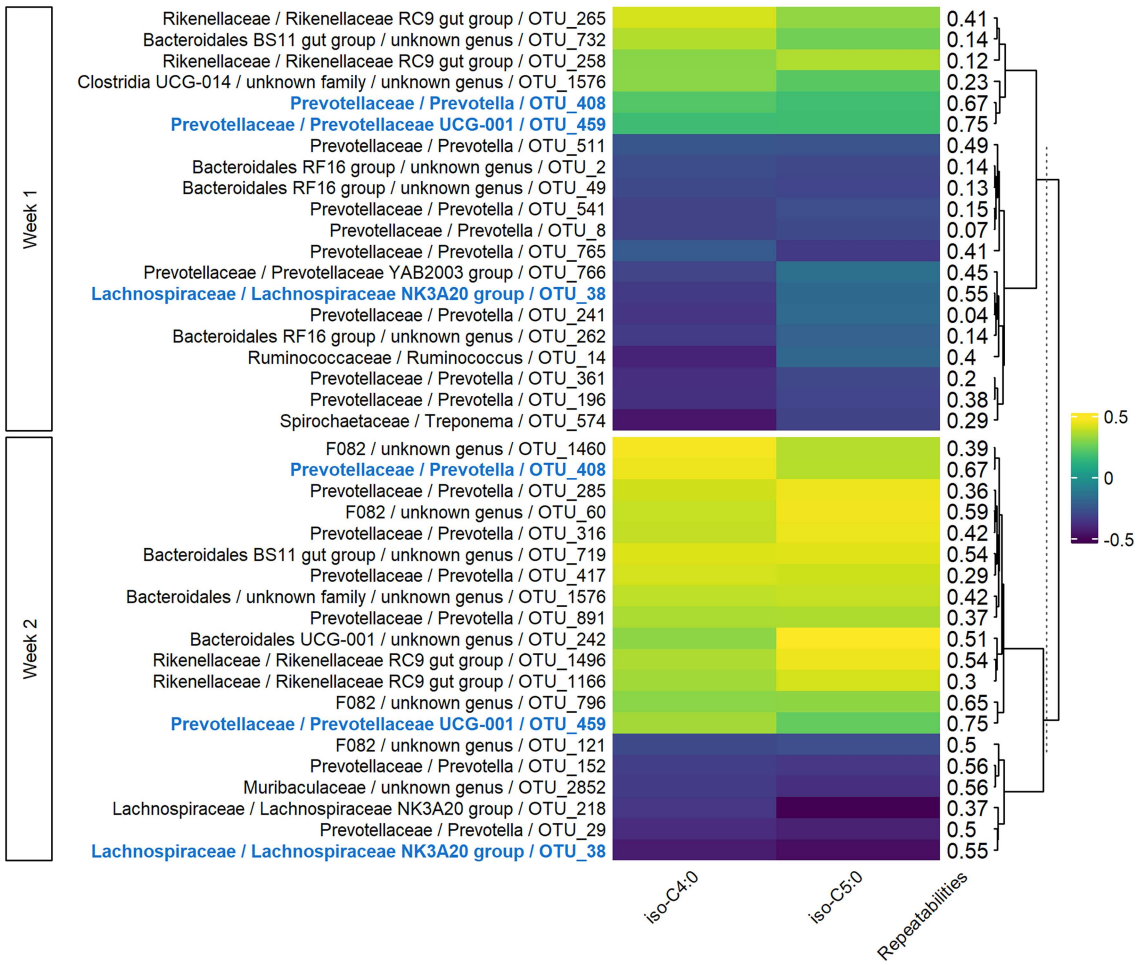


FIGURE 3 | Pearson correlations between bacterial OTUs and VFAs on both weeks of sampling. Only the 20 more correlated OTUs and VFAs having at least one significant correlation in both weeks with them were presented. In bold and blue, OTUs conserved from week 1 to week 2.

rumen bacterial composition is affected by various factors, mainly biological ones. The most known is the diet composition, often represented by the forage/concentrate ratio that can influence family and genus abundances (Henderson et al., 2015). Another one is the feeding time, influencing diurnal variations in bacterial concentrations in sheep fed once daily (Warner, 1966), with the abundance of families belonging to the phylum Firmicutes increasing with time after feeding (de Assis Lage et al., 2020). In addition, both parity and lactation stage affect bacterial composition by modifying the abundance of phyla and genera (Pitta et al., 2014; Bainbridge et al., 2016; Xue et al., 2018). Although many authors have found some bacterial taxa to be common in all the animals in their study (Jami and Mizrahi, 2012; Xue et al., 2018; Huang et al., 2021), interindividual variations in bacterial composition were observed, and evidence of genetic determinism of abundance of some bacteria has been raised by Sasson et al. (2017) and Difford et al. (2018).

Technical and bioinformatics processing factors are also known to greatly influence the observed bacterial abundance (Pollock et al., 2018; Wang and LêCao, 2020). Diversity and relative abundances vary depending on the type of sampling (de Assis Lage et al., 2020) and DNA extraction methods (Henderson et al., 2013; Gerasimidis et al., 2016). Considering sequencing strategies, the choice of primers influences the estimation of bacterial abundances (Tremblay et al., 2015; Fouhy, 2016) while the sequencing platform influences the read length and error rate (D'Amore et al., 2016; Kchouk et al., 2017). The number of observed OTUs depends on the chosen pipeline, in particular, the clustering and filtering steps (Schloss, 2010). As the sequencing depth is highly variable among the samples, at random or depending on the sequencing run, normalization is usually applied to the data (McMurdie and Holmes, 2014; Weiss et al., 2017).

As the current objective was to determine if the bacterial composition was stable over time, the week was the factor of greatest interest in this study. To allow proper observation, breed, diet, and lactation stages were experimentally controlled. Moreover, the same DNA extraction method was used for the samples of both weeks, only one sequencing run was performed, and the bioinformatic analysis included all samples. The other

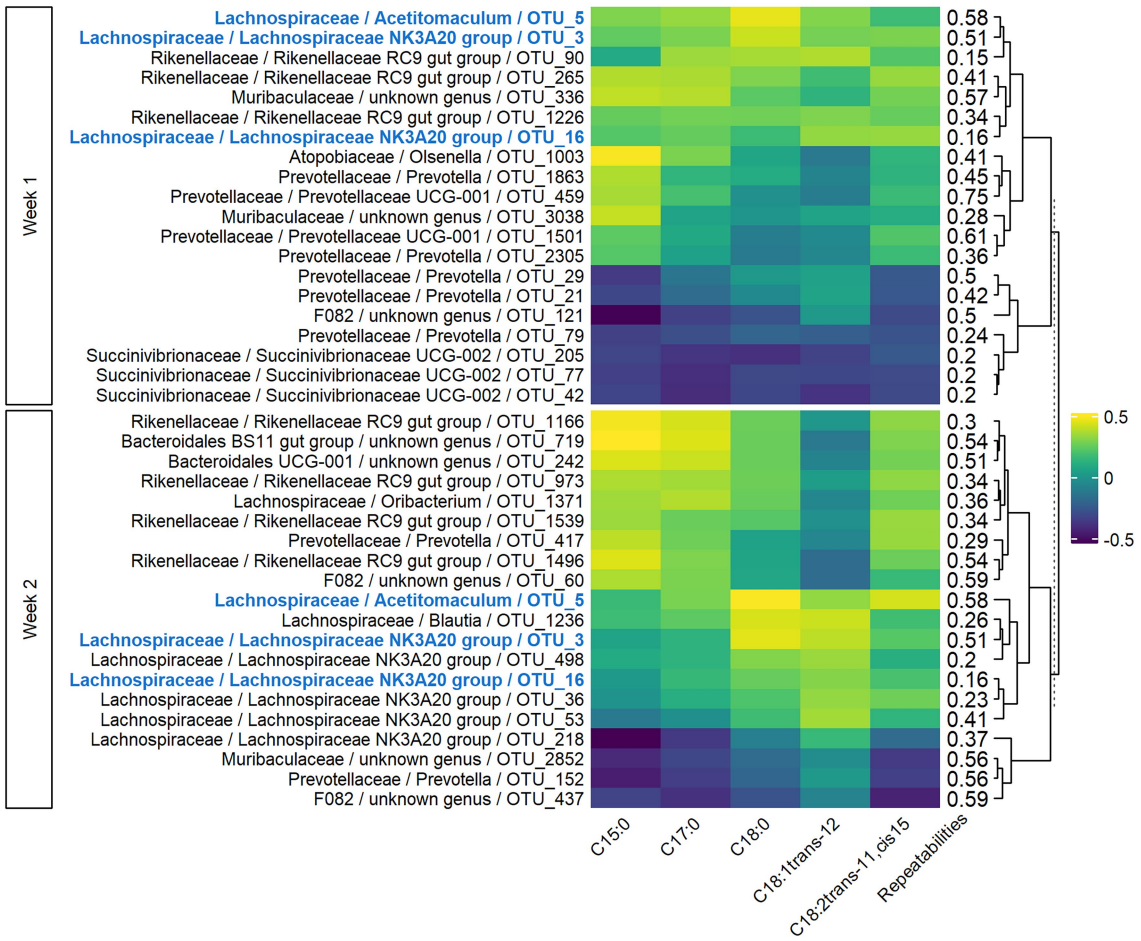


FIGURE 4 | Pearson correlations between bacterial OTUs and rumen long-chain fatty acids on both weeks of sampling. Only the 20 more correlated OTUs and rumen LCFAs having at least one significant correlation in both weeks with them were presented. In bold and blue, OTUs conserved from week 1 to week 2.

effects, including genetic lines, SOCS2 genotype, parity, time after feeding, and sequencing depth were corrected through linear models following the methodology used by Martinez Boggio et al. (2021).

Instability of Bacterial Composition

Assessing the stability of the bacterial composition over a short period of time will support minimizing repeated sampling, as they are invasive for the animals and time-consuming. It is essential to know if the results obtained at one sampling are reliable, whether the conclusions are the same 1 week later. In this study, two criteria were used to assess the stability of the bacterial composition. In the first criterion, bacterial composition was defined as stable when phyla, genera, and OTU proportions were conserved from 1 week to the next for the same animal. In the literature, various methods have been used for assessing bacterial composition stability, such as two one-sided tests (Skarlupka et al., 2019), principal coordinate analysis, and analysis of similarity (Huang et al., 2020; Mamun et al., 2020). In the current study, considering the compositional nature of the data, we used multivariate analyses (working in the Euclidian space), such as sPLS-DA and sPLS. In the second criterion, at the OTU level, the repeatability of the OTUs was used to determine their stability, ranging from 0 (completely unstable) to 1 (completely stable), with OTUs having a repeatability greater than 0.50 being defined as stable. Another method, the Lin's concordance correlation coefficient, which is also conceptually similar, was used by Zhu et al. (2021) to estimate OTU stability.

Although the week effect was the second largest in the ANOVA, it was significant for only 16% of OTUs. A clustering of the samples by week was observed in the sPLS-DA, mainly characterized by different proportions of *Bacteroidetes* and *Firmicutes*, but less than 6% of the variance was explained. Even if the results of these two analyses suggested bacterial composition instability, the high unexplained variance made them inconclusive. As explained previously, various factors influencing bacterial composition cannot be controlled or corrected for. In our study, such factors impacted the variance

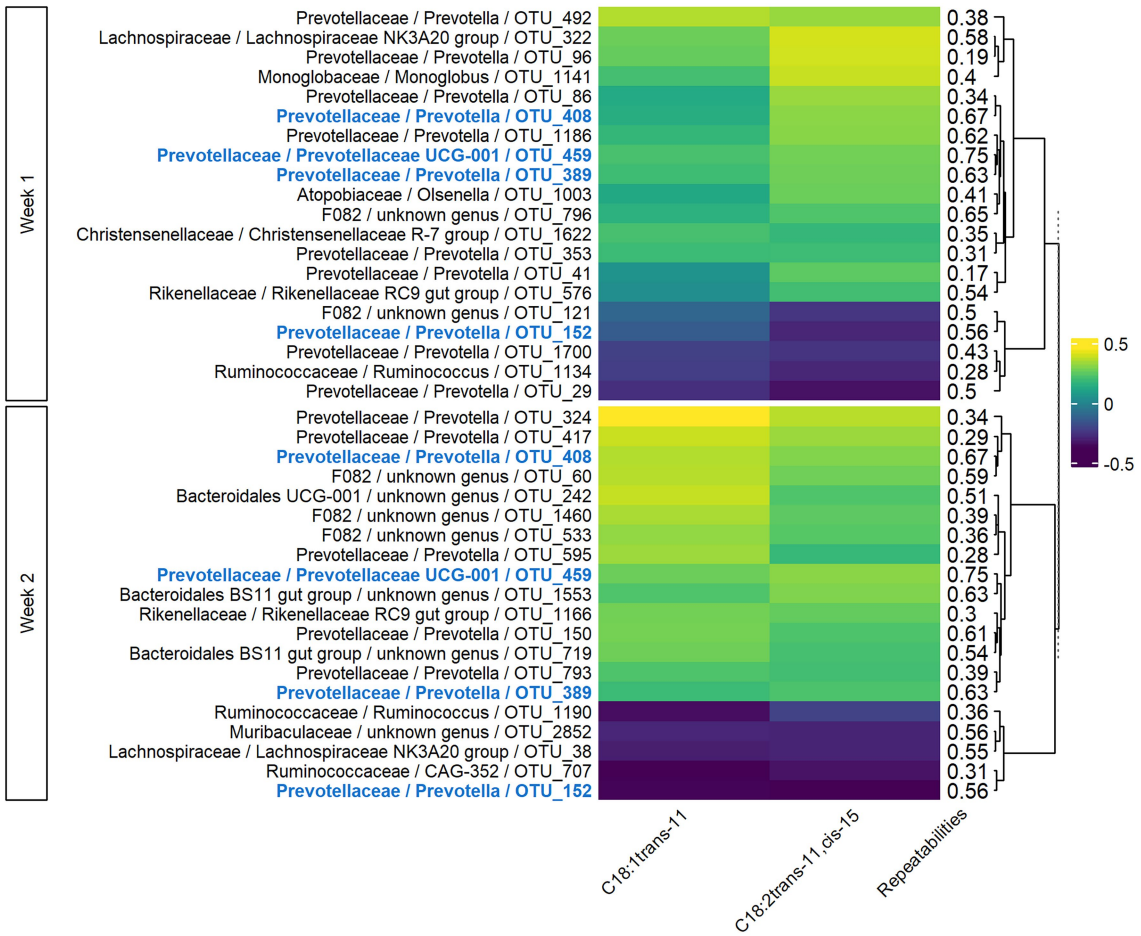


FIGURE 5 | Pearson correlations between bacterial OTUs and milk long-chain fatty acids on both weeks of sampling. Only the 20 more correlated OTUs and the milk LCFAs having at least one significant correlation in both weeks with them were presented. In bold and blue, OTUs conserved from week 1 to week 2.

explained in the ANOVA and the sPLS-DA analyses. They could be unrecorded technical effects such as (1) contamination from the oral cavity or saliva (Lodge-Ivey et al., 2009; Terré et al., 2012), (2) multiple relocation of the gastric tube in the rumen (Geishauser and Gitzel, 1996), and (3) variable proportions of solid and liquid phases (Henderson et al., 2013; Vaidya et al., 2018); or biological effects such as the stress induced by handling before and during sampling (Yoshihara and Ogawa, 2021).

However, the repeatability of 96% of the OTUs was under 0.5, and it was this large proportion of non-stable OTUs (median repeatability of 0.15) that allowed us to state bacterial composition instability. Similar results were obtained by Jewell et al. (2015); Bainbridge et al. (2016); and Zhu et al. (2021) when working over long periods of 75–122 days, which was expected because of the numerous effects that could alter bacterial composition. The 96% of non-stable OTUs (96%) obtained is consistent with the 97% found by Zhu et al. (2021). Contrasting with our results, some authors obtained stability over short periods, from 3 days to 1 week using analyses comparable to ANOVA and sPLS-DA (Skarlupka et al., 2019; Huang et al., 2020; Mamun et al., 2020). However, they did not consider the compositional nature of the data neither computed the OTU repeatability that demonstrated that most of them were moderately repeatable (<0.50). Moreover, the number of animals used in this study was larger than in the previous ones, allowing for revealing small differences between weeks covered up by uncontrolled alterations of the bacterial composition. Those three specificities of our study allowed to clearly reveal bacterial composition instability over short periods.

Instability of Correlations With Fatty Acid Phenotypes

The biological links between OTUs and FA phenotypes were compared between week 1 and week 2 to discuss functional ruminal bacterial stability over time. At the OTU level, only three or four OTUs were conserved from week 1 to week 2, with similar correlations to FA phenotypes (**Figures 3–5**).

However, we found that OTUs belonging to a same family presented all correlations of the same sign for a given FA. This phenomenon was observed for some of the most predominant families in the rumen, namely *Lachnospiraceae*, *Rikenellaceae*, and *Ruminococcaceae*, and may suggest functional redundancy (Wohl et al., 2004; Weimer, 2015; Louca et al., 2018). That is to say, different OTUs belonging to the same family may have the same function, allowing them to maintain the organism function despite a variable microbial composition. In addition, OTUs from *Prevotellaceae* family presented both positive and negative correlations with the FAs. The main hypothesis to explain this observation is the large genetic and functional diversity of this family (Stewart et al., 1997; Matsui et al., 2000). Further investigation of these aspects is not possible with 16S rRNA gene sequencing, because as stated by Plummer and Twin (2015), it does not allow access to the species and strain classification necessary to investigate the function of the bacteria identified.

The OTUs generally presented low repeatability, with only 4% of them defined as stable. It is notable that stable OTUs were overrepresented among the OTUs highly correlated with FA phenotypes in both weeks, ranging from 3 to 11 among 20. If the stable OTUs are considered as those having high genetic determinism, it was not surprising to find them highly associated with FA phenotypes, for which a significant part of the variability is related to animal genetics. The correlations between bacteria and FA phenotypes from week 1 to week 2 were not conserved, but OTUs related to FA phenotypes appeared to be the most repeatable.

Week-to-week variation in the rumen bacterial composition of Lacaune dairy ewes was observed, with only 4% of the OTUs being stable, which seemed to alter the correlations between microbiota and FAs phenotypes, as only a few OTUs were conserved between 1 week and the next. Even if FA phenotypes were not linked to the same OTUs from 1 week to the next, it is noticeable that they were particularly associated with stable OTUs, possibly due to common genetic determinism of those traits and OTUs. In conclusion, we proved with a large dataset of 188 dairy ewes that bacterial composition and its phenotypic correlations with fatty acids are not transposable from 1 week to the next. Further work is necessary to confirm those results, as very few studies were performed on microbiota stability over short periods of time, and none using compositional data approach. We hope that our study will be a first step in identifying microbiota members with high repeatability and thus potential heritability, to then consider the possibility of genetic selection.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://www.ncbi.nlm.nih.gov/, PRJNA765197.

ETHICS STATEMENT

The animal study was reviewed and approved by the appropriate Ethical Committee (APAFIS#6292-2016080214271984 v8). Written informed consent was obtained from the owners for the participation of their animals in this study.

AUTHOR CONTRIBUTIONS

CM-E and AM conceived the experiments and participated in sample collection and processing. GMB performed data processing. SF performed statistical analyses under the supervision of GMB and CM-E. SF, CM-E, GMB, and AM participated in the interpretation of the results. SF wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

The experiment was funded by Génétique Animale and Physiologie Animale et Systèmes d'Elevage divisions of INRAE.

ACKNOWLEDGMENTS

The authors are grateful to Hélène Larroque and Rachel Rupp for access to the genetic resources produced in the La Fage experimental unit. The authors would like to thank the technical staff of the INRAE UE La Fage (doi: 10.15454/1.548325234664425E12) for animal care and rumen sampling, Béatrice Gabinaud for preparing microbiota samples, Get-Plage platform in Toulouse for sequencing, and Yves Farizon for the analyses of the fatty acid composition of rumen fluid and milk samples.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.848518/full#supplementary-material

REFERENCES

Aitchison, J. A. (1986). *The Statistical Analysis of Compositional Data*. New York, USA: Chapman & Hall Ltd

Ali, A. K. A., and Shook, G. E. (1980). An optimum transformation for somatic cell concentration in milk. *J. Dairy Sci.* 63, 487–490. doi: 10.3168/jds.S0022-0302(80)82959-6

Andersson, A. F., Lindberg, M., Jakobsson, H., Bäckhed, F., Nyrén, P., and Engstrand, L. (2008). Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PLoS One* 3:e2836. doi: 10.1371/journal.pone.0002836

Bainbridge, M. L., Cersosimo, L. M., Wright, A.-D. G., and Kraft, J. (2016). Rumen bacterial communities shift across a lactation in Holstein, Jersey and Holstein × Jersey dairy cows and correlate to rumen function, bacterial fatty acid composition and production parameters. *FEMS Microbiol. Ecol.* 92:fiw059. doi: 10.1093/femsec/fiw059

Belanche, A., Kingston-Smith, A. H., Griffith, G. W., and Newbold, C. J. (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. *Front. Microbiol.* 10:122. doi: 10.3389/fmicb.2019.00122

Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, J. I., Knight, R., et al. (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat. Methods* 10, 57–59. doi: 10.1038/nmeth.2276

Buitenhuis, B., Lassen, J., Noel, S. J., Plichta, D. R., Sørensen, P., Difford, G. F., et al. (2019). Impact of the rumen microbiome on milk fatty acid composition of Holstein cattle. *Genet. Sel. Evol.* 51:23. doi: 10.1186/s12711-019-0464-8

D'Amore, R., Ijaz, U. Z., Schirmer, M., Kenny, J. G., Gregory, R., Darby, A. C., et al. (2016). A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC Genomics* 17:55. doi: 10.1186/s12864-015-2194-9

de Assis Lage, C. F., Räisänen, S. E., Melgar, A., Nedelkov, K., Chen, X., Oh, J., et al. (2020). Comparison of two sampling techniques for evaluating ruminal fermentation and microbiota in the planktonic phase of rumen digesta in dairy cows. *Front. Microbiol.* 11:618032. doi: 10.3389/fmicb.2020.618032

Dehority, B. A. (2003). *Rumen Microbiology*. Nottingham: Nottingham University Press

Difford, G. F., Plichta, D. R., Løvendahl, P., Lassen, J., Noel, S. J., Højberg, O., et al. (2018). Host genetics and the rumen microbiome jointly associate with methane emissions in dairy cows. *PLoS Genet.* 14:e1007580. doi: 10.1371/journal.pgen.1007580

Escudié, F., Auer, L., Bernard, M., Mariadassou, M., Cauquil, L., Vidal, K., et al. (2018). FROGS: find, rapidly, OTUs with galaxy solution. *Bioinformatics* 34, 1287–1294. doi: 10.1093/bioinformatics/btx791

Fernando, S. C., Purvis, H. T., Najar, F. Z., Sukharnikov, L. O., Krehbiel, C. R., Nagaraja, T. G., et al. (2010). Rumen microbial population dynamics during adaptation to a high-grain diet. *Appl. Environ. Microbiol.* 76, 7482–7490. doi: 10.1128/AEM.00388-10

Fouhy, F. (2016). 16S rRNA gene sequencing of mock microbial populations-impact of DNA extraction method, primer choice and sequencing platform. *BMC Microbiol.* 16:123. doi: 10.1186/s12866-016-0738-z

Geishauser, T., and Gitzel, A. (1996). A comparison of rumen fluid sampled by oro-ruminal probe versus rumen fistula. *Small Rumin. Res.* 21, 63–69. doi: 10.1016/0921-4488(95)00810-1

Gerasimidis, K., Bertz, M., Quince, C., Brunner, K., Bruce, A., Combet, E., et al. (2016). The effect of DNA extraction methodology on gut microbiota research applications. *BMC. Res. Notes* 9:365. doi: 10.1186/s13104-016-2171-7

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* 8:2224. doi: 10.3389/fmicb.2017.02224

Henderson, G., Cox, F., Ganesh, S., Jonker, A., Young, W., and Janssen, P. H. (2015). Rumen microbial community composition varies with diet and host, but a core microbiome is found across a wide geographical range. *Sci. Rep.* 5:14567. doi: 10.1038/srep14567

Henderson, G., Cox, F., Kittelmann, S., Miri, V. H., Zethof, M., Noel, S. J., et al. (2013). Effect of DNA extraction methods and sampling techniques on the apparent structure of cow and sheep rumen microbial communities. *PLoS One* 8:e74787. doi: 10.1371/journal.pone.0074787

Huang, S., Ji, S., Suen, G., Wang, F., and Li, S. (2021). The rumen bacterial Community in Dairy Cows is Correlated to production traits During freshening period. *Front. Microbiol.* 12:630605. doi: 10.3389/fmicb.2021.630605

Huang, S., Ji, S., Yan, H., Hao, Y., Zhang, J., Wang, Y., et al. (2020). The day-to-day stability of the ruminal and fecal microbiota in lactating dairy cows. *MicrobiologyOpen* 9:e990. doi: 10.1002/mbo3.990

Hungate, R. E. (1966). *The Rumen and its Microbes*. New York/London: Academic Press.

Jami, E., and Mizrahi, I. (2012). Composition and similarity of bovine rumen microbiota across individual animals. *PLoS One* 7:e33306. doi: 10.1371/journal.pone.0033306

Jami, E., White, B. A., and Mizrahi, I. (2014). Potential role of the bovine rumen microbiome in modulating milk composition and feed efficiency. *PLoS One* 9:e85423. doi: 10.1371/journal.pone.0085423

Jewell, K. A., McCormick, C. A., Odt, C. L., Weimer, P. J., and Suen, G. (2015). Ruminal bacterial community composition in dairy cows is dynamic over the course of two lactations and correlates with feed efficiency. *Appl. Environ. Microbiol.* 81, 4697–4710. doi: 10.1128/AEM.00720-15

Kchouk, M., Gibrat, J. F., and Elloumi, M. (2017). Generations of sequencing technologies: from first to next generation. *Biol. Med.* 9:3. doi: 10.4172/0974-8369.1000395

Liu, Z., Lozupone, C., Hamady, M., Bushman, F. D., and Knight, R. (2007). Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res.* 35:e120. doi: 10.1093/nar/gkm541

Liu, X., Sha, Y., Dingkao, R., Zhang, W., Lv, W., Wei, H., et al. (2020). Interactions between rumen microbes, VFAs, and host genes regulate nutrient absorption and epithelial barrier function during cold season nutritional stress in tibetan sheep. *Front. Microbiol.* 11:593062. doi: 10.3389/fmicb.2020.593062

Lodge-Ivey, S. L., Browne-Silva, J., and Horvath, M. B. (2009). Technical note: bacterial diversity and fermentation end products in rumen fluid samples collected via oral lavage or rumen cannula. *J. Anim. Sci.* 87, 2333–2337. doi: 10.2527/jas.2008-1472

Louca, S., Polz, M. F., Mazel, F., Albright, M. B. N., Huber, J. A., O'Connor, M. I., et al. (2018). Function and functional redundancy in microbial systems. *Nat. Ecol. Evol.* 2, 936–943. doi: 10.1038/s41559-018-0519-1

Lourenço, M., Ramos-Morales, E., and Wallace, R. J. (2010). The role of microbes in rumen lipolysis and biohydrogenation and their manipulation. *Animal* 4, 1008–1023. doi: 10.1017/S175173111000042X

Mamun, M. A. A., Sandeman, M., Rayment, P., Brook-Carter, P., Scholes, E., Kasinadhuni, N., et al. (2020). The composition and stability of the faecal microbiota of merino sheep. *J. Appl. Microbiol.* 128, 280–291. doi: 10.1111/jam.14468

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:e0254874. doi: 10.1371/journal.pone.0254874

Matsui, H., Ogata, K., Tajima, K., Nakamura, M., Nagamine, T., Aminov, R. I., et al. (2000). Phenotypic characterization of polysaccharidases produced by four Prevotella type strains. *Curr. Microbiol.* 41, 45–49. doi: 10.1007/s002840010089

McMurdie, P. J., and Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* 10:e1003531. doi: 10.1371/journal.pcbi.1003531

Park, P. W., and Goins, R. E. (1994). In situ preparation of fatty acid methyl esters for analysis of fatty acid composition in foods. *J. Food Sci.* 59, 1262–1266. doi: 10.1111/j.1365-2621.1994.tb14691.x

Pitta, D. W., Kumar, S., Vecchiarelli, B., Shirley, D. J., Bittinger, K., Baker, L. D., et al. (2014). Temporal dynamics in the ruminal microbiome of dairy cows during the transition period. *J. Anim. Sci.* 92, 4014–4022. doi: 10.2527/jas.2014-7621

Plaizier, J. C., Li, S., Tun, H. M., and Khafipour, E. (2017). Nutritional models of experimentally-induced subacute ruminal acidosis (SARA) differ in their impact on rumen and hindgut bacterial communities in dairy cows. *Front. Microbiol.* 7:2128. doi: 10.3389/fmicb.2016.02128

Playne, M. J. (1985). Determination of ethanol, volatile fatty acids, lactic and succinic acids in fermentation liquids by gas chromatography. *J. Sci. Food Agric.* 36, 638–644. doi: 10.1002/jsfa.2740360803

Plummer, E., and Twin, J. (2015). A comparison of three bioinformatics pipelines for the analysis of preterm gut microbiota using 16S rRNA gene sequencing data. *J Proteomics Bioinform.* 8:12. doi: 10.4172/jpb.1000381

Pollock, J., Glendinning, L., Wisedchanwet, T., and Watson, M. (2018). The madness of microbiome: attempting to find consensus "best practice" for 16S microbiome studies. *Appl. Environ. Microbiol.* 84, e02627–e02717. doi: 10.1128/AEM.02627-17

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/

Sasson, G., Kruger Ben-Shabat, S., Seroussi, E., Doron-Faigenboim, A., Shterzer, N., Yaacoby, S., et al. (2017). Heritable bovine rumen bacteria are phylogenetically related and correlated with the cow's capacity to harvest energy from its feed. *MBio* 8, e00703–e00717. doi: 10.1128/mBio.00703-17

Schloss, P. D. (2010). The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. *PLoS Comput. Biol.* 6:e1000844. doi: 10.1371/journal.pcbi.1000844

Skarlupka, J. H., Kamenetsky, M. E., Jewell, K. A., and Suen, G. (2019). The ruminal bacterial community in lactating dairy cows has limited variation on a day-to-day basis. *J. Animal Sci. Biotechnol.* 10:66. doi: 10.1186/s40104-019-0375-0

Stewart, C. S., Flint, H. J., and Bryant, M. P. (1997). "The rumen bacteria," in *The Rumen Microbial Ecosystem*. eds. P. N. Hobson and C. S. Stewart (London, UK: Blackie academic & professional), 10–71.

Terré, M., Castells, L., Fàbregas, F., and Bach, A. (2012). Short communication: comparison of pH, volatile fatty acids, and microbiome of rumen samples from preweaned calves obtained via cannula or stomach tube. *J. Dairy Sci.* 96, 5290–5294. doi: 10.3168/jds.2012-5921

Tremblay, J., Singh, K., Fern, A., Kirton, E. S., He, S., Woyke, T., et al. (2015). Primer and platform effects on 16S rRNA tag sequencing. *Front. Microbiol.* 6:771. doi: 10.3389/fmicb.2015.00771

Vaidya, J. D., van den Bogert, B., Edwards, J. E., Boekhorst, J., van Gastelen, S., Saccenti, E., et al. (2018). The effect of DNA extraction methods on observed microbial communities from fibrous and liquid rumen fractions of dairy cows. *Front. Microbiol.* 9:92. doi: 10.3389/fmicb.2018.00092

Wang, Y., and LêCao, K.-A. (2020). Managing batch effects in microbiome data. *Brief. Bioinform.* 21, 1954–1970. doi: 10.1093/bib/bbz105

Warner, A. C. I. (1966). Diurnal changes in the concentrations of micro-organisms in the rumens of sheep fed limited diets once daily: with an appendix on the kinetics of rumen microbes and flow. *J. Gen. Microbiol.* 45, 213–235. doi: 10.1099/00221287-45-2-213

Weimer, P. J. (2015). Redundancy, resilience, and host specificity of the ruminal microbiota: implications for engineering improved ruminal fermentations. *Front. Microbiol.* 6:296. doi: 10.3389/fmicb.2015.00296

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 5:27. doi: 10.1186/s40168-017-0237-y

Wohl, D. L., Arora, S., and Gladstone, J. R. (2004). Functional redundancy supports biodiversity and ecosystem function in a closed and constant environment. *Ecology* 85, 1534–1540. doi: 10.1890/03-3050

Xue, M., Sun, H., Wu, X., Guan, L. L., and Liu, J. (2018). Assessment of rumen microbiota from a large dairy cattle cohort reveals the pan and core bacteriomes contributing to varied phenotypes. *Appl. Environ. Microbiol.* 84, e00970–e00988. doi: 10.1128/AEM.00970-18

Yoshihara, Y., and Ogawa, Y. (2021). Handling stress-induced ruminal microbiota chnages reduce grass hay degradability in sheep. *J. Anim. Feed Sci.* 30, 159–164. doi: 10.22358/jafs/135543/2021

Zened, A., Troegeler-Meynadier, A., Nicot, M. C., Combes, S., Cauquil, L., Farizon, Y., et al. (2011). Starch and oil in the donor cow diet and starch in substrate differently affect the in vitro ruminal biohydrogenation of linoleic and linolenic acids. *J. Dairy Sci.* 94, 5634–5645. doi: 10.3168/jds.2011-4491

Zhu, Z., Difford, G. F., Noel, S. J., Lassen, J., Løvendahl, P., and Højberg, O. (2021). Stability assessment of the rumen bacterial and archaeal communities in dairy cows within a single lactation and its association with host phenotype. *Front. Microbiol.* 12:636223. doi: 10.3389/fmicb.2021.636223

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Fresco, Marie-Etancelin, Meynadier and Martinez Boggio. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Conclusion of section 4.3

We studied the stability of the bacterial composition as microbial community or as individual OTU. We also evaluated the conservation from one week to the next of the associations between ruminal bacteria and FAs in milk and rumen. Although repeatability was low for most OTUs, 66 bacterial abundances were quite stable and associated with milk and rumen FAs. Associations between microbial abundances and FAs were more conserved working at higher taxonomic levels, such as the family level rather than the OTU level.

To conclude, the rumen bacterial abundances were not stable from one week to the next, but some phenotypic association between bacterial family abundances and milk and rumen FAs were conserved. The instability of rumen bacteria needs to be considered for future interpretations.

4.4 Ruminal bacteria heritabilities: which is the impact of data processing?

72nd Annual Meeting of the European Federation of Animal Science (EAAP).

September 2021 – Davos, Switzerland.

Estimation of genetic parameters for ruminal bacteria in Lacaune dairy ewes

Guillermo Martinez-Boggio[1]*, Annabelle Meynadier[1], Charlotte Allain[2], Christel Marie-Etancelin[1]

*guillermo.martinez-boggio@inrae.fr

[1]GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326 Castanet Tolosan, France

[2]INRAE, Experimental Unit of La Fage, F-12250 Saint-Jean et Saint-Paul, France

Among livestock species, ruminants such as sheep have the ability to digest the plant fiber, thanks to the main action of symbiotic ruminal bacteria. Some authors studied the link between the microbiota and different complex traits in cattle, and more recently in sheep. But what is the level of host impact on ruminal bacterial composition? The aim of this study is to estimate the heritability of Operational Taxonomic Units (OTUs) and bacterial genera in a population of Lacaune dairy ewes.

The data corresponds to 800 dairy Lacaune ewes raised at the INRAE Experimental Farm of La Fage (France). These ewes were adult animals, mostly in their second lactation, feeding the same diet (TMR), and the rumen sampling was done between 2015 and 2019. Ruminal metagenome was sequenced using metabarcoding (V3-V4 of gene of 16S rRNA) with Illumina Miseq technology. The microbiota sequences were analyzed with FROGS pipeline to obtain the relative abundances of 2,059 OTUs, grouped in 112 genera. In the abundance table the zero values were replaced with the GBM method, to allow for Centered Log-Ratio (CLR) transform all values.

The additive host genetic effects on the rumen bacteria abundance were estimated applying a linear mixed model. The general equation included the OTUs and genera abundance as response variable; fixed effects, such as stage of lactation (45 to 133 DIM), lactation number (2 up to 7 lactations), genetic line (SCS+, SCS-, PERS+, PERS- and SOCS genotype), run sequencing effect, date and year of the rumen sampling; and the individual animal random additive genetic effects ~ $N(0,A\sigma^2_a)$, where σ^2_a is the additive genetic variance and A is the pedigree relationship matrix.

With these methodologies, we will estimate the heritability of OTUs and bacterial genera in a population of Lacaune dairy ewes.

72nd Annual Meeting of the European Federation of Animal Science (EAAP) in September 2021 in Davos (Switzerland).

Ruminal bacteria heritabilities: which is the impact of different data processing ?

Guillermo Martinez Boggio[1], Annabelle Meynadier[1], Charlotte Allain[2], Christel Marie-Etancelin[1]
[1]GenPhySE, Université de Toulouse, INRAE, ENVT, France. [2]INRAE, Experimental Unit of La Fage, France

Introduction and objective

Microbiome data is compositional: the information is found in the operational taxonomic unit (OTU) ratios.

As high-dimensional data with zeros, it is necessary to replace the zeros, and apply value transformation.

Conclusions

Ruminal bacteria h^2 are affected by data processing.

Zeros replacement impact the most the number of heritable OTUs.

GBM leads to less difference between the two data transformation.

To compare **data processing methods** in terms of **heritability** estimates (h^2).

Materials and methods



Rumen samples of 795 Lacaune ewes
Metabarcoding 16s rRNA gene

Table 1. Data processing methods to be compare.

	GBM	C1	Z01
CLR	GBM-CLR	C1-CLR	Z01-CLR
Log	GBM-Log	C1-Log	Z01-Log

Zeros replacement:
GBM= Geometric Bayesian Method
C1= Add constant equal 1
Z01= Replace only zeros with 0.001

Data transformation:
CLR= Centred Log-Ratio transformation
Log= Logarithm transformation

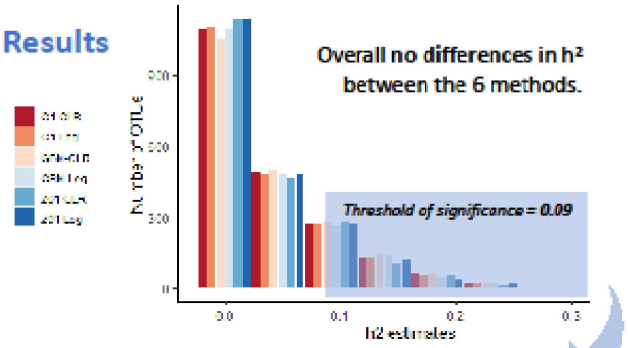
Heritability estimates*:

$$y = Xb + Za + e$$

y: OTU abundances processed with each of the 6 methods.
b: fixed effects lactation stage, litter size, year, number of lactation, sequencing run, sampling order and total nb. sequences.
a: random animal effect ~ $N(0, \sigma^2_a A)$.
e: residual ~ $N(0, \sigma^2_e I)$.

*The threshold of significance obtained by phenotype-animal permutation

Results



But, the number of **heritable OTUs** defined by the methods were **variable**.

Table 2. Number of heritable OTUs by processing data method.*

	GBM	C1	Z01
CLR	326	293	249
Log	294	274	254

GBM > C1 > Z01

By replacing only the zeros we obtain lower number of heritable OTUs.

% OTUs in common between CLR and Log

78% 72%

Using **GBM** the differences between CLR and Log are lesser than with C1.

Conclusion of section 4.4

In general, the heritability estimates obtained with the six methodologies presented a similar distribution with the most OTUs having heritabilities closed to zero. However, methods showed differences in the number of OTUs that passed the threshold of significance (h^2> 0.09). In this regard, the method Z01 which only replace the zero values in data, is the one that shows the greatest difference, and from a theoretical point of view, it distorts the ratios between OTUs. CLR and log-transformation results were quite similar, depending on the imputation of zero values by GBM or C1, but in general CLR retained a higher number of OTUs.

To conclude, the processing of microbiota data has an impact on the estimates of genetic parameter, and it is recommended to use the compositional data approach.

4.5 Conclusion

Due to advances in NGS technologies, it is now very common to work with high-dimensional data, such as metagenomics. This chapter emphasizes the importance of working correctly by applying the most appropriate methodologies available, considering the nature of the data studied. Since microbial abundances are compositional data, we defined a workflow adapted, applied and compared with other data processing methods in terms of phenotypic and genetic estimates. However, the instability of the microbial composition and, to a lesser extent, of their associations with phenotypes over a short period of time will allow for us to put our future results into perspective.

Chapter 5
Host genetic control on ruminal microbiota

5.1 Introduction

In ruminants, it is well known that milk yield and fine composition are derived from nutrients in the diet that are available through digestion by the rumen microbiota (Hungate, 1966). This phenotypic link between rumen microbiota and milk traits was evidenced in Chapter 4, where rumen bacterial abundances were associated mainly with milk composition and rumen FAs. However, the correlations obtained were low to moderate for a small group of bacteria. To go further in understanding of these associations between rumen microbiota and phenotypes, it is first necessary to know what genetic control the host has over both. In sheep, milk composition traits and LSCS were known to be under a moderate to high genetic control of the host. But, there is no information on the heritabilities of rumen bacteria, neither genetic correlations between rumen bacteria and milk traits, nor potential genomic regions that determine the expression of both traits simultaneously. This will be key to answering questions such as: whether the host is able to control the microbial composition? and what mechanisms account for this control as well as associations with dairy traits? And furthermore, what role does the microbiota play in the control of diseases such as mastitis? The objectives of this chapter are to quantify the host genetic effect on rumen bacterial abundances and the genetic correlations between rumen bacterial abundance and milk and udder health traits, as well as to identify QTLs simultaneously associated with bacterial abundances and milk and udder health traits.

Results obtained for bacteria heritabilities, genetic correlations and GWASs for rumen bacterial abundances and dairy traits, were published under the title **"Host genetic control on rumen microbiota and its impact on dairy traits in sheep"** in Genetic Selection Evolution Journal in November 2022. As supplementary results to the aforementioned publication, we included

the distribution of the heritabilities for rumen bacteria, and the estimations of heritability for alpha-diversity indices. Secondly, the results obtained for the genetic correlations between rumen bacterial abundance and LSCS and two-trait GWAS were presented under the title **"Evidence of genetic links between rumen bacteria and milk somatic cell score in Lacaune dairy sheep"** in Poster format at the 12th World Congress on Genetics Applied to Livestock Production in July 2022 in Rotterdam (The Netherlands).

5.2 Host genetic control on rumen microbiota and its impact on dairy traits in sheep

Martinez Boggio *et al.*
Genetics Selection Evolution (2022) 54:77
https://doi.org/10.1186/s12711-022-00769-9

GSE Genetics Selection Evolution

RESEARCH ARTICLE Open Access

Host genetic control on rumen microbiota and its impact on dairy traits in sheep

Guillermo Martinez Boggio[1*], Annabelle Meynadier[1], Albert Johannes Buitenhuis[2] and Christel Marie-Etancelin[1]

Abstract

Background: Milk yield and fine composition in sheep depend on the volatile and long-chain fatty acids, microbial proteins, vitamins produced through feedstuff digestion by the rumen microbiota. In cattle, the host genome has been shown to have a low to moderate genetic control on rumen microbiota abundance but a high control on dairy traits with heritabilities higher than 0.30. There is little information on the genetic correlations and quantitative trait loci (QTL) that simultaneously affect rumen microbiota abundance and dairy traits in ruminants, especially in sheep. Thus, our aim was to quantify the effect of the host genetics on rumen bacterial abundance and the genetic correlations between rumen bacterial abundance and several dairy traits, and to identify QTL that are associated with both rumen bacterial abundance and milk traits.

Results: Our results in Lacaune sheep show that the heritability of rumen bacterial abundance ranges from 0 to 0.29 and that the heritability of 306 operational taxonomic units (OTU) is significantly different from 0. Of these 306 OTU, 96 that belong mainly to the *Prevotellaceae*, *Lachnospiraceae* and *Ruminococcaceae* bacterial families show strong genetic correlations with milk fatty acids and proteins (absolute values ranging from 0.33 to 0.99). Genome-wide association studies revealed a QTL for alpha-lactalbumin concentration in milk on *Ovis aries* chromosome (OAR) 11, and six QTL for rumen bacterial abundances i.e., for two OTU belonging to the genera *Prevotella* (OAR3 and 5), *Rikenellaceae_RC9_gut_group* (OAR5), *Ruminococcus* (OAR5), an unknown genus of order *Clostridia UCG-014* (OAR10), and *CAG-352* (OAR11). None of these detected regions are simultaneously associated with rumen bacterial abundance and dairy traits, but the bacterial families *Prevotellaceae*, *Lachnospiraceae* and *F082* show colocalized signals on OAR3, 5, 15 and 26.

Conclusions: In Lacaune dairy sheep, rumen microbiota abundance is partially controlled by the host genetics and is poorly genetically linked with milk protein and fatty acid compositions, and three main bacterial families, *Prevotellaceae*, *Lachnospiraceae* and *F082*, show specific associations with OAR3, 5, 15 and 26.

Background

Ruminants can digest plant fiber thanks to the symbiotic microbiota in their rumen. This complex microbial community is composed mainly of bacteria but also includes archaea, fungi and protozoa. Bacteria degrade and ferment fibrous feedstuffs to produce volatile fatty acids (FA), microbial proteins and vitamins and to transform dietary lipids, all of which are used by the animal for maintenance, growth and lactation [1]. The rumen microbiota has been shown to be associated with production traits in dairy cows [2–4] and, more recently, in dairy sheep [5]. Furthermore, some authors have demonstrated that rumen microbiota abundance is under low to moderate control by the host genome [2, 4, 6], which may provide the opportunity to select animals with specific

*Correspondence: guillermo.martinez-boggio@inrae.fr

[1] GenPhySE, INRAE, ENVT, Université de Toulouse, 24 Chemin de Borde Rouge, 31326 Castanet-Tolosan, France
Full list of author information is available at the end of the article

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

microbial communities that are associated with dairy traits. However, to date, there are no published studies on the genetic association between the rumen microbiota and milk traits in sheep.

In dairy sheep, fine milk composition traits are extremely important for the production of high-quality cheese. The FA present in milk influence its texture and nutritional value, and the proteins influence its coagulation capacity, which in turn affect its heat stability and the cheese yield [7]. Dairy traits in sheep have moderate to high heritabilities (h^2) ranging from 0.30 to 0.60 [7, 8], and numerous quantitative trait loci (QTL) and major genes associated with dairy traits have been identified in the ovine genome [7, 9–11]. The detection of QTL contributes to the identification of potential candidate genes and of the underlying mechanisms that determine the genetic expression of these relevant traits. However, for rumen microbiota abundance, there is limited information about possible associated genomic regions. Thus, the identification of QTL that are simultaneously associated with rumen bacterial abundance and milk traits can help reveal the basis of the genetic link and the shared metabolic pathways between the rumen microbiota and dairy traits in sheep.

We hypothesised that, in Lacaune ewes, rumen microbiota abundance is affected by the host genetics and is genetically linked to dairy traits through shared metabolic pathways. Thus, based on the use of the dataset that was previously described by Martinez Boggio et al. [5], the objectives of the current study were to (1) quantify the effect of the host genetics on rumen bacterial abundance by estimating its heritability, (2) identify and quantify the genetic links between rumen bacteria and dairy traits by estimating the genetic correlations between the two, and (3) identify the QTL and potential underlying mechanisms that determine the genetic variation in rumen bacterial abundance and in dairy traits by identifying shared genomic regions between the two through genome-wide association studies (GWAS).

Methods

Data structure

Data were obtained from the INRAE Experimental Unit of La Fage (UE 321 agreement A312031, Roquefort, France) between 2015 and 2019. Multiparous Lacaune dairy ewes (mean weight of 77 ± 9 kg) were raised indoors and fed a mixed ration of on average 90% meadow hay and silage plus 10% barley (on a gross matter basis) supplemented with approximately 150 g of a commercial protein concentrate (38% of crude protein on a dry matter basis) distributed in the milking parlor. Adjustment of the percentage of concentrates and forages was done each year according to the feeding value of the forages to cover the needs of the ewes, thus they received the same amount of nutrients during the 5 years. On average during this period, the ewes ingested 3.27 kg of dry matter that contained 16% of crude protein and 30% of crude fiber. The genetic structure of the INRAE La Fage flock consists of two independent groups of ewes, which were both divergently selected, as described by Martinez Boggio et al. [5]. Briefly, genetic selection was based on the estimated breeding values (EBV) for milk somatic cell score (SCS) [8] or on the EBV for milk persistency (PERS), expressed as the coefficient of variation in milk production (CV milk). Ewes belonging to each of these two lines were studied (N = 700). The genetic difference between the ewes from the divergent lines for SCS (SCS+ /SCS−) was 2.19 units of SCS EBV [i.e., a 3.60 genetic standard deviation (SD)], and that for PERS (PERS+/PERS−) was 5.52 units of CV milk EBV (i.e., a 2.10 genetic SD). Ninety-five additional ewes were included in the dataset, which were derived from the oldest SCS line but are currently selected to increase the frequency of the mutant allele (T) of the *suppressor of cytokine signalling 2* (*SOCS2*) gene in the experimental Lacaune population to investigate possible associations with other traits. This allele corresponds to a mutation that was identified in 2015 [12] and explains 12% of the genetic variance in somatic cell count in Lacaune dairy sheep. Thus, the final experimental dataset consisted of data from 795 ewes, including 298 under SCS selection (94 SCS+ and 204 SCS−), 402 under PERS selection (200 PERS+ and 202 PERS−), and 95 under selection for the *SOCS2* mutation.

Rumen sampling and analysis of the bacterial community

Rumen sampling was performed within 3 days of milk recording. Ruminal contents were sampled from each ewe using a vacuum pump and a medical gastric tube. Then, DNA was extracted and purified from a ruminal sample of 80 µL using the QIAamp DNA Stool Mini Kit (Qiagen Ltd, West Sussex, UK) according to the manufacturer's instructions, with a previous bead-beating step in a FastPrep instrument (MP Biomedicals, Illkirch, France). The 16S rRNA V3–V4 regions were amplified by a first round of PCR with 30 cycles using the following primers: forward F343 (5′-CTTTCCCTACACGACGCTCTTCCGATCTACGGRAGGCAGCAG-3′; [13]) and reverse R784 (5′-GGAGTTCAGACGTGTGCTCTTCCGATCTTACCAGGGTATCTAATCCT-3′; [14]). Since the Illumina MiSeq technology results in 250-base paired-end reads, we obtained overlapping reads that generated extremely high-quality, full-length reads of the entire V3 and V4 regions in a single run. Single multiplexing was performed using a 6-base pair (bp) index, which was added to the R784 primer, during a second round of PCR

with 12 cycles with home-made primers including also Illumina adapters: forward (AATGATACGGCGACC ACCGAGATCTACACTCTTTCCCTACACGAC) and reverse (CAAGCAGAAGACGGCATACGAGATGTG ACTGGAGTTCAGACGTGT). The resulting PCR products were purified and loaded onto an Illumina MiSeq cartridge (Illumina, San Diego, CA, USA) at the Genomic and Transcriptomic Platform (INRAE, Toulouse, France) according to the manufacturer's instructions. More details on rumen sampling, DNA extraction and amplicon sequencing are provided in Martinez Boggio et al. [5]. The sequences of the 795 samples were processed using the FROGS 3.0 pipeline [15] as follows: (i) read preprocessing, which consists in the removal of sequences that present a primer mismatch, display an unexpected length i.e., shorter than 300 bp or longer than 500 bp, or that contain at least one ambiguous base; (ii) removal of chimera; (iii) regrouping of sequences by clustering with Swarm in FROGS, and we chose the parameters for a distance equal to 1; (iv) cluster filtering, i.e., removal of the clusters with abundances lower than 0.005% [16]; and (v) taxonomic assignment to operational taxonomic units (OTU) using the SILVA database (version 138) [17] (see Additional file 1: Table S1). The abundance table and the taxonomy files were imported into R (v4.0.2) [18]. The core microbiome was quantified based on the occurrence of OTU across multiple samples, and the proportion of samples over which OTU must occur was set to 90% [19].

Analysis of dairy traits

Official daily records of milk yield (MY), milk somatic cell count as quantified with a Fossomatic cell counter (Foss, Nanterre, France), and milk fat and protein contents (FC and PC, respectively) were obtained on the 795 adult ewes between 28 and 133 days in milk (DIM). Two milk samples collected per animal during morning and afternoon milking were sent for analysis at the Interprofessional Milk Analysis Laboratory (Agrolab's Aurillac, France). Milk FC and PC were analyzed with mid-infrared (MIR) techniques on a Milko-Scan™ FT6000 instrument (Foss, Nanterre, France). Thus, we analyzed MY, FC and PC, with FC and PC taken as averages weighted by morning and afternoon milk yield records.

From these official milk records, the MIR spectra for 563 ewes were retrieved to predict the fine profile of the daily milk proteins, i.e., the four caseins: alpha-S1-casein (α_{s1}-CN), alpha-S2-casein (α_{s2}-CN), beta-casein (β-CN) and kappa-casein (κ-CN), and two whey proteins: alpha-lactalbumin (α-lactalbumin) and beta-lactoglobulin (β-lactoglobulin), and of the FA, i.e., saturated FA (SFA), such as butyric acid (C4:0), caproic acid (C6:0), caprylic acid (C8:0), capric acid (C10:0), lauric acid (C12:0), and palmitic acid (C16:0), and unsaturated FA (UFA), such as oleic acid (*cis-9* C18:1), rumenic acid (*cis-9 trans-11* C18:2) and alpha-linolenic acid (C18:3*n-3*). Milk proteins and FA are expressed in g per 100 mL as averages weighted by morning and afternoon milk yield records. The MIR prediction accuracy, as estimated by the value of the coefficient of determination (R^2), were retrieved from Ferrand et al. [20] and Ferrand-Calmels et al. [21] for the milk proteins and FA, respectively. The R^2 for caseins are higher than 0.82, and for β-lactoglobulin and α-lactalbumin are equal to 0.77 and 0.26, respectively. The R^2 for the SFA and *cis-9* C18:1 are higher than 0.93, and for *cis-9 trans-11* C18:2 and C18:3*n*-3 are equal to 0.91 and 0.74, respectively. Two additional traits, i.e., average lactation somatic cell score (LSCS) and CV milk were included in the analysis to account for the genetic structure of the population that is formed of two divergent lines for SCS and PERS.

Genotyping

DNA extraction from blood samples and genotyping were performed for the 795 ewes. Of these 795 ewes, 743 were genotyped using a medium-density single nucleotide polymorphism (SNP) chip (Illumina Ovine SNP50 BeadChip: 54,241 SNPs), 314 at the Laboratoire d'Analyses Génétiques pour les Espèces Animales (Jouy-en-Josas, France) and 429 at Aveyron-Labo (Rodez, France). The remaining 52 ewes were genotyped with a low-density SNP chip (Illumina Ovine SNP15: 16,681 SNPs) at Neogen (Lansing, USA), followed by imputation to a medium-density SNP chip within the framework of the Lacaune dairy sheep genomic selection programme [22]. Genotypes were subjected to quality control, based on minimum call rates of 90% for SNPs and 95% for individuals and on the exclusion of SNPs with a minor allele frequency lower than 5%. The final dataset included 773 genotyped individuals and 35,492 autosomal SNPs. Markers were positioned on the 26 *Ovis aries* (OAR) autosomes and mapped to the *Ovis aries* genome assembly Oar_v3.1 [23]. The SNP corresponding to the mutation in the *SOCS2* gene was included in the map of OAR3 (129,722,200 bp).

Statistical analyses

Microbiota abundance data are compositional data [24]. The abundances of each sample are constrained to a total sum imposed by the sequencing technology used. Therefore, the information is contained in the ratios between OTU abundances, and the raw counts are irrelevant. This produces an interdependence between abundances, and they need to be considered as compositional data [25]. Thus, we applied an approach based on compositional data theory [26] which consisted in transforming the counts into centered log-ratios. As zeros are not

compatible with the log-ratios transformation, the OTU abundance data were zero imputed with the geometric Bayesian-multiplicative method [27] through the cmultRepl function of the zCompositions package [28] in R (v4.0.2) [18]. Then, they were centered log-ratio (CLR) transformed with the function clr of the compositions package [29] in R (v4.0.2) [18] and standardised to a variance equal to 1. Hereafter, OTU abundance refers to CLR-transformed abundance data.

Estimation of variance components

Environmental factors were tested by analyzing the variance of each trait. Statistical significance was defined at $P < 0.05$. Environmental factors that were significant for the abundance of more than 10% of the OTU were included in the model (Table 1). All traits were tested for the year of sampling (five levels: 2015 to 2019), the number of lactations (three levels: 2, 3, and 4 or more lactations) and litter size (two levels: 1 and 2 or more lambs) as fixed effects, and DIM (28 to 133 DIM) as a covariable. OTU abundances were also tested with the sequencing run for DNA samples (six levels), total sequence number per DNA sample (five levels: ≤ 5000; >5000 and $\leq 10{,}000$; $>10{,}000$ and $\leq 15{,}000$; $>15{,}000$ and $\leq 20{,}000$; and $>20{,}000$ sequences), time of rumen sampling (morning and afternoon) and the order in which the animal was rumen-sampled (eight levels) as fixed effects. For LSCS and CV milk, which are expressed on a lactation basis, the number of milk recording controls was included as a fixed effect (four levels: 4 to 6, 7, 8, and 9 test-days).

In this study, we accounted for the structure of the population under selection through the use of multiple-trait models [30], including the selected traits (LSCS and CV milk) as the first two traits in each model. Two multiple-trait models were used. First, to estimate heritabilities, we used a three-trait model with the abundance of each OTU or a dairy trait as a third trait, and second, to estimate the genetic correlations between OTU abundance and the dairy traits, we used a four-trait model with the abundance of each OTU and a dairy trait as third and fourth traits, respectively. Genetic correlations with absolute values higher than twice the standard error were considered to differ from zero.

The data were analyzed using the following multiple-trait animal model:

$$\begin{bmatrix} \mathbf{y_1} \\ \mathbf{y_2} \\ \mathbf{y_3} \\ \mathbf{y_4} \end{bmatrix} = \begin{bmatrix} \mathbf{X_1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{X_2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{X_3} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{X_4} \end{bmatrix} \begin{bmatrix} \mathbf{b_1} \\ \mathbf{b_2} \\ \mathbf{b_3} \\ \mathbf{b_4} \end{bmatrix} + \begin{bmatrix} \mathbf{W_1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{W_2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{W_3} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{W_4} \end{bmatrix} \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a_2} \\ \mathbf{a_3} \\ \mathbf{a_4} \end{bmatrix} + \begin{bmatrix} \mathbf{e_1} \\ \mathbf{e_2} \\ \mathbf{e_3} \\ \mathbf{e_4} \end{bmatrix}, \quad (1)$$

where $\mathbf{y_1}$, $\mathbf{y_2}$, $\mathbf{y_3}$, and $\mathbf{y_4}$ are the vectors of observations for LSCS, CV milk, OTU abundance or one dairy trait in the three-trait model, and OTU abundance and one dairy trait in the four-trait model; $\mathbf{b_1}$, $\mathbf{b_2}$, $\mathbf{b_3}$, and $\mathbf{b_4}$ are the vectors of fixed effects described in Table 1 for each trait; $\mathbf{a_1}$, $\mathbf{a_2}$, $\mathbf{a_3}$, and $\mathbf{a_4}$ are the vectors of additive genetic effects; and $\mathbf{e_1}$, $\mathbf{e_2}$, $\mathbf{e_3}$, and $\mathbf{e_4}$ are the vectors of residual effects. $\mathbf{X_1}$, $\mathbf{X_2}$, $\mathbf{X_3}$, and $\mathbf{X_4}$ are incidence matrices relating fixed effects to vectors $\mathbf{y_1}$, $\mathbf{y_2}$, $\mathbf{y_3}$ and $\mathbf{y_4}$, respectively; $\mathbf{W_1}$, $\mathbf{W_2}$, $\mathbf{W_3}$, and $\mathbf{W_4}$ are incidence matrices relating additive effects to vectors $\mathbf{y_1}$, $\mathbf{y_2}$, $\mathbf{y_3}$ and $\mathbf{y_4}$, respectively.

Table 1 Environmental effects included in the animal models for operational taxonomic units (OTU) and dairy traits

Trait	DIM	Year	Year lactation	Litter size	Year run	Total sequences	Year time: order	Test-day
OTU[a]	ns	88%	23%	ns	44%	15%	44%	
Milk yield	†	†	ns	ns				
Fat content	†	†	†	ns				
Protein content	†	†	ns	†				
Milk proteins[b]	†	†	ns	†				
Milk FA[b]	†	†	†	†				
LSCS	ns	ns	ns	†				†
CV milk	ns	†	ns	†				†

DIM days in milk, *Year lactation* number of lactation nested within year, *Year run* run of sequencing nested within year, *Total sequences* number of total sequences per DNA sample, *Year time order* sampling order nested in rumen sampling time and year, *FA* fatty acids, *LSCS* lactation somatic cell score, *CV milk* coefficient of variation in milk production

[a] Percentage of OTU with a significant effect

[b] Each model for milk protein and fatty acids includes at least one of the environmental effects

†: The effect is included in the model; ns: the effects are non-significant; blank cell: the effect is not tested

The assumptions of the model are $\mathbf{a} \sim N(0, \mathbf{A} \otimes \mathbf{P})$ and $\mathbf{e} \sim N(0, \mathbf{I} \otimes \mathbf{R})$, where $\otimes$ denotes the direct product between two matrices, **A** is the pedigree relationship matrix, **I** is an identity matrix, and **P** and **R** are the genetic and residual variance–covariance matrices for the random additive and residual effects, respectively. The pedigree of the Lacaune breed traced back to five generations of ancestors (N = 4296). The analyses were performed using the BLUPF90+ software with the OPTION method VCE [31] and by including 100 initial rounds of EM-REML to obtain initial variance components. We used a convergence criterion of $1e^{-10}$ which is the value set by default in the BLUPF90 software.

To test the significance of the heritability estimates for the 2059 OTU, an empirical significance threshold for the null hypothesis of no genetic control was estimated. For OTU, the null hypothesis was obtained by randomly shuffling their abundances among the individuals. We selected two OTU (one with many and one with few zeros) and performed the random shuffling 10,000 times. For each permutation and OTU, we estimated the heritability of OTU abundance using a three-trait model, as presented in Eq. (1). To define an error rate of 5%, we arranged the heritability estimates by increasing order, and retained the lower value of the upper 5% yielding a significance threshold obtained for both OTU of 0.10.

In order to determine whether certain bacterial genera were over- or under-represented among the OTU that had a heritability significantly different from zero compared to all the OTU, we tested the percentage of OTU of the same genus in both groups using a Fisher's exact test at $P < 0.05$.

Genome-wide association studies

Genome-wide association studies (GWAS) of dairy traits and OTU abundances were performed using the single-step genomic best linear unbiased prediction (ssGBLUP) approach [32]. The following single-trait model was used:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Wg} + \mathbf{e}, \tag{2}$$

where **y** is the vector of observations for OTU abundance or a dairy trait, **b** is the vector of fixed effects described in Table 1, **g** is the vector of additive genetic effects, and **e** is the vector of residual effects. **X** and **W** are incidence matrices for **b** and **g**, respectively. The assumptions of the model are $\mathbf{g} \sim N(0, \mathbf{H}\sigma_g^2)$, where **H** is a matrix that combines pedigree- and genome-based relationships [33] and σ_g^2 is the additive variance, and $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$, where **I** is an identity matrix and σ_e^2 is the residual variance. The joint pedigree-genomic relationship matrix **H** was constructed as follows:

$$\mathbf{H} = \begin{pmatrix} \mathbf{A_{11}} - \mathbf{A_{12}}\mathbf{A_{22}^{-1}}\mathbf{A_{21}} + \mathbf{A_{12}}\mathbf{A_{22}^{-1}}\mathbf{G}\mathbf{A_{22}^{-1}}\mathbf{A_{21}} & \mathbf{A_{12}}\mathbf{A_{22}^{-1}}\mathbf{G} \\ \mathbf{G}\mathbf{A_{22}^{-1}}\mathbf{A_{21}} & \mathbf{G} \end{pmatrix}, \tag{3}$$

which projects genomic relationships $\mathbf{G} = \mathbf{ZZ}'/2\sum p_i(1-p_i)$ [34] from genotyped animals (labelled as "2") to non-genotyped animals (labelled as "1"). The matrix $\mathbf{A} = \begin{bmatrix} \mathbf{A_{11}} & \mathbf{A_{12}} \\ \mathbf{A_{21}} & \mathbf{A_{22}} \end{bmatrix}$ is the pedigree-based relationship matrix, and the genomic relationship matrix **G** is constructed as $\mathbf{G} = (1-\alpha)\left(a + b\frac{\mathbf{ZZ}'}{2\sum p_i q_i}\right) + \alpha\mathbf{A_{22}}$, where a and b are chosen to equate average inbreeding and average relationships in **G** and $\mathbf{A}_{22}$ and α is a small value (0.05) [34–36]. The variance components were estimated using BLUPF90+ with the OPTION method VCE [31].

After solving the ssGBLUP model in Eq. (2), we estimated the SNP effects by back-solving the breeding value estimates as in Eq. (4) [32, 37] and the P value of each SNP as in Eq. (5) [38].

$$\hat{a}|\hat{g} = (1-\alpha)b\mathbf{Z}'\frac{1}{2\sum p_i q_i}\mathbf{G}^{-1}\hat{g}_2, \tag{4}$$

$$\text{Pvalue}_i = 2\left(1 - \Phi\left(\left|\frac{\hat{a}_i}{sd(\hat{a}_i)}\right|\right)\right),$$

$$\text{Var}(\hat{a}_i) = \frac{1}{2\sum p_i q_i}(1-t\alpha)bz_i'\mathbf{G}^{-1}\left(\mathbf{G}\sigma_g^2 - \mathbf{C^{u_2u_2}}\right) \times \mathbf{G}^{-1}z_i(1-\alpha)b\frac{1}{2\sum p_i q_i}, \tag{5}$$

where $\hat{a}$ are the estimates of SNP effects, $\hat{a}_i$ is the estimate for each SNP i, $\hat{g}$ are estimates of breeding values, and the $\mathbf{C^{u_2u_2}}$ matrix contains the prediction error covariance of EBV for genotyped animals. The SNP effects and P values were computed using BLUPF90+ and POSTGSF90 [31].

We corrected for multiple testing by using the false discovery rate (FDR) in the p.adjust package in R (v4.0.2) [18]. A genome-wide SNP significance threshold of $P < 0.10$ and a suggestive threshold of $P < 0.30$ were applied to each trait. After identification of the lead SNP in a given region, significant SNPs that were located less than 1000 mega base pairs (Mbp) apart and were included in the upper third of the peak were grouped within the same QTL region. The CMplot package in R (v4.0.2) [18] was used to generate Manhattan plots showing the $-\log_{10}$(P value) on the y-axis.

We estimated the allele substitution effect for significant SNPs that were detected on the same chromosome and at the same position. The following single-trait model was used:

$$\mathbf{y} = \text{SNP} + \mathbf{Xb} + \mathbf{Wa} + \mathbf{e}, \quad (6)$$

where $\mathbf{y}$ is the vector of observations for OTU abundance, SNP is the significant SNP detected for $\mathbf{y}$ as a fixed effect, coded as 0, 1, 2 when the SNP is homozygous for the first allele, heterozygous, and homozygous for the second allele, respectively, $\mathbf{b}$ is the vector of fixed effects described in Table 1 for OTU abundance, $\mathbf{a}$ is the vector of additive genetic effects, and $\mathbf{e}$ is the vector of residual effects. $\mathbf{X}$ and $\mathbf{W}$ are incidence matrices for $\mathbf{b}$ and $\mathbf{a}$, respectively. The assumptions of the model are $\mathbf{a} \sim N(0, \mathbf{A})$ and $\mathbf{e} \sim N(0, \mathbf{I})$, where $\mathbf{A}$ is the pedigree-based relationship matrix, and $\mathbf{I}$ is an identity matrix.

Linkage disequilibrium (LD) along the genome was calculated as the squared correlation of allele counts for two SNPs on each chromosome. The proportion of phenotypic variance explained was calculated for each SNP by defining a window variance of 20 adjacent SNPs and for each QTL region defined in this study. The LD and variance explained were calculated using POSTGSF90 [31].

Genes that were identified for the significant SNPs were retrieved from the Ensembl database using the BioMart web interface based on the *Ovis aries* genome assembly Oar_v3.1 [39]. Then the DAVID functional annotation tool [40] was used to analyse the overrepresented Gene Ontology (GO) biological terms, including biological processes, and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [41].

Results

After bioinformatic analysis, 63% of the 9,552,103 initial DNA sequences obtained from the rumen samples of the 795 ewes were retained. The abundance table included 2059 affiliated OTU, represented by 751 to 168,785 sequences, with a median of 1761 DNA sequences per OTU. The finest taxonomic level was the genus level due to an unknown species frequency of 95%.

Overall, the 2059 OTU from the 795 rumen samples were attributed to 11 phyla, 56 families and 112 genera. Expressed as a percentage of the total sequences for all samples, the most representative phyla were *Bacteroidota* (51%), *Firmicutes* (44%), and *Proteobacteria* (3%). At the next taxonomic level, the most abundant families were *Prevotellaceae* (38%) mainly represented by the *Prevotella* genus, *Lachnospiraceae* (18%) represented by 40 genera, *Ruminococcaceae* (9%) mainly represented by the *Ruminococcus* genus, *Oscillospiraceae* (5%) represented mainly by the NK4A214_group, and three families each represented by one genus: *Christensenellaceae* represented by *Christensenellaceae_R-7_group* (5%), and *Rikenellaceae* represented by *Rikenellaceae_RC9_gut_group* (4%). The average percentage of zeros (OTU not detected in the rumen sample) in the OTU abundance table was 37.5%.

Description of the dairy traits analyzed

The descriptive statistics for the dairy traits analyzed in our study are in Table 2. LSCS and CV milk were included to account for population genetic structure for the variance estimation of each OTU.

Rumen bacterial heritability

The heritability of rumen bacterial abundance ranged from 0 to 0.29 ± 0.07 with a mean of 0.04 ± 0.03 (see Additional file 2: Table S2). Based on these heritability estimates, we found a group of 306 OTU that had a heritability significantly different from 0 ($h^2 > 0.10$, according to the empirical threshold computed) and a mean of 0.15 ± 0.04. Among these 306 OTU, expressed as a percentage of OTU, the main phyla were *Bacteroidota* (61%), *Firmicutes* (34%) and *Spirochaetota* (2%), and the 10 most represented genera are in Table 3.

The *Prevotella* genus was significantly overrepresented (92 of the 306 OTU) among the OTU with a significant heritability based on Fisher's exact test ($P > 0.05$).

Genetic correlations between rumen bacteria and dairy traits

We estimated the genetic correlation between each of the 306 OTU and each of the 18 dairy traits included in this study. We obtained 5508 genetic correlations of which 301 (i.e., 5%) were significant (i.e., those that were greater than twice the standard error). This corresponds to 96 OTU abundances that had a significant genetic correlation with one or several dairy traits: 81 OTU were correlated with FC and milk FA (see Additional file 3: Table S3), and 56 OTU were correlated with PC and milk proteins (see Additional file 4: Table S4), and 41 of these 96 OTU were correlated with both groups of traits. For MY, we did not find significant genetic correlations with ruminal bacteria.

Genetic correlations with milk fat content and fatty acids

Daily fat content and milk FA showed significant genetic correlations with OTU from 26 genera belonging to 17 families. The bacterial families that shared the largest number of significant genetic correlations were *Prevotellaceae*, *Lachnospiraceae* and *Ruminococcaceae* (Fig. 1). The significant genetic correlations ranged from -0.97

Table 2 Descriptive statistics and heritability (h^2) of dairy traits in Lacaune ewes

	Mean	SD	CV (%)	h^2	SE h^2
Traits expressed on a daily basis[a]					
Milk yield (mL)	1946	589	30.3	0.28	0.06
Fat content (g/100 mL)	7.37	1.14	15.5	0.59	0.06
Protein content (g/100 mL)	5.71	0.52	9.1	0.57	0.06
Alpha-S1-casein	1.38	0.15	10.9	0.54	0.08
Alpha-S2-casein	0.66	0.26	39.4	0.68	0.07
Beta-casein	2.10	0.23	10.9	0.41	0.08
Kappa-casein	0.45	0.04	8.9	0.50	0.08
Alpha-lactalbumin	0.13	0.01	7.7	0.36	0.08
Beta-lactoglobulin	0.46	0.50	108.7	0.44	0.08
Butyric acid (C4:0)	0.25	0.03	12.0	0.50	0.08
Caproic acid (C6:0)	0.21	0.03	14.3	0.53	0.07
Caprylic acid (C8:0)	0.20	0.03	15.0	0.55	0.08
Capric acid (C10:0)	0.73	0.12	16.4	0.58	0.07
Lauric acid (C12:0)	0.49	0.08	16.3	0.60	0.07
Palmitic acid (C16:0)	1.96	0.37	18.9	0.54	0.08
Oleic acid (*cis-9* C18:1)	0.80	0.31	38.7	0.44	0.08
Rumenic acid (*cis-9 trans-11* C18:2)	0.04	0.02	50.0	0.45	0.08
Alpha-linolenic acid (C18:3*n-3*)	0.04	0.01	25.0	0.38	0.08
Traits expressed on a lactation basis					
Lactation somatic cell score (LSCS)	3.29	1.50	45.6	0.33 to 0.41[b]	0.07 to 0.09
CV milk	53.93	12.18	22.6	0.19 to 0.26[b]	0.07 to 0.09

SD standard deviation, *CV (%)* coefficient of variation expressed in percentage, *SE* standard error

[a] Milk proteins and fatty acids given in g per 100 mL

[b] Mean heritability of LSCS and CV milk estimated for each three-trait model

Table 3 Mean and maximum heritability, and genus representativeness levels over the operational taxonomic units (OTU)

Genus	N	Heritability			G1 (%)	G2 (%)
		Mean	SD	Max		
Prevotella	92	0.15	0.04	0.27	30*	22
Christensenellaceae_R-7_group	21	0.13	0.03	0.21	7	8
F082/unknown_genus	21	0.15	0.05	0.26	7*	4
Rikenellaceae_RC9_gut_group	21	0.17	0.05	0.29	7	5
Prevotellaceae_UCG-001	15	0.14	0.03	0.20	5*	3
Ruminococcus	11	0.15	0.06	0.26	4	6*
Muribaculaceae/unknown_genus	8	0.15	0.04	0.21	3	3
Acetitomaculum	7	0.15	0.04	0.22	2	2
Treponema	7	0.13	0.04	0.19	2	1
Ruminococcaceae/unknown_genus	6	0.14	0.02	0.17	2	1

N number of OTU grouped in each genus, *SD* standard deviation, *G1(%)* genus representativeness levels over OTU with a significant heritability, *G2(%)* genus representativeness levels over all OTU

*($P < 0.05$) Fisher's exact test

to −0.34 and from 0.35 to 0.99. Milk SFA shared many more correlated OTU than milk UFA, however C18:3*n-3* was the FA that correlated with the largest number of OTU (42 OTU).

Milk SFA (C4:0 to C16:0) correlated with 41 OTU, of which 26 OTU were positively correlated with at least one SFA. Five OTU were linked with all SFA, with variable genetic correlations, and a closely clustered group

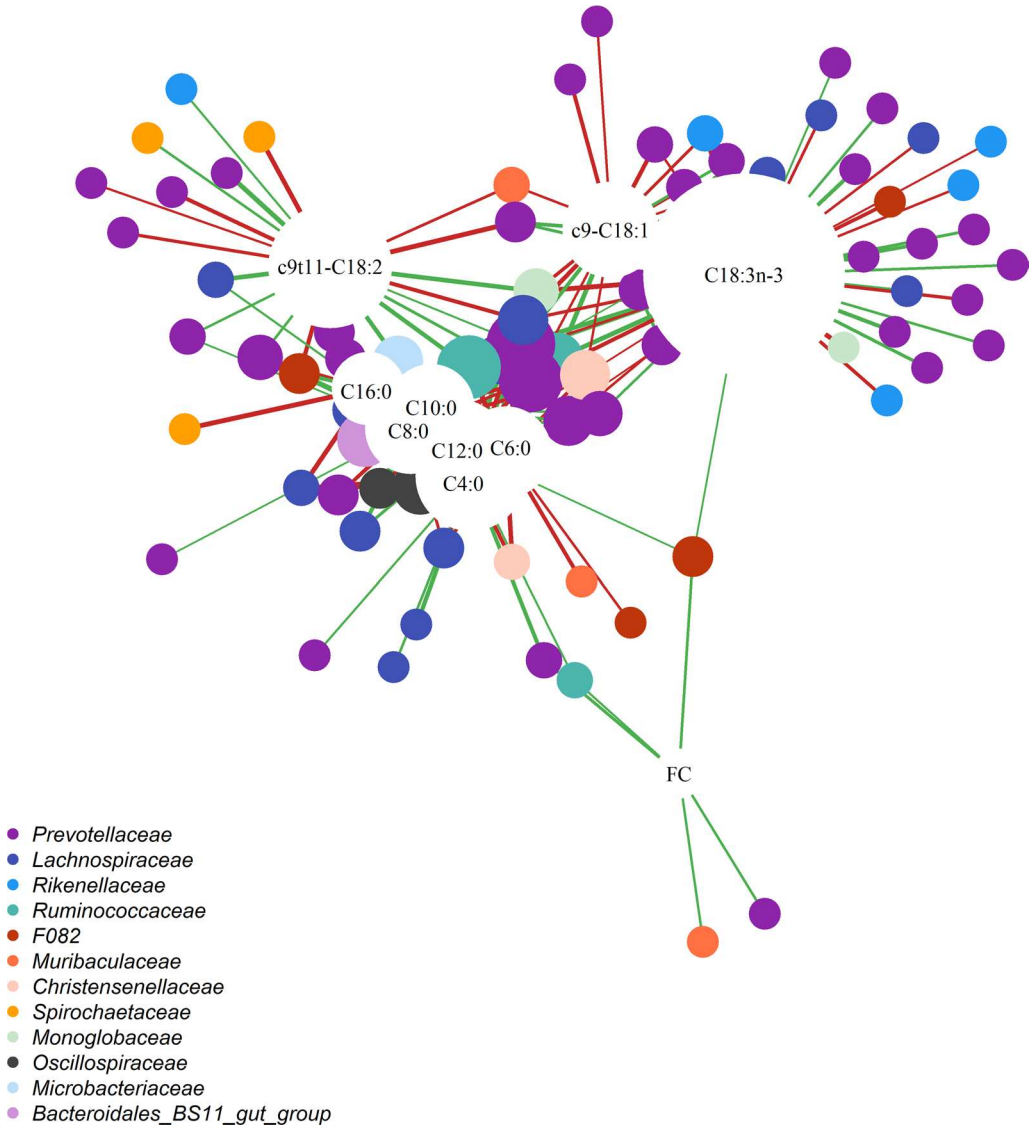


Fig. 1 Genetic correlation network between operational taxonomic units (OTU) and fat content (FC) and milk fatty acids, namely, butyric acid (C4:0), caproic acid (C6:0), caprylic acid (C8:0), capric acid (C10:0), lauric acid (C12:0), palmitic acid (C16:0), oleic acid (c9-C18:1), rumenic acid (c9t11-C18:2) and alpha-linolenic acid (C18:3n-3). The nodes represent OTU (in colors associated to the corresponding families where rare families with one OTU are not represented) and dairy traits (in white). The diameter of the nodes is proportional to the number of genetic correlations, and the thickness and color of the edges represent values (−0.97 to −0.34 and 0.35 to 0.99) and signs (in green positive and in red negative ones) of the genetic correlation, respectively

consisting of C8:0, C10:0 and C12:0 shared positive correlations with 12 OTU from the *Prevotellaceae*, *Lachnospiraceae* and *Oscilllospiraceae* families (Fig. 1). Milk UFA (*cis-9* C18:1, *cis-9 trans-11* C18:2 and C18:3*n*-3) correlated with 62 OTU, half of which belonged to the *Prevotellaceae* and *Rikenellaceae* families. Two OTU from the *Prevotellaceae* family correlated with all the UFA but showed opposite correlation signs.

Genetic correlations with milk protein content and proteins

The daily protein content and milk proteins showed significant genetic correlations with OTU from 23 genera belonging to 17 families. The bacterial families that shared the largest number of significant genetic correlations were *Prevotellaceae* and *Lachnospiraceae* (Fig. 2). The significant genetic correlations ranged from −0.99 to −0.33 and from 0.36 to 0.98. Figure 2 shows two

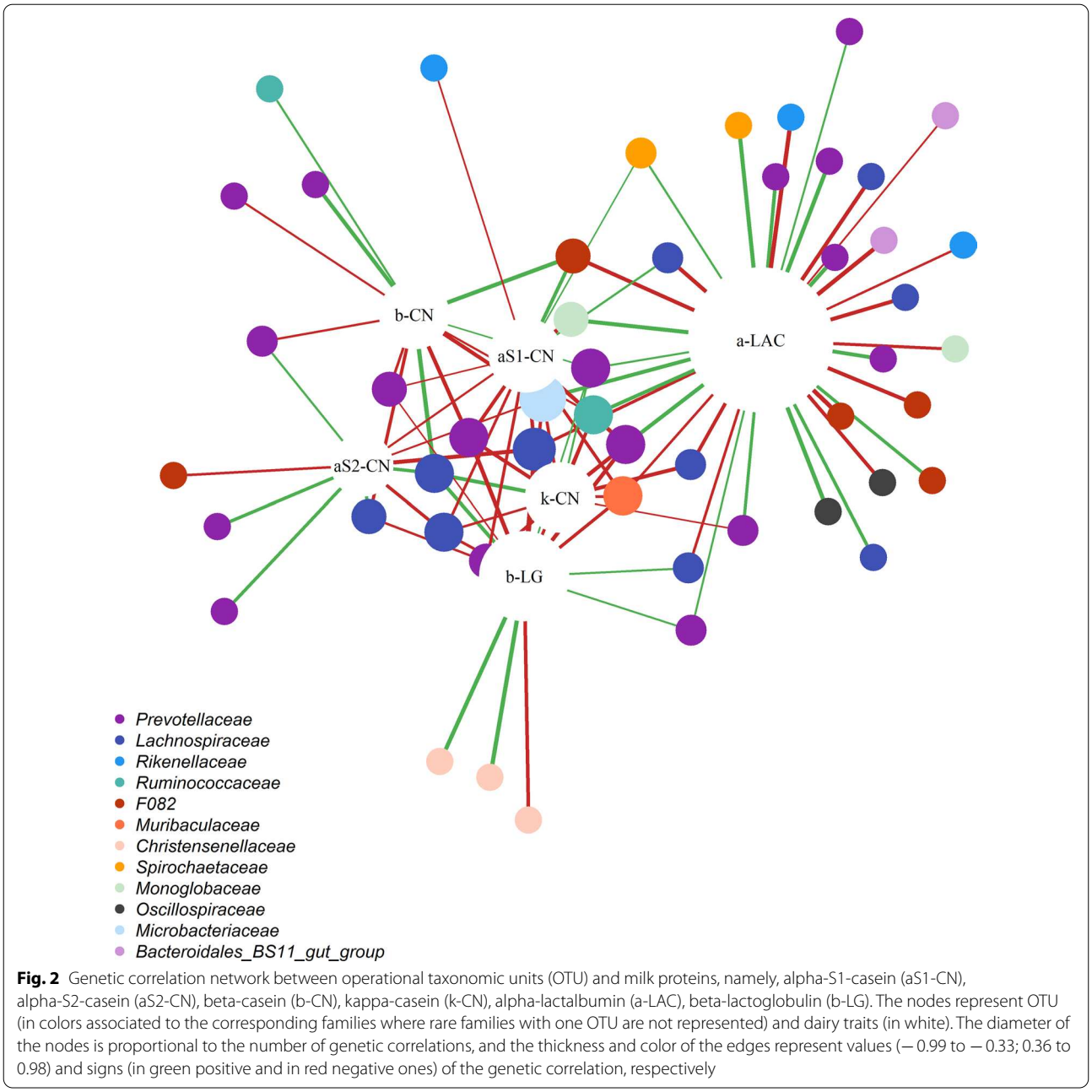


Fig. 2 Genetic correlation network between operational taxonomic units (OTU) and milk proteins, namely, alpha-S1-casein (aS1-CN), alpha-S2-casein (aS2-CN), beta-casein (b-CN), kappa-casein (k-CN), alpha-lactalbumin (a-LAC), beta-lactoglobulin (b-LG). The nodes represent OTU (in colors associated to the corresponding families where rare families with one OTU are not represented) and dairy traits (in white). The diameter of the nodes is proportional to the number of genetic correlations, and the thickness and color of the edges represent values (− 0.99 to − 0.33; 0.36 to 0.98) and signs (in green positive and in red negative ones) of the genetic correlation, respectively

groups of traits: one represented by alpha-lactalbumin, which correlated exclusively with 20 OTU, and a group of caseins (α_{s1}-CN, α_{s2}-CN, β-CN, and κ-CN) and beta-lactoglobulin, which correlated mostly negatively with OTU.

All caseins were correlated with 30 OTU, of which five were negatively correlated with three or more CN. *Leifsonia* OTU1479 belonging to the *Microbacteriaceae* family showed negative correlations with all CN, and four OTU from the *Lachnospiraceae* and *Prevotellaceae* families were also correlated with two CN (α_{s2}-CN and κ-CN). The whey proteins (α-lactalbumin and β-lactoglobulin) correlated with 47 OTU, of which eight were in common. Four OTU from the *Prevotellaceae*, *Lachnospiraceae* and *Muribaculaceae* families showed the same correlation sign with α-lactalbumin and β-lactoglobulin.

Table 4 Significant SNPs from the genome-wide association studies of dairy traits

Trait	SNP name	OAR	Position (bp)	Var (%)	$-\log_{10}$(P value)
α-Lactalbumin	rs426734075	1	147,595,299	0.06	4.79
α-Lactalbumin	rs400013895	2	25,181,840	0.11	4.41
α-Lactalbumin	rs421261402	2	58,918,403	0.23	4.63
C8:0	rs415371608	2	117,736,662	0.03	5.67
α-Lactalbumin	rs418593908	5	103,652,938	0.04	5.03
α_{S2}-CN	rs423428584	6	81,218,896	0.20	8.18
α-Lactalbumin	rs409523937	6	100,115,861	0.22	5.94
α-Lactalbumin	rs399440927	11	32,011,837	0.46	4.43
α-Lactalbumin	**rs429602859**	11	32,652,804	1.05	5.12
α-Lactalbumin	**rs412766461**	11	32,783,984	1.03	4.89
α-Lactalbumin	**rs410865757**	11	33,302,950	1.97	5.40
α-Lactalbumin	**rs402411249**	11	33,379,161	1.75	5.77
α-Lactalbumin	**rs401296484**	11	33,733,902	0.97	5.51
α-Lactalbumin	**rs425950097**	11	34,091,412	0.98	4.43
α-Lactalbumin	**rs420633999**	11	34,146,282	0.88	5.38
α-Lactalbumin	**rs424294429**	11	34,239,169	0.58	4.53
Milk yield	rs402677421	15	7,078,006	0.14	5.82
C8:0	rs405420878	17	51,422,515	0.24	5.79
C6:0	rs405420878	17	51,422,515	0.27	5.60
C6:0	rs410355614	17	52,959,387	0.28	5.30
α-Lactalbumin	rs426375102	20	21,580,756	0.05	4.75
α-Lactalbumin	rs409395180	24	6,819,203	0.07	4.39

SNPs that are located in a quantitative trait locus region are in bold characters

OAR Ovis aries autosome, *Var (%)* phenotypic variance explained for a window of 20 adjacent SNPs

Core microbiome

The core microbiome was represented by 275 OTU, which represent 13% of all OTU, belonging to the most abundant genera, such as *Prevotella*, *Christensenellaceae_R-7_group* and *Ruminococcus*. In the core microbiome, 44 OTU showed a significant heritability ($h^2 > 0.10$), which represented 14% of the heritable OTU, and 13 OTU were genetically correlated with dairy traits.

GWAS of dairy traits and rumen bacteria

GWAS of dairy traits

A GWAS was performed for each of the 18 dairy traits included in this study. The GWAS results are in Table 4, which shows that 22 significant SNPs (FDR < 0.10) were detected for milk yield, alpha-lactalbumin, alpha-S2-casein, caproic acid (C6:0) and caprylic acid (C8:0) concentrations in milk.

The largest number of significant SNPs was detected for alpha-lactalbumin on seven chromosomes, with a QTL on OAR11 (Fig. 3). This QTL on OAR11 includes eight significant SNPs that were located in the genomic region between 32.6 and 34.2 Mbp, with an LD score ranging from 0.24 to 0.52. The QTL region explained 2.5% of the phenotypic variance, with a lead SNP (rs402411249) showing the maximum $-\log_{10}$(P value) of 5.77. On OAR6, we detected one significant SNP and one suggestive SNP in the region between 100.1 and 101.3 Mbp with an LD score of 0.56.

For alpha-S2-casein, we detected a significant SNP (rs423428584) on OAR6, with the highest P value ($-\log_{10}$(P value) of 8.18) obtained in this study for all dairy traits.

For the SFA C6:0 and C8:0, we detected the same significant SNP (rs405420878) on OAR17 (Table 4). Another significant SNP (rs410355614) was detected for C6:0 and as a suggestive SNP for C4:0 with a $-\log_{10}$(P value) of 5.52 (Fig. 4). However, the LD score between these SNPs was lower than 0.10. For C8:0, a significant SNP (rs415371608) was found on OAR2, which was also identified as a suggestive SNP for C10:0 with a $-\log_{10}$(P value) of 5.25.

For milk yield, we detected a significant SNP on OAR15, but we did not identify any signal on OAR3, which included the SNP corresponding to the mutation in the *SOCS2* gene.

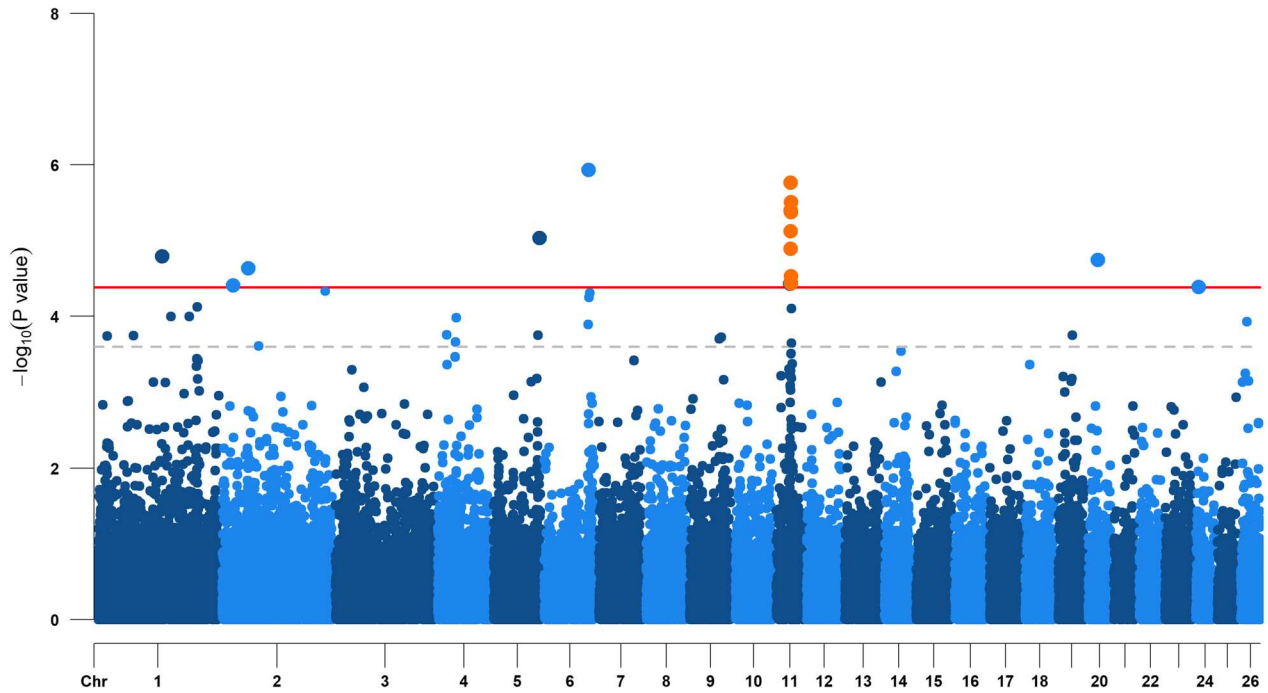


Fig. 3 Manhattan plot for the genome-wide association study of alpha-lactalbumin. The horizontal red line and grey dashed line show the false discovery rate thresholds of 0.10 and 0.30, respectively. Orange dots indicate SNPs in the QTL region

GWAS of rumen bacteria

A GWAS was performed for each of the 306 OTU abundances that had a heritability significantly different from 0. We detected 94 significant SNPs (FDR < 0.10) that were distributed across 22 chromosomes for 56 OTU (see Additional file 5: Table S5). For the 94 SNPs detected, we recovered 42 potential candidate genes that are involved in various GO biological processes and KEGG pathways (see Additional file 6: Table S6). In addition, as shown in Table 5, six QTL were identified on OAR3, 5, 10 and 11.

The group of 56 OTU with significant SNPs belonged to three phyla, 11 families, and 23 genera. Expressed as a percentage of OTU, the *Prevotellaceae* family was the most represented, with 23 OTU (41%) of which 19 were *Prevotella* OTU, followed by *Lachnospiraceae* (20%) and *Ruminococcaceae* (9%).

For the 94 significant SNPs detected in the host genome and associated with rumen bacterial abundance, several genomic regions showed QTL and significant SNPs for OTU abundance and dairy traits (Table 5).

OAR3

On OAR3, we detected a QTL for *Prevotella* OTU196 that includes two significant SNPs in high LD (LD score of 0.67). The SNP rs419358934 presented a colocalized signal with *Prevotellaceae_UCG-001* OTU906, and the allele substitution effect for that SNP varied in the same direction for both *Prevotellaceae* OTU. In addition, for each OTU, we detected two other significant SNPs that were located outside of the QTL for which two candidate genes were identified (Table 5): (1) the *cytotoxic granule-associated RNA binding protein* (*TIA1*) gene that is involved in the negative regulation of cytokines (GO:0001818), negative regulation of translation (GO:0017148) and regulation of mRNA splicing (GO:0048024), and (2) the *interleukin 18 receptor* (*IL18R1*) gene that has a role in natural killer cell activation (GO:0030101), positive regulation of interferon-gamma production (GO:0032729) and cell signalling and inflammatory processes (KEGG: oas04060).

OAR5

On OAR5, we detected three QTL for *Prevotella* OTU1843, *Rikenellaceae_RC9_gut_group* OTU546 and *Ruminococcus* OTU1191. The QTL for OTU1843 includes two significant SNPs, of which one (rs402307868) showed a colocalized signal with the *Possible_genus_sk018* OTU1631. The allele substitution effects for the SNP with colocalized signals showed contrasting signs for the *Prevotella* and *Possible_genus_sk018* OTU. The QTL for *Rikenellaceae_RC9_gut_group*

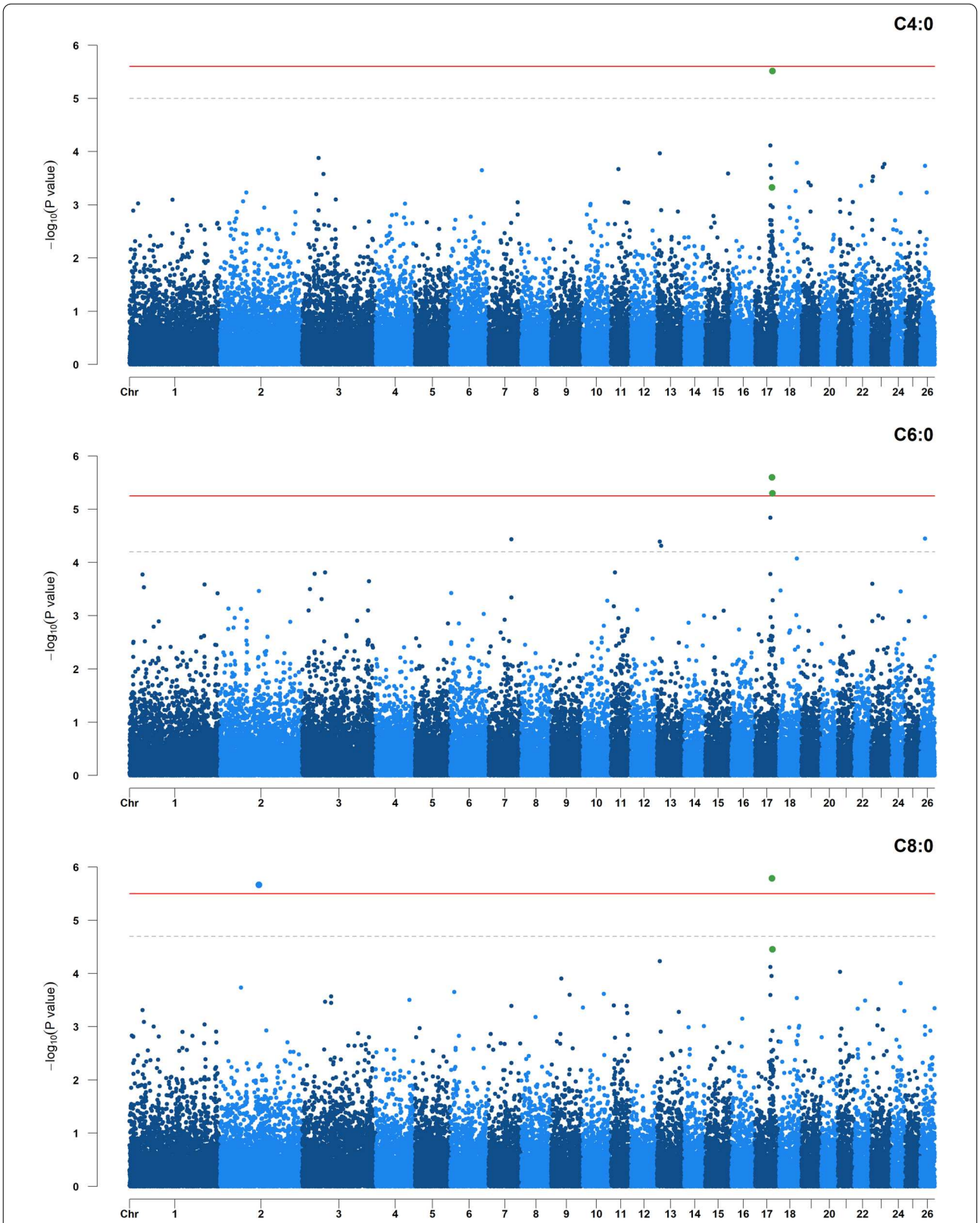


Fig. 4 Manhattan plots for the genome-wide association studies of butyric acid (C4:0), caproic acid (C6:0) and caprylic acid (C8:0). The horizontal red line and grey dashed line show the false discovery rate thresholds of 0.10 and 0.30, respectively. Green dots indicate SNPs in common between traits

Table 5 Significant SNPs from the genome-wide association studies of the rumen bacteria on chromosomes that colocalized with dairy traits

Genus	OTU name	SNP name	OAR	Position (bp)	Var (%)	$-\log_{10}$ (P value)	Gene name
Lachnospiraceae_NK4A136_group	OTU2186	rs422351583	3	5,668,047	0.13	6.09	
[Eubacterium]_coprostanoligenes_group /uG	OTU2438	rs420593841	3	8,693,890	0.05	5.37	*GARNL3*
[Eubacterium]_ventriosum_group	OTU355	rs415665472	3	27,430,018	0.14	5.95	
Prevotella	OTU196	rs399829503	3	38,048,434	0.09	5.09	*TIA1*
Clostridia UCG-014/uF/uG	OTU2515	rs417512713	3	82,810,037	0.14	5.79	
Prevotella	OTU196	rs407226710	3	87,932,154	0.37	5.07	
Prevotellaceae_UCG-001	OTU906	rs419358934[a]	3	89,802,120	0.33	6.24	
Prevotella	OTU196	**rs419358934**[a]	3	89,802,120	0.18	5.32	
Prevotella	OTU196	**rs430243701**	3	89,808,523	0.14	6.78	
Prevotellaceae_UCG-001	OTU906	rs424849255	3	99,246,368	0.09	5.25	*IL18R1*
Prevotellaceae_UCG-001	OTU906	rs404566475	3	112,090,102	0.23	5.92	
Syntrophococcus	OTU509	rs409660427	3	138,657,389	0.14	5.27	*ENDOU*
Prevotella	OTU427	rs402872469	3	221,603,160	0.16	5.40	
Possible_genus_Sk018	OTU1631	rs402307868[a]	5	31,467,389	0.03	5.76	
Prevotella	OTU1843	**rs402307868**[a]	5	31,467,389	0.11	5.11	
Prevotella	OTU1843	**rs413547561**	5	31,477,570	0.10	5.99	
Fibrobacter	OTU1335	rs405557393	5	59,933,921	0.06	6.10	*ANXA6*
Rikenellaceae_RC9_gut_group	OTU546	**rs410315179**	5	72,275,014	0.38	6.75	
Rikenellaceae_RC9_gut_group	OTU546	**rs429334236**	5	72,324,510	0.24	5.81	
Fibrobacter	OTU1335	rs398177414	5	79,732,944	0.18	6.56	
Ruminococcus	OTU1191	**rs411645323**	5	83,683,348	0.20	5.30	
Ruminococcus	OTU1191	**rs418870684**	5	83,724,689	0.12	5.29	ENSOARG00000025325
Lachnospiraceae_UCG-008	OTU655	rs403619685	6	28,689,782	0.08	5.92	
Lachnospiraceae_XPB1014_group	OTU947	rs400970883	6	36,655,091	0.05	6.20	*SPP1*
Prevotella	OTU1843	rs416815759	6	45,906,622	1.43	5.59	*TBC1D19*
Prevotella	OTU440	rs406261149	6	50,058,344	0.20	5.30	
Acetitomaculum	OTU612	rs430647780	6	75,425,021	0.20	5.58	
Acetitomaculum	OTU612	rs415319007	6	85,637,777	0.29	5.51	
Prevotella	OTU399	rs428296445	10	16,139,981	0.33	5.57	*W5PCY9_SHEEP*
Clostridia UCG-014/uF/uG	OTU1386	**rs423990418**	10	16,216,118	0.21	6.28	*W5PCZ6_SHEEP*
Clostridia UCG-014/uF/uG	OTU1386	**rs404163943**	10	16,222,665	0.10	6.28	*W5PCZ6_SHEEP*
Prevotella	OTU13	rs408448649	10	46,573,078	0.25	5.58	*DACH1*
Muribaculaceae/uG	OTU803	rs403822645	10	80,687,261	0.08	4.87	ENSOARG00000026327
Ruminococcaceae/uG	OTU155	rs408532009	10	84,678,999	0.35	6.41	*ANKRD10*
Ruminococcaceae/uG	OTU501	rs422890024	11	7,826,874	0.04	5.40	*MSI2*
Christensenellaceae_R-7_group	OTU596	rs414408260	11	58,151,039	0.14	5.63	
CAG-352	OTU304	**rs421085019**	11	61,946,606	0.23	5.48	
CAG-352	OTU304	**rs429659626**	11	62,034,261	0.04	5.48	*HELZ*
Prevotella	OTU943	rs412819070	15	5,606,060	0.13	5.71	*MMP20*
Prevotella	OTU423	rs409362751[a]	15	37,553,506	0.18	6.97	*PDE3B*
Prevotella	OTU80	rs409362751[a]	15	37,553,506	0.16	5.84	*PDE3B*
Prevotella	OTU399	rs409362751[a]	15	37,553,506	0.19	5.52	*PDE3B*
Prevotella	OTU399	rs426932216	15	44,445,054	0.14	5.13	*LOC101105776*
Blautia	OTU1025	rs425070593	15	80,043,367	0.20	5.34	*TCN1*
F082/uG	OTU167	rs412496804[a]	26	28,316,180	0.31	6.06	
Prevotella	OTU1336	rs412496804[a]	26	28,316,180	0.20	5.98	
Prevotellaceae_UCG-001	OTU906	rs412496804[a]	26	28,316,180	0.20	5.20	

SNPs that are in a quantitative trait locus region are in bold characters

OAR Ovis aries chromosome, *Var (%)* phenotypic variance explained for a window of 20 adjacent SNPs, *uF* unknown family, *uG* unknown genus

[a] SNP shared between different OTU

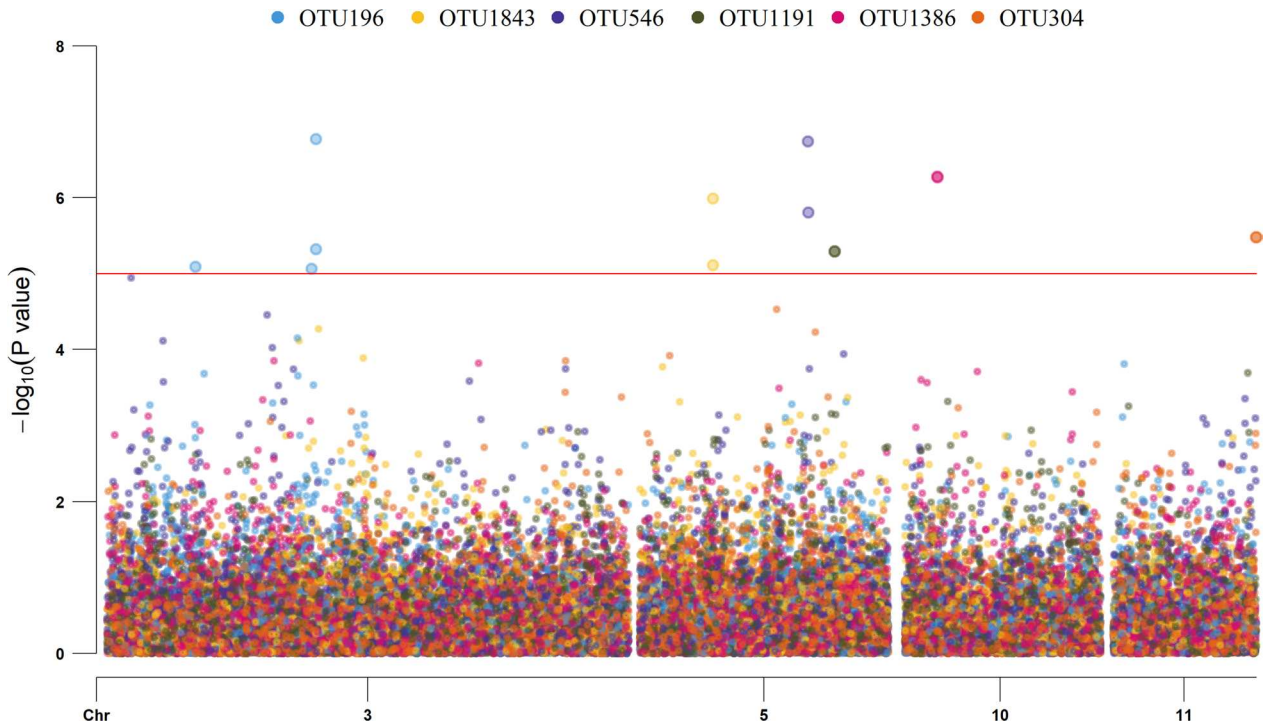


Fig. 5 Manhattan plots for the genome-wide association studies of rumen bacterial abundance showing quantitative trait loci (QTL) on chromosomes 3, 5, 10 and 11. The horizontal red line shows the false discovery rate threshold of 0.10. Operational taxonomic units (OTU): *Prevotella* OTU196, *Prevotella* OTU1843, *Rikenellaceae_RC9_gut_group* OTU546, *Ruminococcus* OTU1191, *Clostridia UCG-014*/unknown family/unknown genus OTU1386, and *CAG-352* OTU304

OTU546 includes two significant SNPs in high LD (LD score of 0.74). The QTL for *Ruminococcus* OTU1191 includes two significant SNPs separated by 0.04 Mbp, but the LD score was very low in that region (< 0.10). The SNP rs418870684 was associated with the novel gene *ENSOARG00000025325*.

OAR6

On OAR6, we detected six significant SNPs for five OTU but no QTL. In the region where a significant SNP was detected for alpha-S2-casein (81.2 Mbp), we identified two significant SNPs for *Acetitomaculum* OTU612, however, the LD score was low (< 0.20), and no shared SNPs were detected for these two traits.

OAR10

On OAR10, we detected a QTL with two significant SNPs at 16.2 Mbp for OTU1386 belonging to the order *Clostridia UCG-014*. The two significant SNPs were located in a region in high LD (LD score of 0.78), and the same gene, *E3 ubiquitin-protein ligase* (*W5PCZ6_SHEEP*), was detected. The *W5PCZ6_SHEEP* gene is involved in the ubiquitin-dependent protein catabolic process (GO:0006511), multicellular organism development (GO:0007275) and regulation of protein stability (GO:0031647).

OAR11

On OAR11, we detected a QTL with two significant SNPs in complete LD for *CAG-352* OTU304 belonging to the *Ruminococcaceae* family. However, this QTL for OTU304 was located far from the QTL detected for alpha-lactalbumin on OAR11 (32.6 to 34.2 Mbp). The candidate gene *helicase with zinc finger* (*HELZ*), that was located in this QTL, was not associated with any GO biological process or KEGG pathway.

OAR15

On OAR15, we detected the same significant SNP, rs409362751, for three *Prevotella* OTU (Table 5), and the allele substitution effects varied in the same way for the three *Prevotella* OTU. For this SNP, we identified the candidate gene *phosphodiesterase 3B* (*PDE3B*), which is involved in the negative regulation of cell adhesion (GO:0007162), negative regulation of lipid

catabolic process (GO:0050995), and metabolic pathways (KEGG: oas01100).

OAR26

On OAR26, we detected the same significant SNP, rs412469804, for *Prevotella* OTU1336, *Prevotellaceae_UCG-001* OTU906 and OTU167 that belongs to the *F082* family. The allele substitution effects for the two *Prevotellaceae* OTU varied in the same direction, but in opposite directions for *F082* OTU167.

As presented above, four SNPs on OAR3, 5, 15 and 26 simultaneously affected the rumen abundance of OTU belonging to the *Prevotellaceae*, *F082*, and *Lachnospiraceae* families (Fig. 5). However, no shared genomic regions were detected with the 18 dairy traits included in our study.

Discussion

In this study, we hypothesised that, at least in part, the rumen microbiota abundance of Lacaune dairy ewes is under the control of the host genome and genetically linked with dairy traits through shared genomic regions coding for common metabolic pathways.

Heritability of rumen bacteria and dairy traits

Our results obtained in Lacaune dairy ewes demonstrate that rumen bacterial abundance is partially controlled by the host genetics. The OTU heritabilities showed an asymmetric distribution with most values close to 0, but with 15% of the OTU (306 of 2059 OTU) having a low to moderate heritability (0.10 to 0.29), which is similar to the results reported for rumen bacteria in other ruminants [2, 4, 6, 42, 43]. However, there are some differences in the heritability range and in the percentage of OTU with a significant heritability. For example, Difford et al. [2] and Zhang et al. [6], using the same dataset of 750 dairy cows, obtained significant heritability estimates for 5 to 10% of genera, with heritabilities ranging from 0.17 to 0.25, and 2 to 6% of OTU with heritabilities ranging from 0.16 to 0.44. These results, which were based on 16S rRNA gene sequencing data, are similar to the range of heritabilities that we found for OTU, but the authors obtained a lower percentage of OTU with a significant heritability. It is also relevant to highlight the differences that were observed by working at different taxonomic levels, i.e., these authors reported lower heritabilities at the genus level. However, Martínez-Álvaro et al. [43], using untargeted "shotgun" sequencing, obtained significant heritabilities for 16% of genera and a range of higher heritabilities, with a maximum heritability of 0.60.

The group of 306 heritable OTU belonged to the most abundant phyla, such as *Bacteroidota* and *Firmicutes*, as reported by Li et al. [4], Wallace et al. [42] and Zhang et al. [6]. Furthermore, in terms of genus representation, *Prevotella*, which was overrepresented among the OTU with a significant heritability and showed an average estimate of 0.15 ± 0.04, is considered as a highly heritable genus in the literature [4, 42]. The main differences in terms of affiliation of the heritable rumen bacteria are with the results presented by Martínez-Álvaro et al. [43], with *Proteobacteria*, *Actinobacteria* and *Firmicutes* being the main represented phyla, but these differences could be due to the "shotgun" sequencing technique that reveals all the microbial genomes within a sample.

In general, the differences in gut bacteria observed between studies can be due to several factors: (1) the sequencing technology used, which does not give enough information to obtain affiliations at lower taxonomic levels (e.g., species), meaning that most studies report results at the genus level [2, 4, 6, 42] and rarely at the OTU level as we have in our study; (2) bioinformatic processing of DNA sequences and the subsequent analyses that have an impact on the total number of OTU to be analyzed and their taxonomic affiliations (e.g., Difford et al. [2] and Li et al. [4] amplified the V1–V3 region of 16S rRNA gene, meaning that some of their OTU were not obtained in our study because we amplified the V3–V4 region); (3) the number of individuals included in the analyses, which usually affects the accuracies of heritability estimates and explains the differences with studies that analyze fewer animals [43]; and (4) the animal species (bovine vs. sheep) investigated, which have differential bacterial abundances, as evidenced by Henderson et al. [44]. Notably, and in spite of the differences observed with other studies in ruminants, we report the highest percentage of bacterial OTU that are genetically controlled by the host.

However, a core microbiome cannot be declared based on these heritable OTU only, as was done by Wallace et al. [42], because we obtained only 44 of the 306 OTU present in most animals, and the same percentage of OTU with a significant heritability was observed in the core and total microbiotas. This result is expected because the core microbiome consists of microbes that are stable between animals in the same environment, which implies potential horizontal transmission of microbes between them. Therefore, our finding that only 14% of the host-genetically controlled OTU were part of the core microbiome means that most microbes come repeatedly from the environment and much less by vertical transmission from their parents.

To the best of our knowledge, our study is the first to estimate the genetic parameters of rumen bacteria from approximately 800 dairy ewes from the same farm under the same housing conditions.

For daily milk traits, we obtained moderate to high heritabilities, with values of 0.28 ± 0.06 for MY, 0.57 ± 0.06 for PC and 0.59 ± 0.06 for FC. Compared to two previous studies [7, 8] that include larger populations of Lacaune ewes in their first lactation, our estimates were similar to those of Rupp et al. [8] for MY (0.28) and PC (0.51) but higher for FC (0.41) on an annual basis, and to those of Boichard et al. [7] (PC 0.39 and FC 0.29) on a daily basis.

For the milk composition traits, milk FA presented moderate to high heritabilities, as reported by Boichard et al. [7] in Lacaune ewes. In addition, SFA had higher heritabilities than UFA, as reported by Boichard et al. [7] and by Buitenhuis et al. [3] who measured milk FA by accepted wet-lab methodologies in dairy cows. Milk proteins showed moderate (0.36 ± 0.08 for alpha-lactalbumin) to high (0.68 ± 0.07 for alpha-S2-casein) heritabilities, and were slightly higher than those reported in Lacaune ewes [7]. The heritability estimate for alpha-lactalbumin was low, as in Boichard et al. [7], which may be due to the low accuracy of this trait based on MIR spectrum prediction, with an R^2 of 0.26 [20]. Thus, the results for alpha-lactalbumin should be interpreted with caution.

Genetic correlations between rumen bacteria and dairy traits

Rumen bacteria are undeniably of crucial importance to their host, providing FA and microbial amino acids that contribute to milk fat and protein production [45]. However, only 96 heritable OTU from the most abundant families *Prevotellaceae*, *Lachnospiraceae* and *Ruminococcaceae* were moderately to highly genetically correlated (absolute values ranging from 0.33 to 0.99) with one or more milk FA and proteins. The overall quantification of these correlations is of interest to understand the relationship between rumen microbiota and phenotypes, in spite of the limitations of our study due to the low precision of 16S rRNA gene sequencing at the species level and the large number of estimated genetic correlations (more than 5500) based on a small dataset of 800 ewes, which, even if those with large standard errors are discarded, may lead to some spurious correlations.

The impact of the heritable OTU on the dairy traits was weak, as evidenced by the small number of significant genetic correlations obtained in our study (on average 301), i.e., only 5% of the total estimates. This figure was even smaller when we evaluated the number of OTU that were significantly associated with dairy traits in the core microbiome since only 13 of 96 OTU were genetically linked. However, interestingly, some of these OTU might have host-relevant functions since they are part of the core microbiome. Although the number of genetic correlations between rumen bacteria and milk traits was small, their values were moderate to high, which suggests that it could be useful to include microbiota composition in the genetic models, when the objective is to determine how the genetic effect impacts the phenotype not directly but with an indirect effect through microbiota abundance. In this regard, some authors have proposed methodologies to account for this indirect genetic effect mediated by the microbiota [46–48].

Our results show that most of the heritable OTU do not impact dairy traits. For this reason, the genetic associations that are observed mainly with milk FA and proteins can be difficult to use as selection criterion to improve fine milk composition traits, but this does not limit their inclusion in genetic models as an additional source of information to estimate the indirect genetic effect on those phenotypes of interest.

GWAS of dairy traits and rumen bacteria

After having demonstrated that a group of heritable OTU was genetically linked to milk composition traits, we performed GWAS to identify host genomic regions that simultaneously affect rumen bacterial abundance and dairy traits. In the GWAS of dairy traits, we detected 22 significant SNPs that were distributed across nine chromosomes. Among these regions, two on OAR6 showed associations with alpha-lactalbumin and alpha-S2-casein. The SNP rs423428574 detected for alpha-S2-casein was located in the region of OAR6 where the major gene *CSN1S2* was reported in the literature on caseins [10, 11, 49]. In addition, in Lacaune sheep, Boichard et al. [7] detected a QTL on OAR6 for alpha-S2-casein, but without reporting a position. We have found no publications of QTL detected on OAR6 for alpha-lactalbumin in dairy ewes. On OAR11, we detected a QTL for alpha-lactalbumin with eight significant SNPs (Fig. 3), where a significant region for caseins and beta-lactoglobulin was also reported [7]. Although the major gene *LALBA*, which encodes alpha-lactalbumin in sheep, was identified on OAR3 [11], we found no significant SNPs for this whey protein in our study. Significant SNPs detected for alpha-lactalbumin on OAR2 and 20 were close to those that García-Gámez et al. [11] reported to be associated with protein and fat yields for dairy sheep. For the milk SFA C6:0 and C8:0, we detected the same SNP on OAR17, where one suggestive SNP was also identified for C4:0. García-Gámez et al. [11] reported a significant SNP (58.8 Mbp) on OAR17 that was associated with fat percentage, and Carta et al. [9] detected a QTL in the backcross population of Lacaune sheep for C6:0 and C8:0 on OAR17 and one additional signal that we did not detect on OAR8. In addition, although we did not detect SNPs for milk FA on OAR11, Marina et al. [10] reported major

genes for milk fat synthesis (*ACACA* and *FASN*) on this chromosome.

In the GWAS of rumen bacterial abundances, we detected signals that were distributed across the host genome, as reported by Li et al. [4], Abbas et al. [50] and Zhang et al. [6]. However, we detected three main regions on OAR3, 15 and 26 with signals for *Prevotellaceae* OTU that colocalized with the same SNP, i.e., colocalizations for *Prevotellaceae* and *Lachnospiraceae* OTU on OAR5, and colocalizations for *Prevotellaceae* and *F082* OTU on OAR26. This suggests that some regions of the host genome are associated with rumen bacterial abundance, but given the limitations raised by Pérez-Enciso et al. [51] about the difficulty in identifying causative SNPs for microbiota abundance, the results should be confirmed with a larger dataset. From the genomic regions in which QTL were detected for OTU abundances, we recovered several potential candidate genes, such as *TIA1*, *IL18R1*, *W5PCZ6_SHEEP* and *PDE3B*, which are involved in host immune system processes, regulation and catabolic processes, as reported by Abbas et al. [50]. However, it was not possible to identify shared metabolic pathways between OTU and dairy traits because we did not detect genomic regions where pleiotropic or closely-related QTL affected both traits simultaneously. One explanation may be that although the dataset is large for a microbial analysis, the number of records is too small for the detection of genome-associated regions and pleiotropic effects compared to other studies [52, 53].

Conclusions

Our findings, based on a large and unique microbiome database, demonstrate that a small proportion of the total bacterial abundance in the rumen of Lacaune ewes is partially controlled by the host genome, for both the total and core microbiome, and that a few of the bacteria are associated with specific genomic regions on OAR3, 5, 15 and 26. Only very few of the genetic correlations between milk fine composition and rumen bacteria were significant, and it was not possible to identify genomic regions and metabolic pathways shared between rumen bacterial abundance and dairy traits. The use of 16S rRNA gene sequencing in this study did not allow for species-level affiliation, and additional work is needed to identify which microbial functions are involved and to associate the functions with milk traits. Then, it will be possible to consider the incorporation of microbiota abundance and microbial functions into genetic evaluation models to account for indirect genetic effects through the microbiota.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12711-022-00769-9.

Additional file 1: Table S1. Operational taxonomic unit (OTU) sequences with corresponding taxonomy affiliation. Consensus sequences of the OTU with corresponding affiliation from Kingdom to species.

Additional file 2: Table S2. Heritability (h^2) and genetic correlations (rg) between lactation somatic cell score (LSCS), coefficient of variation in milk production (CV Milk), and operational taxonomic units (OTU). Heritability and genetic correlations for all 2059 OTU abundances with lactation somatic cell score and the coefficient of variation in milk production.

Additional file 3: Table S3. Genetic correlations between fat content (FC), milk fatty acids and operational taxonomic units (OTU). Genetic correlations between OTU abundances, fat content and milk fatty acids, such as butyric acid (C4:0), caproic acid (C6:0), caprylic acid (C8:0), capric acid (C10:0), lauric acid (C12:0), and palmitic acid (C16:0), oleic acid (*cis-9* C18:1), rumenic acid (*cis-9 trans-11* C18:2) and alpha-linolenic acid (C18:3*n-3*).

Additional file 4: Table S4. Genetic correlations between protein content (PC), milk proteins and operational taxonomic units (OTU). Genetic correlations between OTU abundances, protein content and milk proteins, such as alpha-S1-casein, alpha-S2-casein, beta-casein, kappa-casein, alpha-lactalbumin and beta-lactoglobulin.

Additional file 5: Table S5. Significant SNPs from GWAS of operational taxonomic units (OTU). Significant SNPs from GWAS of operational taxonomic units with candidate genes detected for each SNP.

Additional file 6: Table S6. Functional annotation for gene IDs recovered from detected SNPs for microbiome GWAS. Description of the functional annotations for genes recovered from Ensembl platform and metabolic processes and KEGG pathways from DAVID platform.

Acknowledgements

We are grateful to the technical staff at INRAE experimental unit of La Fage (DOI: https://doi.org/10.15454/1.548325523466425E12) for animal care and help with rumen sampling. We thank Béatrice Gabinaud for preparing microbiota samples and Get-Plage platform in Toulouse for DNA sequencing, and Hélène Larroque and Rachel Rupp for access to the genetic resources produced at INRAE La Fage.

Author contributions

GMB performed all analyses and wrote the manuscript. AM and CME designed and supervised the experimental scheme, managed farm sampling and data collection. CME carried out the genotyping editing and quality control. AM, CME and AJB discussed results and improved the writing. All authors read and approved the final manuscript.

Funding

The financial support from INRAE and CRB-Anim (DOI: https://doi.org/10.15454/1.56137856228273 78E12) for this project is acknowledged, as well as the co-financing of GMB's doctoral thesis by INRAE and the Occitanie Region. The funders had no role in study design and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

The dataset supporting the conclusions of this article is available in the NCBI repository, https://www.ncbi.nlm.nih.gov/sra/PRJNA723543.

Declarations

Ethics approval and consent to participate

The protocol received approval from the French Ministry of Higher Education, Research and Innovation—Animal Ethics Committee with the following approval number: APAFIS#6292-2016080214271984 v8.

Consent for publication

Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]GenPhySE, INRAE, ENVT, Université de Toulouse, 24 Chemin de Borde Rouge, 31326 Castanet-Tolosan, France. [2]Center for Quantitative Genetics and Genomics, Aarhus University, Blichers Allé 20, 8830 Foulum, Denmark.

Received: 3 May 2022 Accepted: 9 November 2022
Published online: 24 November 2022

References

1. Stewart CS, Flint HJ, Bryant MP. The rumen bacteria. In: Hobson PN, Stewart CS, editors. The rumen microbial ecosystem. London: Chapman & Hall; 1997. p. 10–72.
2. Difford GF, Plichta DR, Løvendahl P, Lassen J, Noel SJ, Højberg O, et al. Host genetics and the rumen microbiome jointly associate with methane emissions in dairy cows. PLoS Genet. 2018;14:e1007580.
3. Buitenhuis B, Lassen J, Noel SJ, Plichta DR, Sørensen P, Difford GF, et al. Impact of the rumen microbiome on milk fatty acid composition of Holstein cattle. Genet Sel Evol. 2019;51:23.
4. Li F, Li C, Chen Y, Liu J, Zhang C, Irving B, et al. Host genetics influence the rumen microbiota and heritable rumen microbial features associate with feed efficiency in cattle. Microbiome. 2019;7:92.
5. Martinez Boggio G, Meynadier A, Daunis-I-Estadella P, Marie-Etancelin C. Compositional analysis of ruminal bacteria from ewes selected for somatic cell score and milk persistency. PLoS One. 2021;16:e0254874.
6. Zhang Q, Difford G, Sahana G, Løvendahl P, Lassen J, Lund MS, et al. Bayesian modeling reveals host genetics associated with rumen microbiota jointly influence methane emission in dairy cows. ISME J. 2020;14:2019–33.
7. Boichard D, Govignon-Gion A, Larroque H, Maroteau C, Palhière I, Tosser-Klopp G, et al. Déterminisme génétique de la composition en acides gras et protéines du lait des ruminants, et potentialités de sélection. INRA Prod Anim. 2014;27:283–98.
8. Rupp R, Lagriffoul G, Astruc JM, Barillet F. Genetic parameters for milk somatic cell scores and relationships with production traits in French Lacaune dairy sheep. J Dairy Sci. 2003;86:1476–81.
9. Carta A, Casu S, Usai MG, Addis M, Fiori M, Fraghì A, et al. Investigating the genetic component of fatty acid content in sheep milk. Small Ruminant Res. 2008;79:22–8.
10. Marina H, Gutiérrez-Gil B, Esteban-Blanco C, Suárez-Vega A, Pelayo R, Arranz JJ. Analysis of whole genome resequencing datasets from a worldwide sample of sheep breeds to identify potential causal mutations influencing milk composition traits. Animals (Basel). 2020;10:1542.
11. García-Gámez E, Gutiérrez-Gil B, Sahana G, Sánchez JP, Bayón Y, Arranz JJ. GWA analysis for milk production traits in dairy sheep and genetic support for a QTN influencing milk protein percentage in the *LALBA* gene. PLoS One. 2012;7:e47782.
12. Rupp R, Senin P, Sarry J, Allain C, Tasca C, Ligat L, et al. A point mutation in suppressor of cytokine signalling 2 (Socs2) increases the susceptibility to inflammation of the mammary gland while associated with higher body weight and size and higher milk production in a sheep model. PLoS Genet. 2015;11:e1005629.
13. Liu Z, Lozupone C, Hamady M, Bushman FD, Knight R. Short pyrosequencing reads suffice for accurate microbial community analysis. Nucleic Acids Res. 2007;35:e120.
14. Andersson AF, Lindberg M, Jakobsson H, Bäckhed F, Nyrén P, Engstrand L. Comparative analysis of human gut microbiota by barcoded pyrosequencing. PLoS One. 2008;3:e2836.
15. Escudié F, Auer L, Bernard M, Mariadassou M, Cauquil L, Vidal K, et al. FROGS: find, rapidly, OTUs with galaxy solution. Bioinformatics. 2018;34:1287–94.
16. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, et al. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nat Methods. 2013;10:57–9.
17. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41:D590–6.
18. R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2021.
19. Neu AT, Allen EE, Roy K. Defining and quantifying the core microbiome: challenges and prospects. Proc Natl Acad Sci USA. 2021;118:e2104429118.
20. Ferrand M, Miranda G, Larroque H, Leray O, Guisnel S, Lahalle F, et al. Determination of protein composition in milk by mid-infrared spectrometry. In: Proceedings of the 38th ICAR annual meeting: 28 May 2012, Cork; 2012.
21. Ferrand-Calmels M, Palhière I, Brochard M, Leray O, Astruc JM, Aurel MR, et al. Prediction of fatty acid profiles in cow, ewe, and goat milk by mid-infrared spectrometry. J Dairy Sci. 2014;97:17–35.
22. Larroque H, Chassier M, Saintilan R, Astruc JM. Imputation accuracy from a low density SNP panel in 5 dairy sheep breeds in France. In: Proceedings of the 68th annual meeting of the European Association for Animal Production: 28 August–1 September 2017, Tallinn; 2017.
23. Oar_v3.1—Genome Assembly NCBI. https://www.ncbi.nlm.nih.gov/assembly/GCF_000298735.1/. Accessed 29 Mar 2022.
24. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. Front Microbiol. 2017;8:2224.
25. Quinn TP, Erb I, Richardson MF, Crowley TM. Understanding sequencing data as compositions: an outlook and review. Bioinformatics. 2018;34:2870–8.
26. Aitchison JA. The statistical analysis of compositional data. London: Chapman & Hall Ltd; 1986.
27. Martín-Fernández JA, Hron K, Templ M, Filzmoser P, Palarea-Albaladejo J. Bayesian-multiplicative treatment of count zeros in compositional data sets. Stat Modelling. 2015;15:134–58.
28. Palarea-Albaladejo J, Martín-Fernández JA. zCompositions—R package for multivariate imputation of left-censored data under a compositional approach. Chemom Intell Lab Syst. 2015;143:85–96.
29. van den Boogaart KG, Tolosana-Delgado R. "compositions": a unified R package to analyze compositional data. Comput Geosci. 2008;34:320–38.
30. Hofer A. Variance component estimation in animal breeding: a review. J Anim Breed Genet. 1998;115:247–65.
31. Misztal I, Tsuruta S, Strabel T, Auvray B, Druet T, Lee DH. BLUPF90 and related programs (BGF90). In: Proceedings of the 7th world congress on genetics applied to livestock production: 19–23 August 2002, Montpellier; 2002.
32. Aguilar I, Legarra A, Cardoso F, Masuda Y, Lourenco D, Misztal I. Frequentist p-values for large-scale-single step genome-wide association, with an application to birth weight in American Angus cattle. Genet Sel Evol. 2019;51:28.
33. Legarra A, Aguilar I, Misztal I. A relationship matrix including full pedigree and genomic information. J Dairy Sci. 2009;92:4656–63.
34. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23.
35. Vitezica ZG, Aguilar I, Misztal I, Legarra A. Bias in genomic predictions for populations under selection. Genet Res (Camb). 2011;93:357–66.
36. Christensen OF, Madsen P, Nielsen B, Ostersen T, Su G. Single-step methods for genomic evaluation in pigs. Animal. 2012;6:1565–71.
37. Strandén I, Garrick DJ. Technical note: derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J Dairy Sci. 2009;92:2971–5.
38. Gualdrón Duarte JL, Cantet RJ, Bates RO, Ernst CW, Raney NE, Steibel JP. Rapid screening for phenotype-genotype associations by linear transformations of genomic evaluations. BMC Bioinformatics. 2014;15:246.
39. Ensembl BioMart. https://www.ensembl.org/biomart/martview. Accessed 29 Mar 2022.
40. DAVID: functional annotation tools. https://david.ncifcrf.gov/tools.jsp. Accessed 29 Mar 2022.
41. KEGG PATHWAY database. https://www.genome.jp/kegg/pathway.html. Accessed 29 Mar 2022.
42. Wallace RJ, Sasson G, Garnsworthy PC, Tapio I, Gregson E, Bani P, et al. A heritable subset of the core rumen microbiome dictates dairy cow productivity and emissions. Sci Adv. 2019;5:eaav8391.
43. Martínez-Álvaro M, Auffret MD, Duthie CA, Dewhurst RJ, Cleveland MA, Watson M, et al. Bovine host genome acts on rumen microbiome function linked to methane emissions. Commun Biol. 2022;5:350.

44. Henderson G, Cox F, Ganesh S, Jonker A, Young W, Global Rumen Census Collaborators, et al. Rumen microbial community composition varies with diet and host, but a core microbiome is found across a wide geographical range. Sci Rep. 2015;5:14567.
45. Hungate R. The rumen and its microbes. New York: Acaemic Press; 1966.
46. Tiezzi F, Fix J, Schwab C, Shull C, Maltecca C. Gut microbiome mediates host genomic effects on phenotypes: a case study with fat deposition in pigs. Comput Struct Biotechnol J. 2020;19:530–44.
47. Weishaar R, Wellmann R, Camarinha-Silva A, Rodehutscord M, Bennewitz J. Selecting the hologenome to breed for an improved feed efficiency in pigs—a novel selection index. J Anim Breed Genet. 2020;137:14–22.
48. Christensen OF, Börner V, Varona L, Legarra A. Genetic evaluation including intermediate omics features. Genetics. 2021;219:iyab130.
49. Noce A, Pazzola M, Dettori ML, Amills M, Castelló A, Cecchinato A, et al. Variations at regulatory regions of the milk protein genes are associated with milk traits and coagulation properties in the Sarda sheep. Anim Genet. 2016;47:717–26.
50. Abbas W, Howard JT, Paz HA, Hales KE, Wells JE, Kuehn LA, et al. Influence of host genetics in shaping the rumen bacterial community in beef cattle. Sci Rep. 2020;10:15101.
51. Pérez-Enciso M, Zingaretti LM, Ramayo-Caldas Y, de los Campos G. Opportunities and limits of combining microbiome and genome data for complex trait prediction. Genet Sel Evol. 2021;53:65.
52. Jenko J, McClure MC, Matthews D, McClure J, Johnsson M, Gorjanc G, et al. Analysis of a large dataset reveals haplotypes carrying putatively recessive lethal and semi-lethal alleles with pleiotropic effects on economically important traits in beef cattle. Genet Sel Evol. 2019;51:9.
53. Bolormaa S, Pryce JE, Reverter A, Zhang Y, Barendse W, Kemper K, et al. A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle. PLoS Genet. 2014;10:e1004198.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



5.2.1 Supplementary results

Distribution of OTUs heritabilities

The distribution of heritabilities for each of 2,059 OTUs was positively skewed, with a median of 0.02 and a mean of 0.04 ± 0.03, as shown in Figure 14. The most part (85%) of the rumen bacterial OTUs showed a heritability lower than the threshold of significance (h^2 of 0.10). The 15% represented by 306 OTUs showed a heritability between 0.10 to 0.29 (mean of 0.15 ± 0.04), considering them OTUs with a heritability different from zero.

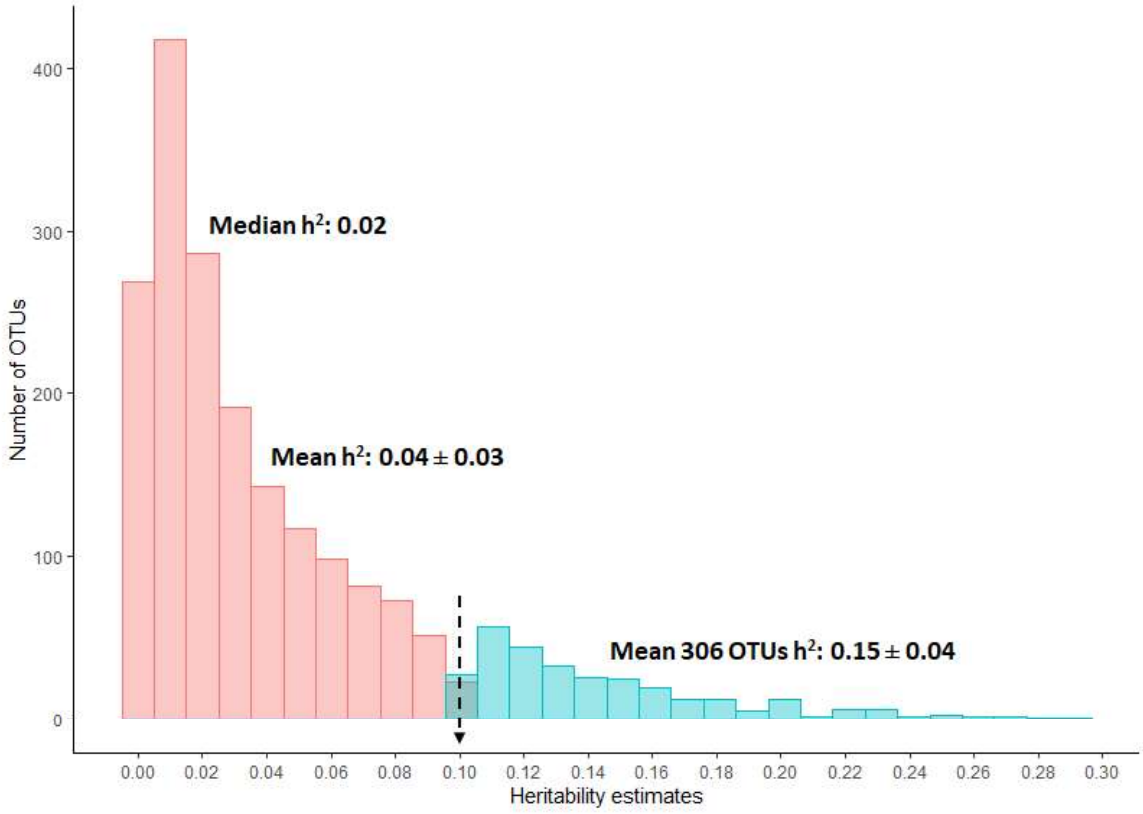


Figure 14. Frequency distribution of OTU heritabilities. In pink, OTUs with heritability non-significantly different from zero and in sky-blue, OTUs with significant heritability significantly different from zero (threshold of significance defined at 0.10).

Alpha-diversity heritability

As additional results of that presented in the chapter, heritability was estimated for the number of observed OTUs and three alpha-diversity indices, such as Inverse Simpson, Shannon and Simpson. The three alpha-diversity indices expressed the mean species diversity in a sample, but whatever the trait, heritabilities were zero (Table 6).

Table 6. Alpha-diversity indices and their heritabilities estimates.

	Observed OTUs	Inverse Simpson	Shannon	Simpson
Value ± SD	1322 ± 238	198 ± 73	6.17 ± 0.43	0.99 ± 0.04
h^2 ± SE	0.003 ± 0.002	0.003 ± 0.008	0.004 ± 0.001	0.003 ± 0.001

SD: standard deviation, h^2: heritability, SE: standard error.

Conclusion of section 5.2

A small proportion (15%) of bacterial abundances in the rumen of Lacaune dairy ewes is partially controlled by host genetics and some of the bacteria were associated with specific genomic regions on chromosomes 3, 5, 10 and 11. OTUs belonging to the core microbiome were not more heritable than other OTUs. In addition, only 96 OTUs were significantly genetically correlated with fine milk composition, and it was not possible to identify genomic regions and metabolic pathways shared between rumen bacterial abundance and dairy traits.

To conclude, we have demonstrated host genetic control over a fraction of rumen bacterial abundance, rather polygenic due to the weakness of the few QTL identified. A small part of these OTUs were genetically linked to fine milk composition traits. However, future work is needed to assess the relevance of incorporating microbiota data into genetic evaluation.

5.3 Evidence of genetic links between rumen bacteria and milk somatic cell score in Lacaune dairy sheep

12th World Congress on Genetics Applied to Livestock Production (WCGALP).

July 2022 - Rotterdam, The Netherlands.

Evidence of genetic links between rumen bacteria and milk somatic cell score in Lacaune dairy sheep

G. Martinez Boggio[1], A. Meynadier[1], R. Rupp[1], C. Allain[2], A.J. Buitenhuis[3] and C. Marie-Etancelin[1*]

[1] GenPhySE - INRAE – ENVT – Université de Toulouse, 24 Chemin de Borde Rouge, 31326, Castanet-Tolosan, France; [2] INRAE – Unité Expérimentale de La Fage, 12250, Saint-Jean et Saint-Paul, France; [3] Center for Quantitative Genetics and Genomics, Aarhus University, Blichers Allee 20, DK-8830 Tjele, Denmark; [*]christel.marie-etancelin@inrae.fr

Abstract

Mastitis is a high prevalence infectious disease in dairy sheep. Even if selection by milk somatic cell score (SCS) has been implemented to increase mastitis resistance, we wonder if rumen bacteria could influence udder inflammatory response. The objectives are to estimate genetic correlations between rumen bacteria and average SCS over lactation (LSCS), and to identify underlying mechanism between traits. 795 Lacaune ewes were studied using metabarcoding (16S rRNA gene) for rumen bacteria characterization and annual SCS records. Two-trait genomic models were used for the parameter estimation and GWAS. Results showed significant genetic correlations, and colocalized QTLs for LSCS and nine bacterial operational taxonomic units (OTUs) on chromosomes 3, 17, and 20. We found a genetic association between resistance/susceptibility to mastitis, using LSCS, and the ruminal bacteria abundance. These preliminary results pave the way for better understanding the biological basis and causalities of an intra-mammary inflammations in sheep.

Introduction

As in cattle, mastitis is one of the most important infectious diseases in small dairy ruminants, due to its high frequency and the price reduction paid to the farmer of up to 10% (95€ per 1,000 litres of ewes' milk). The selection was focused on the milk SCS, as a tool to decrease subclinical intra-mammary infections, which is the main problem in dairy sheep. The existence of a genetic basis for mastitis resistance has been demonstrated, and the underlying mechanisms involved such as the *suppressor of cytokine signalling 2* (Socs2) gene (Rupp *et al*., 2015). Recently, the influence of the rumen microbiome on subclinical ketosis has been demonstrated in cows (Gebreyesus *et al*., 2020), but no work is available in sheep. Therefore, could rumen bacteria impact on intra-mammary inflammations in sheep? To address this, and after quantifying host control over rumen bacteria, we 1) estimated genetic correlations between rumen bacteria and LSCS, and 2) identified possible underlying mechanisms with genome-wide association analysis (GWAS).

Materials & Methods

Animals phenotyping. A total of 795 Lacaune dairy ewes were reared at INRAE Experimental Unit of La Fage (UE 321 agreement A312031, Roquefort, France), between 2015 to 2019. All ewes (weighing 77 kg on average) raised indoors, and fed 93% meadow hay and silage plus 7% of concentrates (on dry matter basis). The daily somatic cell count (SCC) was quantified with a Fossomatic cell counter (Nanterre, France), as part of the official milk recordings of the flock. Then, the average LSCS was calculated as log of somatic cell count, including at least four milk test-days. Rumen sampling of adult ewes between 28 to 133 days in milk was done once per ewe using a vacuum pump and a medical gastric tube. The protocol received approval from the French Ministry of Higher Education, Research and Innovation – Animal Ethics Committee

(approval number: APAFIS#6292-2016080214271984 v8). The RNA sequences of the 795 samples were processed with the FROGS 3.0 pipeline, and OTUs with abundances below 0.005% were removed (Bokulich *et al*., 2013). The resulting abundance table included 2,059 OTUs, but we focused on the 306 OTUs with heritability significantly different from zero (empirical significant threshold of 0.10 obtained after running 10,000 analyses where OTUs abundancies were shuffling randomly across individuals) using an additive relationship matrix.

Animals genotyping. DNA extraction from blood samples and genotyping were performed for 795 Lacaune ewes. From those ewes, 743 were genotyped using a medium-density SNP chip (Illumina Ovine SNP50 BeadChip, 54,241 SNPs). The remaining 52 ewes were genotyped with a low-density SNP chip (Illumina Ovine SNP15, 16,681 SNPs) and imputed to a medium-density SNP chip genotypes in the framework of the Lacaune dairy sheep genomic selection program (Larroque *et al*., 2017). SNPs were removed using a call rate <0.99 for SNPs, and <0.95 for individuals, and minor allele frequency <0.05. In total, for the 26 *ovis aries* autosomal chromosomes (OAR), 35,496 SNPs remained for the analyses, plus one SNP corresponding to the Socs2 gene mutation (OAR3).

Parameter estimation. To estimate the heritability of the 306 rumen OTUs and LSCS, and the genetic correlations between each OTU and LSCS, a two-trait animal model was used:

$$\begin{bmatrix}\mathbf{y_1}\\ \mathbf{y_2}\end{bmatrix} = \begin{bmatrix}\mathbf{X_1} & \mathbf{0}\\ \mathbf{0} & \mathbf{X_2}\end{bmatrix}\begin{bmatrix}\mathbf{b_1}\\ \mathbf{b_2}\end{bmatrix} + \begin{bmatrix}\mathbf{Z_1} & \mathbf{0}\\ \mathbf{0} & \mathbf{Z_2}\end{bmatrix}\begin{bmatrix}\mathbf{a_1}\\ \mathbf{a_2}\end{bmatrix} + \begin{bmatrix}\mathbf{e_1}\\ \mathbf{e_2}\end{bmatrix} \quad (1)$$

where $\mathbf{y_1}$ and $\mathbf{y_2}$ are the vectors of observations for LSCS and one OTU at a time, respectively; $\mathbf{b_1}$ and $\mathbf{b_2}$ are the vectors of fixed effects, $\mathbf{b_1}$ including for LSCS: litter size, number of milking test-day, and $\mathbf{b_2}$ for the OTUs: sampling year, total number of DNA sequences per sample, and three effects nested on year sampling: lactation number, sequencing run, and order-time of sampling; $\mathbf{a_1}$ and $\mathbf{a_2}$ are the vectors of random additive effects; and $\mathbf{e_1}$ and $\mathbf{e_2}$ are the vectors of residual effects. $\mathbf{X_1}$ and $\mathbf{X_2}$ are incidence matrices relating fixed effects to vector $\mathbf{y_1}$ and $\mathbf{y_2}$, respectively; $\mathbf{Z_1}$ and $\mathbf{Z_2}$ are incidences matrices relating the additive effects to vector $\mathbf{y_1}$ and $\mathbf{y_2}$, respectively. The distributional assumptions are $\mathbf{a} \sim N(0, \mathbf{G} \otimes \mathbf{W})$, and $\mathbf{e} \sim N(0, \mathbf{I} \otimes \mathbf{R})$; $\mathbf{G}$ is a genomic relationships matrix as $\mathbf{G} = \mathbf{ZZ}'/2\sum p_i(1-p_i)$ (VanRaden, 2008); $\mathbf{I}$ is an identity matrix; $\mathbf{W}$ and $\mathbf{R}$ are the variance-covariance matrices for the random additive and residual effects, respectively. The heritability was estimated as $h^2 = \sigma_a^2/\sigma_y^2$, and the genetic correlations as $r_g = \sigma_{a12}/\sqrt{(\sigma_{a1}^2\sigma_{a2}^2)}$. The variance components were estimated using AIREMLF90 (Misztal *et al*., 2002). The significance was considered as genetic correlation value greater than the standard error.

Genome-Wide Association. The GWAS was performed using the two-trait model presented in Equation 1. After solving the single step GBLUP model, all SNP effects were estimated back solving the breeding values estimates as proposed by Aguilar *et al*. (2019) using POSTGSF90 (Misztal *et al*., 2002). To correct for multiple testing the false discovery rate (FDR) was used as implemented in the p.adjust package in R software. A SNP was considered significant at an FDR of P<0.10.

Results

The average heritability estimated for LSCS was 0.39 ± 0.07. Genetic parameters of the 14 rumen bacteria having significant genetic correlations with LSCS are presented in Table 1.

Table 1. Heritability (h^2) and genetic correlations (r_g) between lactation somatic cell score and ruminal bacteria abundance.

Genus	OTU name	h^2 (SE)	r_g (SE)
Lachnospiraceae_NK3A20_group	OTU506	0.10 (0.06)	0.68 (0.63)
Prevotella	OTU871	0.17 (0.07)	0.58 (0.43)
Monoglobus	OTU1066	0.13 (0.06)	0.51 (0.40)
Leifsonia	OTU1479	0.14 (0.07)	0.48 (0.43)
Lachnospiraceae/unknown genus	OTU309	0.15 (0.07)	0.47 (0.41)
Mailhella	OTU2331	0.15 (0.06)	0.39 (0.34)
F082/unknown genus	OTU1009	0.14 (0.06)	0.35 (0.29)
Acetitomaculum	OTU612	0.23 (0.07)	0.31 (0.21)
Lachnospiraceae NK3A20 group	OTU278	0.20 (0.07)	0.30 (0.24)
Lachnospiraceae/unknown genus	OTU1666	0.19 (0.07)	-0.21 (0.20)
Prevotella	OTU1496	0.22 (0.07)	-0.32 (0.24)
Prevotella	OTU196	0.16 (0.06)	-0.35 (0.28)
Prevotella	OTU136	0.13 (0.06)	-0.36 (0.34)
F082/unknown_genus	OTU2431	0.20 (0.07)	-0.37 (0.22)

SE= standard error.

The GWAS for LSCS detected on OAR3 four significant QTLs: in decreasing order of significance, the SNP for Socs gene (129,722,200 bp), SNP rs412514556 (129,423,378 bp), SNP rs425386363 (23,336,791 bp) and SNP rs416770789 (133,977,235 bp). The microbiome GWAS detected 387 significant SNPs for 67 OTUs among the 306 OTUs with heritability significantly different from zero. SNPs were scattered on all chromosomes, with the highest p-values and highest number of signals on OAR1, OAR2, OAR3, and OAR5. But, in this study we focus on the OTUs with significant signals located close to the LSCS QTLs on OAR3 (Table 2).

Table 2. Summary of the significant results obtained from the microbiome GWAS.

Genus	OTU name	SNP	Position (bp)	-log10 (P-value)
Prevotella	OTU1419	rs425386363	23,336,791	5.59
Lachnospiraceae/unknown genus	OTU1666[1]	rs425386363	23,336,791	6.90
Ruminococcus	OTU602	Socs	129,722,200	5.70

[1] OTUs with significant genetic correlations with LSCS.

Discussion

The LSCS heritability was greater than the 0.15 reported by Rupp *et al.* (2009), using the whole Lacaune population. Even if the genetic correlations have a high standard error which can lead to spurious ones, most genetic correlations between LSCS and OTUs were positive, meaning that higher bacterial abundance is genetically associated with susceptibility to mastitis. There is also a higher variety of genera within the positive correlations than for the negative correlations where the most frequently associated is *Prevotella*, a bacteria belonging to the *Bacteroidetes* phylum. Zhong *et al*. (2018) obtained *Bacteroidetes* as the most abundant in the group with low SCC, even if they showed that the rumen microbiome in cows was globally stable at different SCC levels. The link between gut bacteria and mastitis is probably not direct but would be through bacterial metabolites absorbed by the host, and which would have pro- or anti-inflammatory properties (Hu *et al.*, 2019). A QTL on OAR3 was detected for

Ruminococcus OTU602 at the Socs gene position, and the Socs2 genotype known to have more LSCS in milk is associated with highest abundance of OTU602 (results not shown). In contrast, Chuang *et al*. (2021) found that *Ruminococcus* genus was phenotypically reduced in the rumen of cows with clinical mastitis. As well on OAR3, two QTLs were detected, one for *Prevotella* OTU1419 and one for *Lachnospiraceae* OTU1666. The SNP genotype associated with more LSCS in milk is linked with more *Prevotella* OTU1419 abundance, which is contrary with the main genetic correlations for this genus, but on others OTUs. The same SNP genotype associated with mastitis susceptibility is also linked with less *Lachnospiraceae* OTU1666 abundance, which is consistent with the negative but weak genetic correlation of -0.21 with LSCS. This result on OTU1666 is in agreement with Wang *et al*. (2021) finding that this family is phenotypically associated with anti-inflammatory molecules such as 2-Phenylbutiric acid. But, for the three others OTUs belonging to the *Lachnospiraceae* family, an increase of their abundances is genetically linked with more LSCS. The identification in this study of QTLs on OAR3 common for LSCS and OTUs in addition to the results obtained for genetic correlations, contribute to the hypothesis of a genetic link between rumen bacteria and mastitis. We can conclude that host genetic may influence ruminal bacteria abundances and udder inflammation in sheep via mainly positive genetic correlations, but the underlying metabolic pathways are still difficult to identify with only the identification of bacterial genera. The genetic study of these bacteria's functions is underway in order to unravel the various results obtained at the OTU level.

References

Aguilar I., Legarra A., Cardoso F., Masuda Y., Lourenco D., *et al*. (2019). Genet Sel Evol 51:28. https://doi.org/10.1186/s12711-019-0469-3

Bokulich N.A., Subramanian S., Faith J.J., Gevers D., Gordon J.I., *et al.* (2013). Nat Methods 10:57–59. https://doi.org/10.1038/nmeth.2276

Chuang S.-T., Li K.-Y., Tu P.-W., Ho S.-T., Hsu C.-C., *et al.* (2021). Animals 11:3108. https://doi.org/10.3390/ani11113108

Gebreyesus G., Difford G.F., Buitenhuis B., Lassen J., Noel S.J., *et al.* (2020). J. Dairy Sci. 103:4557–4569. https://doi.org/10.3168/jds.2019-17824

Hu X., Li S., Fu Y., Zhang N., (2019). Eur. J. Clin. Microbiol. Infect. Dis. 38:1409–1423. https://doi.org/10.1007/s10096-019-03549-4

Larroque H., Chassier M., Saintilan R., Astruc J.M. (2017). Proc. of EAAP, Tallinn, Estonia.

Misztal I., Tsuruta S., Strabel T., Auvray B., Druet T., *et al.* (2002). Proc. of 7th WCGALP, Montpellier, France.

Rupp R., Bergonier D., Dion S., Hygonenq M.C., Aurel M.R., *et al.* (2009). J. Dairy Sci. 92:1203–1219. https://doi.org/10.3168/jds.2008-1435

Rupp R., Senin P., Sarry J., Allain C., Tasca C., *et al*. (2015). PLoS Genet 11:e1005629. https://doi.org/10.1371/journal.pgen.1005629

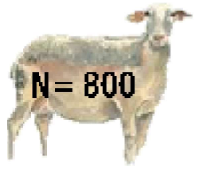
VanRaden P.M., (2008). J. Dairy Sci. 91:4414–4423. https://doi.org/10.3168/jds.2007-0980

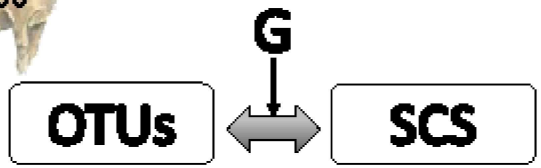
Wang Y., Nan X., Zhao Y., Jiang L., Wang M., *et al*. (2021). J Animal Sci Biotechnol 12:36. https://doi.org/10.1186/s40104-020-00543-1

Zhong Y., Xue M., Liu J. (2018). Front. Microbiol. 9:3217. https://doi.org/10.3389/fmicb.2018.03217

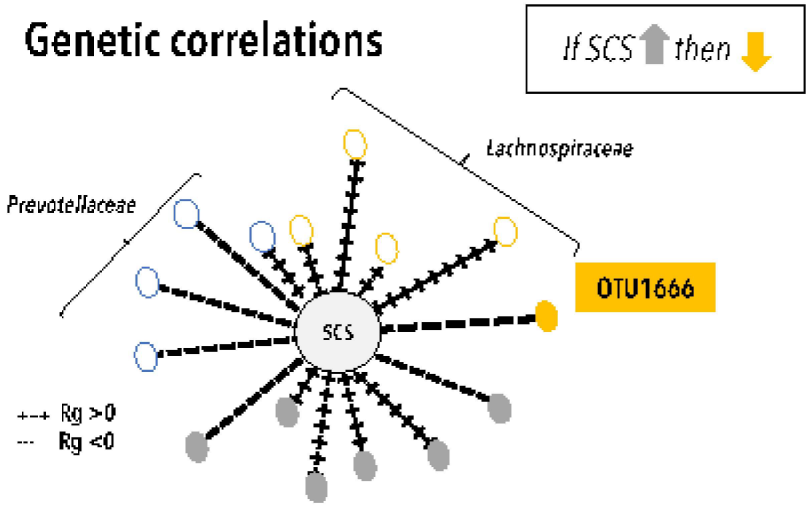
Evidence of genetic links between rumen bacteria and milk somatic cell score in Lacaune dairy ewes

G. Martinez Boggio, A. Meynadier, R. Rupp, C. Allain, A.J. Buitenhuis, C. Marie-Etancelin
guillermo.martinez-boggio@inrae.fr

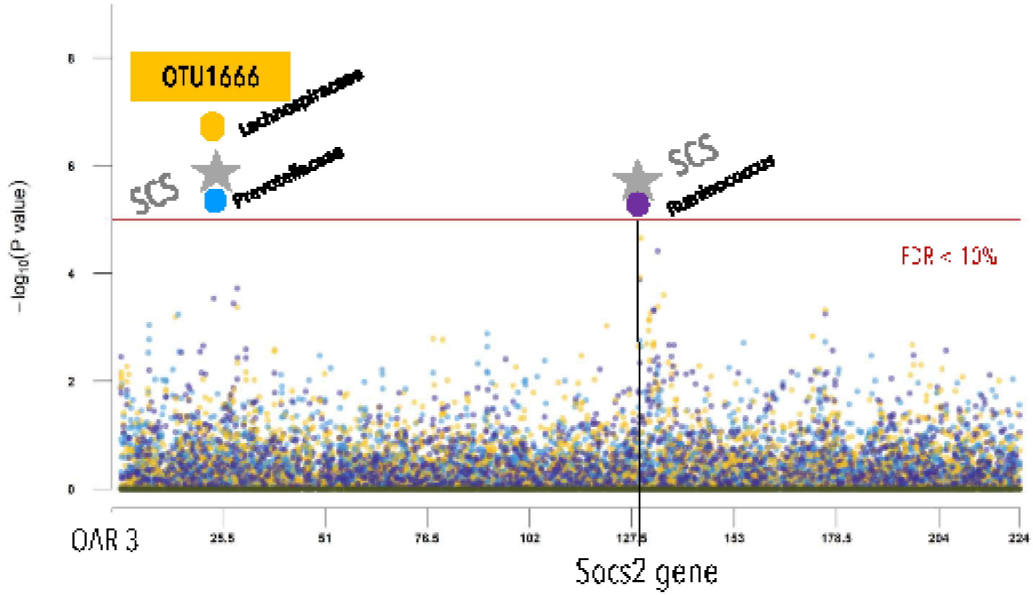




Genetic correlations



GWAS (54 k SNP) on two-traits model



Two-traits GWAS and genetic correlations are two consistent but above all complementary approaches, which allow us to characterize genetic links between milk SCS and rumen bacteria

Conclusion of section 5.3

The host genetics may influence ruminal bacteria abundances and udder inflammation in sheep: using a bi-trait GWAS we identified two QTLs associated with SCS and three OTU abundances. But the links were very few and the underlying metabolic pathways are still difficult to identify with only the identification of bacteria at the genera level.

To conclude, the combination of genetic correlations and GWAS are complementary approaches to characterize the first genetic links between milk somatic cell score and some rumen bacterial abundances.

5.4 Conclusion

As for others livestock species, we demonstrated the modulation of gut bacterial abundances by sheep host genes. The genetic links found between rumen bacterial abundances and fine milk composition traits were moderate to high but for a very small proportion of OTUs of the total microbial community. Combination of methodologies such as genetic correlations and multiple trait GWAS allowed for evaluate more precisely the associations between microbial abundances and phenotypes, but a further step is still needed to see if these links are of interest in animal breeding.

Chapter 6
Impact of rumen microbiota on phenotypes

6.1 Introduction

The high-dimensional omics data can be analyzed as direct trait, as well as complementary information in the models to explained the expression of complex traits. The classical genetic models account the direct genetic effect on the phenotype, but using omics data, it is possible to account also for the indirect genetic effect on the phenotype, that is the genetic effect mediated by RNAs (transcriptomics), proteins (proteomics), metabolites (metabolomics), or metagenomics (microbial genes). In that way, Christensen *et al.* (2021) proposed a joint model for intermediate omics and phenotypes, and a method for prediction of breeding values developed. Our study is focus on microbial abundances, so even if it is not strictly an intermediate trait as proposed the authors, it is still interesting to quantify which part of the host genetic effect is mediated by the rumen bacterial abundances. Not only can the addition of the microbial component have an impact on the accuracies of EBVs, but it will increase knowledge about the potential host genetic control over the microbial composition that impact the phenotype of interest.

The work began with the application of the method proposed by Christensen *et al.* (2021), using a joint model in order to calculated the total heritability of dairy traits. As a first step it was estimated the microbiability and residual heritability using a model which includes dairy traits as phenotype, and the microbiota and genomics as random effects. And as second step it was estimated the heritability of microbial community in the rumen using a model which includes the microbiome effect predicted in first model for each animal and dairy trait, and genomics as random effect. Moreover, we estimated the gain in fit of the genetic model by adding the microbiota effect.

A first part of the results obtained from these analyses were presented as Oral presentation at the 12th World Congress on Genetics Applied to Livestock Production in July 2022 in

Rotterdam (The Netherlands), and the complete results plus MWAS analyses were submitted for publication under the title **"Microbiability of milk composition and genetic control of microbiota effects in sheep"** in Journal of Dairy Science in October 2022.

6.2 Microbiability of milk composition and genetic control of microbiota effects in sheep

Microbiability of milk composition and genetic control of microbiota effects in sheep

G. Martinez Boggio[1], O. F. Christensen[2], A. Legarra[1], A. Meynadier[1], and C. Marie-Etancelin[1]

[1]GenPhySE - Université de Toulouse - INRAE - ENVT, 24 Chemin de Borde Rouge, 31326, Castanet Tolosan, France

[2]Center for Quantitative Genetics and Genomics, Aarhus University, Blichers Alle 20, DK-8830 Tjele, Denmark

Corresponding authors:
Guillermo Martinez Boggio, 24 Chemin de Borde Rouge, Castanet-Tolosan, guillermo.martinez-boggio@inrae.fr.
Christel Marie-Etancelin, 24 Chemin de Borde Rouge, Castanet-Tolosan, christel.marie-etancelin@inrae.fr.

Abstract

Recently, high-dimensional omics data are becoming available in larger quantities, and models have been developed that integrate them with genomics to understand in finer details the relationship between genotype and phenotype, and thus improve the performance of genetic evaluations. Our objectives are to quantify the impact of the inclusion of metagenomic data in the genetic evaluation for dairy traits in sheep, through the estimation of the heritability, microbiability, and how the metagenomic effect on dairy traits decomposes into genetic and non-genetic parts. In this study we analyzed milk and rumen samples of 795 Lacaune dairy ewes. We included, as phenotype, the dairy traits and milk fatty acids and proteins composition; as omics measurements, 16S rRNA rumen bacterial abundances; and as genotyping, 54k SNP chip for all ewes. Two nested genomic models were used: a first model to predict the individual contributions of the genetic and microbial abundances to phenotypes, and a second model to predict the additive genetic effect of the microbial community. In addition, microbiome-wide association studies for all dairy traits was applied using the 2,059 rumen bacterial abundances. Results showed that in general the inclusion of both genetic and microbial effect did not improve the fit of the model compare to the model with the genetic effect only. In addition, for all dairy traits the total heritability was equal to the residual heritability after fitting microbiota effects, due to a microbiability being almost zero for most dairy traits and heritability of the microbial community as a whole was very close to zero. MWAS did not show OTUs with major effect for any of the dairy traits evaluated. So far, we can conclude that, using a substantial dataset of 795 Lacaune dairy ewes, rumen bacterial abundances do not provide improved genetic evaluation for dairy traits in sheep.

Key words

Rumen microbiota, heritability, microbiability, MWAS, dairy sheep

Introduction

In plant and animal breeding, it has recently become possible to obtain high-dimensional omics data in larger quantities, such as metabolites, gene expression, proteins, microbial genes, in addition to the already available genotypes. This has enabled the integration of different types of data to uncover the genotype-phenotype relationship (Morgante et al., 2020), which will be beneficial for the development of an optimized breeding strategy to improve complex traits (Fernie and Schauer, 2009; Guo et al., 2016).

However, to use omics information in genomic evaluations there must be methods and models available that can account for these high-dimensional data. To this end, first, Weishaar et al. (2020) used two models: a microbial model to describe the microbial abundances effect on phenotypes, *i.e.* the microbiability defined by Difford et al. (2018) for gut microbiome in cattle, and a genetic model for microbial abundances, which quantify the host genetic control on the microbiota. The genetic effect on the trait, not mediated by the microbiome, is included in a final step in a selection index. However, in this regard, Christensen et al. (2021) proposed a joint model including the genetic effect not mediated by the microbiome in the first model. This joint model was proposed for intermediate traits, which are those traits between the DNA action and phenotype expression. Although microbiome data is not strictly an intermediate trait from a metabolic point of view, in practice it is a source of information between genotype and phenotype.

In ruminants, the rumen microbiome is crucial to digest plant fiber and is closely associated to productive traits (Difford et al., 2018; Xue et al., 2018; Matthews et al., 2019). Particularly in dairy ruminants, the rumen fermentation products, such as volatile fatty acids, and the microbial amino acids are used by the mammary gland to produce milk with a high solid content which allows the production of high-quality cheese. Our previous results showed that rumen bacterial abundances are controlled by host genetics in Lacaune dairy ewes (Martinez Boggio et al., 2022) and that fine milk composition is linked with particular rumen bacteria (Martinez Boggio et al., 2021). So, we hypothesized that there is an indirect genetic effect on dairy traits mediated by the bacterial abundances in the ewe's rumen. Therefore, we propose to apply the method defined by Christensen et al. (2021) on a Lacaune dairy sheep dataset including targeted metagenomics (16S rRNA gene), and fine milk composition traits as phenotypes, aiming to quantify the impact of the inclusion of the rumen microbial abundance in the genetic evaluation of dairy traits, estimating the residual heritability (h_r^2) and microbiability (c_m^2), and the

heritability of the predicted microbiome effect (h_m^2), which amounts to a general heritability of the microbial community (Christensen et al., 2021).

Materials and Methods

The study was conducted at the INRAE Experimental Unit of La Fage (UE 321 agreement A312031, Roquefort, France) with a protocol for rumen sampling approved by the French Ministry of Higher Education, Research and Innovation – Animal Ethics Committee (approval number: APAFIS#6292-2016080214271984 v8).

Animals' phenotyping

Data of dairy traits and rumen bacterial abundance from 795 multiparous Lacaune dairy ewes were collected from 2015 to 2019. For details, see Martinez Boggio et al. (2022). The data consisted of records of milk yield (MY), and mid-infrared (MIR) spectra predictions of the milk fat content (FC), protein content (PC) and fine profile of daily milk fatty acids (FAs) and proteins. The accuracies of the predictive equations of ewe milk FAs and proteins were retrieved from Ferrand-Calmels et al. (2014) and Ferrand et al. (2012), respectively. The coefficients of determination for the FAs retained for this study, were higher than 0.91, with the exception of *c9t11* C18:2 and C18:3*n-3* with values of 0.74. Meanwhile, for caseins, the coefficients of determination were higher than 0.82, and equal to 0.77 and 0.26 for beta-lactoglobulin and alpha-lactalbumin, respectively. The fatty acids predicted by MIR were saturated FAs, such as butyric acid (C4:0), caproic acid (C6:0), caprylic acid (C8:0), capric acid (C10:0), lauric acid (C12:0), and palmitic acid (C16:0), and unsaturated FAs, such as oleic acid (*c9* C18:1), rumenic acid (*c9t11* C18:2) and alpha-linolenic acid (C18:3*n-3*). The proteins predicted by MIR were caseins, namely, alpha-S1-casein (α_{s1}-CN), alpha-S2-casein (α_{s2}-CN), beta-casein (β-CN) and kappa-casein (κ-CN), and two whey proteins, namely, alpha-lactalbumin (α-LAC) and beta-lactoglobulin (β-LG).

Rumen sampling

Rumen sampling was performed in morning or afternoon within 3 days around the official milk recording of the flock. Ruminal contents were sampled from each ewe using a vacuum pump and a medical gastric tube. After extraction, the DNA strands of the 795 samples were sequenced using V3-V4 region of 16S rRNA gene with Illumina MiSeq technology (Illumina, USA) at the Genomic and Transcriptomic Platform (INRAE, Toulouse, France). More details

on rumen sampling, DNA extraction, amplicon sequencing and bioinformatic process are provided in Martinez Boggio et al. (2022). We obtained an abundance table with 2,059 operational taxonomic units (OTUs) as result of the bioinformatic process of DNA sequences using FROGS 3.0 pipeline (Escudié et al., 2018).

Genotyping

Among the 795 ewes, 743 were genotyped using Illumina Ovine SNP50 BeadChip (54,241 SNPs), and 52 ewes were genotyped with Illumina Ovine SNP15 (16,681 SNPs) followed by imputation to a medium-density SNP chip as part of the Lacaune dairy sheep genomic selection program (Larroque et al., 2017). Genotypes were subjected to quality control, including minimum call rates of 90% for SNPs and 95% for individuals and exclusion of SNPs with a minor allele frequency lower than 5%. The final dataset included 773 genotyped individuals and 35,492 autosomal SNPs.

Statistical analyses

The methodology applied for inference with complete omics data was the one proposed by Christensen et al. (2021), where a joint model for omics and phenotypes was proposed, and a method for prediction of breeding values developed. The method consisted of the following two steps: 1) prediction of the individual contributions of the genetic and microbiome to phenotypes, and 2) prediction of the additive genetic effect of the microbial community. As an additional analysis, in the first step, we compared two models to assess the relevance of including the microbiome effect to the genetic one. The individual dairy traits were analyzed separately using single trait models.

Step 1

The first step consisted on fitting a single trait model including the residual additive genetic effect and the microbiome effect (Eq. 1).

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Z_1 a_r} + \boldsymbol{Z_2 u} + \boldsymbol{e}, \quad (1)$$

where $\boldsymbol{y}$ is the vector of observations, $\boldsymbol{b}$ is the vector of fixed effects, $\boldsymbol{a_r}$ is the vector of random residual additive genetic effects, $\boldsymbol{u}$ is the vector of individual effects due to microbiome, and $\boldsymbol{e}$ is the vector of random residual effects, $\boldsymbol{X}$ is the incidence matrix for $\boldsymbol{b}$, and $\boldsymbol{Z_1}$ and $\boldsymbol{Z_2}$ are incidence matrices for $\mathbf{a_r}$ and $\boldsymbol{u}$, respectively. The distributional assumptions are $\boldsymbol{a_r} \sim N(0, \boldsymbol{G}\sigma^2_{a_r})$, $\boldsymbol{u} \sim N(0, \boldsymbol{O}\sigma^2_u)$, $\boldsymbol{e} \sim N(0, \boldsymbol{I}\sigma^2_e)$, where $\boldsymbol{G}$ is a genomic relationship matrix

computed based on the first method proposed by VanRaden (2008), $\boldsymbol{O}$ is the microbial similarity matrix, and $\boldsymbol{I}$ is an identity matrix; $\sigma^2_{a_r}$ is the residual additive genetic variance, σ^2_u is the microbiome effect variance, and σ^2_e is the residual variance. The microbiability was estimated as $c^2_m = \sigma^2_u/\sigma^2_y$, and the residual heritability as $h^2_r = \sigma^2_{a_r}/\sigma^2_y$, where σ^2_y is the phenotypic variance.

The fixed effects considered in $\boldsymbol{b}$ (Eq. 1) were defined by ANOVA test with P <0.05. For all dairy traits, we included the effects of days in milk (28 to 133 DIM) as a covariate and the sampling year (with five levels: 2015 to 2019). For FC and milk FAs, we included the number of lactations nested in the sampling year (with seven levels: in 2015, ewes in second, third or fourth and more lactations, and in 2016 to 2019 ewes in second lactation only), and for PC, milk proteins and FAs, we also included litter size, with two levels (1, or 2 and more lambs).

The microbial similarity matrix ($\boldsymbol{O}$) for the 795 ewes was computed based on the rumen bacterial abundance similarities between animals using the method proposed by Ross et al. (2013):

$$\boldsymbol{O} = \frac{\boldsymbol{M}\boldsymbol{M}'}{n}$$

The matrix was computed as a variance-covariance matrix from rumen bacterial abundances as where $\boldsymbol{M}$ is the abundance matrix with $n = 2{,}059$ OTUs, with zeros corrected with the geometric Bayesian-multiplicative (GBM) method (Martín-Fernández et al., 2015) and centered log-ratio (CLR) transformed. OTU abundances were pre-corrected for the significant fixed effects (for more than 10% of OTUs; P <0.05) not included in $\boldsymbol{b}$ (Eq. 1), such as total number of sequences per rumen sample, sequencing run, and sampling time and order, all of them nested on the sampling year. Note that the microbial matrix $\boldsymbol{O}$, is called $\boldsymbol{M}$ by other authors (Difford et al., 2018; Weishaar et al., 2020), but here we use $\boldsymbol{M}$ for the abundance matrix to have the same notation as Christensen et al. (2021). Similarly, microbiability, that we call c^2_m, was named m^2 in Difford et al. (2018).

To assess the relevance of including the microbiome effect in the prediction model, we used the full model with genetic and microbiome effects (Eq. 1), and a model with the genetic effect only as follows:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z_1}\boldsymbol{g} + \boldsymbol{e}, \quad (2)$$

where $\boldsymbol{g}$ is the vector of random additive genetic effects, and $\boldsymbol{y}$, $\boldsymbol{b}$, $\boldsymbol{e}$, $\boldsymbol{X}$ and $\boldsymbol{Z_1}$ are described above.

This model with the genetic effect only (Eq. 2), defined as the null hypothesis, was compared to the full one (Eq. 1) on the basis of a likelihood ratio test. The *P*-values were computed considering that the distribution of likelihood ratio test asymptotically is a mixture of Chi-squared distributions (Visscher, 2006). The significance for Chi-squared test with one degree of freedom was defined at $P < 0.05$.

Step 2

The second step consisted on fitting a single trait model using the estimates of microbiome effect (obtained from Eq. 1) as phenotype:

$$\widehat{\boldsymbol{u}} = \boldsymbol{W\beta} + \boldsymbol{Z_1 g} + \boldsymbol{\varepsilon} \quad (3)$$

where $\widehat{\boldsymbol{u}}$ is the predicted effects $\boldsymbol{u}$ in Equation 1 for each individual, $\boldsymbol{\beta}$ is the vector of fixed effects (same effects as in $\boldsymbol{b}$); $\boldsymbol{g}$ is the vector of random additive genetic effects on microbial OTUs; and $\boldsymbol{\varepsilon}$ is the vector of random residual effects; $\boldsymbol{W}$ is the incidence matrix for $\boldsymbol{\beta}$. The distributional assumptions are $\boldsymbol{g} \sim N(0, \boldsymbol{G}\sigma_g^2)$ and $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{I}\sigma_\varepsilon^2)$, where, $\boldsymbol{G}$ is a genomic relationship matrix, and $\boldsymbol{I}$ is an identity matrix; σ_g^2 is additive genetic variance and σ_ε^2 is residual variance. The heritability of the microbial community was calculated as $h_m^2 = \sigma_g^2/(\sigma_g^2 + \sigma_\varepsilon^2)$. Here, contrary to our previous work (Martinez Boggio et al., 2022) we assume (or force) the same heritability for all OTUs, which is a simplifying assumption in Christensen et al. (2021). In theory, using the Equation 3 in all traits should result in the same h_m^2 . However, first, when $\sigma_u^2 \to 0$ there is no variation in $\widehat{\boldsymbol{u}}$ and h_m^2 (although well defined) is not estimable. Second, because the variance components of Equation 1 are themselves estimates, the estimate of variance components in Equation 3, which uses inferred $\widehat{\boldsymbol{u}}$, will suffer from incorrect estimation of variance components in Equation 1.

Thus, the total heritability of a phenotype is decomposed into a microbiome mediated heritability ($c_m^2 h_m^2$), and a residual heritability (h_r^2) according to the formula $h^2 = c_m^2 h_m^2 + h_r^2$ presented by Christensen et al. (2021). All variance components were estimated using BLUPF90+ (Misztal et al., 2002) with OPTION method VCE.

Microbiome-wide association study (MWAS)

Single-OTU regression analyses were applied to test the effect of the 2,059 OTUs, one at a time, and obtain the associated *P*-value. The model used was the Equation 2 plus the specific OTU as a fixed covariate. The *P*-value of the estimate of the regression coefficient for the fitted OTU-covariate was obtained by converting the estimate and its standard error to Z-score and

applying a Chi-squared test. The model was fitted using BLUPF90+ (Misztal et al., 2002) with OPTION method VCE.

Results

Description of phenotypes

A data summary on dairy traits and rumen bacterial composition is available in Martinez Boggio et al. (2022). Briefly, MY averaged 1.95 ± 0.59 liters, with a FC of 7.37 ± 1.14 g/100ml and a PC of 5.71 ± 0.52 g/100ml. Among the fine milk components measured by MIR spectra, beta-casein was the most abundant protein (2.10 ± 0.23 g/100ml) and palmitic acid (C16:0) the most abundant FA (1.96 ± 0.37 g/100ml).

Comparison between full model and genetic model

The inclusion of the microbiome effect was significant for beta-lactoglobulin (P <0.01), alpha-S2-casein and kappa casein (P <0.05). However, we obtained non-significant effects of the inclusion of the microbiome effect in all FAs (Table 1).

Microbiability and heritability of the microbial community

The microbiability estimated from the full model was almost zero for most dairy traits (Table 1), with exception of beta-lactoglobulin (0.06 ± 0.05), alpha-S2-casein (0.07 ± 0.05), kappa-casein and alpha-S1-casein (0.04 ± 0.04) and PC (0.03 ± 0.03) where values were barely higher than zero. For all the dairy traits with a microbiability almost equal to zero, the model in step 2 did not run because the predicted phenotypes $\widehat{\boldsymbol{u}}$ have zero variance and h_m^2 is therefore not estimable. For the remaining traits having a nonzero microbiability (albeit small in all cases), h_m^2 was very close to zero. So, in practice, the total heritability ($c_m^2 h_m^2 + h_r^2$) of dairy traits was equal to the residual heritability (from step 1) as shown in Table 1. Furthermore, we tested the contribution of each specific OTU on dairy traits using a MWAS, but no OTU for any dairy trait showed a significant effect (results not shown).

Table 1. Estimates of residual heritability (h_r^2), microbiability (c_m^2), heritability of the microbial community (h_m^2), total heritability as product of ($c_m^2 h_m^2$) plus (h_r^2), and heritability (h^2) from the model including only the genetic effect, and *P*-values from the comparison between the full model (Eq. 1) and model with genetic effect only (Eq. 2). †model in Equation 4 not run**.**

Trait	h_r^2	c_m^2	h_m^2	$c_m^2 h_m^2 + h_r^2$	h^2	*P*-value
MY	0.25±0.07	0.005±0.02	†	0.25	0.25±0.07	0.33
FC	0.54±0.07	<10e-4 (<10e-3)	†	0.54	0.54±0.07	0.50
PC*	0.51±0.07	0.03±0.03	0.004±0.003	0.51	0.52±0.07	0.06*
α_{s1}-CN*	0.49±0.09	0.04±0.04	0.006±0.004	0.49	0.50±0.09	0.05*
α_{s2}-CN**	0.70±0.08	0.07±0.05	0.03±0.007	0.70	0.69±0.08	0.02**
β-CN	0.42±0.09	<10e-4 (<10e-3)	†	0.42	0.42±0.09	0.50
κ-CN**	0.47±0.08	0.04±0.04	0.006±0.003	0.47	0.48±0.08	0.03**
α-LAC	0.28±0.10	<10e-4 (<10e-3)	†	0.28	0.28±0.10	0.50
β-LG***	0.44±0.09	0.06±0.05	0.003±0.002	0.44	0.46±0.09	0.005***
C4:0	0.46±0.09	<10e-4 (<10e-3)	†	0.46	0.46±0.09	0.50
C6:0	0.44±0.09	<10e-4 (<10e-3)	†	0.44	0.44±0.09	0.50
C8:0	0.47±0.09	<10e-4 (<10e-3)	†	0.47	0.47±0.09	0.50
C10:0	0.52±0.09	<10e-4 (<10e-3)	†	0.52	0.52±0.09	0.50
C12:0	0.55±0.08	<10e-4 (<10e-3)	†	0.55	0.55±0.08	0.50
C16:0	0.47±0.08	0.01±0.03	†	0.47	0.47±0.08	0.41
c9 C18:1	0.43±0.09	<10e-4 (<10e-3)	†	0.43	0.43±0.09	0.50
c9t11 C18:2	0.35±0.09	<10e-4 (<10e-3)	†	0.35	0.35±0.09	0.50
C18:3*n-3*	0.36±0.09	<10e-4 (<10e-3)	†	0.36	0.36±0.09	0.50

Discussion

In this study, we evaluated the impact of rumen bacterial abundance on milk composition traits in dairy ewes. Rumen microbiota was considered as an intermediate trait in the model of Christensen et al. (2021), although microbiota is not strictly intermediate between DNA action and expression of the phenotype of interest. Despite that, it is interesting to assess the potential impact of including metagenomic data in the genetic evaluation of milk composition traits closely associated with rumen microbiota (Matthews et al., 2019). Moreover, a genetic control of bacterial abundances by the host, and genetic associations with the fine milk composition, had been shown Martinez Boggio et al. (2022). This close link of some rumen microbiota with milk FAs is due to rumen fermentation products, such as volatile fatty acids, which influence the milk FA profile through *"de novo"* synthesis of FAs in the mammary gland and biohydrogenation of unsaturated FAs in the rumen (Lourenço et al., 2010; Osorio et al., 2016). Furthermore, the link between rumen microbiota and milk proteins is mainly due to the fact that the passage of microbial proteins to the intestine supplies the mammary gland with essential amino acids used for protein production (Osorio et al., 2016). Although microbiota provides precursors for the production of milk FAs and proteins, the microbiome effect obtained in this study was only slightly significant for milk protein models.

For all dairy traits included in this study, total heritability was moderate to high (0.25 to 0.70). However, the contribution to total heritability was entirely due to the direct genetic effect on the phenotype, *i.e.* total heritability ($c_m^2 h_m^2 + h_r^2$) was equal to the residual heritability (h_r^2). This was due, first, to the weak impact of the microbiota on most dairy traits, as shown by low values of microbiability (c_m^2) estimated. In the literature, a zero microbiability was reported for most milk FAs in dairy cows, except for C15:0 (0.42 ± 0.18) and C18:3*n*-3 (0.31 ± 0.14) (Buitenhuis et al., 2019), and there are no references for milk proteins, for which we obtained a very low impact of the microbiota (*e.g.* alpha-S2-casein 0.07 ± 0.05 and beta-lactoglobulin 0.06 ± 0.05). Second, the heritability of the microbial community (h_m^2) as a whole estimated in step 2 was close to zero, and similar to the average heritability of 0.04 ± 0.03 obtained for the 2,059 OTUs (Martinez Boggio et al., 2022). Thus, those elements evidence a non-contribution of the genetic effect on the phenotype mediated by the rumen microbial abundance ($c_m^2 h_m^2$).

Despite the results obtained with dairy traits, the methodology proposed by Christensen et al. (2021) allowed to separate the direct genetic effect on phenotypes from the indirect genetic

effect on phenotypes mediated by the rumen microbiome. However, regarding the indirect genetic effect, we observed that at the extreme when the predicted microbiome effects $\hat{\boldsymbol{u}}$ were zero, the model in step 2 did not run. That imposed limitations on the use of the resulting predicted phenotype for the quantification of host genetic control of the microbial community. The results here showed that including metagenomics as an additional information source of data will not improve accuracy of estimated breeding values. It would be interesting to evaluate some elements that may improve this accuracy. Firstly, the use of traits more closely associated with rumen microbiota, such as rumen FAs or methane emissions, which are direct products of the rumen microbial community (Hurtaud et al., 1993; Dehority, 2003). In the literature for those traits, moderate but significant impacts of the rumen microbiota were obtained. For metabolic traits, such as milk acetone and beta-hydroxybutyric acid, a microbiability of 0.15 ± 0.09 were estimated for both traits (Gebreyesus et al., 2020). In addition, for methane emissions Difford et al. (2018) and Ramayo-Caldas et al. (2020) in dairy cows and Hess et al. (2020) in meat sheep, reported microbiabilities of 0.13 ± 0.09, 0.16 ± 0.09 and 0.19 ± 0.07, respectively. Secondly, replacing the use of rumen microbial abundances by rumen microbial functions to construct the microbial similarity matrix would avoid functional redundancy at the rumen microbiota (Weimer, 2015). Furthermore, we suggest much further work is needed to evaluate more complex models that fit the biological nature of the microbial community in the rumen, since the methodology used in this study was designed for intermediate -omics traits between DNA and the phenotype of interest, which is not really the case with microbiota.

Conclusions

Rumen microbial abundance does not contribute to the phenotypic variance of most of the fine milk composition traits and no causative OTUs were detected. In turn, there was no genetic control of the microbial community. So far, we can conclude that, using a substantial dataset of 795 Lacaune dairy ewes, rumen bacterial abundances do not provide improved genetic evaluation for dairy traits in sheep. Further work will be needed to evaluate traits that are a direct product of the microbiota (*e.g.* rumen volatile fatty acids or methane emissions) as well as the substitution of microbial abundances for microbial functions.

Acknowledgments

We are grateful to the technical staff at INRAE experimental unit of La Fage (DOI: 10.15454/1.548325523466425E12) for animal care and help with rumen sampling. We thank Béatrice Gabinaud for preparing microbiota samples and Get-Plage platform in Toulouse for DNA sequencing, and Hélène Larroque and Rachel Rupp for access to the genetic resources produced at INRAE La Fage.

References

Buitenhuis, B., Lassen, J., Noel, S. J., Plichta, D. R., Sørensen, P., Difford, G. F., & Poulsen, N. A. (2019). Impact of the rumen microbiome on milk fatty acid composition of Holstein cattle. *Genetics Selection Evolution*, *51*(1), 23. https://doi.org/10.1186/s12711-019-0464-8

Christensen, O. F., Börner, V., Varona, L., and Legarra, A. 2021. Genetic evaluation including intermediate omics features. Genetics. 219:iyab130. https://doi.org/10.1093/genetics/iyab130

Dehority, B. A. 2003. Gross anatomy, physiology and environment of the ruminant stomach. Pages 19-42 in Rumen microbiology. Nottingham Univ. Press. UK

Difford, G. F., Plichta, D. R., Løvendahl, P., Lassen, J., Noel, S. J., Højberg, O., Wright, A.-D. G., Zhu, Z., Kristensen, L., Nielsen, H. B., Guldbrandtsen, B., and Sahana, G. 2018. Host genetics and the rumen microbiome jointly associate with methane emissions in dairy cows. PLOS Genet. 14:e1007580. https://doi.org/10.1371/journal.pgen.1007580

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. https://doi.org/10.1093/bioinformatics/btx791

Fernie, A. R., and Schauer, N. 2009. Metabolomics-assisted breeding: A viable option for crop improvement? Trends Genet. 25:39–48. https://doi.org/10.1016/j.tig.2008.10.010

Ferrand, M., Miranda, G., Larroque, H., Leray, O., Guisnel, S., Lahalle, F., Brochard, M., and Martin, P. 2012. Determination of protein composition in milk by mid-infrared spectrometry. ICAR 38 Annual Meeting. 5.

Ferrand-Calmels, M., Palhière, I., Brochard, M., Leray, O., Astruc, J. M., Aurel, M. R., Barbey, S., Bouvier, F., Brunschwig, P., Caillat, H., Douguet, M., Faucon-Lahalle, F., Gelé, M.,

Thomas, G., Trommenschlager, J. M., and Larroque, H. 2014. Prediction of fatty acid profiles in cow, ewe, and goat milk by mid-infrared spectrometry. J. Dairy Sci. 97:17–35. https://doi.org/10.3168/jds.2013-6648

Gebreyesus, G., Difford, G. F., Buitenhuis, B., Lassen, J., Noel, S. J., Højberg, O., Plichta, D. R., Zhu, Z., Poulsen, N. A., Sundekilde, U. K., Løvendahl, P., and Sahana, G. 2020. Predictive ability of host genetics and rumen microbiome for subclinical ketosis. J. Dairy Sci. 103:4557–4569. https://doi.org/10.3168/jds.2019-17824

Guo, Z., Magwire, M.M., Basten, C.J., Xu, Z., and Wang, D. 2016. Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. Theor. Appl. Genet. 129:2413–2427. https://doi.org/10.1007/s00122-016-2780-5.

Hess, M. K., Rowe, S. J., Van Stijn, T. C., Henry, H. M., Hickey, S. M., Brauning, R., McCulloch, A. F., Hess, A. S., Kirk, M. R., Kumar, S., Pinares-Patiño, C., Kittelmann, S., Wood, G. R., Janssen, P. H., and McEwan, J. C. 2020. A restriction enzyme reduced representation sequencing approach for low-cost, high-throughput metagenome profiling. PLOS ONE. 15:e0219882. https://doi.org/10.1371/journal.pone.0219882

Hurtaud, C., Rulquin, H., and Verite, R. 1993. Effect of Infused Volatile Fatty Acids and Caseinate on Milk Composition and Coagulation in Dairy Cows. J. Dairy Sci. 76:3011–3020. https://doi.org/10.3168/jds.S0022-0302(93)77640-7

Larroque, H., Chassier, M., Saintilan, R., and Astruc, J. M. 2017. Imputation accuracy from a low density SNP panel in 5 dairy sheep breeds in France. Annual Meeting of the European association for animal production. 513. https://hal.inrae.fr/hal-02737593

Lourenço, M., Ramos-Morales, E., and Wallace, R. J. 2010. The role of microbes in rumen lipolysis and biohydrogenation and their manipulation. Animal. 4:1008–1023. https://doi.org/10.1017/S175173111000042X

Martinez Boggio, G., Meynadier, A., Buitenhuis, A. J., & Marie-Etancelin, C. (2022). Sheep host genetics shape the rumen microbiota associated with dairy traits (In revision). *Genetics Selection Evolution*.

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:e0254874. https://doi.org/10.1371/journal.pone.0254874

Martín-Fernández, J.-A., Hron, K., Templ, M., Filzmoser, P., and Palarea-Albaladejo, J. 2015. Bayesian-multiplicative treatment of count zeros in compositional data sets. Stat. Modelling. 15:134–158. https://doi.org/10.1177/1471082X14535524

Matthews, C., Crispie, F., Lewis, E., Reid, M., O'Toole, P. W., and Cotter, P. D. 2019. The rumen microbiome: A crucial consideration when optimising milk and meat production and nitrogen utilisation efficiency. Gut Microbes. 10:115–132. https://doi.org/10.1080/19490976.2018.1505176

Misztal, I., Tsuruta, S., Strabel, T., Auvray, B., Druet, T., and Lee, D. H. 2002. BLUPF90 AND RELATED PROGRAMS (BGF90). Proceedings of the 7th World Congress on Genetics Applied to Livestock Production. 7.

Morgante, F., Huang, W., Sørensen, P., Maltecca, C., and Mackay, T. F. C. 2020. Leveraging Multiple Layers of Data To Predict *Drosophila* Complex Traits. G3: Genes Genomes Genet. 10:4599–4613. https://doi.org/10.1534/g3.120.401847

Osorio, J. S., Lohakare, J., and Bionaz, M. 2016. Biosynthesis of milk fat, protein, and lactose: Roles of transcriptional and posttranscriptional regulation. Physiol. Genomics. 48:231–256. https://doi.org/10.1152/physiolgenomics.00016.2015

Ramayo-Caldas, Y., Zingaretti, L., Popova, M., Estellé, J., Bernard, A., Pons, N., Bellot, P., Mach, N., Rau, A., Roume, H., Perez-Enciso, M., Faverdin, P., Edouard, N., Ehrlich, D., Morgavi, D. P., and Renand, G. 2020. Identification of rumen microbial biomarkers linked to methane emission in Holstein dairy cows. J. Anim. Breed. Genet. 137:49–59. https://doi.org/10.1111/jbg.12427

Ross, E. M., Moate, P. J., Marett, L., Cocks, B. G., and Hayes, B. J. 2013. Investigating the effect of two methane-mitigating diets on the rumen microbiome using massively parallel sequencing. J. Dairy Sci. 96:6030–6046. https://doi.org/10.3168/jds.2013-6766

VanRaden, P. M. 2008. Efficient Methods to Compute Genomic Predictions. J. Dairy Sci. 91:4414–4423. https://doi.org/10.3168/jds.2007-0980

Visscher, P. M. 2006. A Note on the Asymptotic Distribution of Likelihood Ratio Tests to Test Variance Components. Twin Res. Hum. Genet. 9:490–495. https://doi.org/10.1375/twin.9.4.490

Weimer, P. J. 2015. Redundancy, resilience, and host specificity of the ruminal microbiota: Implications for engineering improved ruminal fermentations. Front. Microbiol. 6. https://doi.org/10.3389/fmicb.2015.00296

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. J. Anim. Breed. Genet. 137:14–22. https://doi.org/10.1111/jbg.12447

Xue, M., Sun, H., Wu, X., Guan, L. L., and Liu, J. 2018. Assessment of Rumen Microbiota from a Large Dairy Cattle Cohort Reveals the Pan and Core Bacteriomes Contributing to Varied Phenotypes. Appl. Environ. Microbiol. 84:e00970-18. https://doi.org/10.1128/AEM.00970-18

6.2.1 Supplementary results

In this chapter we estimated the microbiability using a mixed model with a genomic effect and microbial effect. We accounted the microbial effect through a microbial similarity matrix based on the microbial abundances for each individual. We constructed the $\boldsymbol{O}$ matrix using the method proposed by Ross *et al.* (2013b):

$$\boldsymbol{O} = \frac{\boldsymbol{M}\boldsymbol{M}'}{n} \quad (9)$$

where, $\boldsymbol{M}$ is the abundance matrix with $n =$ total number of OTUs. However, we have tried including different information and number of OTUs in the $\boldsymbol{M}$ matrix:

- $\boldsymbol{O}_\mathbf{1}$: All 2,059 OTU CLR-transformed abundances, GBM-imputed zeros, and values corrected only for the number of sequences.
- $\boldsymbol{O}_\mathbf{2}$: All 2,059 OTU CLR-transformed abundances, GBM-imputed zeros, and values corrected for all significant fixed effects.
- $\boldsymbol{O}_\mathbf{3}$: All 2,059 OTU log10-transformed abundances, zeros imputed by adding one, and values corrected for all significant fixed effects.
- $\boldsymbol{O}_\mathbf{4}$: 306 'heritable' OTU CLR-transformed abundances, GBM-imputed zeros, and values corrected for all significant fixed effects (to note GBM and CLR were applied to the all 2,059 OTUs and then 'heritable' OTUs were selected).

We chose beta-lactoglobulin as the trait to test the different types of microbial matrices, because it was one of the traits that showed the highest microbial effect in the abovementioned estimation models. As shown in Table 7, the microbiabilities were quite similar among the different matrices evaluated, ranging from 0.03 ± 0.02 to 0.10 ± 0.07. However, the matrix including CLR-transformed abundances corrected for all fixed effects ($\boldsymbol{O}_\mathbf{2}$) was the one that showed the average of the diagonal ($\overline{\mathrm{diag}(\mathbf{O})}$) closest to one, and average relationship ($\overline{\boldsymbol{O}}$) equal zero. This is relevant, as proposed by Legarra (2016), for the relationship matrices used to obtain comparable estimates ($\overline{\mathrm{diag}(\mathbf{O})}$ of 1 and $\overline{\boldsymbol{O}}$ of 0).

Table 7. Estimates of variance components (σ^2_{ar} residual genetic, σ^2_u microbial, and σ^2_e residual variances), heritability (h^2), microbiability (c^2_m), statistics of the relationship matrices (average of the diagonal $\overline{diag(\boldsymbol{O})}$ and average relationship $\bar{\boldsymbol{O}}$) for beta-lactoglobulin, for four different types of the relationships ($\boldsymbol{O}_\mathbf{1}$, $\boldsymbol{O}_\mathbf{2}$, $\boldsymbol{O}_\mathbf{3}$, $\boldsymbol{O}_\mathbf{4}$).

	$\boldsymbol{O}_\mathbf{1}$	$\boldsymbol{O}_\mathbf{2}$	$\boldsymbol{O}_\mathbf{3}$	$\boldsymbol{O}_\mathbf{4}$
σ^2_{ar}	0.35±0.08	0.36±0.08	0.36±0.08	0.36±0.08
σ^2_u	0.05±0.03	0.05±0.04	0.08±0.07	0.02±0.02
σ^2_e	0.40±0.07	0.40±0.07	0.41±0.07	0.42±0.07
h^2_r	0.44±0.09	0.44±0.09	0.42±0.09	0.45±0.09
c^2_m	0.06±0.04	0.06±0.05	0.10±0.08	0.03±0.02
$\overline{diag(\boldsymbol{O})}$	1.28	1.05	0.47	1.14
$\bar{\boldsymbol{O}}$	1.26e-06	1.26e-06	1.26e-06	1.26e-06

6.3 Conclusion

This work was the first implementation of the methodology proposed by Christensen *et al.* (2021) for dairy traits, which allows for estimate the direct and indirect genetic effect mediated by the metagenomics on the phenotype. Unfortunately, microbiabilities for fine milk composition traits were low or equal to zero. These very low microbiabilities generate problems in estimating the associated genetic effect, which revealed some limitations of the method used. The heritability of the microbial community was expected due to the mean of heritability obtained for the 2,059 OTUs. Then, to conclude, with the dataset used and the low genetic variance explained by the rumen microbiota, the rumen bacteria abundances do not provide improved genetic evaluation for dairy traits in sheep.

Chapter 7

General discussion

In recent years, with the microbiome revolution, as defined by Blaser (2014), more attention has been given to the microbiota composition of animals and their crucial role in their host. In that sense, the concept of a holobiont, *i.e.,* the assemblage between the host and their microbial communities was proposed as a single unit. Considering the organism as a whole allows for a better understanding of its functioning and therefore for the improvement of traits of interest in animal breeding. The microbiota is associated with some of these traits of interest (*e.g.,* milk composition traits, methane emissions, nutrient digestibility, growth traits); in ruminants, this occurs through the rumen symbiotic microorganisms responsible for converting nondigestible plant mass into energy (Hungate, 1966). This capability is of enormous importance to mankind because ruminants are intermediaries between the light energy harvested in green parts of plants and the production of digestible compounds for humans such as milk and meat. Fermentation products of the rumen microbiota, mainly bacteria, have a direct effect on the composition of fatty acids and proteins in milk (Hurtaud *et al.*, 1993). The combination of this information on the impact of the microbiota on productive traits and the host and microbial genome based on recent advances in NGS technologies allows for us to better understand the relationship between the host and its microbiome.

The overall objective of this thesis was to integrate the relationship between the host genome and the rumen metagenome of the animal in modeling the variability of milk composition and udder health in dairy sheep.

7.1 What impact does the rumen microbiota have on phenotypic variability?

The main results of this thesis that contribute to this question are presented in Figure 15. These are integrated and discussed throughout this section for a better understanding of the impact of the rumen microbiota on phenotypes.

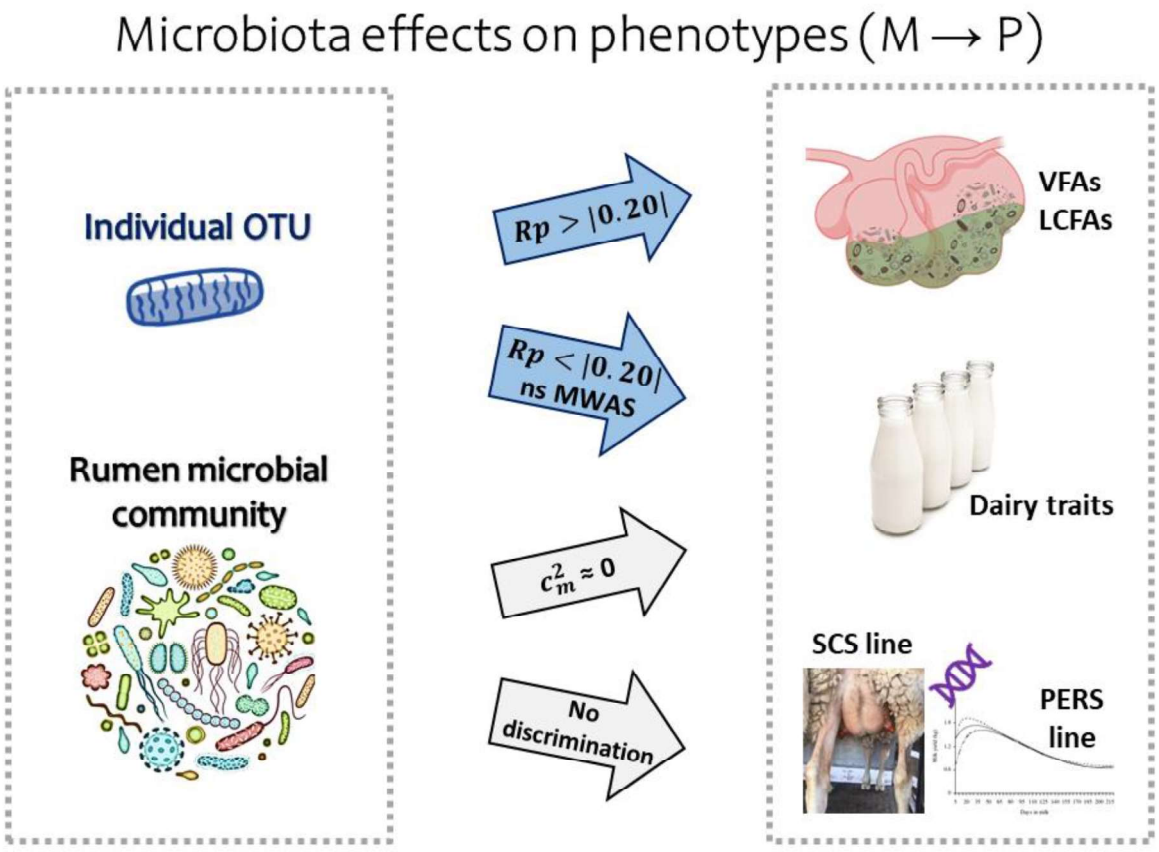


Figure 15. Main results obtained for individual and microbiota effects on phenotypes. OTU: operational taxonomic unit, Rp: phenotypic correlation, c_m^2: microbiability, VFAs: volatile fatty acids in rumen, LCFAs: long-chain fatty acids in rumen, SCS line: genetic lines selected for somatic cell score, PERS: genetic lines selected for persistency in milk production. The microbiome-wide association study (MWAS) did not show any significant OTU for any of the dairy traits. The rumen microbial community was not able to discriminate through a discriminant analysis the genetic lines SCS (SCS+ and SCS-), nor PERS (PERS+ and PERS-).

7.1.1 Links between OTU abundances and phenotypes

The rumen bacterial abundance was low to moderately associated with phenotypes. In general, for traits that are a direct product of bacteria in the rumen, such as VFAs and LCFAs, OTUs showed moderate phenotypic correlations (|0.20-0.50|). However, for milk composition traits, such as FAs and proteins, OTUs showed low phenotypic correlations (|0.08-0.20|), and were even lower for MY and SCS (<|0.11|). These results evidence different degrees of association between rumen bacteria and their direct fermentation products and final milk components. The low correlations observed between bacteria and milk composition traits are logical given the many genetic (Osorio *et al.*, 2016) and nongenetic factors (Hurtaud *et al.*, 1993) that mediate

the synthesis of FAs and proteins in the mammary gland. In this sense, ruminal bacteria play a key role in the production of FAs that determine the milk composition. This intermediate process is evidenced by the identification of the same predominant bacterial genera in the rumen, such as *Prevotella*, *Ruminococcus*, and *Rikenellaceae_RC9_gut_group*, associated with traits in the rumen and milk. However, we also identified a few bacterial genera, such as *Succinivibrionaceae_UCG-002, Lachnospiraceae_NK3A20_group* and *Acetitomaculum*, that were associated with rumen or milk FAs but not with both. The approach used in this study allows for initial evaluation of the associations with direct rumen fermentation products such as VFAs and LCFAs of the rumen bacteria and milk composition traits, allowing for us to partially unravel the intermediate process from the rumen to the mammary gland. These results need to be confirmed given the functional redundancy in the rumen (Weimer, 2015), which results in many bacteria being involved in the same metabolic function.

Do the low phenotypic correlations obtained for dairy traits mean that there is no direct link between rumen microbial abundances and fine milk composition traits? Or are they due to the variance of these traits, which are strongly influenced by how they are predicted or measured in milk? In this regard, there may be two ways to address this question:

1. Perform standard laboratory measurements (*e.g.* wet lab method) for milk components as in Buitenhuis *et al.* (2019). This allows for us to compare the association with rumen microbial abundances using different methodologies such as wet lab techniques, MIR spectra predictions and gas chromatography, but implies higher costs and sample processing time for a large number of animals.

2. Confirm the results obtained in this work using MIR spectra predictions and gas chromatography with a larger database. This would allow for us to obtain more precise estimates, as there are no published studies to refer to with a large number of ruminants, such as those used in our work, that include milk traits and microbial abundances.

Furthermore, we did not detect any association between rumen bacterial OTUs and the fine milk composition using a single OTU-regression analysis, *i.e.,* MWAS. These results are even less promising than the phenotypic correlations with milk FAs and proteins. We performed MWASs for all 2,059 OTUs available without preselection based on the variance explained for the two principal components (as in the sparse Partial Least-Square analysis), as for the phenotypic correlations. In addition, in the MWAS, we included an extra effect in the regression model, which is the genetic effect that controls the risk of associations based on the genetic

population structure. In general, in livestock species using varied traits, most OTUs in the MWAS failed to pass the defined significance threshold (Difford *et al.*, 2018; Gebreyesus *et al.*, 2020; Vollmar *et al.*, 2020; Aliakbari *et al.*, 2022). In our study, we applied a compositional data process of microbial abundances, which was suggested by Vollmar *et al.* (2020) to improve the detection of OTU causalities in the MWAS but we did not identify any significant signal.

7.1.2 Microbial community effect on phenotypes

Usually, the number of OTUs is larger than the number of animals sampled, so a solution could be to estimate all microbial effects simultaneously rather than OTU by OTU. For this, the rumen bacterial community could be considered through the use of an abundance matrix with animals and OTUs or through the construction of a similarity matrix that expresses the similarities between individuals in terms of microbial abundances.

First, we used the global rumen microbial abundances in order to discriminate two lines divergently selected by years for two specific traits: somatic cell score in milk (SCS lines) and the coefficient of variation of milk production (PERS lines). Despite the clear genetic differences between these lines (>2 genetic standard deviations), no discrimination was possible using the microbial community abundance. These results could mean that genetic selection occurs, which is based on the transmission of the best combination of alleles that will improve the phenotype of interest (*e.g.,* SCS or PERS), and any of those alleles are transmitted jointly with alleles associated with the rumen bacterial abundances.

Then, using a microbial similarity matrix ($\boldsymbol{O}$), we estimated the variance explained by each dairy trait by the microbiota abundance, *i.e.,* the microbiability. We obtained no phenotypic variance for milk FAs and very weak variance for milk proteins (<0.06). It can be assumed that the phenotypic correlations of dairy traits with the OTUs presented above are diluted when studying the overall microbial effect on the phenotype. These results are thus not far those obtained in dairy cows for milk FAs (Buitenhuis *et al.*, 2019), where the microbiabilities of the fine milk composition were very low and had large standard errors.

Then, the question of **what impact does the rumen microbiota have on the phenotypic variability of dairy traits was examined**. The rumen microbial abundance has a weak association with dairy traits, mainly evidenced by a group of OTUs with some significant phenotypic correlations, but these OTU effects on the phenotype are diluted when using the microbial community as a whole. It is also noteworthy that rumen VFAs and LCFAs were more closely associated with OTUs than fine milk composition traits, although these results should

be confirmed. The first step in considering the potential use of genetic selection on the traits that showed an impact of the rumen microbiota, however low, is to evaluate the genetic control on microbial composition.

7.2 What degree of control do host genetics have over microbial composition?

The main results obtained from the genetic analyses were integrated to determine the host genetic control on the microbial composition in the rumen of dairy sheep (Figure 16).

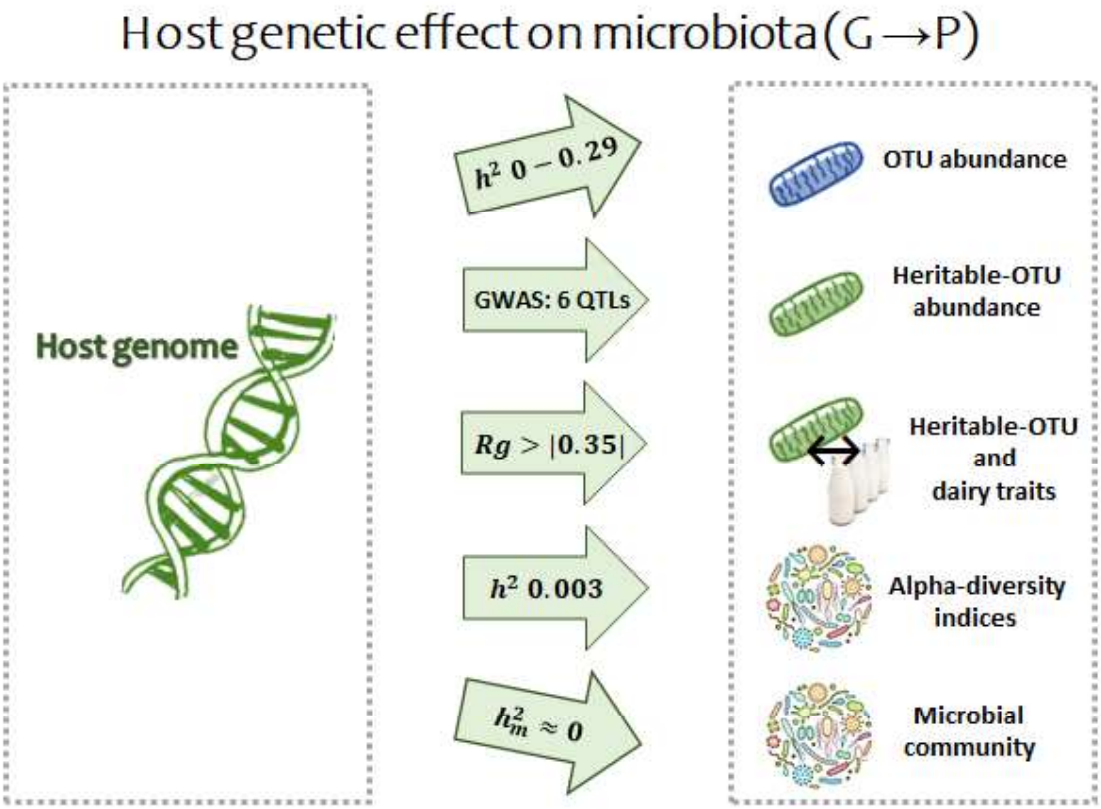


Figure 16. Main results obtained for host genetic effect on individual and community microbial composition. OTU: operational taxonomic unit, Heritable-OTU: operational taxonomic units that with a heritability significantly different from zero (N=306 OTUs), Rg: genetic correlation, h^2: heritability, h^2_m: heritability of the microbial community. For the genome-wide association study (GWAS) of OTUs with significant genetic variance were identified six quantitative trait loci (QTLs).

7.2.1 Host genetic control on OTU abundance

The rumen bacteria represented by 2,059 OTUs, without applying the rarefication process, presented a proportion of 15% (306 OTUs) with low to moderate host genetic control (h^2 from

0.10 to 0.29). These results are higher than the 2%, 4%, 6% and 8% presented in cows by Zhang *et al.* (2020), Zang *et al.* (2022), Difford *et al.* (2018) and Wallace *et al.* (2019), respectively, and there is thus far no reference in sheep. Therefore, in general, in the rumen microbiota, there are a large number of OTUs with no genetic control and a few OTUs with moderate to high genetic control; in our work, the maximum heritability was 0.29, but for some other authors, it was as high as 0.60 (Wallace *et al.*, 2019).

Considering GWASs, we determined the abundance of each OTU and found that the host genetic control was very low. In fact, major effects on OTU abundances were detected on only four autosomal chromosomes, and many of the signals found throughout the host genome were unclear. The low number of QTLs for the OTU abundances detected contributes to the hypothesis of Pérez-Enciso *et al.* (2021) that it is very difficult to find causative SNPs for bacterial abundances with GWAS analysis alone. First, numerous factors may have biased the results obtained, such as the number of GWASs and independent statistical analyses performed with a relatively small dataset of ~800 ewes. Moreover, the low heritability of OTUs leads to problems in estimating the effects and variances of SNPs with the methodology of single-step GWAS (Aguilar *et al.*, 2019).

The genetic links between OTU abundances with significant genetic variance and dairy traits were high (|0.35–0.99|), and for the SCS were moderate (|0.21-0.68|). However, a very small proportion of the microbiota (one-third of heritable OTUs *i.e.,* 5% of the total) showed a significant correlation with traits, even if they were affiliated with the most abundant microbial families and genera in the rumen. The fact that the most important genera are represented is misleading, since many OTUs belonging to these genera presented opposite correlations with the same dairy trait in many cases.

Despite the methodological limits, the partial genetic control of the host over bacterial abundances could be explained by the functional redundancy in the rumen. We hypothesize that the host controls key microbial functions in the rumen that have a role in its metabolism rather than OTU abundance. Therefore, from one animal to another, different OTUs could be involved for the same function, which would produce that the variance of this microbial function is divided in the abundance of different OTUs, explaining the weak relationship between abundances and traits. In this regard, we attempted to determine the metabolic functions by which the host controls over the microbial abundances through the QTL identification of OTU abundances. Most of the microbial functions detected were associated with the immune system.

Future work could identify microbial functions that are not possible to determine using short sequencing reads, as they do not contain enough information to support highly resolved phylogenetic classification. A promising option could be the use of short-read sequencers to generate accurate synthetic long reads. This sequencing technology (LoopSeq 16S Long Read) amplifies the full-length 16S rRNA gene and achieves enough accuracy for species-level assignment from the 16S rRNA gene (Callahan *et al.*, 2021). Another option is to use the high-dimensional data from shotgun sequencing, as Martínez-Álvaro *et al.* (2022) did, which allowed for them to work with microbial functions and estimate the heritabilities of microbial genes, for which they obtained higher values than for the abundances (27% of microbial genes with h^2 from 0.2 to 0.6). However, and despite being finer than 16S rRNA gene sequencing for the study of microbial functions and LoopSeq 16S when large amounts of nontarget DNA are present in the samples, shotgun sequencing costs five times more per sample (133.6€ per sample) than targeting the 16S rRNA gene.

7.2.2 Host genetic control over the microbial community

The genetic control of the host on OTU abundance was low to moderate, and if we consider the rumen bacterial community as a whole, this genetic control was even lower. In general, the results showed zero heritability of alpha-diversity indices and no genetic control over the microbial community estimated with the Christensen *et al.* (2021) method for different dairy traits.

First, an alpha-diversity index provides a measure of species diversity within each sample, thus allowing for all OTUs to be condensed into a single value per animal. As higher microbial diversity is often considered an attribute of gut health (Fouhse *et al.*, 2016), some authors considered these measures for potential use in genetic selection for associated traits in monogastric species (Aliakbari *et al.*, 2021; Déru *et al.*, 2022). However, in our work, the results obtained for Simpson's and Shannon's indices were zero, which differs from published results using nonrecommended preprocessing of microbial abundances such as rarefaction (McMurdie and Holmes, 2014). For the alpha-diversity indices in dairy cows, F. Li *et al.* (2019) obtained a heritability of 0.19 ± 0.08 for the Simpson index and 0.23 ± 0.09 for the Shannon index. In addition, for beta-diversity indices such as Bray-Curtis, for the two first principal coordinates of the dissimilarity matrix, Difford *et al.* (2018) obtained a heritability of 0.20 ± 0.10 and 0.02 ± 0.05, and F. Li *et al.* (2019) obtained 0.12 ± 0.07 and 0.25 ± 0.09.

Second, as mentioned above, the microbial effect on dairy traits predicted using a microbial similarity matrix was equal to or very close to zero for all traits. Then, we estimated the genetic control over the microbial community as the heritability of the predicted microbial effect, and it was similar to the mean of 0.04 ± 0.03 obtained for the heritability of the 2,059 OTUs. However, in this work, we reveal a limitation of the Christensen *et al.* (2021) methodology; when the model uses the microbial effect predicted from the previous model as the phenotype and this derived phenotype has very low variance (values close to zero for all animals), then the model does not run.

The results obtained from community analyses are less promising than for individual OTUs, which could be explained by the information contained in the microbial similarity matrix. This microbial matrix was proposed by Ross *et al.* (2013b), and it is the most widely used for microbiome studies in different livestock species (Camarinha-Silva *et al.*, 2017; Difford *et al.*, 2018; Vollmar *et al.*, 2020; Khanal *et al.*, 2021). In this work, we evaluated different ways of constructing the microbial similarity matrix ($\boldsymbol{O}$) such as the use of CLR-transformed abundances of all OTUs or only the 'heritable' ones (N=306), but in any case, we observed significant differences in microbiabilities for the beta-lactoglobulin. Thus, whether the microbial matrix focuses only on the 306 OTUs or considers all OTUs did not change the estimated effects, meaning that OTUs with significant heritability do not have a differential impact on the phenotype compared to the others. One additional element to consider in the microbial matrix is the information that it contains. We know that the microbiota composition is highly impacted by the surrounding environment; thus, the lack of host genetic control over the microbial community could be explained by the fact that the similarity between individuals could be mostly due to the shared environment between the contemporary group of animals. For example, He *et al.* (2022) obtained greater microbiabilities in pigs when they did not include pen and sire effects in the prediction models. The way to reduce this phenomenon is to limit the experimental variation factors and, if not controlled, record them for correction by statistical models. However, the sources of variation are so large and there is possible horizontal microbial transmission between animals, thus it would be impossible to eliminate the environmental effects.

What degree of control do host genetics have over microbial composition? Host genetics have low to moderate genetic control over a small group of OTUs in the rumen which are also associated with dairy traits and udder health traits. However, the use of microbial community information for breeding purposes is not promising due to the null results obtained in this work.

7.3 How could the holobiont concept be implemented in animal breeding?

The results obtained in this thesis allow for us to better understand the relationship between the rumen microbiota and its host from a phenotypic and genetic points of view, as well as to elaborate possible scenarios for the use of microbiota in animal breeding.

7.3.1 Genetic selection of microbiota

Genetic selection for a particular rumen microbial community could be seen as an opportunity to improve the phenotype of interest. To date, the only published experience of selection on microbial genera is from Larzul *et al.* (2021). The authors conducted a two-generation directional selection based on the abundance of four bacterial genera. These genera are part of two gut enterotypes in pigs associated with growth traits, such as *Prevotella* and *Mitsuokella* and *Ruminococcus* and *Treponema*. This study shows that genetic selection for bacterial genus abundance is possible based on the moderate heritability estimated (h^2 from 0.3 to 0.4) and the genetic differences observed for those genera under selection after two generations. However, some points deserve consideration. The selection criteria used by Larzul *et al.* (2021) were well defined in a previous study that evidenced these two enterotypes in the pig intestine (Ramayo-Caldas *et al.*, 2016). However, beyond the four genera, we know little about the impact of this selection on the microbial community or how to carry out such experiments in nonuniform environmental conditions?

In dairy ewes, the search for ruminotypes has not yet been conducted. However, the zero heritability of alpha-diversity indices and the low to moderate heritability of rumen bacteria, of which only the *Prevotella* genus was identified as abundant and overrepresented over heritable OTUs, as well as the few OTUs genetically associated with fine milk composition do not allow for us to consider selection on some specific OTUs or genera considering only genetic effects.

7.3.2 Short-term changes in microbiota composition

Considering that the microbiota is highly influenced by environmental effects and is dynamic throughout the animal's life, we could imagine, beyond the selection, making innovative short-term changes in the composition of the microbiota. This means that at any time of the animal's life, a rapid change in the rumen microbial composition could occur for a short period of time to impact the phenotype of interest until the animal re-establishes the original composition. There are long-used methods to change the microbiota in the rumen, and the knowledge generated in our work indicates that it could be interesting to re-evaluate some of them (*e.g.,* rumen microbial transplantation and bacteriophages).

Rumen microbial transplantation (or also known as transfaunation) refers to transferring rumen fluid containing microbes and nutrients from one animal to another. However, to date, the studies using this technique included a low number of animals, two to eight animals (Weimer *et al.*, 2010; Zhou *et al.*, 2018; Santos *et al.*, 2021), and did not evaluate the impact on traits directly affected by the rumen bacteria. Previous experiences have shown that there is generally high interindividual variability in terms of time to restore the original microbial composition, but it commonly occurs within a period of 28 days (Zhou *et al.*, 2018) to 63 days (Weimer *et al.*, 2010). Therefore, it could be an interesting window of time to modify the rumen microbiota and thus impact the phenotype of interest. Another option is the use of bacteriophages, which are viruses that infect and kill bacterial cells through a lytic process (Patra, 2012). They were applied to control potential pathogens naturally occurring in the rumen, *e.g.*, two days before cattle slaughter, to reduce the risk of meat contamination (Rozema *et al.*, 2009; Montso *et al.*, 2021) and were also proposed to control methane emissions in cattle (Mirzaei-Aghsaghali and Maheri-Sis, 2015). However, the results obtained by the authors using bacteriophages are too variable, and further work is needed to find a suitable way to use them.

In conclusion, and based on the weak genetic control of the host over its rumen microbial composition, it would be interesting to use some of the long-used methods to rapidly improve the production of rumen traits directly associated with the microbiota. As an experimental scenario, we could impact methane emissions for a short time by rumen microbial transplantation of animals from very efficient lines.

7.3.3 Selection of microbiota as a nongenetic effect

For years, animal selection schemes have been performed using only genetic effects, where genes are inherited by the transmission of DNA material that determines the expression of phenotype. Nongenetic sources of inheritance (*e.g.,* epigenetics, microbiota, behavior effects) have thus far not been considered in the selection of livestock species, which today is a limitation to achieving the maximum progress in selection (David and Ricard, 2019). One of the first references is Tal *et al.* (2010), who proposed a mathematical model that was later used by Varona *et al.* (2015) to estimate the variance components of epigenetic marks. More recently, David and Ricard (2019) proposed a transmissibility model to estimate the transmissible values instead of EBVs, including all sources of genetic and nongenetic variance, such as the microbiota composition.

Then, the microbiota could also be considered a nongenetic effect that is physically transmitted across generations (Sandoval-Motta *et al.*, 2017), contributing to the phenotypic resemblance between relatives. For example, direct physical contact with the mother during birth and during the suckling period is crucial to the establishment of the microbial community in ruminants (Dehority, 2004). However, the mechanisms of microbiota transmission or the possibility of simultaneously enhancing the transmission of microbiota effects controlled and not controlled by host genetics are not discussed in a recent review by Pérez-Enciso *et al.* (2021). We proposed to include this transmissible subset of OTUs that are not genetically controlled but have an impact on the phenotype, as well as the subset of OTUs under genetic control that impact the phenotype.

As we now know that microbiota has host genetic control, though only 15% of the rumen microbiota in our study, is it possible to select microbiota for the nongenetic but transmissible microbial effect in animal breeding? This will depend on three factors: the transmission fidelity of the microbiota due to the variety of transmission mechanisms, the environmental colonization and within-host proliferation during development periods of the host, and the number of microbial generations within one host generation (Bruijning *et al.*, 2022). Then, considering the proposal of David *et al.* (2019), we can perform an intervention at a specific moments in the animal's life, such as birth, to generate a change in the composition of the rumen microbiota of the mother that will be transmitted vertically to the offspring. In this way, we can have an impact on the expression of the phenotype of future generations.

Chapter 8

Conclusions and perspectives

The analyses performed in this thesis produced multiple results that have greatly improved our knowledge of the dairy sheep holobiont. First, we defined a suitable method to work with microbial abundances considering the compositional nature of the microbiome community and its high susceptibility to environmental factors. Host genetic selection based on rumen bacterial composition for milk quality improvement is not promising due to mainly null microbiability as well as the low number of OTUs genetically controlled and genetically associated (even highly) with the fine milk composition and the absence of genomic regions in common between bacterial abundances and dairy traits. Milk quality could not be improved through host genetic selection for microbiota-mediated effects. However, if the microbiota had influenced milk traits and was transmissible, we could select for those traits even if the microbiota is not genetically controlled. Knowledge about the relationship between the host, rumen microbiome and traits will allow for us to assess the future use of "holobiont" breeding.

Based on the data in this project, the **perspectives** for future work are as follows:

The microbial abundances were useful as a first approach to the microbiome analysis. However, they contain redundant information about functions in the rumen, then, it might be better to use microbial functions instead of abundances. In our study, we evidenced through GWASs of bacterial abundances that host genome was associated through genes that were involved in immune system pathways. Then, an interesting point to explore could be through microbial functions confirm these association with the immune system more precisely, and evidence potential associations of rumen bacteria with markers of disease or immune system measures.

The microbiome is a complex and dynamic community interacting with its host and environment. At present, most studies provide a 'snapshot' of the microbial population present at the time of collection, so longitudinal microbiome studies which collected data across multiple time points is highly recommended. In this way, these studies capture temporal variation within microbiome to gain mechanistic insights in microbial system, and to assess

phenotypic and genetic associations with phenotypes of interest over different periods of the animal's life.

The microbiome in the holobiont will play a fundamental role in adapting to changing environmental conditions (*e.g.,* climate change). This is because even with low host genetic control, the microbiota can improve the phenotypic means of producing specialized individuals for a production system or increase the phenotypic variance of producing animals adapted to changing environmental situations. Then, the incorporation in the genetic evaluations system of nongenetic effect could be an interesting approach to refining the breeding strategies for the future agroecology by focusing on the holobiont rather than the animal.

References

Aguilar, I. *et al.* (2019) Frequentist p-values for large-scale-single step genome-wide association, with an application to birth weight in American Angus cattle. *Genetics Selection Evolution* 51: 28. https://doi.org/10.1186/s12711-019-0469-3.

Aitchison, J.A. (1986) *The Statistical Analysis of Compositional Data*. London: Chapman & Hall Ltd.

Aliakbari, A. *et al.* (2021) Genetic relationships between feed efficiency and gut microbiome in pig lines selected for residual feed intake. *Journal of Animal Breeding and Genetics* 138: 491–507. https://doi.org/10.1111/jbg.12539.

Aliakbari, A. *et al.* (2022) Microbiability and microbiome-wide association analyses of feed efficiency and performance traits in pigs. *Genetics Selection Evolution* 54: 29. https://doi.org/10.1186/s12711-022-00717-7.

Barillet, F. *et al.* (2016) Objectifs de sélection et stratégie raisonnée de mise en œuvre à l'échelle des populations de brebis laitières françaises. *INRA Productions Animales* 29: 19-40. https://hal.archives-ouvertes.fr/hal-01519345/

Bertan Membrive, C.M. (2016) Anatomy and Physiology of the Rumen. In *Rumenology*. Switzerland: Springer. pp. 1–38.

Blaser, M.J. (2014) The microbiome revolution. *Journal of Clinical Investigation* 124: 4162–4165. https://doi.org/10.1172/JCI78366.

Bokulich, N.A. *et al.* (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature Methods* 10: 57–59. https://doi.org/10.1038/nmeth.2276.

Bourdon, R.M. (2000) *Understanding animal breeding*. Second edition. Upper Saddle River, New Jersey: Prentice-Hall.

Bruijning, M. *et al.* (2022) Natural selection for imprecise vertical transmission in host–microbiota systems. *Nature Ecology & Evolution* 6: 77–87. https://doi.org/10.1038/s41559-021-01593-y.

Buitenhuis, B. *et al.* (2019) Impact of the rumen microbiome on milk fatty acid composition of Holstein cattle. *Genetics Selection Evolution* 51: 23. https://doi.org/10.1186/s12711-019-0464-8.

Callahan, B.J. *et al.* (2021) Ultra-accurate microbial amplicon sequencing with synthetic long reads. *Microbiome* 9: 130. https://doi.org/10.1186/s40168-021-01072-3.

Camarinha-Silva, A. *et al.* (2017) Host Genome Influence on Gut Microbial Composition and Microbial Prediction of Complex Traits in Pigs. *Genetics* 206: 1637–1644. https://doi.org/10.1534/genetics.117.200782.

Christensen, O.F. *et al.* (2021) Genetic evaluation including intermediate omics features. *Genetics* 219: iyab130. https://doi.org/10.1093/genetics/iyab130.

Collomb, M. *et al.* (2008) Seasonal variation in the fatty acid composition of milk supplied to dairies in the mountain regions of Switzerland. *Dairy Science and Technology* 88: 631–647. https://doi.org/10.1051/dst:2008029.

Danchin, É. *et al.* (2011) Beyond DNA: integrating inclusive inheritance into an extended theory of evolution. *Nature Reviews Genetics* 12: 475–486. https://doi.org/10.1038/nrg3028.

David, I. *et al.* (2019) Intergenerational Transmission of Characters Through Genetics, Epigenetics, Microbiota, and Learning in Livestock. *Frontiers in Genetics* 10: 1058. https://doi.org/10.3389/fgene.2019.01058.

David, I. and Ricard, A. (2019) A Unified Model for Inclusive Inheritance in Livestock Species. *Genetics* 212: 1075–1099. https://doi.org/10.1534/genetics.119.302375.

Dehority, B.A. (2004) *Rumen microbiology*. Reprinted. Nottingham: Nottingham Univ. Press.

Déru, V. *et al.* (2022) Genetic relationships between efficiency traits and gut microbiota traits in growing pigs being fed with a conventional or a high-fiber diet. *Journal of Animal Science* 100: skac183. https://doi.org/10.1093/jas/skac183.

Difford, G.F. *et al.* (2018) Host genetics and the rumen microbiome jointly associate with methane emissions in dairy cows. *PLOS Genetics* 14: e1007580. https://doi.org/10.1371/journal.pgen.1007580.

Escudié, F. *et al.* (2018) FROGS: Find, Rapidly, OTUs with Galaxy Solution. *Bioinformatics* 34: 1287–1294. https://doi.org/10.1093/bioinformatics/btx791.

Falconer, D.S. (1986) *Introduction to Quantitative Genetics*. New York: Longman Scientific & Technical.

FAOSTAT (2020) Crops and livestock products. *Food and Agriculture Organization of the United Nations*. https://www.fao.org/faostat/en/#data/QCL (Accessed: 3 October 2022).

Ferrand, M. *et al.* (2012) Determination of protein composition in milk by mid-infrared spectrometry. *ICAR 38 Annual Meeting* 5.

Ferrand-Calmels, M. *et al.* (2014) Prediction of fatty acid profiles in cow, ewe, and goat milk by mid-infrared spectrometry. *Journal of Dairy Science* 97: 17–35. https://doi.org/10.3168/jds.2013-6648.

Fouhse, J.M. *et al.* (2016) The role of gut microbiota in the health and disease of pigs. *Animal Frontiers* 6: 30–36. https://doi.org/10.2527/af.2016-0031.

Gebreyesus, G. *et al.* (2020) Predictive ability of host genetics and rumen microbiome for subclinical ketosis. *Journal of Dairy Science* 103: 4557–4569. https://doi.org/10.3168/jds.2019-17824.

Gloor, G.B. *et al.* (2016) It's all relative: analyzing microbiome data as compositions. *Annals of Epidemiology* 26: 322–329. https://doi.org/10.1016/j.annepidem.2016.03.003.

Gloor, G.B. *et al.* (2017) Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology* 8: 2224. https://doi.org/10.3389/fmicb.2017.02224.

Goodwin, S. *et al.* (2016) Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* 17: 333–351. https://doi.org/10.1038/nrg.2016.49.

Graeve, M. and Greenacre, M.J. (2020) The selection and analysis of fatty acid ratios: A new approach for the univariate and multivariate analysis of fatty acid trophic markers in marine pelagic organisms. *Limnology and Oceanography: Methods* 18: 196–210. https://doi.org/10.1002/lom3.10360.

Greenacre, M. *et al.* (2021) Compositional Data Analysis of Microbiome and Any-Omics Datasets: A Validation of the Additive Logratio Transformation. *Frontiers in Microbiology* 12: 727398. https://doi.org/10.3389/fmicb.2021.727398.

Grieneisen, L. *et al.* (2021) Gut microbiome heritability is nearly universal but environmentally contingent. *Science* 373: 181–186. https://doi.org/10.1126/science.aba5483.

Guo, Z. *et al.* (2016) Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. *Theortical and Applied Genetics* 129: 2413–2427. https://doi.org/10.1007/s00122-016-2780-5.

Hayes, B.J. *et al.* (2017) Accelerating wheat breeding for end-use quality with multi-trait genomic predictions incorporating near infrared and nuclear magnetic resonance-derived phenotypes. *Theortical and Applied Genetics* 130: 2505–2519. https://doi.org/10.1007/s00122-017-2972-7.

He, Y. *et al.* (2022) Use of Host Feeding Behavior and Gut Microbiome Data in Estimating Variance Components and Predicting Growth and Body Composition Traits in Swine. *Genes* 13: 767. https://doi.org/10.3390/genes13050767.

Henderson, G. *et al.* (2013) Effect of DNA Extraction Methods and Sampling Techniques on the Apparent Structure of Cow and Sheep Rumen Microbial Communities. *PLOS ONE* 8: e74787. https://doi.org/10.1371/journal.pone.0074787.

Henry, L.P. *et al.* (2021) The microbiome extends host evolutionary potential. *Nature Communications* 12: 5141. https://doi.org/10.1038/s41467-021-25315-x.

Hess, M.K. *et al.* (2020) A restriction enzyme reduced representation sequencing approach for low-cost, high-throughput metagenome profiling. *PLOS ONE* 15: e0219882. https://doi.org/10.1371/journal.pone.0219882.

Hobson, P.N. and Stewart, C.S. (1997) *The Rumen Microbial Ecosystem*. Dordrecht: Springer Netherlands.

Hu, T. *et al.* (2021) Next-generation sequencing technologies: An overview. *Human Immunology* 82: 801–811. https://doi.org/10.1016/j.humimm.2021.02.012.

Hungate, R. (1966) *The Rumen and Its Microbes*. New York: Acaemic Press.

Hurtaud, C. *et al.* (1993) Effect of Infused Volatile Fatty Acids and Caseinate on Milk Composition and Coagulation in Dairy Cows. *Journal of Dairy Science* 76: 3011–3020. https://doi.org/10.3168/jds.S0022-0302(93)77640-7.

Khanal, P. *et al.* (2021) Microbiability of meat quality and carcass composition traits in swine. *Journal of Animal Breeding and Genetics* 138: 223–236. https://doi.org/10.1111/jbg.12504.

Lagriffoul, G. *et al.* (2016) Panorama de la production de lait de brebis en France et son évolution depuis 50 ans. *INRA Productions Animales* 29: 7-18. https://hal.inrae.fr/hal-02638129

Larroque, H. *et al.* (2017) Imputation accuracy from a low density SNP panel in 5 dairy sheep breeds in France. *Annual Meeting of the European Association for Animal Production (EAAP)*, Tallinn, Estonia, 513. https://hal.inrae.fr/hal-02737593.

Larzul, C. *et al.* (2021) Response to selection on fecal microbiota composition in Large White piglets. *ISAG*, France. https://hal.archives-ouvertes.fr/hal-03464516.

Legarra, A. (2016) Comparing estimates of genetic variance across different relationship models. *Theoretical Population Biology* 107: 26–30. https://doi.org/10.1016/j.tpb.2015.08.005.

Léonil, J. *et al.* (2013) Les structures supramoléculaires du lait : structure et impact nutritionnel de la micelle de caséine et du globule gras. *INRA Productions Animales* 26: 129–144. https://doi.org/10.20870/productions-animales.2013.26.2.3142.

Li, F. *et al.* (2019) Host genetics influence the rumen microbiota and heritable rumen microbial features associate with feed efficiency in cattle. *Microbiome* 7: 92. https://doi.org/10.1186/s40168-019-0699-1.

Li, Z. *et al.* (2019) Integrating Gene Expression Data Into Genomic Prediction. *Frontiers in Genetics* 10: 126. https://doi.org/10.3389/fgene.2019.00126.

López-García, A. *et al.* (2022) Fungal and ciliate protozoa are the main rumen microbes associated with methane emissions in dairy cattle. *GigaScience* 11: giab088. https://doi.org/10.1093/gigascience/giab088.

Marchesi, J.R. and Ravel, R. (2015) The vocabulary of microbiome research: a proposal. *Microbiome* 3: 31. https://doi.org/10.1186/s40168-015-0094-5.

Margulis, L. (1993) *Symbiosis in Cell Evolution: Microbial Communities in the Archean and Proterozoic Eons*. Second edition. New York: W.H. Freeman and Co.

Martínez-Álvaro, M. *et al.* (2022) Bovine host genome acts on rumen microbiome function linked to methane emissions. *Communications Biology* 5: 350. https://doi.org/10.1038/s42003-022-03293-0.

Martín-Fernández, J.-A. *et al.* (2015) Bayesian-multiplicative treatment of count zeros in compositional data sets. *Statistical Modelling* 15: 134–158. https://doi.org/10.1177/1471082X14535524.

Matthews, C. *et al.* (2019) The rumen microbiome: a crucial consideration when optimising milk and meat production and nitrogen utilisation efficiency. *Gut Microbes* 10: 115–132. https://doi.org/10.1080/19490976.2018.1505176.

Mazza Rodrigues, P.H. (2016) Control and Manipulation of Ruminal Fermentation. In *Rumenology*. Switzerland: Springer. pp. 157–188.

McMurdie, P.J. and Holmes, S. (2014) Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLOS Computational Biology* 10: e1003531. https://doi.org/10.1371/journal.pcbi.1003531.

Meuwissen, T.H.E. *et al.* (2001) Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics* 157: 1819–1829. https://doi.org/10.1093/genetics/157.4.1819.

Mirzaei-Aghsaghali, A. and Maheri-Sis, N. (2015) Factors affecting mitigation of methane emission from ruminants: Microbiology and biotechnology strategies. *Journal of Animal Behaviour and Biometeorology* 4: 22–31. https://doi.org/10.14269/2318-1265/jabb.v4n1p22-31.

Mohapatra, A. *et al.* (2019) Sheep milk: A pertinent functional food. *Small Ruminant Research* 181: 6–11. https://doi.org/10.1016/j.smallrumres.2019.10.002.

Montso, P.K. *et al.* (2021) An Assessment of the Viability of Lytic Phages and Their Potency against Multidrug Resistant Escherichia coli O177 Strains under Simulated Rumen Fermentation Conditions. *Antibiotics* 10: 265. https://doi.org/10.3390/antibiotics10030265.

Morgante, F. *et al.* (2020) Leveraging Multiple Layers of Data To Predict *Drosophila* Complex Traits. *G3 Genes|Genomes|Genetics* 10: 4599–4613. https://doi.org/10.1534/g3.120.401847.

Nagaraja, T.G. (2016) Microbiology of the Rumen. In *Rumenology*. Switzerland: Springer. pp 39–62.

OECD/FAO (2022) *OECD-FAO Agricultural Outlook 2022-2031*. Paris: OECD Publishing. https://doi.org/10.1787/f1b0b29c-en.

Ono, T. and Obata, T. (1989) A model for the assembly of bovine casein micelles from F2 and F3 subunits. *Journal of Dairy Research* 56: 453–461. https://doi.org/10.1017/S0022029900028934.

Osorio, J.S. *et al.* (2016) Biosynthesis of milk fat, protein, and lactose: roles of transcriptional and posttranscriptional regulation. *Physiological Genomics* 48: 231–256. https://doi.org/10.1152/physiolgenomics.00016.2015.

Park, P.W. and Goins, R.E. (1994) In situ preparation of fatty acid methyl esters for analysis of fatty acid composition in foods. *Journal of Food Science* 59: 1262–1266. https://doi.org/10.1111/j.1365-2621.1994.tb14691.x.

Park, Y.W. *et al.* (2007) Physico-chemical characteristics of goat and sheep milk. *Small Ruminant Research* 68: 88–113. https://doi.org/10.1016/j.smallrumres.2006.09.013.

Patra, A.K. (2012) Enteric methane mitigation technologies for ruminant livestock: a synthesis of current research and future directions. *Environmental Monitoring and Assessment* 184: 1929–1952. https://doi.org/10.1007/s10661-011-2090-y.

Pérez-Enciso, M. *et al.* (2021) Opportunities and limits of combining microbiome and genome data for complex trait prediction. *Genetics Selection Evolution* 53: 65. https://doi.org/10.1186/s12711-021-00658-7.

Pitta, D.W. *et al.* (2014) Temporal dynamics in the ruminal microbiome of dairy cows during the transition period. *Journal of Animal Science* 92: 4014–4022. https://doi.org/10.2527/jas.2014-7621.

Playne, M.J. (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. https://doi.org/10.1002/jsfa.2740360803.

Poretsky, R. *et al.* (2014) Strengths and Limitations of 16S rRNA Gene Amplicon Sequencing in Revealing Temporal Microbial Community Dynamics. *PLOS ONE* 9: e93827. https://doi.org/10.1371/journal.pone.0093827.

Quast, C. *et al.* (2012) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* 41: D590–D596. https://doi.org/10.1093/nar/gks1219.

Quince, C. *et al.* (2017) Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology* 35: 833–844. https://doi.org/10.1038/nbt.3935.

Ramayo-Caldas, Y. *et al.* (2016) Phylogenetic network analysis applied to pig gut microbiota identifies an ecosystem structure linked with growth traits. *The ISME Journal* 10: 2973–2977. https://doi.org/10.1038/ismej.2016.77.

Ramos-Morales, E. *et al.* (2014) Use of stomach tubing as an alternative to rumen cannulation to study ruminal fermentation and microbiota in sheep and goats. *Animal Feed Science and Technology* 198: 57–66. https://doi.org/10.1016/j.anifeedsci.2014.09.016.

Raynal-Ljutovac, K. *et al.* (2008) Composition of goat and sheep milk products: An update. *Small Ruminant Research* 79: 57–72. https://doi.org/10.1016/j.smallrumres.2008.07.009.

Riedelsheimer, C. *et al.* (2012) Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nature Genetics* 44: 217–220. https://doi.org/10.1038/ng.1033.

Ritchie, M.D. *et al.* (2015) Methods of integrating data to uncover genotype–phenotype interactions. *Nature Genetics* 16: 85-97. https://doi.org/10.1038/nrg3868.

Ross, E.M. *et al.* (2013) Investigating the effect of two methane-mitigating diets on the rumen microbiome using massively parallel sequencing. *Journal of Dairy Science* 96: 6030–6046. https://doi.org/10.3168/jds.2013-6766.

Ross, E.M. *et al.* (2013b) Metagenomic Predictions: From Microbiome to Complex Health and Environmental Phenotypes in Humans and Cattle. *PLOS ONE* 8: e73056. https://doi.org/10.1371/journal.pone.0073056.

Rothschild, D. *et al.* (2018) Environment dominates over host genetics in shaping human gut microbiota. *Nature* 555: 210–215. https://doi.org/10.1038/nature25973.

Rothschild, D. *et al.* (2022) An atlas of robust microbiome associations with phenotypic traits based on large-scale cohorts from two continents. *PLOS ONE* 17: e0265756. https://doi.org/10.1371/journal.pone.0265756

Rozema, E.A. *et al.* (2009) Oral and Rectal Administration of Bacteriophages for Control of Escherichia coli O157:H7 in Feedlot Cattle. *Journal of Food Protection* 72: 241–250. https://doi.org/10.4315/0362-028X-72.2.241.

Rupp, R. *et al.* (2003) Genetic Parameters for Milk Somatic Cell Scores and Relationships with Production Traits in French Lacaune Dairy Sheep. *Journal of Dairy Science* 86: 1476–1481. https://doi.org/10.3168/jds.S0022-0302(03)73732-1.

Rupp, R. *et al.* (2009) Response to somatic cell count-based selection for mastitis resistance in a divergent selection experiment in sheep. *Journal of Dairy Science* 92: 1203–1219. https://doi.org/10.3168/jds.2008-1435.

Rupp, R. *et al.* (2015) A Point Mutation in Suppressor of Cytokine Signalling 2 (Socs2) Increases the Susceptibility to Inflammation of the Mammary Gland while Associated with Higher Body Weight and Size and Higher Milk Production in a Sheep Model. *PLOS Genetics* 11: e1005629. https://doi.org/10.1371/journal.pgen.1005629.

Sanchez, M.-P. *et al.* (2019) Analyse génétique de la « fromageabilité » du lait de vache prédite par spectrométrie dans le moyen infrarouge en race Montbéliarde. *INRA Productions Animales* 32: 379–398. https://doi.org/10.20870/productions-animales.2019.32.3.2950.

Sandoval-Motta, S. *et al.* (2017) The Human Microbiome and the Missing Heritability Problem. *Frontiers in Genetics* 8: 80. https://doi.org/10.3389/fgene.2017.00080.

Sanger, F. *et al.* (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* 74: 5463–5467. https://doi.org/10.1073/pnas.74.12.5463.

Santos, A. *et al.* (2020) Computational methods for 16S metabarcoding studies using Nanopore sequencing data. *Computational and Structural Biotechnology Journal* 18: 296–305. https://doi.org/10.1016/j.csbj.2020.01.005.

Santos, B.P. *et al.* (2021) Transfaunation of the ruminal fluid from cows alters ruminal microbiota structure but not dominant protozoa in healthy sheep. *Small Ruminant Research* 194: 106283. https://doi.org/10.1016/j.smallrumres.2020.106283.

Shapira, M. (2016) Gut Microbiotas and Host Evolution: Scaling Up Symbiosis. *Trends in Ecology & Evolution* 31: 539–549. https://doi.org/10.1016/j.tree.2016.03.006.

Tal, O. *et al.* (2010) Epigenetic Contribution to Covariance Between Relatives. *Genetics* 184: 1037–1050. https://doi.org/10.1534/genetics.109.112466.

Tiezzi, F. *et al.* (2021) Gut microbiome mediates host genomic effects on phenotypes: a case study with fat deposition in pigs. *Computational and Structural Biotechnology Journal* 19: 530–544. https://doi.org/10.1016/j.csbj.2020.12.038.

Varona, L. *et al.* (2015) A Bayesian Model for the Analysis of Transgenerational Epigenetic Variation. *G3 Genes|Genomes|Genetics* 5: 477–485. https://doi.org/10.1534/g3.115.016725.

Vollmar, S. *et al.* (2020) The Gut Microbial Architecture of Efficiency Traits in the Domestic Poultry Model Species Japanese Quail (*Coturnix japonica*) Assessed by Mixed Linear Models. *G3 Genes|Genomes|Genetics* 10: 2553–2562. https://doi.org/10.1534/g3.120.401424.

Wallace, R.J. *et al.* (2019) A heritable subset of the core rumen microbiome dictates dairy cow productivity and emissions. *Science Advances* 5: eaav8391. https://doi.org/10.1126/sciadv.aav8391.

Wang, J. and Jia, H. (2016) Metagenome-wide association studies: fine-mining the microbiome. *Nature Reviews* 14: 508-522. https://doi.org/10.1038/nrmicro.2016.83

Wang, Y. and LêCao, K.-A. (2019) Managing batch effects in microbiome data. *Briefings in Bioinformatics* 21: 1-17. https://doi.org/10.1093/bib/bbz105.

Weimer, P.J. *et al.* (2010) Host specificity of the ruminal bacterial community in the dairy cow following near-total exchange of ruminal contents. *Journal of Dairy Science* 93: 5902–5912. https://doi.org/10.3168/jds.2010-3500.

Weimer, P.J. (2015) Redundancy, resilience, and host specificity of the ruminal microbiota: implications for engineering improved ruminal fermentations. *Frontiers in Microbiology* 6: 296. https://doi.org/10.3389/fmicb.2015.00296.

Weishaar, R. *et al.* (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. https://doi.org/10.1111/jbg.12447.

Wen, C. *et al.* (2019) The gut microbiota is largely independent of host genetics in regulating fat deposition in chickens. *The ISME Journal* 13: 1422–1436. https://doi.org/10.1038/s41396-019-0367-2.

Whipps, J.M. *et al.* (1988) Mycoparasitism and plant disease control. In *Fungi in Biological Control Systems*. Manchester: University Press. pp. 161–187.

Xue, M. *et al.* (2018) Assessment of Rumen Microbiota from a Large Dairy Cattle Cohort Reveals the Pan and Core Bacteriomes Contributing to Varied Phenotypes. *Applied and Environmental Microbiology* 4: e00970-18. https://doi.org/10.1128/AEM.00970-18.

Zang, X.-W. *et al.* (2022) Heritable and Nonheritable Rumen Bacteria Are Associated with Different Characters of Lactation Performance of Dairy Cows. M*systems* 14: e00422-22. https://doi.org/10.1128/msystems.00422-22.

Zened, A. *et al.* (2011) Starch and oil in the donor cow diet and starch in substrate differently affect the in vitro ruminal biohydrogenation of linoleic and linolenic acids. *Journal of Dairy Science* 94: 5634–5645. https://doi.org/10.3168/jds.2011-4491.

Zhang, Q. *et al.* (2020) Bayesian modeling reveals host genetics associated with rumen microbiota jointly influence methane emission in dairy cows. *The ISME Journal* 14: 2019–2033. https://doi.org/10.1038/s41396-020-0663-x.

Zhao, T. *et al.* (2022) Extend mixed models to multilayer neural networks for genomic prediction including intermediate omics data. *Genetics* 221: iyac034. https://doi.org/10.1093/genetics/iyac034.

Zhou, M. *et al.* (2018) Assessment of microbiome changes after rumen transfaunation: implications on improving feed efficiency in beef cattle. *Microbiome* 6: 62. https://doi.org/10.1186/s40168-018-0447-y.

Zhou, X. and Stephens, M. (2014) Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods* 11: 407–409. https://doi.org/10.1038/nmeth.2848.

Zilber-Rosenberg, I. and Rosenberg, E. (2008) Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution. *FEMS Microbiology Reviews* 32: 723–735. https://doi.org/10.1111/j.1574-6976.2008.00123.x.

Zilber-Rosenberg, I. and Rosenberg, E. (2021) Microbial-driven genetic variation in holobionts. *FEMS Microbiology Reviews* 45: fuab022. https://doi.org/10.1093/femsre/fuab022.

Training and education during the PhD thesis

Year	Training course	Hours
2022	• Programming and computer algorithms in animal breeding with focus on single-step GBLUP and reality of genomic selection. UGA, USA, 16th May - 5th Jun. 2022. Dr. Ignacy Misztal and Collaborators.	80
2021	• Biometrical Methods in Breeding. Norwegian University of Life Science, Norway (Online), 9th Sept. - 6th Dec. 2021. Dr. Tormod Ådnøy.	30
	• Writing and presenting scientific papers. 72º Annual Meeting of European Federation of Animal Science, Davos, Switzerland, 29th Aug. 2021.	8
	• TOEIC score 870, 20th Apr. 2021.	
	• Python course. INRAE - Themanis, Toulouse, France, 29th - 31th Mar. 2021.	21
	▪ MOOC: Speak English Professionally: In person, online and on phone. Georgia Tech Language Institute, 19th Feb. 2021.	12
	▪ Ethics and Scientific integrity. Université Fédérale de Toulouse Midi Pyrénées, Toulouse, France, 10th Feb 2021.	6
2020	▪ French as a foreign language course, level A2. Alliance Française, Toulouse, France, 2nd Nov. 2020.	18
	▪ Linux, Modify and extract information from large text files and AWK programming. Genotoul bioinformatics platform, INRAE Toulouse, France, 12th - 14th Oct. 2020.	24
	▪ Compositional data analysis with CoDaPack and R. Universitat de Girona, Spain, 19th Jun. - 10th Sept. 2020. Dr. Josep Antoni Martin-Fernandez and CoDa group.	25
	▪ French as a foreign language course, level A1. Alliance Française, Toulouse, France, 6th Feb. 2020.	18
	Total hours of training courses	242

Publications during the PhD thesis

Peer-reviewed journal articles:

Year	Authors – Publication title - Journal
2022	**Martinez Boggio**, G., Christensen, O.F., Legarra, A., Meynadier, A., Marie-Etancelin, C. Microbiability of milk composition and genetic control of microbiota effects in sheep. Journal of Dairy Science (In review).
	Martinez Boggio, G., Marie-Etancelin, C., Menras, J-M., Tomas, R., Chemit, M-L., Gabinaud, B., Pascal, G., Meynadier, A. A large database linking the rumen bacterial composition and milk traits in Lacaune sheep. Scientific Data (Accepted).
	Martinez Boggio, G., Meynadier, A., Buitenhuis, A.J., Marie-Etancelin, C. Host genetic control on rumen microbiota and its impact on dairy traits in sheep. Genetic Selection Evolution, 54, 77.
	Fresco, S., Marie-Etancelin, C., Meynadier, A., **Martinez Boggio**, G. Variation in rumen bacteria of Lacaune dairy ewes from one week to the next. Frontiers in Microbiology, 13: 848518.
2021	Marie-Etancelin, C., Tortereau, F., Gabinaud, B., **Martinez Boggio**, G., Le Graverand, Q., Marcon, D., De Almeida, M-L., Weisbecker, J-L, Meynadier, A. 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: 3265.
	Martinez Boggio, G., Meynadier, A., Daunis-i-Estadella, P., Marie-Etancelin, C. Compositional analysis of ruminal bacteria from ewes selected for somatic cell score and milk persistency. PLOS ONE, 16(7): e0254874.

International Conferences:

Year	Authors – Publication title – Congress - Format
2022	**Martinez Boggio**, G., Christensen, O.F., Legarra, A., Allain, C., Meynadier, A., Marie-Etancelin, C. Rumen bacteria do not provide improved genetic evaluation of dairy traits in sheep. In: 12th World Congress on Genetics Applied to Livestock Production. Oral presentation.
	Martinez Boggio, G., Meynadier, A., Rupp, R., Allain, C., Buitenhuis, A.J., Marie-Etancelin, C. Evidence of genetic links between rumen bacteria and milk somatic cell score in Lacaune dairy sheep. In: 12th World Congress on Genetics Applied to Livestock Production. Poster.
2021	**Martinez Boggio**, G., Meynadier, A., Allain, C., Marie-Etancelin, C. Ruminal bacteria heritabilities: Which is the impact of different data processing? In: 72st EAAP Annual Meeting. Poster.
2020	**Martinez Boggio**, G., Meynadier, A., Rupp, R., Larroque, H., Allain, C., Marie-Etancelin, C. Effects of Lacaune ewe selection for somatic cell score and milk persistency on rumen bacteria. In: 71st EAAP Annual Meeting. Oral presentation.

Résumé

La sélection animale a montré ces dernières années un intérêt croissant pour l'utilisation des informations du microbiote digestif étant donné l'association avec les caractères physiologiques de l'hôte. Dans ce contexte, les hypothèses de cette thèse sont : 1) le microbiote du rumen influence les caractéristiques de composition fine du lait et de la santé de la mamelle, et 2) le microbiote ruminal est contrôlé par les gènes de l'hôte ce qui affecte les cararactères de composition du lait et de santé de la mamelle. Pour répondre à ces hypothèses, nous avons exploré la relation entre le génome de l'hôte et le métagénome ruminal dans la modélisation de la variabilité des phénotypes chez la brebis laitière. Le projet "MicroGenOL" était basé sur 795 brebis Lacaune provenant de deux lignées génétiques élevées dans les mêmes conditions de logement et d'alimentation à l'Unité Expérimentale INRAE de La Fage. Pour tous les animaux, la composition fine du lait a été prédite par les spectres dans le moyen infrarouge, et pour 118 brebis, les acides gras (AG) ont été mesurés par chromatographie en phase gazeuse. En outre, tous les animaux ont été génotypés à l'aide d'une puce SNP 54k, et un prélèvement de rumen a été effectué pour tous les animaux, ou deux à une semaine d'intervalle pour 118 brebis. Le gène de l'ARN 16S a été séquencé pour tous les échantillons de jus de rumen. Pour travailler avec les abondances microbiennes en raison de leur nature compositionnelle, un process de traitement a été défini : après l'analyse bioinformatique des séquences brutes, les dénombrements des Unités Taxonomiques Opérationnelles (OTU) ont été transformés en logratios (après imputation des zéros) et corrigés par l'environnement en utilisant des modèles linéaires. La variabilité des abondances microbiennes d'une semaine à l'autre et l'impact du traitement des données sur les estimations génétiques nous ont permis de définir la meilleure façon de travailler et les limites de l'interprétation des résultats. Les premières analyses ont porté sur l'influence du microbiote du rumen sur le phénotype. Nous avons trouvé des associations phénotypiques faibles (<|0.20|) avec les AG et les protéines du lait, et modérées (|0.20-0.50|) avec les AG du rumen. De plus, la matrice microbienne explique un très faible pourcentage de la variabilité fine de la composition du lait (microbiabilité nulle pour les AG et inférieure à 0,06 pour les protéines). L'utilisation des données du microbiote dans son ensemble n'a pas permis de mettre en évidence des relations avec aucun des caractères laitiers, alors que certains OTU sont apparus associés aux AG du rumen et dans une moindre mesure aux AG et protéines du lait. En ce qui concerne l'existence de liens entre le microbiote ruminal et la génétique de l'hôte, nous avons trouvé un faible impact des gènes de la brebis sur son microbiote avec seulement 15% des abondances bactériennes du rumen présentant des héritabilités significatives (0.10-0.29). Les associations génétiques pour ces OTU héritables avec la composition fine du lait étaient élevées (|0.35-0.99|) mais pour moins de 5% des OTU. Six régions du génome de l'hôte localisées sur les chromosomes 3, 5, 10 et 11 ont été détectées comme impactant des abondances bactériennes, mais aucune de ces régions n'était commune avec celles des caractères laitiers. Enfin, comme la plupart des OTU n'était pas contrôlée par la génétique de l'hôte et que les liens entre le microbiote et la composition fine du lait étaient faibles, la sélection génétique des brebis pour l'abondance microbienne ne semble pas prometteuse. Cependant, la prise en compte de la transmissibilité d'effets non génétiques portés par le microbiote pourrait être une opportunité intéressante pour augmenter le progrès génétique. Ce projet a ouvert de nouvelles perspectives pour l'intégration de l'hôte et de son microbiote dans la sélection animale, grâce à une meilleure compréhension de la biologie sous-jacente aux modèles.

Mots clés : Microbiote, Rumen, Génétique quantitative, Brebis laitière, Holobionte

Summary

Animal breeding has shown in recent years an increasing interest in the use of digestive microbiota information, given the association with the physiological traits of the host. In this context, the hypotheses of this thesis are: 1) the rumen microbiota influence milk composition and udder health traits, and 2) a genetic determinism exist on rumen microbiota that affects milk composition and udder health traits. To address these hypotheses, we explored the relationship between the host genome and ruminal metagenome in the modeling of dairy ewe's phenotypes. The "MicroGenOL" project was based on 795 Lacaune ewes from two genetic lines reared under the same housing and feeding conditions at the INRAE Experimental Unit of La Fage. For all animals the fine milk composition was predicted by mid-infrared spectra, and for 118 ewes, fatty acids (FAs) were measured by gas chromatography. In addition, all animals were genotyping with 54k SNP chip, and one rumen sampling for all animals, or two one week apart for 118 ewes, were collected. Rumen samples were sequencing of 16S RNA gene. To work with microbial abundances due to their compositional nature, a workflow was defined: after bioinformatics process of the raw sequences, operational taxonomic unit (OTU) counts in the abundance table were transformed in logratios (after zeros imputation) and corrected by environmental factors using linear models. The variability of microbial abundances between two weeks and the impact of data processing on genetic estimations allowed for us to defined the best way to work with, and the limits of results interpretation. The first analysis focused on the influence of the rumen microbiota on the phenotype. We found low (<|0.20|) phenotypic associations with milk FAs and proteins, and moderate (|0.20–0.50|) associations with rumen FAs. In addition, the microbial matrix explains a very low percentage of the fine milk composition variability (microbiability null for FAs and lower than 0.06 for proteins). Using the microbiota data as a whole didn't allow for evidence relationships with none of the dairy traits, but at the OTU level, associations have been demonstrated with direct products such as rumen FAs and in a lesser extent with milk FAs and proteins. Regarding the existence of host genetic control over the rumen microbiota composition, we found a low impact on the microbiota since only 15% of rumen bacterial abundances showed significant heritabilities (0.10-0.29). The genetic associations for those heritable OTUs with the fine milk composition was high (|0.35-0.99|) but for less than 5% of OTUs. Six quantitative trait loci regions were detected for bacterial abundances in autosomal chromosomes 3, 5, 10 and 11, but none were in common with those of dairy traits. Finally, given that most OTUs were not controlled by host genetics and the weak associations between microbiota and fine milk composition, genetic selection of sheep for microbial abundance does not seem promising. However, considering the transmissibility of non-genetic effect possibly carried by the microbiota could be an interesting opportunity to increase the genetic progress. This project produced new perspectives to incorporate of the host and its microbiota in animal breeding through a better understanding of the biology behind the models.

Keywords: Microbiota, Rumen, Quantitative genetics, Dairy sheep, Holobiont