



UNIVERSITÀ DEGLI STUDI DI MILANO

PhD Course in Veterinary and Animal Sciences (Year XXXVI)

Dipartimento di Veterinaria e Scienze Animali (DIVAS)

Bioinformatic pipelines and microbial databases
for the analysis of dairy cows 16S rRNA-gene
sequencing data: a benchmarking study.

VET/03

Chiara Gini

R12929

Supervisor:

Professor Fabrizio Ceciliani

PhD Program Coordinator:

Professor Fabrizio Ceciliani

Academic Year 2022/2023

*Knowledge
is the only treasure
that increases when shared.*

Index

Index.....	3
Index of figures.....	8
Index of tables.....	12
Abbreviations.....	13
Abstract	15
Sommario.....	16
Chapter 1: Introduction	17
1.1: The study of microbial communities and their interactions	17
1.1.1: Microbial communities: who is there?	18
1.1.1.1: Bacteria	18
1.1.1.2: Archaea and Protists.....	19
1.1.2: Omics sciences: definitions and fields of interest.....	19
1.1.3: Amplicon sequencing and shotgun sequencing.....	23
1.1.4: Sequencing generations	24
1.2: Bioinformatic analysis.....	25
1.2.1: General outline of bioinformatic pipelines of amplicon sequencing outputs	25
1.2.1.1: Clustering methods.....	26
1.2.1.2: Denoising methods	27
1.2.2: Data analysis	28
1.2.2.1: Commonly used analysis	28
1.2.2.2: Commonly used software	30
1.3: Use and usefulness of microbial communities' study through bioinformatic analysis in dairy cows.....	30
1.3.1: Focusing on the issues to find the solutions.....	30
1.3.2: The holobiont.....	31
1.3.3: Dairy cows milk microbiota.....	31
1.3.4: Dairy cows skin microbiota.....	33
1.3.5: Dairy cows rumen microbiota.....	34
1.3.6: Dairy cows fecal microbiota.....	36
Chapter 2: Potential significance of the study	37
Chapter 3: Material and methods	39

3.1: First section: testing of 1 pipeline and 1 database on 1 dataset and comparing the results with the same database in a newest version and other 3 databases (of which 1 in 2 different versions).....	39
3.1.1: Detailed description of the dataset included in the study	39
3.1.2: Sequencing applied	40
3.1.3: Detailed description of the pipelines and databases included in the study ...	42
3.1.4: Biostatistics applied	43
3.1.5: Description of the other databases included in the study.....	44
3.1.5.1: 16S-ITGDB.....	44
3.1.5.2: Greengenes (version 13.8 and 2)	44
3.1.5.3: RDP	44
3.1.5.4: SILVA version 138	45
3.2: Second section: testing of other 3 pipelines (of which 2 in 2 different approaches - clustering vs denoising) and 3 databases (of which 2 in 2 different versions) on 1 dataset.....	46
3.2.1: Other pipelines chosen details	46
3.2.1.1: MICCA	48
3.2.1.2: QIIME version 2	48
3.2.1.3: VSEARCH.....	49
3.3: Third section: testing of other 3 pipelines (of which 2 in 2 different approaches - clustering vs denoising) and 3 databases (of which 2 in different versions) on 5 datasets	50
3.3.1: Other Dataset chosen details	50
3.3.1.1: Mock community from Prodan et al., 2020 [105]	50
3.3.1.2: Mock community from Callahan et al., 2016 [106].....	51
3.3.1.3: Skin samples from the MILKQUA project	52
3.3.1.4: Rumen samples from Ranilla et al., 2023.....	53
3.3.1.5: Fecal samples from Andrés et al., 2024	54
Chapter 4: Results	56
4.1: First section: testing of 1 pipeline and 1 database on 1 dataset and comparing the results with the same database in a newest version and other 3 databases (of which 1 in 2 different versions).....	56
4.1.1: Sequencing results	56
4.1.2: Taxonomy obtained using SILVA version 132	56
4.1.2.1: Results at phylum level	56
4.1.2.2: Results at genus level.....	58
4.1.3: Alpha and beta diversity using SILVA version 132.....	61

4.1.4: Performance metrics using SILVA version 132 and with the same database in a newest version and other 3 databases (of which 1 in 2 different versions).....	63
4.1.4.1: Alpha diversity indices from 4 databases (of which 2 in 2 different versions) on the same dataset analyzed with the same pipeline	63
4.1.4.2: Beta diversity indices from 4 databases (of which 2 in 2 different versions) on the same dataset analysed with the same pipeline	65
4.2: Second section: testing of other 3 pipelines (of which 2 in 2 different approaches - clustering vs denoising) and 3 databases (of which 2 in 2 different versions) on 1 dataset.....	68
4.2.1: Filtered and unfiltered results	68
4.2.2: Correlations and P values of the comparisons among pipeline x database results.....	70
4.3: Third section: testing of other 3 pipelines (of which 2 in 2 different approaches - clustering vs denoising) and 3 databases (of which 2 in different versions) on 5 datasets	72
4.3.1: Mock communities from Prodan et al., 2020.....	72
4.3.2: Mock communities from Callahan et al., 2016 (Mockrobiota project)	78
4.3.3: Skin samples from the MILKQUA project	83
4.3.4: Rumen samples from Ranilla et al., 2023.....	85
4.3.5: Fecal samples from Andr��s et al., 2024.....	87
4.3.6: Cumulative analysis of the benchmarking results	88
Chapter 5: Discussion	91
5.1: Introduction	91
5.2: First section: testing of 1 pipeline and 1 database on 1 dataset and comparing the results with the same database in a newest version and other 3 databases (of which 1 in 2 different versions).....	93
5.3: Second section: testing of other 3 pipelines (of which 2 in 2 different approaches - clustering vs denoising) and 3 databases (of which 2 in 2 different versions) on 1 dataset.....	94
5.4: Third section: testing of other 3 pipelines (of which 2 in 2 different approaches - clustering vs denoising) and 3 databases (of which 2 in different versions) on 5 datasets	97
5.5: Limitations of the presented work.....	99
Chapter 6: Final remarks and conclusions	100
References.....	101
Appendix.....	114
Metagenomics pipelines command lines.....	114
Common steps: cutadapt and sickle	114
Cutadapt	114

Sickle	114
QIIME version 1.9	115
Joining paired-end reads.....	115
Quality filtering.....	115
OTU picking.....	115
Format conversion	116
QIIME version 2	117
Common steps.....	117
Data import	117
Joining paired-end reads.....	117
Denoising approach (classic approach)	117
Denoising	117
Classify	117
ASV table production and exportation.....	118
Clustering approach (alternative approach)	118
Clustering	118
Classify	118
OTU table production and exportation.....	118
Micca	120
Common steps.....	120
Clustering approach (closed reference approach)	120
Denoising approach (denovo unoise approach)	120
Vsearch.....	120
Joining paired-end reads.....	120
Clustering.....	121
R scripts key commands	121
Phyloseq	121
PERMANOVA.....	123
Rarefaction curves.....	123
List of publications related to the PhD	125
Published papers.....	125
Submitted manuscripts	125
Manuscripts to be submitted	125
Conference proceedings	126
Posters.....	126

List of other publication not related to the present thesis project	127
Acknowledgements	128
Ringraziamenti.....	130

Index of figures

Figure 1: From Holland and DeAngelis, 2009. The figure represent the so-called interaction compass designed by the authors, which is used for the classification of interspecific interactions based on interaction outcomes, for which the positive (+), neutral (0), or negative (-) signs represent the effect of one species on the other. The first sign represents the effect of species i on species j, and the second sign represents the effect of species j on species i.....	17
Figure 2: Figure from Berg et al. briefly representing how the microbiome encompasses microbiota, as well as metabolites, structural elements, and environmental conditions (Berg et al., 2020).	20
Figure 3: From Morgan et al., 2012. The figure gives an overview of the pipelines and outcomes of the two major sequencing approaches: amplicon and shotgun. From DNA extraction, with different pathways, both approaches can provide information regarding OTU phylogeny. Amplicon sequencing performs instead better in exposing community abundances, whereas shotgun metagenomic approach allows to more efficiently detect sequences variants.	22
Figure 4: From Ronholm et al., 2016. The figure compare the three generation of sequencing in temporal order, giving hints of the main advantages and disadvantages of each of the sequencing generation. Whereas Shotgun sequencing was highly accurate and had long reads, it was penalised by the small amount of data generated. On the other hand, Parallel sequencing increased the amount of data that could be obtained by sequencing, at cost of read length but keeping the accuracy that instead decreased again when gaining long reads within Single-molecule sequencing.	24
Figure 5: Approaches to analyse amplicon sequencing outputs. This tree summarizes the approaches applicable to sequencing outputs that need to undergo taxonomic classification: the first branching oppose clustering and denoising approach, to further investigate the clustering approach characteristics and features.	26
Figure 6: From Callahan et al., 2017. This figure aims to graphically represent the validity of de novo OTUs, closed-reference OTUs, and ASVs derived from a specific dataset.	28
Figure 7: From Liu et al., 2020. This scheme aims to give a general overview of the most used statistical analysis performed on microbiota and their visualization method. Analysis will in fact start from a taxonomic table (OTU or ASV table, depending on the approach used) and, depending on research targets and statistical tools available, deliver a variety of information.	29

Figure 8: From Guo et al., 2023. Scheme of microorganisms' probable derivation when found in dairy cows' milk, including external sources, entero-mammary and rumen-mammary pathways, and mammary resident microbes.....	33
Figure 9: From Xu et al., 2021. The figure provides an overview of the more common metabolic diseases of dairy cows' digestive system related to their microflora. Unbalance in the composition of Bacteria, Archaea, Fungi and viruses can in fact heavily influence dairy cows welfare and production.	35
Figure 10: Infographic of the thesis structure. Starting from the centre of the bull eye and going to the external circles, this figure represents how the thesis aimed to enlarge the frame of the study from considering just the variation on 1 dataset analysed with 1 pipeline and 1 database, to the same analysis swiping databases, to the inclusion of different pipelines, to the application of the pipeline + databases combination (36 of them) to other datasets.	37
Figure 11: From Caporaso et al., 2011. The figure graphically shows the protocol to be applied when using barcoded Illumina pyrosequencing, including primer annealing, amplification and sequencing starting from primer annealing sites.....	41
Figure 12: 16S Library Preparation Workflow as described in Illumina 16S Metagenomic Sequencing Library Preparation Protocol.	42
Figure 13: This organization chart shows the main steps required by each pipeline to deliver the OTU/ASV table. Pipelines with 2 different approaches included in their commands were flanked and branch accordingly. The command use is given by the bold names on the left, and pipelines are aligned by command similarity. Created with BioRender.com	47
Figure 14: From Ranilla et al., 2023. Infographic of the study regarding the evaluation of microbiota on rumen fluids combined with essential oils compounds or pure. Created with BioRender.	54
Figure 15: Boxplots of the distribution of relative abundance higher than 5% in the milk microbiota (all 64 samples together) at the phylum level.....	57
Figure 16: Significantly abundant taxa in milk microbiota at phyla level: on the left, the heatmap of the significance of taxa relative abundance, while on the right, the bar plot representing the behaviour of the significantly different abundant phyla along time points: the scale in the x-axis is the ratio of Treated against Controls. Please note that some changes are not visible due to the contraposition in scale between Proteobacteria and the other phyla. Therefore, values are reported on the bars.....	57
Figure 17: Boxplots of the distribution of relative abundance higher than 5% in the milk microbiota (all 64 samples together) at the genus level.	58

Figure 18: Significantly abundant taxa in milk microbiota at genus level: on the left, the heatmap of the significance of taxa relative abundance, while on the right, the bar plot representing the behaviour of the significantly different abundant phyla along time points: the scale in the x-axis is the ratio of Treated against Controls.....	60
Figure 19: Scatterplots for alpha diversity of milk samples evaluating treatment effect alone (top left), time effect alone (top right), and treatment + time effect (bottom left). MDS represents the Multidimensional scaling (MDS) of Bray-Curtis dissimilarities for milk samples. Multidimensional scaling (MDS) plot of Bray-Curtis distances based on the OTU table from the milk microbiota (bottom right).	62
Figure 20: Comparison of the scatterplots for alpha diversity of milk samples evaluating treatment effect alone, time effect alone, and treatment + time effect derived for every database using QIIME VERSION 1.9 pipeline on milk dataset.	64
Figure 21: Comparison of the MDS plots for beta diversity of milk samples evaluating for every database using QIIME version 1.9 pipeline on milk dataset.	66
Figure 22: Pie charts of the unfiltered results regarding the microbial composition of the milk samples belonging to the MILKQUA project at the phylum level	69
Figure 23: Pie charts of the filtered results regarding the microbial composition of the milk samples belonging to the MILKQUA project at the phylum level	69
Figure 24: Heatmap of the correlation values of each pipeline + database combination outputs against each other in the milk dataset. Color range from black (negative correlation, -1) to white (positive correlation, +1). Numbers inside the tiles are the p values of the correlation. Tiles containing a p value lower than 0.05 were contoured in red.....	70
Figure 25: Heatmap of the percentages of correspondence of family taxonomic level expected versus obtained by each pipeline + database combination on mock communities from Prodan et al., 2020.	73
Figure 26: Heatmap of the percentages of correspondence of genera expected versus obtained by each pipeline + database combination on mock communities from Prodan et al., 2020.	74
Figure 27: Rarefaction curves of the results of pipeline + database combinations over Prodan et al., 2020 mock community. Results were grouped per approach: the figure above presents clustering methods rarefaction curves, while the figure below presents denoising method rarefaction curves. Triplicates per each combination represent the 3 mock communities contained in Prodan et al., 2020 dataset.	77

Figure 28: Heatmap of the percentages of correspondence of family expected versus obtained by each pipeline + database combination on Mockrobiota mock communities.	79
Figure 29: Heatmap of the percentages of correspondence of genera expected versus obtained by each pipeline + database combination on Mockrobiota mock communities.	80
Figure 30: Rarefaction curves of the results of pipeline + database combinations over Callahan et al., 2016 mock community. Results were grouped per approach: the figure above presents clustering methods rarefaction curves, while the figure below presents denoising method rarefaction curves.	83
Figure 31: Heatmap of the correlation values of each pipeline + database combination outputs against each other in the skin dataset. Color range from black (negative correlation, -1) to white (positive correlation, +1). Numbers inside the tiles are the p values of the correlation. Tiles containing a p value lower than 0.05 were contoured in red.	84
Figure 32: Heatmap of the correlation values of each pipeline + database combination outputs against each other in the rumen dataset. Color range from black (negative correlation, -1) to white (positive correlation, +1). Numbers inside the tiles are the p values of the correlation. Tiles containing a p value lower than 0.05 were contoured in red.	86
Figure 33: Heatmap of the correlation values of each pipeline + database combination outputs against each other in the fecal dataset. Color range from black (negative correlation, -1) to white (positive correlation, +1). Numbers inside the tiles are the p values of the correlation. Tiles containing a p value lower than 0.05 were contoured in red.	87
Figure 34: Bubble plot of the counts of correlation with a percentage higher than 90, only considering pipelines.	88
Figure 35: Bubble plot of the counts of correlation with percentages higher than 90, only considering databases.	89
Figure 36: Bubble plot of the counts of correlation with a percentage higher than 90, considering pipelines and database pairings.	90

Index of tables

Table 1: Composition of the mock community from Prodan et al., 2020, obtained by the sequencing of Genomic DNA from the Microbial Mock Community B (Even, Low concentration), v5.1L (Catalog no. HM-782D, obtained through BEI Resources, NIAID, NIH as part of the Human Microbiome Project.....	51
Table 2: Composition of the mock community from Callahan et al., 2016. It was included in the Mockrobiota project as Mock-12.	52
Table 3: Table of the results obtained by testing with PERMANOVA with 999 permutations every database result when applied to QIIME version 1.9 pipeline on the milk dataset. Results include P values for significance, F values as distribution values, R squared values as determination coefficient and omega squared values as effect size values.	67
Table 4: Table of the percentages of error at genera and species level, number of OTU/ASV found, and three alpha diversity indices (Chao1, Shannon and Fisher) derived from the results of the 36 pipeline + database combinations on the mock communities from Prodan et al., 2020.	74
Table 5: Table of the percentages of error at genera and species level, number of OTU/ASV found, and three alpha diversity indices (Chao1, Shannon and Fisher) derived from the results of the 36 pipeline + database combinations on the mock communities from Callahan et al., 2016 (Mockrobiota project).	80

Abbreviations

ACE: Abundance-Based Coverage Estimator

ADG: Average Daily Gain

ANOVA: ANalysis of VAriance

ANR: French National Research Agency

ASV: Amplicon Sequence Variants

bp: base pairs

CH₄: methane

COI: Cytochrome C Oxidase Subunit I

CSIC: Consejo Superior de Investigaciones Cientificas, Spain

EO: Essential Oil

ESV: Exact Sequence Variant

F:B: Firmicutes:Bacteroidetes ratio

FAO: Food and Agriculture Organization of the United Nations

FISH: Fluorescence In Situ Hybridization

GHG: GreenHouse Gases

GS: Gas Chromatography

HTS: High-Throughput Sequencing

INRAE: France's National Research Institute for Agriculture, Food and Environment

ITS: internal transcribed space

kg CO₂ eq/kg: kg of carbon dioxide per kg

LBW: live body weight

Mb: mega bytes

MDS: MultiDimensional Scaling

MILKQUA: Milk Quality along the Dairy Chain for a Safe and Sustainable MILK

MR: milk replacer

MS: Mass Spectrometry

mtpa: millions of tons per year

MUR: Ministero dell'Università e della Ricerca, Italian Ministry for University and Research

NEO: Natural Essential Oil

NGS: Next Generation Sequencing

OTU: Operational Taxonomic Units

PBMCs: Peripheral Blood Mononuclear Cells

PCA: Principal Component Analysis

rRNA: ribosomal RNA

SCC: Somatic Cells Count
SCFA: Short-Chain Fatty Acids
SEO: Synthetic Essential Oil
SOP: Standard Operating Procedures
TCEO: *Thymus capitatus* Essential Oil
TMR: total mixed ration
tsv: tab separated values
VFA: Volatile Fatty Acids
zOTU: zero-OTU

Abstract

Background

Understanding the microbiomes in different organisms is crucial across various scientific fields. To achieve this, specialized computer programs called bioinformatic pipelines have been developed to convert raw data into meaningful biological results. However, the rapid growth of this field in dairy cows' research has led to a lack of guidelines on how these pipelines differ and what they can do.

Methods

This study aims to benchmark four different software - MICCA, QIIME (versions 1.9 and 2), and VSEARCH - by examining their features and evaluating their outputs using three different databases, two of which in two different releases. The performances of these combinations will be assessed using two mock communities and four dairy cows-related datasets: milk, skin, rumen, and feces.

Results

Mock communities' analysis outcomes focus on evaluating the pipelines' ability to accurately predict the composition of bacterial communities previously documented in the literature. Real-life dataset instead allow to investigate the influence of pipelines and databases on dataset analysis at 3 stages: firstly, a pilot experiment with 1 dataset (milk), 1 pipeline (QIIME version 1.9), and 1 database (SILVA version 132) was used to compare with the results of the same pipeline on the same dataset using other 3 different databases, of which 1 in double release and 1 as the updated version of the one previously used. Then, the application of 4 pipelines (of which 2 with the 2 different approaches) and 4 databases (of which 2 in a double release, for an overall of 36 combinations) on the same dataset, and later, the benchmarking of the same 36 combinations on other datasets will allow to gain a broader view of the effects of the pipelines of analysis and microbial databases available for dairy cows 16S rRNA-gene sequencing data.

Conclusions

Ultimately, this study aimed to establish a standardized approach and enhance decision-making in 16S sequencing analysis, thereby improving the scientific rigor of microbiome analysis studies in dairy cows. MICCA and QIIME version 2, in conjunction with the SILVA version 138 database, demonstrated the most promising results in achieving this goal.

Sommario

Contesto

La comprensione dei microbiomi in diversi organismi è cruciale in vari campi scientifici. Per raggiungere questo obiettivo, sono stati sviluppati programmi informatici specializzati chiamati pipeline bioinformatiche per convertire i dati grezzi in risultati biologici significativi. Tuttavia, la rapida crescita di questo settore nella ricerca zootecnica ha portato a una mancanza di linee guida su come queste pipeline differiscano tra di loro e come queste differenze possano influenzare i risultati finali.

Metodi

Questo studio mira a confrontare quattro diversi programmi informatici - MICCA, QIIME (versioni 1.9 e 2) e VSEARCH - esaminando le loro caratteristiche e valutando i loro risultati utilizzando tre diversi database - 16S-ITGDB, RPD, SILVA e Greengenes -, di cui due in due differenti aggiornamenti (SILVA versione 132 e versione 138, Greengenes versione 13.9 e versione 2). Le prestazioni di queste combinazioni saranno testate utilizzando due comunità fittizie e quattro set di dati relativi al bovino: latte, pelle, rumine e feci.

Risultati

I risultati dell'analisi delle comunità modello si concentrano sul valutare la capacità delle pipeline di predire con precisione la composizione delle comunità batteriche precedentemente documentate in letteratura. I dati reali, invece, consentono di indagare l'influenza delle pipeline e dei database sull'analisi dei dati in tre fasi: inizialmente, con un esperimento pilota con un set di dati (latte), una pipeline e un database, utilizzato per confrontare i risultati della stessa pipeline sullo stesso set di dati utilizzando gli altri database. Successivamente, con l'applicazione di 6 pipeline e 6 database (36 combinazioni in totale) sullo stesso set di dati (latte), e infine il confronto delle stesse 36 combinazioni su altri set di dati, per ottenere una visione più ampia sugli effetti delle pipeline di analisi e dei database microbici disponibili per i dati di sequenziamento del gene 16S rRNA del bovino.

Conclusioni

In definitiva, questo studio mira a stabilire uno standard e migliorare il processo decisionale nell'analisi di sequenziamento 16S, aumentando il rigore scientifico degli studi di analisi del microbioma nelle vacche da latte. MICCA e QIIME versione 2, in combinazione con il database SILVA versione 138, hanno mostrato i risultati più promettenti nel raggiungimento di questo obiettivo.

Chapter 1: Introduction

1.1: The study of microbial communities and their interactions

Microscopic organisms, encompassing Bacteria, Archaea, and Protists, collectively form diverse and dynamic communities across all the environments. These microbial communities are not merely bystanders in the biosphere but play crucial roles in shaping the ecology and evolution of macroscopic organisms and ecosystems [1]. In fact, survival in natural ecosystems hinges on the intricate web of interactions and co-existence between species. These interspecific interactions can be reciprocal, where both parties benefit, or unilateral, where one species benefits at the expense of the other. However, these relationships are never absolutes, meaning that their outcomes can vary with time and context [2]. To better understand the complexities of these relationships, scientists have classified them into distinct categories: mutualistic, commensal, amensal, parasitic, competitive, predatory, and neutral [3] (Figure 1). By delving deeper into these various interaction types within microbial communities, it is possible to gain a fundamental understanding of the processes that shape the health and functioning of individual organisms and ecosystems.

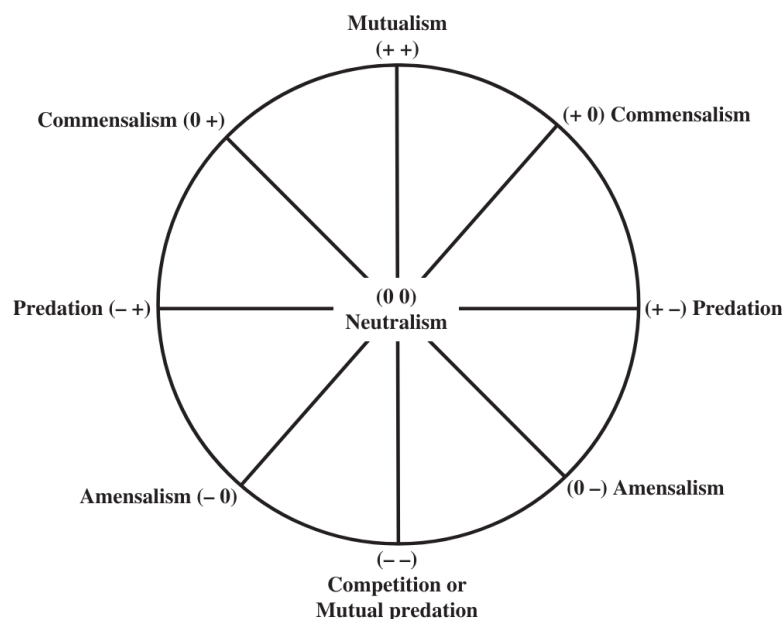


Figure 1: From Holland and DeAngelis, 2009. The figure represents the so-called interaction compass designed by the authors, which is used for the classification of interspecific interactions based on interaction outcomes, for which the positive (+), neutral (0), or negative (-) signs represent the effect of one species

on the other. The first sign represents the effect of species *i* on species *j*, and the second sign represents the effect of species *j* on species *i*.

The impact of microbial communities' presence and therefore interaction is reinforced by their ubiquitous presence. Microbes are in fact known to colonize environments from the most extremes such as Antarctic [4], deep down to ocean blue holes [5], and up to Mount Everest [6], as well as closer to us as our own guts and our pets and livestock. Nonetheless, microbial presence always relates to environmental conditions, as previously discussed by van der Gast [7], [8].

1.1.1: Microbial communities: who is there?

As mentioned above, the ubiquitous microbes are mainly composed by prokaryotes (which include Bacteria and Archaea), that are estimated to number around 10^{30} , of which 10^{12} microbial species (1 trillion) are known. About 1400 species are the cause of infectious diseases that are responsible for human morbidity, mortality, pandemics, and the resulting substantial economic losses [9], [10], [11]. Nonetheless, the remaining and their roles and functions are still undisclosed [12]. A great effort has been made throughout the last two decades to develop technologies that could help better understand the microbial communities and, in general, the microbiome.

Ma et al. published a review underlining the impact of microbial communities on hosts and environment in the perspective of One Health, being microbes fundamental linking nodes of carbon cycle, nitrogen cycle, as well as energy flow through ecosystems [13], and this scientific opinion was further supported (in the human perspective) by Trinh et al. review [14].

1.1.1.1: Bacteria

Microbes are reported to influence reproductive biology as modified by captivity and fertility control events rather than wild events [15]. For example, some Bacterial species (as well as some Archaeal ones, such as *Thaumarchaeota*) support nitrogen metabolism in sponges [16], [17] and molluscs [18], but can also function as growth promoters, nutrient suppliers, heavy metals transformers and immune system regulators. Eisthen and Theis showed instead how microbes influence animals' communication, behaviours and therefore probably the nervous system evolution itself [19]. Ruminants and methanogenic Bacteria interweave relationship that allow the gain of life-sustaining sugar from fibres to the first while feeding the seconds during rumination. Interestingly, this relationship

accounts for almost the 30% of all anthropogenic methane emissions [20], [21] and results also shared in insects [22].

Moreover, Bacteria influence on animals is that relevant to make McGall-Ngai et al. entitle their review, cited more than 2500 times (2665 citations reported by Google Scholar, accessed June 13th, 2024), “Animals in a Bacterial world, a new imperative for the life sciences” [23]. In fact, the authors sustain the theory that not only Bacteria influence animals nowadays but have also shaped and simultaneously have been shaped by animal development [24]. Consequently, research directed towards the investigation of the interactions between animals and their microbial communities is of uttermost importance to unravel current relations and future evolutions of the species.

1.1.1.2: Archaea and Protists

However, a mention is due on microbial communities: even though Bacteria are probably the better-known a most extensively studied part of them, microbes include other 2 of the 6 kingdoms of taxonomy: Archaea, and Protists. Archaea have raised scientific interest in reason of their fundamental role in host-associated microbial community, so to be defined archaeome, described by Borrel et al. [25]. Among the most mentioned Archaea, *Methanobrevibacter* is the predominant, because of its role in gastrointestinal tracts, followed at a distance by *Methanosphaera*, *Methanomassiliicoccus*, and *Thaumarchaeota* species: in fact, *Methanobrevibacter* can efficiently contribute to polysaccharides digestion through methane production [26]. *Haloarchaea* and *Thaumarchaeota* were instead found to be species specific and skin-related in animals.

Protists, on the other hand, seems to be the most neglected in scientific literature regarding animal microbiome: however, articles report protists presence in cow gastrointestinal tract and manure [27], [28], along with minor roles as parasites in plants or mutualists as *Symbiodinium*, that helps synthesizing coral food from light [29], [30].

In conclusion, the microbial communities are deeply tangled with animal species that host them, and dairy cows in particular benefit of bacterial presence. Enhancement in the comprehension of these relationship could help solving growing issues related to animal wellness and handling and needs to be investigated with advanced tools.

1.1.2: Omics sciences: definitions and fields of interest

The impact of microbial communities on animal welfare, particularly when related to dairy cows, has been under the lens since the end of last century. In fact, the discover of new evaluation tools can actively tackle the pollution and production issues that dairy

cows' industries daily face. In this frame, the advancement of science and technology has set the new horizon on microbiome study.

The microbiome is the total information about the microorganisms, their products (e.g., metabolites, proteins, and nucleic acids), and the environmental conditions of a specific ecological organism or ecosystem [31] (Figure 2).

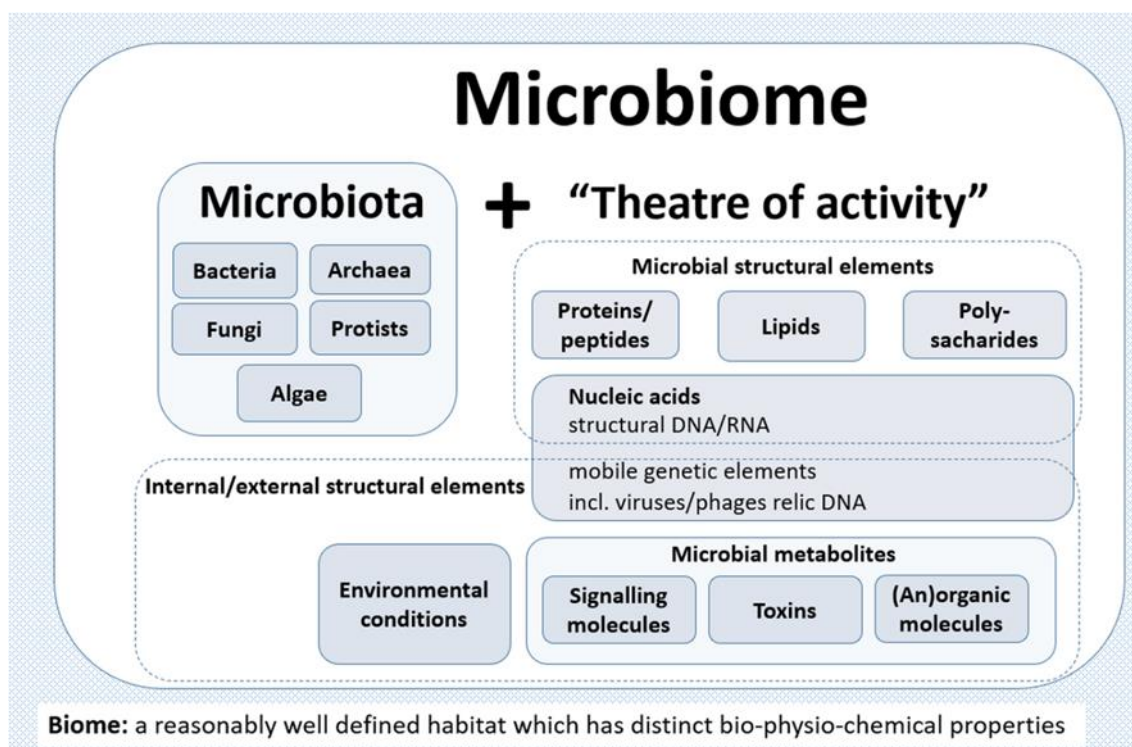


Figure 2: Figure from Berg et al. briefly representing how the microbiome encompasses microbiota, as well as metabolites, structural elements, and environmental conditions (Berg et al., 2020).

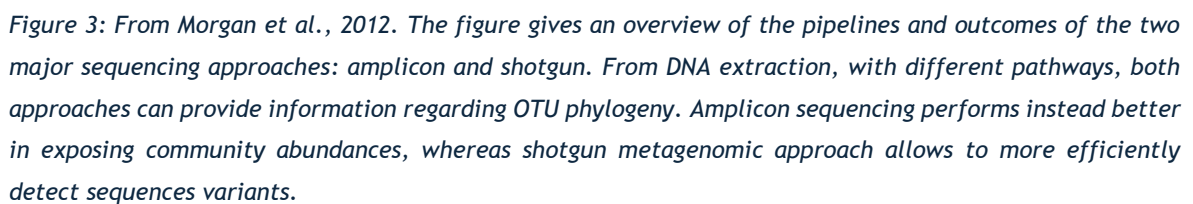
To study it, great technique efforts are made and collected under the definition of omics sciences. As reported by Yadav [32], the suffix -omics denotes a comprehensive and extensive exploration within the life sciences, concentrated on the analysis of extensive datasets. The aim is to gain insights into various aspects of life encapsulated in terms such as "omes" and "omics", to then relate them as layers of systems biology [33]. Metagenomics, metataxonomics, metatranscriptomics, metaproteomics, and metabonomics are probably the most popular omics currently. While metagenomics aims at the microbiota, metataxonomics can create the taxonomy tree for the organism or ecosystem. Metatranscriptomics explores the regulation and expression of the mRNAs

(messenger RNAs) present, while metaproteomics characterizes instead the protein profiles, and metabonomics describes the profile of the metabolites of the organism or ecosystem [34].

This thesis aims to focus on the microbiota of dairy cows: therefore, the collection of genes and genomes of this branch of the dairy cows will be studied with the aim to unravel how the microbial communities of 4 body sites can vary, in different experimental settings (as the experiments are mostly unrelated between each other), putatively influencing then dairy cows health. The outcomes of these observation could influence both the way of farming and medically treating dairy dairy cows, as experiments considered include treatments applied or fed to the animals.

As mentioned above, the gold standard to study it is metagenomics, which allows unravelling the unculturable part of the microbes of the organism or ecosystem. Classical microbiology has been in fact strongly limited by the impossibility of cultivating some microbes due to their growth condition (e.g., extreme environments). However, it should be noted that an extra effort has been made to flank metagenomics to reduce the so-called microbial dark matter, that is, the microbes remaining uncharacterized and uncultured. Culturomics was deployed for this effort; nonetheless, some microbes remain undiscovered due to their premature death and destruction [35], [36].

The distinction between metagenomics and metataxonomics to describe microbiota is related to the techniques used. Microbiota studies primarily rely on high-throughput sequencing (**HTS**), meaning sequencing technologies that succeeded Sanger sequencing and outdid it per speed and number of DNA molecules simultaneously sequenced [37]. HTS is known to be applied with two approaches: amplicon sequencing and metagenomic or shotgun sequencing [38], [39] (Figure 3).



1.1.3: Amplicon sequencing and shotgun sequencing

The 16S ribosomal RNA (rRNA) gene is used in amplicon sequencing as a reference for Bacteria (and Archaea) identification. In contrast, the 18S rRNA gene and internal transcribed spaces (ITS) are used for Fungi. For animals instead, cytochrome c oxidase subunit I (COI) gene is preferred [40]. The use of the 16S rRNA gene is encouraged by the presence in its structure of highly conserved regions that, along its 1500 base pairs (bp) of length, are spaced by variable regions, also called V-regions, and numbered from V1 to V9. While the conserved regions allow the building of primers for regional amplification, hypervariable ones allow Bacterial taxa discrimination [41]. The outcomes of amplicon sequencing generally describe the operational taxonomic units (OTU) present in the community and their phylogeny.

Similarly, this happens for the 18S rRNA gene, which also owns 9 hypervariable and ITS regions. ITS regions are generally ~600 bp long and can be distinguished between ITS1 and ITS2, characterized by different molecular weights (18S and 5.8S and between 5.8S and 28S, respectively) and among which ITS1 is considered more taxonomically accurate than ITS2 [42].

These regions, each on its Kingdom of reference, have been primarily employed to overcome the previously mentioned “unculturable” problem, which was the impossibility of culturing and, therefore, studying certain organisms out of their environment. However, it should be noted that other sequencing alternatives are available, such as serotyping and phage typing [43].

Shot-gun sequencing instead aims for complete genome sequencing by cloning the DNA sequence of interest. Its costs and data handling (including software availability and computing resources needed) are higher than amplicon sequencing but can provide a more comprehensive framework of the microbial community, unravelling gene profiles and reaching higher resolution in taxonomic annotation [38]. Additionally, these features allow the identification of novel genes that exhibit significant divergence from currently recognized genes, while it would be challenging to amplify them through polymerase chain reaction (PCR) or to express them in standard hosts [44]. Therefore, shotgun sequencing outcomes can be summarized in OTU phylogeny, variants reports, and gene pathways.

1.1.4: Sequencing generations

Whichever of the approaches is chosen, sequencing techniques offer increasingly cheaper, faster, and more reliable options. This development started with First-generation sequencers, known as Sanger sequencing, which allowed high accuracy and a relatively long read length. Second-generation sequencing, used by Ion Torrent Personal Genome Machine, Illumina platforms, and 454 pyrosequencing, improved the technique by using a massively parallel sequencing to achieve high base-calling accuracy, even though it is penalized by a relative shortness of the reads. Third-generation sequencers shifted from clonal template generation to single-molecule real-time sequencing [43]. Outcomes present very long read length at high throughput, paired with a relatively higher error rate than other technologies. Second-generation machines have a lower error rate (from 0.1 to 1%) than third-generation ones that, on the other hand, can sequence longer reads (up to 60,000 bp for Pacific Biosciences and 1 Megabyte (Mb) for Oxford Nanopore technology against the 35-700 bp) (Figure 4).

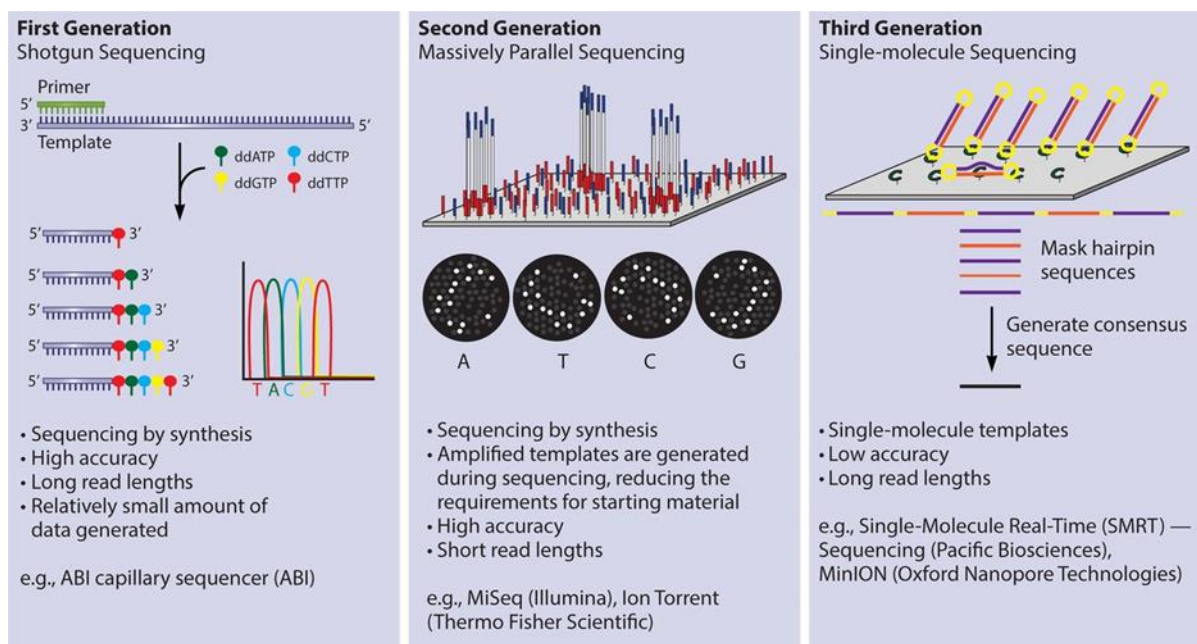


Figure 4: From Ronholm et al., 2016. The figure compares the three generations of sequencing in temporal order, giving hints of the main advantages and disadvantages of each of the sequencing generations. Whereas Shotgun sequencing was highly accurate and had long reads, it was penalized by the small amount of data generated. On the other hand, Parallel sequencing increased the amount of data that could be obtained by sequencing, at the cost of read length but keeping the accuracy that instead decreased again when gaining long reads within Single-molecule sequencing.

Machines using second and third-generation sequencing methods are generally defined as Next-Generation Sequencing (NGS) platforms. They are the most used methods to study microbiota via the V-regions of the 16S rRNA gene.

1.2: Bioinformatic analysis

The output of metagenomics, especially of 16S rRNA gene sequencing, are sequences that allow the identification of the microbial communities of the dairy cows' samples taken into analysis. Several steps are needed to adequately process these sequences and describe the microbial community. Sequences need to be processed before they can be assigned to microbial taxa at different taxonomic levels, from phylum to species. Researchers can build a specific analysis pipeline with scripts that integrate multiple software packages in a predetermined sequence to execute a complex analysis task [38]. In the case of 16S rRNA-gene sequencing, multiple options are available, all based on shared principles.

1.2.1: General outline of bioinformatic pipelines of amplicon sequencing outputs

Results from sequencing platforms are generally delivered as FastQ files, which include quality scores for each base identified in each sample. Results can come from the platform already demultiplexed or would need to be demultiplexed, meaning that sequences would need to be divided by sample based on barcodes before proceeding with further steps. Once demultiplexed, the steps of the analysis pipelines can vary based on the software used but commonly include primer and adapter removal, plus quality filtering. Cleaned sequences will commonly be used as proxies for taxonomic classification. This task is mainly performed using two approaches: clustering and denoising (Figure 5), even though some recent publications support the point that these two methods should be complementary rather than alternative to each other [45].

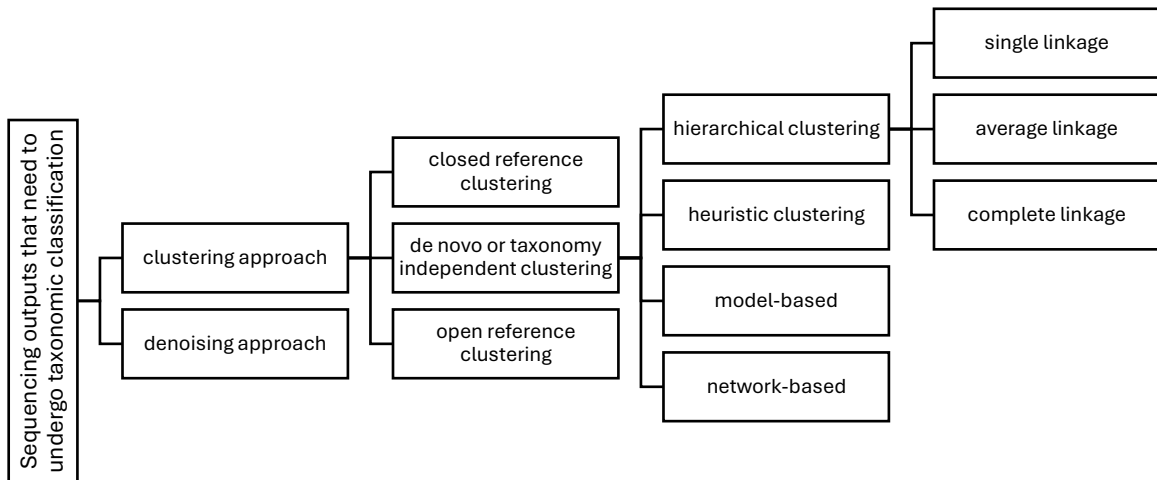


Figure 5: Approaches to analyse amplicon sequencing outputs. This tree summarizes the approaches applicable to sequencing outputs that need to undergo taxonomic classification: the first branching oppose clustering and denoising approach, to further investigate the clustering approach characteristics and features.

1.2.1.1: Clustering methods

The clustering approach involves aligning the sequences against a reference database. The output will be several OTUs, where each OTU represents a sequence clustered against the database of reference, allowing a certain percentage of uncertainty defined as an identity threshold, generally set at 97% (hence allowing a 3% mismatch). Three methods are commonly used for this task: closed reference clustering, *de novo* (also known as taxonomy-independent) clustering, and open reference clustering [46].

The closed reference clustering approach [47] maps the reads against the database and assigns them to the best-matching reference sequence, allowing to specify the threshold required; therefore, reads failing to match with a sufficiently high score are discarded, while the ones matching the same reference sequence are classified as an OTU. This method causes a high dependence of the outcomes on the reference database used [48]. With the *de novo* approach, sequences are grouped into OTUs by considering pairwise sequence distances instead of comparing them to a reference database [49]. *De novo* strategies [47] can be broadly categorized into hierarchical clustering, single, average, and complete linkage, or heuristic strategy [50], [51]. In hierarchical strategies are included all those methods that cluster sequences basing on a chosen distance metric

with the aim of grouping them, with various level of completeness, aiming to build a tree-such as structure. Single linkage involves placing a sequence in a cluster if it shares a similarity above a certain threshold with at least one other sequence in the cluster. However, this method often leads to large clusters with considerable heterogeneity and is infrequently employed, except for rapidly evolving genes such as the fungal ITS region [52]. Conversely, in complete linkage, sequences within a cluster must exhibit a similarity above the threshold within each other, thus resulting in smaller but more numerous OTUs and a tendency to overestimate community richness. Moreover, average linkage requires the average similarity between a sequence and all others in the same cluster to surpass the threshold. Heuristic strategies instead approximate complete- or average-linkage methods by comparing each sequence against a "centroid" sequence within its respective cluster [53]. An approximation of average-linkage clustering is achieved by establishing a distance cut-off between the centroid sequence and potential sequences. Alternatives to hierarchical and heuristic methods, such as model-based and network-based clustering, are available. Still, their computational demand and their unsuitability for big datasets make them scarce in popularity [50]. Lastly, the open reference approach uses the to-be-discarded reads of the closed reference to undergo a *de novo* strategy.

1.2.1.2: Denoising methods

Denoising or binning methods aim to deduce biological sequences before amplification and sequencing errors, discerning sequence variants that differ by as little as one nucleotide. This method aims to generate amplicon sequence variants (**ASV**), also known as sub-OTUs, zero-radius OTUs (**zOTUs**), or exact sequence variants (**ESV**) [54]. It should be highlighted that ASVs differ from *de novo* OTUs, as ASVs consistently represent the sequence present in the sample. In contrast, *de novo* OTUs represent the clustering of similar sequences. Callahan et al. schematized this difference in a compelling figure, here reported as Figure 6, in which biological variations are contrasted with the amplicon data generated in a specific locus, highlighting how the method's region of validity varies based on the approach [55].

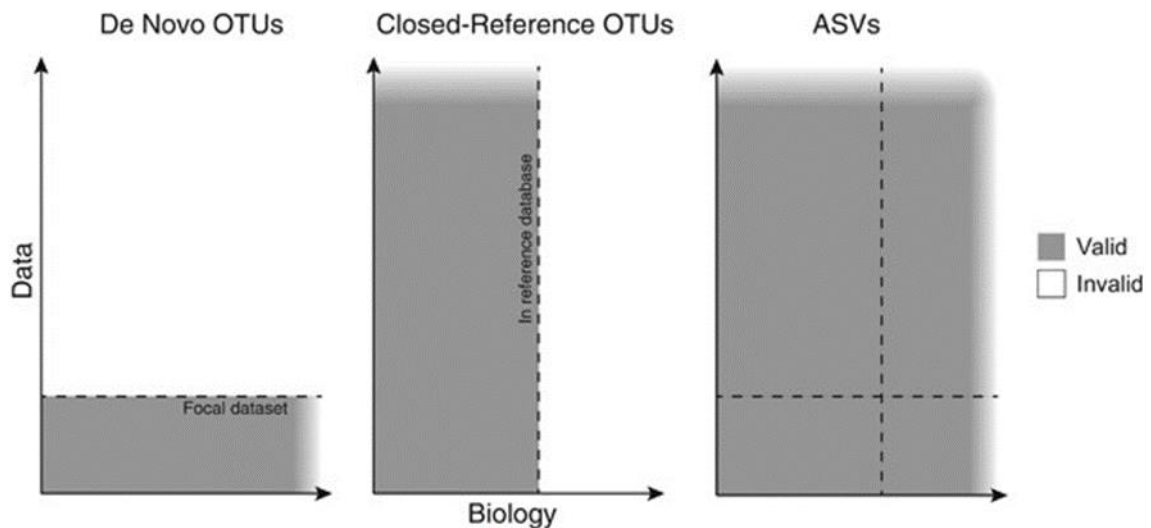


Figure 6: From Callahan et al., 2017. This figure aims to graphically represent the validity of de novo OTUs, closed-reference OTUs, and ASVs derived from a specific dataset.

1.2.2: Data analysis

Both OTU clustering and ASV denoising approaches will generate a table that relates each sample with a taxon identified by reporting its abundance, which is then linked to a database of choice to assign the correct taxonomy. It is good practice, sometimes as well embedded in the analysis software, to normalize the results, aiming to eliminate uneven amplification and sequencing depth. A simple and common way to do it is to calculate the ratio between the single sample Bacterial count and the total Bacterial count of the sample.

1.2.2.1: Commonly used analysis

Downstream analysis includes a variety of approaches, such as alpha and beta diversity analysis and taxonomic composition. Furthermore, biomarker identification can be achieved using comparison, correlation analysis, network analysis, and machine learning [38]. Alpha diversity assesses the diversity within a given sample, encompassing species richness and evenness measures. The most commonly used alpha-diversity indices are A) Chao1, estimating true sample richness by considering the tail of the species or OTUs abundance distribution, especially considering singletons and doubletons; B) **ACE** (abundance-based coverage estimator), which enlarges Chao1 properties to all OTUs observed, C) Simpson, which calculates the odds that two randomly sampled microbes

belong to the same taxon and D) Shannon, which predicts the uncertainty of species identification. Beta diversity examines variations in the microbial composition across different samples by measuring the distance between samples based on OTUs abundances. It is typically utilized with dimensionality reduction techniques such as principal coordinate analysis (**PCA**) and multidimensional scaling (**MDS**) to generate visual representations.

Other commonly used analysis include the study of the taxonomic composition, which refers to the microbiota within a microbial community, F:B ratio (Firmicutes to Bacteroides ratio) used to assess some pathological condition regarding gastrointestinal tracts), gamma diversity, phylogenetic trees, etc. (Figure 7) [50].

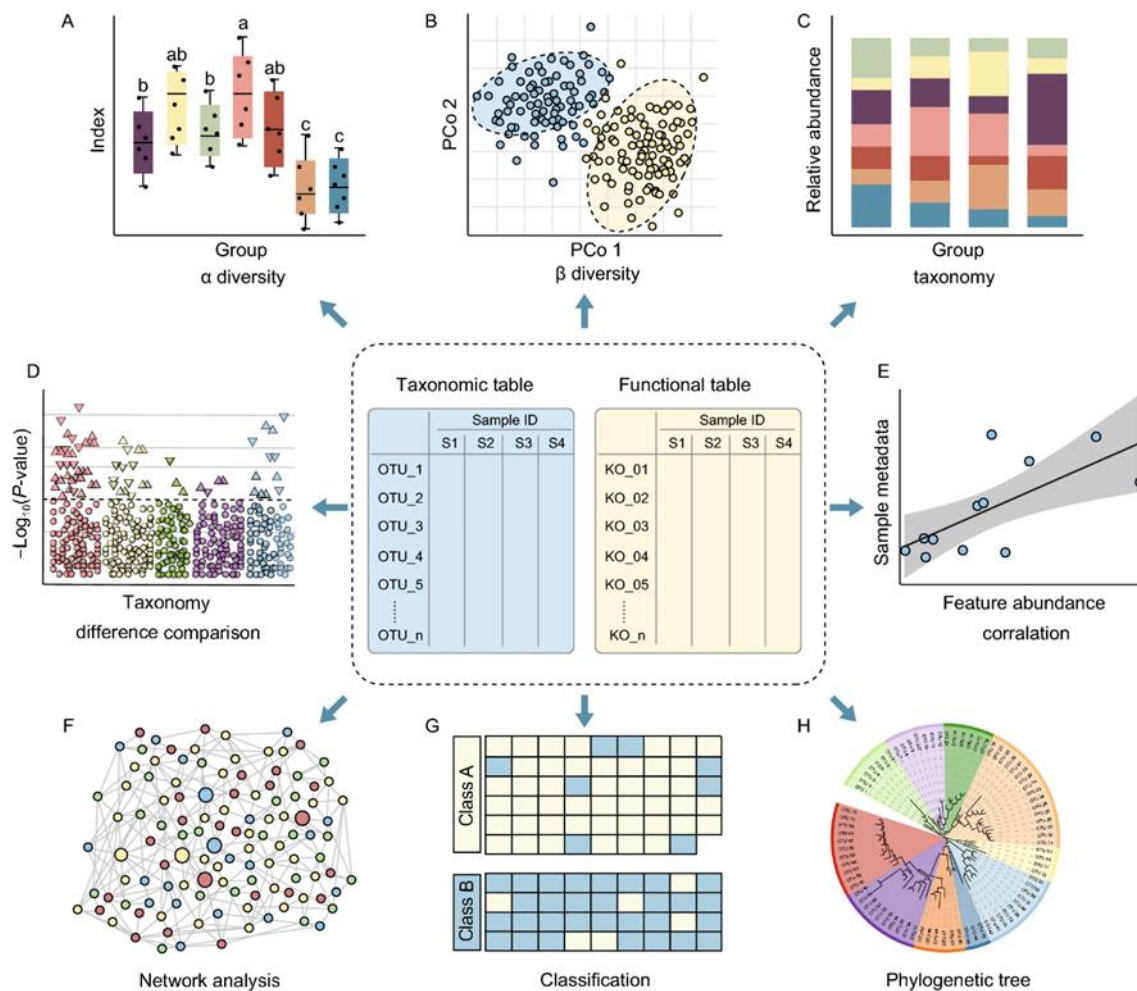


Figure 7: From Liu et al., 2020. This scheme aims to give a general overview of the most used statistical analysis performed on microbiota and their visualization method. Analysis will in fact start from a taxonomic table (OTU or ASV table, depending on the approach used) and, depending on research targets and statistical tools available, deliver a variety of information.

1.2.2.2: Commonly used software

Several software are available to perform the above-mentioned analysis. For sure, the most popular one is R [56]. R is a programming language developed for statistical computing and graphics. It is most often used in its graphical interface, R Studio, where the user can perform analysis starting by a wide variety of statistical and graphical techniques [57]. Moreover, what is not in the basic version of the software can be implemented through the installation of the so-called packages, which are often developed and provided by the community, therefore creating a wiki for R: an example are phyloseq package [58] and ggplot [59] package, that will be mentioned further in the Material and Methods chapter of this work. Manuals and blogs of support, such as StackOverflow, are active and in continuous improvement to support inexperienced users as well. R is provided for free and without registration, and is frequently updated: however, the user is requested to learn some basic skills of programming also to interact with the graphic version.

Alternatively, software such as SPSS, SAS, GraphPad Prism and XLStat are available to run statistical analysis. All this software however require a subscription on payment, which pairs with the availability of a specific assistance equipped by the developers but often discourages users that do not have fundings for assets such as that.

1.3: Use and usefulness of microbial communities' study through bioinformatic analysis in dairy cows

1.3.1: Focusing on the issues to find the solutions

The outcomes of bioinformatic analysis supporting genomics, and therefore the extensive description of the microbial communities present in dairy cows' samples, have strongly enhanced the comprehension of the dairy cows' industry in the last decade. Indeed, a deeper knowledge of the microbiome involved in animal wellness and production allows a broadening of tools available for enhanced health and sustainability of dairy cows.

First, it helps the tackling of pollutant issues such as methane production. Methane (CH_4) is in fact is formed by the methano-archaeal components of the rumen and protist microbiome, and place cows first over termites, sheep, goats, pigs, humans and chickens in CH_4 in millions of tons per year (mtpa). Moreover, the Food and Agriculture Organization of the United Nations (FAO) calculated that farm-gate greenhouse gas emissions (GHG) had a 14% increase between 2000 and 2021, of which its 53% are related

to livestock. The global emissions intensity for cow milk was 1 kg of carbon dioxide per kg of milk (**kg CO₂ eq/kg**). Secondly, it supports dairy cows monitoring before, during and after lactation. In fact, world milk production (data are reported for dairy cows' milk, accounting for cow and buffalo milk) had in 2021 an increase of 328 million tonnes compared to 2000, for a total of 883 million tonnes [60].

However, as research by Koeck et al. suggests, higher-producing cows may experience a higher incidence of clinical mastitis, lameness, and other diseases [61]. Additionally, increased production can lead to higher antibiotic consumption. Lactation stress can weaken cow immune defences and induce mastitis, as well as increased susceptibility to other diseases, shortened lifespans, and reduced environmental fitness, as well as increased susceptibility to diseases, short life expectancy and reduced environmental fitness [62], [63].

1.3.2: The holobiont

Stated that microbial communities of the host and the host itself compose a so called "holobiont" in which the complex and mutualistic interactions is capable of homeostasis maintenance, literature mostly focuses on specific anatomical sites to be characterized, enlarging the field by connecting omics and building the system biology in a bottom-up perspective. Regarding dairy cows, scientific research has delved into various domains, including easily accessible and sampleable targets, such as milk, skin, and gastrointestinal tract. However, investigations have also extended to less conventional regions such as the airways, as well as unexpected sites such as the placenta and foetus [64]. This thesis investigates four of them, namely milk, skin, rumen and feces, with the goal to gain an overview over the status of microbial communities in cows (or cows derived samples) healthy or specifically treated.

1.3.3: Dairy cows milk microbiota

As the first study that will be exposed further in this thesis regards milk samples, an overview on this topic is hereby exposed. It should be brought to the reader attention that the interest towards milk microbiota is related to dairy cow health and productivity: as mentioned beforehand, milk production causes a stress on the gland that, flanked to mechanical, physical, chemical, or other biological causes could lead to mastitis, which is the inflammation of the mammary gland and causes an increase of somatic cell counts (European Union adopted by law the threshold of 400 000 cells/mL in bulk tank milk to

define subclinical mastitis, that decreases to 100 000 cells/mL per individual) and visible signs such as swelling, heat, hardness, redness or pain. The milk such as a watery appearance, flakes, clots or pus, that generally define a clinical mastitis [65], [66]. Mastitis is therefore considered the most frequent disease in dairy cows, and influences dairy farm profitability directly, and antibiotic resistance indirectly, considering that pharmaceutical treatments are provided to the animals to heal but typically only the 10 to 30% is consumed. The rest remains unmetabolized and is excreted in urine or faeces [67]. Microbial communities' comprehension plays therefore a primary role in keeping the mammary gland healthy, and milk constitutes the least invasive and most effective mean to assess mammary gland microbiota.

It should however be mentioned that dairy cows' mammary gland was long considered a sterile environment, and therefore microorganisms found in milk were assumed as external contamination due to sampling or culturing methods. Instead, core microorganisms in dairy cows' milk were discovered to contribute to the udder wellness and to the milk quality, raising the interest towards their impact on animal health, and characteristics and safety of dairy products. Most importantly, two theories lead to explain the presence of microbes in milk: entero-mammary pathways and rumen-mammary pathways, both encompassed in endogenous pathways, counterposed to the retrograde pathways, that points to the calves as mean of transfer for microorganism but is discredited by the breeders' habit to remove calves shortly after birth. Endogenous pathway suggests that microorganisms are sequestered and transported to the udder by the immune cells through the lymphatic system as previously demonstrated in mice and human [68], [69], [70], [71]: among entero-mammary and rumen-mammary pathways, the second is empowered in dairy cows by the established microbial communities of the rumen, as shown by Metzger et al. [72] (Figure 8).

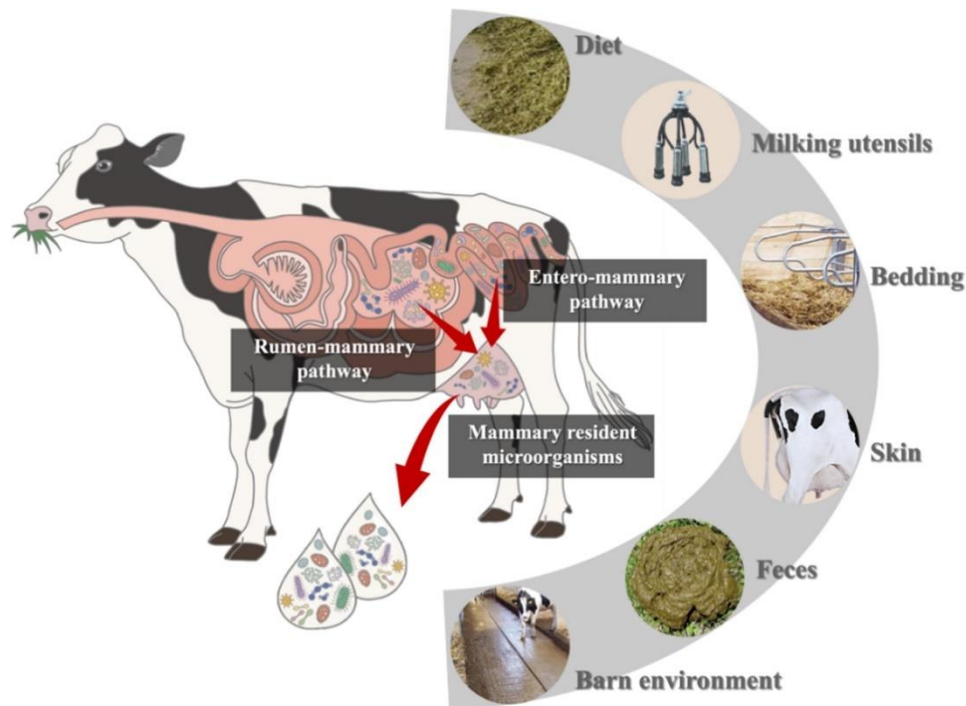


Figure 8: From Guo et al., 2023. Scheme of microorganisms' probable derivation when found in dairy cows' milk, including external sources, entero-mammary and rumen-mammary pathways, and mammary resident microbes

Literature reports that Proteobacteria and Firmicutes collectively constitute approximately 90% of milk microorganisms, followed by Bacteroidetes and Actinobacteria. Within the Proteobacteria group, the more frequent genera are *Acinetobacter*, *Pseudomonas*, *Escherichia*, *Vibrio*, *Erwinia*, and *Pantoea*. Firmicutes phylum instead is predominantly represented by *Streptococcus*, *Enterococcus*, *Staphylococcus*, and *Bacillus* [73]. The genera that are considered as representing an healthy quarter are *Staphylococcus*, *Ruminococcaceae*, *Lachnospiraceae*, *Propionibacterium*, *Stenotrophomonas*, *Corynebacterium*, *Pseudomonas*, *Streptococcus*, *Comamonas*, *Bacteroides*, *Enterococcus*, *Lactobacillus*, and *Fusobacterium* [66]. *Staphylococcus aureus*, *Streptococcus dysgalactiae*, *Streptococcus uberis*, *Streptococcus agalactiae*, *Klebsiella* spp., and *Escherichia coli* are instead the most known and investigated pathogens related to mastitis.

1.3.4: Dairy cows skin microbiota

For the same reasons listed in the previous paragraph, skin microbiota is investigated principally as related to udder health. However, interest has been arisen also by the

microbiota related to digital dermatitis. Indeed, bacteria along with their emissions of volatile odorants and gases, have been demonstrated to attract stable flies and stimulate oviposition that causes dairy cows' lesions that cause lameness. By reducing animal welfare through painful lesions, dairy industry suffers economic losses. Moreover, the same bacteria related to digital dermatitis were also found in other dairy cows' foot lesions as toe necrosis, non-healing white line disease and non-healing sole ulcer [74].

Nielsen et al. identified *Treponema*, *Mycoplasma*, *Fusobacterium* and *Porphyromonas* as the best reporters at genus level for dairy cows with digital dermatitis against controls: in details, the most prevalent genera was *Treponema*, followed by *Fusobacterium* and *Porphyromonas* [75]. *Borrelia*, *Bacteroides*, *Campylobacter* and *Mycoplasma* were instead reported in Evan et al. [74].

Udder skin microbiota on the other hand was reported to be composed by Proteobacteria, Firmicutes, Actinobacteria and Bacteroides at phyla level, flanked by two archaeal phyla: Euryarchaeota and Thaumarchaeota. Genus level is instead dominated by *Acinetobacter*, which increases related to stress conditions such as postpartum at the expenses of *Enhydrobacter* and *Bradyrhizobium*. Other commonly identified genera of the udder skin include *Moraxella*, *Staphylococcus*, *Lactobacillus*, *Atopostipes*, *Alysiella*, and *Fusobacterium*. *Staphylococcus*, *Ruminococcaceae*, *Bacteroidales*, *Clostridiales*, and *Pseudomonas* are also the predominant constituents of healthy colostrum microbiota [76], [77].

1.3.5: Dairy cows rumen microbiota

The interest in the microbial composition of ruminal environment stems in its life-sustaining function: in fact, ruminants, although being mostly herbivorous, are not able to digest fibres. Therefore, they solely stand on the byproducts of the microbes able to establish symbiotic relationships and degrade plant polymers into simpler sugars by utilizing their own secreted enzymes [78], [79]. This symbiosis guarantees the dairy cows proper sustainment, but on the other hand is the main cause of the previously mentioned GHG emissions. GHG emissions are in fact ascribable to eructation or enteric fermentation and can be modulated through dietary modification to reach efficient emission mitigations in commercial farms[80].

Both under the welfare and the pollution point of view, microbial communities of the rumen have been therefore previously investigated. Xu et al. reported that rumen contains about 10^3 bacteria/mL, of which predominant phyla are Proteobacteria,

Bacteroidetes and Firmicutes, along Archaea (such as Methanobrevibacter and Methanosphaera), Fungi, and Protozoa [79], [81].

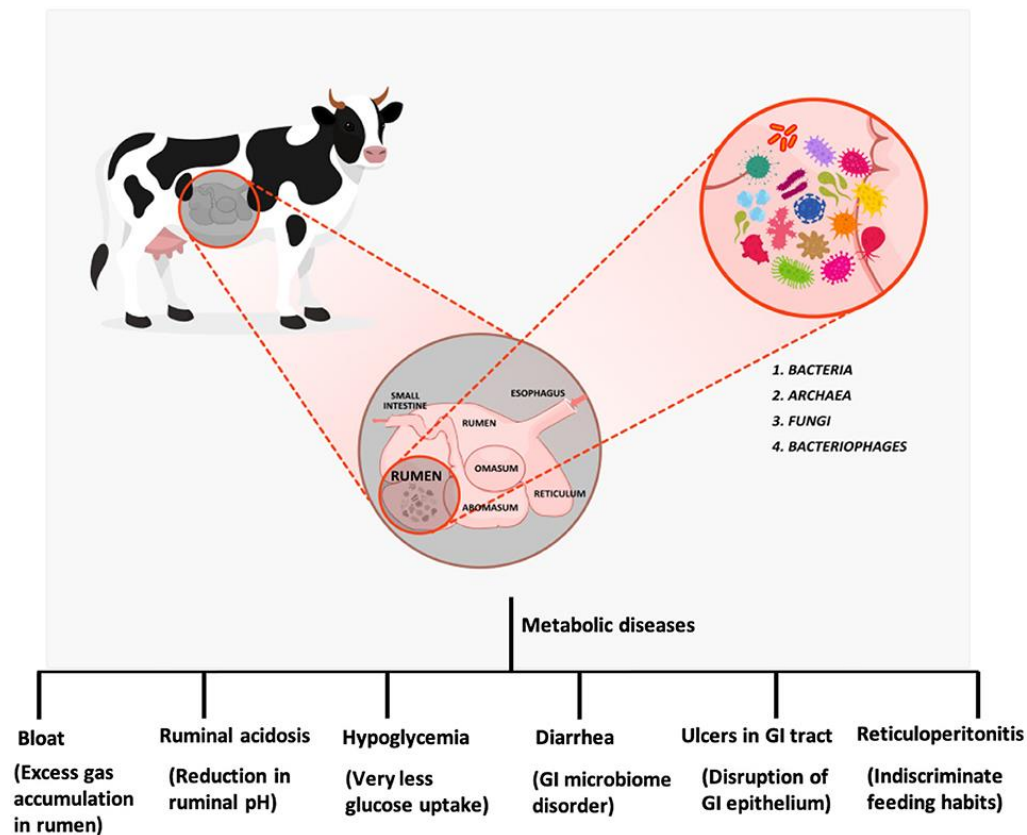


Figure 9: From Xu et al., 2021. The figure provides an overview of the more common metabolic diseases of dairy cows' digestive system related to their microflora. Unbalance in the composition of Bacteria, Archaea, Fungi and viruses can in fact heavily influence dairy cows welfare and production.

Nonetheless, it is known that microbial composition can variate basing on feed consumption [82], [83], and food imbalances can lead to a range of metabolic disorders, including bloat (consisting in an excess of gas accumulation in the rumen, leading to animal distress and), ruminal acidosis, hypoglycemia, diarrhea, gastrointestinal ulcers, and reticuloperitonitis [81] (Figure 9). Therefore, knowing the structure of ruminal microbiota can optimize its functioning guaranteeing animal welfare and consistent production. Moreover, it constitutes the key in allowing fast adaptation to new diets, improving dairy cows' resiliency [84].

1.3.6: Dairy cows fecal microbiota

Dairy cows' fecal microbiota significantly impacts not only the welfare and productivity of the animals themselves, but is also critical to food safety, pathogens spreading, and the effectiveness of methods used to detect fecal pollution. In fact, dairy cow faeces represent an easily collectable proxy of the gut microbiota, whose activities range from nutrients production to host metabolism regulation, energy balancing and immune response [85], [86]. On the other hand, however, they also represent manure, which can be considered one of the plainest ways of antibiotic resistance shedding to environment. Fecal microbiota composition was outlined from Kim et al., who reported Firmicutes as predominant phylum, followed by Proteobacteria, Bacteroidetes, TM7 and Actinobacteria [87]. *Actinobacteria*, *Spirochaetes*, *Verrucomicrobia*, *Fibrobacteres*, and *Lentisphaerae* are as well reported in literature [88], along with *Anaerosporebacter* that negatively relates to cow age, *Treponema*, *Alistipes*, *Prevotellaceae* and *Bacteroides* [89].

Chapter 2: Potential significance of the study

The potential significance of this thesis outcomes is to provide an overview of the differences among some of the most used pipelines of analysis and microbial databases available for dairy cows 16S rRNA-gene sequencing data. To achieve this result, the study was articulated in 3 main progressive sections: i) a pilot experiment with 1 dataset, 1 pipeline and 1 database, and comparing the results with the same database in a newest version and other 3 databases (of which 1 in 2 different versions), and then ii) the application of other 3 pipelines (of which 2 in 2 different approaches - clustering vs denoising) and 3 databases (of which 2 in 2 different versions) on the same dataset (36 combination in total), and lastly iii) the benchmarking of the same 36 combination on 3 pipelines (of which 2 in 2 different approaches - clustering vs denoising) and 3 databases (of which 2 in different versions) on 5 other datasets (Figure 10).

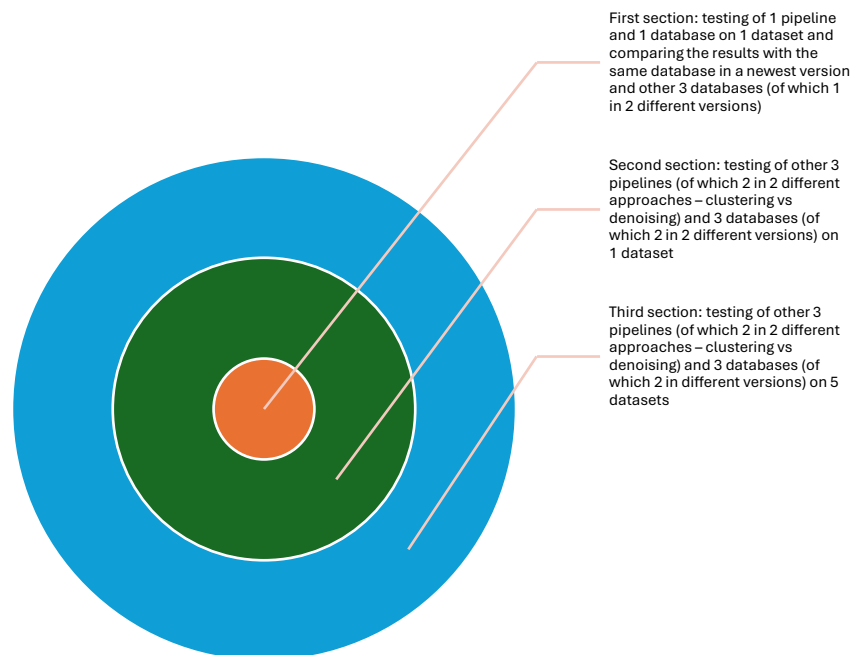


Figure 10: Infographic of the thesis structure. Starting from the centre of the bull eye and going to the external circles, this figure represents how the thesis aimed to enlarge the frame of the study from considering just the variation on 1 dataset analysed with 1 pipeline and 1 database, to the same analysis swiping databases, to the inclusion of different pipelines, to the application of the pipeline + databases combination (36 of them) to other datasets.

In details, the first step involved the analysis of one real-life dataset, the milk dataset, belonging to MILKQUA project. The sequencing results of this dataset were analysed using

QIIME version 1.9 paired with SILVA version 132 database. The output obtained was investigated through biostatistics to evaluate alpha and beta diversity indices, as well as taxonomy. Alpha and beta diversity results were then compared to those obtained combining QIIME VERSION 1.9 with the other databases, namely 16S-ITGDB, Greengenes version 13.8 and 2, RDP, and SILVA version 138, to study the impact of microbial databases on the outputs.

The second step consisted of repeating the analysis of the milk dataset using other pipelines available to reach the same results, namely QIIME VERSION 2, VSEARCH and MICCA. It should be noted that MICCA and QIIME VERSION 2 were used with both the closed reference and the denoising approach, which generated a total of 5 pipelines to test beyond QIIME VERSION 1.9.

Furthermore, the author aimed to ascertain whether and how the combination of pipelines and databases was influencing the results: therefore, once assessed the results of the pipelines and databases combinations on milk dataset as pilot experiment, the study was extended to other four dataset: two of them, namely mock community from Prodan et al. 2020 and mock community from Callahan et al. 2016, were used to test the combinations on datasets limited and known in Bacterial composition. The other three were used to mimic the results of milk dataset on other dairy cows' datasets, as skin, rumen, and stool datasets.

The final aim of the thesis has been to provide a scientific consensus about the pipelines and databases best combination to analyse 16S rRNA sequencing results from cow samples.

Chapter 3: Material and methods

In this chapter, the three sections of the study are extensively explained. As bioinformatics can only exist thanks to the work of planning, animal care and treatment, sample collection, and storage until reaching the sequencing facility, study designs are reported for completeness, even if not performed by this author. All the datasets, except for the mock communities downloaded from public repositories, shared the sequencing facility and platform, Illumina MiSeq, as described in the first section below (subchapter 3.1.2) and recalled in each study.

3.1: First section: testing of 1 pipeline and 1 database on 1 dataset and comparing the results with the same database in a newest version and other 3 databases (of which 1 in 2 different versions)

The first section of the study describes the pilot experiment, in which 1 pipeline, QIIME version 1.9, was paired with database SILVA version 132 and used to analyse 1 dataset, the milk dataset belonging to the MILKQUA project. The results were then compared with those obtained by combining the same dataset and pipelines with the same database in a newest version, that is SILVA version 138, and other 3 databases (of which 1 in 2 different versions), namely RDP, Greengenes version 13.8 and Greengenes version 2, and 16S-ITGDB.

3.1.1: Detailed description of the dataset included in the study

The following study was conducted in the frame of the **MILKQUA** (Milk Quality along the Dairy Chain for a Safe and Sustainable MILK) project, identified as MILKQUA-H2020-PRIMA 2018—Section 2 and funded by **MUR** (Ministero dell'Università e della Ricerca -Direzione Generale dell'internazionalizzazione e della comunicazione Programma cooperazione internazionale PRIMA - Decreto Ministeriale 26 luglio 2016 n. 593) and **ANR** (French National Research Agency).

The experimental group comprised twelve Holstein dairy cows at the end of their lactation period, housed in the Méjusseaume **INRAE** (France's National Research Institute for Agriculture, Food and Environment) experimental dairy farm. None presented clinical signs of infection, received antibiotic treatment 5 months before the experiment, or had previous clinical mastitis history. The animals were divided into two groups (control and EO). Milk was collected twice daily in a rotary milking parlour for the entire study period. The treatment was constituted by a prepared shot of 10% TCEO dissolved in 4.5 g of milking grease, stored at 4°C, and applied to the udder after each milking for a week. Throughout the experimental duration, the EO group (comprising 6 animals, 2 of which were sampled on 2 quarters instead of only one) received massages using the mixture. In contrast, the Control group (again consisting of 6 animals of which 2 were sampled on 2 quarters instead of only one) was subjected to massages using milking grease only. Milk samples were collected to study the TCEO treatment effect on microbiota (milk and skin), as well as somatic cell count (SCC), bacteriological examination, fat and protein content assessment, evaluation of casein micelles, sensory analysis (milk only) on fresh samples. Viscosity analysis, IL-8, acute phase proteins (haptoglobin and lactoferrin), and lipidome were studied on milk after storage at -80 °C. On days 0 (T0), 7 (T7), 21 (T21), and 28 (T28), milk samples were collected aseptically before the morning milking session. After discarding the initial three foremilk streams, teats were disinfected using 70% alcohol. Milk DNA extraction was performed following an in-house protocol described by Cremonesi et al. [90].

3.1.2: Sequencing applied

Before sequencing, samples required DNA quality and quantity assessment, which was performed using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA). Afterward, samples were amplified, targeting the V3 and V4 regions of the 16S rRNA. Illumina guide suggests primers 5'-CCTACGGGNGGCWGCAG-3' and 5'-GACTACHVGGGTATCTAATCC-3' to target the regions [91]. Those primers were designed in-house with a dedicated program for a broad range of 16S rRNA primer design by Herlemann et al. [92] and tested for coverage and specificity on the Archaea and Bacteria domains by Klindworth et al. [93]. Additionally, Illumina equips the primers with adapter overhang nucleotide sequences to obtain 5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG-3' and 5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC-3' as full-length primer sequences forward and reverse, respectively (Figure 11).

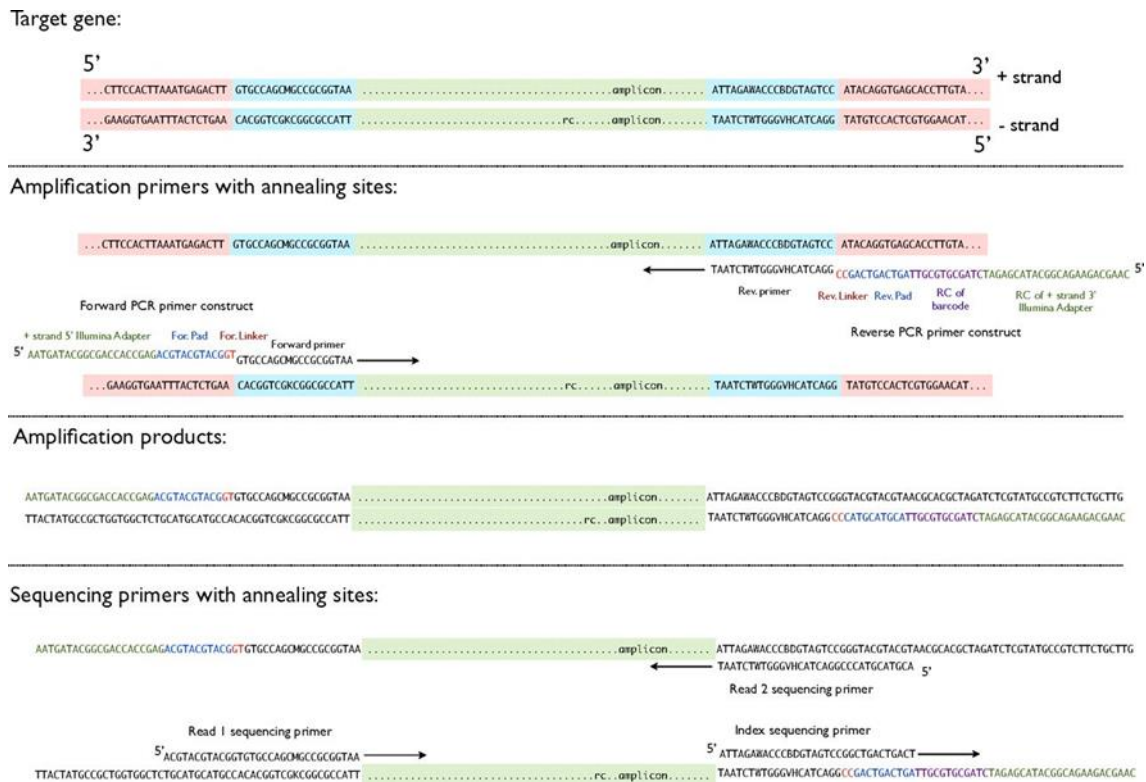


Figure 11: From Caporaso et al., 2011. The figure graphically shows the protocol to be applied when using barcoded Illumina pyrosequencing, including primer annealing, amplification and sequencing starting from primer annealing sites.

As previously described in Biscarini et al. [94], PCR was performed in a 25 μ l volume, of which 1:2 of the total volume of KAPA HIFI Master Mix 2 $\times$ (Kapa Biosystems, Inc., MA, USA), 5 μ l of Amplicon PCR Forward Primer 1 μ M and 5 μ l of Amplicon PCR Reverse Primer 1 μ M, plus 2.5 μ l of genomic DNA 5ng/ μ l. Blank controls (non-DNA templates) were included. Following the 16S mentioned above Metagenomic Sequencing Library Preparation Protocol from Illumina, the first amplification step had 25 cycles of denaturation at 98 $^{\circ}$ C for 30 s, annealing at 56 $^{\circ}$ C for 1 min, and extension at 72 $^{\circ}$ C for 1 min, forerun by a 95 $^{\circ}$ C for 3 min denaturation and followed by a 72 $^{\circ}$ C for 7 min extension. Amplification was performed in an Applied Biosystem 2,700 thermal cycler (ThermoFisher Scientific). An amplicon cleaning step was performed using Agencourt AMPure XP (Beckman, Coulter Brea, CA, USA). Libraries were prepared following the protocol's instructions and quantified by Real-Time PCR with KAPA Library Quantification Kits (Kapa Biosystems, Inc., MA, USA). Obtained libraries were then pooled in equimolar portions and sequenced using an Illumina MiSeq machine with 2 $\times$ 250-base paired-end reads (Figure 12).

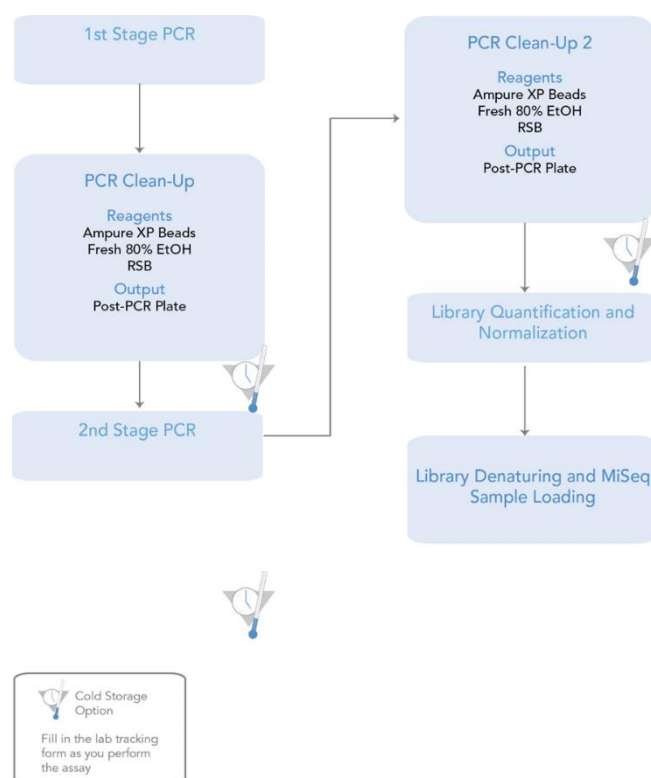


Figure 12: 16S Library Preparation Workflow as described in Illumina 16S Metagenomic Sequencing Library Preparation Protocol.

3.1.3: Detailed description of the pipelines and databases included in the study

Samples from sequencing were delivered already demultiplexed. Quality control analysis was performed before proceeding with downstream analysis, e.g., by using MULTIQC [95], an expansion of the FASTQC software [96] that allows the overview of the complete set of samples instead of one by one.

Reads can be analyzed with several software packages and tools available to the community: for the studies mentioned above, QIIME VERSION 1.9 was used [97]. Therefore, the first step included joining the forward and reverse reads into single reads using the C++ package SeqPrep as embedded in the pipeline. Merged sequences underwent quality filtration according to the following criteria: (1) exclusion of more than three consecutive low-quality base calls (Phred < 19); (2) prohibition of bases labeled as "N" (unidentified); (3) a prerequisite of a minimum 75% of consecutive bases in a read with Phred > 19. Sequences failing to meet these filtering conditions were eliminated.

The output of the filtering was a single Fasta file that was aligned to the database of choice using the specific command (and therefore approach) of the selected pipeline. As mentioned, SILVA database version 132 was used for closed-reference OTU identification and quantification, with 97% cluster identity. This step allowed us to obtain a phylum-to-species level taxonomic identification based on a predefined map of reference sequences to each taxon, generating an OTU table. The OTU table was cleaned from OTUs with a total count lower than 10 in less than 2 samples. As for the databases, SILVA is considered by many a top-tier one, justifying its inclusion in this study. Fasta and taxonomy files for version 132 were obtained from the official website (<https://www.arb-silva.de/download/archive/>) [98]. A custom-made classifier for version 132, compatible with QIIME VERSION 2, was created locally.

3.1.4: Biostatistics applied

Downstream statistical analyses of the results were performed using R and Rstudio [57]. R package phyloseq was used to manipulate the OTU table for further analysis: the OTU table abundances were normalized by cumulative sum scaling (CSS) [99] and then alpha and beta diversity indices were calculated using the phyloseq R package [58]. Microbial richness and diversity within the samples were assessed through various indices, including Chao1 and ACE for richness and Shannon, Simpson, and Fisher's alpha for diversity. Additionally, evenness was evaluated using Simpson E and Pielou's J (Shannon's evenness). One-way ANalysis of VAriance (ANOVA) [100] was applied, to analyze all the data, employing a linear model different for each study, depending on the variables available. A general equation would look as the following:

$$Y_{ij} = \mu + T_i + C_j + e_{ij}$$

where y_{ij} is the individual value for each variable, μ is the mean of the depending variable on the independent ones, T_i is the fixed effect of the i treatment (any of the groups tested), C_j is the fixed effect of the j time point, and e_{ij} is the residual error. The models were therefore evaluated as follows: model (1) was defined as the one considering the treatment effect alone, model (2) considered the time effect alone, and model (3) considered both effects combined. Beta diversity was calculated using Bray-Curtis dissimilarities instead, and ANOVA was later applied to assess the significance. Plots and tables were produced using the R package ggplot2 [59]. Rarefaction curves were calculated using the vegan package [101].

3.1.5: Description of the other databases included in the study

3.1.5.1: 16S-ITGDB

The 16S-ITGDB database, as cited [102], is designed to enhance the taxonomic classification of 16S ribosomal RNA. It is proposed as an integrating database merging from RDP (version NO.18 trainset) [103], SILVA version 138 (99% clustering similarity) [98], and Greengenes (version 13_8, 99% clustering similarity) [104]. The inclusion of this study was motivated by the intention to compare results and performance with the original databases subsequently. The database files were obtained from the official GitHub page (<https://drive.google.com/drive/folders/1UVRXhawL95a8kiBrPxpH-kTbfJWr3Oq4>) and were utilized in all pipelines, except for QIIME VERSION 2, where a locally custom-made file format was necessary.

3.1.5.2: Greengenes (version 13.8 and 2)

Despite not being updated since 2019, Greengenes version 13.9 remains of interest due to its popularity and incorporation into 16S-ITGDB, prompting its inclusion in our testing. Furthermore, the latest release was evaluated because of its very recent publication. Database files (fasta and taxonomy) for versions 13.8 and the classifier for version 2 were obtained from the official website (http://Greengenes.microbio.me/Greengenes_release/). The classifier needed for version 13.8 was sourced from the QIIME2 official website (<https://docs.QIIME2.org/2022.2/data-resources/>), while the database files for version 2 were generated locally from the official classifier following the instructions provided at <https://docs.QIIME2.org/2022.11/tutorials/exporting/>.

3.1.5.3: RDP

The RDP database was selected for its high citation frequency, ranking third after SILVA version 132 and Greengenes. version 13.8 of the database was employed, and the necessary data was acquired from the mothur website (https://mothur.org/wiki/rdp_reference_files/). A custom classifier was generated using the provided instructions.

3.1.5.4: SILVA version 138

Similarly to SILVA version 132 mentioned in subchapter 3.1.4, fasta and taxonomy files for SILVA version 138 were obtained from the official website (<https://www.arb-silva.de/download/archive/>). The files, as required by VSEARCH, were instead downloaded from the link provided here: <https://github.com/PacificBiosciences/pb-16S-nf/blob/main/README.md>, and the export process was executed following the instructions available at <https://docs.QIIME2.org/2022.11/tutorials/exporting/>.

3.2: Second section: testing of other 3 pipelines (of which 2 in 2 different approaches - clustering vs denoising) and 3 databases (of which 2 in 2 different versions) on 1 dataset

3.2.1: Other pipelines chosen details

The second section extends the study of the outcomes of the first study by adding the results obtained by the other pipeline and database combinations on the same dataset for comparison. It should be noted that sequencing and biostatistics were the same as described in the first section. Besides QIIME version 1.9, many pipelines are available for 16S rRNA gene analysis. In this study, 3 were selected and the studies were MICCA, QIIME VERSION 2, and VSEARCH. As mentioned previously, some of the pipelines allowed both the clustering and the denoising approach, and both were used when available (Figure 13).

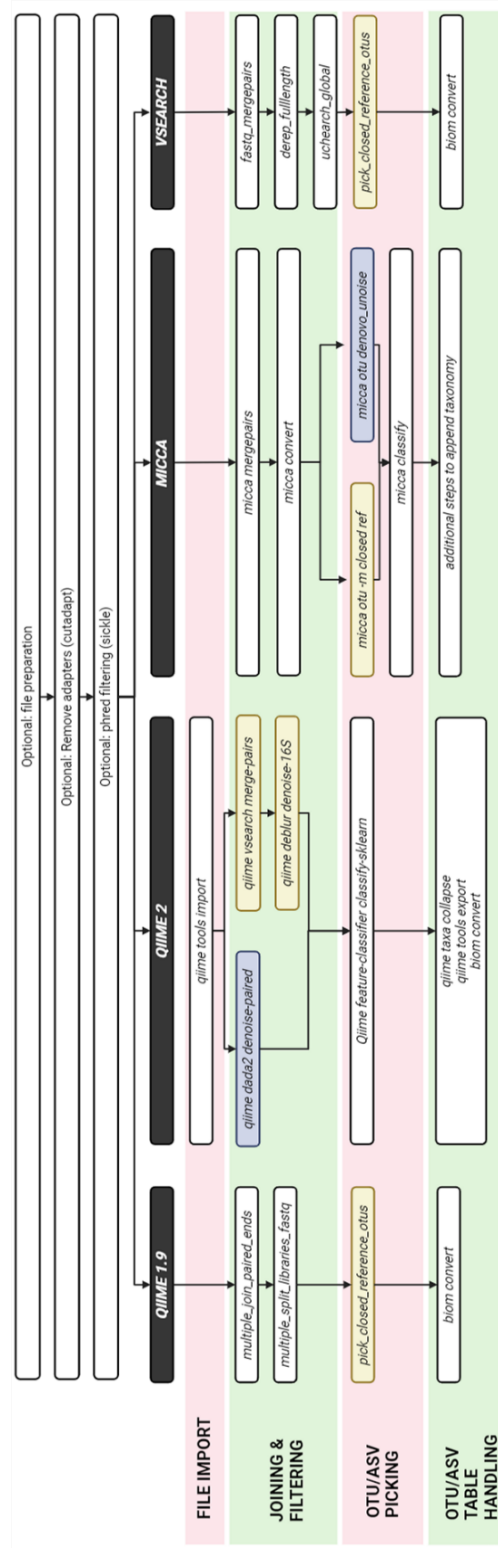


Figure 13: This organization chart shows the main steps required by each pipeline to deliver the OTU/ASV table. Pipelines with 2 different approaches included in their commands were flanked and branch accordingly. The command use is given by the bold names on the left, and pipelines are aligned by command similarity. Created with BioRender.com

3.2.1.1: MICCA

The MICCA pipeline was constructed following the guidelines outlined in the “Paired-end sequencing - 97% OTU tutorial” on the official website. The software container was obtained from Docker Hub at the link <https://hub.docker.com/r/compmetagen/MICCA/>. For joining reads, the MICCA *mergepairs* command was utilized. To maintain consistency across pipelines, parameters were set at *-l 32* (minimum overlapping length) and *-d 8* (maximum number of mismatches in the overlap region). Both parameters were adjusted to 0 for the mock community datasets. To circumvent the filtering step, merged fastq files were converted to fasta using the MICCA *convert* command, which does not require specific features. MICCA provides both clustering and binning functionalities, and both techniques were employed using their respective commands. OTU clustering was executed with the MICCA *otu* command using the closed reference method, denoted by the feature *-m closed_reference*. The reference database for each pipeline-database combination was specified, and a 99% identity threshold was set with the feature *-d 0.99*. Simultaneously, ASV denoising was performed using the same command but with *-m denovo_unoise* specified. Classification for the results of both commands was performed using the MICCA *classify* command with the feature *-m cons*, maintaining consistency across databases. The parameter *-cons-id* was set at 0.99 (default is 0.9). Taxonomy and OTU/ASV counts were subsequently merged using a customized Rscript.

3.2.1.2: QIIME version 2

The QIIME version 2 pipeline for this study was constructed based on the “Atacama soil microbiome” tutorial provided by the QIIME VERSION 2 development team. The QIIME VERSION 2 container was built following the instructions on the official website (<https://docs.QIIME2.org/2023.2/install/>). The initial step involved importing samples in a QIIME VERSION 2-compatible format using the QIIME VERSION 2 *import* command. The *PairedEndFastqManifestPhred33V2* was specified as the input-format parameter, and *SampleData[PairedEndSequencesWithQuality]* was set as the type parameter. The manifest for this step was custom-made in the script using basic UNIX commands. The pipeline then branched into two approaches: QIIME VERSION 2 with a classical approach, which followed the “Atacama soil microbiome” tutorial, and QIIME VERSION 2 with an alternative approach, provided separately by QIIME VERSION 2 developers on GitHub. The alternative approach allows users to bypass the dada2 approach and merge reads more similarly to other pipelines before proceeding, basically mimicking a clustering approach

when the QIIME VERSION 2 standard uses denoising (<https://github.com/QIIME2/docs/blob/master/source/tutorials/read-joining.rst>). In the classical pipeline, the QIIME *dada2 denoise-paired* command was employed, with features such as *--p-trim-left-f*, *--p-trim-left-r*, *--p-trunc-len-f*, and *--p-trunc-len-r* set at 0 to align with results from other pipelines.

The alternative pipeline merged the reads using the QIIME VSEARCH *merge-pairs* command. Subsequently, merged reads were deblurred using QIIME *deblur denoise-16S*, with the feature *--p-trim-length* set to 250, as suggested in Prodan et al. (Prodan et al., 2020).

For both alternatives of the QIIME VERSION 2 pipeline, the QIIME *feature-classifier classify-sklearn* command was used to train the classifier and then to classify the samples against the desired database using the QIIME *metadata tabulate* command. The ASV table was obtained through *taxa collapse* and *export*, with support from biom convert to convert the biom format into a tab-separated values (tsv) format.

3.2.1.3: VSEARCH

The VSEARCH pipeline utilized a container obtained from the link <https://hub.docker.com/r/biomehub/VSEARCH>, and guidelines were sourced from Torbjørn Rognes' GitHub in the Wiki section of the VSEARCH project (<https://github.com/torognes/VSEARCH/wiki>). Sequences were merged using the *fastq_mergepairs* command with the features *--fastq_maxdiffs* 8 and *--fastq_minovlen* 32. Subsequently, dereplication was performed sample by sample using the *--derep_fulllength* command and samples were merged using UNIX basic commands. Alignment of the samples to the databases was executed using the *usearch_global* command. Although *search_exact* would have better met the author's requests, it has the limitation of seeking perfectly identical sequences with the same length and sequence, as noted in Issue #295 on GitHub torognes/VSEARCH (<https://github.com/torognes/VSEARCH/issues/295>). To achieve the most identical alignment possible, resembling other pipelines, the *usearch_global* command was used with *--id* 1.0. Taxonomy information was appended using a customized Rscript.

3.3: Third section: testing of other 3 pipelines (of which 2 in 2 different approaches - clustering vs denoising) and 3 databases (of which 2 in different versions) on 5 datasets

The study's third section consisted of applying the full set of pipelines and databases (36 combinations) on datasets other than the milk dataset. Two kinds of datasets were used: mock communities and dairy cows' datasets. Mock communities aimed to test in replicate the combinations on the dataset that could remove the bias regarding experimental samples. Their characteristic is to be constituted by a defined set of microbes, already ascertained in other studies. Dairy cows' datasets, composed of skin, rumen, and fecal samples, aimed to repeat the same analysis on datasets that differed in time and methods of the sample collection and manipulation, as well as in the original aim for which they were collected, to mimic real-life research environment and necessities. Results were reported for mock communities as percentages of coherence of the results with the actual composition of the community, while for dairy cows' datasets correlation, calculated as Spearman correlation, between each pipeline and dataset results and every other, was reported, with its p-value as well. To the best author knowledge, no previous study was found reporting a similar approach to compare the results derived from the pipelines or the databases or the combination of both, therefore this should be an innovative although simple approach in benchmarking them. Lastly, to give the study a comprehensive overview of the influence of pipelines and databases on the results, dataset information was removed from the correlations results of milk, skin, rumen, and stool, and results were aggregated to analyze the pipeline effect alone, dataset effect alone and then pipeline and dataset effect together.

3.3.1: Other Dataset chosen details

3.3.1.1: *Mock community from Prodan et al., 2020 [105]*

Raw sequence data results from Prodan et al., 2020 [105] were retrieved from the publicly available repository Github (<https://github.com/andreiprodan/mock-sequences>). Data consists of three different sequencing runs of the Microbial Mock Community B (Even, Low concentration), v5.1L (Catalog no. HM-782D, obtained through BEI Resources, NIAID, NIH as part of the Human Microbiome Project). The authors do not further specify the

sequencing platform. The simulated sample comprises 22 variants (ASVs) from the V4 region of the 16S gene. When clustered at 97% identity, these sequences represent 19 OTUs. (Table 1).

Table 1: Composition of the mock community from Prodan et al., 2020, obtained by the sequencing of Genomic DNA from the Microbial Mock Community B (Even, Low concentration), v5.1L (Catalog no. HM-782D, obtained through BEI Resources, NIAID, NIH as part of the Human Microbiome Project.

Species	
<i>Acinetobacter baumannii</i>	<i>Listeria monocytogenes</i>
<i>Actinomyces odontolyticus</i>	<i>Neisseria meningitidis</i>
<i>Bacillus cereus</i>	<i>Propionibacterium acnes</i>
<i>Bacteroides vulgatus</i>	<i>Pseudomonas aeruginosa</i>
<i>Clostridium beijerinckii</i>	<i>Rhodobacter sphaeroides</i>
<i>Deinococcus radiodurans</i>	<i>Staphylococcus aureus</i>
<i>Enterococcus faecalis</i>	<i>Staphylococcus epidermidis</i>
<i>Escherichia coli</i>	<i>Streptococcus agalactiae</i>
<i>Helicobacter pylori</i>	<i>Streptococcus mutans</i>
<i>Lactobacillus gasseri</i>	<i>Streptococcus pneumoniae</i>

3.3.1.2: Mock community from Callahan et al., 2016 [106]

Since 2016, Mockrobiota has aimed to constitute a public resource for sharing mock community data or research aims [107]. The continuously updated collection of datasets and related metadata can be found on the project's GitHub page (<http://caporaso-lab.github.io/mockrobiota/>). For the aim of this study, Mock-12 was chosen for simplicity as it comes already demultiplexed. Mock-12 was published in Callahan et al., 2016 [106] under the name “Extreme”, and it is composed of 27 human gastrointestinal tract bacterial isolates, spanning six orders of magnitude in abundance. These isolates exhibited high genomic similarity, with some differing by as little as a single nucleotide within the quantified and pooled 16S rRNA gene amplicons derived from individual cultures. The bacterial strains that are detailed in Supplementary Table 2 of the above-cited paper and were amplified for V4 region (515f/806r primers) and sequenced on an Illumina MiSeq platform (Table 2). The Extreme dataset can also be found deposited by the European Nucleotide Archive (ENA) under the accession number PRJNA292120.

Table 2: Composition of the mock community from Callahan et al., 2016. It was included in the Mockrobiota project as Mock-12.

Species	
<i>Bacteroides cellulosilyticus</i>	<i>Coprococcus comes</i>
<i>Bacteroides eggerthii</i>	<i>Eubacterium rectale</i>
<i>Bacteroides fragilis</i>	<i>Howardella ureilytica</i>
<i>Bacteroides massiliensis</i>	<i>Parabacteroides distasonis</i>
<i>Bacteroides ovatus</i>	<i>Parabacteroides merdae</i>
<i>Bacteroides uniformis</i>	<i>Parabacteroides sp. D13</i>
<i>Bacteroides vulgatus</i>	<i>Paraprevotella clara</i>
<i>Barnesiella intestinihominis</i>	<i>Prevotella buccalis</i>
<i>Clostridium celatum</i>	<i>Prevotella copri</i>
<i>Clostridium cocleatum</i>	<i>Roseburia intestinalis</i>
<i>Clostridium methylpentosum</i>	<i>Roseburia inulinivorans</i>
<i>Clostridium phytofermentans</i>	<i>Ruminococcus gnavus</i>
<i>Clostridium xylanovorans</i>	

3.3.1.3: Skin samples from the MILKQUA project

Skin samples belong to the above-cited MILKQUA project from which milk samples belong. For skin samples instead, the collection was performed at three points: before TCEO application (T0A), after TCEO application (T0B), and once on T7. In detail, an area of 1 cm² of the quarter was wiped with a humid tissue, then dried and rubbed using a skin swab (R. Langenbrinck GmbH, Emmendingen, Germany) by slightly rotating the swab through an angle of 45° for 10 seconds. Samples were then stored at -80°C, as cited above. Skin samples are required using the QIAamp PowerFecal Pro DNA Kit (Qiagen, Hilden, Germany), with minor modifications according to the manufacturer's protocol. Sequencing was performed with primers targeting the V3-V4 hypervariable regions of the 16S rRNA gene (forward: 5'-CCTACGGGNGGCWGCAG-3', reverse: 5'-GACTACHVGGGTATCTAATCC-3') [97] on an Illumina MiSeq platform. The 16S rRNA gene raw sequences obtained were deposited in UNIMI Dataverse (https://doi.org/10.13130/RD_UNIMI/GMPFOB, https://doi.org/10.13130/RD_UNIMI/OPMKOC).

3.3.1.4: Rumen samples from Ranilla et al., 2023

This dataset derives from the hereby presented work published in Scientific Reports: the study design is provided comprehensively to elucidate the source of the analyzed samples [108]. The samples in this study were constituted by rumen fluids collected from four different Holstein dairy cows at the slaughterhouse, filtered, mixed with 4 times their buffer volume, and then incubated with controlled doses of testing compounds plus diet. As testing compounds, the natural essential oil of *Thymbra capitata* (**NEO**), synthetic essential oil (**SEO**), plus three pure compounds, namely carvacrol, p-cymene, and γ -terpinene, were tested against a control (no additive). The final mixes of rumen fluids with the compound were sealed in bottles using rubber stoppers and aluminum caps and incubated at 39 °C for 24 hours. Two replicates per group were assayed (Figure 14).

Samples were evaluated for total gas production, volatile fatty acids (**VFA**), and ammonia-N concentration and analyzed by gas chromatography. Moreover, the content of the vials was collected and frozen at -80°C. DNA extraction from those samples was performed using QIAamp Fast DNA Stool Mini Kit (Qiagen) following the manufacturer's instructions. Bacterial DNA was amplified by targeting the V3-V4 hypervariable regions of the 16S rRNA gene, using V3-V4 targeting primers 5'-CCTACGGGNGGCWGCAG-3' as forward and 5'-GACTACHVGGGTATCTAATCC-3' as reverse [109]. The PCR amplification of each sample was performed in a 25- μ L volume, using 12.5 μ L of KAPA HIFI Master Mix 2 $\times$ (Kapa Biosystems, Inc., MA, USA). Then, 0.2 μ L of each primer (100 μ M) was added to 2 μ L of genomic DNA (5 ng/ μ L). Non-DNA templates (blank controls) were included. Amplification and library quantification were carried out as described previously [94]. Sequencing was performed using an Illumina MiSeq platform.

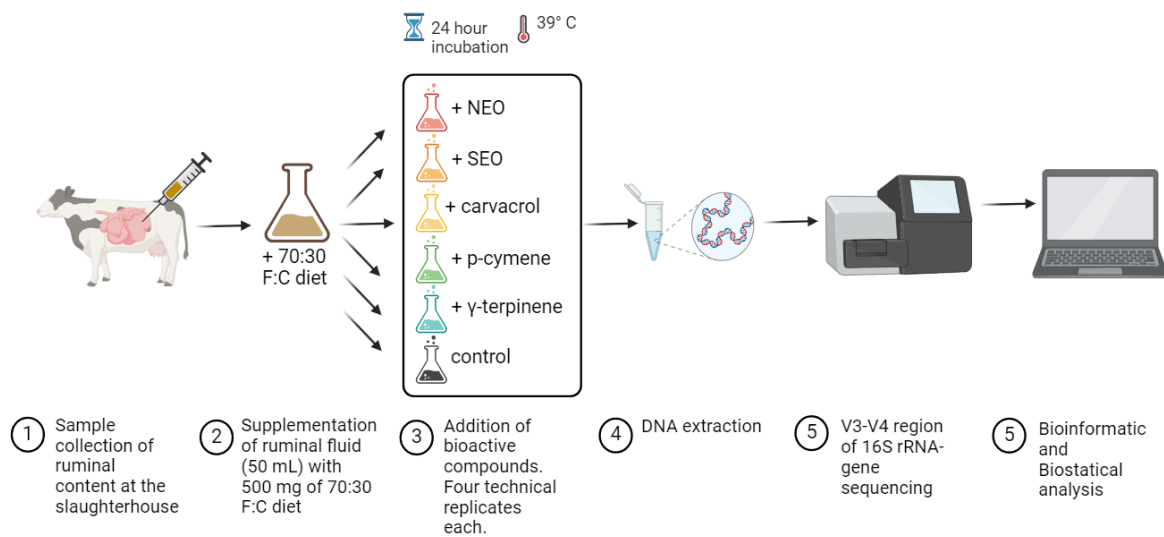


Figure 14: From Ranilla et al., 2023. Infographic of the study regarding the evaluation of microbiota on rumen fluids combined with essential oils compounds or pure. Created with BioRender.

3.3.1.5: Fecal samples from Andrés et al., 2024

The hereafter presented investigation was carried out by the Spanish National Research Council (Consejo Superior de Investigaciones Científicas, **CSIC**). The study design is reported for completeness to describe the origin of the samples analyzed. All data belong to Andrés et al., 2024 and were obtained following the recommendations of Directive 2010/63/EU of the European Parliament, the Council on Protecting Animals used for scientific purposes, and the IGM-CSIC Animal Experimentation Committee (protocol 736/2018), along with those in the ARRIVE guidelines [110].

In brief, CSIC performed a trial using sixteen new-born dairy calves that were homogeneous per live body weight (**LBW**) and day of birth. The selected animals were separated from the cows within the first day of life and fed 3 L of colostrum daily until day 3 of life. Then, they were housed individually in straw-bedded hutches and equally divided into two groups (control group and essential oil, or EO, group) while being fed milk replacer (**MR**, 145 g of Novilac Turbostart, Schils/L) twice a day until day 45, increasing the dose coherently with the average weight growth. In this frame, animals in the EO group received MR enriched with 0.23 ml of oregano EO, accounting for 200 mg of carvacrol (Zane Hellas 100%) diluted in the first 100 mL of MR. After 45 days of life, calves were managed in a feedlot using an automatic feeding machine until weaned at 70 days and fed MR until accumulating a dry matter intake of 54 kg per day, plus being allowed

ad libitum to a starter feed compound. After weaning, animals were fed a total mixed ration (TMR-1) formulated to cover their requirements during the first post-weaning period (up to 6 months of life) and a second TMR (TMR-2) with a higher amount of fiber during the replacement phase (6-12 months of life), both without EO enrichment. Monthly weighting of the animals was carried out during the replacement phase.

Throughout the study, three-time points were selected for blood and two for fecal collection: 3 days (T0), 45 days (T1), and 13 months old (T2), and 45 days (T1) and 13 months old (T2), respectively. Blood samples were used for metabolomics analysis, biochemical profiling, and evaluation of peripheral blood mononuclear cells (PBMCs) markers such as CD4+, CD8+, CD21+, WC1+, and CD14+. Fecal samples were instead collected for microbiota analysis and total apparent digestibility evaluation and, therefore, collected as rectal grab samples before the morning meal and stored at -80°C until they were freeze-dried.

According to the manufacturer's protocol, the QIAamp Fast DNA Stool Mini Kit (Qiagen, Germany) was used to extract DNA from frozen fecal samples. Sequencing was performed on an Illumina MiSeq platform, using 5'-CCTACGGGNGGCWGCAG-3' and 5'-GACTACHVGGGTATCTAATCC-3' primers as cited above, as forward and reverse respectively.

Chapter 4: Results

4.1: First section: testing of 1 pipeline and 1 database on 1 dataset and comparing the results with the same database in a newest version and other 3 databases (of which 1 in 2 different versions)

4.1.1: Sequencing results

This subchapter outlines the results regarding the pilot experiment: in detail, the QIIME version 1.9 pipeline was coupled with the SILVA version 132 database to analyze the milk dataset from the MILKQUA project. Subsequently, the outcomes were compared to results obtained by applying the same dataset and pipelines with the other databases, including SILVA version 138, RDP, Greengenes version 13.8 and 2, and 16S-ITGDB.

Sequencing the V3-V4 regions of the bacterial 16S rRNA gene from 64 milk samples resulted in 6 264 103 assembled reads (joined R1-R2 paired-end reads). After quality filtering, 341 996 sequences were excluded, leaving 5 922 107 sequences for subsequent analyses. The average retention rate was 94.6%, with a maximum of 99.8% and a minimum of 86.3%. The operational taxonomic units (OTUs) that were identified were 28712.

4.1.2: Taxonomy obtained using SILVA version 132

4.1.2.1: Results at phylum level

To guarantee the best database functioning, milk microbiota was filtered for taxa belonging to the Bacteria kingdom. Analysis at the phyla level exhibited four main phyla with a relative abundance greater than 5% across different time points: Firmicutes, Proteobacteria, Actinobacteria, and Bacteroidetes (Figure 15).

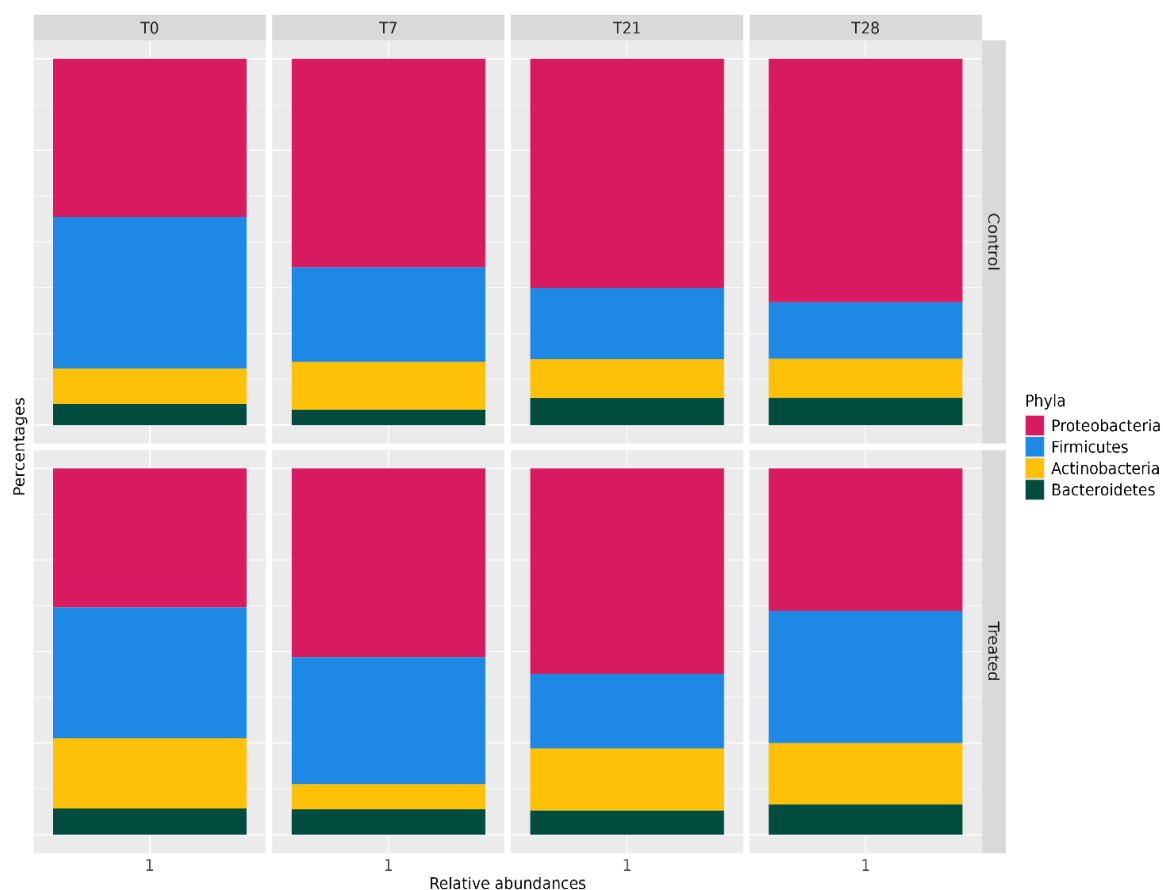


Figure 15: Boxplots of the distribution of relative abundance higher than 5% in the milk microbiota (all 64 samples together) at the phylum level

However, six other phyla showed significant differences between treatments over time. Interestingly, four phyla, Proteobacteria, Gemmatimonadetes, Thermotogae, and Aerophobetes, decreased at T21 in the Treated group compared to the Control group (Figure 16), while at T7, two phyla, Nitrospirae and Armatimonadetes, were found increased.

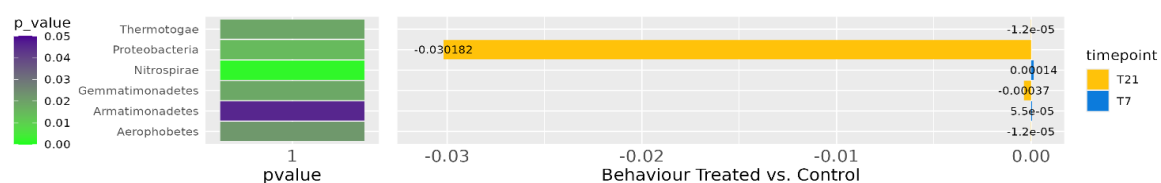


Figure 16: Significantly abundant taxa in milk microbiota at phyla level: on the left, the heatmap of the significance of taxa relative abundance, while on the right, the bar plot representing the behaviour of the significantly different abundant phyla along time points: the scale in the x-axis is the ratio of Treated against Controls. Please note that some changes are not visible due to the contraposition in scale between Proteobacteria and the other phyla. Therefore, values are reported on the bars.

4.1.2.2: Results at genus level

At the genus level instead, samples showed 24 genera with a relative abundance greater than 5% across different time points. Among them, a group denominated "Uncultured or unknown" group was artificially composed for statistical analysis, comprehending all the genera that the database (SILVA version 132) retrieved as "uncultured" "uncultured bacterium," "Other," or "uncultured organism" and similar (Figure 17).

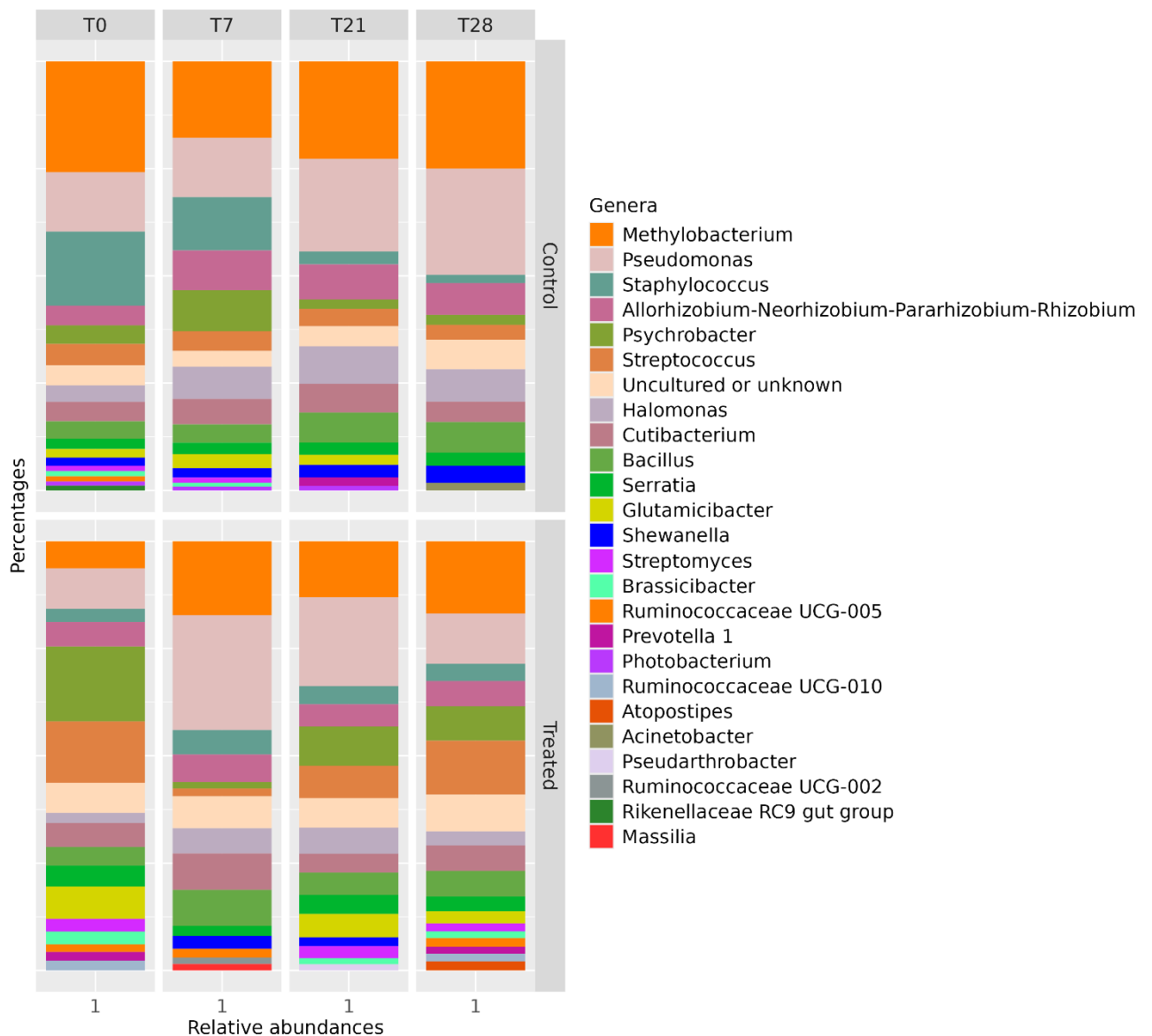


Figure 17: Boxplots of the distribution of relative abundance higher than 5% in the milk microbiota (all 64 samples together) at the genus level.

Of the 72 genera showing significant differences between treatments over time, 22 belonged to T0. It was found to increase in the Treated group compared to the Control

group, with exceptions made for *Aestuariibacter*, *Deinococcus*, and *Saccharofermentans*, which were found to decrease. Similarly, at T0, 3 out of 4 genera were found to increase (*Brevibacterium*, *Candidatus Amoebophilus*, and *Catenuloplanes*), and 1 out of 4 decreased (*Fron Dihabitans*). Of the 22 genera found significantly different at T21, 8 decreased and 14 increased, while 12 decreased and 12 increased at T28 (Figure 18).

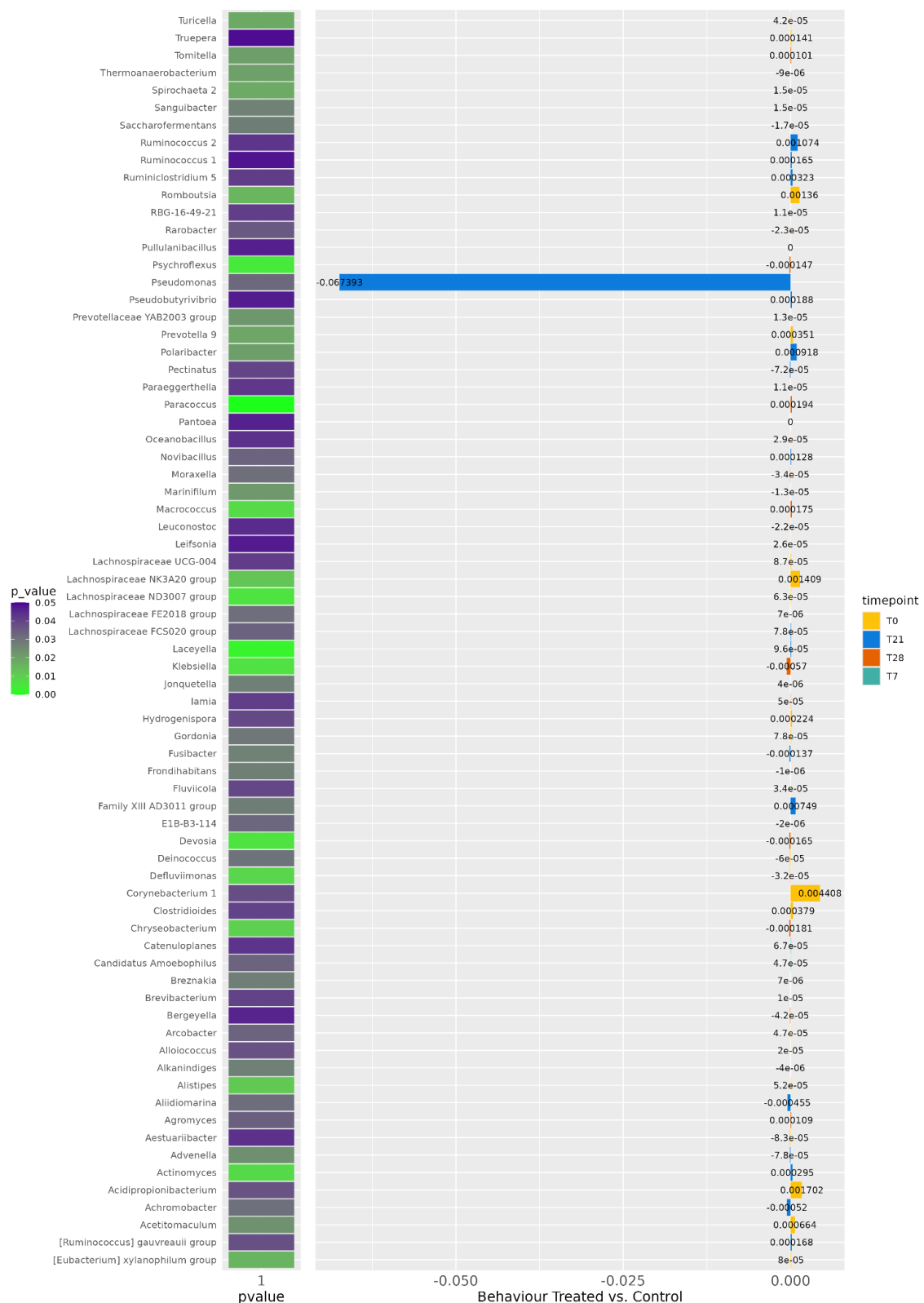


Figure 18: Significantly abundant taxa in milk microbiota at genus level: on the left, the heatmap of the significance of taxa relative abundance, while on the right, the bar plot representing the behaviour of the significantly different abundant phyla along time points: the scale in the x-axis is the ratio of Treated against Controls.

Overall, the microbial milk community showed significant differences between the Treated and Control groups across multiple time points.

4.1.3: Alpha and beta diversity using SILVA version 132

As previously described, alpha and beta diversity indices were generated using the phyloseq R package. These indices were then further analyzed within the R environment. For alpha diversity, two variations of the model were applied: one considering both time and treatment effects together, and another considering only the time effect or only the treatment effect. On the other hand, Beta diversity was calculated using Bray-Curtis dissimilarities and subsequently assessed for significance using ANOVA.

Once corrected to a baseline equal to T0 (milk samples collected before treatment application), alpha indices were compared in the three previously mentioned (Material and Methods section) linear models to assess their behavior. Figure 19 reports the scatterplots of i) the significance (P) of treatment at T7, T21, and T28 (Figure 19, top left), ii) the significance (P) of the time point within treatment (Figure 19, top right), iii) the significance (P) of time point and treatment from the model (3) (Figure 19, bottom left), for the different alpha diversity indices.

From model (1), the effect of time is significant: 1 index (Shannon) has a $P < 0.05$ on T21, and 1 index in the significance range between 0.05-0.10 (Simpson), plus five indices with a $P < 0.05$ in T28 (Observed_OTUs, Chao1, ACE, Shannon, and Fisher) and 1 index in the significance range between 0.05-0.10 (Simpson). From model (2), indices clusters are similar for Treated and Control groups, with only 2 indices in the Control group found in the 0.05-0.10 range of significance (Observed_OTUs and Fisher). From model (3), treatment has a clear effect over time points, with two diversity indices found with P lower than 0.05 (Shannon and Simpson) and 4 indices in the 0.05-0.10 range (Observed_OTUs, Chao1, ACE, and Fisher) for the Treated group compared with the Control, influenced by time effect. In contrast, only 1 index (Fisher) had $P < 0.05$ at T28. Regarding beta diversity evaluation, the relationships between samples were assessed based on Bray-Curtis dissimilarities, as mentioned above. Figure 19 (bottom right) shows the distribution of samples along the first two dimensions from the multidimensional scaling (MDS) of Bray-Curtis dissimilarities: a clustering was observed between time points ($P = 0.002$), while no clear clustering by treatment nor treatment + time was detected (P

= 0.07 and $P = 0.06$ from PERMANOVA between treatment, time, and treatment + time effects combined, 999 permutations).

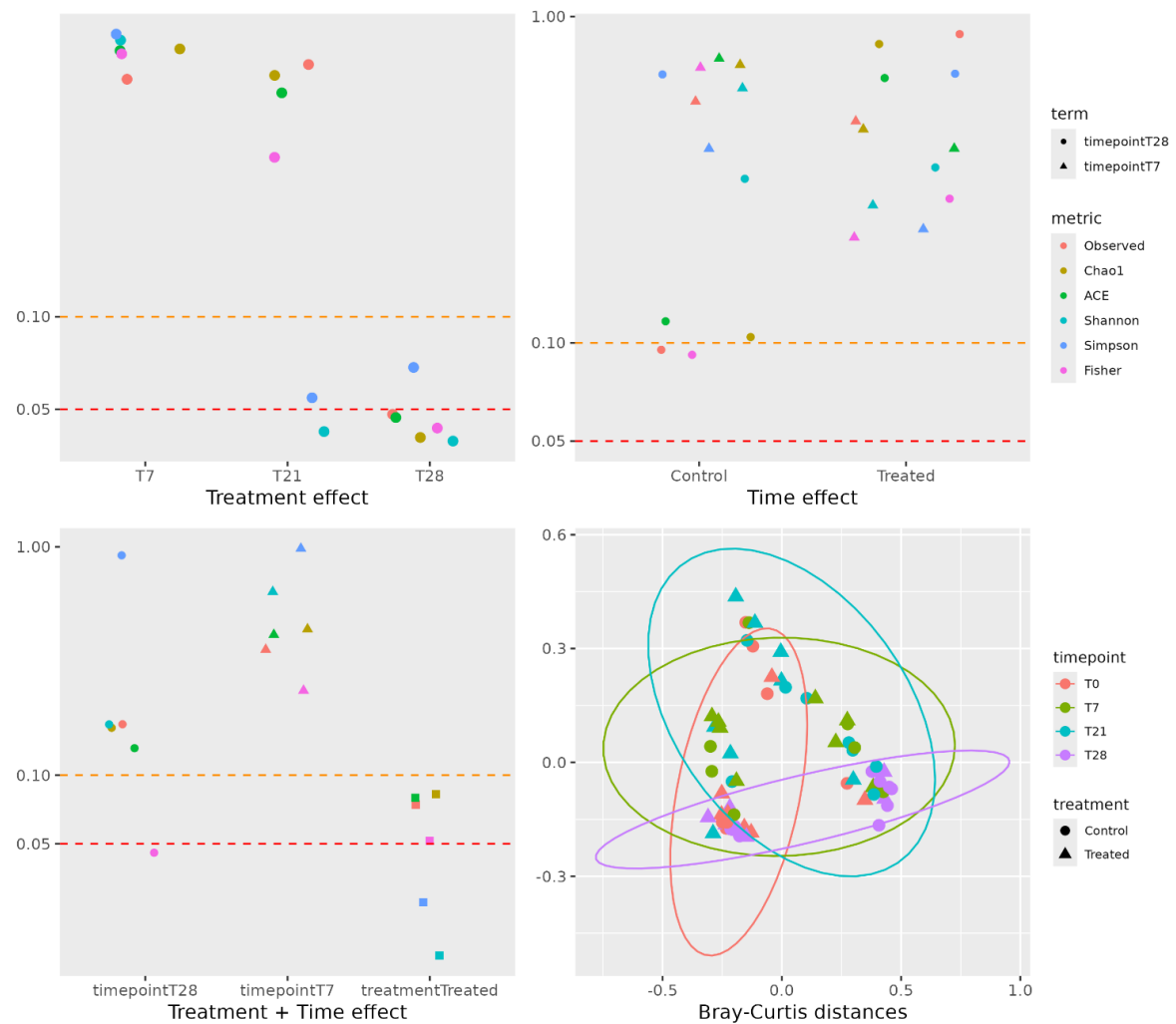


Figure 19: Scatterplots for alpha diversity of milk samples evaluating treatment effect alone (top left), time effect alone (top right), and treatment + time effect (bottom left). MDS represents the Multidimensional scaling (MDS) of Bray-Curtis dissimilarities for milk samples. Multidimensional scaling (MDS) plot of Bray-Curtis distances based on the OTU table from the milk microbiota (bottom right).

4.1.4: Performance metrics using SILVA version 132 and with the same database in a newest version and other 3 databases (of which 1 in 2 different versions)

4.1.4.1: Alpha diversity indices from 4 databases (of which 2 in 2 different versions) on the same dataset analyzed with the same pipeline

The alpha diversity analysis was reproduced using the same dataset, milk samples from the MILKQUA project, and the same pipeline, QIIME VERSION 1.9, but different databases: namely, 16S-ITGDB, Greengenes version 13.8 (gg_13_8), RDP, Greengenes version 2, and SILVA version 138. The models applied to the results obtained were used for SILVA version 132 (Figure 20).

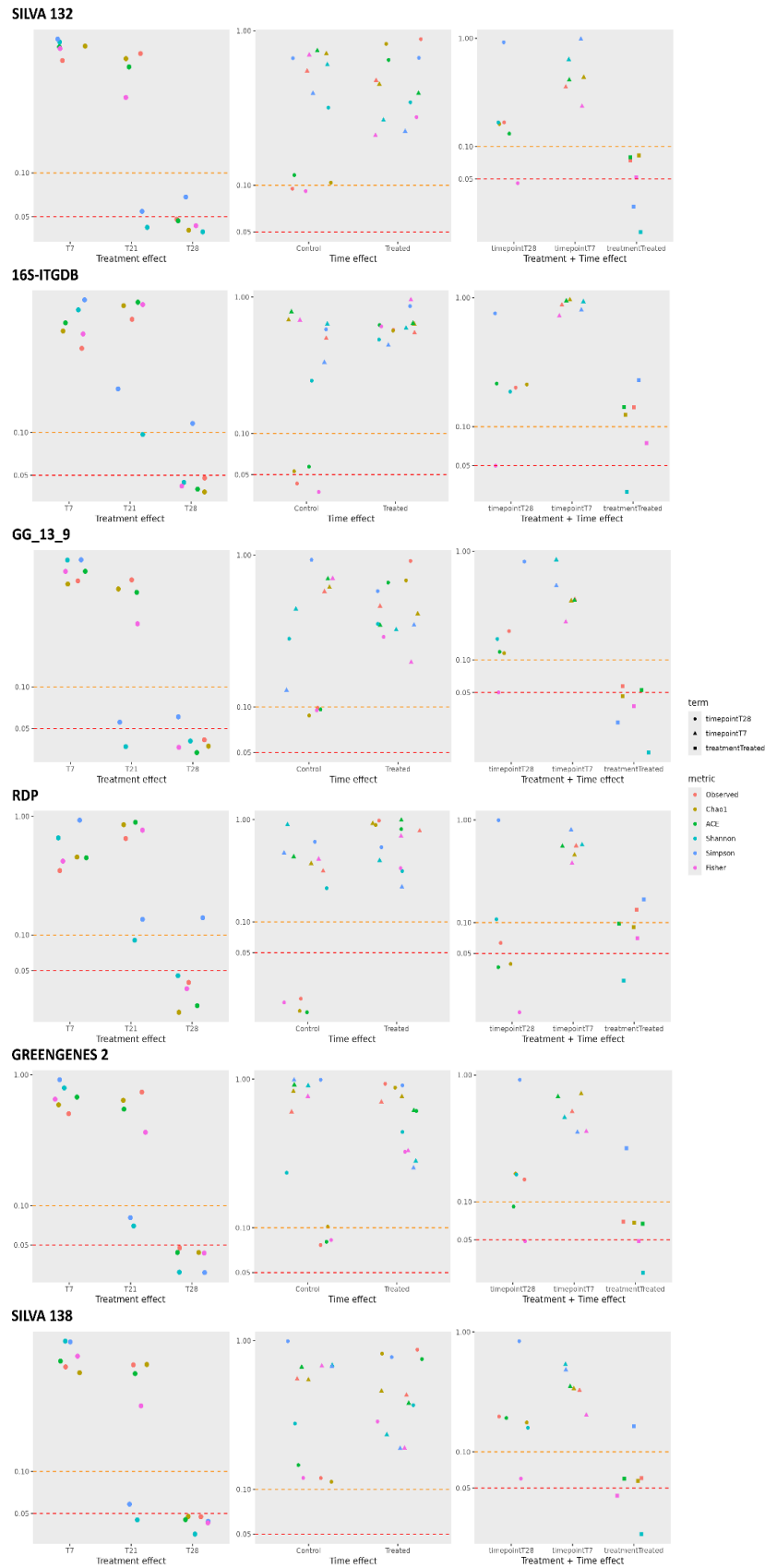


Figure 20: Comparison of the scatterplots for alpha diversity of milk samples evaluating treatment effect alone, time effect alone, and treatment + time effect derived for every database using QIIME VERSION 1.9 pipeline on milk dataset.

It can be observed that model (1) delivered the same results with no significance found at T7, while at T21, the results are not homogeneous. RDP and Greengenes version 13.8 reported no index with $P < 0.05$, while others reported one (Shannon). On the other hand, results were similar regarding T28, where all the results reported 5 out of 6 indices with $P < 0.05$.

Model (2), considering time effect alone, reported different outcomes depending on the database used, with SILVA version 132, Greengenes version 13.8, and SILVA version 138 identifying no index with $P < 0.05$, contrary to RDP and 16S-ITGDB that found 4 and 2 indices with $P < 0.05$, respectively. At the same time, coherent results regarding no index found significant in the Treated group were shared by all results.

Lastly, model (3) results were shared regarding only 1 index (Fisher) in the significance area ($P < 0.05$) for timepoint T28 by all the databases except SILVA version 138, even though RDP also had other 2 (ACE and Chao1). Results for timepoint T7 were instead coherently showing no significance, while diverged on when considering time influence in the model: in fact, while Greengenes version 13.8 reported 4 indices with $P < 0.05$, the other databases had 1 or 2 indices in the same area, showing a deep variation of the database influence on the results in this case.

4.1.4.2: Beta diversity indices from 4 databases (of which 2 in 2 different versions) on the same dataset analysed with the same pipeline

Bray-Curtis similarity indices were used to study beta diversity results as delivered by the outputs of different databases on the same dataset using the same pipeline. Results reported in Figure 21 show that along the different databases, samples tend to cluster similarly, especially by time point rather than by treatment.

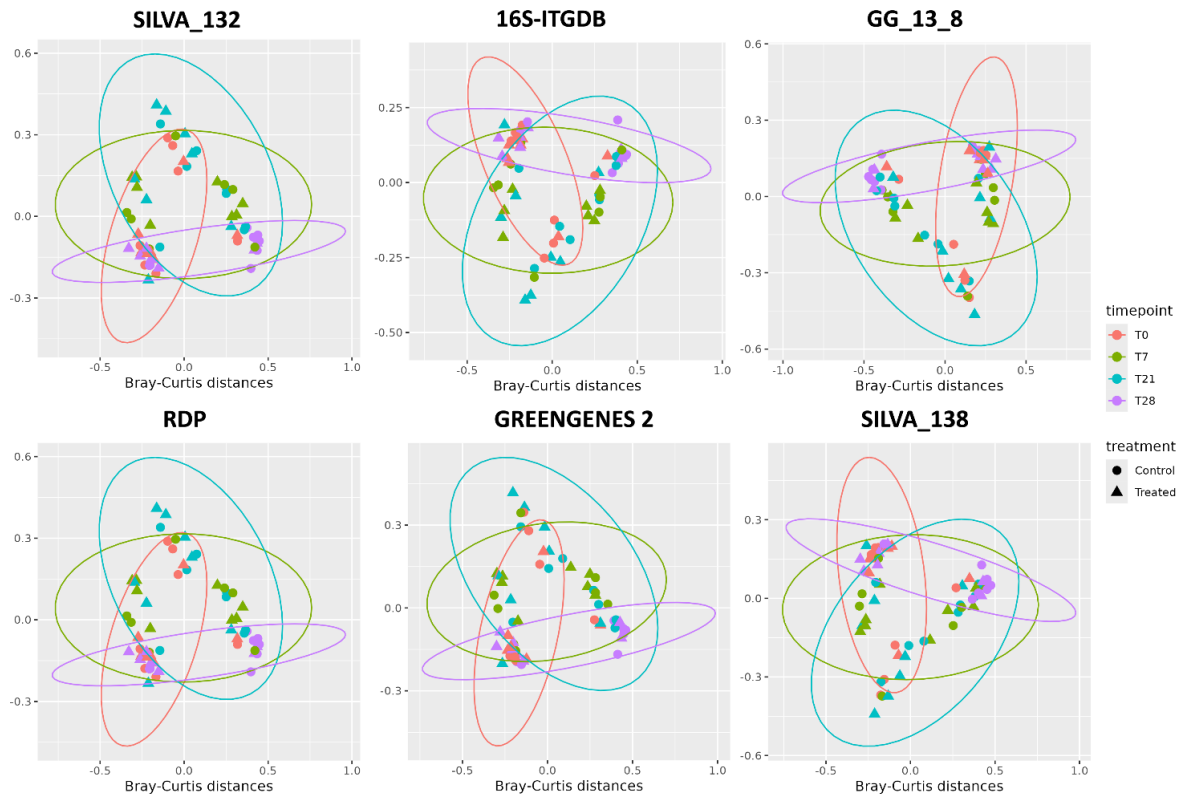


Figure 21: Comparison of the MDS plots for beta diversity of milk samples evaluating for every database using QIIME version 1.9 pipeline on milk dataset.

The observation was tested using PERMANOVA with 999 permutations, and the results are reported in Table 3. It can be observed that while the significance of clustering between time points crosses all the databases, P values related to treatment clustering differ, with 16S-ITGDB results alone reporting significance. Similarly, P values result significant for treatment + time effect only in 16S-ITGDB and Greengenes version 13.8, while other results do not highlight a $P < 0.05$.

Table 3: Table of the results obtained by testing with PERMANOVA with 999 permutations every database result when applied to QIIME version 1.9 pipeline on the milk dataset. Results include P values for significance, F values as distribution values, R squared values as determination coefficient and omega squared values as effect size values.

	16SITGDB				Greengenes version 13.8				Greengenes version 2				RDP				SILVA version 132				SILVA version 138			
	P value	F value	R ²	Ω ²	P value	F value	R ²	Ω ²	P value	F value	R ²	Ω ²	P value	F value	R ²	Ω ²	P value	F value	R ²	Ω ²	P value	F value	R ²	Ω ²
Between treatment	0,03	2,66	0,04	0,02	0,07	2,15	0,03	0,02	0,05	2,21	0,03	0,02	0,08	2,05	0,03	0,02	0,07	1,93	0,03	0,01	0,08	1,86	0,02	0,01
Between timepoints	0,001	3,03	0,13	0,08	0,002	2,83	0,12	0,08	0,002	3,06	0,13	0,08	0,002	3,1	0,13	0,08	0,001	2,87	0,13	0,08	0,001	2,77	0,12	0,07
Between treatment + time points	0,15	1,4	0,05	0,02	0,11	1,44	0,06	0,75	0,09	1,46	0,06	0,02	0,18	1,37	0,06	0,02	0,12	1,41	0,06	0,02	0,09	1,52	0,06	0,02

4.2: Second section: testing of other 3 pipelines (of which 2 in 2 different approaches - clustering vs denoising) and 3 databases (of which 2 in 2 different versions) on 1 dataset

4.2.1: Filtered and unfiltered results

The Milk dataset belonged to the project MILKQUA mentioned above and was used to study the coherence among results between the different pipelines and dataset combinations.

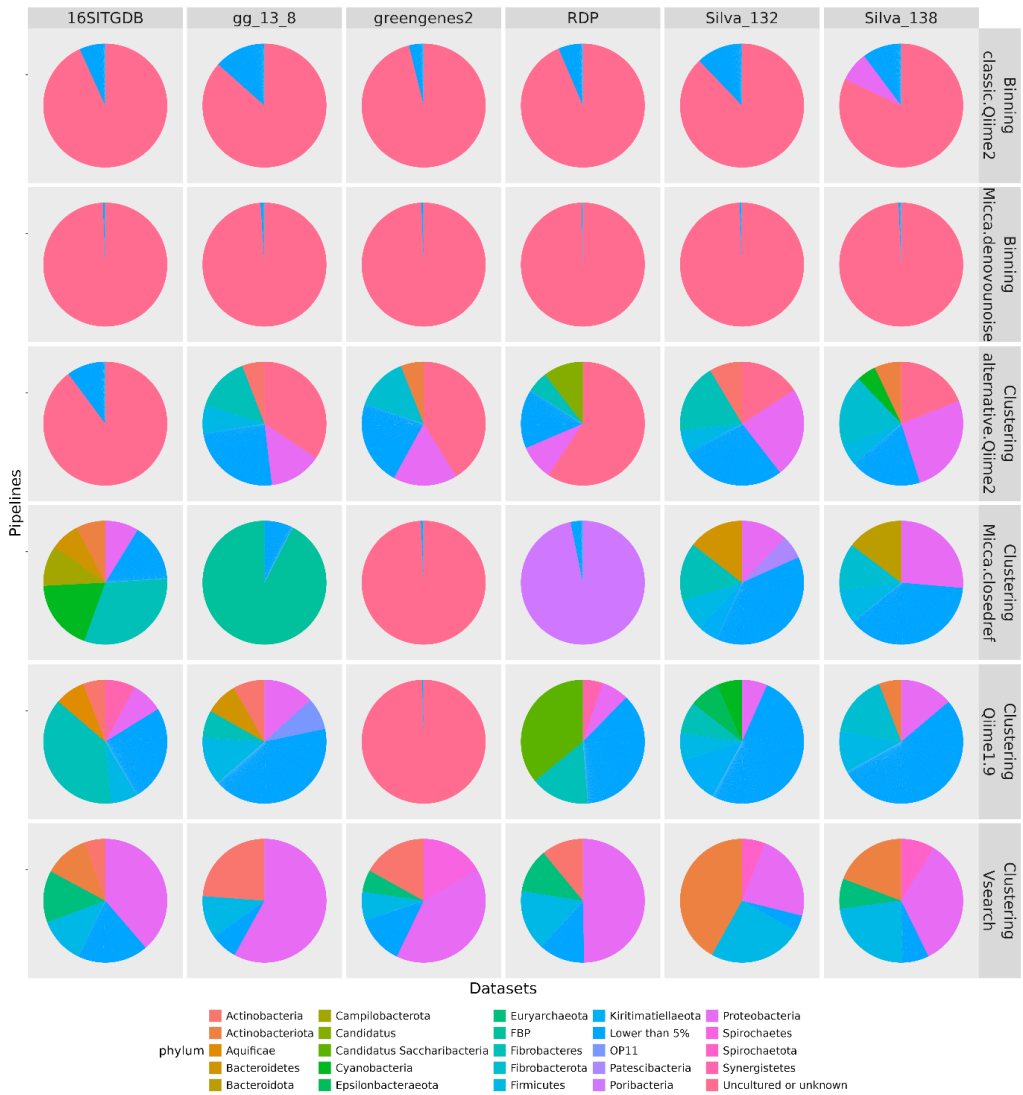


Figure 22: Pie charts of the unfiltered results regarding the microbial composition of the milk samples belonging to the MILKQUA project at the phylum level

Figure 22 represents the unfiltered result at the phyla level, while Figure 23 presents filtered results for Uncultured or Unknown taxa that were removed. It is interesting to observe that pipelines related to denoising methods were the ones who most benefited from this cleaning step.

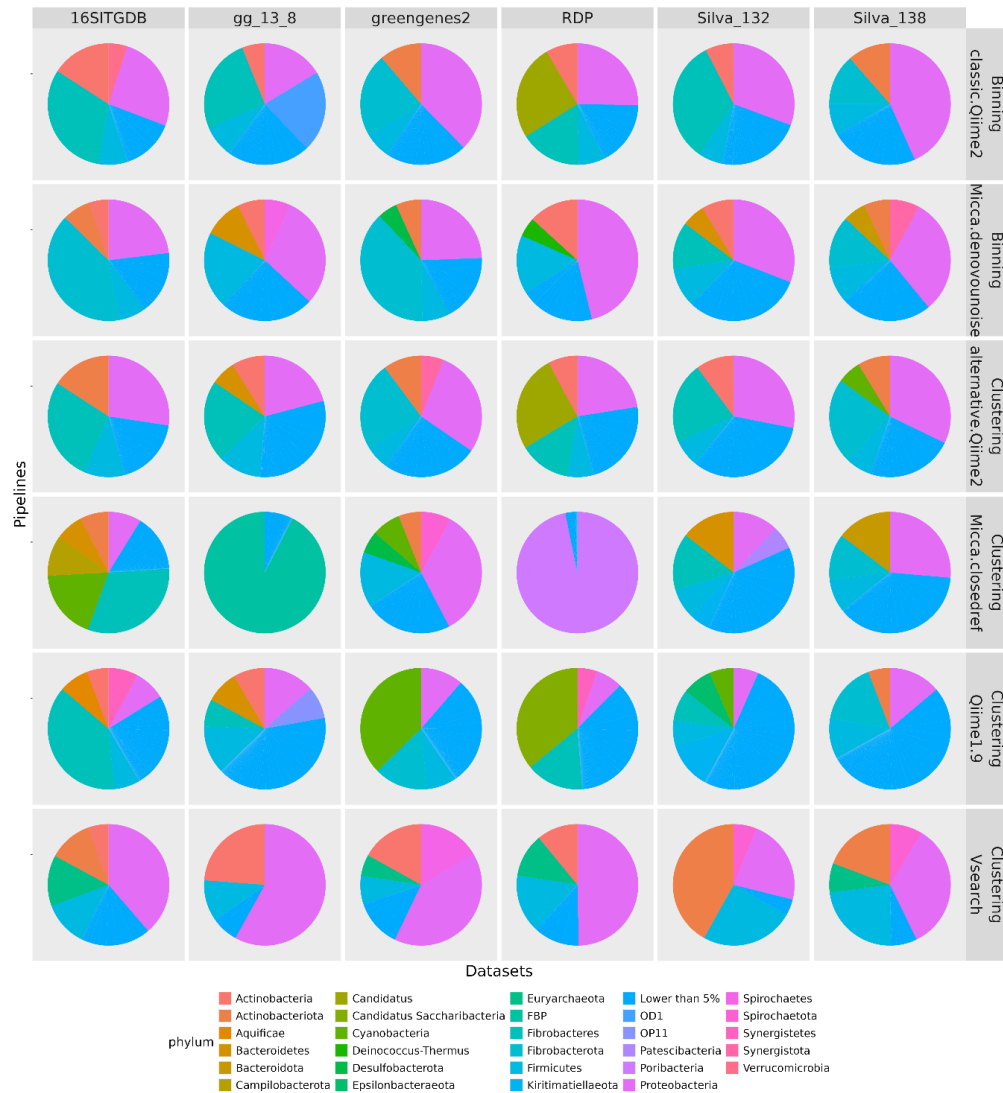


Figure 23: Pie charts of the filtered results regarding the microbial composition of the milk samples belonging to the MILKQUA project at the phylum level

4.2.2: Correlations and P values of the comparisons among pipeline x database results

The Spearman correlation test was performed to test each pipeline and dataset combination against each other. Figure 24 shows the correlation coefficients obtained, and the p-value of the correlations.

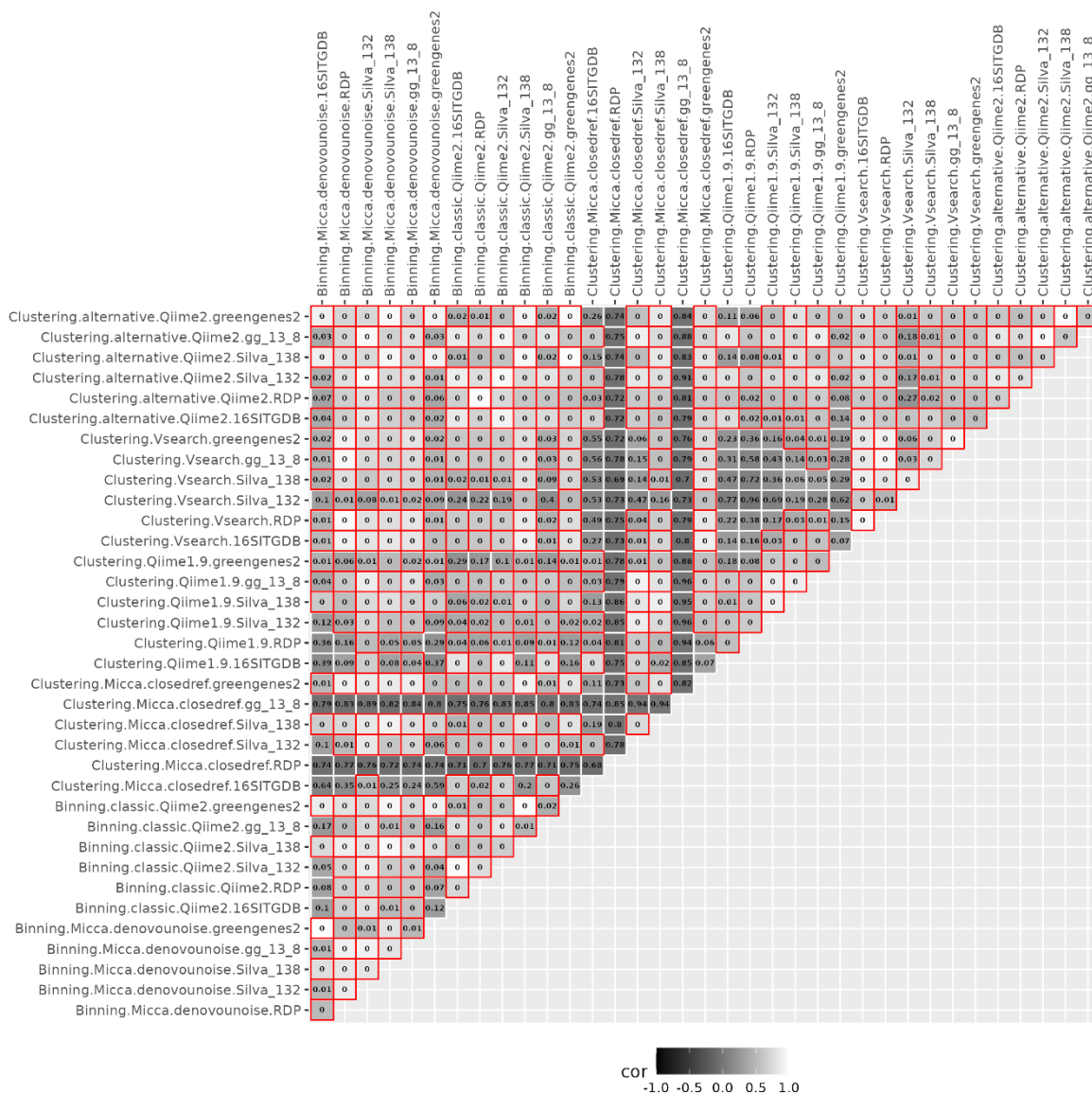


Figure 24: Heatmap of the correlation values of each pipeline + database combination outputs against each other in the milk dataset. Color range from black (negative correlation, -1) to white (positive correlation, +1). Numbers inside the tiles are the p values of the correlation. Tiles containing a p value lower than 0.05 were contoured in red.

It is noticeable that the MICCA pipeline with a closed reference feature paired with RDP and Greengenes version 13.9 database retrieved the lowest correlations and, therefore, the less significant p values with all the other pipelines. Interestingly, when correlation values were filtered for results higher than 90%, two pipelines had higher correlations: MICCA with the denovo-unoise approach and QIIME2 with the classic approach, both belonging to the binning methods. Among the databases, Greengenes 2 seemed to favor a higher correlation, followed by SILVA version 138, SILVA version 132, and 16SITGDB. The highest number of correlations higher than 90% was achieved by MICCA with the de novo approach paired with SILVA version 138 (5 hits), followed by QIIME version 2 classic paired with SILVA version 138 (4 hits).

4.3: Third section: testing of other 3 pipelines (of which 2 in 2 different approaches - clustering vs denoising) and 3 databases (of which 2 in different versions) on 5 datasets

4.3.1: Mock communities from Prodan et al., 2020

The mock communities from Prodan et al. [105] were expected to represent 21 bacterial strains, of which genera and species are known, and therefore family was deducted. To assess each pipeline + database performance, the output was compared with the expected taxa at both taxonomic levels. The comparison results are reported in Figure 25 for the family level and Figure 26 for the genus level.



Figure 25: Heatmap of the percentages of correspondence of family taxonomic level expected versus obtained by each pipeline + database combination on mock communities from Prodan et al., 2020.

Among all the pipelines + database combinations, it is interesting to notice that at family level the clustering approaches outperformed the binnings, especially over QIIME version 2.

it is interesting that VSEARCH + Greengenes2 could not identify any expected genera and had the lowest performance at family leve. Outputs from QIIME1.9 and MICCA seemed to offer higher completeness in identifying the genera present, especially when equipped with 16SITGDB, SILVA version 132, and SILVA version 138 databases. Instead, QIIME2

classic and alternative pipelines seemed to deliver results that represent scarcely the accurate sample composition.



Figure 26: Heatmap of the percentages of correspondence of genera expected versus obtained by each pipeline + database combination on mock communities from Prodan et al., 2020.

Moreover, the percentage of error at genera and species level were calculated, along with the number of OTU/ASV found, and three alpha diversity indices (Table 4).

Table 4: Table of the percentages of error at genera and species level, number of OTU/ASV found, and three alpha diversity indices (Chao1, Shannon and Fisher) derived from the results of the 36 pipeline + database combinations on the mock communities from Prodan et al., 2020.

Pipeline	Database	% error genera level	% error family level	N° of OTU/ASV found	Alpha index: Chao1	Alpha index: Shannon	Alpha index: Fisher
QIIME version 1.9	16S-ITGDB	66,67%	0,00%	1166	1634,9	2,3	121,5
	Greengenes version 13.9	50,00%	0,00%	2653	3985,3	2,7	305,7
	Greengenes version 2	8,33%	0,00%	9578	16145,6	2,8	1327,9
	RDP	41,67%	0,00%	488	651,9	2,2	46,2
	SILVA version 132	33,33%	0,00%	108	4862,5	2,3	372
	SILVA version 138	50,00%	0,00%	119	4540,6	2,1	345,3
QIIME version 2 classic	16S-ITGDB	16,67%	77,78%	3	3	0,9	0,3
	Greengenes version 13.9	16,67%	66,67%	3	3	0,9	0,3
	Greengenes version 2	16,67%	66,67%	3	3	0,9	0,3
	RDP	16,67%	61,11%	3	3	0,9	0,3
	SILVA version 132	16,67%	88,89%	3	3	0,9	0,3
	SILVA version 138	16,67%	83,33%	3	3	0,9	0,3
QIIME version 2 alternative	16S-ITGDB	16,67%	83,33%	10	10	0,8	0,7
	Greengenes version 13.9	33,33%	66,67%	16	16	1,2	1,2
	Greengenes version 2	16,67%	66,67%	16	16	1,2	1,2
	RDP	41,67%	66,67%	16	16	1,3	1,2
	SILVA version 132	33,33%	88,89%	12	12	0,4	0,8
	SILVA version 138	41,67%	83,33%	12	12	0,4	0,8
Micca closedref	16S-ITGDB	50,00%	0,00%	119	156,4	1,9	15,2
	Greengenes version 13.9	33,33%	0,00%	407	532,1	2,3	63,7
	Greengenes version 2	16,67%	0,00%	346	475,1	2,1	52,6
	RDP	41,67%	0,00%	28	31	1,5	2,9
	SILVA version 132	33,33%	0,00%	510	637,6	2,5	83,4
	SILVA version 138	41,67%	0,00%	464	573,1	2,5	74,5
Micca denovounoise	16S-ITGDB	50,00%	5,56%	16	16	1,6	1,6
	Greengenes version 13.9	33,33%	66,67%	16	16	1,6	1,6

	Greengenes version 2	16,67%	22,22%	16	16	1,6	1,6
	RDP	33,33%	38,89%	16	16	1,6	1,6
	SILVA version 132	33,33%	5,56%	16	16	1,6	1,6
	SILVA version 138	41,67%	11,11%	16	16	1,6	1,6
Vsearch	16S-ITGDB	50,00%	0,00%	36	51	1,7	4,1
	Greengenes version 13.9	16,67%	50,00%	108	125,1	1,7	14,2
	Greengenes version 2	8,33%	88,89%	124	148,8	1,7	16,6
	RDP	41,67%	5,56%	21	27	1,5	2,2
	SILVA version 132	16,67%	22,22%	108	138	1,7	14,2
	SILVA version 138	16,67%	16,67%	101	125,4	1,7	13,1

Coherently with the expectations, the percentage of error increased at genus level, whereas it was already high at genera level. It is interesting to notice that denoising approaches had higher error percentages, while clustering approaches matched better with the expectations: however, this result is due to the higher number of OTUs retrieved by clustering approaches. In fact, clustering approaches reported very high number of OTUs found in comparison to those expected, and this enhance the probability of finding a correct, but quite fortuitous, match with the expected taxa. On the other hand, denoising approaches identified a much lower number of ASVs, that is closer to the expected, although as said had a worse matching.

Alpha diversity indices reflected the OTU/ASV number trend, showing higher values for clustering approaches than for denoising ones.

The OTU/ASV tables from each combination were also fed to vegan R package to visualize the rarefaction curves. Results were grouped by clustering and denoising approaches (Figure 27).

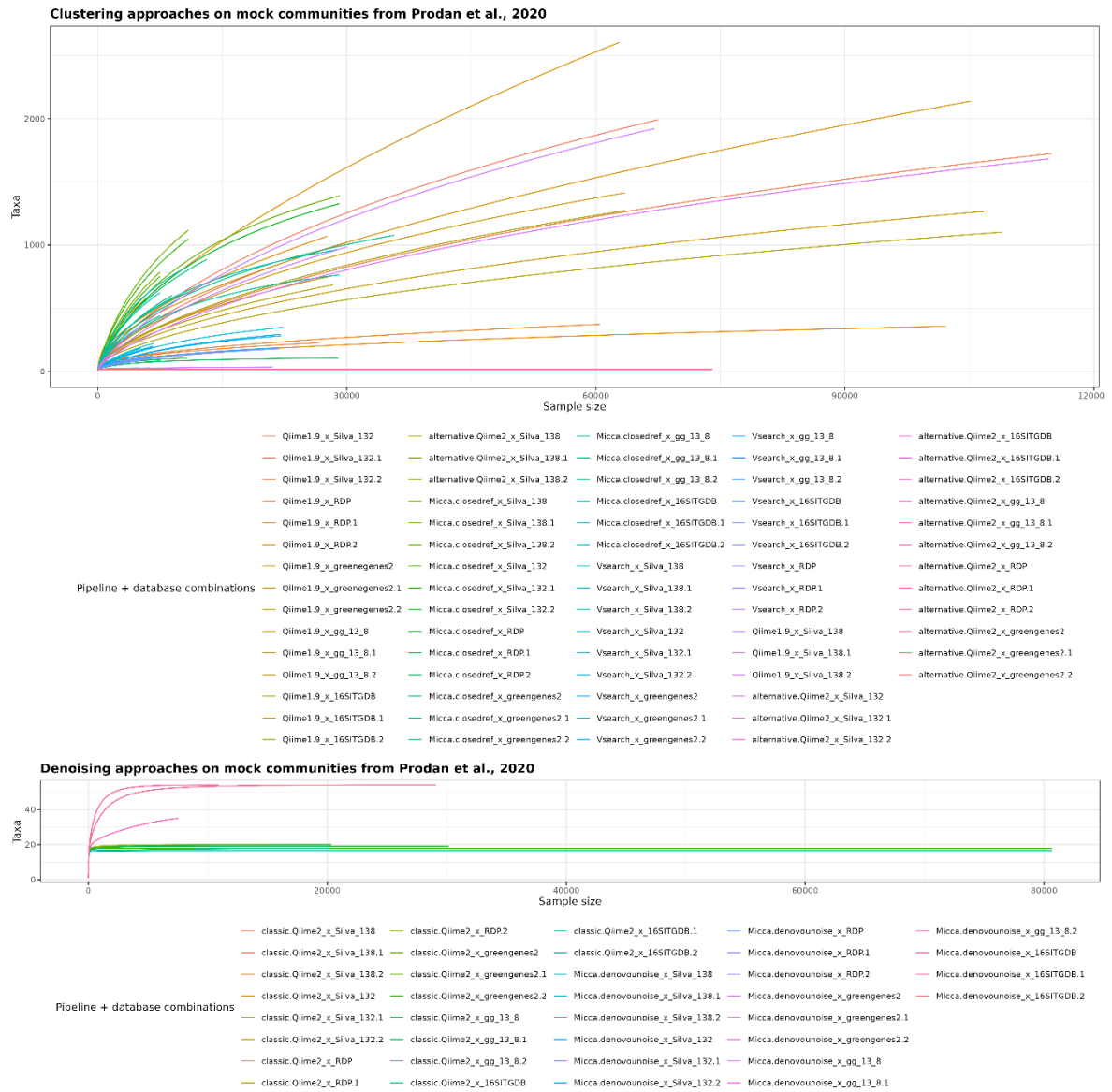


Figure 27: Rarefaction curves of the results of pipeline + database combinations over Prodan et al., 2020 mock community. Results were grouped per approach: the figure above presents clustering methods rarefaction curves, while the figure below presents denosing method rarefaction curves. Triplicates per each combination represent the 3 mock communities contained in Prodan et al., 2020 dataset.

The outcomes of the rarefaction curves showed the denosing approaches reached the plateau much faster than clustering approaches, where clustering approaches seemed instead to require a deeper inspection although they present a much higher number of species identified.

4.3.2: Mock communities from Callahan et al., 2016 (Mockrobiota project)

The mock community from the Mockrobiota project was Mock 12, from the Extreme dataset in Callahan et al., 2016 [106], and contained 27 bacteria, of which genera and species are known and family information were deducted. It must be noted that *Clostridium xylanovorans* was found to be *Anaerocolumna xylanovorans* because of the recent reclassification [111]. Most of the pipelines performed well in *Bacteroides*, *Roseburia*, *Anaerocolumna*, *Clostridium*, and *Parabacteroides* identification at the genus level, exceptions made for QIIME version 2 with a classical approach, and the pipelines related to the Greengenes version 2 database (Figure 28).



Figure 28: Heatmap of the percentages of correspondence of family expected versus obtained by each pipeline + database combination on Mockrobiota mock communities.

At family level, the pipeline + database combinations performed the best in identifying Bacteroidaceae and Lachnospiraceae, while all failed on Barnesiellaceae and Eubacteriaceae. Clustering approaches presented slightly better results than binning approaches, although throughout identifications did not correspond to the expected.

The pipeline + database combinations identified none of the other bacteria at the genus level by the pipeline + database combinations, except for *Parabacteroides*, which were determined by 7 combinations out of the 36 studied (Figure 29).

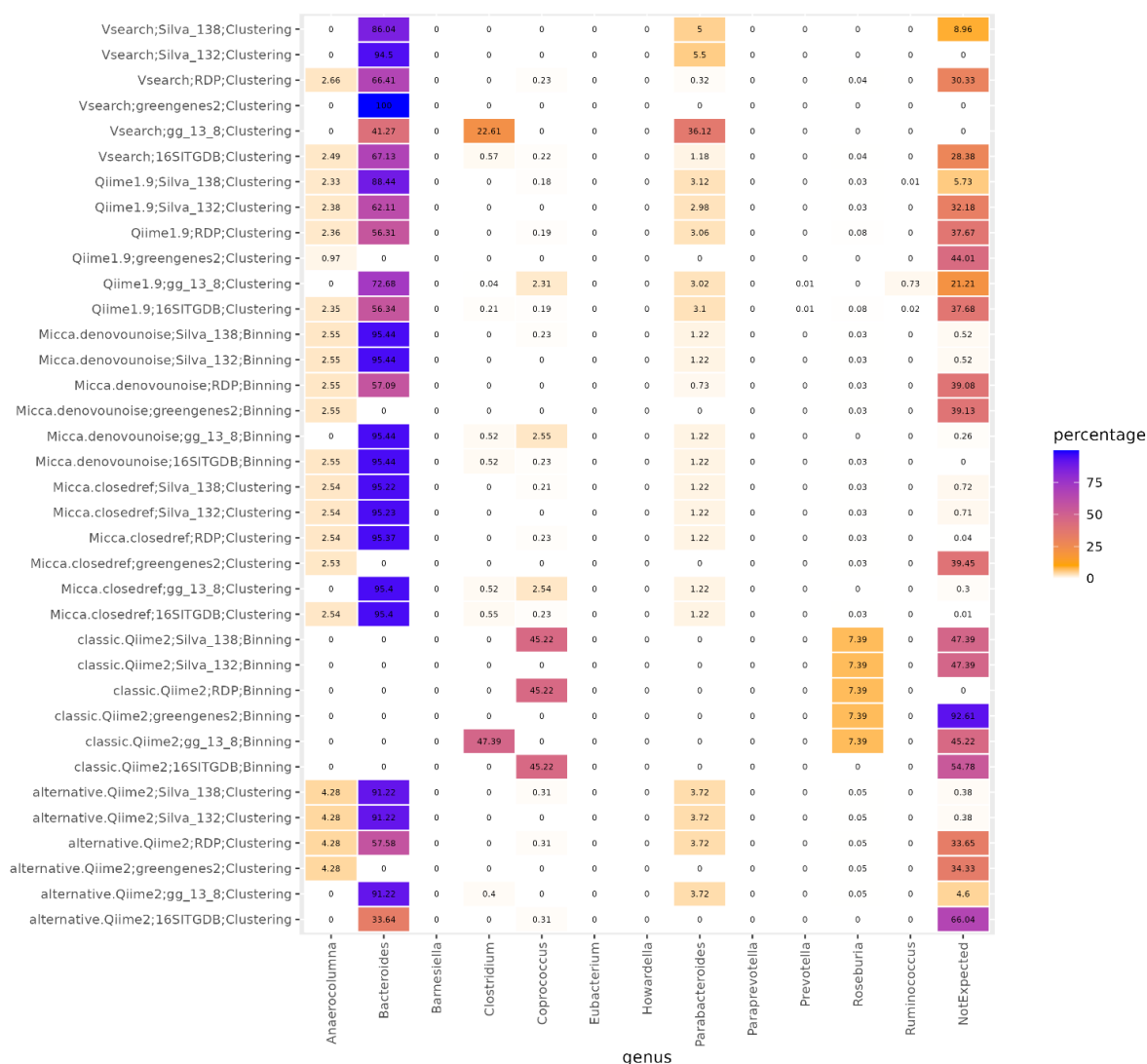


Figure 29: Heatmap of the percentages of correspondence of genera expected versus obtained by each pipeline + database combination on Mockrobiota mock communities.

Moreover, the percentage of error at genera and species level were calculated, along with the number of OTU/ASV found, and three alpha diversity indices (Table 5).

Table 5: Table of the percentages of error at genera and species level, number of OTU/ASV found, and three alpha diversity indices (Chao1, Shannon and Fisher) derived from the results of the 36 pipeline + database combinations on the mock communities from Callahan et al., 2016 (Mockrobiota project).

Pipeline	Database	% error genera level	% error family level	N° of OTU/ASV found	Alpha index: Chao1	Alpha index: Shannon	Alpha index: Fisher
QIIME version 1.9	16S-ITGDB	66,67%	33,33%	1916	1851,7	3,9	174,2

(continues)

	Greengenes version 13.9	50,00%	33,33%	2154	2076,6	3,9	200,6
	Greengenes version 2	8,33%	44,44%	4158	4764,5	4,5	383,3
	RDP	41,67%	33,33%	557	504,1	3,6	44,5
	SILVA version 132	33,33%	22,22%	3037	2933,6	3,8	292,6
	SILVA version 138	50,00%	22,22%	2945	2893,6	3,5	281,4
QIIME version 2 classic	16S-ITGDB	16,67%	77,78%	19	17	2,5	1,7
	Greengenes version 13.9	16,67%	77,78%	21	18,6	2,7	1,9
	Greengenes version 2	16,67%	77,78%	20	19	2,7	1,9
	RDP	16,67%	77,78%	20	19	2,7	1,9
	SILVA version 132	16,67%	77,78%	20	19	2,7	1,9
	SILVA version 138	16,67%	77,78%	19	19	2,7	1,9
QIIME version 2 alternative	16S-ITGDB	16,67%	55,56%	16	16	2,5	1,6
	Greengenes version 13.9	33,33%	55,56%	18	18	2,7	1,8
	Greengenes version 2	16,67%	44,44%	19	19	2,7	1,9
	RDP	41,67%	55,56%	19	19	2,7	1,9
	SILVA version 132	33,33%	44,44%	19	19	2,7	1,9
	SILVA version 138	41,67%	44,44%	18	19	2,7	1,9
Micca closedref Micca denovounoise	16S-ITGDB	50,00%	44,44%	1286	1167,4	3,7	187,4
	Greengenes version 13.9	33,33%	55,56%	973	883,6	3,5	130,2
	Greengenes version 2	16,67%	44,44%	1460	1369	3,7	189,9
	RDP	41,67%	55,56%	137	117,4	2,8	14,3
	SILVA version 132	33,33%	33,33%	1778	1622,6	3,9	260,1
	SILVA version 138	41,67%	33,33%	1873	1686,3	3,9	279,2
	16S-ITGDB	50,00%	66,67%	53	48,3	2,8	6,2
	Greengenes version 13.9	33,33%	66,67%	53	48,3	2,8	6,2
	Greengenes version 2	16,67%	55,56%	53	48,3	2,8	6,2
	RDP	33,33%	66,67%	53	48,3	2,8	6,2
	SILVA version 132	33,33%	55,56%	53	48,3	2,8	6,2

(continues)

	SILVA version 138	41,67%	55,56%	53	48,3	2,8	6,2
Vsearch	16S-ITGDB	50,00%	44,44%	232	199,3	2,8	22,1
	Greengenes version 13.9	16,67%	77,78%	240	216,2	2,8	23,1
	Greengenes version 2	8,33%	88,89%	397	390,7	2,9	38,7
	RDP	41,67%	66,67%	42	44,1	2,6	4
	SILVA version 132	16,67%	77,78%	392	396,3	2,9	39,4
	SILVA version 138	16,67%	66,67%	448	420,8	2,9	46,6

The percentages of error showed, as expected, an increase of error at species level when compared to genera level, as it is well known that a deeper insight into taxonomic level require technique efforts that at today still are not matched completely. On the other hand, identified OTU and ASV number show that high variability that pipelines can add to their results. It is interesting to notice that clustering approaches gave higher number of OTUs, while denoising had lower numbers coming much closer to the expected ones. Similarly, alpha indices reported higher values for clustering approaches then for denoising ones.

The OTU/ASV tables from each combination were also fed to vegan R package to visualize the rarefaction curves. Results were grouped by clustering and denoising approaches (Figure 30).

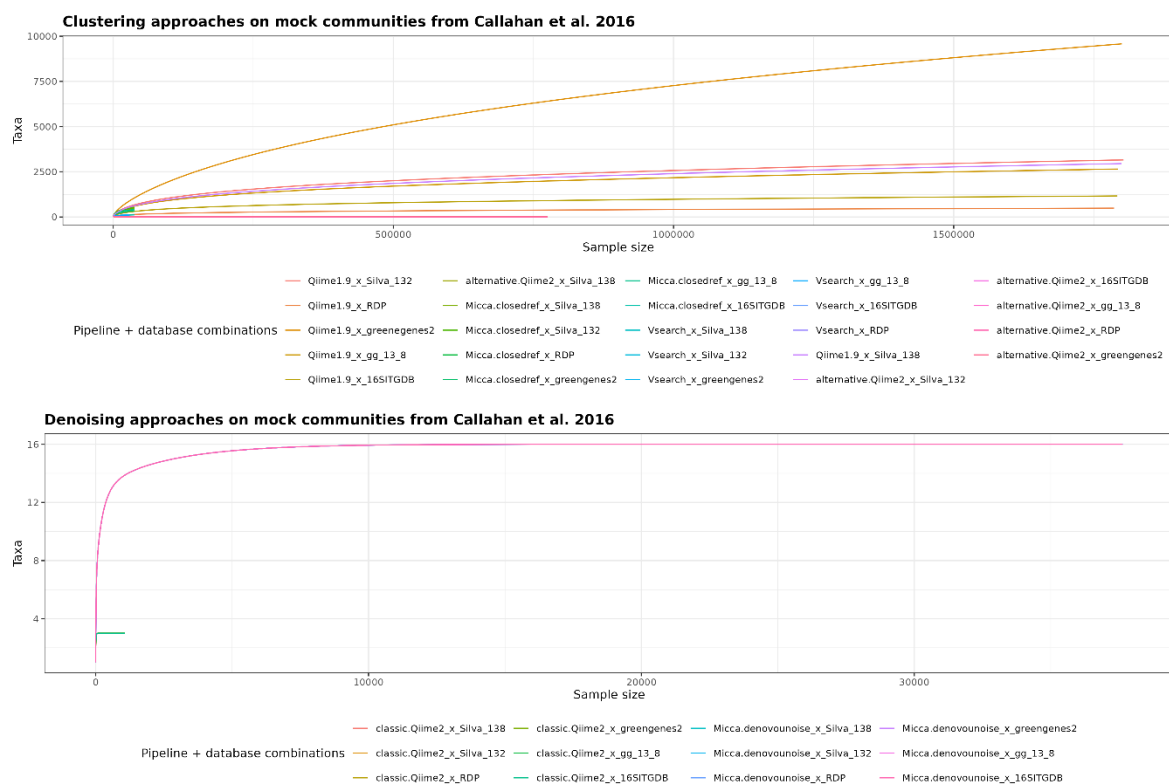


Figure 30: Rarefaction curves of the results of pipeline + database combinations over Callahan et al., 2016 mock community. Results were grouped per approach: the figure above presents clustering methods rarefaction curves, while the figure below presents denoising method rarefaction curves.

Similarly to the other mock community results, rarefaction curves regarding denoising approaches reached a plateau much faster and with a lower number of species than clustering approaches.

4.3.3: Skin samples from the MILKQUA project

The skin dataset belongs to the MILKQUA project as the milk dataset., but sampling was more closed-up and restricted to the first 7 days. Samples were sequenced in the same Illumina Miseq platform and underwent the pipelines + databases comparison. To compare the outcome of each pipeline and dataset combination against every other, the Spearman correlation test was used. Figure 31 shows the correlation coefficients obtained and the p-value of the correlations.

From the heatmap, it is quite straightforward that most of the pipelines and dataset combinations correlated positively, apart from QIIME VERSION 2 in the classic approach

paired with RDP. However, QIIME VERSION 2 in the alternative approach reported slightly worse results in correlation than other combinations, showing several 0.4-scored correlations. On the other hand, MICCA, using the denovo approach, and QIIME version 2 showed good correlations. It is interesting to notice some patterns present in correlations regarding MICCA and QIIME version 2: in these two cases, pipelines paired with the same databases but using a different approach (clustering versus binning) still correlated better than other combinations, suggesting that the pipeline algorithm strongly influences the outcomes of the results. Parallely, SILVA 138 and Greengenes version 13.8 seem keener in guaranteeing good correlations along different pipelines, as expected.



Figure 31: Heatmap of the correlation values of each pipeline + database combination outputs against each other in the skin dataset. Color range from black (negative correlation, -1) to white (positive correlation, +1). Numbers inside the tiles are the p values of the correlation. Tiles containing a p value lower than 0.05 were contoured in red.

The correlations are confirmed by the significant p-values spread all over the combinations. Again, the non-significant combinations are lined on the row representing the QIIME version 2 with binning approach paired with RDP database, the VSEARCH paired with RDP compared with QIIME version 1.9 paired with the same database. Interestingly, even though the dataset and the database correspond, the pipeline highly influences the results, so the p-value is less significant among the observed.

4.3.4: Rumen samples from Ranilla et al., 2023

The samples from Ranilla et al. were analyzed using the pipelines mentioned above combined with different databases, and correlation was assessed using the Spearman method.

Negative correlations are mostly reported related to comparisons with combinations including VSEARCH pipeline, especially when paired with SILVA version 138, but are moderate (lowest correlation value was -0.26 in QIIME version 2 with alternative approach paired with Greengenes version 13.8 compared against Vsearch paired with SILVA 138 database). Moreover, patterns of positive correlation were present along QIIME version 2 in the two approaches and followed the same database. Nonetheless, QIIME version 2 in the clustering approach also showed good correlations with MICCA in the clustering approach, but again based on using the same database (SILVA version 132, SILVA version 138, and Greengenes version 13.8). Interestingly, the same pipelines and databases showed similar behavior (high correlation) even when the approach was shifted to binning (Figure 32).

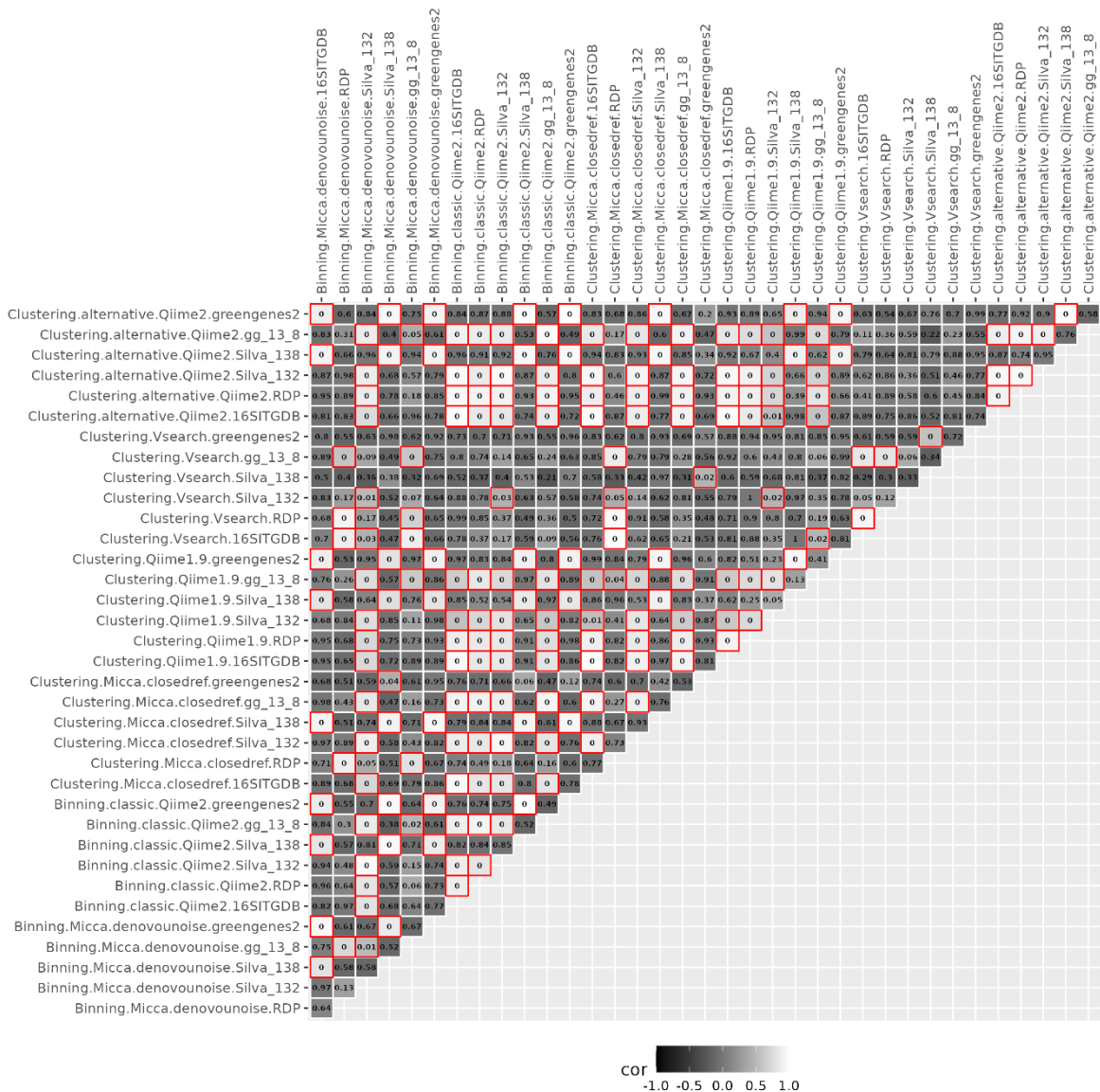


Figure 32: Heatmap of the correlation values of each pipeline + database combination outputs against each other in the rumen dataset. Color range from black (negative correlation, -1) to white (positive correlation, +1). Numbers inside the tiles are the p values of the correlation. Tiles containing a p value lower than 0.05 were contoured in red.

The heatmap regarding the p-values of the correlations obtained comparing each pipeline and dataset outcome against every other allows us to observe a pattern of significant p-values in the correlation with QIIME VERSION 2 with the binning approach and QIIME VERSION 1.9 and Micca with the clustering approach. Indeed, any database except SILVA version 138 seemed to favor high p-values in all three pipelines. Moreover, VSEARCH in the clustering approach reported the lowest number of significant p-values of all the correlations.

4.3.5: Fecal samples from Andrès et al., 2024

The results of the Spearman correlation test applied to the fecal dataset outlined a clear influence of the database on the positivity of the correlation. This pattern was shown both by SILVA version 138 and Greengenes version 13.8. MICCA reported a high correlation between different approaches when led by the same database (Figure 33).



Figure 33: Heatmap of the correlation values of each pipeline + database combination outputs against each other in the fecal dataset. Color range from black (negative correlation, -1) to white (positive correlation, +1). Numbers inside the tiles are the p values of the correlation. Tiles containing a p value lower than 0.05 were contoured in red.

The p-values were fairly distributed across the combinations. Still, it should be considered that they both represent the significance of low and high correlations, therefore being spread in the comparisons between pipelines and databases.

4.3.6: Cumulative analysis of the benchmarking results

To gain further insight into the comparisons of the pipelines and the databases and their pairing and benchmarking, the results of the Spearman correlation from the milk, skin, rumen, and stool datasets were grouped, excluding the information regarding the dataset of origin. This approach aimed to homogenize the results, masking the influence of the dataset on the outputs.

Pipelines alone were examined first (Figure 34). Although all the pipelines reported at least one match with a correlation above 0.9, most matched preferably with themselves. The exception is QIIME version 1.9, which matches with MICCA in the closed reference approach, other than with itself.

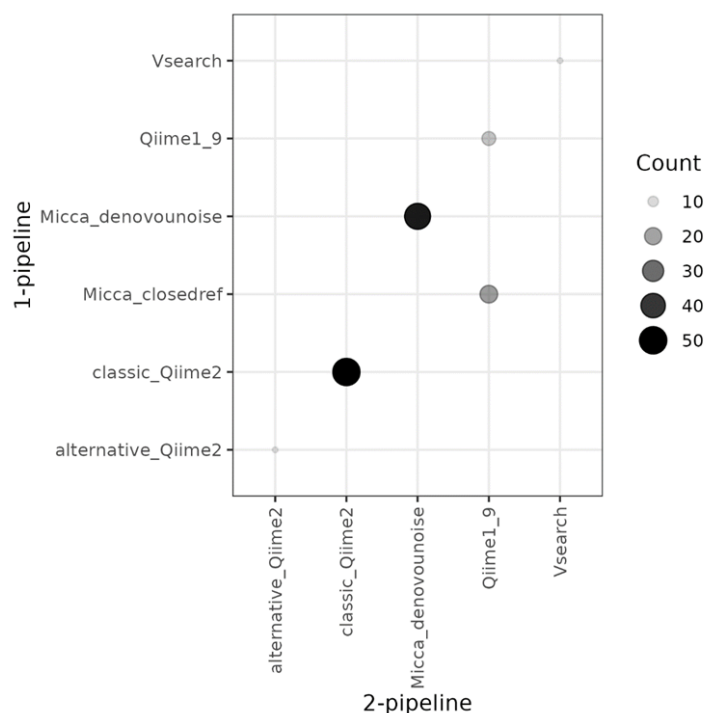


Figure 34: Bubble plot of the counts of correlation with a percentage higher than 90, only considering pipelines.

Similarly, databases alone were considered when reporting a correlation higher than 0.9 in the combinations (Figure 35). In this case, the results of the databases were correlated

with themselves only in the case of SILVA version 138, while Greengenes version 2 performed the best, matching with both 16SITGDB and Greengenes version 13.9. Nonetheless, it should be noted that SILVA version 138 and Greengenes version 2 correlated, too.

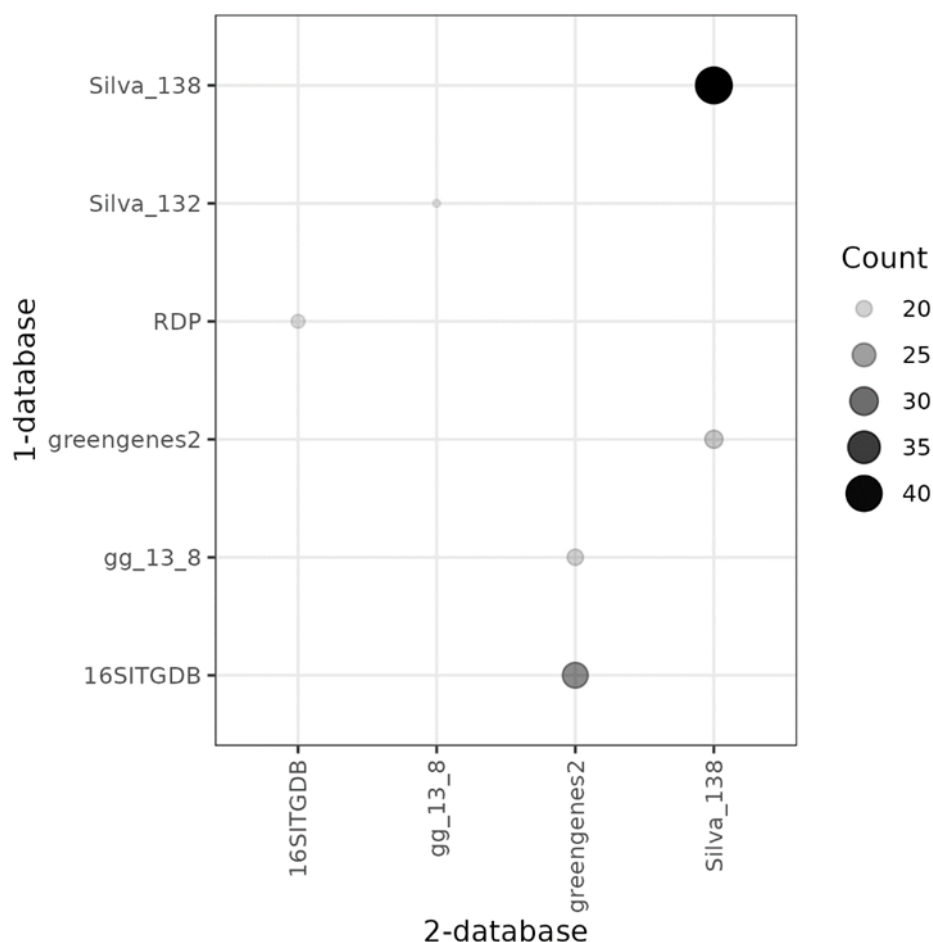


Figure 35: Bubble plot of the counts of correlation with percentages higher than 90, only considering databases.

In conclusion, pipelines and database pairings were considered, but only those reporting a correlation higher than 0.9 (Figure 36). First, it is interesting that SILVA's two highest score counts are SILVA version 138. The pairing that cumulated more correlations higher than 0.9 is QIIME version 2 in the classic approach paired with SILVA version 138 matched with the same pipeline paired with Greengenes version 2, and MICCA with the denovo approach paired with SILVA version 138 matched with QIIME version 2 in the classic approach paired with SILVA version 138. It stands out also that VSEARCH results only matched with themselves, even if not with the same databases. Overall, QIIME version 2 and MICCA reported the highest count of correlations higher than 0.9 when paired with SILVA version 138 and Greengenes version 2, followed at a distance by SILVA version 132.

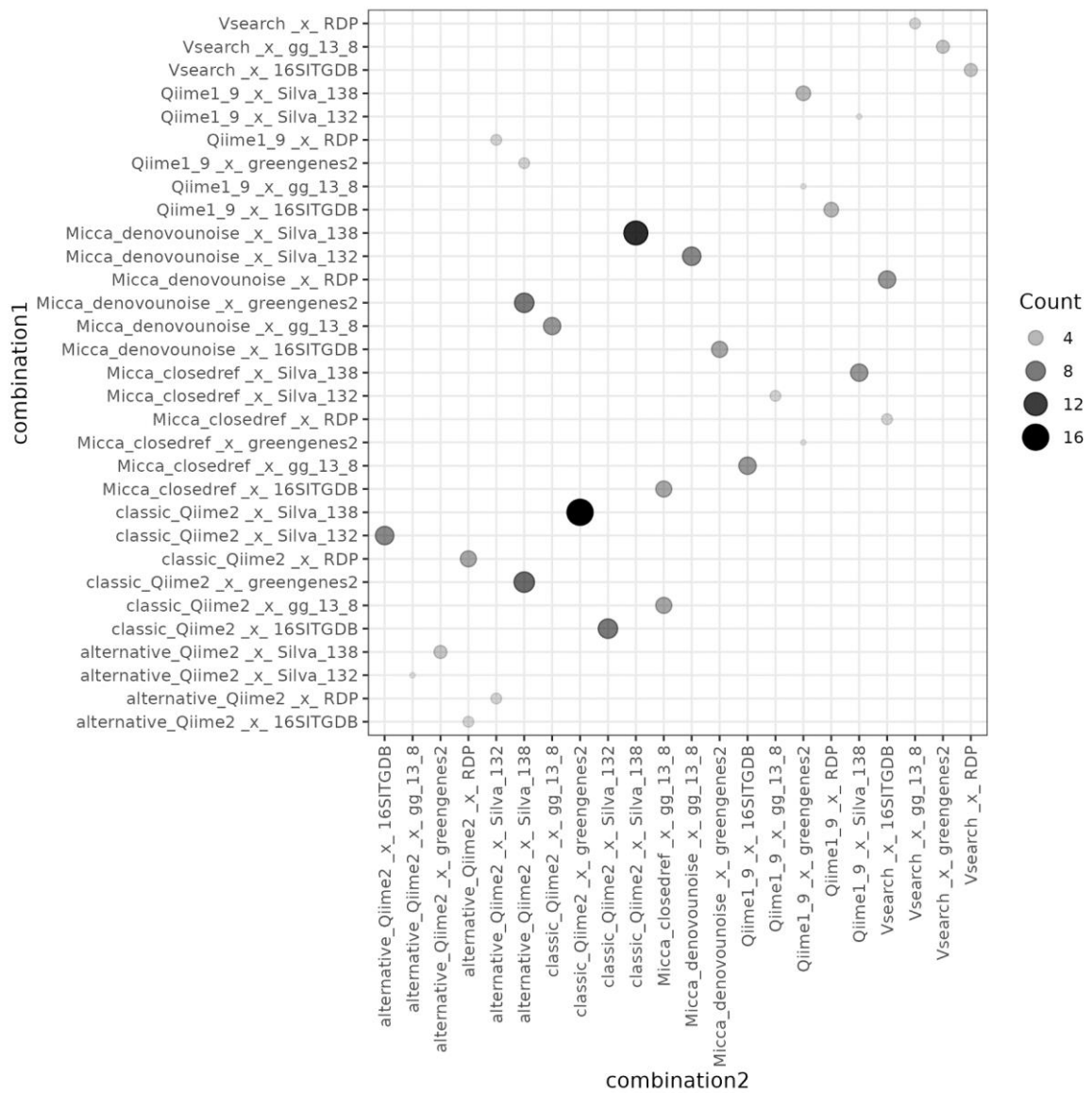


Figure 36: Bubble plot of the counts of correlation with a percentage higher than 90, considering pipelines and database pairings.

Chapter 5: Discussion

5.1: Introduction

Microbial communities are probably the major biological shapers of the environment as it is. Their ubiquity, along with the interactions they can build, are the fundamentals of ecosystem functioning. More, they can direct also other species evolution when they are not crucial for their survival. In this frame the activity of Bacteria, along with Archaea and Protists, is a wide field of knowledge just partially explored, while a better understanding of microbial activities and interactions, that are strictly correlated to many present issues such as methane emissions and antimicrobial resistance in dairy cows, would help improve interventions in these directions. When considering microbial communities, it is fundamental to focus on the plurality of information that microbiome can offer. In fact, as stated by Berg et al. [31], the microbiome encompasses all the microscopic organisms, the substances they produce (such as metabolites, proteins, and nucleic acids), and the surrounding environmental conditions that exist within a particular host or ecosystem. This means that at least one branch should be dedicated to each of these aspects, generating then, among others, metagenomics, metataxonomics, metatranscriptomics, metaproteomics, and metabonomics.

The use of focusing on metagenomics as was done in this thesis is linked to the urge of changes in the dairy cows' industry that can actively mitigate the effects of the increasing production pressure. Sequencing is an entrusted method to study microbial communities: in fact, due to the relatively low cost of its procedures and its portability and applications into a variety of scientific disciplines has seen an exponential growth. 16S rRNA gene sequencing is certainly one of the most popular options to study microbial communities, with bearable costs and discrete precision. Its affordability is primarily due to the fast evolution of the techniques, which nowadays are settled among next-generation sequencing and third-generation sequencing. However, this last is still quite expensive in terms of reagents and service, as well as whole genome sequencing. Parallely, the amount of data these platforms produce tends to be huge and need specialized software and computational facilities to handle. Roles as bioinformaticians and computational biologists has emerged and become of primary importance at this point, representing the joining link between biology and informatics to be able to extrapolate results with biological meaning from samples and develop appropriate and efficient software.

Luckily enough, this popularity has resulted in a consequent broadening of the availability of tools dedicated to sequencing outputs analysis on the market as well. Bioinformatics in fact develops and uses pipelines of commands able to approach sequencing data and delve into their details, also considering human- or machine-derived errors. To serve this purpose, databases are flanked to the pipelines and in continuous expansion and refinement.

Therefore, among these efforts in software development, this thesis focuses on benchmarking some of the most popular approaches for 16S rRNA gene sequence analysis pipelines, taking into consideration both OTU clustering and ASV denoising. As is well known, metagenomics results can be highly influenced by the choice of software, approach, and database, resulting in a biased taxonomic classification. The pipelines selected in this thesis included QIIME VERSION 1.9, QIIME VERSION 2, MICCA, and VSEARCH, where QIIME VERSION 2 and MICCA allowed both clustering and denoising approaches, while VSEARCH and QIIME VERSION 1.9 were strictly offering a clustering approach. Each pipeline was paired with each one of the selected databases: SILVA version 132 and SILVA version 138, RDP, Greengenes version 13.9 and Greengenes version 2, and 16SITGDB. Docker containers guaranteed a fresh start, as software could not be biased during downloading or environment building. Instead, they came already tested as available on the online repository, and Singularity supported their functioning efficiently. Lastly, a total of 36 pipeline and database pairings were tested. A double choice was made to reach the goal of robust benchmarking: two mock communities were used as sparring partners to test how fit the outcomes were with the accurate results. Secondly, several dairy cows-related datasets were included: in detail, milk, skin, rumen, and feces. Many studies on this species in the veterinary field suggested the choice of dairy cows' species. Indeed, because of their worldwide distribution and exploitation, dairy cows use results in a great economic impact as well as being one of the key players for animal health in the One Health perspective, therefore requiring constant monitoring.

5.2: First section: testing of 1 pipeline and 1 database on 1 dataset and comparing the results with the same database in a newest version and other 3 databases (of which 1 in 2 different versions)

The first section of this thesis was dedicated to the investigation of 1 dairy cows' related dataset, milk dataset. Derived from MILKQUA experiment, the milk dataset allowed a first approach to a well-known pipelines of amplicon sequencing analysis, QIIME version 1.9, and the above-mentioned databases.

Although being surpassed, at least technologically talking, by QIIME version 2 [112], QIIME version 1.9 was consistently found in literature oriented to the comparison of many pipelines of analysis, like Straub et al., Prodan et al., and Marizzoni et al. [105], [113], [114]. Pioneering tool in its field, QIIME version 1.9 offers a clustering algorithm based on UPARSE [53].

The databases paired to QIIME version 1.9 were compared by contrasting their alpha diversity level results. This method allowed us to visualize statistically and graphically the impact of the database on the results. For example, alpha diversity showed that the inference of databases made the T21 (compared with a T0 baseline) alpha indices in Treatment effect variables across the databases. Similarly, RDP led to a significant increase in control group alpha indices in time effect, which is absent from other databases. Moreover, the beta diversity presented slightly different MDS plots and PERMANOVA outcomes: even though results did not vary drastically, some databases showed significant p-values, while the same terms did not.

This analysis was a unicum until Ceccarani and Severgnini preprint deposited in BioRxiv [115], and agrees with it on the outcomes, although Ceccarani and Severgnini included NCBI database and excluded Greengenes version 2 and SILVA version 138: while alpha and beta diversity values showed variation, sometimes even statistically relevant, the databases consistently recovered the original experimental group differences.

The choice of including SILVA, Greengenes and RDP seems to be quite common in the literature, although every author seem then to include other databases: GTDB [116], ConTax [116] and the above-mentioned NCBI [115].

5.3: Second section: testing of other 3 pipelines (of which 2 in 2 different approaches - clustering vs denoising) and 3 databases (of which 2 in 2 different versions) on 1 dataset

The coherence of choice between this thesis and literature is kept in the pipelines included in this study, beyond databases choice. Studies including QIIME version 2 [105], [113], [114], [117] and USEARCH-UPARSE [105], [114], [117] are in fact quite common, whereas MICCA seems to be neglected similarly to Bioconductor [114], MEGAN [113], MicrobiomeAnalyst [118], MG-RAST [119], and VSEARCH [120]. On the other hand, other popular pipelines like mothur and DADA2 were here excluded to favour an old-versus-new comparison with QIIME version 1.9, whose popularity is decreasing. The study excluded MOTHUR due to technical difficulties encountered during pipeline building, as the software structure seemed not to allow the removal of steps that were not in common with other pipelines. DADA2 would also represent a valid alternative: this approach operates as an R package, and being R already popular as statistical software, researchers are increasingly approaching this method. Moreover, a broad community and clear guidelines are provided on GitHub: the different approaches required made it out of range for the benchmarking presented here. Time reasons also influenced these decisions. A further point should be made about the selection of the pipelines: in fact, all the pipelines considered are preferably run on servers. Computers with adequate computational power would sustain them as well, but on average, personal computers are machines equipped with up to 16 threads, and the computational time would be excessive and put at risk the machine itself, making it useless for other tasks. Additionally, those pipelines are functional only from the command line (QIIME version 2 alone owns a graphical interface), posing a further challenge for researchers not already in bioinformatics.

Nonetheless, literature results meagre regarding pipelines manuals of the here cited pipelines. This study required in fact to seek instructions in literature, developers' guidelines, and software-related forums: a point must be made about the absence of comparable outlines about software usage. Each developers' team offers guidelines and manuals with a non-standardized approach, making it difficult to find information when benchmarking multiple software simultaneously. QIIME version 1.9, for example, provides a website (<http://QIIME.org/install/index.html>) that has been discontinued in favor of QIIME version 2 (therefore, it is harder to find it on the internet), where single commands with related specific features can be found. However, the website lacks complete analysis tutorials, which can be found in the literature. QIIME version 2 instead does not offer a

straightforward manual for each command but instead has a website (<https://docs.QIIME2.org/YYYY.V/>, with YYYY representing the release year and V the release version) with documents suggesting which commands to use depending on the dataset to analyze, considering the sequencing method (single or paired-end sequencing, platform used). On the bright side, the developers and users' community maintain an active and responsive community on the QIIME version 2 forum (<https://forum.QIIME2.org/>). However, it is still up to the researcher to figure out which commands fit the most with the requirements of the study taken into consideration. It should be noted that, however, the QIIME version 2 community strongly enhances the use of the denoising approach. In contrast, the clustering approach, equally made available by the developers on GitHub, tends to receive less attention and is much less advertised. Compared with the previously mentioned two versions of QIIME, VSEARCH lacks clear guidelines for using the software. The developers made available a detailed command manual and two examples of pipeline usage on GitHub (<https://github.com/torognes/vsearch>), but no further explanation or website is available. The user community is much less populated than QIIME one, even though the number of papers using this software is not scarce. In the author's opinion, it is possible that authors slavishly follow the commands mentioned above offered by the developers or the best practices of their lab without further discussing the usage of the software. Finally, MICCA software was taken into consideration. MICCA is probably the least popular among the software mentioned above, but in the author's opinion, it deserves better attention in the scientific community. MICCA owns a well-organized website (<https://micca.readthedocs.io/en/latest/>) that is similar in structure to the one proposed by QIIME version 1.9 developers. On the website, researchers can find both tutorials detailed step-by-step and single commands explained in their aims and detailed with all the features available, supplied with examples. MICCA was the most user-friendly among all the software, requiring little programming experience and simple syntax. QIIME version 2 is strongly advantaged by the software's active feedback when a command has a negative outcome. VSEARCH, instead, was probably the most demanding in terms of coding skills and time spent problem-solving, as it does not give feedback but proceeds with empty files until the end of the pipeline. Three databases of which two in a double release were included: it should be noted that QIIME version 2 required a unique format for the database to be used, and as for some it was available online (SILVA version 138, 16SITGDB, Greengenes2), some others needed to be custom-made (SILVA version 132, RDP, Greengenes 13.9). Equally, SILVA version 138 had to be returned to the non-QIIME version 2 format to be used in other databases. The

aim of using six databases was to compare their performances related to the novelty of the version. The continuous updating of knowledge regarding the microbial community is responsible for an adequate outcome of the sequencing-related analysis.

Among the outcomes, denoising methods presented an exponentially higher number of unknowns from the milk dataset than clustering methods. However, when filtering for unknowns, results return a similar behaviour to the results of the clustering methods. This sensitivity towards unknowns can undoubtedly be attributed to the definition of denoising methods, causing a non-tolerance towards mismatching. Still, it could also be affected by the Illumina sequencing techniques. Nakamura et al. previously reported a weakness of Illumina platforms toward position-dependent errors affecting sequence patterns [121]. Nonetheless, this plasticity has shown a higher correlation of the results among those from other pipelines, even those clustering related.

5.4: Third section: testing of other 3 pipelines (of which 2 in 2 different approaches - clustering vs denoising) and 3 databases (of which 2 in different versions) on 5 datasets

Differently from other studies, this thesis aimed to test the pipeline + database dyad on dairy cows' datasets, as well as on mock communities. Whereas mock communities are quite popular in benchmarking studies, although their variety is wide, it seems to be innovative to focus on multiple datasets in a single study. Previous articles in fact preferred human datasets [105], [114], [117] or environment samples [113], and just in one case, chicken cecum [117].

The choice of these datasets was guided by the need to test the results produced by the 36 combinations: mock communities were used to assess the accuracy each prediction could make against the known composition of the community. On the other hand, a real-life dataset was used to evaluate the reproducibility of the results and, therefore, tested for correlation using the Spearman correlation test. It should be noted that the dataset considered for this study was paired-end sequenced operating Illumina platforms: future studies should also look at improving single-end sequencing, as mentioned by Maki et al. [122].

Mock communities from Prodan et al. [105] and Callahan et al. [123] were enrolled and showed similar behaviour in the results: although those produced by Prodan et al. resulted in more hits than the ones from Callahan et al., it was observed that in both scenarios, QIIME version 2, both in the clustering and denoising approach, similar patterns were presented paired with high percentages of unknowns regardless of the dataset used. Similarly, greengenes2 was the database that, with the most difficulties, identified the correct species and genera. On the other hand, QIIME version 1.9 and MICCA, with the clustering approach, presented good accuracy compared to other pipelines. This statement leads to the observation that the clustering methods seem to outperform the denoising ones, which can also be observed in the unfiltered results of the milk dataset. However, the values obtained considering the number of OTU and ASV identified show that the number of OTU is much higher than ASV, therefore raising the probability of founding the correct match with the expected taxa. Yet, the probability that this is a random event rather than a correct identification is higher. Moreover, rarefaction curves showed a better performance of denoising approaches than clustering approaches in both mock communities outcomes, reaching plateaux more efficiently.

Regarding mock communities, a further comment is due: the choice of reporting in this work both genus level and species level results for the mock communities, although it is known that deepening the taxonomic level analysis impoverish the results, was linked to the fact that original studies presented the mock communities at species level. Being aware of the technique limitations, results still can be defined as satisfactory.

The four additional datasets included in the study were used to investigate pipeline + database performance on dairy cows' dataset as MILKQUA milk and skin dataset, rumen dataset from Ranilla et al. [108], and stool dataset from Andres et al. [124]. This expansion of the study aimed to reproduce the same results obtained on the milk dataset with different starting datasets, to merge them later and assess the performances of single pipeline, single database, and single pipeline and database pair against every other, unbiased by dataset influences. The results reported were selected for positive correlations higher than 90%, as the background idea of the benchmarking would be to propose a pipeline and database pairing to obtain the most inclusive and performing results possible.

Across the different datasets, trends were noticed when the same pipeline was used in a double approach, whatever the database, as one would expect. About the pipelines, the most outstanding results were QIIME version 2 in both approaches, for which results quality seemed to be dependent on the database, flanked by MICCA with the same caveat. On the other hand, VSEARCH was the carrier of negative correlations on multiple occasions. Regarding databases instead, SILVA version 138 and Greengenes version 2 performed the best, as one would expect, considering that an earlier version would include the previous information in the same database. Disappointingly, however, in the frame of a one-by-one dataset, 16S-ITGDB did not excel as expected by a database that has been recently developed and aimed to include all the most popular previously published as mentioned above.

Nonetheless, for completeness, the correlation results were considered cumulatively, excluding the dataset of origin. As mentioned, the goal here was to benchmark the pipelines, the databases, and the pairings to unravel a dataset-independent observation. Pipelines showed an obvious matching within themselves, where MICCA and QIIME version 2 had the higher counts: the most outstanding result was the matching between MICCA with the closed reference approach and QIIME version 1.9, showing that the algorithm is not dissimilar between the two when the approach stays the same. In the author's opinion, these results empower MICCA, confirming the portability of the software already described in the commands section.

Database outcomes were, instead, probably heavily influenced by the algorithm effects, as the counts tend to be very low, except for SILVA version 138. This last database and Greengenes version 2 retrieved the highest count of positive correlations, probably due to its updates. The pipelines and database pairings merely reflected the above-mentioned results, partially redeeming VSEARCH, showing a better ductility linked to databases, and giving more space to SILVA version 132.

5.5: Limitations of the presented work

This study in its completeness carries some technical limitations. First, the attention was focused on bacteria: as explained in the introduction of this work, microbial communities are instead known to be composed by archaea and protists as well, whereas some authors consider viruses as well. This choice is justified by reasons of cleanliness of the analysis: sequencing of the dairy cows' datasets was performed using V3-V4 primers, which can as well locate and identify archaea and protists but are not the gold standard for them and could produce inaccurate results. Therefore, filters were applied to remove taxa identified as belonging to these two kingdoms. Secondly, the datasets included in the benchmarking were restricted to dairy cows related dataset: a more comprehensive study in the veterinary field could include also other species, such as pigs, chicken and fishes, and would probably be a first, as to the best of the author knowledge literature similar to this study is limited to mock communities, human and environmental datasets. Thirdly, a higher number of pipelines and databases could have been included: as cited above, Bioconductor [114], MEGAN [113], MicrobiomeAnalyst [118], MG-RAST [119], mothur [125] and DADA2 [123] could have been included as pipelines, while GTDB [116], ConTax [116], NCBI [115] and GSR-DB [126] could have been included as databases, just to mention some. Time limits restricted this study to the ones presented above. Lastly, at comparison with other similar studies, computational timing was not reported. This information was in fact discarded as the analysis were run on servers equipped with scheduling policies, which could have returned a running time dependent on the amount of work in that moment on the server. Therefore, the information would have resulted biased and not useful to a comparison.

Chapter 6: Final remarks and conclusions

Nowadays, with the enhancement at a high rate of research and its products and by-products, bioinformatics primarily enhances the comprehension of several fields, such as dairy cows' health and production.

From that, this thesis aims to benefit the scientific community by benchmarking some of the most popular software dedicated to 16S rRNA gene sequencing analysis paired with up-to-date databases and tested on different datasets. A clear outline of the commands to build essential yet reliable pipelines was defined, along with testing their performance and dependability. Furthermore, examples of the effectiveness of the methods were provided via tests on mock communities and real-life databases related to dairy cows, such as cow milk, skin, ruminal communities, and stool.

The results point to encouraging denoising approaches as more reliable, with MICCA and QIIME version 2 presenting the higher scoring in terms of matching results. Among databases, SILVA version 138 is currently considered the gold standard, with the tip of keeping up with SILVA updates. Therefore, the present study would suggest using QIIME version 2 and MICCA in the denoising approach with the most updated databases, as denoising approach constitutes the future of research. In fact, databases are currently not able to identify ASV that have not yet been cultured or in any way identified yet but will improve in the next decade: hence, the 2 above-mentioned software will guarantee clear, straightforward pipelines to deliver great quality results.

In conclusion, as the reliability of the approach depends on the availability of trustworthy sources, this study align with the previous ones in the invocation for harmonization of guidelines, software and parameters regarding sequencing data analysis [114], [127].

References

- [1] C. Wagg, K. Schlaeppi, S. Banerjee, E. E. Kuramae, and M. G. A. van der Heijden, “Fungal-bacterial diversity and microbiome complexity predict ecosystem functioning,” *Nat Commun*, vol. 10, no. 1, p. 4841, Oct. 2019, doi: 10.1038/s41467-019-12798-y.
- [2] A. Mougi, “The roles of amensalistic and commensalistic interactions in large ecological network stability,” *Sci Rep*, vol. 6, no. 1, p. 29929, Jul. 2016, doi: 10.1038/srep29929.
- [3] N. J. Holland and D. L. DeAngelis, “Consumer-resource theory predicts dynamic transitions between outcomes of interspecific interactions,” *Ecol Lett*, vol. 12, no. 12, pp. 1357-1366, Dec. 2009, doi: 10.1111/j.1461-0248.2009.01390.x.
- [4] V. V. Doytchinov and S. G. Dimov, “Microbial Community Composition of the Antarctic Ecosystems: Review of the Bacteria, Fungi, and Archaea Identified through an NGS-Based Metagenomics Approach,” *Life*, vol. 12, no. 6, p. 916, Jun. 2022, doi: 10.3390/life12060916.
- [5] S. Zhou *et al.*, “Unique bacterial communities and lifestyles in deep ocean blue holes: Insights from the Yongle Blue Hole (South China Sea),” *Front Mar Sci*, vol. 10, Jan. 2023, doi: 10.3389/fmars.2023.1086117.
- [6] N. B. Dragone, L. B. Perry, A. J. Solon, A. Seimon, T. A. Seimon, and S. K. Schmidt, “Genetic analysis of the frozen microbiome at 7900 m a.s.l., on the South Col of Sagarmatha (Mount Everest),” *Arct Antarct Alp Res*, vol. 55, no. 1, Dec. 2023, doi: 10.1080/15230430.2023.2164999.
- [7] C. J. van der Gast, “Microbial biogeography: the end of the ubiquitous dispersal hypothesis?,” *Environ Microbiol*, vol. 17, no. 3, pp. 544-546, Mar. 2015, doi: 10.1111/1462-2920.12635.
- [8] C. S. Cockell, “Are microorganisms everywhere they can be?,” *Environ Microbiol*, vol. 23, no. 11, pp. 6355-6363, Nov. 2021, doi: 10.1111/1462-2920.15825.
- [9] K. J. Locey and J. T. Lennon, “Scaling laws predict global microbial diversity,” *Proceedings of the National Academy of Sciences*, vol. 113, no. 21, pp. 5970-5975, May 2016, doi: 10.1073/pnas.1521291113.

- [10] W. B. Whitman, D. C. Coleman, and W. J. Wiebe, "Prokaryotes: The unseen majority," *Proceedings of the National Academy of Sciences*, vol. 95, no. 12, pp. 6578-6583, Jun. 1998, doi: 10.1073/pnas.95.12.6578.
- [11] R. Rappuoli, P. Young, E. Ron, S. Pecetta, and M. Pizza, "Save the microbes to save the planet. A call to action of the International Union of the Microbiological Societies (IUMS)," *One Health Outlook*, vol. 5, no. 1, p. 5, Mar. 2023, doi: 10.1186/s42522-023-00077-2.
- [12] M. Kowarsky *et al.*, "Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA," *Proceedings of the National Academy of Sciences*, vol. 114, no. 36, pp. 9623-9628, Sep. 2017, doi: 10.1073/pnas.1707009114.
- [13] L. Ma, H. Zhao, L. B. Wu, Z. Cheng, and C. Liu, "Impact of the microbiome on human, animal, and environmental health from a One Health perspective," *Science in One Health*, vol. 2, p. 100037, 2023, doi: 10.1016/j.soh.2023.100037.
- [14] P. Trinh, J. R. Zaneveld, S. Safranek, and P. M. Rabinowitz, "One Health Relationships Between Human, Animal, and Environmental Microbiomes: A Mini-Review," *Front Public Health*, vol. 6, Aug. 2018, doi: 10.3389/fpubh.2018.00235.
- [15] P. Comizzoli, M. L. Power, S. L. Bornbusch, and C. R. Muletz-Wolz, "Interactions between reproductive biology and microbiomes in wild animal species," *Anim Microbiome*, vol. 3, no. 1, p. 87, Dec. 2021, doi: 10.1186/s42523-021-00156-7.
- [16] T. Pape, F. Hoffmann, N.-V. Quéric, K. von Juterzenka, J. Reitner, and W. Michaelis, "Dense populations of Archaea associated with the demosponge *Tentorium semisuberites* Schmidt, 1870 from Arctic deep-waters," *Polar Biol*, vol. 29, no. 8, pp. 662-667, Jul. 2006, doi: 10.1007/s00300-005-0103-4.
- [17] H. Margot, C. Acebal, E. Toril, Amils R., and J. Fernandez Puentes, "Consistent association of crenarchaeal Archaea with sponges of the genus *Axinella*," *Mar Biol*, vol. 140, no. 4, pp. 739-745, Apr. 2002, doi: 10.1007/s00227-001-0740-2.
- [18] E. Vortsepneva *et al.*, "Microbial associations of shallow-water Mediterranean marine cave *Solenogastres* (Mollusca)," *PeerJ*, vol. 9, p. e12655, Dec. 2021, doi: 10.7717/peerj.12655.
- [19] H. L. Eisthen and K. R. Theis, "Animal-microbe interactions and the evolution of nervous systems," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 371, no. 1685, p. 20150052, Jan. 2016, doi: 10.1098/rstb.2015.0052.

- [20] B. J. Finlay, G. Esteban, K. J. Clarke, A. G. Williams, T. M. Embley, and R. P. Hirt, "Some rumen ciliates have endosymbiotic methanogens," *FEMS Microbiol Lett*, vol. 117, no. 2, pp. 157-161, Apr. 1994, doi: 10.1111/j.1574-6968.1994.tb06758.x.
- [21] M. Herrero *et al.*, "Biomass use, production, feed efficiencies, and greenhouse gas emissions from global livestock systems," *Proceedings of the National Academy of Sciences*, vol. 110, no. 52, pp. 20888-20893, Dec. 2013, doi: 10.1073/pnas.1308149110.
- [22] J. H. P. Hackstein and T. A. van Alen, "Methanogens in the Gastrointestinal Tract of Animals," 2018, pp. 121-152. doi: 10.1007/978-3-319-98836-8_8.
- [23] M. McFall-Ngai *et al.*, "Animals in a bacterial world, a new imperative for the life sciences," *Proceedings of the National Academy of Sciences*, vol. 110, no. 9, pp. 3229-3236, Feb. 2013, doi: 10.1073/pnas.1218525110.
- [24] T. Pradeu, "A Mixed Self: The Role of Symbiosis in Development," *Biol Theory*, vol. 6, no. 1, pp. 80-88, Dec. 2011, doi: 10.1007/s13752-011-0011-5.
- [25] G. Borrel, J.-F. Brugère, S. Gribaldo, R. A. Schmitz, and C. Moissl-Eichinger, "The host-associated archaeome," *Nat Rev Microbiol*, vol. 18, no. 11, pp. 622-636, Nov. 2020, doi: 10.1038/s41579-020-0407-y.
- [26] C. Moissl-Eichinger, M. Pausan, J. Taffner, G. Berg, C. Bang, and R. A. Schmitz, "Archaea Are Interactive Components of Complex Microbiomes," *Trends Microbiol*, vol. 26, no. 1, pp. 70-85, Jan. 2018, doi: 10.1016/j.tim.2017.07.004.
- [27] S. Neupane, C. Saski, and D. Nayduch, "House fly larval grazing alters dairy cattle manure microbial communities," *BMC Microbiol*, vol. 21, no. 1, p. 346, Dec. 2021, doi: 10.1186/s12866-021-02418-5.
- [28] B.-A. T. Nguyen, Q.-L. Chen, J.-Z. He, and H.-W. Hu, "Livestock manure spiked with the antibiotic tylosin significantly altered soil protist functional groups," *J Hazard Mater*, vol. 427, p. 127867, Apr. 2022, doi: 10.1016/j.jhazmat.2021.127867.
- [29] S. A. Wooldridge, "Is the coral-algae symbiosis really 'mutually beneficial' for the partners?," *BioEssays*, vol. 32, no. 7, pp. 615-625, Jul. 2010, doi: 10.1002/bies.200900182.

- [30] A. L. Lind and K. S. Pollard, "Accurate and sensitive detection of microbial eukaryotes from whole metagenome shotgun sequencing," *Microbiome*, vol. 9, no. 1, p. 58, Dec. 2021, doi: 10.1186/s40168-021-01015-y.
- [31] G. Berg *et al.*, "Microbiome definition re-visited: old concepts and new challenges," *Microbiome*, vol. 8, no. 1, p. 103, 2020, doi: 10.1186/s40168-020-00875-0.
- [32] S. P. Yadav, "The wholeness in suffix -omics, -omes, and the word om.," *J Biomol Tech*, vol. 18, no. 5, p. 277, Dec. 2007.
- [33] C. Bedia, "Experimental Approaches in Omic Sciences," 2018, pp. 13-36. doi: 10.1016/bs.coac.2018.07.002.
- [34] J. R. Marchesi and J. Ravel, "The vocabulary of microbiome research: a proposal," *Microbiome*, vol. 3, no. 1, p. 31, 2015, doi: 10.1186/s40168-015-0094-5.
- [35] S. Bellali *et al.*, "Running after ghosts: are dead bacteria the dark matter of the human gut microbiota?," *Gut Microbes*, vol. 13, no. 1, pp. 1-12, 2021, doi: 10.1080/19490976.2021.1897208.
- [36] J.-C. Lagier *et al.*, "Culture of previously uncultured members of the human gut microbiota by culturomics," *Nat Microbiol*, vol. 1, no. 12, p. 16203, Nov. 2016, doi: 10.1038/nmicrobiol.2016.203.
- [37] J. M. Churko, G. L. Mantalas, M. P. Snyder, and J. C. Wu, "Overview of High Throughput Sequencing Technologies to Elucidate Molecular Pathways in Cardiovascular Diseases," *Circ Res*, vol. 112, no. 12, pp. 1613-1623, Jun. 2013, doi: 10.1161/CIRCRESAHA.113.300939.
- [38] Y. X. Liu *et al.*, "A practical guide to amplicon and metagenomic analysis of microbiome data," *Protein and Cell*, vol. 12, no. 5. Higher Education Press Limited Company, pp. 315-330, May 01, 2021. doi: 10.1007/s13238-020-00724-8.
- [39] P. Rausch *et al.*, "Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms," *Microbiome*, vol. 7, no. 1, p. 133, 2019, doi: 10.1186/s40168-019-0743-1.
- [40] B. E. Deagle, S. N. Jarman, E. Coissac, F. Pompanon, and P. Taberlet, "DNA metabarcoding and the cytochrome C oxidase subunit I marker: not a perfect match," *Biol Lett*, vol. 10, no. 9, p. 20140562, Sep. 2014, doi: 10.1098/rsbl.2014.0562.

- [41] A.-S. Isabel *et al.*, “Primer, Pipelines, Parameters: Issues in 16S rRNA Gene Sequencing,” *mSphere*, vol. 6, no. 1, pp. 10.1128/msphere.01202-20, Feb. 2021, doi: 10.1128/msphere.01202-20.
- [42] C. E. Ritland, K. Ritland, and N. A. Straus, “Variation in the ribosomal internal transcribed spacers (ITS1 and ITS2) among eight taxa of the *Mimulus guttatus* species complex,” *Mol Biol Evol*, vol. 10, no. 6, pp. 1273-1288, Nov. 1993, doi: 10.1093/oxfordjournals.molbev.a040067.
- [43] J. Ronholm, N. Neda, P. Nicholas, and P. Franco, “Navigating Microbiological Food Safety in the Era of Whole-Genome Sequencing,” *Clin Microbiol Rev*, vol. 29, no. 4, pp. 837-857, Aug. 2016, doi: 10.1128/cmr.00056-16.
- [44] K. Chen and L. Pachter, “Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities,” *PLoS Comput Biol*, vol. 1, no. 2, pp. e24-, Jul. 2005, [Online]. Available: <https://doi.org/10.1371/journal.pcbi.0010024>
- [45] A. Antich, C. Palacin, O. S. Wangenstein, and X. Turon, “To denoise or to cluster, that is not the question: optimizing pipelines for COI metabarcoding and metaphylogeography,” *BMC Bioinformatics*, vol. 22, no. 1, p. 177, Dec. 2021, doi: 10.1186/s12859-021-04115-6.
- [46] S. L. Westcott and P. D. Schloss, “De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units,” *PeerJ*, vol. 3, p. e1487, Dec. 2015, doi: 10.7717/peerj.1487.
- [47] J. A. Navas-Molina *et al.*, “Advancing Our Understanding of the Human Microbiome Using QIIME,” 2013, pp. 371-444. doi: 10.1016/B978-0-12-407863-5.00019-8.
- [48] P. D. Schloss and S. L. Westcott, “Assessing and Improving Methods Used in Operational Taxonomic Unit-Based Approaches for 16S rRNA Gene Sequence Analysis,” *Appl Environ Microbiol*, vol. 77, no. 10, pp. 3219-3226, May 2011, doi: 10.1128/AEM.02810-10.
- [49] D. Forster, M. Dunthorn, T. Stoeck, and F. Mahé, “Comparison of three clustering approaches for detecting novel environmental microbial diversity,” *PeerJ*, vol. 4, p. e1692, Feb. 2016, doi: 10.7717/peerj.1692.
- [50] L. W. Hugerth and A. F. Andersson, “Analysing microbial community composition through amplicon sequencing: From sampling to hypothesis testing,” *Frontiers in Microbiology*, vol. 8, no. SEP. Frontiers Media S.A., Sep. 04, 2017. doi: 10.3389/fmicb.2017.01561.

- [51] P. D. Schloss and J. Handelsman, "Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness," *Appl Environ Microbiol*, vol. 71, no. 3, pp. 1501-1506, Mar. 2005, doi: 10.1128/AEM.71.3.1501-1506.2005.
- [52] B. D. Lindahl *et al.*, "Fungal community analysis by high-throughput sequencing of amplified markers - a user's guide," *New Phytologist*, vol. 199, no. 1, pp. 288-299, Jul. 2013, doi: 10.1111/nph.12243.
- [53] R. C. Edgar, "Search and clustering orders of magnitude faster than BLAST," *Bioinformatics*, vol. 26, no. 19, pp. 2460-2461, Oct. 2010, doi: 10.1093/bioinformatics/btq461.
- [54] S. I. Glassman and J. B. H. Martiny, "Broadscale Ecological Patterns Are Robust to Use of Exact Sequence Variants versus Operational Taxonomic Units," *mSphere*, vol. 3, no. 4, Aug. 2018, doi: 10.1128/mSphere.00148-18.
- [55] B. J. Callahan, P. J. McMurdie, and S. P. Holmes, "Exact sequence variants should replace operational taxonomic units in marker-gene data analysis," *ISME J*, vol. 11, no. 12, pp. 2639-2643, Dec. 2017, doi: 10.1038/ismej.2017.119.
- [56] R Core Team (2022), "R: A language and environment for statistical ## computing."
- [57] RStudio Team (2020), "RStudio: Integrated Development for R," <http://www.rstudio.com/>.
- [58] P. J. McMurdie and S. Holmes, "phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data," *PLoS One*, vol. 8, no. 4, pp. e61217-, Apr. 2013, [Online]. Available: <https://doi.org/10.1371/journal.pone.0061217>
- [59] H. Wickham, *Ggplot2: Elegant graphics for data analysis*. Springer International Publishing, 2016.
- [60] Food and Agriculture Organization of the United Nations, *World Food and Agriculture FAO - Statistical Yearbook*. Rome, Italy: FAO, 2023. Accessed: Jun. 26, 2024. [Online]. Available: <https://openknowledge.fao.org/handle/20.500.14283/cc8166en>
- [61] A. Koeck, S. Loker, F. Miglior, D. F. Kelton, J. Jamrozik, and F. S. Schenkel, "Genetic relationships of clinical mastitis, cystic ovaries, and lameness with milk

- yield and somatic cell score in first-lactation Canadian Holsteins,” *J Dairy Sci*, vol. 97, no. 9, pp. 5806-5813, Sep. 2014, doi: 10.3168/jds.2013-7785.
- [62] V. Bronzo *et al.*, “The Role of Innate Immune Response and Microbiome in Resilience of Dairy Cattle to Disease: The Mastitis Model,” *Animals*, vol. 10, no. 8, p. 1397, Aug. 2020, doi: 10.3390/ani10081397.
- [63] W. N. Cheng and S. G. Han, “Bovine mastitis: risk factors, therapeutic strategies, and alternative treatments – A review,” *Asian-Australas J Anim Sci*, vol. 33, no. 11, pp. 1699-1713, Nov. 2020, doi: 10.5713/ajas.20.0156.
- [64] M. F. Addis, A. Tanca, S. Uzzau, G. Oikonomou, R. C. Bicalho, and P. Moroni, “The bovine milk microbiota: insights and perspectives from -omics studies,” *Mol Biosyst*, vol. 12, no. 8, pp. 2359-2372, 2016, doi: 10.1039/C6MB00217J.
- [65] N. G. Khasapane, Z. T. H. Khumalo, S. Kwenda, S. J. Nkhebenyane, and O. Thekisoe, “Characterisation of Milk Microbiota from Subclinical Mastitis and Apparently Healthy Dairy Cattle in Free State Province, South Africa.,” *Vet Sci*, vol. 10, no. 10, Oct. 2023, doi: 10.3390/vetsci10100616.
- [66] H. Derakhshani *et al.*, “Invited review: Microbiota of the bovine udder: Contributing factors and potential implications for udder health and mastitis susceptibility,” *J Dairy Sci*, vol. 101, no. 12, pp. 10605-10625, Dec. 2018, doi: 10.3168/jds.2018-14860.
- [67] M. Zubair, Z. Li, R. Zhu, J. Wang, X. Liu, and X. Liu, “The Antibiotics Degradation and Its Mechanisms during the Livestock Manure Anaerobic Digestion,” *Molecules*, vol. 28, no. 10, p. 4090, May 2023, doi: 10.3390/molecules28104090.
- [68] I. Azagra-Boronat *et al.*, “*Lactobacillus fermentum* CECT5716 supplementation in rats during pregnancy and lactation affects mammary milk composition,” *J Dairy Sci*, vol. 103, no. 4, pp. 2982-2992, Apr. 2020, doi: 10.3168/jds.2019-17384.
- [69] T. Jost, C. Lacroix, C. P. Braegger, F. Rochat, and C. Chassard, “Vertical mother-neonate transfer of maternal gut bacteria via breastfeeding,” *Environ Microbiol*, vol. 16, no. 9, pp. 2891-2904, Sep. 2014, doi: 10.1111/1462-2920.12238.
- [70] J. M. Rodríguez, “The Origin of Human Milk Bacteria: Is There a Bacterial Entero-Mammary Pathway during Late Pregnancy and Lactation?,” *Advances in Nutrition*, vol. 5, no. 6, pp. 779-784, Nov. 2014, doi: 10.3945/an.114.007229.

- [71] J. E. Williams *et al.*, “Strong Multivariate Relations Exist Among Milk, Oral, and Fecal Microbiomes in Mother-Infant Dyads During the First Six Months Postpartum,” *J Nutr*, vol. 149, no. 6, pp. 902-914, Jun. 2019, doi: 10.1093/jn/nxy299.
- [72] S. A. Metzger, L. L. Hernandez, J. H. Skarlupka, T. M. Walker, G. Suen, and P. L. Ruegg, “A Cohort Study of the Milk Microbiota of Healthy and Inflamed Bovine Mammary Glands From Dryoff Through 150 Days in Milk,” *Front Vet Sci*, vol. 5, Oct. 2018, doi: 10.3389/fvets.2018.00247.
- [73] W. Guo *et al.*, “Bovine milk microbiota: Key players, origins, and potential contributions to early-life gut development,” *J Adv Res*, Jul. 2023, doi: 10.1016/j.jare.2023.06.016.
- [74] N. J. Evans *et al.*, “Association between bovine digital dermatitis treponemes and a range of ‘non-healing’ bovine hoof disorders,” *Veterinary Record*, vol. 168, no. 8, pp. 214-214, Feb. 2011, doi: 10.1136/vr.c5487.
- [75] M. W. Nielsen *et al.*, “Potential bacterial core species associated with digital dermatitis in cattle herds identified by molecular profiling of interdigital skin samples,” *Vet Microbiol*, vol. 186, pp. 139-149, Apr. 2016, doi: 10.1016/j.vetmic.2016.03.003.
- [76] J. Zhang *et al.*, “Comparison of Dynamics of Udder Skin Microbiota From Grazing Yak and Cattle During the Perinatal Period on the Qinghai-Tibetan Plateau,” *Front Vet Sci*, vol. 9, May 2022, doi: 10.3389/fvets.2022.864057.
- [77] S. F. Lima *et al.*, “The bovine colostrum microbiome and its association with clinical mastitis,” *J Dairy Sci*, vol. 100, no. 4, pp. 3031-3042, Apr. 2017, doi: 10.3168/jds.2016-11604.
- [78] T. A. Mcallister, L. M. Rode, D. J. Major, K.-J. Cheng, and J. G. Buchanan-Smith, “Effect of ruminal microbial colonization on cereal grain digestion,” *Can J Anim Sci*, vol. 70, no. 2, pp. 571-579, Jun. 1990, doi: 10.4141/cjas90-069.
- [79] K. Liu *et al.*, “Ruminal microbiota-host interaction and its effect on nutrient metabolism,” *Animal Nutrition*, vol. 7, no. 1, pp. 49-55, Mar. 2021, doi: 10.1016/j.aninu.2020.12.001.
- [80] A. Hardan, P. C. Garnsworthy, and M. J. Bell, “Detection of Methane Eructation Peaks in Dairy Cows at a Robotic Milking Station Using Signal Processing.,” *Animals (Basel)*, vol. 12, no. 1, Dec. 2021, doi: 10.3390/ani12010026.

- [81] Q. Xu *et al.*, “Gut Microbiota and Their Role in Health and Metabolic Disease of Dairy Cow,” *Front Nutr*, vol. 8, Aug. 2021, doi: 10.3389/fnut.2021.701511.
- [82] T. R. Callaway *et al.*, “Evaluation of bacterial diversity in the rumen and feces of cattle fed different levels of dried distillers grains plus solubles using bacterial tag-encoded FLX amplicon pyrosequencing1,” *J Anim Sci*, vol. 88, no. 12, pp. 3977-3983, Dec. 2010, doi: 10.2527/jas.2010-2900.
- [83] S. C. Fernando *et al.*, “Rumen Microbial Population Dynamics during Adaptation to a High-Grain Diet,” *Appl Environ Microbiol*, vol. 76, no. 22, pp. 7482-7490, Nov. 2010, doi: 10.1128/AEM.00388-10.
- [84] M. Schären *et al.*, “Interrelations between the rumen microbiota and production, behavioral, rumen fermentation, metabolic, and immunological attributes of dairy cows,” *J Dairy Sci*, vol. 101, no. 5, pp. 4615-4637, May 2018, doi: 10.3168/jds.2017-13736.
- [85] L. Zhao, X. Li, E. R. Atwill, S. S. Aly, D. R. Williams, and Z. Su, “Dynamic changes in fecal bacterial microbiota of dairy cattle across the production line,” *BMC Microbiol*, vol. 22, no. 1, p. 132, Dec. 2022, doi: 10.1186/s12866-022-02549-3.
- [86] C. B. Welch, V. E. Ryman, T. D. Pringle, and J. M. Lourenco, “Utilizing the Gastrointestinal Microbiota to Modulate Cattle Health through the Microbiome-Gut-Organ Axes,” *Microorganisms*, vol. 10, no. 7, p. 1391, Jul. 2022, doi: 10.3390/microorganisms10071391.
- [87] M. Kim and James. E. Wells, “A Meta-analysis of Bacterial Diversity in the Feces of Cattle,” *Curr Microbiol*, vol. 72, no. 2, pp. 145-151, Feb. 2016, doi: 10.1007/s00284-015-0931-6.
- [88] O. C. Shanks *et al.*, “Community structures of fecal bacteria in cattle from different animal feeding operations,” *Appl Environ Microbiol*, vol. 77, no. 9, pp. 2992-3001, May 2011, doi: 10.1128/AEM.02988-10.
- [89] J. R. Williamson, T. R. Callaway, J. M. Lourenco, and V. E. Ryman, “Characterization of rumen, fecal, and milk microbiota in lactating dairy cows,” *Front Microbiol*, vol. 13, Sep. 2022, doi: 10.3389/fmicb.2022.984119.
- [90] P. Cremonesi *et al.*, “Raw Milk Microbiota Modifications as Affected by Chlorine Usage for Cleaning Procedures: The Trentingrana PDO Case,” *Front Microbiol*, vol. 11, Oct. 2020, doi: 10.3389/fmicb.2020.564749.

- [91] Illumina, "16S Metagenomic Sequencing Library Preparation." Accessed: Dec. 14, 2023. [Online]. Available: https://support.illumina.com/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf
- [92] D. P. R. Herlemann, M. Labrenz, K. Jürgens, S. Bertilsson, J. J. Waniek, and A. F. Andersson, "Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea," *ISME J*, vol. 5, no. 10, pp. 1571-1579, 2011, doi: 10.1038/ismej.2011.41.
- [93] A. Klindworth *et al.*, "Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies," *Nucleic Acids Res*, vol. 41, no. 1, pp. e1-e1, Jan. 2013, doi: 10.1093/nar/gks808.
- [94] F. Biscarini *et al.*, "A Randomized Controlled Trial of Teat-Sealant and Antibiotic Dry-Cow Treatments for Mastitis Prevention Shows Similar Effect on the Healthy Milk Microbiome," *Front Vet Sci*, vol. 7, Sep. 2020, doi: 10.3389/fvets.2020.00581.
- [95] P. Ewels, M. Magnusson, S. Lundin, and M. Käller, "MultiQC: summarize analysis results for multiple tools and samples in a single report," *Bioinformatics*, vol. 32, no. 19, pp. 3047-3048, Oct. 2016, doi: 10.1093/bioinformatics/btw354.
- [96] S. Andrews, "FastQC: A Quality Control Tool for High Throughput Sequence Data." Accessed: Dec. 21, 2023. [Online]. Available: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- [97] J. G. Caporaso *et al.*, "Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample," *Proceedings of the National Academy of Sciences*, vol. 108, no. supplement_1, pp. 4516-4522, Mar. 2011, doi: 10.1073/pnas.1000080107.
- [98] C. Quast *et al.*, "The SILVA ribosomal RNA gene database project: improved data processing and web-based tools," *Nucleic Acids Res*, vol. 41, no. D1, pp. D590-D596, Nov. 2012, doi: 10.1093/nar/gks1219.
- [99] J. N. Paulson, O. C. Stine, H. C. Bravo, and M. Pop, "Differential abundance analysis for microbial marker-gene surveys," *Nat Methods*, vol. 10, no. 12, pp. 1200-1202, Dec. 2013, doi: 10.1038/nmeth.2658.
- [100] J. Kaufmann and A. Schering, "Analysis of Variance ANOVA," in *Wiley StatsRef: Statistics Reference Online*, Wiley, 2014. doi: 10.1002/9781118445112.stat06938.

- [101] J. Oksanen, F. G. Blanchet, and R. Kindt, “vegan: Community Ecology Package.” 2012.
- [102] Y.-P. Hsieh, Y.-M. Hung, M.-H. Tsai, L.-C. Lai, and E. Y. Chuang, “16S-ITGDB: An Integrated Database for Improving Species Classification of Prokaryotic 16S Ribosomal RNA Sequences,” *Frontiers in Bioinformatics*, vol. 2, Aug. 2022, doi: 10.3389/fbinf.2022.905489.
- [103] B. L. Maidak, G. J. Olsen, N. Larsen, R. Overbeek, M. J. McCaughey, and C. R. Woese, “The Ribosomal Database Project (RDP),” *Nucleic Acids Res*, vol. 24, no. 1, pp. 82-85, Jan. 1996, doi: 10.1093/nar/24.1.82.
- [104] T. Z. DeSantis *et al.*, “Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB,” *Appl Environ Microbiol*, vol. 72, no. 7, pp. 5069-5072, Jul. 2006, doi: 10.1128/AEM.03006-05.
- [105] A. Prodan, V. Tremaroli, H. Brolin, A. H. Zwinderman, M. Nieuwdorp, and E. Levin, “Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing,” *PLoS One*, vol. 15, no. 1, p. e0227434, Jan. 2020, doi: 10.1371/journal.pone.0227434.
- [106] B. J. Callahan, P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes, “DADA2: High-resolution sample inference from Illumina amplicon data,” *Nat Methods*, vol. 13, no. 7, pp. 581-583, Jul. 2016, doi: 10.1038/nmeth.3869.
- [107] N. A. Bokulich *et al.*, “mockrobiota: a Public Resource for Microbiome Bioinformatics Benchmarking,” *mSystems*, vol. 1, no. 5, pp. 10.1128/msystems.00062-16, Oct. 2016, doi: 10.1128/msystems.00062-16.
- [108] M. J. Ranilla *et al.*, “Effects of Thymbra capitata essential oil on in vitro fermentation end-products and ruminal bacterial communities,” *Sci Rep*, vol. 13, no. 1, p. 4153, 2023, doi: 10.1038/s41598-023-31370-9.
- [109] J. G. Caporaso *et al.*, “QIIME allows analysis of high-throughput community sequencing data,” *Nat Methods*, vol. 7, no. 5, pp. 335-336, May 2010, doi: 10.1038/nmeth.f.303.
- [110] N. Percie du Sert *et al.*, “Reporting animal research: Explanation and elaboration for the ARRIVE guidelines 2.0,” *PLoS Biol*, vol. 18, no. 7, pp. e3000411-, Jul. 2020, [Online]. Available: <https://doi.org/10.1371/journal.pbio.3000411>

- [111] A. Ueki, Y. Ohtaki, N. Kaku, and K. Ueki, "Descriptions of *Anaerotaenia torta* gen. nov., sp. nov. and *Anaerocolumna cellulositica* gen. nov., sp. nov. isolated from a methanogenic reactor of cattle waste and reclassification of *Clostridium aminovalericum*, *Clostridium jejuense* and *Clostridium xylanovorans* as *Anaerocolumna* species," *Int J Syst Evol Microbiol*, vol. 66, no. 8, pp. 2936-2943, Aug. 2016, doi: 10.1099/ijsem.0.001123.
- [112] E. Bolyen *et al.*, "Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2," *Nat Biotechnol*, vol. 37, no. 8, pp. 852-857, 2019, doi: 10.1038/s41587-019-0209-9.
- [113] D. Straub, N. Blackwell, A. Langarica-Fuentes, A. Peltzer, S. Nahnsen, and S. Kleindienst, "Interpretations of Environmental Microbial Community Studies Are Biased by the Selected 16S rRNA (Gene) Amplicon Sequencing Pipeline," *Front Microbiol*, vol. 11, Oct. 2020, doi: 10.3389/fmicb.2020.550420.
- [114] M. Marizzoni *et al.*, "Comparison of Bioinformatics Pipelines and Operating Systems for the Analyses of 16S rRNA Gene Amplicon Sequences in Human Fecal Samples," *Front Microbiol*, vol. 11, Jun. 2020, doi: 10.3389/fmicb.2020.01262.
- [115] C. Ceccarani and M. Severgnini, "A comparison between Greengenes, SILVA, RDP, and NCBI reference databases in four published microbiota datasets," *bioRxiv preprint*, 2023, doi: 10.1101/2023.04.12.535864.
- [116] M. P. Ramakodi, "Influence of 16S rRNA reference databases in amplicon-based environmental microbiome research," *Biotechnol Lett*, vol. 44, no. 3, pp. 523-533, Mar. 2022, doi: 10.1007/s10529-022-03233-2.
- [117] I. Allali *et al.*, "A comparison of sequencing platforms and bioinformatics pipelines for compositional analysis of the gut microbiome," *BMC Microbiol*, vol. 17, no. 1, p. 194, Dec. 2017, doi: 10.1186/s12866-017-1101-8.
- [118] A. Dhariwal, J. Chong, S. Habib, I. L. King, L. B. Agellon, and J. Xia, "MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data," *Nucleic Acids Res*, vol. 45, no. W1, pp. W180-W188, Jul. 2017, doi: 10.1093/nar/gkx295.
- [119] K. P. Keegan, E. M. Glass, and F. Meyer, "MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function," 2016, pp. 207-233. doi: 10.1007/978-1-4939-3369-3_13.

- [120] T. Rognes, T. Flouri, B. Nichols, C. Quince, and F. Mahé, “VSEARCH: a versatile open source tool for metagenomics,” *PeerJ*, vol. 4, p. e2584, 2016, doi: 10.7717/peerj.2584.
- [121] K. Nakamura *et al.*, “Sequence-specific error profile of Illumina sequencers,” *Nucleic Acids Res*, vol. 39, no. 13, p. e90, Jul. 2011, doi: 10.1093/nar/gkr344.
- [122] K. A. Maki, B. Wolff, L. Varuzza, S. J. Green, and J. J. Barb, “Multi-amplicon microbiome data analysis pipelines for mixed orientation sequences using QIIME2: Assessing reference database, variable region and pre-processing bias in classification of mock bacterial community samples,” *PLoS One*, vol. 18, no. 1, p. e0280293, Jan. 2023, doi: 10.1371/journal.pone.0280293.
- [123] B. J. Callahan, P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes, “DADA2: High-resolution sample inference from Illumina amplicon data,” *Nat Methods*, vol. 13, no. 7, pp. 581-583, Jun. 2016, doi: 10.1038/nmeth.3869.
- [124] S. Andrés *et al.*, “Essential oil supplementation in milk replacers: short- and long-term impacts on feed efficiency, the faecal microbiota and the plasma metabolome in dairy calves,” *J Dev Orig Health Dis*, vol. 15, p. e5, Apr. 2024, doi: 10.1017/S2040174424000084.
- [125] P. D. Schloss *et al.*, “Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities,” *Appl Environ Microbiol*, vol. 75, no. 23, pp. 7537-7541, Dec. 2009, doi: 10.1128/AEM.01541-09.
- [126] L.-A. G. Molano, S. Vega-Abellana, and C. Manichanh, “GSR-DB: a manually curated and optimized taxonomical database for 16S rRNA amplicon analysis,” *mSystems*, vol. 9, no. 2, Feb. 2024, doi: 10.1128/msystems.00950-23.
- [127] A. Nekrutenko and J. Taylor, “Next-generation sequencing data interpretation: enhancing reproducibility and accessibility,” *Nat Rev Genet*, vol. 13, no. 9, pp. 667-672, Sep. 2012, doi: 10.1038/nrg3305.

Appendix

Metagenomics pipelines command lines¹

Common steps: cutadapt and sickle

To process evenly the datasets considered in this study, the raw data underwent 2 common steps before entering the specific pipeline. Mock communities skipped these steps.

Cutadapt

Cutadapt was meant to remove the sequencing primers, as the samples came from the facility demultiplexed but not cleaned.

```
cutadapt -g <forward_primer_sequence> -G <reverse_primer_sequence> -o  
<output_folder>/cutadapt/${file}_R1.fastq.gz -p  
<output_folder>/cutadapt/${file}_R2.fastq.gz <sample_path>/${file}_R1.fastq.gz  
<sample_path>/${file}_R2.fastq.gz --discard-untrimmed
```

Sickle

Sickle was instead used to filter the sequences with a Phred quality score (Q score) quality lower than 19 to guarantee reads with decent quality. Sickle required to specify parameter *-t* as Type of quality values. Although the reads here considered were derived from Illumina platform, sanger option allowed a better functioning of the operation.

```
sickle pe -f <output_folder>/cutadapt /${file}_R1.fastq.gz -r <output_folder>/cutadapt  
/${file}_R2.fastq.gz -o <output_folder>/sickle/${file}_R1.fastq -p  
<output_folder>/sickle/${file}_R2.fastq -s <output_folder>/sickle/${file}_singles -t  
sanger -q 19
```

¹ As mentioned beforehand in this thesis, all the pipelines were used using singularity containers: therefore, all the commands were launched using *singularity run <singularity_container> command*. Mock communities required specific threshold, specified per command.

QIIME version 1.9

Instruction for QIIME version 1.9 usage were retrieved from Biscarini et al., 2018 and modified when necessary.

Joining paired-end reads

```
multiple_join_pairied_ends.py      --input_dir=<output_folder>/sickle      --
output_dir=<output_folder>/QIIME1.9/joined  --parameter_fp=  join.parameters  --
read1_indicator=_R1 --read2_indicator=_R2
```

For paired-ends reads joining, “SeqPrep” method (<https://github.com/jstjohn/SeqPrep>) was used via the parameter file (QIIME_parameters):

```
join_pairied_ends:pe_join_method SeqPrep
```

Quality filtering

```
multiple_split_libraries_fastq.py  --input_dir=<output_folder>/QIIME1.9/joined  --
read_indicator _R1 -p filter.parameters --output_dir=<output_folder>/QIIME1.9/filter --
include_input_dir_path
```

Quality filter parameter were specified via the parameter file (QIIME_parameters):

```
split_libraries_fastq:max_bad_run_length          3
split_libraries_fastq:min_per_read_length_fraction 0.01
split_libraries_fastq:sequence_max_n 0
split_libraries_fastq:phred_quality_threshold 19
```

Mock communities required parameters *split_libraries_fastq:sequence_max_n 0* and *split_libraries_fastq:phred_offset 33*.

OTU picking

```
pick_closed_reference_otus.py --input_fp =<output_folder>/QIIME1.9/filter/seqs.fna --
output_dir =<output_folder>/QIIME1.9/out_picking --reference_fp $fasta --taxonomy_fp
$taxonomy -force
```

Format conversion

```
biom      convert      -i      <output_folder>/QIIME1.9/otu_table.biom      -o  
=<output_folder>/QIIME1.9/otu_table.tsv --to-tsv --header-key taxonomy
```

QIIME version 2

Common steps

QIIME version 2 requires the creation of a specific manifest that contain information regarding the location of the input raw reads. More details on how to compile it can be found at <https://docs.QIIME2.org/2024.5/tutorials/metadata/>. QIIME2 developers also made available a webtool to validate metadata files called Keemei. More recently, a tool was made available as metadata generator, and is publicly available on GitHub: <https://github.com/charliechen912ilovbash/QIIME2-manifest-metadata-generator>.

Data import

```
QIIME tools import --type 'SampleData[PairedEndSequencesWithQuality]' --input-path
manifest.tsv --output-path <output_dir>/QIIME2-imported.qza --input-format
PairedEndFastqManifestPhred33V2
```

Joining paired-end reads

```
QIIME vsearch merge-pairs --i-demultiplexed-seqs <output_dir>/QIIME2-imported.qza --o-
merged-sequences <output_dir>/imported-joined.qza
```

Denoising approach (classic approach)

Denoising

```
QIIME dada2 denoise-paired --i-demultiplexed-seqs <output_dir>/imported-joined.qza --
p-trim-left-f 0 --p-trim-left-r 0 --o-representative-sequences
<output_dir>/representative-sequences.qza --o-table <output_dir>/table.qza --o-
denoising-stats <output_dir>/denoising-stats.qza --p-trunc-len-f 0 --p-trunc-len-r 0
```

Classify

```
QIIME feature-classifier classify-sklearn --i-classifier $classifier --i-reads
<output_dir>/representative-sequences.qza --o-classification <output_dir>/imported-
dada2-taxonomy.qza
```

ASV table production and exportation

```
QIIME taxa collapse --i-table <output_dir>/table.qza --i-taxonomy
<output_dir>/imported-dada2-taxonomy.qza --p-level 7 --output-dir
<output_dir>/collapsed_biom_table
```

```
QIIME tools export --input-path <output_dir>/collapsed_biom_table/collapsed_table.qza
--output-path <output_dir>/collapsed_biom_table/converted_to_biom
biom convert -i <output_dir>/collapsed_biom_table/converted_to_biom/feature-
table.biom -o <output_dir>/feature-table.tsv --to-tsv
```

Clustering approach (alternative approach)

Clustering

```
QIIME deblur denoise-16S --i-demultiplexed-seqs <output_dir>/imported-joined.qza --p-
trim-length 250 --p-sample-stats --o-representative-sequences
<output_dir>/representative-sequences.qza --o-table <output_dir>/table.qza --o-stats
<output_dir>/deblur-stats.qza
```

For mock communities, *--p-trim-length* was set at 0.

Classify

```
QIIME feature-classifier classify-sklearn --i-classifier $classifier --i-reads
<output_dir>/representative-sequences.qza --o-classification <output_dir>/imported-
dada2-taxonomy.qza
```

OTU table production and exportation

```
QIIME tools export --input-path <output_dir>/collapsed_biom_table/collapsed_table.qza
--output-path <output_dir>/collapsed_biom_table/converted_to_biom
```

```
biom convert -i <output_dir>/collapsed_biom_table/converted_to_biom/feature-  
table.biom -o <output_dir>/feature-table.tsv --to-tsv
```

Micca

Common steps

```
micca mergepairs -i <output_dir>/sickle/*_R1.fastq -o <output_dir>/merged.fastq -l 32 -d 8
micca convert -i <output_dir>/merged.fastq -o <output_dir>/merged.fasta
```

Clustering approach (closed reference approach)

```
micca otu -m closed_ref -i <output_dir>/merged.fasta --ref $fasta -o <output_dir>/closed_ref
micca classify -m cons -i <output_dir>/closed_ref/otus.fasta -o <output_dir>/closed_ref/taxa_closedref.txt -r $fasta -x $taxonomy --cons-id 0.99
```

Denoising approach (denovo unoise approach)

```
micca otu -m denovo_unoise -i <output_dir>/merged.fasta --ref $fasta -o <output_dir>/denovo_unoise
micca classify -m cons -i <output_dir>/denovo_unoise/otus.fasta -o <output_dir>/denovo_unoise/taxa_denovounoise.txt -r $fasta -x $taxonomy --cons-id 0.99
```

Vsearch

Joining paired-end reads

Vsearch requires a loop on the first step, that allows to call the samples one by one to undergo the joining of the paired-ends reads.

```
for f in <output_dir>/sickle/*_R1.fastq; do
  r=$(sed -e "s/_R1/_R2/" <<< "$f")
  s=$(cut -d "/" -f4 <<< "$f")
  n=$(cut -d "_" -f1 <<< "$s")
  echo Processing $n
  vsearch --fastq_mergepairs $f --reverse $r --fastq_maxdiffs 8 --fastq_minovlen 32 --fastout <output_dir>/$n.merged.fasta
```

```
vsearch --derep_fulllength <output_dir>/${n}.merged.fasta --strand plus --output
<output_dir>/${n}.merged.derep.fasta --uc <output_dir>/${n}.derep.uc --relabel ${n}. --
sizeout --fasta_width 0
done
```

```
cat <output_dir>/**/*.merged.derep.fasta > <output_dir>/all.merged.derep.fasta
```

Clustering

```
vsearch --usearch_global <output_dir>/all.merged.derep.fasta --db $fasta --otutabout
<output_dir>/otutab.txt --id 1.0
```

R scripts key commands

Phyloseq

Please note that this script was originally developed by Dr. Filippo Biscarini.

```
fname = "<your_otuasvtable>.biom"
conf_file = "<your_mapping_file>.csv"
min_tot_counts = 0 ## minimum number of total counts per sample to be included in
the analysis
outdir = file.path(results_folder)

repo = file.path(HOME, "<your_folder>")
source(file.path(repo, "support_functions/dist2list.R")) ##
from: https://github.com/vmikk/metagMisc/
source(file.path(repo, "support_functions/phyloseq_transform.R")) ##
from: https://github.com/vmikk/metagMisc/

writeLines(" - reading the filtered (OTU-wise) biom file into phyloseq")
biom_otu_tax <- phyloseq::import_biom(BIOMfilename =
file.path(prj_folder, biom_folder, fname))

writeLines(" - removing samples with too few total counts")
biom_otu_tax = prune_samples(sample_sums(biom_otu_tax)>=min_tot_counts,
biom_otu_tax)
```

```

otu = otu_table(biom_otu_tax, taxa_are_rows = TRUE)
taxa = tax_table(biom_otu_tax)

print(paste("N. of OTUs read from biom file is:", nrow(otu)))
print(paste("N .of samples retained after filtering is:", ncol(otu)))

colnames(otu) <- paste("sample",colnames(otu),sep="")
writeLines(" - change the names of taxonomic levels to Kingdom, Class etc.")
colnames(taxa) <- c("Kingdom","Phylum","Class","Order","Family","Genus","Species", "", "",
"", "", "", "", "") # colnames(taxa) <-
c("Kingdom","Phylum","Class","Order","Family","Genus","Species")

## metadata
writeLines(" - reading the metadata")
metadata = fread(file.path(prj_folder, analysis_folder, conf_file))
names(metadata)[1] <- "sample-id"
if(is.numeric(metadata$`sample-id`)) metadata$`sample-id` =
paste("sample",metadata$`sample-id`,sep="") # in case your sample-id are not only
numeric, remove or comment if(is.numeric(metadata$`sample-id`))
metadata <- as.data.frame(metadata)
row.names(metadata) <- metadata$`sample-id`
metadata$`sample-id` <- NULL

## read into phyloseq
writeLines(" - add metadata to the phyloseq object")
samples = sample_data(metadata)
otu_tax_sample = phyloseq(otu,taxa,samples)
sample_data(otu_tax_sample)

# Alpha diversity
alpha = estimate_richness(otu_tax_sample, split = TRUE)
alpha$`sample-id` = row.names(alpha)
alpha = relocate(alpha, `sample-id`)
fname = file.path(prj_folder, results_folder, "alpha.csv")
fwrite(x = alpha, file = fname)

## Preprocessing: e.g. filtering
otu_tax_sample_norm = phyloseq_transform_css(otu_tax_sample, norm = TRUE, log =
FALSE)

## add taxonomy to normalised counts
otu_css_norm = base::as.data.frame(otu_table(otu_tax_sample_norm))
otu_css_norm$tax_id = row.names(otu_css_norm)
otu_css_norm <- relocate(otu_css_norm, tax_id)
taxonomy = as.data.frame(tax_table(otu_tax_sample_norm))
taxonomy$tax_id = row.names(taxonomy)

```

```

taxonomy <- relocate(taxonomy, tax_id)
otu_css_norm = otu_css_norm %>% inner_join(taxonomy, by = "tax_id")
fname = file.path(prj_folder, results_folder, "otu_norm_CSS.csv")
fwrite(x = otu_css_norm, file = fname)

## relative abundances
otu_relative = transform_sample_counts(otu_tax_sample, function(x) x/sum(x) )
otu_rel_filtered = filter_taxa(otu_relative, function(x) mean(x) > 5e-3, TRUE)
nrow(otu_table(otu_rel_filtered))

# Beta diversity: distances matrices
writeLines("- beta diversity: distance matrices")
writeLines("- available distance metrics")
dist_methods <- unlist(distanceMethodList)
print(dist_methods)

## bray-curtis
writeLines("- calculate Bray-Curtis distances")
distances = distance(otu_tax_sample_norm, method="bray", type = "samples")
iMDS <- ordinate(otu_tax_sample_norm, "MDS", distance=distances)
dd = dist2list(distances, tri = FALSE)
dx = spread(dd, key = "col", value = "value")
fname = file.path(prj_folder, results_folder, "bray_curtis_distances.csv")
fwrite(x = dx, file = fname)

```

PERMANOVA

```

source(file.path(repo, "support_functions/adonis_OmegaSq.R")) ##
from: https://stackoverflow.com/questions/74446879/how-can-the-effect-size-of-a-permanova-be-calculated

```

```

PERMANOVA1 = adonis2(<Your_squared_matrix_without_metadata> ~
<your_matrix_with_metadata>$<column(s)_of_interest>)

```

```

PERMANOVA1
adonis_OmegaSq(PERMANOVA1)

```

Rarefaction curves

```

library(phyloseq)
library(tidyverse)
library(vegan)
library(data.table)

```

```

otuasvtable <- fread("~/<your_rarefaction_curve>")
otuasvtable[is.na(otuasvtable)] <- 0
otuasvtable <- as.data.frame(t(otuasvtable))
metadata <- as.data.frame(colnames(otuasvtable))
test_rare <- rarecurve(otuasvtable, step = 2, tidy = T)
plot <- ggplot(test_rare) + geom_line(aes(x = Sample, y = Species, colour =
Combination)) + theme_bw() + ggtitle("Clustering/Denoising approaches on X dataset" )
+
  xlab("Sample size") + ylab("Taxa") + theme(plot.title = element_text(face = "bold"),
legend.position = "bottom") + labs(color = "Pipeline + database combinations")
ggsave("~/<your_rarefaction_curve>.png", plot = plot)

```

List of publications related to the PhD

Published papers

1. Ranilla MJ, Andrés S, **Gini C**, Biscarini F, Saro C, Martín A, Mateos I, López S, Giráldez FJ, Abdennebi-Najar L, Pereira D, Falleh H, Ksouri R, Cremonesi P, Castiglioni B, Ceciliani F. Effects of *Thymbra capitata* essential oil on in vitro fermentation end-products and ruminal bacterial communities. *Sci Rep*. 2023 Mar 13;13(1):4153. doi: 10.1038/s41598-023-31370-9. PMID: 36914736; PMCID: PMC10011596 [Paper, co-author].
2. Andrés S, **Gini C**, Ceciliani F, Gutiérrez-Expósito D, Arteche-Villasol N, Martín A, Cremonesi P, Faré F, Ghaffari MH, Giráldez FJ, Abdennebi-Najar L. Essential oil supplementation in milk replacers: short- and long-term impacts on feed efficiency, the faecal microbiota and the plasma metabolome in dairy calves. *J Dev Orig Health Dis*. 2024 Apr 2;15:e5. doi: 10.1017/S2040174424000084. PMID: 38563206.

Submitted manuscripts

1. Nehme R† , **Gini C†** , Vanbergue E, Even S, Biscarini F, Andrés S, Rault L, Noel F, Hardit V, Bouhallab S, Pereira DM, Ksouri R, Roussel P, López S, Cremonesi P, Castiglioni B, Caruso D, Faré F, Fontana M, Ceciliani F, Abdennebi-Najar L. The effects of *Thymus capitatus* essential oil topical application on milk quality: a systems biology approach. [Submitted to Scientific Reports, first author]

Manuscripts to be submitted

1. **Gini C**, Ceciliani F, Cánovas A, Biscarini F. Benchmarking bioinformatic pipelines and microbial databases for the analysis of 16S rRNA-gene sequencing data

Conference proceedings

1. **Gini C**, Ceciliani F, Abdennebi-Najar L, Cánovas A, Biscarini F. Behind the scenes of metataxonomy: how bioinformatic pipelines and microbial databases compare in the analysis of 16S rRNA-gene sequencing data. In: Ontario Ministry of Agriculture, Food and Rural Affairs (OMAFRA) animal biosciences graduate student symposium, Guelph, Ontario, Canada, 08-12-2023 [Oral presentation]
2. **Gini C**, Ceciliani F, Abdennebi-Najar L, Lam S, Cánovas A, Biscarini F. The analysis of 16S rRNA-gene sequencing data through bioinformatic pipelines and microbial databases: what is inside the "black boxes"? Dairy Cattle Breeding and Genetics Committee. Guelph, Ontario, Canada, 5-10-2023. [Technical report, Oral presentation]
3. **Gini C**, Ceciliani F, Abdennebi-Najar L, Cánovas A, Biscarini F. Benchmarking Bioinformatic Pipelines and Microbial Databases for the Analysis of Bovine 16S rRNA-Gene Sequencing Data. International Plant & Animal Genome / PAG 31, San Diego, CA, USA. From 12-01-2024 to 17-01-2024 [Oral presentation]

Posters

1. **Gini C**, Biscarini F, Andrés S, Abdennebi-Najar L, Hamrouni I, Ranilla MJ, Cremonesi P, Castiglioni B, Ceciliani F. In vitro evaluation of Thymus capitatus essential oil (and compounds) effects on rumen microbiota. 73rd Annual Meeting of European Federation of Animal Science (EAAP) in Porto, Portugal. From 02-09-2022 to 08-09-2022. [Poster]
2. **Gini C**, Ceciliani F, Abdennebi-Najar L, Cánovas A, Biscarini F. Benchmarking Bioinformatic Pipelines for the Analysis of Bovine 16S rRNA-Gene Sequencing Data: from your best herds to our best results. Annual Dairy Research Showcase, Elora, Ontario, Canada, 2 November 2023 [Poster]
3. **Gini C**, Ceciliani F, Abdennebi-Najar L, Cánovas A, Biscarini F. Benchmarking Bioinformatic Pipelines and Microbial Databases for the Analysis of Bovine 16S rRNA-Gene Sequencing Data. International Plant & Animal Genome / PAG 31, San Diego, CA, USA, January 12-17, 2024 [Poster]

List of other publication not related to the present thesis project

1. Lecchi C, Cecilian F, Petrini S, Cappelli G, Grassi C, Balestrieri A, Galiero G, DeCarlo E, Salvi G, Panzeri F, **Gini C**, Cafiso A, Agazzi A, Martucciello A. Endogenous and viral microRNAs in nasal secretions of water buffaloes (*Bubalus bubalis*) after Bubaline alphaherpesvirus 1 (BuHV-1) challenge infection. *Vet Res.* 2023 Jun 5;54(1):44. doi: 10.1186/s13567-023-01175-9. PMID: 37277883; PMCID: PMC10242922. [Paper, co-author]
2. Matos LG, da Silva Abreu AC, Pereira Perez Alonso V, Gonçalves JL, Nascimento MS, Bertelli Pflanzner Jr S, Rezende de Souza JH, **Gini C**, Faraj Murad N, Mendes Brandão M, Cirone Silva NC. Comparison of Bacterial Diversity of Wet- and Dry-Aged Beef Using Traditional Microbiology and Next Generation Sequencing. <https://doi.org/10.1016/j.microb.2024.100035> [Paper, co-author].
3. de Oliveira Duarte FA, Ramos KK, **Gini C**, Morasi RM, Silva NCC, Efraim P. Microbiological characterization of kombucha and biocellulose film produced with black tea and cocoa bean shell infusion. *Food Res Int.* 2024 Aug;190:114568. doi: 10.1016/j.foodres.2024.114568. Epub 2024 Jun 4. PMID: 38945598.

Acknowledgements

*This is ten percent luck
Twenty percent skill
Fifteen percent concentrated power of will
Five percent pleasure
Fifty percent pain
And a hundred percent reason to remember the name
~Fort Minor, 2005*

To Dr. Filippo Biscarini, my "wizard master", for patiently guiding me along the initiatory path of bioinformatics and biostatistics.

To Professor Fabrizio Ceciliani, for the challenges he has offered and helped me to face, and for training me in the world of research.

To Professor Angela Cànovas, for welcoming me at the University of Guelph and allowing me to get the most out of this life-changing experience.

To Professor Cristina Lecchi, Gabriela, Susanna, for providing me with a dynamic research group in which to spend these years.

To Shuvo for arriving without pretensions and showing me that friendships are built slowly, even in academia. For sharing with me the office, food, cultural exchange, and the most intense experiences of my doctorate: ধন্যবাদ বস

To my Canadian team, Alice, Ricarda, Victoria, Steph, Sirlene, Isis, Larissa, Ana Carolina, Bruno, Walmir, Lucas, Colin, Kristin, Samla, Krishani, Coralie, Julia, for making my first experience at UoG unforgettable both professionally and personally. See you at the conference, rising stars!

To Carlotta and Valentina, without whose support I would never have started this experience.

To Gabriele, who has been the older brother of this doctorate: I could have never imagined that life would have honoured me with a friend like you.

To Andrea, for making me laugh and eat when I needed it most, and there's nothing else we need to add.

To Laura, Delower, Barbara and the Deed gang, for keeping me company in the evening fog of Lodi.

To Ivan, Alessandro, Selene, Antonio, Andrea, Baldo, Riky and Chiara, Stefania and Stefano for making me feel at home every time I came back and for always being that place where I could come back as if I had never left.

To Christian, Claudia, Christian and Mapi of the Croce Verde Fino Mornasco Volunteer Group for keeping alive the fire of a shared ideal.

To Shannon and Odin, Jesùs, Rowena, Deep, Shourin and Manuel of the Salvation Army Young Adults and their amazing Tuesday nights, for making me the honorary Canadian who can no longer stop thinking about traveling.

To Sarah, for always supporting me "from a distance": because the Este-Como axis bends, stretches, dives into the ocean and then returns, but never breaks.

To Cristina, Gigi and Rock, Makeba, Frantoio and the newcomer Solea, for making me feel at home when I unplugged. To Francesca, for patiently enduring the bad days and helping me to flood the good ones with sunshine, growing up with me.

To my parents, Marisa and Francesco, for tirelessly supporting me along my path, every day, and looking with confidence to its continuation, even though sometimes they couldn't see it, "because it's your way, not mine (Point Break, 2015)" but "we are the safe harbour". I hope I have made you nothing less than proud.

To Andrea and Napoleone, for showing me that a brother's love does not know what it means to be tired.

Thank you for being a part of my life.

Virtute Siderum Tenus,

I finished the PhD.

Ringraziamenti

*Questo è dieci per cento fortuna
Venti per cento abilità
Quindici per cento concentrato di forza di volontà
Cinque per cento piacere
Cinquanta per cento sofferenza
E fa cento per cento dei motivi per ricordare il mio nome*

~Fort Minor, 2005

Al dottor Filippo Biscarini, il mio “maestro stregone”, per avermi guidato con pazienza lungo il percorso iniziatico della bioinformatica e della biostatistica.

Al professor Fabrizio Ceciliani, per le sfide che mi ha offerto ed aiutato ad affrontare e per avermi formato al mondo della ricerca.

Alla professoressa Angela Cànovas, per avermi accolto presso l’Università di Guelph ed avermi consentito di trarre il massimo da quest’esperienza che mi ha cambiato la vita.

Alla professoressa Cristina Lecchi, a Gabriela, a Susanna, per avermi offerto un dinamico gruppo di ricerca in cui trascorrere questi anni.

A Shuvo per essere arrivato senza pretese ed avermi dimostrato che le amicizie si costruiscono piano, anche nel mondo accademico. Per aver condiviso con me l’ufficio, il cibo, lo scambio culturale e le esperienze più intense del mio dottorato: ধন্যবাদ বস

Al mio gruppo canadese, Alice, Ricarda, Victoria, Steph, Sirlene, Isis, Larissa, Ana Carolina, Fatima, Bruno, Walmir, Lucas, Colin, Kristin, Samla, Krishani, Coralie, Julia, per aver reso la mia prima esperienza in UoG indimenticabile a livello lavorativo e personale. Ci vediamo a congresso, astri nascenti, eh!

A Carlotta e Valentina, senza il cui apporto non avrei mai cominciato questa esperienza.

A Gabriele, che è stato il fratello maggiore di questo dottorato: mai avrei pensato di potermi fregiare dell’onore di un amico come te.

Ad Andrea, per avermi fatto ridere e mangiare quando più ce n’era bisogno, e non c’è altro che abbiamo bisogno di aggiungere.

A Laura, Barbara ed alla Deed gang, per avermi tenuto compagnia in mezzo alle serali nebbie lodigiane.

A Ivan, Alessandro, Selene, Antonio, Andrea, Baldo, Riky e Chiara, Stefania e Stefano, per avermi fatto sentire a casa ogni volta che ci tornavo e per essere sempre quel posto in cui tornare come se non fossi mai andata via.

A Christian, Claudia, Christian e Mapi del Gruppo Volontari Croce Verde Fino Mornasco per avermi tenuto vivo il fuoco di un ideale condiviso.

A Shannon e Odin, Gesù, Rowena, Deep, Shourin e Manuel degli Young Adults della Salvation Army e ai loro martedì sera, per avermi reso la canadese onoraria che ora non riesce più a smettere di pensare a viaggiare.

A Sarah, per avermi sempre sostenuto “a distanza”: perché l’asse Este-Como si piega, si allunga, si tuffa nell’oceano e poi ritorna, ma comunque, mai si spezza.

A Cristina, Gigi e Rock, Makeba, Frantoio e alla nuova arrivata Solea, per avermi fatto sentire a casa quando staccavo la spina. A Francesca, per aver pazientemente sopportato le giornate storte ed avermi aiutato ad inondare di sole quelle giuste, crescendo insieme a me.

Ai miei genitori, Marisa e Francesco, per avermi sostenuto instancabilmente lungo il mio percorso, ogni giorno, ed aver guardato con fiducia alla sua continuazione, anche se “è la tua via, non la mia” (Point Break, 2015) ma “siamo il porto sicuro”. Spero di non avervi reso niente meno che fieri.

Ad Andrea e a Napoleone, per avermi dimostrato che l’amore di un fratello non sa cosa voglia dire essere stanchi.

Grazie di essere parte della mia vita.

Virtute Siderum Tenuis,

Ho finito il dottorato.