University of São Paulo
"Luiz de Queiroz" College of Agriculture

Detection of mastitis-causing pathogen by sequencing different regions of the 16S rRNA gene and machine learning

Luan Gaspar Clemente

Dissertation presented to obtain the degree of Master in Science. Area: Animal Science and Pastures

Piracicaba
2022

Luan Gaspar Clemente
Agronomist

Detection of mastitis-causing pathogen by sequencing different regions of the 16S rRNA gene and machine learning

Advisor:
Prof. Dr. **LUIZ LEHMANN COUTINHO**

Dissertation presented to obtain the degree of Master in Science. Area: Animal Science andPastures

Piracicaba
2022

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Clemente, Luan Gaspar

Detection of mastitis-causing pathogen by sequencing different regions of the 16S rRNA gene and machine learning / Luan Gaspar Clemente.- - Piracicaba, 2022.

37 p.

Dissertação (Mestrado)- - USP / Escola Superior de Agricultura “Luiz de Queiroz”.

1. Aprendizado de máquina 2. Mastite bovina 3. 16S rRNA 4. *Staphylococcus aureus* 5. *Escherichia coli* I. Título

ACKNOWLEDGMENTS

I would like to thank to my family, Alzira and Tania, for alwas being there for me. Even when disagree with my choices, they created a safe space for persuit my dreams. To all of my friends, for sharing pleasant moments, allowing me to always move forward with energy and serenity in the face of problems.

I would also like to thank my advisor, Prof. Dr. Luiz Lehmann Coutinho, for his guidance, help and support, and confidence in me. I always learn something from you, even in the most informal, day-to-day conversations. To my LBA team, in particular Ingrid and Barbara.

Finally, I wish to express my gratitude to University of São Paulo – "Luiz de Queiroz" College of Agriculture (USP/ESALQ) for the privilege granted and for aggregating so much knowledge in my academic history. A special acknowledgement to CAPES for supporting me with a scholarship.

SUMMARY

RESUMO

Detecção de patógenos causadores de mastite pelo seqüenciamento de diferentes regiões do gene 16S rRNA e aprendizado de máquina

A correta identificação de patógenos causadores de mastite é um fator chave para o sucesso do manejo das fazendas leiteiras. Técnicas como meio de cultura, qPCR e sequenciamento de 16S rRNA têm sido utilizadas para detectar microrganismos importantes em amostras de leite bovino cru. No entanto, devido aos custos, alguns desafios permanecem. Os métodos de aprendizado de máquina têm se mostrado uma alternativa atraente, pois podem integrar diferentes fontes de dados, com diversas finalidades. Novos estudos com foco na detecção de mastite clínica e subclínica destacam o potencial de métodos de aprendizado de máquina aplicados ao manejo da mastite em fazendas leiteiras. Neste trabalho, avaliamos o desempenho de três métodos de aprendizado de máquina para detectar o patógeno causador de mastite mais abundante em amostras individuais de leite cru de bovinos integrando dados de composição do leite e sequenciamento de 16S rRNA. Mostramos o potencial para a identificação de *Escherichia coli* e *Staphylococcus aureus*. Para abundância superior a 3% em amostras individuais, uma precisão de 100% e 86% foi alcançada, respectivamente. Esses resultados mostram que não apenas a mastite subclínica e clínica pode ser detectada por métodos de aprendizado de máquina, mas também alguns patógenos causadores de mastite. Além disso, para maximizar as informações obtidas do sequenciamento do gene 16S rRNA, avaliamos a diversidade genética in silico para diferentes regiões do gene 16S rRNA e validamos os resultados pelo sequenciamento Illumina. Mostramos que para melhor detecção de microrganismos associados à mastite bovina, a região V2-V3 detecta maior prevalência com maior abundância relativa. Esperamos que este trabalho possa contribuir para um melhor manejo das propriedades leiteiras bem como o desenvolvimento de novas ferramentas para o controle da mastite bovina.

Palavras-chave: Aprendizado de máquina, Mastite bovina, 16S rRNA, *Staphylococcus aureus*, *Escherichia coli*

ABSTRACT

Detection of mastitis-causing pathogen by sequencing different regions of 16S rRNA gene and machine learning

The correct identification of mastitis-causing pathogens is a key factor in the successful management of dairy farms. Techniques such as culture medium, qPCR, and 16S rRNA sequencing have been used to detect important microorganisms in raw bovine milk samples. However, due to costs, some challenges remain. Machine learning methods have been shown as an attractive alternative, as they can integrate different sources of data, with a diversity of purposes. New studies focusing on the detection of clinical and subclinical mastitis highlight the potential of applied machine learning methods to the management of mastitis in dairy farms. In this work, we evaluate the performance of three machine learning methods to detect the most abundant mastitis-causing pathogen in individual raw milk bovine samples integrating data from milk composition and 16S rRNA sequencing. We show the potential for the identification of *Escherichia coli* and *Staphylococcus aureus*. For abundance greater than 3% in individual samples, an accuracy of 100% and 86% was achieved, respectively. These results show that not only subclinical and clinical mastitis can be detected by machine learning methods, but some mastitis-causing pathogens either. Moreover, to maximize the information obtained from 16S rRNA sequencing, we evaluate in silico genetic diversity for different regions of the 16S rRNAgene and validate the results by Illumina sequencing. We show that for better detection of microorganisms associated with bovine mastitis, the V2-V3 region detects a higher prevalence with more relative abundance. We hope that this work can contribute to better management of dairy farms as well as the development of new tools for the control of bovine mastitis.

Keywords: Machine learning, Bovine mastitis, 16S rRNA, *Staphylococcus aureus*, *Escherichia coli*

1. INTRODUCTION

Mastitis is considered to be the most costly disease in the dairy industry in the world mainly due to reduced milk production and longevity of animals, changes in milk quality, labor costs, diagnosis, and treatments (Oliveira et al., 2013). It can be classified as clinical and subclinical mastitis. Clinical mastitis can be detected by inspecting changes in the appearance of milk and local signs in the mammary gland. Subclinical mastitis, on the other hand, does not show obvious signs of infection and requires specific methods to detect.

The most used technique for the determination of subclinical mastitis is the evaluation of somatic cell count (SCC) (Sharma et al., 2011). However, the somatic cell count can't identify the causative agent. Earlier detection of the pathogen involved can lead to a better choice of treatment method, antibiotic selection, and better management strategies (Ashraf and Imran, 2018). To detect microorganisms, methods based on culture medium and guide antibiotic treatment were developed. However, approximately 25% of samples from clinical mastitis are culture-negative or do not present significant pathogens (Bradley et al., 2007). Another possibility of identification is through the use of the quantitative PCR technique (Masco et al., 2007; Malorny et al., 2008; Le Dréan et al., 2010).

The progress of next-generation sequencing (NGS) technologies and the consequent reduction in sequencing costs has revolutionized human medicine and can add value to agribusiness. One of its contributions is through the use of genetic markers (Johnson et al., 2019). The 16S small subunit ribosomal RNA (16S rRNA) gene has been established as the most widely used genetic marker in modern microbiome studies due to its conservation and universal presence in prokaryotes (Amit Roy et al., 2014).

However, the use of large-scale DNA sequencing to identify mastitis-causing pathogens is restricted to a few reports in the academic literature comparing healthy and infected udders (Porcellato et al., 2020). Moreover, recent studies have shown that the choice of 16S rRNA gene region may affect estimates of taxonomic diversity, leading to unreliable estimated proportions of different taxa between regions (Bukin et al., 2019).

Some advances have been made in low-cost alternatives for the detection of clinical and subclinical mastitis using machine learning methods. As stated by Bobbo 2021, different studies have applied machine learning techniques to diagnose mastitis. Some studies relied on the presence of high milk somamtic cell counts (SCC) (Ebrahimie et al., 2018) or mastitis pathogens (Esener et al., 2018), while others have established SCC-independent models for mastitis prediction using milking traits (Ebrahimi et al, 2019). The developed studies for the detection of mastitis-causing pathogens relies only upon few species or genera (e.g. strains of *Streptococcus uberis* in Esener et al., 2018), and a broader detection has not yet been developed.

In this study, we explore the in silico genetic diversity for different regions of the 16S rRNA gene to find the regions that maximize genetic diversity for common mastitis-causing pathogens and validate these results by metabarcoding Illumina sequencing. Moreover, we investigate the potential of applied machine learning to raw milk composition to detect the most abundant mastitis-causing pathogen.

With our results, we show that sequencing of the 16S rRNA V2-V3 region may lead to a better taxonomic assignment at the species level for mastitis-causing pathogens. Moreover, machine learning methods applied to the detection of mastitis-causing pathogens may lead to the identification of *Escherichia coli* and *Staphylococcus aureus* on individual raw milk samples, with low-cost data (e.g. somatic cell counts, total bacteria count, fat, protein, lactose, total solids, and defatted dry extract).

1.1. General Objectives

Identify the 16S rRNA region that maximizes the detection of mastitis-causing pathogens.

Evaluate different machine learning models to predict mastitis-causing pathogens in individual raw milk samples by low-cost data, such as somatic cell count, total bacteria count, fat, protein, lactose, total solids, and defatted dry extract.

1.2. Specific Objectives

(a) In silico genetic diversity evaluation of V2-V3, V4, and V5-V6 regions to find the region that maximizes species identification for mastitis-causing pathogens.
(b) Validate the in silico genetic diversity evaluation results by metabarcoding 16S rRNA Illumina sequencing.
(c) Evaluate three different models to predict the most abundant mastitis-causing pathogens in individual raw milk samples.

REFERENCES

Amit Roy, S. R. (2014). Molecular Markers in Phylogenetic Studies - A Review.Journal of Phylogenetic & Revolutionary Biology. 2. doi: 10.4172/2329-9002.1000131.

Ashra, A., Imran, M. (2018). Diagnosis of bovine mastitis: from laboratory to farm. Tropical Animal Health and Production. 50. doi: 10.1007/s11250-018-1629-0.

Bobbo, T., Bifanni, S., Taccioli, C., Penasa, M., Cassandro, M. (2021). Comparison of machine learning methods to predict udder health status based on somatic cell counts in dairy cows. Scientific Reports. 11. doi: 10.1038/s41598-93056-4.

Bradley, A. J., Leach, K. A., Breen, J. E., Green, L. E., Green, M. J. (2007). Survey on the incidence and aetiology of mastitis on dairy farms in England and Wales. Vet. Rec. 160.doi: 10.1136/vr.160.8.253.

Bukin, Y. S., Galanchyants, Y. P., Moronov, I. V., Bukin, S. V., Zakharenko, A. S., Zemskaya, T. I. (2019). The effect of 16S rRNA region choice on bacterial community metabarcoding results. Sci. Data. 6. doi: 10.1038/sdata.2019.7.

Ebrahime, M., Mohammadi-Dehcheshmeh, M., Ebrahime, E., Petrovski, K. R. Comprehensive analysis of machine learning models for prediction of sub-clinical mastitis: Deep Learning and Gradient-Boosted Tress outperform other models. Comput. Biol. Med. 114. doi: 10.1016/j.compbiomed.2019.103456.

Ensener, N. et al. (2018). Discrimination of contagious and environmental strains of Streptococcus uberis in dairy herds by means of mass spectrometry and machine-learning.Scientific Reports. 8. doi: 10.1038/s41598-018-35867-6.

Johnson, J. S., Spakowicz, D. J., Hong, B. Y., Petersen, L. M., Demkowicz, P., Chen, L., Leopold, S. R., Hanson, B. M., Agresta, H. O., Gerstein, M., Sodergren, E., Weinstock, G. M. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis.Nature Communications. 10. doi: 10.1038/s41467-019-13036-1.

Le Dréan, G., Mounier, J., Vasseur, V., Arzur, D., Habrylo, O., Barbier, G. (2009). Quantification of Penicillumcamemberti and P. roqueforti mycelium by real-time PCR to assess their growth dynamics during ripening cheese.Int J Food Microbiol. 100. doi: 10.1016/j.ijfoodmicro.2009.12.013.

Malorny, B., Lofstrom, C., Wagner, M., Kramer, N., Hoorfar, J. (2007).Enumeration of salmonella bacteria in food and feed samples by real-time PCR for quantitative microbial risk assessment.Appl Environ Microbiol. 74. doi: 10.1128/AEM.02489-07.

Masco, L., Vanhoutte, T., Temmerman, R., Swings, R., Huys, G. (2006). Evaluation of real-time PCR targeting 16S rRNA and recA genes for the enumeration of bifidobacteria in probiotic products.Int J Food Microbiol. 113. doi: 10.1016/j.ijfoodmicro.2006.07.021.

Oliveira, L., Hulland, C., Ruegg, P. L. (2013). Characterization of clinical mastitis occurring in cows on 50 large dairy herds in Wisconsin.Journal of Dairy Science. 96. doi: 10.3168/jds.2012-6078.

Porcellato, D., Meisal, R., Bombelli, A., Narvhus, J. A. (2020). A core microbiota dominates a rich microbial diversity in the bovine udder and may indicate presence of dysbiosis. Scientific Reports. 10. doi: 10.1038/s41598-020-77054-6.

Sharma, N., Singh, N. K., Bradwal, M. S. (2011). Relationship of Somatic Cell Count and Mastitis: An Overview. Asian-Australian Journal of Animal Science. 24. doi: 10.5713/ajas.2011.10233.

2. IN SILICO GENETIC DIVERSITY EVALUTION OF 16S RRNA REGIONS FOR PRE-SELECTED SPECIES CAN IMPROVE TAXONOMIC ASSIGNMENT

Abstract

Modern microbiome studies relies on the correct identification of microbial communities and their impact on different life phenomena. Historically, conventional techniques, such as classical Gram staining, were used for the identification of cultured bacteria from clinical, food, and environmental origins. However, cultured bacteria communities represent only a fraction of the real diversity. This limitation was partially overcome with the advent of next-generation sequencing (NGS) technologies and 16S rRNA marker gene, but challenges remain. The choice of a reference database for classification can be tricky. Larger databases make it potentially more difficult to assign taxonomy at genus and species-level as the likelihood of ambiguous assignment increases, but smaller databases possibly do not contain a sufficient representation of species. In an attempt to overcome these limitations, some dedicated reference databases were constructed for a certain niche, such as HITdb (Human Intestinal 16S rRNA database) and DAIRYdb. We hypothesize that in cases when only a few species or genera are needed to be detected, targeting a region of 16S rRNA gene marker that maximizes genetic diversity for this species would lead to a better taxonomic assignment. Here, we evaluate the genetic diversity for the V2-V3, V4, and V5-V6 regions for the most common genera associated with mastitis in bovine milk, and contrast the results of the lowest and highest genetic diversity regions by Illumina sequencing. We show that our approach increases the number of species-level assigned sequences.

2.1. Introduction

One of the main goals of modern microbiome studies is the correct identification of microbial communities and their impact on different life phenomena. Microbial communities, also known as microbiota, are groups of microorganisms that share a common living space and are present in virtually all known environments, such as oceans (Suganawa et al., 2015), soils (Thompson et al., 2017), ice (Christner et al., 2008), food (Yeluri et al., 2018), and other living organisms (Cho andBlaser, 2012).

Historically, conventional techniques, such as classical Gram staining, were used for the identification of cultured bacteria from clinical, food, and environmental samples. However, cultured bacteria communities represent only a fraction of the real diversity, being about 0.1% of complex communities as the human intestinal microbiota (Cao et al., 2017). Overcoming these limitations, microbiome studies have experienced an enhancement during the last decade with the advent of next-generation sequencing (NGS) technologies (Porter et al., 2018) and the use of genetic markers (Johnson et al., 2019).

The 16S small subunit ribosomal RNA (16S rRNA) gene has been established as the most widely used genetic marker in modern microbiome studies due to its conservation and universal presence in prokaryotes (Amit Roy et al., 2014). The generation of millions of reads per single run, increased read length, sample multiplexing, and reduced costs of high-throughput sequencing platforms, lead to an accumulation of 16S sequence data from various microbial situs, and reference databases like Silva (Quast et al., 2013) and RDP (Ribosomal Database Project) (Cole et al., 2014) have been built to enable a phylogenetic analysis of these data.

A major step in analyzing 16S rRNA data is the taxonomic assignment of the sequences. Taxonomic assignment depends on several factors, including sequence length, the target region of the 16S gene, classification method, and reference database (Huse et al., 2008; Bowen et al., 2012; Werner et al., 2011).

Larger databases make it potentially more difficult to assign taxonomy at genus and species-level as the likelihood of ambiguous assignment increases due to highly similar sequences. Furthermore, the 16S rRNA gene

shows higher ambiguous assignment at lower taxonomic levels compared with other taxonomic marker genes (Mende et al., 2013). In an attempt to overcome these limitations, some dedicated reference databases have been proposed in the last years, such as DAIRYdb(Meola et al., 2019) and HITdb (Ritari et al., 2015).

These authors hypothesize that reducing the size of the reference database to encompass only the sequences innate to the environment under study would lead to improved taxonomic classifications at lower taxonomic levels due to less competition among targets (Meola et al., 2019; Ritari et al., 2015).

Moreover, the choice of 16S rRNA gene region may affect estimates of taxonomic diversity, leading to unreliable estimated proportions of different taxa between regions and the known true composition (Bukin et al., 2019). One of the possible approaches to solving this problem is to use the experience of studying the same communities employing different 16S rRNA regions (Bukin et al., 2019).

However, it's not always necessary to detect as many taxa as possible and their real proportions. In some cases, as in the diagnosis of mastitis-causing pathogens, only a few species are correlated with the disease. Species such as *Staphylococcus aureus*, *Streptococcus agalactiae*, *Streptococcus uberis*, and *Mycoplasma uberis* are targets for mastitis-causing pathogens and commercial kits for real-time PCR are available. Despite being accurate in their purpose, these kits detect only a few species, making the 16S sequencing approach more useful due to its greater ability to detect microbial diversity.

Here, we hypothesize that in cases when only a few species or genera are needed to be detected, targeting a region of 16S rRNA gene marker that maximizes genetic diversity for these species would lead to a better taxonomic assignment. For that, we evaluate the *in silico* genetic diversity for the V2-V3, V4, and V5-V6 regions for the most common genera associated with bovine mastitis, and compare the results of the lowest and highest genetic diversity regions by Illumina sequencing.

We have shown that choosing the region that maximizes the genetic diversity for species of interest leads to greater detection. We also provide a pipeline, publicly available at (https://github.com/clementeluangaspar/Derep16S)to target the region of interest based on a list of species or genera.

2.2. Materials and Methods

2.2.1. In silico genetic diversity evaluation of 16S rRNAtarget regions

To determine the genetic diversity captured by different amplicons, we quantify the identity between each pair of sequences for the DAIRYdb 16S rRNA reference database for the V2-V3, V4, and V5-V6 regions. Moreover, we evaluate the number of unique sequences for the V2-V3, V4, and V5-V6 regions (For primer sequence, see Table 1). Unique sequences were considered the ones with less than 100 percent identity compared to all the remaining sequences.

Table 1. Primer pairs used for in silico evalution of genetic diversity between V2-V3, V4 and V5-V6 regions.

Gene	Amplicon	Primer	Sequence
16S rRNA	V2-V3	V2-V3F	ANTGGCGGACGGGTGAGTAA
		V2-V3R	GTGCCAGCAGCCGCGG
	V4	V4F	GTGNCAGCNGCCGCGGTAA
		V4R	GGACTACNNGGGTNTCTAAT
	V5-V6	V5-V6F	ATTAGATACCCNGGTAG
		V5-V6R	CGACAGCCATGCANCACCT

To find unique sequences we construct a pipeline based on three main steps: primer mapping, trimming, and clustering. In the primer mapping step, a global alignment (Bodenhofer et al., 2015) is performed between the set of primers and the sequences present in the reference fasta. Alignments of both primers with 100 percent identity and expected positions are considered valid and the sequences are trimmed in the annealing positions, excluding the primer sequence.

The trimmed sequences are clustered based on identity. Sequences that present 100 percent identity and have the same length are considered duplicated. All duplicate sequences are dereplicated and a new identification is created, providing the specie-level information of all sequences present in the cluster. All non-duplicate sequences are maintained. For this study, we considered regions with the lowest number of pairs of sequences with 100 percent of identity and the highest number of unique sequences to be the ones with the most genetic diversity.

2.2.2. Samples

We select 96 bulk tank milk samples obtained from herds located in São Paulo State, Brazil. Homogenized samples were collected from the tank before transportation of the milk to the industry. Samples were collected and stored in tubes with Bronopol® conservative with a concentration between 0,004% - 0,005%. The tubes were stored refrigerated for five days maximum until DNA extraction and library generation.

2.2.3. DNA extraction and library generation

For DNA extraction and library generation, 2mL of milk was transferred to a 2mL microcentrifuge tube and centrifuged at 14.000 rpm for 5 minutes. Supernatant and fat were discarded and the pellet was stored at -20ºC until DNA extraction. DNA extraction was done using MagMAX™ CORE combined with MagMAX™ CORE Mechanical Lysis Module (ThermoFisher™) according to the manufacturer's instructions.

The 16S Metagenomic Sequencing Library Preparation Guidelines (Illumina Inc., San Diego, CA) were performed for library construction. One library was prepared for each of the primer pairs (Table 2) to amplify hypervariable regions of the 16S rRNA gene by PCR. Equimolar quantities of each library were pooled and sequencing by MiniSeq High Output reagent kit (300 cycles) on the MiSeq platform (Illumina Inc., San Diego, CA).

Table 2. Primer pairs used for Illumina sequencing.

Gene	Amplicon	Primer	Sequence
16S rRNA	V2-V3	V2-V3F	ANTGGCGGACGGGTGAGTAA
		V2-V3R	GTGCCAGCAGCCGCGG
	V4	V4F	GTGNCAGCNGCCGCGGTAA
		V4R	GGACTACNNGGGTNTCTAAT

2.2.4. Sequencing data analysis

The DADA2, an open package implemented in the R language (Callahan et al., 2016), was used for modeling and error correction of amplicons, with the construction of ASVs (Amplicon Sequencing Variants). Studies have shown that in several simulated communities, DADA2 identified more real variants and produced fewer spurious sequences than other methods (Callahan et al., 2016). Filtering of fastq files was performed to remove sequences from the primers and low-quality bases at the end of the reads (Q<30). After filtering, the forward and reverse reads were joined to reassemble the complete fragment.

The DADA2 algorithm makes use of a parametric error model, incorporating the different error rates in each amplicon dataset. The "learnErrors" method adjusts the error model from the data, alternating between estimating error rates and inferring sample composition until convergence. Through dereplication, the list of unique sequences was obtained, with relative abundances, and then the chimeras were removed.

After initial data processing, taxonomies were assigned to each ASV using a DADA2 implementation of the Naive Bayesian classifier method (Wang et al., 2007), using the DAIRYdb reference database (Alishum, 2019) with minimum bootstrap confidence of 90.

2.2.5. Statistical comparison of species diversity

To compare community biodiversity, we calculated Shannon (Hil, 1973) Simpson (Hil, 1973) Chao1 (O'Hara, 2005), and ACE index (Chiu et al., 2014) and tested for significant differences using a paired modification of the Wilkinson-Mann-Whitney nonparametric criterion (Bauer, 1972).

To check if read counts were sufficient for characterizing community diversity we determined the correlation between sample sizes and diversity index (Shannon and Simpson) for V2-V3 and V4 regions by non-parametric Spearman correlation (Zar, 1972). All diversity indexes were measured at ASV-level by "estimate richness" function of the phyloseq R package (McMurdie and Holmes, 2013), and visualized by density plot using ggplot2 (Wickham, H. 2016).

For qualitative comparison, we evaluate the species spectrum (shared and non-shared species) of the common species associated with bovine mastitis. We also evaluate the percentage of samples that detect these species and the abundance average of each species for V2-V3 and V4 regions.

2.3. Results and discussion

2.3.1. In silico V2-V3 sequences shows higher genetic diversity compared to V4 and V5-V6 sequences for common mastitis-causing pathogens

For the evalution of the genetic diversity of common species associated with bovine mastitits, we select all sequences classified to species-level for genera *Staphylococcus*, *Sterptococcus*, *Escherichia*, *Enterococcus*, *Klebsiella*, *Serratia*, *Corynebacterium*, *Trueperella* and *Mycoplasma* available at DAIRYdb reference database.The selection of species was based on a list of species of the commercial kit real-time PCR VetMAXMastiType Multi Kit (ThermoFisher™). A total of 288 near-full 16S rRNA sequences were selected for 217 species.

For primmer mapping step, we select a primer pair for the V2-V3, V4, and V5-V6 regions of the 16S rRNA gene (Table 1). Next, we align the sequence of forward and reverse primers for each of the 288 sequences, and trimmed the sequences based on annealing sites, regardless of primer sequence. Only sequences with 100% identity at the primer site were considered valid. For the evaluation of the annealing step, we construct a *boxplot* for the position of the alignment of forward and reverse primer (Figure 1) and compare the location according to a scheme of ribosome complex and 16S rRNA gene available at (Fukuda et al., 2016).

As can be seen in Figure 1, we obtained reliable annealing for the forward and reverse primers for the V2V3, V4, and V5-V6 primers. Although we have variability in the annealing positions, this may be due to an incompleteness of the sequences, especially in the V1 region. The 288 sequences have a mean length of 1486 bases, with the first quarter of 1452 and third quarter of 1483, a minimum of 1051 and a maximum of 2230, so it is expected to have variability on the annealing position, however, the distance between forward and reverse primers must be similar.

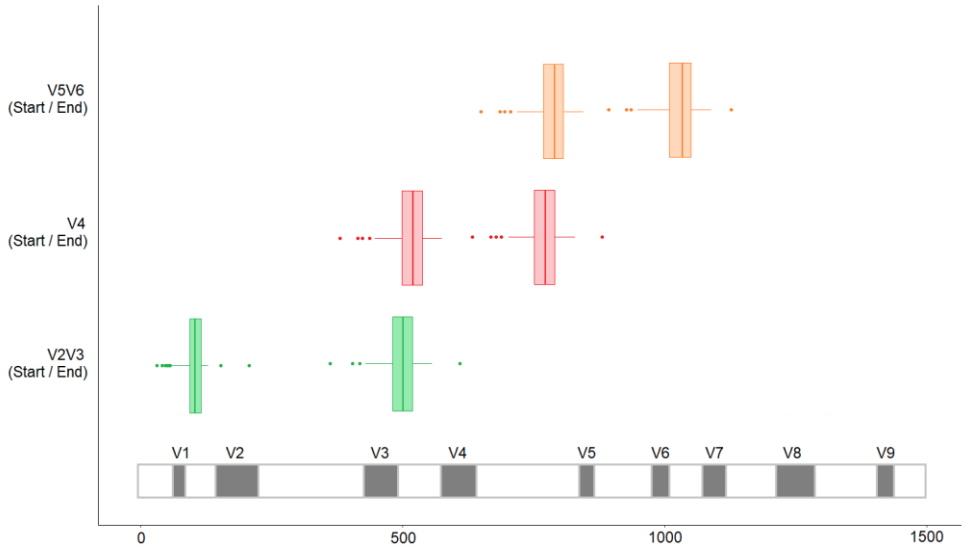

Figure 1. Boxplot with annealing position of forward and reverse primers for V2-V3, V4, V5-V6 regions with scheme of ribosome complex and 16S rRNA gene adapted from (Fukuda, K., 2016). The X-axis represents a position within the 16SrRNA gene, and the Y-axis, the V2-V3, V4, and V5-V6 primer pairs. The values were colored by primers pairs. The green color represents the values for V2-V3 regions, the red color represents the V4 regions and the orange represent the V5-V6 region.

For discarding inaccurate alignments, we select only trimmed sequences with length around the median length for the region. For the V2-V3 regions, we considered a length of 400 bases, 254 bases for V4, and 240 bases for the V5-V6 regions. We obtained a total of 273 sequences for the V2-V3 regions, 279 sequences for V4, and 277 sequences for V5-V6.

To evaluate the genetic diversity of the sequences in the DAIRYdb, we calculate the identity for each pair of sequences for V2-V3, V4, and V5-V6 regions using "pid" function of the Biostrings R package (Pagès et al., 2022). After the construction of a matrix of identity for each pair of sequences, we determine the percentage of pairs with 100 percent identity. A total of 415 (0.55%) were obtained for V2-V3 regions, 1475 (1.89%) for V4 region and 1011 (1.32%) for V5-V6 regions. To illustrate the genetic diversity we construct a heatmap of the identity value of each pair of sequences (Figure 2).

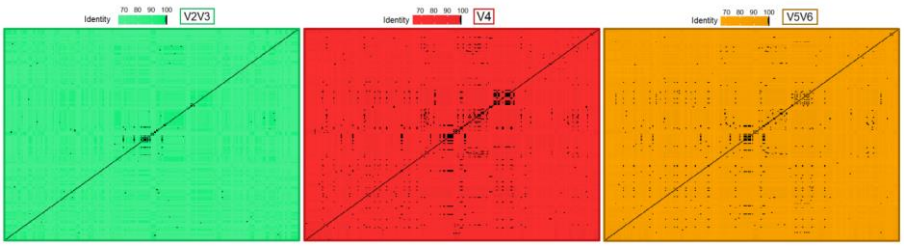

Figure 2. Heatmap of the identity value for each pair of sequences from the 288 sequences of DAIRYdb reference database. The colored squares represent values with identity lower than 100. The green color were used for V2-V3 regions, red color for V4 regions and orange color for V5-V6 regions. The black squares represent pairs of sequences with 100 percent identity.

Moreover, we compare the remaining sequences based on identity and remove duplicated sequences using the function"unique"of r-base. A total of 239 unique sequences were obtained for the V2-V3 region, 170 unique sequences for the V4 region, and 189 unique sequences for the V5-V6 region. Based on these results we select the V2-V3 (highest genetic diversity) and V4 (lowest genetic diversity) regions for validation by 16S rRNA sequencing of milk samples.

2.3.2. Taxonomical classification is greatly impacted by target region.

The sequencing of 16S rRNA fragments from the 96 raw milk samples yielded a total of 2,907,146 paired-end reads (mean of 30,462 paired-end reads with sd (Standard deviation) of 4,072.48) for V2-V3 regions and 7,757,884 paired-end-reads (mean of 80,811 paired-end reads with sd of 30,284) for V4 region.After quality control, denoising, and exclusion of chimeras we retained a total of 2,149,292 (73.93%) paired-end reads resolved in 16,418 ASVs for V2-V3 regions and 6.977.815 (89.94%) paired-end reads resolved in 18,651 ASVs for V4 regions.

To measure the community diversity of V2-V3 and V4 regions we calculate Shannon and Simpson indices and test for significant differences using a paired modification of the Wilkinson-Mann-Whitney-nonparametric test (Table 3). The values for Shannon diversity indices in all samples for both regions ranged from 0.3776 to 5.947, while the Simpson index ranged from 0.0935 (V2-V3 region) to 0.9953561 (V4 region). Testing with the Wilkinson-Mann-Whitney index showed that the average of both indices differs significantly between the regions (Shannon p-value <

0.03, Simpson p-value < 0.03). Indicating that metabarcoding V4 16S rRNA region captures higher values for species diversity.

Table 3 Comparison of community diversity indices.

Community diversity Indicex	Mean value for V2-V3 region	Mean value for V4 region	P value from Wikinson-Mann-Whitney test
Shannon index	0.89	0.90	0.03
Simpson index	3.97	4.21	0.03
Chao1 index	305.64	392.66	7.92e-05
ACE index	303.44	392.77	5.73e-05

In terms of microbial community resolution, an analysis of the Chao1 and ACE indices shows that, in most cases, the hidden species richness of the V4 region is higher than in the V2-V3 regions. For the V4 region, the Chao1 index varied from 46 to 815.5 (mean value of 392.66) while for the V2-V3 regions it varied from 8 to 764 (mean value of 305.64). The Wilkinson-Mann-Whitney test confirmed that the V4 region for metabarcoding studies will lead to an overall greater resolution (Chao1 p-value <.7.925e-05, ACE p-value < 5.73e-05). More details of diversity indices distribution can be seen in Figure 3.

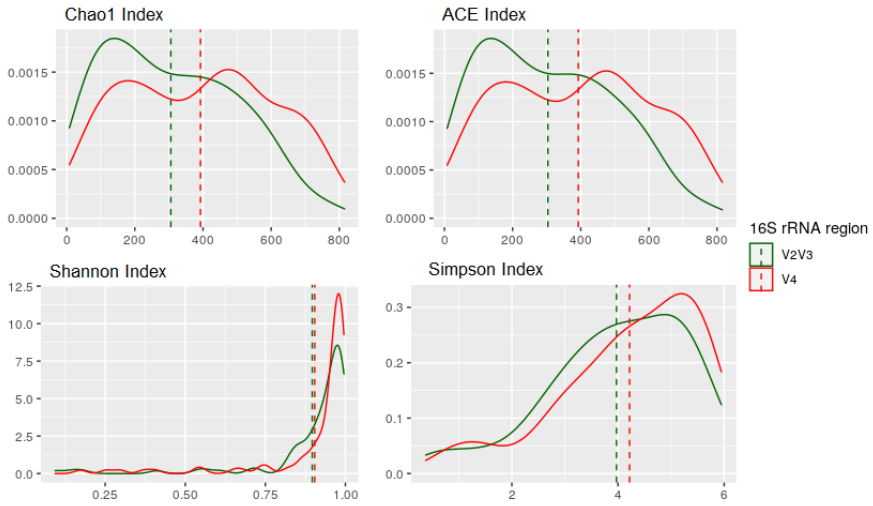

Figure 3. Density plot for diversity indices of the V2-V3 and V4 regions. The vertical line represents the mean value for each of the regions. Green color represents V2-V3 region. Red color represents V4 region.

To test if these results were biased by read counts, we have done a Spearman correlation test between read counts and diversity indices (Simpson and Shannon). Spearman correlation analysis shows that there is no correlation between diversity indices and read counts (Table 4). If the coverage of the community were insufficient, diversity indices would increase with the read count and there would be a significant positive ($r > 0$) correlation between these values.

Table 4. Correlation between diversity indices (Shannon and Simpson) and read counts.

Community diversity Indices	16S rRNA region	Correlation coefficient	P value from Spearman's rank correlation test
Shannon index	V2-V3	0.01	0.99
Shnanon index	V4	-0.01	0.94
Simpson index	V2-V3	0.02	0.84
Simpson index	V4	-0.06	0.55

To investigate the qualitative differences between the V2-V3 and V4 regions (considering ASVs that were classified at species-level), we assign taxonomy to each ASVs, using the "assignTaxonomy" function of the dada2 R package with DAIRYdb as reference database (min bootstrap confidence of 90).

A total of 7.474 ASVs were classified at species-level for V2-V3 regions and 10,168 ASVs for V4 regions. After merging ASVs with the same taxonomy at the species level, we obtained 1,127 species for the V2-V3 regions and 1.096 species for the V4 region. Among the species identified, 694 (45.38%) were common to both regions, 433 (28.32%) were specific to V2-V3 regions and 402 (26.30%) for V4 region (Figure 4).

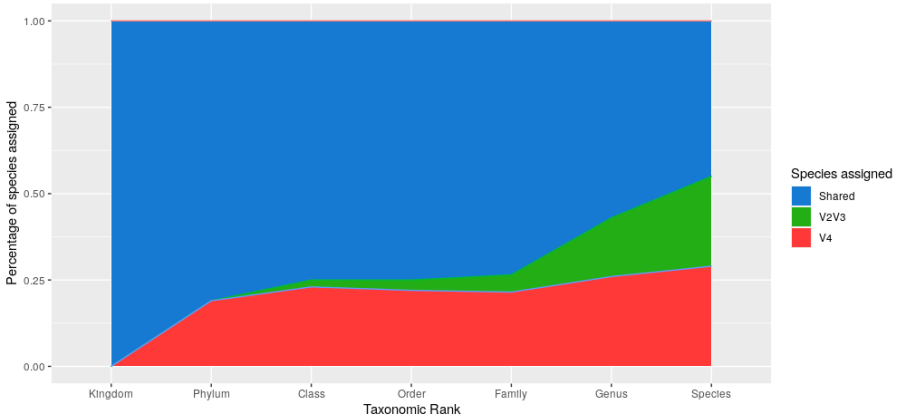

Figure 4. Taxonomic assignment spectrum for ASVs identified at species-level for regions V2-V3 and V4 regions. The X-axis represents the seven taxonomic ranks, going from Kingdom to Species. The Y-axis represents the comparison of the assignments between V2-V3 and V4 regions. Area colored by blue represents shared assignments betweens regions, while colored green represents assignments specific to V2-V3 regions and red for assignments specific to V4 region.

It's interesting to note that this difference is most prominent at higher taxonomic ranks (higher than genus), as can be seen in Figure 4. V4 primers are known to be the most "universal" primer, with an exception for only a few specific taxons (Mao et al., 2012), so, as expected, it has a greater capacity for coverage of more taxa present in the sample. However, when lowering the taxonomic ranks we see that this does not remain for genus and species rank. This may be explained by some reasons.

As discussed in the Bukin, 2019, we can consider the possibility that the dissimilarity in species spectra detected by different 16S rRNA fragments is related to primer specificity and PCR artifacts (Wang et al., 2009). Primers with lower specificity would disrupt the distribution uniformity and decrease the Shannon indices.

In some cases, it may happen that for some species, genes were not amplified at all. And this is what we find in our results. When we look at the results of higher taxonomic ranks for the V2-V3 and V4 regions we find 7 phyla that were not detected at all (e.g.*Chlamydiae*and*Gracilibacteria*) for the V2-V3 region. Moreover, the length of the regions may also affect the assignment at lower taxonomic ranks. The larger the region, the more likely there is a mutation that discriminates against the taxons.

Another interesting reason, also discussed in the Bukin, 2019, is the biological function of these fragments. The V2 region is responsible for maintaining the structural stability of the 16S rRNA gene, while the V4 region takes part in the translation process, responsible for binding tRNAs and interacting with the 23S rRNA gene (Van de Peer et al., 1999, Schluenzen et al., 2000, Morosyuk et al., 2001). So, it's expected that the V4 region will accumulate mutation at a slower pace compared to the V2 region.

Some studies compared diversity estimates between different regions of the 16S rRNA gene (Bukin et al., 2019, Fadeev et al., 2021). However, differences in the degenerated bases in primers sequence, sequencing platform, type of samples, and framework applied may impair direct comparison to our study.

Finally, to check if our initial hypothesis was correct, we compare the assignment for mastitis-causing pathogens between V2-V3 and V4 regions. We compare the species-level assignment for*Staphylococcus aureus*,*Streptococcus agalactiae*,*Streptococcus dysgalaciae, Streptococcus uberis, Escherichiacoli*,*Klebsiellaoxytoca*,*Klebsiella pneumonia, Serratiamarcescens*,*Corynebacterium bovis*and*Mycoplasma bovis.*

For V2-V3 we were able to detect *Staphylococcus aureus, Streptococcus dysgalactiae, Streptococcus agalactiae, Corynebacterium bovis, Streptococcus uberis, and Mycoplasma bovis,* while for the V4 region we were able to detect only *Streptococcus dysgalactiae* and *Mycoplasma bovis*. Details of abundance percentage for each species can be found in Figure 5.

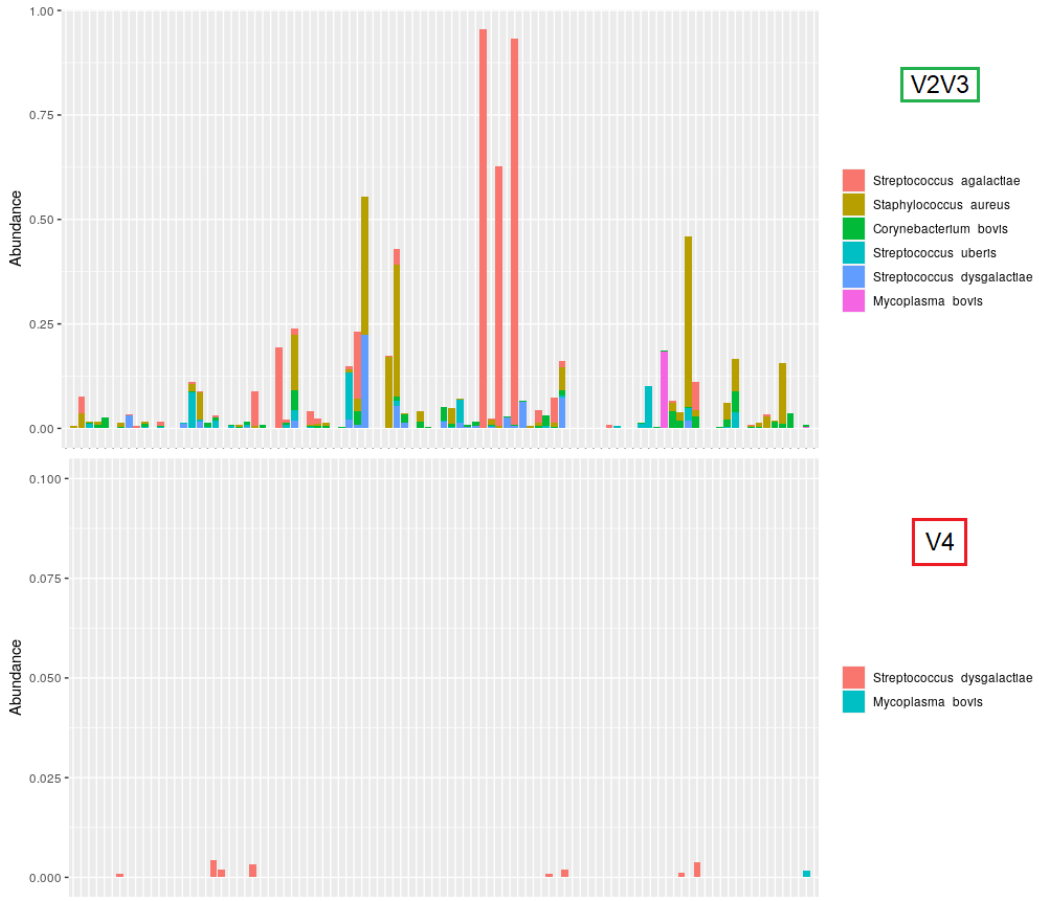

Figure 5. Abundance and prevalence of mastitis-causing pathogens for 96 raw milk samples sequencing for V2-V3 and V4 regions by Illumina sequencing. The Y-axis for V2-V3 regions ranged from 0.00 to 1.00. The Y-axis for V4 regions ranged from 0.00 to 0.10.

The sequencing of the V2-V3 region was not only able to detect a higher number of species (6 species) associated with bovine mastitis, but it also detected mastitis-causing pathogens in a larger number of samples (76 samples) compared to the V4 region (9 samples) with an important difference between the abundance estimates.

All these results show the potential of sequencing 16S rRNA V2-V3 fragment to be used as a tool for diagnostic of mastitis infection. Although it's not properly determined if the abundance estimates are accurate. It can increase the knowledge of the microbiome profile of dairy farms qualitatively. To the best of our knowledge, this is the first work that focuses on the target 16S rRNA region that maximizes the genetic diversity of mastitis-causing pathogens for raw bovine milk samples.

Some disadvantage for the V2-V3 region is the lack of primer specificity compared to a more "universal" primer, such as V4. However, this can be partially overcome by a more detailed study of the V2-V3 primer region for raw bovine milk samples. A possible approach is to evaluate targeting bases to be degenerated, increasing the coverage of the primer for prokaryote species. We also acknowledge that diagnosis of mastitis-causing pathogens it's not the only reason for microbiome profiling of raw bovine milk and more studies need to be done to account for a diversity of purposes.

2.4. Conclusions

The goal of this work was to find a 16S rRNA region that maximizes the detection of mastitis-causing pathogens on raw bovine milk samples. For that, we evaluate in silico the genetic diversity for V2-V3, V4, and V5-V6 regions, and find V2-V3 as a potential region. To validate our in silico results wth real data, we sequenced 96 raw bovine milk samples for the V2-V3 and V4 regions and compared the results. The result shows that although the V4 region was better to determinate the microbial diversity, it lacks taxonomic assignment at lower ranks, such as genus and species. Moreover, the V2-V3 region was able to detect in a large number of samples, a greater number of species correlated with bovine mastitis, with more abundance estimation. Here we demonstrated that targeting a region that maximizes genetic diversity according to the purpose of the study can have an important impact on the results.

References

Ali Alishum. (2019). DADA2 formatted 16SrRNA gene sequences for botch bactéria &archeaZenodo. doi:10.1128/aem.00062-07.

Amit Roy, S. R. (2014). Molecular Markers in Phylogenetic Studies - A Review.Journal of Phylogenetic & Revolutionary Biology. 2. doi: 10.4172/2329-9002.1000131.

Bauer, D. F. (1972). Constructing confidence sets using rank statistics. Journal of the American Statistical Association. 62. 687-690.

Bodenhofer, U., Bonatesta, E., Horejs-Kainrath, C., Hochreiter, S. (2015).msa: an R package for multiple sequence alignment. Bioinformatics. 31. doi: 10.1093/bioinformatics/btv494.

Bowen de Leon, K., Ramsay, B. D., Fields, M. W. (2012). Quality-score refinement of SSU rRNA gene pyrosequencing differs across gene region for environmental samples. Microb.Ecol. 64.doi: 10.1007/s00248-012-0043-9.

Bukin, Y. S., Galanchyants, Y. P., Moronov, I. V., Bukin, S. V., Zakharenko, A. S., Zemskaya, T. I. (2019). The effect of 16S rRNA region choice on bacterial community metabarcoding results. Sci. Data. 6. doi: 10.1038/sdata.2019.7.

Callahan, B. J., McMurdie., P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. Nat. Methods. 13. doi: 10.1038/nmeth.3869.

Cao. Y., Fanning. S., Proos, S., Jordan, K., Srikumar, S. (2017). A Review on the Applications of Next Generation Sequencing Technologies as Applied to Food-Related Microbiome Studies.Frontiers in Microbiology. 8. doi: 10.3389/fmicb.2017.01829.

Chiu, C. H., Wang. Y. T., Walther, B. A., Chao, A. An improved nonparametric lower bound of species richness via a modified good-turing frequency formula.Biometrics. 70. 617-682.

Cho, I., Blaser, M. J. (2012). The human microbiome: at the interface of health and disease. Nat Genet. 4. doi: 10.1038/nrg3182.

Christner, B., Skidmore, M., Priscu, J., Tranter, M., Foreman, C. (2008).Bacteria in subglacial envoriment.Psy.: from Biodiv. toBiotec. 51-71.

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., Brown, C. T., Porras-Alfaro, A., Kuske, C. R., Tiedje, J. M. (2013). Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acid Research.42. doi:10.1093/nar/gkt1244.

Fadeev, E., Cardozo-Mino, M. G., Rapp, J. Z., Beinhold, C., Salter, I., Salman-Carvalho, V., Molari, M., Tegetmeyer, H. E., Buttigieg, P. L., Boetius, A. (2021). Comparison of Two 16s rRNA Primers (V3-V4 and V4-V5) for Studies of Artic Microbial Communities. Frontiers in Microbiology. 12. doi: 10.3389/fcmicb.2021.637526.

Fukuda, K., Ogawa, M., Taniguchi, H., Saito, M. Molecular Approaches to Studying Microbial Communities: Targeting the 16S Ribosomal RNA Gene. Journal of UOEH. 36. doi: 10.7888/juoeh.38.223.

Hill, M. O. (1973). Diversity and evenness: a unifying notation and its consequences. Ecology. 54. 427-432.

Huse, S. M., Dethlefsen, L., Huber, J. A., Welch, D. M., Relman, D. A., Sogin, M. L. (2008). Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. PloS Genetic. 4. doi: 0.1371/journal.pgen.1000255.

Johnson, J. S., Spakowicz, D. J., Hong, B. Y., Petersen, L. M., Demkowicz, P., Chen, L., Leopold, S. R., Hanson, B. M., Agresta, H. O., Gerstein, M., Sodergren, E., Weinstock, G. M. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis.Nature Communications. 10. doi: 10.1038/s41467-019-13036-1.

Mao, D. P., Zhou, Q., Chen, C. Y., Quan, Z. X. (2012). Coverage evaluation of universal bacterial primers using the metagenomic datasets.BMC microbiology. 12. doi: 10.1186/1471-2180-12-66.

Mende, D. R., Sunagawa, S., Zeller, G., Bork, P. Accurate and universal delineation of prokaryotic species.Nat. Methods. 10. doi: 10.1038/nmeth.2575.

Meola, M., Rifa, E., Shani, N., Delbés, C., Berthoud, H., Chassard, C. (2019).DAIRYdb: a manually curated reference database for improved taxonomy annotation of 16S rRNA gene sequences from dairyproducts. BMC genomics. 20. doi: 10.1186/s12864-019-5914-8.

Morosyuk, S. V., Cunningham, P. R. SantaLucia, J. Jr. (2001).Structure and function of the conserved 690 harpin in Escherichia coli 16s ribosomal RNA.Journal of Molecular Biology. 307. 197-211.

O'Hara, R. B. (2005). Species richness estimators: how many species can dance on the head of a pin?.Journal of Animal Ecology. 74. 375-386.

Porter, T. M., Hajibabaei, M. (2018).Scalling up: A guide to high-throughput genomic approaches for biodiversity analysis. Mol. Ecol. 27. doi: 10.1111/mec.14478.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glockner, F. O. (2012).The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. NucleicAcids Res. 41. doi: 10.1093/nar/gks1219.

Ritari, J.,Salojarvi, J., Lahti, L., Vos, W. M. (2015). Improved taxonomic assignment of human intestinal 16S rRNA sequences by dedicated reference database. BMC Genomics. 16. doi: 10.1186/s12864-0152265y.

Schluenzen, F., Tocilj, A., Zarivach, R., Harms, J., Gluehmann, M., Janeli, D., Bashan, A., Bartels, H., Agmon, I., Franceschi, F., Yonath, A. (2000). Structure of Functionally Activated Small Ribosomal Subunit at 3.3 A Resolution. Cell. 102. doi: 10.1016/S0092-8674(00)00084-2.

Sunagawa, S. et al., (2015).Ocean plankton.Structure and function of the global ocean microbiome.Science. 348. doi: 10.1126/science.1261359.

Thompson, L. R. et al (2017). A communal catalogue reveals Earth's multiscale microbial diversity. Nature. 551. doi: 10.1038/nature24621.

Van de Peer, Y. et al. (1999). Database on the structure of small subunit ribosomal RNA.Nucleic Acid Research.27.179-183.

Wang, Q., Garrity, G. M., Tiedje, J. M., Cole, J. R. (2007). Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy.Applied and Env. Microb. 73. doi: 10.1128/aem.00062.07.

Werner, J. J., Koren, O., Hugenholtz, P., DeSantis, T. Z., Walters, W. A., Caproso, J. G., Angenent, L. T., Knight, R., Ley, R. E. (2011). Impact of training sets on classification of high-throughput bacterial 16S rRNA gene surveys.ISME Journal. 6. doi; 10.138/ismej.2011.82.

Wickham, H. (2016). Elegant Graphics for Data Analysis.https://ggplot2.tidyverse.org.

Yeluri, J. B. R., McSweeney, P. L. H., Cotter, P. D. (2018). Sequencing of the cheese microbiome and its relevance to industry.Frontiers in Microbiology. 9. doi: 10.3389/fmicb.2018.01020.

Zar, J. H. (1972). Significance testing of the Spearman rank correlation coefficient.Journal of the American Statistical Association. 67. 578-580.

3. MACHINE LEARNING APPLICATION FOR MASTITIS-CAUSING PATHOGEN DETECTION ON INDIVIDUAL SAMPLES OF RAW BOVINE MILK

Abstract

The correct identification of mastitis-causing pathogens is an important step in management of dairy farms. However, due to costs, techniques such as qPCR and 16S rRNA sequencing remain unfeasible for some herds. Machine learning methods have been shown as an attractive alternative. New studies focusing on the detection of clinical and subclinical mastitis highlight the potential of applied machine learning methods to the management of mastitis in dairy farms. Few studies of machine learning models were applied to detection mastitis-causing pathogen. In this work, we evaluate the performance of three machine learning methods to detect the most abundant mastitis-causing pathogen in individual raw milk bovine samples integrating data from milk composition and 16S rRNA sequencing. For abundance greater than 3% in individual samples, an accuracy of 100% for *Escherichia coli*and 86% for *Staphylococcus aureus*were achieved. These results show that not only subclinical and clinical mastitis can be detected by machine learning methods, but also some mastitis-causing pathogens

3.1. Introduction

Milk is a widely consumed food, both in its *in natura* form and its products. The milk contains water and nutritional compounds, such as vitamins, minerals, and macronutrients. The milk has about 0.9% of vitamins and minerals, such as calcium, phosphorus, sodium, potassium, B vitamins (B2, B6, and B12), vitamin A, and 12% of macronutrients, with 4% protein, 4.2% fat, and 4.6% carbohydrates (Mansoon, 2008). The content of the nutritional compounds can be affected by several factors such as diet, diseases, stage of lactation, age of the animal, and breed (Walstra, 1999).

In addition to water and nutritional compounds, microorganisms are commonly found in milk. Due to its composition rich in nutrients, water, and neutral pH (6.2 to 6.8), they obtain the vital media for their proliferation (Quigley et al. ., 2013b; Vithanageet al., 2016). Specific profiles of the microbiota of milk can pose dangers to human health if consumed raw. Moreover, some species directly impact the profitability of producers, technological processes of food products, and the quality of the final product directed to the consumer (Maréchal et al., 2011; Quigley et al., 2013b).

Raw milk can be a vehicle for certain microorganisms that cause diseases in humans, such as*Brucella spp.* (Brucellosis) (Galinska and Zagorski, 2013),*Listeria monocytogenes*(Listeriosis) (Swaminathan and Gerner-Smith, 2007),and*Mycobacterium tuberculosis*(Tuberculosis) (Russel, 2001). In addition, microorganisms of the genera*Bacillus*and*Clostridium*can form spores and become resistant to heat treatments, leading to a late deterioration of dairy products (Scheldeman et al., 2006). Moreover, psychrotrophicbacterias are capable of producing and releasing thermoresistant proteolytic and lipolytic enzymes, which compromise the quality of the dairy product even after heat treatments (Ribeiro Junior et al., 2017).

The disease in the mammary gland with greater importance for milk production is mastitis (Watts, 1988). Mastitis is considered to be the most costly disease in the dairy industry in the world, leading to economic losses of up to 26 billion dollars annually (www.dairy.ahdb.org.uk), mainly due to reduced milk production, reduced longevity of animals, changes in milk quality, labor costs, diagnosis and treatments (Oliveira et al., 2013).

Mastitis can be classified into two types; the clinic, when the animal presents physical symptoms, such as the presence of pus or lumps in the milk, redness, and swelling of the breasts; and the subclinical, where there are

alterations in the quality of the milk without the presence of physical symptoms in the animal. Data presented by Marcelo Busanello (2017) show that of the 517 herds studied in five Brazilian states, 46% of the animals had subclinical mastitis, with an incidence of 0.17 new cases per month.

Clinical mastitis can be detected by inspecting changes in the appearance of milk, local signs in the mammary gland, such as swelling, pain, redness, or signs in the animal, such as fever, apathy, anorexia, and dehydration. Subclinical mastitis, on the other hand, does not show obvious signs of infection and requires specific methods to detect. The most used technique for the determination of this type of mastitis is the evaluation of somatic cell count (SCC) (Sharma et al., 2011). Somatic cell counts are mainly formed by leukocytes (neutrophils, macrophages, lymphocytes, erythrocytes) and epithelial cells, serving as a useful predictor of intramammary infection (Sharma et al., 2011).

However, the somatic cell count can't identify the causative agent. Earlier detection of the pathogen involved can lead to a better choice of treatment method, antibiotic selection, and better management strategies to control the spread of disease in the case of a contagious organism (Ashraf and Imran, 2018). To detect microorganisms, methods based on culture medium and guide antibiotic treatment were developed. However, approximately 25% of samples from clinical mastitis are culture-negative or do not present significant pathogens (Bradley et al., 2007). Likewise, more than 30% of samples from cows or udders with high SCC were reported to be culture-negative (Bradley et al., 2007). In addition, traditional culture and microbiological identification by biochemical tests present some other limitations, such as analysis time, differences in reliability between tests from different laboratories, and a large number of erroneously identified mastitis bacteria, and the impossibility of identifying microorganisms at the strain level. (Paszynska-WesolowskaandBartoszcz, 2009; Gunasekea et al., 2009).

Another possibility of identification is through the use of the PCR technique. In the last decade, the amplification technique has evolved towards quantitative PCR (qPCR) (Masco et al., 2007; Malorny et al., 2008; Le Dréan et al., 2010) and the ISO 2012 and 2013 guidelines describe the use of qPCR for the detection of microorganisms in food. In this context, the molecular identification of microorganisms can replace the conventional characterization, based on clonal cultures, providing a more precise, sensitive and less laborious genomic definition, but with a high cost.

The progress of next-generation sequencing (NGS) technologies and the consequent reduction in sequencing costs has revolutionized human medicine and can add value to agribusiness contributing to the solution of several problems. These technologies allow the study of highly complex biological samples, enabling the taxonomic and functional characterization of microbial communities that practically colonize all ecological niches. However, the use of large-scale DNA sequencing to identify mastitis-causing pathogens is restricted to a few reports in the academic literature comparing healthy and infected udders(Porcellato et al., 2020). Thus, the use of sequencing data from the 16S rRNA hypervariable regions of the 16S gene for the characterization of the microbiota has been a great bet for the future development of a diagnostic tool for the disease in its clinical and subclinical phases. Compared to the multiplex qPCR molecular methodology, the sequencing methodology is not limited to the number of previously selected pathogens, but it is also expensive.

The standard control of the bovine herd occurs by the evaluation of the somatic cell count in the milk tank. In Brazil, this practice is mediated by Normative Instructions 76 and 77, of the Ministry of Agriculture and Supply (MAPA), which determine the maximum limits for commercialization of refrigerated raw milk in the dairy industry. However, the ideal would be to perform individual weekly control of SCC, as well as the identification of

microorganisms that cause mastitis in all animals with mastitis and with high SCC counts. Unfortunately, this practice comes at an unfeasible cost for most farms.

Some advances have been made in low-cost alternatives for the detection of clinical and subclinical mastitis using machine learning methods. As stated by Bobbo 2021, different studies have applied machine learning techniques to diagnose mastitis. Some studies relied on the presence of high milk SCC (Ebrahimie et al., 2018) or mastitis pathogens(Esener et al., 2018), while othershave established SCC-independent models for mastitis prediction using alternative milking traits (e.g., milk volume, fat, protein, lactose) (Ebrahimi et al, 2019). However, when the mastitis-causing pathogen detectionwas developed, it was only for a few species or genera (e.g. strains of *Streptococcus uberis* in Esener et al., 2018), and a broader detection has not yet been developed.

Due to this, the present work aims to integrate tankmilk composition traits to 16S rRNA sequencing data to be able to predict the most abundant mastitis-causing pathogen present in individual sample milk.For that, we applied 3 machine learning methods to 442 individual raw bovine milk samples that had been previously sequenced for 16S rRNA V4 regions, and evaluate the accuracy ofthe prediction of *Staphylococcus aureus*, *Escherichia coli*, *Streptococcus agalactiae,* and *Streptococcus dysgalactiae,* based on fat, protein, lactose, total solids, defatted dry extract, SCC and total bacterial countmeasured for tank samples obtained in the same week.

3.2. Materials and Methods

3.2.1. Samples

We select 442individual milk samples and 9 (one for each month) tank milk samples obtained from 2 herds located in São Paulo State, Brazil. (282 individual samples from herd A for June, July, October, November of 2019, and 160 samples from herd B for May, July, August, September of 2019, and January of 2020). Samples were collected and stored in tubes with Bronopol® conservative with a concentration between 0,004% - 0,005%. The tubes were stored refrigerated for five days maximum until DNA extraction and library generation.

3.2.2. Somatic cell count, total bacterial count and composition analysis

The analyzes of somatic cell count (CCS) total bacterial count (CBT), fat, protein, lactose, total solids (TS) and defatted dry extract (DDE) were performed by flow cytometry methodology according to ISO 13366-2:2006/ IDF 148-2:2006 / ISO 16297: 2013/ IDF 161: 2013 by Clínica do Leite.

The somatic cell count and the total bacterial count was measured for all individual samples while fat, protein, lactose, total solids and deffated dry extract was measured for the tank milk samples collected on the same week of individual samples.

3.2.3. DNA extraction, library preparation and sequencing

For storage, 2 ml aliquots were made in 2 ml eppendorf tubes and subjected to centrifugation for 5 minutes at 14,000 rpm. After centrifugation, fat and supernatant were discarded and the pellet stored at -20º C until DNA

extraction. DNA extraction was performed according to the protocol of the MagMax CORE Nucleic Acid Purification kit together with the mechanical lysis module (both from Applied Biosystems, Foster City, CA, USA), in the KingFisher equipment (ThermoFisher). The quality and quantity of the extracted DNA was evaluated by 1% agarose gel. The extracted DNA was stored in a freezer until use.

The extracted DNA was subjected to library construction as suggested by the Illumina protocol for 16S libraries. The chosen region of the 16S gene for amplification was accordingly to works published by Quigley et al (2013a) and Bonsaglia et al (2017). In Quigley's and Bonsaglia's works they used the V4 region to identify the microbial profile of bovine milk.

The Illumina protocol is based on two PCR reactions. The first with primers 515B/806B, which are universal locus-specific primers for the V4 hypervariable region of bacteria. The second PCR consists of the ligation of Illumina adapters that allow the multiplexing of the samples and the hybridization of the sequences in the sequencing slide. After being purified, the libraries were pooled with equimolar concentrations. The pool was quantified in qPCR and proceded to the denaturation and sequencing steps in MiSeq System equipment (Illumina, San Diego, CA, USA).

3.2.4. Analysis of 16S sequencing data

The DADA2 program, an open package implemented in the R language (Callahan et al., 2016), was used for modeling and error correction of amplicons, with the construction of ASVs (Callahan et al., 2016). Filtering of fastq files was performed to remove sequences from the primers and filter the ends due to quality decay (Q<30). After filtering, the forward and reverse reads were joined to reassemble the complete fragment.

The DADA2 algorithm makes use of a parametric error model, incorporating the different error rates in each amplicon dataset. The "learnErrors" method adjusts the error model from the data, alternating between estimating error rates and inferring sample composition until convergence. Through dereplication, the list of unique sequences was obtained, with relative abundances, and then the chimeras were removed.Taxonomies were assigned to each ASVs (Amplicon Sequencing Variants) using a DADA2 implementation of the Naive Bayesian classifier method(Wang et al., 2007)wirhDAIRYdb reference database (Alishum, 2019) (min bootstrap confidence of 90).

Afterwards, the data generated by the DADA2 program were imported into the phyloseq program (Murdie& Holmes, 2013), also implemented in R, for tax agglomeration and abundance estimation at species-level taxomical rank.

3.2.5. Machine learning methods

In machine learning, the term "learning" refers to running a computer program to induce a model using training data. Machine learning techniques use statistical theory in building computational models to make inferences from a sample. The learning process consists of several steps. In the first step, we integrate and merge different sources of information into a single format (*e.g,* data from an experiment and respective metadata). In the second step, it is necessary to select, clean, and transform the data. To perform this step, we need to eliminate or correct the data, as well, as decide the strategy to impute the missing data (if present). In this step, we can also select the relevant variables. In the third step, we take into account the objectives of the study to choose the most appropriate analysis

for the data (regression, for quantitative prediction, or classification for qualitative prediction). Once the model is obtained, it must be evaluated and, if necessary, return to the previous steps for a new iteration. The chosenmodel is then used to solve the problem (Larranaga, P. et. al., 2006).

Topredictmastitis-causing pathogens we used the caret v.6.0 package (Kuhn, M. et al., 2020), by screening threemachine learning techniques of classification task (Conditional Random Forests, Naïve Bayes, andBoosted Logistic Regression). The goal of the classification task was to determine the most abundant mastitis-causing pathongen present in the individual raw milk sample. The effectiveness of the prediction was evaluated through the overall accuracy and accuracy for each pathogen.

To construct the input dataset we combine the relative abundance for *Staphylococcus aureus*, *Escherichia coli*, *Streptococcus agalactiae,* and *Streptococcus dysgalactiae*, with values from milk traits composition into a "data.frame". Each sample has its unique somatic cell count and total bacterial count and shared the fat, protein, lactose, total solids and defatted dry extract values from its tank's sample. Following, we determine the most abundant specie for each sample (from pre-selected species) and used this information as a predicted variable in the machine learning methods. Samples with zero relative abundance for all selected species were labeled as "None". To differentiate the ability of the model to detect low relative abundance and high relative abundance, we created a second predicted variable, considering only relative abundances higher than 3%. Samples that did not meet these criteria were labeled as "None".

To start the machine learning analyzes, we divided the total sample set into two subsets, the training subset and the validation subset, in ratio of 1:4. The learner's efficiency (parameter tuning) was optimized through cross-validation. Once the model was optimized we validated and calculate the accuracies for the validation subset.

3.3. Results and discussion

The sequencing of 16S rRNA fragments from the 451 raw milk samples yielded a total of 25,497,249 paired-end reads (mean of 57,686 paired-end reads with sd (Standard deviation) of 25,510). After quality control, denoising, and exclusion of chimeras we retained a total of 19,829,789 (77.77%) paired-end reads resolved in 50,660 ASVs.To assign taxonomic to each ASV we used the Naïve Bayes Classifier method implemented in the DADA2 R package with DAIRYdb as a reference dataset (min bootstrap confidence of 90). After taxon agglomeration and relative abundance estimates, we retrievea total of 1,762 species.

Following the construction of input data (milk trait composition and relative abundance for selected species),we trained the models on 356 samples and tested them on 86 samples to identify the most prevalent mastitis-causing pathogen using milk traits composition values.Theperformances of each machine learning method are detailed in Table 1.

Table 1. Performance metrics on external validation set.

Algorithm	Variables	Abundance+	Kappa	E.coli[2]	S.aureus[3]	S. dysgalactiae*	S.agalatiae**	All[1]
Naïve Bayes	SCCTBC	0%	0.08	0.07	0.94	0.00	0.23	0.44
		3%	0.04	0.19	0.00	0.00	0.00	0.51
	Full	0%	0.21	0.71	0.51	0.00	0.15	0.48
		3%	0.71	1.00	0.87	0.00	0.14	0.81
Conditioal Random Forest	SCCTBC	0%	0.09	0.32	0.77	0.00	0.08	0.42
		3%	0.12	0.19	0.06	0.00	0.00	0.55
	Full	0%	0.29	0.57	0.86	0.00	0.15	0.55
		3%	0.77	1.00	0.86	0.00	0.00	0.85
Logic Boost	SCCTBC	0%	0.01	0.14	0.26	0.00	0.08	0.20
		3%	0.17	0.44	0.06	0.00	0.00	0.54
	Full	0%	0.22	0.43	0.40	0.00	0.08	0.38
		3%	0.72	1.00	0.81	0.00	0.00	0.83

SCC/TBC – Models construct considering only somatic cell counts and total bacterial counts
Full – Models construct considering somatic cell counts, total bacterial counts, fat, protein, lactose, total solids and defatted dry extract.
0% - Labels representing the most abundant mastitis-causing pathogens with relative abundance greather than 0 percent.
3% - Labels representing the most abundant mastitis-causing pathogens with relative abundance greather than 3 percent.
[1]Overall accuracy
[2] Accuracy for classification of *Escherichia coli*
[3] Accuracy for classification of *Staphylococcus aureus*
* Accuracy for classification of *Streptococcus dysgalactiae*
** Accuracy for classification of *Streptococcus agalactiae*

The best configuration for mastitis-causing pathogen was achieved using Conditional Random Forest models for full milk trait composition (kappa value of 0.77 and overall accuracy of 0.85) with abundance greater than 3%. In all scenarios, it wasn't possible to detect *Streptococcus dysgalactiae* and *Streptococcus agalactiae*. This may be due to an underrepresentation of this species 0inthe training and validation set. In training set, we had 30 samples for *Streptococcus agalactiae* and 6 samples for *Streptococcus dysglactiae*, while in the validation set we had 7 samples for *Streptococcus agalactiae* and 1 sample for *Streptococcus dysgalactiae*.

Furthermore, we obtained a good classification accuracy of *Escherichia coli*(100%) and *Staphylococcus aureus*(86%). In Sharifi 2018 they were able to detect with 83% accuracy (on average), *Escherichia coli* induced mastitis using gene expression data of bio-signature genes (e.g ZC3H12A, CXCL2).Whileinteresting in highlighting genes related to immune response and inflammation, transcriptomic profile it's too expensive to be used as a tool for mastitis diagnosis in dairy farms.

Although the models were not able to detect all selected species, these results are relevant.*Staphylococcus aureus* mastitis is known to have a lower cure rate for treatments with antibiotics (Barkema et al, 2006) and cause infections that can persist through lactation with antibiotic pressure (Brouilette et al, 2004). Moreover, *Staphylococcus aureus*has the ability to invade epithelial cells (Bardiau et al, 2014) and form small colony variants (SCVs) in dairy cows with chronic intramammary*S. aureus* infection history (Attala et al, 2008), leading to a more challenging control, compared to *Streptococcus agalactiae*, that has a cure rate above 90% with antibiotic treatments (Barkema et al, 2006).

As discussed in Barkema 2006, for successful implementation of mastitis control program, it is important to identify *Staphylococcus aureus*-infected cows to isolate the animal and to minimize the opportunity for spread of the pathogen in the herd. Some antibiotic can be applied, such as penicillin; however the cure levels will depende on host-levels factors, patrhogens factors and strains with antibiotic resistance (Barkema et al, 2006). To control *Escherichia-coli*-infected cows some scientific evidences for fluoroquinolones and cephalosporins are available and antimicrobial resistance is generally not a limiting factor for the success of these treatments. (Suojala et al, 2013).

Our results show the potential of using compositional milk traits for detecting mastitis-causing pathogens with machine learning methods. However, some limitation needs to be pointed out. Our study was conducted based on two herds located in São Paulo, Brazil with a dataset of 442 individual raw milk samples. The dataset size was considerably shorter compared to other studies (> 18,000 samples for Bobbo et al., 2021) with fewerfeatures (> 250 features for Hyde et al., 2020) and the compositional values (e.g. fat and protein) was measured for the tank and not individual samples. In addition, the 16S rRNA sequenced region may not be the most appropriateforthe detection of mastitis-causing pathogens, as described in Chapter 01 (In silico genetic diversity evaluation of 16S rRNA regions for pre-selected species can improve taxonomic assignement). All of these limitations may impact the accuracy and the generalization of the models for other herds. So, more studies need to be conducted with larger sample size, a greater number of herds,individual compositional values,and a different 16S rRNA target region.

3.4. Conclusion

The goal of this work was to evaluate the performance of three machine learning methods for mastitis-causing pathogen detection. We select the species *Staphylococcus aureus*, *Streptococcus agalactiae*, *Streptococcus dysgalactiae*, and *Escherichia cole* to be detected. For relative abundance greater than 3% we were able to detect *Staphylococcus aureus* with 86% of accuracy and *Escherichia coli* with 100% of accuracy. These results can assist dairy farms, since *Staphylococcus aureus* is one of the most complicated species associated with bovine mastitis. It doesn't respond well to antibiotic treatments, and has a long-lasting infection in herds. We hope that our work highlight the potential of using machine learning methods for detection of mastitis-causing pathogens and more studies are carried out to construct a more robust and generalized model.

References

Ashra, A., Imran, M. (2018). Diagnosis of bovine mastitis: from laboratory to farm. Tropical Animal Health and Production. 50. doi: 10.1007/s11250-018-1629-0.

Atalla, H., Gyles, C., Mallard, B. (2011). Staphylococcus aureus small colony variants (SCVs) and their role in disease.Anim Health Res Rev. 12.doi: 10.1017/S1466252311000065.

Bardiau, M., Detileux, J., Farnir, F., Mainil, J. G., Ote, I. (2014). Associations between properties linked with persistence in a collection of Staphylococcus aureus isolates from bovine mastitis. Vet. Microbiol. 169. doi: 10.1016/j.vetmic.2013.12.010.

Barkema, H. W., Schukken, Y. H., Zadoks, R. N. (2006). The role of cow, pathogen, and treatment regimen in the therapeutic success of bovine Staphylococcus aureus mastitis.Journal of Dairy Science. 6. doi: 10.3168/jds.S0022-0302(06)72256-1.

Bobbo, T., Bifanni, S., Taccioli, C., Penasa, M., Cassandro, M. (2021).Comparison of machine learning methods to predict udder health status based on somatic cell counts in dairy cows. Scientific Reports. 11. doi: 10.1038/s41598-93056-4.

Bonsaglia, E. C. R., et al. (2017). Milk microbiome and bacterial load following dry cow therapy without antibiotics in dairy cows with healthy mammary gland.Scientific Reports. 7. doi: 10.1038/s41598-017-08790-5.

Bradley, A. J., Leach, K. A., Breen, J. E., Green, L. E., Green, M. J. (2007). Survey on the incidence and aetiology of mastitis on dairy farms in England and Wales. Vet. Rec. 160.doi: 10.1136/vr.160.8.253.

Brouillette, E., Martinez, A., Boyll, B. J., Allen, N. E., Malouin, F. (2004).Persistence of a Staphylococcus aureus small-colony variant under antibiotic pressure in vivo.FEMS Immunol Med Microbiol. 41. doi: 10.1016/j.femsim.2003.12.007.

Busanello, M., et al. (2017).Estimation of prevalence and incidence of subclinical mastitis in a large population of Brazilian dairy herds.Journal of Dairy Science. 100. doi: 10.3168/jds.2016-12042.

Callahan, B. J., et al. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. 13. doi: 10.1038/nmeth.3869.

Ebrahime, E., Ebrahime, F., Ebrahimi, M., Tomlinson, S., Petroviski, K. R. (2018). Hierarchical pattern recognition in milking parameters predicts mastitis prevalence. Comput.Electron.Agric. 147.doi: 10.1016/j.compag.2018.02.003.

Ebrahime, M., Mohammadi-Dehcheshmeh, M., Ebrahime, E., Petrovski, K. R. Comprehensive analysis of machine learning models for prediction of sub-clinical mastitis: Deep Learning and Gradient-Boosted Tress outperform other models. Comput. Biol. Med. 114. doi: 10.1016/j.compbiomed.2019.103456.

Ensener, N. et al. (2018). Discrimination of contagious and environmental strains of Streptococcus uberis in dairy herds by means of mass spectrometry and machine-learning.Scientific Reports. 8. doi: 10.1038/s41598-018-35867-6.

Galinska, E. M., Zagorski, J. (2013). Brucellosis in humans – etiology, diagnostics, clinical forms.Annals of Agricultural and Environmental Medicine.20 (2).233-238.

Gunasekera, T. S., et al. (2002). Inducible Gene Expression by Nonculturable Bacteria in Milk after Pasteurization.Appl Environ Microbiol. 68. doi: 10.1128/AEM.68.4.1988-1993.2002.

Hyde, R. M. et al. (2020). Automated prediction of mastitis infection patterns in dairy herds using achine learning. Scientific Reports. 10. doi: 10.1038/s41598-020-61126-8.

Kuhn, M. (2008).Building Predictive Models in R using the caret package.Journal of Statistical Software. 25. doi: 10.18637/jss.v028.i05.

Larranaga, P., et al. (2006). Machine learning in Bioinformatics.Briefings in Bioinformatics. 7. 86-112.

Le Dréan, G., Mounier, J., Vasseur, V., Arzur, D., Habrylo, O., Barbier, G. (2009). Quantification of Penicillumcamemberti and P. roqueforti mycelium by real-time PCR to assess their growth dynamics during ripening cheese.Int J Food Microbiol. 100. doi: 10.1016/j.ijfoodmicro.2009.12.013.

Malorny, B., Lofstrom, C., Wagner, M., Kramer, N., Hoorfar, J. (2007).Enumeration of salmonella bacteria in food and feed samples by real-time PCR for quantitative microbial risk assessment.Appl Environ Microbiol. 74. doi: 10.1128/AEM.02489-07.

Manson, H. L. (2008). Fatty acids in bovine milk fat.Food.Nutr.Res. 52.doi: 10.3402/fnr.v52i0.18.21

Maréchal, C. L., et al. (2011). Molecular Basis of Vilurence in Staphylococcus aureus Mastitis. Plos One. 6. doi: 10.1371/journal.pone.0027354.

Masco, L., Vanhoutte, T., Temmerman, R., Swings, R., Huys, G. (2006). Evaluation of real-time PCR targeting 16S rRNA and recA genes for the enumeration of bifidobacteria in probiotic products.Int J Food Microbiol. 113. doi: 10.1016/j.ijfoodmicro.2006.07.021.

McMurdie, P. J., Holmes, S. (2013). phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. Plos One.doi: 10.1371/journal.pone.0061217.

Oliveira, L., Hulland, C., Ruegg, P. L. (2013). Characterization of clinical mastitis occurring in cows on 50 large dairy herds in Wisconsin.Journal of Dairy Science. 96. doi: 10.3168/jds.2012-6078.

Paszynska-Wesolowska, I., Bartoszcz, M. (2009).Bacteria in the state of VBNC - A threat to human health.MedycynaWeterynaryjna. 65. 228-231.

Porcellato, D., Meisal, R., Bombelli, A., Narvhus, J. A. (2020). A core microbiota dominates a rich microbial diversity in the bovine udder and may indicate presence of dysbiosis. Scientific Reports. 10. doi: 10.1038/s41598-020-77054-6.

Quigley, L., et al. (2013)a. The microbial content of raw and pasteurized cow milk as determined by molecular approaches.Journal of Dairy Science. 96. doi: 10.3168/jds.2013-6688.

Quigley, L., et al. (2013)b. The complex microbiota of raw milk.FEMS MicrobiologyReviews. 37. doi? 10.1111/1574-6976.12030.

Ribeiro Júnio, J. C., de Oliveira, A. M. Silva, F de G., Tamanini, R., de Oliveira, A. L. M., Beloti, V. (2018). The main spoilage-related psychrotrophicbactéria in refrigerated raw milk.Journal of Dairy Science. 101. doi: 10.3168/jds.2017-13069.

Russel, D. G. (2001). Mycobacterium tuberculosis: here today, and here tomorrow. Nat Rev Mol Cell. 2. doi: 10.1038/35085034.

Scheldeman, P., Herman, L., Foster, S., Heyndrickx, M. (2006). Bacillus sporothermodurans and other highly heat-resistance spore formers in milk. J Applied Microbiol. 101 (3).doi: 10.1111/j.1365-2672.2006.02964.x.

Sharma, N., Singh, N. K., Bradwal, M. S. (2011). Relationship of Somatic Cell Count and Mastitis: An Overview. Asian-Australian Journal of Animal Science. 24. doi: 10.5713/ajas.2011.10233.

Suojala, L., Kaartinen, L., Pyorala, S. (2013). Treatment for bovine Escherichia coli mastitis – an evidence-based approach.J Vet Pharm Therap. 36.doi: 10.1111/jvp.12057.

Swaminathan, B., Gerner-Smidt, P. (2007).The epidemiology of human listeriosis.Microbes and Infection.9 (10).doi: 10.1016/j.micinf.2007.05.011.

Vithanage, N. R., et al. (2016). Biodiversity of culturablepsychrotrophic microbiota in raw milk attributable to refrigeration conditions, seasonality and their spoilage potential. International Dairy Journal. 57. doi: 10.1016/j.dairyj.2016.02.042.

Wang, Q., Garrity, G. M., Tiedje, J. M., Cole, J. R. (2007). Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Applied and Env. Microb. 73. doi: 10.1128/aem.00062.07.

Wastra, P. (1999). Dairy technology: principles in milk properties and processes. CRC Press.

Watts, J. L. (1988). Etiological agents of bovine mastitis.Veterinary Microbiology. 16. doi: 10.1016/0378-1135(88)90126-5.

4. CONCLUSIONS

The goal of this work was to explore mastitis-causing pathogen detection on raw milk bovine individual samples. For that, first we find a 16S rRNA region that maximizes the detection of mastitis-causing pathogens on raw bovine milk samples and validate these results by sequencing. Second, weevaluate the performance of three machine learning methods for mastitis-causing pathogen detection.

We find that although the V4 region was better to determinate the microbial diversity, it lacks taxonomic assignment at lower ranks, such as genus and species. Moreover, the V2-V3 region was able to detect in a large number of samples, a greater number of species correlated with bovine mastitis, with more abundance estimation. Moreover, using Conditional Random Forest for relative abundance greater than 3% we were able to detect *Staphylococcus aureus* with 86% of accuracy and *Escherichia coli* with 100% of accuracy. We hope that our work highlight the potential of using machine learning methods for detection of mastitis-causing pathogens and more studies are carried out to construct a more robust and generalized model.