UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Dept. of Biotechnology

Optimization of the methodology for processing intestinal biopsies in microbiota studies.

Master's Thesis

Master's Degree in Biomedical Biotechnology

AUTHOR: Sánchez Rumí, Manuel José

Tutor: Giraldo Reboloso, Esther

External cotutor: Lloréns Rico, Verónica

ACADEMIC YEAR: 2023/2024

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to all the people who, in one way or another, have contributed to the completion of this Master's Thesis.

First and foremost, I would like to thank the Polytechnic University of Valencia for providing me with the opportunity to learn and develop my skills in its academic environment, as well as for allowing the use of its facilities to carry out this work.

I would also like to express my gratitude to the Prince Felipe Research Center for allowing me to use its facilities to develop this project and for warmly welcoming me as an intern. I am especially grateful to Dr. Verónica Lloréns, who leads the research group of which I had the honor to be a part. Her advice and knowledge have been invaluable throughout this time, and I am deeply thankful for her guidance.

I would like to extend my thanks to the other members of Group I-79, who were always willing to help whenever I needed it. Sharing this experience with you has been one of the most enriching and wonderful stages of my life.

My gratitude also extends to Esther Giraldo, who, with great dedication, accepted to be my internal tutor. I greatly appreciate all her revision work and the valuable advice she has provided.

Finally, I wish to express my deepest thanks to my family and friends. They have unconditionally supported me in all the decisions I have made during this time. Without their support, it would not have been possible for me to reach this point.

Lastly, I would like to dedicate this work to my grandmother, whose memory and example have inspired me throughout this journey.

For all of this, thank you very much.

TABLE OF CONTENTS

ABSTRACT……………………………………………………………………..…………I
RESUMEN……………………………………………………………………………………II

1 INTRODUCTION 2

1.1 DEFINITION OF MICROBIOTA AND GUT MICROBIOTA. FUNCTIONS AND RELATED DISEASES 2
1.2 ANALYSIS METHODS AND STUDY MODELS OF THE MICROBIOTA 3
1.2.1 Analysis methods *3*
1.2.1.1 16S rRNA amplicon sequencing 4
1.2.1.2 Metagenomics 5
1.2.1.3 Metatranscriptomics 7
1.2.1.4 Other Analysis. Metabolomics and metaproteomics 10
1.2.2 Study models *10*
1.2.2.1 *In vitro* models 10
1.2.2.2 Animal Models 11
1.2.2.3 Human Studies 12
1.3 CHALLENGES IN GUT MICROBIOTA FUNCTIONAL STUDIES 13
1.4 OBJECTIVES 14

2 MATERIALS & METHODS **15**

2.1 SAMPLE COLLECTION AND STORAGE 15
2.2 DNA/RNA EXTRACTION 16
2.2.1 Common sample lysis procedure *17*
2.2.2 Qiagen DNA/RNA Extraction *18*
2.2.3 Promega DNA/RNA Extraction *18*
2.2.4 NZYTech DNA/RNA Extraction *19*
2.2.5 Macherey-Nagel DNA/RNA Extraction *20*
2.3 EVALUATION OF THE CONCENTRATION AND QUALITY OF DNA AND RNA 20
2.4 REVERSE TRANSCRIPTION OF RNA TO CDNA 21
2.5 QPCR 21
2.6 16S RRNA AMPLICON SEQUENCING 22
2.7 DATA ANALYSIS 23

3. RESULTS **24**

3.1 TOTAL NUMBER OF SAMPLES ANALYZED 24
3.2 COMPARISON OF DIFFERENT SAMPLE TYPES 25
3.2.1 Comparison Between Mucosal Scrapings and Intestinal Biopsies *26*
3.3 COMPARISON OF DIFFERENT EXTRACTION KITS 27
3.4 COMPARISON OF DIFFERENT RNA PURIFICATION METHODS 28
3.5 ANALYSIS OF THE 16S SEQUENCING RESULTS 32
3.6 IDENTIFICATION OF ENVIRONMENTAL CONTAMINANTS 37
3.7 MICROBIOME COMPOSITION IN MOUSE COLONIC SAMPLES 42

4.DISCUSSION **45**

5. CONCLUSIONS **48**

6. BIBLIOGRAPHY **49**

SUPPLEMENTARY MATERIAL

Abstract

The gut microbiota, composed of microorganisms residing in the gastrointestinal tract, plays a fundamental role in the health and functioning of the organism. The variation in its composition and functionality is influenced by factors such as diet, lifestyle, and genetics; and requires comprehensive analysis to understand and treat metabolic disorders and other microbiota-associated diseases. Microbiota analyses are mostly conducted through 16S rRNA gene amplicon sequencing in fecal samples. However, microbial composition alone often proves insufficient to fully understand a condition, necessitating a deeper focus on microbial functions. To do so, metatranscriptomics emerges as a tool that allows profiling the expression of the different functions of the gut microbiota. However, due to the short half-life of bacterial RNA, fecal samples frequently used in 16S studies are not representative for studying active functions in specific colon regions. As an alternative to fecal samples, colonic samples may be used for these purposes. Despite their suitability, biopsies pose additional challenges, such as limited biomass and an elevated risk of contamination. Moreover, the lack of standardized protocols for biopsy processing and conducting metatranscriptomics adds further complexity to the field.

In this study, comparative analyses of different DNA and RNA extraction protocols from mouse colon biopsies were conducted using four different commercial kits. Variations in RNA purification treatments and the type of sample used were also evaluated. Concentration and quality of extracted DNA and RNA were measured, RNA was reverse transcribed to cDNA, and qPCRs were performed to quantify the genetic material of bacteria present in each sample. The data were analyzed and visualized using different R packages. Additionally, analyses were conducted through 16S rRNA gene sequencing to identify bacterial communities as well as potential contaminants that might be present in the various kits.

These analyses aim to determine the most suitable protocol for processing this type of samples, both for 16S rRNA gene sequencing and metatranscriptomics, with significant implications for understanding and treating diseases associated with the microbiota.

Keywords: intestinal microbiota, 16S gene amplicon sequencing, fecal samples, metatranscriptomics, colon biopsies, limited biomass, DNA and RNA extraction, RNA purification treatments, cDNA, R.

Resumen

La microbiota intestinal, compuesta por microorganismos que residen en el tracto gastrointestinal, desempeña un papel fundamental en la salud y el funcionamiento del organismo. La variabilidad en su composición y funcionalidad, influenciada por factores como la dieta, el estilo de vida y la genética, requiere un análisis exhaustivo para comprender y abordar trastornos metabólicos y enfermedades asociadas. Los análisis de microbiota intestinal se realizan mayormente mediante secuenciación de amplicones del gen 16S en muestras fecales. Sin embargo, la composición microbiana por sí sola resulta insuficiente para entender plenamente una condición, lo que requiere un enfoque más profundo en las funciones microbianas. Para ello, la metatranscriptómica emerge como una herramienta que permite estudiar la expresión de las distintas funciones de la microbiota. No obstante, debido a la breve vida media del RNA bacteriano, las muestras fecales usadas frecuentemente en estudios de microbiota intestinal son poco representativas para estudiar las funciones activas en regiones concretas del colon, como por ejemplo tumores. Esto conduce a la elección de biopsias de colon como alternativa. A pesar de su idoneidad, las biopsias plantean desafíos adicionales, como la limitada biomasa y el riesgo elevado de contaminación. Además, la falta de protocolos estandarizados para el procesamiento de biopsias y la realización de metatranscriptómica conlleva una complejidad adicional.

En este estudio, se realizaron análisis comparativos de distintos protocolos de extracción de DNA y RNA a partir de biopsias de colon de ratón, empleando cuatro kits comerciales diferentes. Se evaluaron también variaciones en los tratamientos de purificación del RNA y en el tipo de muestra utilizada. Se midió la concentración y calidad del DNA y RNA extraídos, se realizó la retrotranscripción del RNA a cDNA y se llevaron a cabo qPCRs para cuantificar el material genético de las bacterias presentes en cada muestra. Los datos se analizaron y visualizaron usando diferentes paquetes de R. Además, se efectuaron análisis mediante secuenciación del gen 16S para identificar comunidades bacterianas así como posibles contaminantes que pudieran estar presentes en los diversos kits.

Estos análisis buscan determinar el protocolo más idóneo para el procesamiento de este tipo de muestras, tanto para 16S como para metatranscriptómica, con implicaciones significativas para la comprensión y el tratamiento de enfermedades asociadas a la microbiota.

Palabras clave: microbiota intestinal, secuenciación de amplicones 16S, muestras fecales, metatranscriptómica, biopsias de colon, limitada biomasa, extracción de DNA y RNA, tratamientos de purificación de RNA, cDNA, R.

1 Introduction

1.1 Definition of microbiota and gut microbiota. Functions and Related Diseases

Within the framework of human biology study, the microbiota emerges as a highly diverse and dynamic community of microorganisms that colonize various anatomical niches, such as the gastrointestinal tract, skin, and other regions of the human body, playing a key role in the proper functioning and maintenance of the organism.

Among the many communities inhabiting different body niches, the gut microbiota stands out as one of the most abundant and diverse communities. It is commonly known as the set of microorganisms (bacteria, viruses, fungi, among others) that reside in the delicate balance of the gastrointestinal tract, playing an essential role in health and the optimal functioning of the organism. Bacteria are the most studied group of organisms of these communities, and the focus of this work.

The gut microbiome (or bacteriome, when referring only to the bacterial component) composition is influenced by a variety of factors such as diet, genetics, and lifestyle. A comprehensive analysis of this symbiotic interaction is imperative to understand its contribution to health and disease (Ottman et al., 2012). The visual representation provided in Figure 1 offers a panoramic view of the observable changes and diversity in the composition of the microbiota under different pathophysiological conditions.

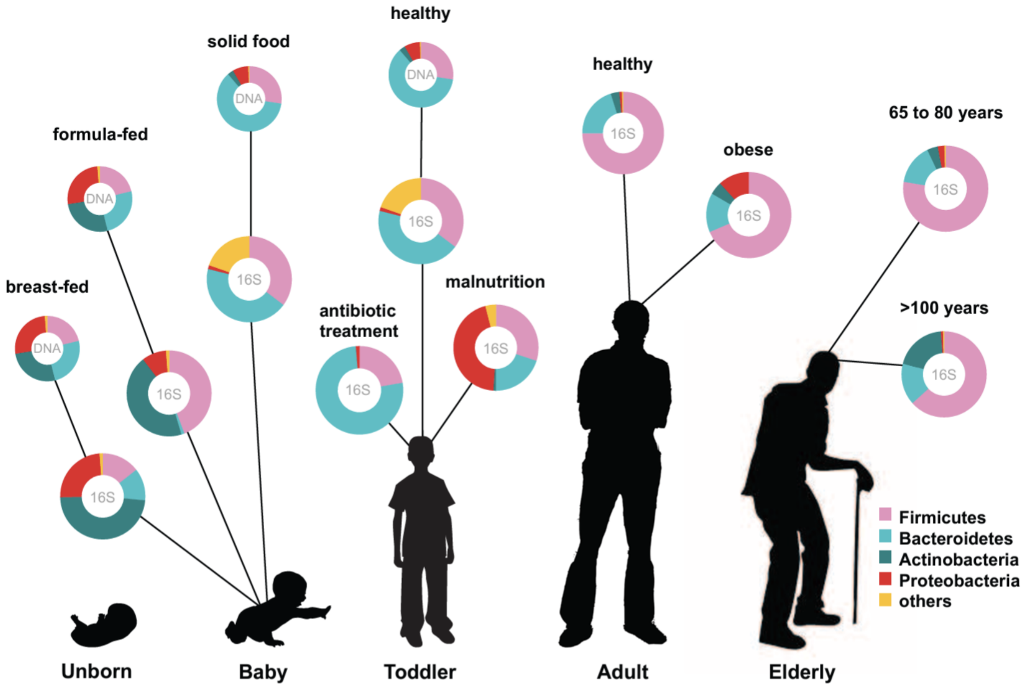


Figure 1: Variation in the predominant bacterial composition of the human microbiota according to age and altered metabolic conditions. The relative abundance of major bacterial phyla in the human microbiota and

their variations at different stages of life and in compromised metabolic states, such as obesity or malnutrition, are illustrated. Data were obtained through 16S rRNA gene sequencing. Ottman et al., 2012.

The impact of the microbiota on human health and the pathogenesis of various diseases manifests in a variety of forms. This microbial community has been associated with the development of metabolic disorders, neurodegenerative and autoimmune diseases, and may modulate the host immune response and interactions with pharmacological agents. However, in most cases, it is still unknown whether the microbiota is the cause or consequence of these diseases and associations.

In this context, the gut microbiota, with its diversity of bacterial species, plays critical roles in nutrient digestion, immune system modulation, and resistance against invasive pathogens. This role is mainly accomplished through the functions performed by these bacteria. For instance, metabolites produced by specific microorganisms in the gut microbiota, such as short-chain fatty acids (Ratajczak et al., 2019), tryptophan (Siying, 2023), and bile acid metabolites (Funabashi et al., 2020), exert an influence not only on genetic and epigenetic regulation but also on the metabolism of immune cells, including both immunosuppressive and inflammatory cells. Different receptors for short-chain fatty acids, tryptophan, and bile acid metabolites from various microbial species have been identified in diverse immunecells. Activation of these receptors not only stimulates the differentiation and function of immunosuppressive cells but also inhibits inflammatory cells, leading to a reprogramming of the local and systemic immune system to maintain homeostasis (Wang et al., 2023).

Additionally, it has been observed that some bacteria can influence the development of certain types of cancer, amplifying or mitigating their effects and evolution. An example of how microbes contribute to carcinogenesis is Fusobacterium nucleatum. It has been observed (in vitro) that the FadA protein produced by this bacterium binds to the E-cadherin receptor, activating the β-catenin pathway and inducing cell proliferation. This interaction highlights the potential role of specific bacterial functions in the development of cancer by directly influencing host cell behavior (Rubinstein et al., 2013).

As demonstrated by these examples, a detailed understanding of the complex interactions between the microbiota and the host, as well as their role in health and disease, has been the subject of intense research in recent decades. Since the importance of the gut microbiota for human health lies in the functions performed by these bacteria, methods that can identify and quantify these functions are required. Different methods and models used in the study of the gut microbiota, together with their capability to quantify microbial functions, are discussed in the next section.

1.2 Analysis methods and study models of the microbiota

To achieve a greater understanding of the composition and functions of the gut microbiota, different analysis methodologies are currently available. Additionally, the choice of an appropriate study model is crucial to obtain results as close to reality as possible.

1.2.1 Analysis methods

Advances in sequencing and bioinformatics techniques have enabled the development of detailed analysis methods, such as 16S rRNA amplicon sequencing, metagenomics, and metatranscriptomics for the study of microbial communities. Additionally, proteomics and metabolomics are used to study the functions of these communities, providing a comprehensive understanding of microbial activity and interactions. This section will focus on cultivation-free methods, as they allow for the study of bacteria without the need for cultivation, offering a more thorough insight into microbial diversity and function.

1.2.1.1 16S rRNA amplicon sequencing

The 16S rRNA gene encodes for the 16S ribosomal RNA, which is a component of the small subunit of the bacterial ribosome. Due to the very high conservation of this gene across different bacterial species, this gene is considered an ideal tool for performing the taxonomic identification and classification of bacteria. Since the 16S rRNA gene sequence contains conserved regions, it is possible to design quasi-universal primers that allow for amplification and sequencing of variable regions, which can be used to assign bacterial taxonomy 16S amplicon sequencing is important for the identification and taxonomy of bacteria for the following reasons:

- **Universality and conservation:** Based on the conserved regions of the 16S rRNA gene, it is possible to design universal primers for a broader range of bacteria that can be amplified. This facilitates the comparison of experiments that have used the same set of primers, and the creation of broad databases for 16S sequences such as SILVA, RDP, Greengenes, etc (Martínez-Porchas and Vargas-Albores, 2017), facilitating computational analysis and taxonomic assignation.
- **Specific variability:** The variable regions of the 16S rRNA gene provide enough variability to allow distinction between different bacterial genera, and between different species in some cases. Sufficient variability is necessary for conducting comprehensive phylogenetic analyses and thus for appropriate taxonomic classification (Yang et al., 2016).
- **Wide range of applications:** 16S rRNA sequencing is performed in many settings beyond the study of the gut microbiota composition. For instance, it is used in clinical applications to identify pathogens in infections that cannot be easily diagnosed through conventional culture-based methods. (Shokralla et al., 2012; Duvallet et al., 2017). Additionally, 16S rRNA sequencing has been applied in environmental and food safety studies, allowing for the detection and identification of bacteria in various contexts.

Although 16S rRNA gene sequencing is a powerful tool for the characterization of microbial communities, it has several limitations and disadvantages that must be considered in gut microbiome studies. These limitations can influence the accuracy and interpretation of the results obtained (Abellan-Schneyder et al., 2021). These drawbacks include:

- **Influence of primer design:** The choice of primers and the variable (V) region of the 16S rRNA gene that is amplified can significantly impact the microbial profiles obtained. Different primers may not uniformly amplify all bacterial species present in the sample, leading to the underrepresentation or even absence of certain taxa (Abellan-Schneyder et al., 2021; Apprill et al., 2015; Walters et al., 2015). There can also be variability between different studies: results may not be comparable between studies that use different primer combinations or target different variable regions, complicating cross-study data validation (Abellan-Schneyder et al., 2021; Klindworth et al., 2013; Walters et al., 2014).
- **Limitations of reference databases:** The accuracy of taxonomic assignment largely depends on the databases used for analysis. Outdated databases, due to the lack of new taxa or modifications to existing ones (Abellan-Schneyder et al., 2021), or differences in nomenclature used in each database (McDonald et al., 2012), can provide unrepresentative results.
- **Lack of standardization:** In addition to the choice of primers, which has already been discussed, other factors such as sample processing, extraction kits, and sequencing protocols play crucial roles in the outcome of microbiome studies. However, there are no established standards in the field, making it challenging to compare results across different studies and reducing the reproducibility of findings. The lack of standardization in these critical steps can lead to variability and inconsistencies in the data, further complicating efforts to draw reliable conclusions from microbiome research.
- **Problems resulting from the complexity of microbial communities:** The complexity of microbial communities can affect the accuracy of taxonomic identification. Primers and databases that work well with simple microbial communities may not be ideal for more complex ones. Studies have shown that certain primers and databases do not correctly identify various taxa in more complex communities, leading to potential misinterpretations (Abellan-Schneyder et al., 2021; Schloss et al., 2011). Additionally, contamination poses a significant challenge, particularly in low-biomass samples, as it can introduce spurious taxa not originally present in the sample. Other issues such as PCR bias and the variable number of ribosomal operons across different bacterial species can further complicate the analysis, potentially skewing the representation of certain taxa. Moreover, 16S rRNA sequencing does not provide information about the functional roles of bacteria within the microbiota. While it is possible to predict functions based on detected taxa (Langille et al., 2013), these predictions do not indicate whether these functions are actively being performed. This limitation underscores the importance of methods like metatranscriptomics for studying active bacterial functions.

1.2.1.2 Metagenomics

Metagenomic sequencing is defined as the sequencing of the entire genetic material extracted directly from environments without prior laboratory cultivation, allowing for microbial community analysis and characterization of its functional potential, that is, the characterization of the functions that are encoded by the genomes of the community. Besides being applied to the study of the intestinal microbiome, it is also particularly effective for investigating complex microbial populations from various environments such as soil and water, and is also widely used in clinical settings (Purushothaman et al., 2022).

Metagenomic analysis follows several key steps, each crucial for obtaining an accurate and detailed characterization of the microbial communities present in the sample. The first step is the extraction of DNA from the sample to be analyzed. This genetic material must be of the highest possible quality and quantity to ensure accuracy in subsequent analyses. The second step involves sequencing the extracted material. Shotgun metagenomics focuses on sequencing all the DNA present in the sample. Unlike amplicon sequencing, shotgun metagenomics is not limited to specific regions of the genome. Instead, it provides a comprehensive view of the complete genetic content of a sample, including bacteria, viruses, archaea, and eukaryotes. This method is more expensive and complex, but it offers much higher resolution and allows for functional inferences (Purushothaman et al., 2022). Shotgun metagenomics facilitates the identification of specific genes associated with metabolic functions, antibiotic resistance, and virulence.

Finally, after obtaining the raw sequencing data, bioinformatic analyses must be performed to infer the microbiota composition and functional capabilities, such as antibiotic resistance and the degradation of organic compounds (Bortolaia et al., 2020). However, since metagenomics involves sequencing of the entire genomes, rather than a specific region, bioinformatic analysis of metagenomic data becomes more complex than that of 16S amplicon sequencing datasets. Despite its complexity, metagenomics has several advantages and important applications in bacterial identification and the study of microbial communities. These include the following:

- **Detection of non-cultivable microorganisms:** There is a great diversity of microorganisms in the environment that are not cultivable in the laboratory. Both metagenomics and 16S amplicon sequencing allow for the detection and analysis of these directly from the sample, while culture-dependent techniques remain limited for this reason (Lagier et al., 2015). A significant advantage of metagenomics over 16S amplicon sequencing is that it does not rely on the efficacy of primers, allowing for a more comprehensive analysis. Additionally, metagenomics enables the exploration of a broader range of organisms beyond bacteria, including viruses, fungi, and other microbes.
- **Higher taxonomic resolution:** Metagenomics enables the analysis of the structure of complex microbial communities by sequencing the entire genome rather than just a region. This allows for resolution at the species level and sometimes even at the strain level, which is crucial for understanding microbial interactions and their influence on the ecosystem or human health (Purushothaman et al., 2022).
- **Analysis of functional potential:** In general, the major advantage of metagenomics, aside from its resolution, is that it allows for the identification of the functions encoded by the genes of the detected bacteria. Therefore, in gut microbiome samples, metagenomics can be used to detect metabolic, virulence or antibiotic resistance genes, among others. For instance, detecting antibiotic resistance genes. is extremely important for epidemiological monitoring and infection control (Bortolaia et al., 2020).

Despite these advantages, metagenomic sequencing is not exempt from limitations and difficulties. Among them are:

- **Complexity of data analysis:** Metagenomics generates a large volume of data, which requires significant computational power for analysis and processing. (Liu et al., 2020; Langmead et al., 2012).
- **Incomplete taxonomic assignment:** Despite achieving a much higher resolution than 16S amplicon sequencing, accurate taxonomic assignment of sequenced DNA fragments can be complicated due to the diversity of microorganisms present in complex communities and the limitations of reference databases or *de novo* genome assembly methods. This can result in incorrect or incomplete assignments (Liu et al., 2020; Truong et al., 2015).
- **DNA contamination and background noise:** Metagenomics is highly susceptible to DNA contamination, typically from the environment or the laboratory. This issue is also similar in 16S rRNA sequencing. This can introduce biases or errors in the results obtained. To minimize this risk as much as possible, it is necessary to perform negative controls (Liu et al., 2020; Fresia et al., 2019).
- **High costs:** Metagenomic sequencing is more expensive than amplicon sequencing methods, which can ultimately limit the number of samples that can be processed in the study (Liu et al., 2020; Bolger et al., 2014).
- **Functional potential:** It is important to note that the functions detected through metagenomics represent the "functional potential" of the microbial community. This means that the functions are encoded in the genomes of these organisms, but it does not imply that these functions are being expressed in a specific sample. This distinction is crucial when comparing metagenomics with metatranscriptomics, as the latter allows for the study of actively expressed functions (Xing et al. 2020).

1.2.1.3 Metatranscriptomics

Metatranscriptomics involves the study of the transcriptome of all microorganisms present in a specific biological niche. (Reigstad & Purna, 2013). The transcriptome represents the complete set of RNAs expressed by each microorganism, thereby reflecting the functions that microorganisms in a community are actively performing. In the context of the intestinal microbiota, metatranscriptomics can uncover the functions carried out by this community under various conditions, such as during different diseases.

This technique not only identifies the active microorganisms present in an ecosystem but also seeks to understand how these microorganisms function and interact within their biological environment. This understanding is crucial for identifying functions that may influence human health, as well as the microorganisms responsible for these functions (Sánchez-Rumí & Lloréns-Rico, 2024). This approach allows the discovery of microbial activities related to disease states efficiently.

Metatranscriptomics reveals biological information that is not easily accessible through conventional genomic profiling methods and complements metagenomic and metataxonomic assessments, which generally do not distinguish between active, inactive, and dead members of the community. To perform a metatranscriptomic analysis, a general process is followed that encompasses several key stages. First, proper handling and processing of samples are crucial for accurate metatranscriptomic analysis. RNA stabilizing solutions, such as RNAlater, allow preservation at room temperature for several days, facilitating sample collection and transport. Alternatively, snap-freezing can be used (Reck et al., 2015; Liu et al., 2020). Although proper sample preservation is significantly more important in metatranscriptomics due to the short half-life of RNA, it is also essential in metagenomics or 16S amplicon sequencing. If samples are not well-preserved, oxygen-tolerant bacteria can grow, distorting the relative abundance of other bacteria.

Efficient RNA extraction is necessary to achieve proper cell lysis and release their contents. Mechanical lysis methods like bead beating are popular due to their ability to increase the detection of greater bacterial diversity (Gangadoo et al., 2021). While this is not specific to metatranscriptomics, the key here is to ensure RNA preservation throughout the entire process, from sample collection to sequencing.

For metatranscriptomic sequencing, RNA must be of the highest possible quality (Giannoukos et al., 2012). Data analysis resembles that of metagenomics, including preprocessing taxonomic and functional assignment, but differential transcript abundance or differential activity tests can also be performed (Franzosa et al., 2014).

Metatranscriptomics offers multiple benefits. Among them are:

- **Detection of active bacteria, active genes and their functions:** It allows the identification of which genes are being expressed in the sample and in which bacteria, as well as the function of these genes (Filiatrault, 2011; Zhang et al., 2021).
- **Greater sensitivity in detecting infectious diseases:** Metatranscriptomics has proven to be more efficient in detecting infectious diseases compared to metagenomics, particularly because it can be used to detect infections caused by RNA viruses, such as SARS-CoV-2, for which metagenomics are not as appropriate (Tao et al., 2022).

Regarding the limitations of metatranscriptomic studies, the following stand out:

- **Contamination and Background Noise:** The susceptibility to contamination from environmental and laboratory RNA can introduce biases in the results. It is crucial to use negative controls and rigorous cleaning techniques (Glassing et al., 2016).
- **Data Analysis Complexity:** As metagenomics, it requires advanced computational resources and specialized bioinformatics tools for processing and analyzing large volumes of data (Giannoukos et al., 2012).
- **RNA Stability:** RNA is an unstable molecule prone to degradation, requiring strict storage and handling conditions (Deutscher, 2006). This is more important in the case of bacteria since the half-life of RNA in prokaryotes is significantly shorter than in eukaryotic cells.
- **Challenges in Sample Preparation:** The efficiency of RNA extraction methods can vary between organisms, sample materials, and RNA species, which can result in uneven yields (Ali et al., 2017). This issue is similar for other techniques such as 16S

rRNA sequencing and metagenomics. High-quality RNA is needed to perform sequencing.

Table 1 more clearly summarizes all the advantages and limitations or drawbacks of the explained analysis methods.

Analysis Method	Advantages	Limitations
16S rRNA Sequencing	- Rapid taxonomic identification (Abellan-Schneyder et al., 2021). - Conservation allow for the design of quasi-universal primers (Martínez-Porchas and Vargas-Albores, 2017). - Standardization in microbiological studies (Hamady & Knight, 2009). - Clinical and environmental applications for identifying pathogens and hard-to-culture bacteria (Shokralla et al., 2012; Duvallet et al., 2017).	- Primer design influence can bias results (Abellan-Schneyder et al., 2021; Apprill et al., 2015). - Limitations of reference databases can affect accuracy (McDonald et al., 2012). - Issues with the complexity of microbial communities can hinder identification (Schloss et al., 2011). - Does not provide functional information about microorganisms.
Metagenomics	- Allows detection of uncultivable microorganisms (Lagier et al., 2015). - Analysis of complex communities and their dynamics (Purushothaman et al., 2022). - Public health studies to identify emerging pathogens and their resistances (Li et al., 2020). - Detection of active genes and their functions in the sample (Filiatrault, 2011; Zhang et al., 2021).	- Data analysis complexity requires advanced computational resources (Liu et al., 2020; Langmead et al., 2012). - Taxonomic assignment can be incomplete or incorrect due to database limitations (Truong et al., 2015). - High susceptibility to DNA contamination and background noise (Fresia et al., 2019). - High sequencing costs can limit the number of samples processed (Bolger et al., 2014). - Variability in relative abundance of microorganisms can hinder detection of less abundant microorganisms (Arumugam et al., 2011).
Metatranscriptomics	- Greater sensitivity in detecting infectious diseases (Tao et al., 2022). - Detection of active bacteria, active genes and their functions:	- Data analysis complexity requires advanced computational resources (Giannoukos et al., 2012). - RNA stability is low, requiring strict storage and handling conditions (Deutscher, 2006). - Challenges in sample preparation due to variability in RNA extraction efficiency (Ali et al., 2017). - Library preparation bias can affect result accuracy (Grünberger et al., 2019).

Table 1: Summary of the advantages and limitations of current microbiota analysis methods. *This table highlights the advantages and limitations of the three main methods for analyzing the intestinal microbiota: 16S rRNA gene sequencing, metagenomics, and metatranscriptomics. Each method offers unique benefits and specific challenges that must be considered when choosing the most appropriate technique for a particular study.(Own elaboration)*

1.2.1.4 Other Analysis. Metabolomics and metaproteomics

Other important methods in the study of the intestinal microbiota include metabolomics and proteomics. Metabolomics focuses on the analysis of metabolites, the small molecules produced by metabolic processes, thus allowing the study of the active functions of the microbiota by identifying the metabolites produced (Puljiz, et al., 2023). Metaproteomics, on the other hand, is dedicated to the study of the proteins present, providing information on the functions these proteins perform within the microbial ecosystem (Ruiz, et al., 2016). Although both techniques offer a detailed view of the active functions and metabolic processes in the microbiota, they present the challenge of accurately identifying which specific members of the microbiota are producing these metabolites and proteins, which may require complementary techniques and integrative analyses to obtain a complete understanding.

1.2.2 Study models

In order to evaluate the functions of the gut microbiota in the organism, the methodologies detailed in the previous section are applied to different study models such a*s in vitro* models, animal models, or human studies.

1.2.2.1 *In vitro* models

In vitro models have become essential tools for studying specific members of the gut microbiota and their interaction with the human intestine. These models can compensate for some of the limitations of animal models or human studies. *In vitro* models can include the host component (such as in cell cultures, or organoids) or not (such as in *in vitro* fermentation assays). When including the host, in vitro models can help clarify how microorganisms interact with the human intestinal epithelium, facilitating high-throughput studies that are crucial for better understanding human intestinal biology and its microbial interactions. Additionally, these models are indispensable for selecting effective probiotics and designing therapeutic interventions, providing a robust and versatile platform for biomedical research (Qi et al., 2023).

These models present a series of advantages and limitations compared to animal models. They provide a more controlled environment for studying complex interactions, eliminating in some cases the need for costly and ethically complex *in vivo* studies (Qi et al., 2023). Additionally, advanced models such as organoids and microfluidic systems allow for precise simulation of human physiological conditions, including oxygen gradients and mechanical forces, enhancing the biological relevance of the studies (Nikolaev et al., 2020; Puschhof et al., 2021). Regarding limitations, most *in vitro* models lack essential components of the human microenvironment, such as connective tissues or immune cells, which limit the model ability to replicate the complexity of the human gastrointestinal system (Puschhof et al., 2021). Moreover, the costs and time required to establish and maintain the most advanced models can be significantly high, which often limits their accessibility for large-scale studies (Qi et al., 2023).

In these models, the techniques mentioned earlier are often used in combination with culture-based methods. For instance, in the fermentation of fecal samples, techniques like 16S rRNA sequencing or metagenomics can be employed to study the dynamics of the microbial community.

1.2.2.2 Animal Models

Conducting animal studies is essential for analyzing the gut microbiota for various reasons. Animal models allow for a comprehensive understanding of the complex interactions between the microbiota and the host in a living organism, something that is not fully replicable with *in vitro* models. These studies enable the observation of the systemic effects of the microbiota, including immune, metabolic, and physiological responses, in a complete biological environment (Backhed et al., 2005). Additionally, animal models can accurately simulate human pathological and physiological conditions, providing valuable insights into the role of the microbiota in various diseases (Turnbaugh et al., 2007). The ability to genetically manipulate animals also allows for the investigation of specific mechanisms and causalities, which is crucial for developing targeted therapies and effective probiotics (Kau et al., 2011).

Among animal models, one stands out above the rest: the mouse. This is due to its physiological and genetic similarity to humans. Although there are significant differences in gut microbiota composition between mice and humans due to differences in behavior, intestinal transit time, and intestinal structure, mice still provide valuable insights. The major advantage of using mice is the control over diet, genotype, and other variables. Additionally, the relative ease of genetic manipulation in mice (Turnbaugh et al., 2009) makes them an ideal model organism for studying the microbiota.

One of the main ways to analyze the intestinal microbiota in mice is through the use of fecal samples for 16S, metagenomic or metatranscriptomic studies. Besides fecal samples, another widely employed strategy is to obtain samples from the mouse cecum, as this section of the intestine harbors a high bacterial load. This high microbial content greatly facilitates the obtaining of metagenomic and metatranscriptomic results, as a large portion of the extracted and analyzed genetic material will predominantly belong to bacterial species rather than the host (Just et al., 2018). Furthermore, the cecum provides an *in situ* sample, which can be of great interest for certain studies, given that fecal sample collection presents inconveniences; for example, prolonged exposure of the sample to oxygen can alter its composition and functions.

Although animal models, especially mice, have proven to be invaluable tools for studying the intestinal microbiome, it is important to recognize that they also have certain limitations. These limitations can influence the accuracy and applicability of the results obtained to human biology. Below are some of the main limitations of animal models in the context of studying the intestinal microbiome:

- **Complexity and Diversity of the Microbiome:** Although animal models are useful, they do not fully capture the complexity and diversity of the human microbiome. There are significant differences in gut microbiota composition between mice and humans due to differences in behavior, intestinal transit time, and intestinal structure (Walter et al., 2020).

- **Variability between Experiments:** There is significant concern regarding biological reproducibility. Variations in the growth of individual strains can lead to substantial differences in community composition in experiments conducted on different days, affecting the architecture of the living community (Walter et al., 2020). Moreover, it is important to consider not only the intrinsic variability of the microbiota but also the conditions of the animal facility, the operator, and other external factors that can influence experimental results.
- **Ethical Evaluation Needed to Begin Animal Research**: Ethical regulations on the use of experimental animals can significantly increase the cost and time of studies in certain instances. Additionally, obtaining cecal samples requires the sacrifice of the mice, which prevents the collection of longitudinal data, something that would be possible with fecal samples.

1.2.2.3 Human Studies

In the same way that animal studies are necessary to begin understanding how the microbiota impacts the proper functioning of the organism, human studies are essential because they eliminate any existing variability between species when translating findings from animals to humans, resulting in more reliable outcomes. In the field of gut microbiota, fecal samples are predominantly used, and in some cases, intestinal biopsies. Each type of sample has its own set of challenges and limitations, which must be considered when interpreting results.

Fecal samples have been used to evaluate changes in the microbial ecosystem in relation to several inflammatory bowel disease s such as ulcerative colitis or Crohn's disease (Lloyd-Price et al., 2019). While these samples provide a non-invasive way to study the gut microbiome, they come with certain limitations. For instance, in human studies, fecal samples may not fully represent the microbiome of other gut regions, which can be a limitation compared to animal models where more invasive sampling is possible. This is true especially for metatranscriptomics studies, where it is important to capture the functionality of the microbiota in specific regions of the colon, such as tumors or inflamed regions. In this case, biopsy samples may be preferrable, but in this case sampling is much more invasive and complicated.

Additionally, it is more challenging to conduct interventional studies in humans, regardless of the type of sample used, due to ethical and logistical constraints.

Table 2 provides a clearer compilation of all the advantages and limitations or drawbacks of the study models employed in microbiota analysis as explained above.

Study Model	Advantages	Limitations
	- Controlled environment allows precise manipulation of variables. - Cost-effective and relatively quick to set up and run. - High reproducibility	- Lack of complexity compared to the in vivo environment. - Limited representation of the complex interactions in a living organism.

In Vitro Models	- Useful for high-throughput screening and mechanistic studies	- Cannot fully replicate the immune responses and systemic effects observed in vivo. - Often lack the full diversity of microbiota present in the human gut.
Animal Models (e.g., mice)	- Similar physiological and genetic characteristics to humans (Ley et al., 2005). - Ability to study systemic interactions and immune responses (Turnbaugh et al., 2007). - Genetically modifiable to study specific genes and pathways (Kau et al., 2011). - Can mimic human disease conditions and study their progression.	- Ethical concerns and regulatory limitations. - Differences in microbiota composition between animals and humans. - High cost and longer timelines compared to in vitro models. - Environmental factors and housing conditions can affect microbiota composition.
Human Studies	- Direct relevance to human health and disease. - Comprehensive understanding of microbiota interactions in the human body. - Ability to study the direct impact of interventions on human microbiota. - Provides real-world data on the effects of diet, lifestyle, and medication.	- High variability due to genetic, environmental, and lifestyle factors among individuals. - Difficulty in obtaining longitudinal samples and maintaining consistent conditions. - Ethical and logistical challenges in conducting controlled studies. - High cost and complexity in study design and execution.

Table 2: Summary of the advantages and limitations of current microbiota study models. *This table highlights the advantages and limitations of the main study models for analyzing the intestinal microbiota: in vitro models, animal models, and human studies. Each model offers unique benefits and specific challenges that must be considered when choosing the most appropriate approach for a particular study. (Own elaboration)*

1.3 Challenges in gut microbiota functional studies

As mentioned in the previous section, it is crucial to study the functions of the microbiota to understand their precise contribution to health and disease. Techniques like 16S and metagenomics do not allow us to discern which functions are actively being expressed by the bacteria in the microbiota. Compared to 16S and metagenomics, there are still very few metatranscriptomic studies in gut microbiota, which are the ones that allow to distinguish the active functions being expressed by the microbial communities. The remainder of this section will focus on the specific challenges associated with metatranscriptomics. These challenges include the difficulty in capturing the full range of active microbial transcripts, the potential for RNA degradation during sample collection and processing, and the need for high-quality, high-throughput sequencing technologies to accurately analyze the complex and dynamic nature of the microbiome's transcriptome.

Most metatranscriptomic studies in the context of the human gut microbiota use fecal samples as a source of microbial RNA. Fecal samples are highly valued for their ease of collection, as patients or healthy volunteers can collect them without medical intervention. Moreover, they contain a large amount of microbial biomass, which facilitates metatranscriptomic analyses, making them useful for some microbiota studies (Sánchez-Rumí & Lloréns-Rico, 2024).

Most metatranscriptomic studies in humans have been conducted using fecal samples, largely because they are easy to obtain in a non-invasive manner. However, in many cases fecal samples are not ideal for metatranscriptomic analysis. Due to short bacterial mRNA half-lives (Rauhut, et al., 1999), microbial RNAs in fecal samples may not be representative of the functions occurring in specific areas of the intestine, such as tumors or lesions located in the proximal colon (Sánchez-Rumí & Lloréns-Rico, 2024). Additionally, short mRNA half-lives make aspects such as immediate sample preservation even more crucial in metatranscriptomics than in other techniques.

Therefore, an alternative approach to the use of fecal samples involves using colon biopsies to analyze the functionality of the gut microbiota using metatranscriptomics. Colonic biopsies are more representative of the specific sections of the intestine being studied in terms of microbial transcription. This direct contact with the affected area allows for a more precise analysis of microbial transcriptional alterations in specific areas, being particularly useful for detailed colon studies (Sánchez-Rumí & Lloréns-Rico, 2024).

However, biopsies also have their disadvantages (Sánchez-Rumí & Lloréns-Rico, 2024). The collection of colonic biopsies requires invasive endoscopic procedures, limiting the accessibility to samples, especially in healthy individuals (Granata *et al*., 2020 The lower microbial biomass in biopsies compared to fecal samples entails additional limitations, such as a higher risk of contamination during collection and processing (Sánchez-Rumí & Lloréns-Rico, 2024). Although the impact of contamination in low biomass samples has been studied in the context of 16S and metagenomics, less is known about its impact in metatranscriptomics. Contaminants may arise from the collection procedures, extraction protocols or library preparation methodologies. Additionally, the low biomass in biopsies requires either specific methods for host RNA and rRNA depletion or a high sequencing depth, resulting in high costs of sequencing (Mahmoudabadi *et al*., 2022).

Despite these challenges, some studies have successfully conducted metatranscriptomic analyses on intestinal biopsies in humans. For instance, studies aimed at determining if the gut microbiota is related to the development of obesity or other metabolic diseases have been conducted (Granata et al., 2020). The goal of these studies is to identify alterations in gene expression in both human and microbial subjects in severely obese individuals compared to lean individuals. By analyzing duodenal biopsy samples using next-generation sequencing, researchers aim to better understand how changes in microbial and human genes contribute to dysregulated metabolic pathways, affecting energy metabolism and contributing to the obese phenotype. This study is the first report on duodenal metatranscriptomic profiles in obese subjects and could provide valuable insights for developing therapeutic strategies aimed at modifying microbial composition and/or function to favorably impact host metabolism (Granata et al., 2020).

1.4 Objectives

Understanding the functions of the gut microbiota is essential, and metatranscriptomics is a powerful technique that allows us to study these functions and identify which microbes are responsible for them. However, when using fecal samples, the relevance of the observed functions is limited, and studies using biopsies are scarce due to technical challenges. Therefore, the overarching question of this study is: What is the best protocol for metatranscriptomic analysis of intestinal biopsies? This question encompasses the type of sample, the extraction kit used, and variations within the kits.

To address this question, a pilot study was conducted using mouse colonic biopsies. Mouse models were used instead of human samples because they allow for controlled experimental conditions and are widely accepted in preclinical research. Moreover, using mice enables us to perform more invasive sampling and to replicate the study under consistent conditions, which is challenging in human studies. Although there are species differences, the fundamental processes and interactions in the gut microbiota are similar between mice and humans. This approach provides a basis for reasonably assuming that the findings from this study can be translated to human applications.

Using these samples, different DNA and RNA extraction protocols were evaluated using four commercial kits. Variations in RNA purification treatments and the type of sample used were also assessed. The aim of this study is to determine if a standardized protocol can be established for processing samples for metatranscriptomic analyses, and if so, identify the best approach. To answer this question, the following objectives were outlined in this study:

1. Generate a collection of intestinal samples of different mouse models that can be used as a proxy for human intestinal biopsies in the evaluation of processing protocols.
2. Conduct comparative analyses of different DNA and RNA extraction protocols using four different commercial kits on the collected mouse samples.
3. Perform 16S amplicon sequencing on the DNA of the extracted samples, to detect DNA contaminants (as a cost-effective proxy to detect potential contaminants present in the RNA).
4. Identify the most appropriate sample type and extraction protocol based on criteria such as the quantity of RNA obtained, the proportion of bacterial biomass in the sample (measured by qPCR, RT-qPCR, and 16S amplicon sequencing), the RNA quality, and the proportion of contaminants detected in actual samples (determined by 16S amplicon sequencing).

2 Materials & Methods

2.1 Sample collection and storage

For this study, mice from several research groups associated with the CIPF were obtained. These mice belong to different strains and research groups with various genotypes and different diets (see supplementary table S1).

Different euthanasia methods were used to sacrifice mice (see supplementary table S1). It is important to note that the euthanasia was conducted as part of research projects that received the corresponding approval from the ethics committee. This procedure was performed by a professional qualified in handling live animals. Access to the samples was granted only after the mice had been sacrificed; no intervention was made prior to this. All animal procedures were conducted in strict compliance with the European Community Directive (2010/63/EU) and Spanish legislation (RD53/2013).

Samples were collected from different parts of the mouse lower intestinal tract, specifically from the distal colon (biopsy and mucosal scraping), the proximal colon (biopsy and mucosal scraping) and the cecum (cecal contents). Figure 2 shows a schematic of different sections of the murine colon and the areas from which samples were extracted.

Specific dissection tools were used to obtain biopsies swiftly, with the goal of preventing genetic material degradation and minimizing tissue damage. Immediately after animal euthanasia, the mouse was secured to a dissection board using surgical needles, and the area was sterilized with 70% ethanol. An incision was made in the abdomen with round-tipped surgical scissors, and the outer skin was separated from the peritoneum. The peritoneum was opened using fine-tipped or precision surgical scissors. With the help of precision tweezers, the colon was located, separated, and placed in a glass Petri dish. Using different sterile scalpel blades, various samples were taken from the cecum, distal colon, and proximal colon, as well as mucosal scrapings. After obtaining all samples, they were stored on dry ice until transfer to an ultrafreezer at -80°C to prevent RNA degradation.

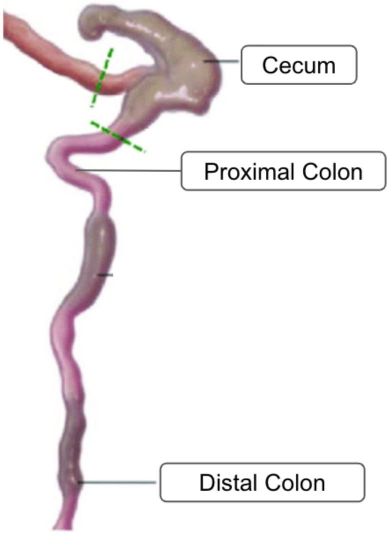


Figure 2: Diagram of the mouse intestine highlighting the areas where biopsies were obtained. (Own elaboration)

2.2 DNA/RNA Extraction

For the extraction of both DNA and RNA, different commercial extraction kits were employed. This is important because, when using biopsies that are difficult to obtain, the use of kits that extract both DNA and RNA simultaneously minimizes sample requirements. This way, it is not necessary to perform two separate extractions, as the same result can be achieved with a single extraction. The following kits were used: AllPrep DNA/RNA Mini (ID: 80204) from Qiagen, ReliaPrep™ RNA Tissue Miniprep System (ID: Z6110) from Promega, DNA/RNA/Protein Isolation Kit from NZYTech (ID: MB45901), and NucleoSpin TriPrep from Macherey-Nagel (ID: 740966.50). These kits share a common sample lysis procedure while also featuring specific differences in the following extraction steps. After the extractions, DNA was stored at -20°C and RNA at -80°C. Each batch of nucleic acid extractions included a blank, which consists of an empty tube processed following the same protocol as specified for each kit. The inclusion of a blank serves as a control to detect potential contaminants that may originate from the extraction kits themselves. This is essential because kit-specific contaminants, often referred to as the 'kitome,' can introduce biases and confound results in metagenomic and metatranscriptomic analyses (Salter et al., 2014).

Figure 3 summarizes the extraction procedures for all kits, highlighting the similarities and differences in the mechanisms of action of each of the kits used.

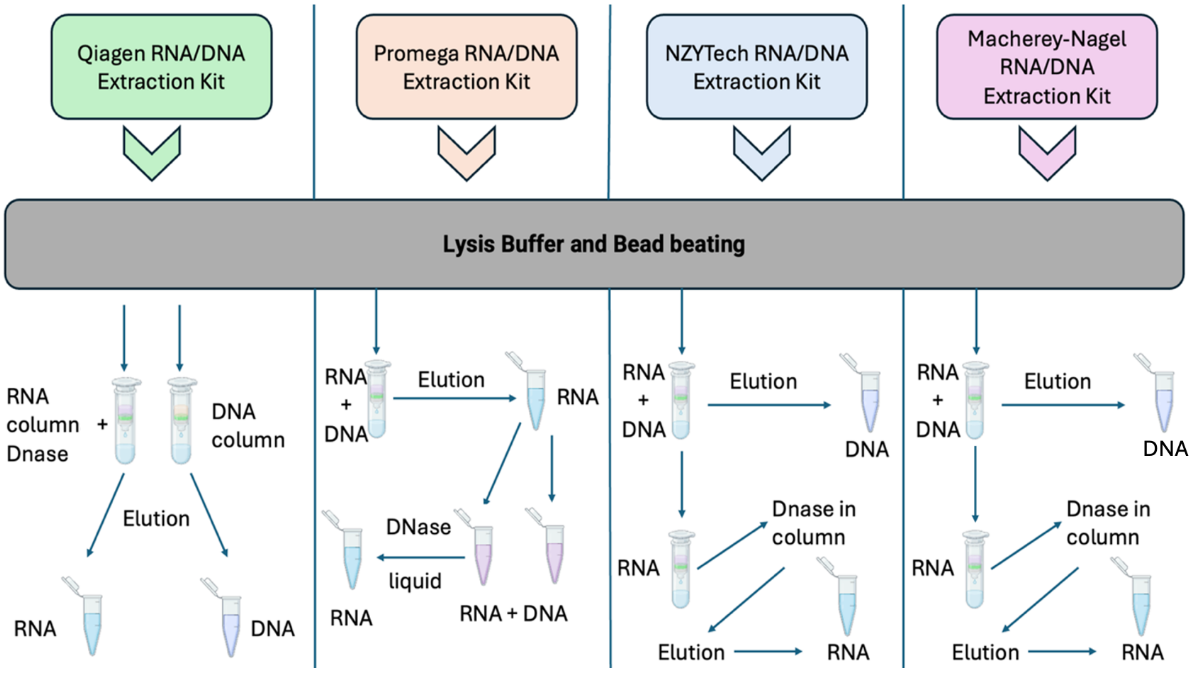


Figure 3: A diagram summarizing the differences in protocols used for various RNA/DNA extraction kits, including Qiagen, Promega, NZYTech, and Macherey-Nagel. The diagram illustrates the specific steps and methodologies, such as column DNase treatment, and liquid solution DNase treatment, highlighting the unique processes and elution points for RNA and DNA extraction across the different kits.

2.2.1 Common sample lysis procedure

The four extraction kits share a common initial procedure. This involves thawing the sample in the presence of the lysis buffer recommended by each kit's manufacturer (e.g., RLT Plus for Qiagen, which requires β-Mercaptoethanol), and DX Reagent in Pathogen Lysis Tubes (ID: 19092), which include the tubes with the beads and the DX Reagent to prevent foaming. Tissue disruption was carried out using Bead Beating, utilizing a Bead Beater at a speed of 2400 rpm for one minute, followed by one minute on ice, and repeating this cycle twice.

2.2.2 Qiagen DNA/RNA Extraction

In the nucleic acid extraction using the AllPrep kit from Qiagen, the procedure adhered closely to the manufacturer's instructions but included specific adaptations. This protocol separates DNA and RNA purification processes using dedicated columns.

Firstly, a working solution of RLT buffer was prepared by combining 350 µL of buffer RLT, 3.5 µL of β-mercaptoethanol, and 1.75 µL of DX Reagent per sample in a 5 mL tube. The tissue sample was then added to each pathogen lysis tube along with 300 µL of the prepared RLT buffer solution. The samples underwent homogenization using the bead beating protocol mentioned above.

After homogenization, the lysate underwent brief centrifugation, and the supernatant was transferred to an AllPrep DNA Mini column for DNA purification. Centrifugation was performed at 8000 g for 30 seconds to bind DNA to the column matrix. The DNA column was stored at 4°C for subsequent processing.

For RNA purification, the residual flow-through from the DNA extraction step was mixed with 50 µL of Proteinase K and 200 µL of 100% ethanol, thoroughly mixed and incubated for 10 minutes. After these 10 minutes, 400 µL of 100% ethanol were added and applied to a RNeasy Mini column. The column was washed with 350 µL of Buffer RW1 and treated with a DNase I and Buffer RDD mixture, incubated at room temperature for 15 minutes. Subsequently, the column was washed again with Buffer RW1 and proceeded with further washes and RNA elution steps as per the manufacturer instructions. In certain cases, DNase I treatment was omitted, or RNA purification was conducted using magnetic beads-based purification, performed by the Genomics Facility at CIPF.

For DNA purification, 350 µL of Buffer AW1 were added to the AllPrep DNA Mini column and centrifuged for 15 seconds at 8000 x g. Then, 20 µL of Proteinase K mixed with 60 µL of Buffer AW1 (per sample) were added, mixed gently, and incubated for 5 minutes at room temperature. Next, 350 µL of Buffer AW1 were added to the AllPrep DNA Mini column and centrifuged for 15 seconds at maximum speed. The eluate was discarded, and the column was washed with 500 µL of Buffer AW2, followed by a 2-minute centrifugation. Subsequently, the column was transferred to a new collection tube, 50 µL of Buffer EB were added, incubated for 1 minute, and centrifuged to elute the DNA.

2.2.3 Promega DNA/RNA Extraction

For the RNA extraction using the ReliaPrep™ RNA Tissue Miniprep System from Promega, the procedure followed the manufacturer's instructions. Since this protocol is designed specifically for RNA extraction, we followed vendor’s recommendations and omitted the DNase treatment from the protocol, conducting the RNA purification post-extraction, either using DNase treatment of the eluate, or magnetic bead-based purification.

Tissue samples were homogenized according to the protocol mentioned above (500 µL of LBA + Thioglycerol Buffer and 170 µL of 100% Isopropanol). After brief centrifugation (15 seconds at maximum speed) to collect the supernatant, it was transferred to a new tube and isopropanol was added for RNA precipitation.

The resulting lysate was then applied to a ReliaPrep™ column and centrifuged at 12,000 x g for 1 minute at room temperature.

The column was washed successively with 500 µL of RNA Wash Solution followed by 200 µL of Column Wash Solution, each with corresponding centrifugation steps. A final wash with 500 µL of RNA Wash Solution was conducted before transferring the column to a new collection tube. Here, 300 µL of RNA Wash Solution was added and centrifuged at maximum speed for 2 minutes.

For total nucleic acid elution, the column was placed in an elution tube and RNase-free water was added directly to the membrane, followed by centrifugation at 12,000 x g for 1 minute. Purified DNA/RNA was stored at -80°C.

To remove residual DNA and obtain purified RNA, two aliquots of the eluted RNA were separated, and DNase treatment or magnetic-bead based purification was applied to one of the aliquots. DNase treatment was conducted using 12.5 µL of total nucleic acid solution, 2.5 µL of Buffer RDD, 0.625 µL of DNase I stock solution, 0.625 µL of Qiagen RNase protector, 1.25 µL of Superase RNase Inhibitor, and 7.5 µL of Nuclease Free Water per sample. For some samples (see supplementary table S2) RNA purification with beads was performed by genomics service.

2.2.4 NZYTech DNA/RNA Extraction

For the simultaneous extraction of DNA and RNA using the NZY DNA/RNA/Protein Isolation kit, the following procedure was followed based on the manufacturer's instructions:

350 µL of Buffer NDRPL and 3.5 µL of β-mercaptoethanol were added to the pathogen lysis tubes containing the beads and DX Reagent. The samples were then homogenized using a Bead Beater with the settings mentioned above. The lysate was then transferred to an NZYSpin filtration column to reduce viscosity and clarify the lysate through centrifugation at 11,000 x g for 1 minute.

Subsequently, 350 µL of 70% ethanol was added to the homogenized lysate and thoroughly mixed to facilitate binding to the column material. The lysate was loaded onto an NZYSpin DRP column designed for the simultaneous binding of DNA and RNA, followed by centrifugation at 11,000 x g for 30 seconds. The columns were washed twice with 500 µL of Buffer NDW, centrifuging at 11,000 x g for 1 minute each time. After washing, the columns were left open for 3 minutes to allow the residual ethanol to evaporate.

DNA was then eluted by adding 100 µL of Buffer NDE directly onto the membrane, incubating for 1 to 5 minutes to ensure efficient elution, followed by centrifugation at 11,000 x g for 1 minute.

To remove residual DNA from the sample, a reaction mixture of rDNase was prepared by reconstituting 10 µL of rDNase with 90 µL of Digestion Buffer. This mixture (95 µL) was applied to the membrane and incubated at room temperature for 15 minutes to digest any remaining DNA bound to the column matrix.

The membrane was subsequently washed with 200 µL of Buffer NDRPW1 and centrifuged, followed by two additional washes with 600 µL and 250 µL of Buffer NDRPW2, respectively, to ensure thorough removal of contaminants.

Finally, RNA was eluted from the membrane using 60 µL of RNase-free water, followed by centrifugation at 11,000 x g for 1 minute. Protein purification was not performed as it was not the objective of this study.

The entire DNA and RNA purification process was performed using a single column, first eluting DNA and subsequently eluting RNA after DNase treatment.

2.2.5 Macherey-Nagel DNA/RNA Extraction

For the extraction of DNA and RNA, the NucleoSpin® TriPrep kit was used with some adaptations. The tissue was lysed by adding 350 µL of Buffer RP1 and 3.5 µL of β-mercaptoethanol to the Pathogen Lysis Tubes containing the beads and DX Reagent. The samples were then homogenized using a Bead Beater using the aforementioned settings. The lysate was then filtered and centrifuged at 11,000 x g for 1 minute. Subsequently, 350 µL of 70% ethanol was added to the homogenized lysate and mixed well. The mixture was loaded onto a NucleoSpin® TriPrep column and centrifuged for 30 seconds at 11,000 x g.

The column was washed twice with 500 µL of Buffer DNA Wash, and then the membrane was air-dried for 3 minutes. DNA was eluted with 100 µL of Buffer DNA Elute, incubating for 1 minute and centrifuging for 1 minute at 11,000 x g.

For RNA purification, a mixture of rDNase and rDNase Reaction Buffer, following the instructions of the fabricant, was prepared and applied to the membrane, incubating at room temperature for 15 minutes. The membrane was washed with Buffer RA2 and Buffer RA3, and RNA was eluted with 60 µL of RNase-free water, centrifuging for 1 minute at 11,000 x g.

Protein purification was not performed as it was not the focus of the study. The eluates of DNA and RNA were stored at -20°C and -80°C, respectively, for further analysis.

The entire DNA and RNA purification procedure was performed using a single column, eluting DNA first and then, after DNase treatment, eluting RNA.

2.3 Evaluation of the concentration and quality of DNA and RNA

The concentration of DNA and RNA was measured using NanoDrop and Qubit. NanoDrop measurements require only 1 µL of the genetic material, which is placed into the device for absorbance measurement, providing the nucleic acid concentration and quality parameters such as the 260/280 and 260/230 ratios, of which low values may be indicative of poor nucleic acid quality or contamination with chemicals or proteins.

For measurement with the Qubit fluorometer, a 1:200 dilution of the Qubit DNA or RNA BR Reagent 200X (for DNA or RNA, respectively) is first prepared by mixing 1 µL of the reagent with 199 µL of buffer per sample to be analyzed. This constitutes the working solution of the kit. Two standards are prepared to create a calibration curve: for each, 10 µL of standard reagent is mixed with 190 µL of the working solution. For the samples, 2 µL of each sample is diluted in 198 µL of the working solution. After mixing and homogenizing on a vortex, the solutions are incubated for 3 minutes and then fluorescence is measured using the Qubit fluorometer. The calibration curve is first established using the standards, followed by the measurement of the different samples.

The quality of both DNA and RNA was assessed using the TapeStation automated electrophoresys system by the Genomics facility at CIPF. These quality metrics are reported as DIN (DNA Integrity Number) for DNA and RIN (RNA Integrity Number) for RNA. Values range from 1 to 10, with higher values indicating better quality and integrity of the genetic material.

2.4 Reverse transcription of RNA to cDNA

The PrimeScript™ RT reagent Kit (Perfect Real Time) from Takara (ID: RR037A) was used for reverse transcription, primarily following the manufacturer's instructions. First, the starting RNA amounts were normalized, using 300 ng of RNA per sample for the reverse transcription whenever possible. A Master Mix was prepared by combining 2 µL of 5X PrimeScript Buffer, 0.5 µL of PrimeScript Enzyme Mix, and 0.5 µL of Random hexamers, aiming to amplify both bacterial and host cDNA, per sample (an additional 10% volume was added to the total Master Mix to avoid pipetting errors) and completed with H_2O to reach a total volume of 10 µL. The reverse transcription was performed in a thermocycler with the following program: a single cycle of 15 minutes at 37ºC, followed by a 5-second period at 85ºC, and finally holding at 4ºC until the tubes were removed from the thermocycler.

2.5 qPCR

The kit used for qPCR was the TB Green Premix Ex Taq (Tli RNaseH Plus) from Takara (ID: RR420A). For the preparation of the Master Mix used in the qPCR, the reagents per sample were as follows: 5 µL of TB Green Premix Ex Taq (2X), 0.2 µL of Forward PCR primers (10 µM), 0.2 µL of Reverse PCR primers (10 µM), and 3.6 µL of sterile purified water, resulting in a total reaction volume of 9 µL, to which 1 µL of DNA/cDNA 20 ng/ µL template was added. Two master mixes were prepared since two different primer pairs were tested: a universal primer pair for the bacterial 16S gene and another for the mouse actin gene (ACTB), see supplementary Table S3 for the sequences of the primers used. The qPCRs for DNA and cDNA were performed on 384-well plates, with a volume of 10 µL per well. The qPCR was carried out in a Light Cycler 480 II thermocycler from Roche using the program recommended by the kit manufacturer: an initial denaturation phase at 95ºC for 30 seconds (1 cycle), followed by a PCR phase with 40 cycles of 95ºC for 5 seconds and 60ºC for 30 seconds. Subsequently, a melting phase was conducted with one cycle at 95ºC for 5 seconds, 60ºC for 1 minute, and then 95ºC. Finally, a cooling phase at 50ºC for 30 seconds (1 cycle) was performed.

For the DNA qPCR, final concentrations of the template oligonucleotide for the 16S and Actin (ACTB) DNA were used as positive controls at 10 pg/µL and 4 pg/µL. Extraction blanks and sterile purified water served as negative controls. For cDNA qPCRs, the same positive controls were used as in the DNA qPCR. Additionally, RNA was included as a negative control to check for genomic contamination in the RNA samples.

A total of three technical replicates were used for each sample, including controls, to control for dispersion in the measurement of Cp values. This approach allows for the calculation of a mean Cp value for each sample. The Cp value was calculated using the Light Cycler 480 II software. Fit Points analysis was employed with a noiseband that was automatically generated by the software.

2.6 16S rRNA amplicon sequencing

The V4 region of the 16S gene were amplified using 515F modified and 806 modified primers (Walters et al., 2016) and the 16S Metagenomic Sequencing Library Preparation kit from Illumina (ID: 15044223). The template DNA concentration was normalized to 5ng/µL as stipulated by the original protocol, resulting in adequate amplification in not all samples. Many of the samples that failed were biopsies, which contain only a small portion of bacterial biomass. As a result, the total amount of starting DNA was increased to improve the likelihood of successful amplification. For the samples that did not amplify in this second attempt, the amplification was repeated using the non-normalized sample as a template, increasing the number of samples amplified. However, there were still samples showing no signs of amplification, so a nested PCR was decided upon, using the previous PCR as the template.

Prior to sample pooling, calculations were made to normalize the PCR products of each sample to 10nM, despite the MiSeq requirements stipulating that denaturation should be done at 4nM. This was done because it was considered that with the potential calculation errors due to the differences in amplification yields (very variable averages) and the inclusion of blanks in the pool, it was possible that the pool would be too diluted for the run. With this information, the pool was made and rechecked with the Tape and Qubit, resulting in concentrations of 3.91nM on the Tape and 6.5 ng/µL on the Qubit. Given the difference between the two measurements, the nanomolarity of the pool was calculated using the formula:

$$[ng/\mu L]/(660g/mol\ *\ average)\ *10^6$$

- **[ng/µL]**: This represents the concentration of DNA in the sample, as measured by a device like a Qubit fluorometer. In this context, let's assume it is 6.5 ng/µL.

- **660 g/mol**: This is the approximate molecular weight of one base pair (bp) of double-stranded DNA. It is used to convert the mass of DNA into the number of moles.

- **average**: This typically represents the average length of the DNA amplicon in base pairs. For example, if the average length of the DNA amplicon is 1000 base pairs, it would be used in this position.

10^6: This factor converts the result into the adequate units, to obtain the concentration in nanomoles per liter (nanomolar, nM). The result of the calculation was 15 nM. Based on that, the calculations for the run were performed. The MiSeq was loaded using the 16S Metagenomic Sequencing Library Preparation kit from Illumina (ID: 15044223) at 4 pM, and PhiX was added at 15%. The sequencing run was configured to obtain reads of 300 base pairs in length.

Library preparation was performed by the Genomics Facility at the CIPF.

2.7 Data Analysis

The data obtained after applying the mentioned techniques were analyzed using various R packages in R version 4.4.1 (R Core Team, 2024). Specifically, the libraries tidyverse (Wickham, et al., 2019), tidyr (Wickham, et al., 2024), ggplot2 (Wickham, 2016), ggpubr (Kassambara, 2023), readr (Wickham, et al., 2024), and dplyr (Wickham, et al., 2023) were used for creating graphics.

The analysis of sequencing data was carried out using the DADA2 (Callahan, et al., 2016) package in R, executed on the Computing Cluster of the Príncipe Felipe Research Center (CIPF). Sample preprocessing included quality control, chimera removal, and taxonomic assignment based on the GTDB v.202 database. For the trimming process in DADA2, the parameters of trimming 10 bases from the start and 100 bases from the end of the reads were used, without changing other default parameters.

For subsequent analyses, such as abundance calculation and contaminant identification, the phyloseq (McMurdie, et al., 2013), microViz (- Barnett, et al., 2021), and decontam (Davies, et al., 2018) packages in R were used. The contaminant analysis in decontam was performed using the prevalence method. This method calculates the prevalence of all identified bacteria in the different samples. Depending on the prevalence of each bacterium in the blanks and in the actual samples, it classifies a bacterium as a contaminant if its prevalence is significantly higher in the blanks.

In addition to 16S analyses, the quality and relative quantity of DNA/RNA were evaluated. Pairwise comparisons between the methods or sample types were performed using Wilcoxon rank-sum tests. Overall comparisons across all extraction methods and/ samples were made using the Kruskal-Wallis test to determine statistical significance. Additional analyses and statistical tests included the calculation and visualization of microbial abundance facilitated by the phyloseq and microViz packages.

3. Results

3.1 Total Number of Samples Analyzed

In this study, a total of 96 samples from mice of different origins, strains, and diets were analyzed. The samples included biopsies from the distal and proximal colon, cecum, and mucosal scrapings from the distal and proximal colon, as well as extraction blanks for each DNA/RNA extraction performed. Additionally, two positive controls were incorporated: one from a pure culture of *Lacticaseibacillus rhamnosus* GG and another control was obtained from a mock community (Zymo, ID: D6305) with a known bacterial composition, including *Pseudomonas aeruginosa, Escherichia coli, Salmonella enterica, Lactobacillus fermentum, Enterococcus faecalis, Staphylococcus aureus, Listeria monocytogenes, and Bacillus subtilis*. It is important to note that the DNA of this mock community was purchased directly, rather than extracting it from the bacterial community.

The distribution of these samples is presented in Figure 4, which shows the total number of samples separated by sample type and extraction kit used.

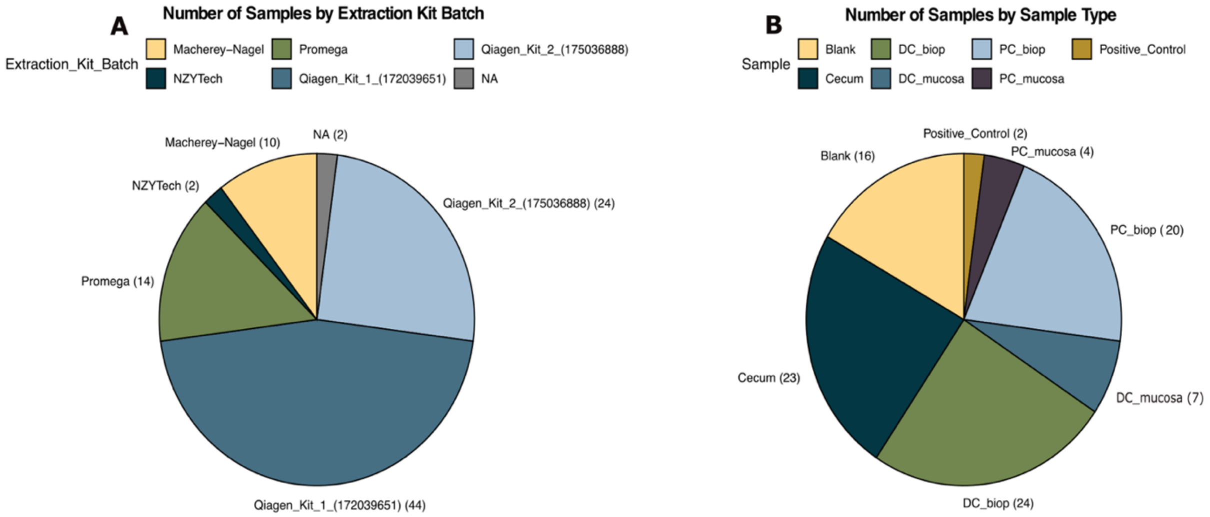


Figure 4: Total number of samples organized by Extraction Kit batch (A) on the left and by sample type on the right (B). Own elaboration.

The distribution of the 96 samples was organized according to the extraction kits used. For Qiagen, 2 different batches of the same kit were used (Figure 4A); 44 samples were processed with batch 1 and 24 with batch 2 of the same kit. Ten extractions were performed with the Macherey-Nagel kit, 2 with the NZYTech kit, and 14 with the Promega kit. Additionally, two positive controls were included, a mock community (only DNA) provided by Zymo, which was extracted using an unknown kit (indicated as NA in Figure 4A) and DNA extracted from a pure culture of *L.rhamnosus* using a modification of the ReliaPrep™ gDNA Tissue Miniprep

System protocol (ID: TM345) performed by Laura Sola, an international PhD Student in this research group.

Regarding the distribution by sample type as it can be seen in Figure 4B, 24 samples were obtained from distal colon biopsies (DC_biop) and 7 from distal colon mucosal scrapings (DC_mucosa). Twenty samples were obtained from proximal colon biopsies (PC_biop), while 4 samples came from proximal colon mucosal scrapings (PC_mucosa). Furthermore, 23 cecum samples were analyzed, and 16 blanks were included, making a total of 96 samples, including the 2 positive controls.

3.2 Comparison of Different Sample Types

A comparison of RNA quality, concentration, and bacterial quantity was performed for the different types of samples used: cecum, biopsies, and mucosal scrapings. The results are visible in Figure 5.

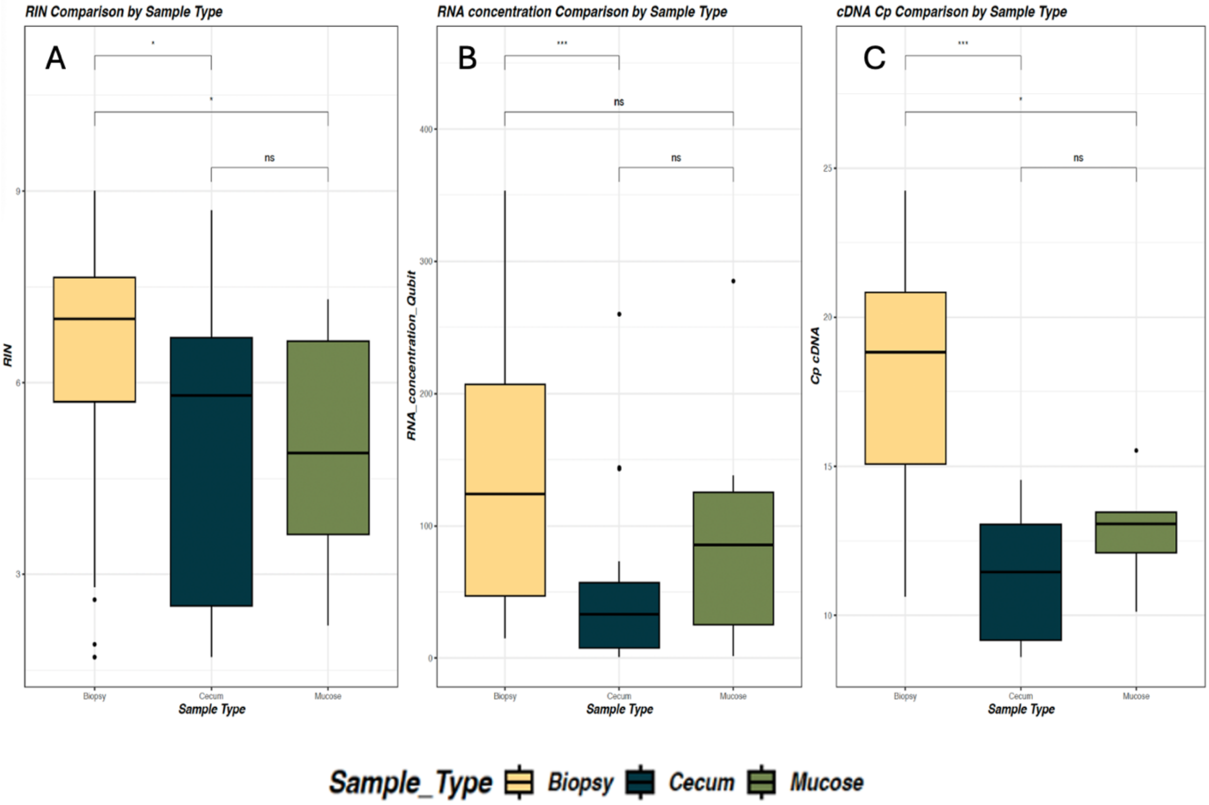


Figure 5: Comparison of RNA Quality, Concentration, and Quantity by Sample Type. This figure displays the comparison of RNA quality (**A**), RNA concentration (**B**), and RNA quantity (**C**) across different sample types: Biopsy, Cecum, and Mucosa. Significant differences, calculated using Wilcoxon rank-sum test, are indicated by asterisks (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). Own elaboration.

Significant differences in RNA quality (Figure 5.A) were observed when comparing the RIN values obtained from biopsies with those from cecum samples and mucosal scrapings, with biopsy samples showing a higher average RIN ($p < 0.05$ Wilcoxon rank-sum). The RNA quality is significantly higher in colon biopsies compared to cecum samples or colon mucosal scrapings due to the inclusion of a higher proportion of intact and well-preserved host tissue in the biopsies. Additionally, the biopsy collection process minimizes the exposure of RNA to degradative factors present in the intestinal lumen, thereby preserving the integrity and stability of the extracted RNA. Regarding RNA concentration (Figure 5.B), significant differences were found between biopsies and cecum samples, with biopsies presenting a higher RNA concentration ($p < 0.001$ Wilcoxon rank-sum). This is because biopsies contain a greater amount of host RNA compared to other samples, resulting in a significantly higher RNA concentration. However, in terms of bacterial RNA relative abundance (Figure 5.C), biopsies showed significantly less RNA compared to cecum samples ($p < 0.001$ Wilcoxon rank-sum) and mucosal scrapings ($p < 0.05$ Wilcoxon rank-sum) for the same reasons; the higher amount of host RNA dilutes the bacterial RNA content. The Wilcoxon rank-sum test was used to compare between sample types and determine statistical significance.

3.2.1 Comparison Between Mucosal Scrapings and Intestinal Biopsies

Within the biopsies and mucosa, differences between the proximal and distal colon in terms of RNA quality, concentration, and quantity have been studied. These differences can be observed in Figure 6.

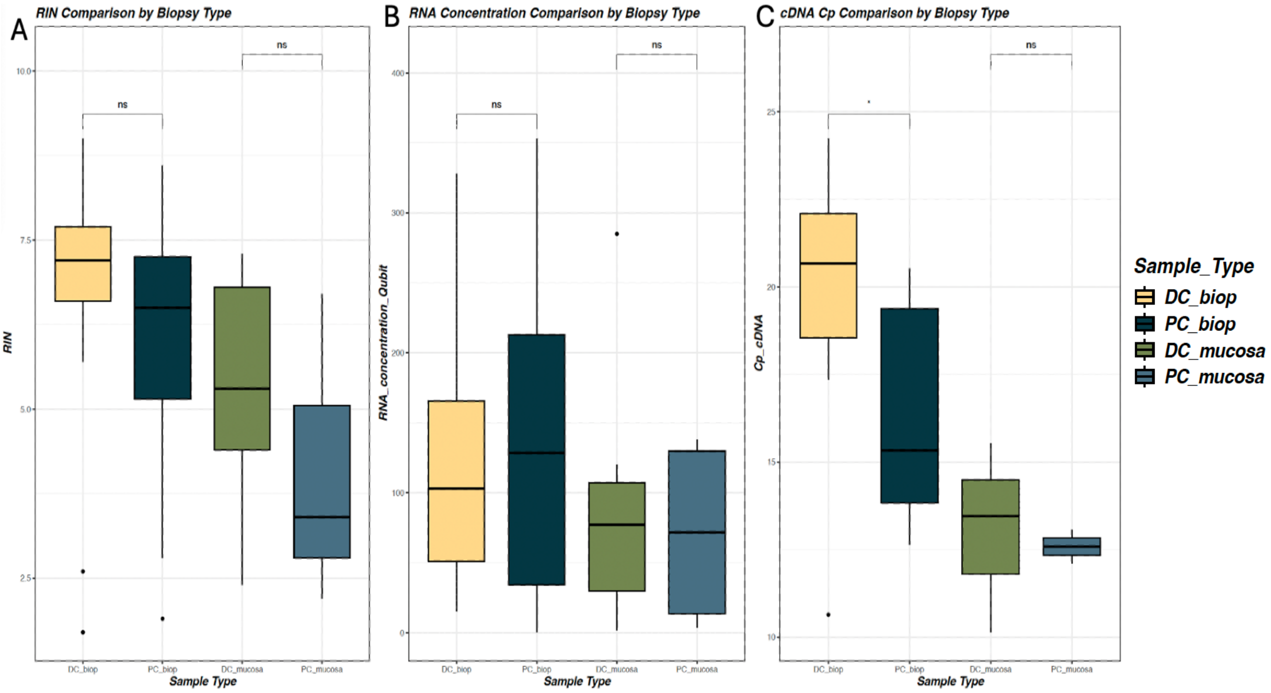


Figure 6: Comparison of RNA by biopsy type. The chart illustrates the statistical differences between RNA extractions performed on samples from various sources, specifically distal colon biopsy and mucosal scrap, as well as proximal colon biopsy and mucosal scrap. The comparisons are made in terms of RNA quality (**A**), RNA concentration (**B**), and RNA amount based on 16S qPCR (**C**). Statistically significant differences, calculated using Wilcoxon rank-sum test, are indicated by asterisks (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). Own elaboration.

RIN values were compared in Figure 6.A among the different types of samples to assess RNA quality. The analysis indicated that there were no statistically significant differences in RNA quality between distal colon biopsy samples and proximal colon biopsy samples, and the same occurs with distal colon and proximal colon mucosal scrapings. This suggests that RNA quality remains consistent regardless of sample type and origin. Nevertheless, a trend is observed. It seems that distal colon samples have higher quality than proximal colon samples. RNA concentration (Figure 6.B), measured in terms of quantification by Qubit, did not show significant differences between sample types. The relative amount of bacterial RNA (Figure 6.C), evaluated through the quantification of Cp from cDNA, showed significant differences between colon biopsies. Proximal colon biopsy samples had a significantly higher amount of bacterial RNA compared to distal colon biopsy samples ($p < 0.05$ Wilcoxon rank-sum). However, no significant differences were found in the relative amount of bacterial RNA between proximal and distal colon mucosal scrapings. All statistical analyses were performed using the Wilcoxon rank-sum test.

3.3 Comparison of Different Extraction Kits

As for the different sample types, a comparison of the quality, concentration, and bacterial quantity of RNA extracted with the different extraction kits was performed. These results are shown in Figure 7.

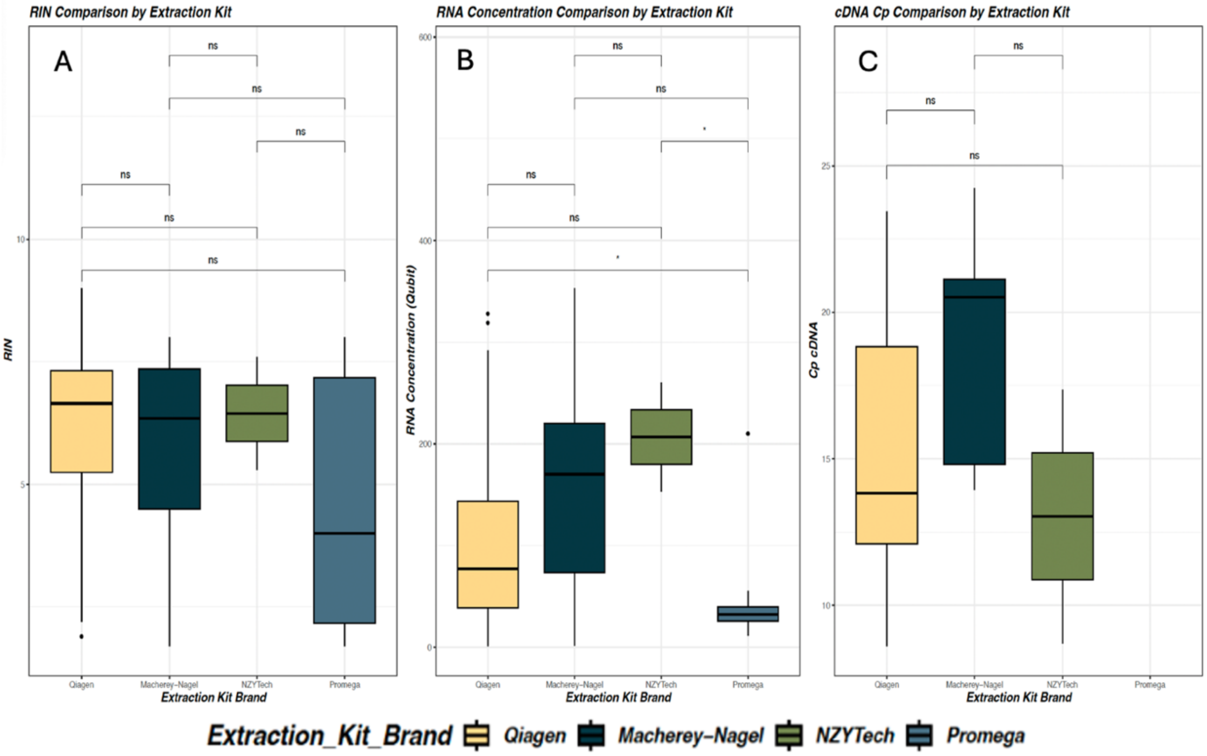


Figure 7: Comparison of RNA Extraction kits. The chart shows the comparison of different commercial DNA/RNA extraction kits methods in terms of RNA quality (**A**), concentration (**B**) and amount of RNA based on 16S qPCR (**C**). The evaluated methods include Qiagen extraction kit, Macherey-nagel extraction kit, NZYTech extraction kit and Promega extraction kit. Statistically significant differences are indicated with asterisks (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$) using Wilcoxon rank-sum test. Own elaboration.

It has been observed that there are no significant differences in the quality of RNA obtained (Figure 7.A), regardless of the extraction kit used. However, the Promega kit shows greater variability in RNA quality compared to the other kits, probably because of the DNase treatment performed on these samples, which affects RNA integrity (see section 3.4. Comparison of different RNA purification methods). In terms of RNA concentration obtained (Figure 7.B), significant differences were found in the Qiagen-Promega and NZYTech-Promega comparisons, with the Promega kit showing a lower RNA concentration in both cases ($p < 0.05$ Wilcoxon rank-sum). The Wilcoxon rank-sum test was used to determine statistical significance. No significant differences were observed in the amount of bacterial cDNA in any of the comparisons made (Figure 7.C). However, extractions performed with the Macherey-Nagel kit tended to show higher Cp values compared to the other kits which is indicative of a lower relative amount of bacterial cDNA. It is worth noting that no qPCRs were performed on the samples extracted with the Promega kit due to the high number of samples with low quality and RNA concentration. The Wilcoxon rank-sum test was used to compare between methods and determine statistical significance.

Although an effort was made to randomize the extractions across sample types and kits, different kits were applied to samples obtained in different days and from different mice. Therefore, to ensure that the differences in RNA (but also DNA) concentration obtained are not due to variations in the weight of the samples, thus confounding the effect of the kit, an additional analysis was conducted. Below is a scatter plot (Figure 8) showing the relationship between the sample weight and the extraction kit yield, for both RNA and DNA.

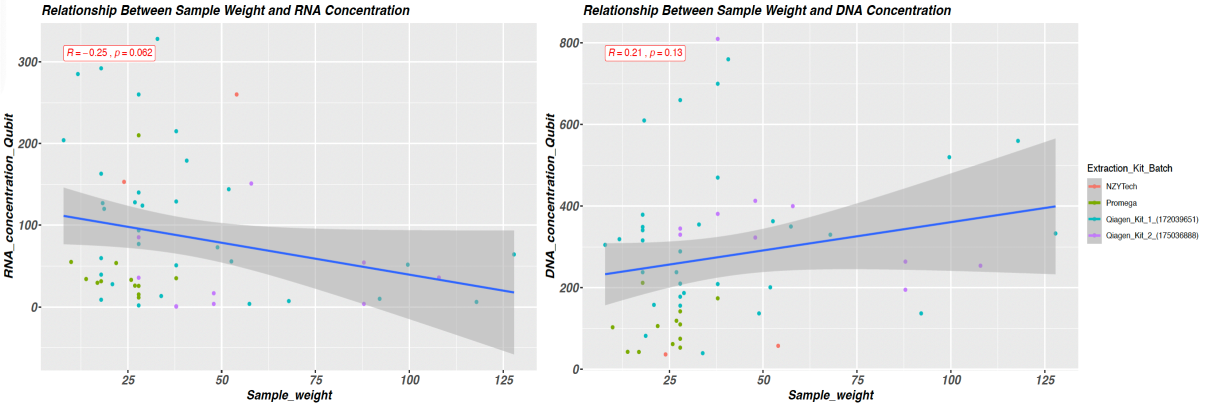


Figure 8: Scatter Plot Comparing RNA and DNA concentration and Sample Weight. This Figure illustrates the linear relationship between sample weight and the concentrations of RNA and DNA obtained using Qubit. The gray area represents the 95% confidence interval. Each sample is color-coded to differentiate between the various extraction kits.

In Figure 8, it can be seen that no linear correlation exists between RNA (R RNA = -0.25) and DNA (R DNA = 0.21) concentration and the sample weight, demonstrating that the differences in RNA/DNA concentration are not due to differences in sample mass. It is important to note that samples extracted using the Macherey-Nagel kit are not represented due to a measurement error. The samples were weighted before bead beating, making it impossible to determine the original weight of the sample due to the irregular weight of the different components (buffer, beads, etc.).

3.4 Comparison of Different RNA Purification Methods

One of the aims of this work was to compare how different RNA purification methods affected its concentration, quality, and bacterial quantity compared to not purifying the RNA during extraction. The results of these comparisons are shown in Figure 9.

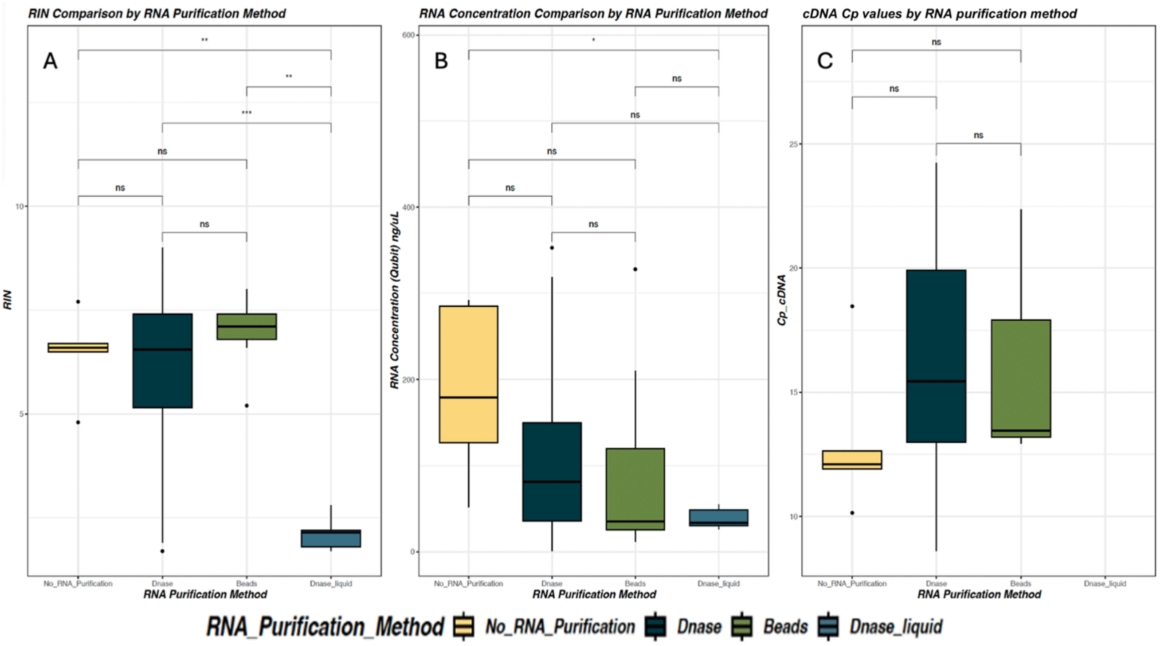


Figure 9: Comparison of RNA Purification Methods. The chart shows the comparison of different RNA purification methods in terms of RNA quality (**A**), concentration (**B**) and amount of RNA based on 16S qPCR (**C**). The evaluated methods include no additional purification (No RNA Purification), DNase treatment (column), bead-based purification, and liquid DNase treatment. Statistically significant differences, using Wilcoxon rank-sum test, are indicated with asterisks (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). Own elaboration.

It was observed that there were no significant differences in RNA quality (measured by RIN) regardless of the RNA purification method used (Figure 9.A), except in the case of liquid DNase treatment performed after the extraction procedure, which presents significantly lower quality compared to the other methods. We speculate that this is probably due to the heat inactivation step of the enzyme. This heat inactivation would compromise the RNA structure by degrading it, thereby decreasing its quality and concentration. Alternatively, this could be due to the DNase I degrading also the RNA in these samples.

Regarding the RNA concentration obtained (Figure 9.B), although no significant differences were detected between most methods, there is a trend suggesting higher RNA concentrations in the samples without purification, followed by those treated with DNase in column, beads, and DNase in liquid. This trend is likely due to DNA contamination in the unpurified samples, which might be detected by the measurement device. the RNA concentration obtained through liquid DNase treatment is significantly lower than that obtained without RNA purification. Similarly to the effect in RNA quality, this significant decrease in RNA concentration with the liquid DNase treatment is likely due to the enzyme inactivation with heat, which may degrade the RNA. These results suggest that while unpurified samples may show artificially high RNA concentrations due to DNA contamination, the method involving DNase in liquid, despite effectively removing DNA, may also result in RNA degradation, thereby lowering the overall RNA yield.

Finally, no significant differences were observed in the relative amount of bacterial cDNA, measured by Cp, between the different methods (Figure 9.C). Nonetheless, it is noted that the absence of RNA purification tends to result in lower Cp values, which indicates a high amount of bacterial cDNA in the sample. This can be explained by the lack of a method that could damage RNA. No qPCRs were performed on samples purified using liquid DNase.

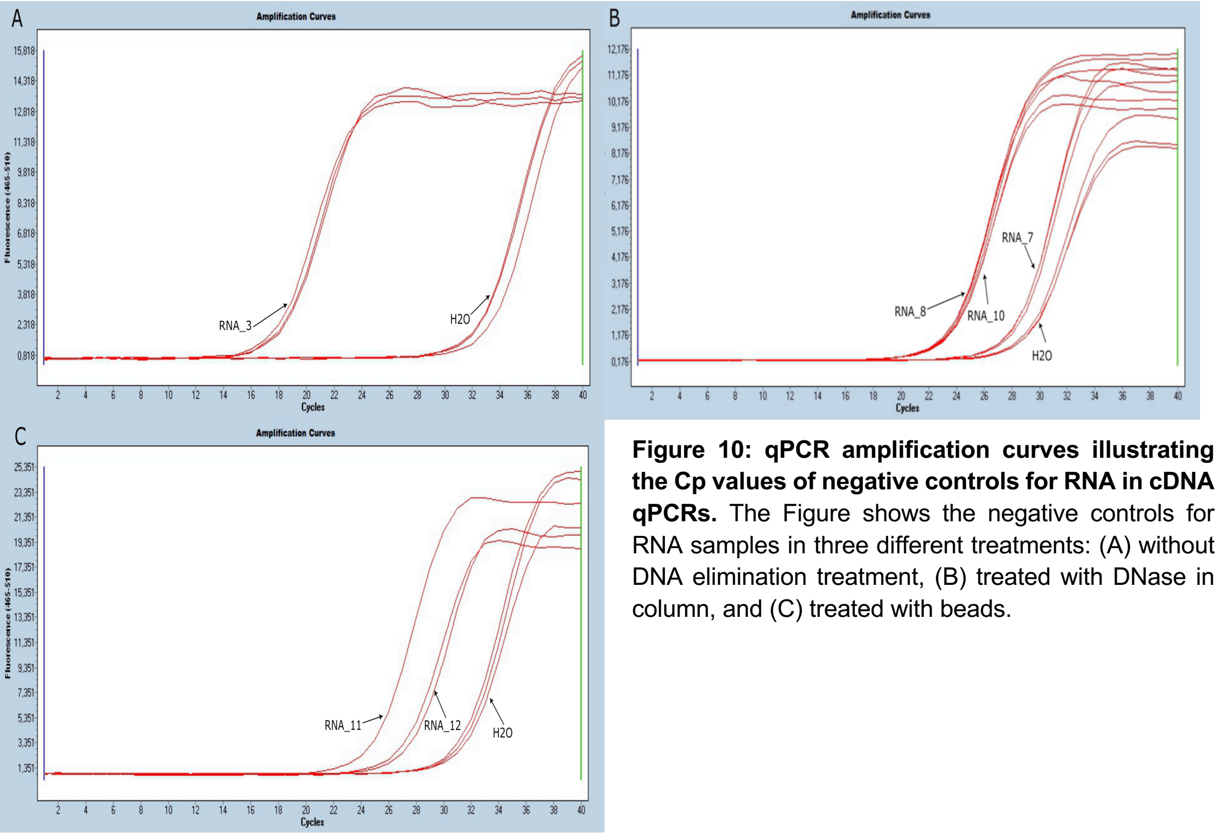


Figure 10: qPCR amplification curves illustrating the Cp values of negative controls for RNA in cDNA qPCRs. The Figure shows the negative controls for RNA samples in three different treatments: (A) without DNA elimination treatment, (B) treated with DNase in column, and (C) treated with beads.

DNase treatments are required due to DNA contamination in RNA samples. This is shown in Figure 10, where it can be seen that in qPCRs of negative controls of RNA, there is DNA contamination in the samples where no RNA purification was done, indicated by a low Cp value, which signifies the presence of genomic DNA in the sample. In Figure 10.A, the Cp values of RNA samples without DNA elimination treatment are represented. In Figure 10.B, the Cp values of RNA samples treated with DNase in column are shown. Finally, in Figure 10.C, the Cp values of RNA samples treated with beads are depicted. Although the Cp values from different qPCR assays should not be directly compared, the Cp of water is shown as a reference to facilitate comparison of the results. This demonstrates that to avoid genomic contamination in RNA samples, it is necessary to perform DNA elimination treatments during extraction, such as column DNase treatment or bead-based treatment.

In contrast, in the qPCRs of the RNA obtained with samples processed with different RNA purification methods, such as DNase in column and beads, this DNA contamination is lower. Since the different purification methods were performed using two specific extraction kits, Figure 11 represents in greater detail the difference between the purification methods stratified by the extraction kit used, either the kit from Qiagen or the one from Promega.

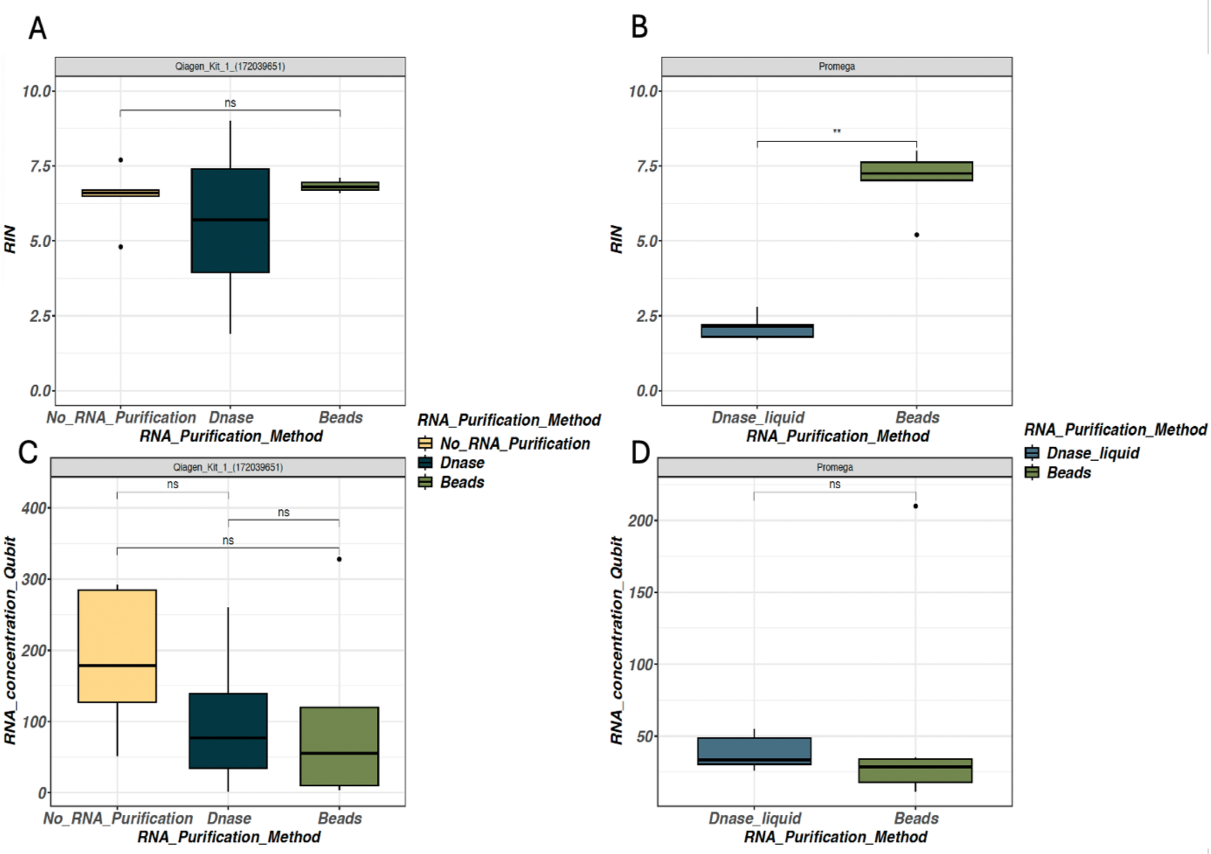


Figure 11: Comparison of purification methods separated by extraction kit. The chart shows RNA quality filtered by extraction kit, Qiagen (**A**) and Promega (**B**), and RNA concentration filtered by extraction kit, Qiagen (**C**) and Promega (**D**). Statistically significant differences are indicated with asterisks (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). Own elaboration.

No significant differences in RNA quality were found with any purification method when using the Qiagen kit. (Figure 11.A). Significant differences were observed in RNA quality only when using the Promega kit and the bead method versus treatment with liquid DNase (Figure 11.B). These differences are because treating RNA with DNase in liquid solution, and then inactivating the DNase with heat, compromises the integrity of the RNA, reducing its quality.

This phenomenon is specific to this kit, as it is the only one of the four evaluated that is not designed for DNA extraction, rendering the purification necessary after the extraction procedure. Furthermore, no significant differences were found in RNA concentration regardless of the purification method used, although we observed a trend showing a decrease in total RNA concentration in the purified samples compared to the non-purified ones in the Qiagen kit (Figure 11.C and 11.D).

3.5 Analysis of the 16S Sequencing Results

16S amplicon sequencing was conducted to verify whether the levels of bacterial DNA/RNA measured with PCRs are reflected in the 16S sequencing results, which in this study are used as a proxy for potential metatranscriptomic outcomes. In this way, 16S results can be considered a cost-effective approach to guide sample selection for metatranscriptomic analysis, allowing to select only samples with enough bacterial biomass to warrant enough bacterial functional coverage. Additionally, this sequencing aimed to identify potential contaminants in the various extraction kits.

The sequencing data from the 96 pilot samples, including extraction blanks for identifying environmental and kit contaminants, were analyzed. The raw results of the taxonomic analysis and the abundance of bacteria in each sample can be observed in Figure 12. This Figure shows the absolute bacterial abundance (in number of reads sequenced and assigned to bacterial taxa) and indicates the 10 most abundant bacteria in all samples, organized by type of sample, including distal and proximal colon biopsies, distal and proximal colon mucosal scrapings, cecum samples, blanks, and positive controls (a pure culture of *Lacticaseibacillus rhamnosus* GG and a *Mock community* from Zymo).

The sequencing results showed considerable variability in the number of reads obtained among the different sample types. These differences are shown in Figure 13. In biopsy samples, a significantly lower number of reads is observed, compared to mucosal scraping samples ($p < 0.001$ Wilcoxon rank-sum) and cecum samples ($p < 0.01$ Wilcoxon rank-sum). No significant differences were found in the number of reads obtained from cecum samples and mucosal scrapings (Figure 13.A). Furthermore, significant differences were observed in the number of reads between distal colon biopsies and proximal colon biopsies ($p < 0.01$ Wilcoxon rank-sum), with proximal colon biopsies having a higher average number of reads compared to distal colon biopsies (Figure 13.B). In Figure 13.C, the total number of reads from mucosal scrapings (both distal and proximal) is shown. No significant differences were found between the mucosal scrapings from the proximal colon and those from the distal colon. The Wilcoxon rank-sum test was used to determine statistical significance.

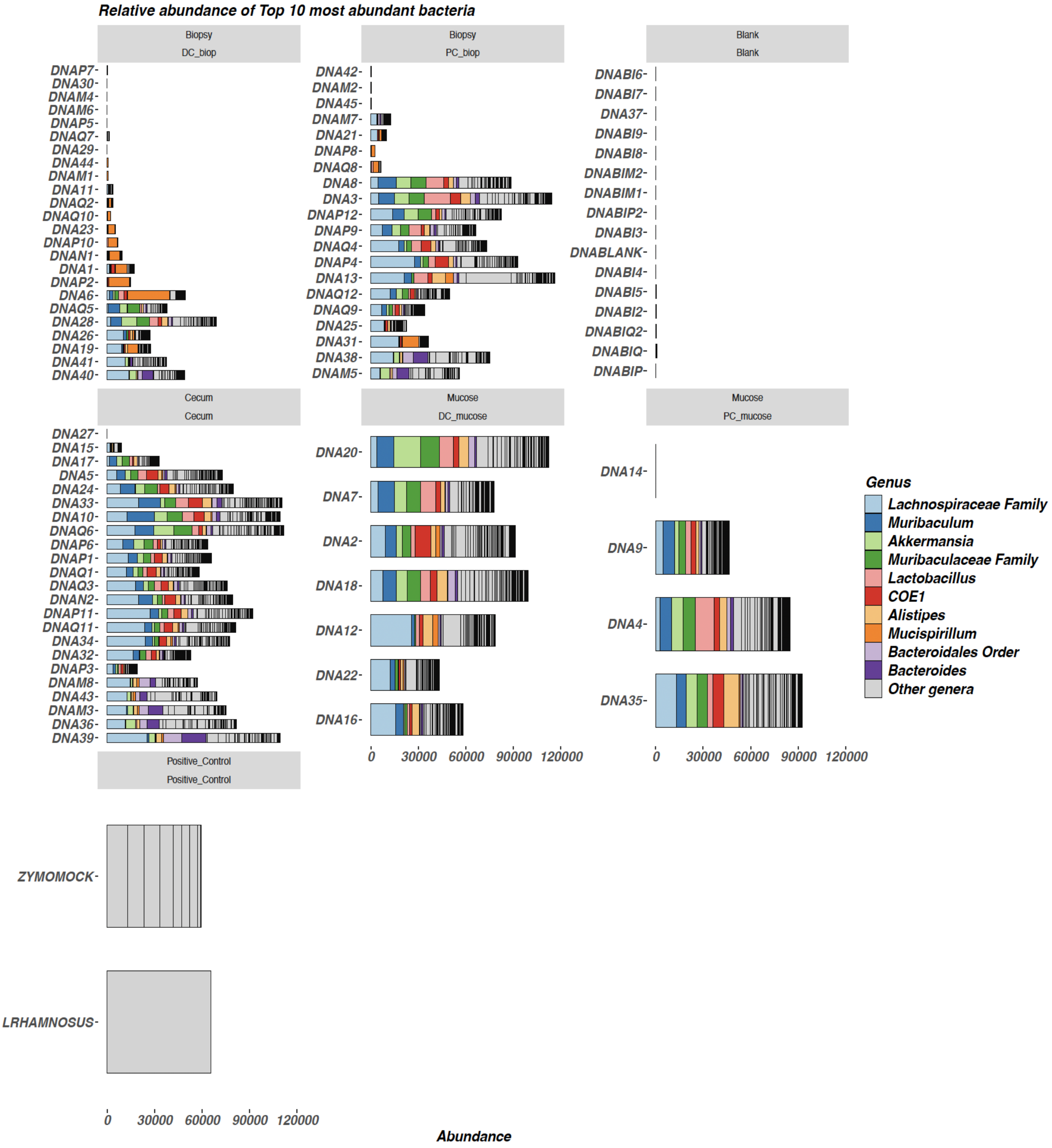


Figure 12: Top 10 Most Abundant Bacteria by DNA Sample. This chart displays the relative abundance of the 10 most prevalent bacterial genera across each of the 96 DNA samples, categorized by biopsy type. The analysis includes all blanks and positive controls. Bacteria not among the top 10 most abundant are represented in grey. Taxonomic assignment was performed using the GTDB v.202 database. Own elaboration.

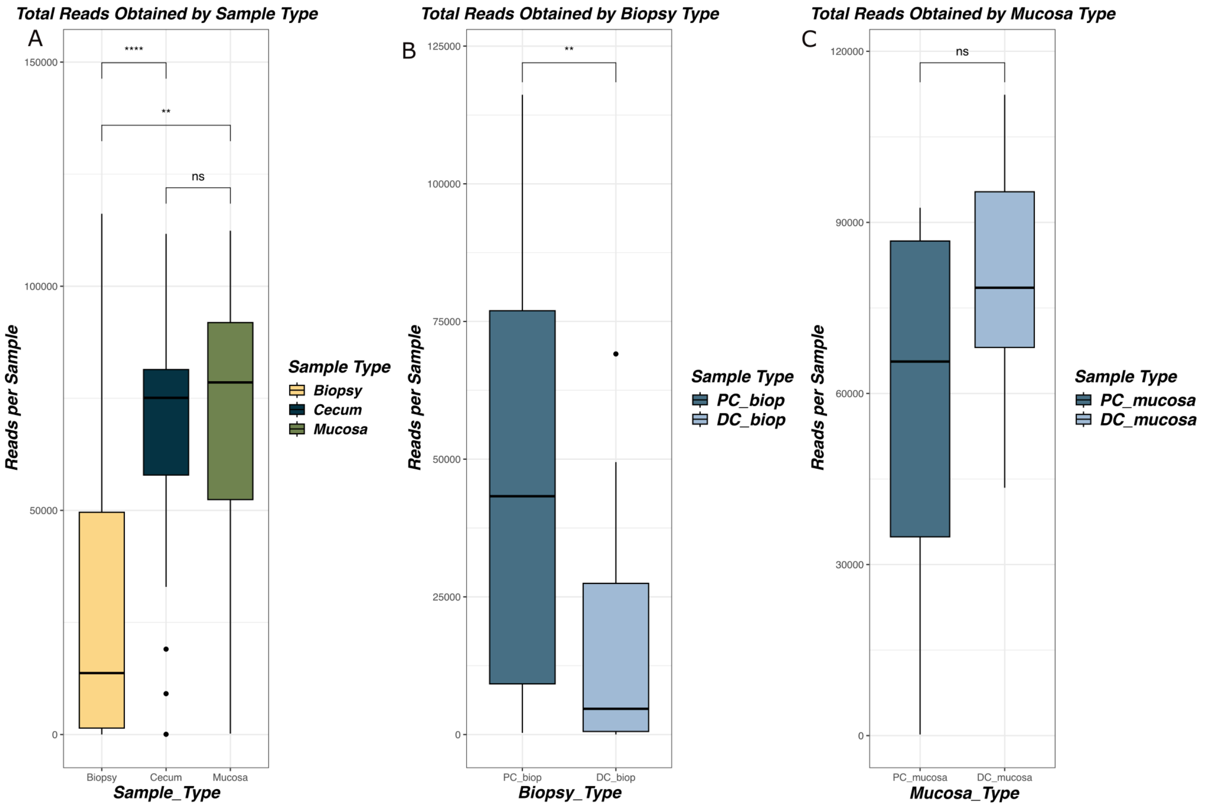


Figure 13: Total number of Reads by sample and biopsy type. This chart depicts the total number of bacterial reads, categorized by sample type (**A**), biopsy type (**B**) and mucosal scraping type (**C**). Statistically significant differences calculated using Wilcoxon rank-sum test, are indicated by asterisks (* p < 0.05, ** p < 0.01, *** p < 0.001, ****: p < 0.0001). Own elaboration.

These results are consistent with the observations made at the RNA level in the previous sections. For example, Cp values for the cDNA qPCRs in biopsies were higher than in cecum and mucosa, and correspondingly, fewer bacterial reads were observed in biopsies compared to other samples. Similarly, proximal colon biopsies had lower Cp values than distal ones, which is reflected in the higher number of reads in proximal biopsies. In the mucosa, differences in Cp values were not significant, which aligns with the similar number of reads observed here. This suggests an apparent correlation between the RNA-level observations and the DNA-level reads.

To further confirm this correlation, the relationship between the relative quantity of bacterial DNA and RNA, as measured by (RT-)qPCR, and the number of reads obtained was examined (Figure 14). In this figure, the total number of reads in each sample is represented on the Y axis, and the Cp value of each sample is represented on the X axis. Each point represents a different sample, and the shaded area represents the 95% confidence interval. Although there is variability, likely due to qPCRs being conducted in small batches on different days, a consistent linear relationship is observed: the lower the Cp value, the higher the number of reads in a sample (R Cp DNA= -0.76; R Cp cDNA = -0.81). This suggests that Cp values could potentially be used as a proxy for selecting samples for metatranscriptomics, potentially eliminating the need for the 16S step and saving costs.

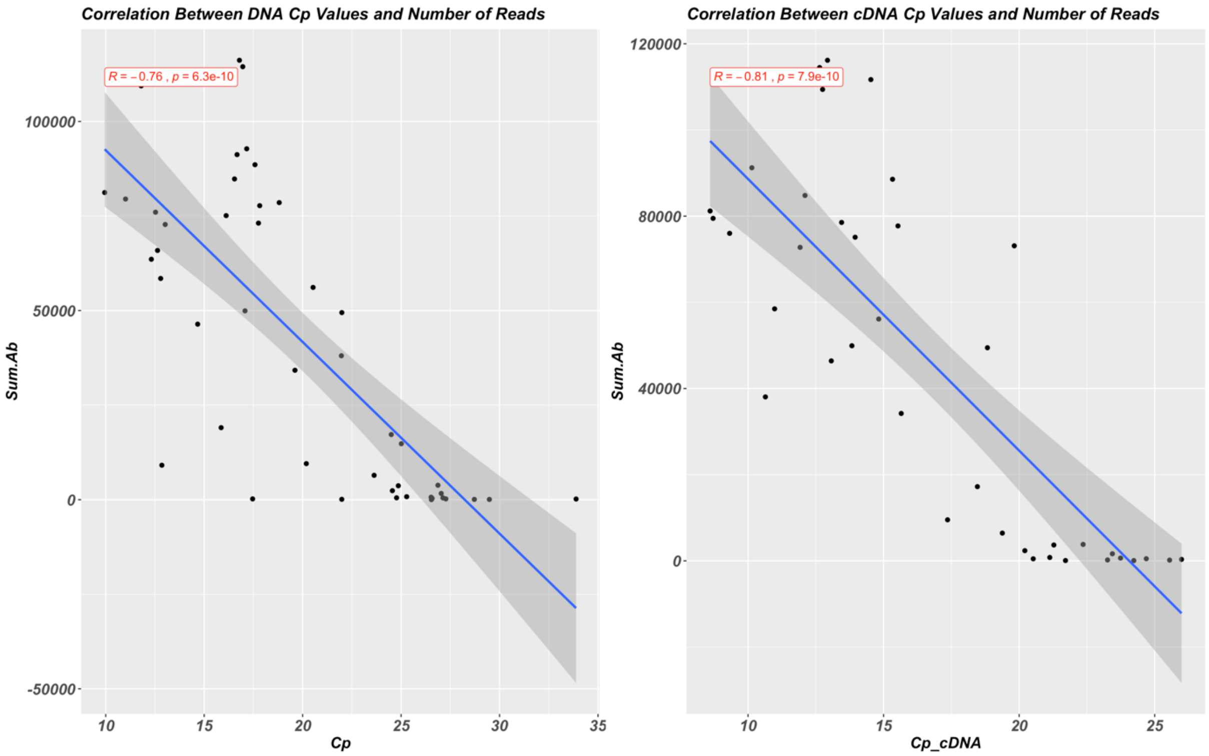


Figure 14: Scatter Plot Comparing Number of Reads per Sample with Cp and Cp cDNA from qPCR. This chart shows the linear correlation between two variables: Cp (Left) and cDNA Cp (Right) with the total number of reads obtained in the sample during sequencing. The gray area represents the 95% confidence interval. Own elaboration.

For example, the blanks have a very high Cp value (around 30-35) and few reads (<1000), while other samples, such as cecum samples or some mucosal scrapings, have a very low Cp value (10-15) and a high number of reads (>60000).

Regarding the controls, the number of reads in the blanks is low (always below 1000 reads) indicating no massive contamination is expected. Regardless, a thorough study of the potential contaminants identified was performed (see section 3.6. Identification of Environmental Contaminants). The positive controls show an acceptable (>10000 reads) number of reads, and the absolute abundances of these positive controls were examined to see if the taxonomic assignment was correct. These results are shown in Figure 15. It must be noted that the bacterial genera assigned do not match the theoretical bacteria in the community. However, this is only caused by the reference database used for the taxonomic assignation (GTDB v.202), and the assigned bacteria are indeed the ones that should be present. To confirm the identity of these bacteria, the mapped sequences were obtained, and a BLAST was performed to determine the correct species. The results indicated that *Pseudocitrobacter* was actually *Escherichia coli* and *Agrilactobacillus* corresponded to *Lacticaseibacillus rhamnosus* GG. This variability underscores the importance of sequence verification and careful database selection to ensure accuracy in bacterial identification.

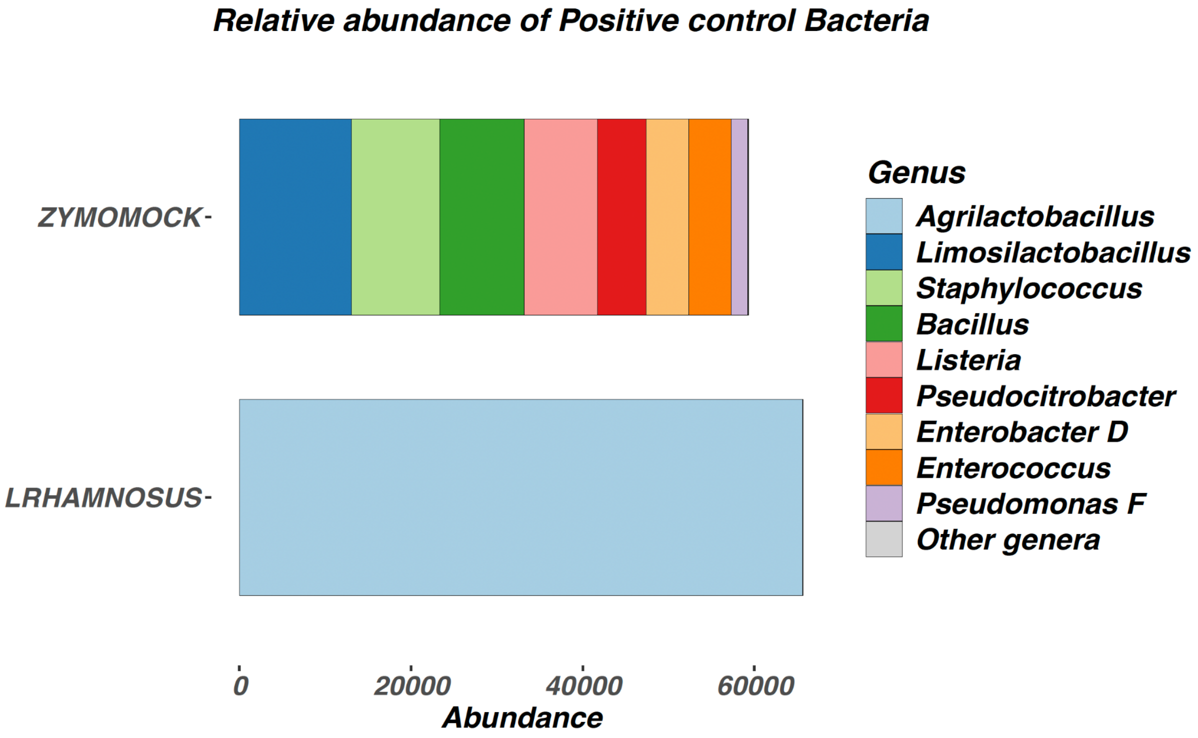


Figure 15: Absolute abundance of Positive control bacteria. This chart illustrates the relative abundance of each bacterial species present in the positive control samples. In the Zymo Mock community positive control, a diverse array of bacterial species is observed, whereas the pure culture of *Lacticaseibacillus rhamnosus* contains only a single bacterial species. Own elaboration.

The total number of reads obtained by each extraction kit for each type of sample was examined (Figure 16). Significant differences are observed in the total number of reads obtained by each extraction kit across the different types of samples, as indicated by the Kruskal-Wallis test comparing the sample types. However, assessing the variation between extraction kits within each type of sample would require a separate statistical test to determine if there are significant differences in performance among the kits for each specific sample type. Although it cannot be conclusively stated that there are significant differences between the kits, a trend is noticeable: the Qiagen kit tends to yield a higher average number of reads in both biopsies and cecal samples compared to the other kits. This trend is partly due to the larger number of samples extracted with this kit. The results from the NZYTech kit are inconclusive, as only two extractions were performed, one biopsy and one cecal sample. However, given the number of reads obtained from each of these samples, the results appear promising. As a result, the Qiagen kit tends to provide a higher number of reads compared to the other kits.

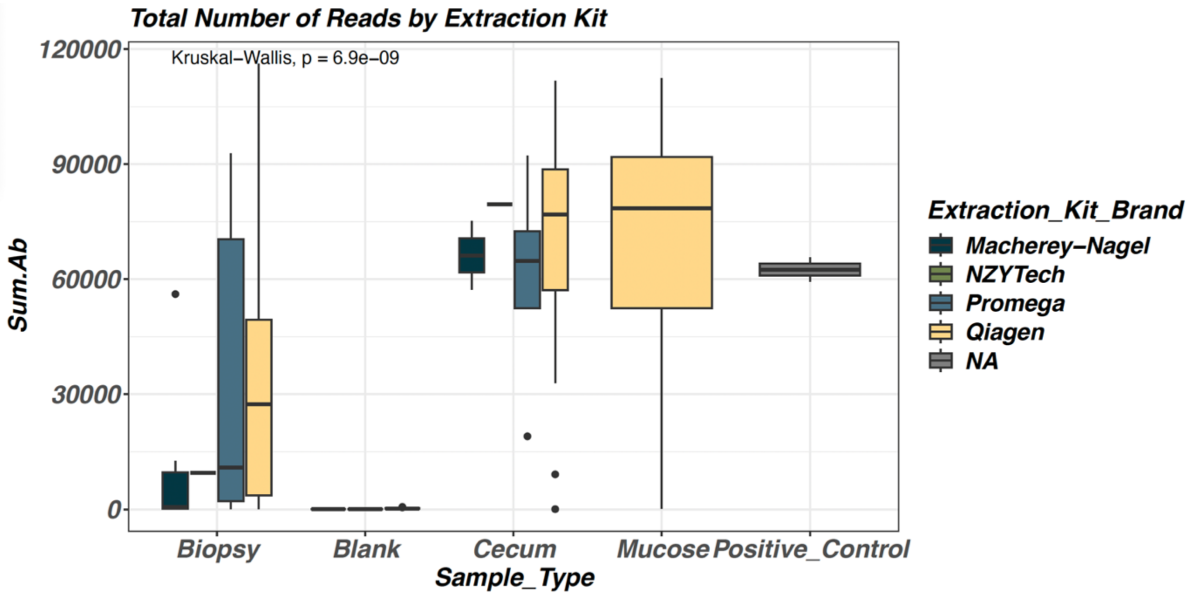


Figure 16: Total number of Reads by extraction kit and sample type. This chart highlights the total number of reads obtained by each Extraction kit in each sample type. Significative differences were calculated using Kruskal-Wallis test showing significant differences between Extraction kit and sample type. Own elaboration.

3.6 Identification of Environmental Contaminants

With the information obtained from the sequencing of the extraction blanks, it was possible to identify contaminants from the different extraction kits. For this, the R package *decontam* was used, which identifies potential contaminants by measuring the prevalence of a bacterium in samples and negative controls within a study.

Before applying *decontam*, the absolute abundance of mapped bacterial reads in the blanks was observed, stratifying by extraction kit. The number of reads obtained in the blanks was very low, indicating that, as expected, the bacterial biomass present in the kits is minimal, though not entirely absent. These contaminants could originate from the extraction kits themselves, the environment where the extraction was performed, or the operator conducting the extraction. However, given the marked differences between kits, despite the extractions being performed by the same operator and under the same conditions in the same cabinet, it is likely that the contaminants are primarily coming from the kits. As seen in Figure 17.A, the bacterial composition of the blanks within the same extraction kit is very similar, while there are remarkable differences across kits, indicating that each kit, or even each batch, seems to have specific contaminants. Therefore, samples extracted with different kits will exhibit different contaminants.

Figure 17.B shows a PCoA, where the data were normalized using a CLR (Centered Log-Ratio) transformation, and Euclidean distance was used to calculate the PCoA. Differences between the various batches of the extraction kits are clearly visible, particularly highlighting the grouping of blanks within the same batch.

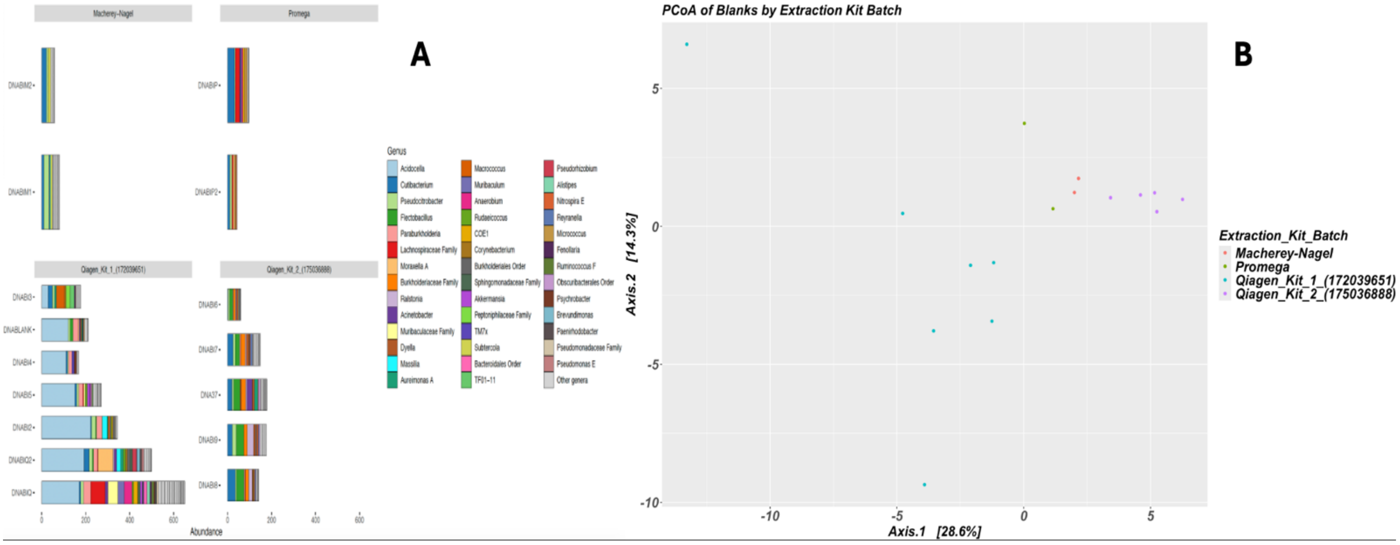


Figure 17: Relative abundance of Negative control (Blanks) bacteria (A) and PCoA grouping by Extraction kit batch (B). This chart illustrates the relative abundance of each bacterial species present in the negative control samples. The bacterial composition of the negative controls belonging to the same batch or extraction kit is very similar. Own elaboration.

As a first step in the contaminant analysis, the distribution of library sizes (the number of reads obtained in each sample) in the samples was observed. These results can be seen in Figure 18. The presented Figure shows the distribution of library sizes (Y axis), while the X axis shows the indices of the samples and blank controls, ordered from 1 to 96, depending on the total number of reads in each sample and differentiating between real samples and blanks. The variable "Sample_or_Blank" is coded as "FALSE" for samples and "TRUE" for blanks.

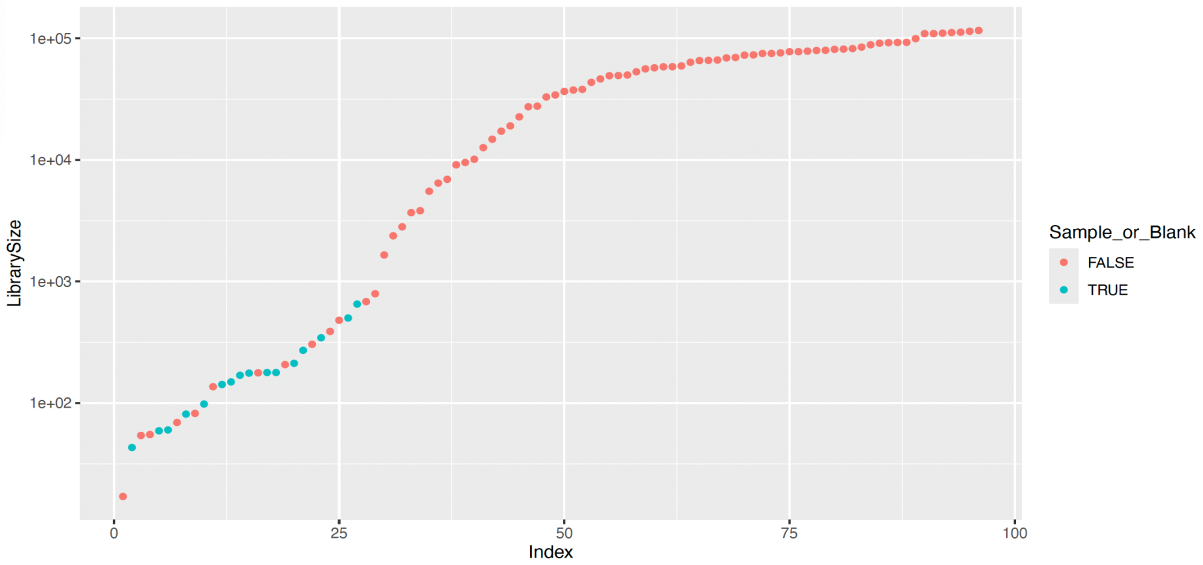


Figure 18: Library Size Distribution for Samples and Blanks. This Figure displays the library size distribution on a logarithmic scale across sample indices, differentiating between actual samples (Sample_or_Blank = FALSE), in red, and blank controls (Sample_or_Blank = TRUE), in blue. The x-axis represents the indices of samples and blanks, ordered from 0 to 100, while the y-axis shows the library size, ranging from 1e+02 to 1e+05. The results demonstrate a wide variability in library sizes for actual samples, reflecting differences in sequence yields, while

the blank controls exhibit consistently lower library sizes, indicating effective contamination control measures. Own elaboration.

The real samples (Sample_or_Blank = FALSE) show a wide variability in library size, with values ranging from 1e+02 to 1e+05, which is expected given the diverse origin of the samples. In contrast, the blank controls (Sample_or_Blank = TRUE) exhibit lower library sizes and less dispersion, which is consistent with a lower amount of genetic material in these controls. The presence of some real samples interspersed with blanks on the left side indicates that a few samples had very low biomass.

To identify bacterial contaminants and assess the extent to which they appear in true samples, particularly in biopsies where low biomass could lead to greater distortion of results, a study on the prevalence of all bacteria detected across different samples, including blanks, was conducted (Figure 19). This is crucial because the presence of contaminants in low-biomass samples like biopsies can significantly affect the accuracy and interpretation of the results.

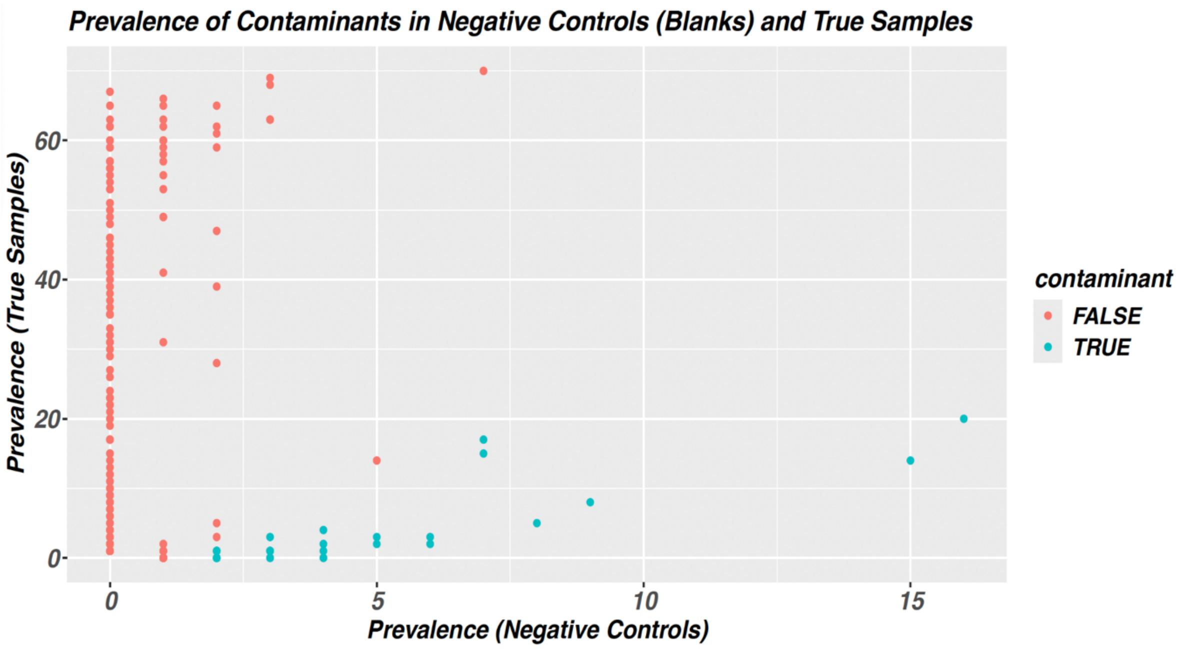
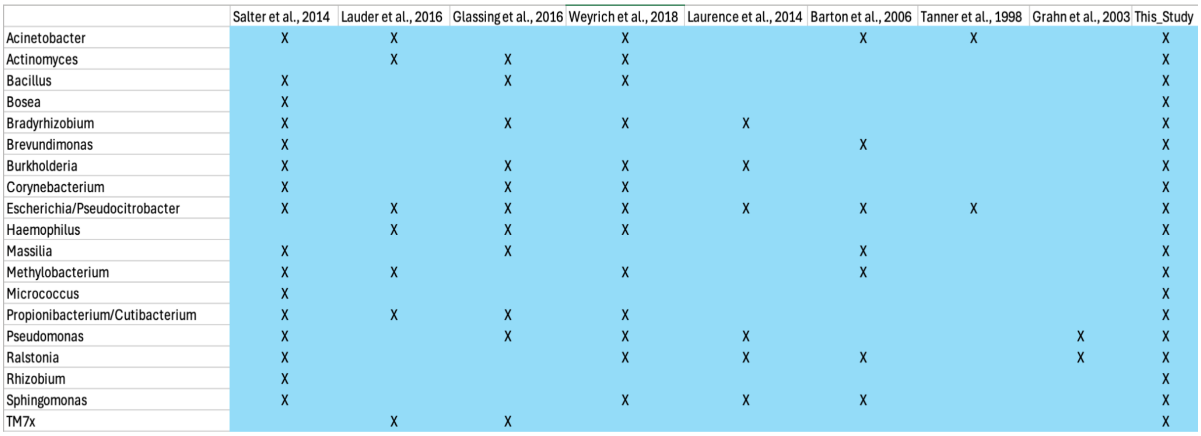


Figure 19: Prevalence of Bacteria in Negative Controls and True Samples. This Figure illustrates the relationship between the prevalence of all bacteria present in negative controls (x-axis) and their prevalence in true samples (y-axis). Data points are color-coded based on the presence of contaminants (TRUE for bacterial genera deemed as contaminants by the decontam method, FALSE for non-contaminants). A high prevalence value on the Y-axis indicates that the bacterium in question is present in many samples, while a high prevalence value on the X-axis indicates a high prevalence in blank controls, suggesting that the bacterium is a contaminant. Own elaboration.

Once contaminant bacterial genera were identified, a table was created that listed the genera of these bacteria. This table includes a total of 42 contaminant bacteria: 39 were detected by *decontam* with a higher prevalence in blanks than in actual samples (supplementary table S4) and 3 were manually added from other studies, as they were present in our samples but missed by the *decontam* method [*Actinomyces (Salter et al., 2014; Lauder et al., 2016; Weyrich et al., 2018), Haemophilus (Lauder et al., 2016; Glassing et al., 2016; Weyrich et al., 2018)* and *Methylobacterium (Salter et al., 2014; Lauder et al., 2016; Weyrich et al., 2018; Barton et al., 2006)*].

To validate if the bacteria identified by *decontam* are truly contaminants, a comparison between the contaminant bacteria listed in Table S4 and other studies has been conducted. The results of this comparison can be seen in Table 3. This table shows that out of the 39 contaminants identified in this study, 19 have also been reported in other studies.



	Salter et al., 2014	Lauder et al., 2016	Glassing et al., 2016	Weyrich et al., 2018	Laurence et al., 2014	Barton et al., 2006	Tanner et al., 1998	Grahn et al., 2003	This_Study
Acinetobacter	X	X		X		X	X		X
Actinomyces		X	X	X					X
Bacillus	X		X	X					X
Bosea	X								X
Bradyrhizobium	X		X	X	X				X
Brevundimonas	X					X			X
Burkholderia	X		X	X	X				X
Corynebacterium	X		X	X					X
Escherichia/Pseudocitrobacter	X	X	X	X	X	X	X		X
Haemophilus		X	X	X					X
Massilia	X		X			X			X
Methylobacterium	X	X		X		X			X
Micrococcus	X								X
Propionibacterium/Cutibacterium	X	X	X	X					X
Pseudomonas	X		X	X	X			X	X
Ralstonia	X			X	X	X		X	X
Rhizobium	X								X
Sphingomonas	X			X	X	X			X
TM7x		X	X						X

Table 3: Comparison of Contaminants Identified in This Study with Those Found in Other Studies. This table lists the bacterial contaminants identified in this study and compares them with those reported in previous studies. Each row represents a different bacterial genus, and each column represents a different study. The presence of a specific contaminant in a study is indicated by an "X". This comparison highlights common bacterial contaminants across various studies, providing a comprehensive overview of potential sources of contamination in microbiome research. Own elaboration.

After compiling the table and identifying the contaminants, the proportion of contaminants in each type of sample was determined. The results are shown in Figure 20.

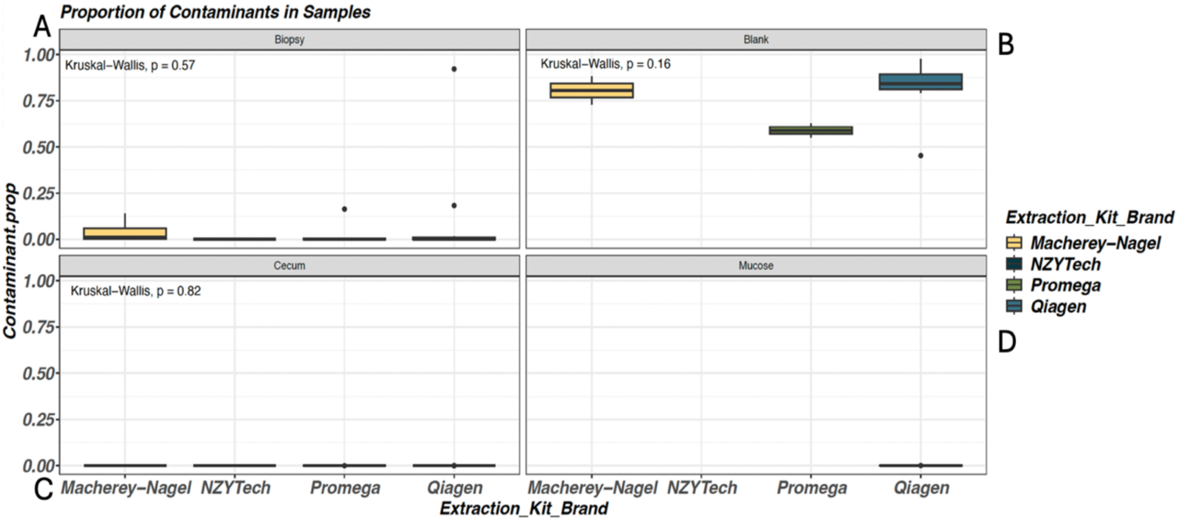


Figure 20: Proportion of contaminant bacterial Reads in each sample filtered by Extraction Kit. The boxplots illustrate the proportion of contaminants across different sample types (Biopsy, Blank, Cecum, and Mucose) and extraction kit brands (Macherey-Nagel, NZYTech, Promega, and Qiagen). Statistical significance was assessed using the Kruskal-Wallis test. Own elaboration.

Figure 20 shows the proportions of contaminant bacteria in each sample type separated by extraction kits. The Kruskal-Wallis test p-values indicate no significant differences in contaminant proportions among the extraction kit brands for each sample type.

Specifically, biopsy samples show no significant difference in contaminant proportions across the extraction kits (Figure 20.A). Similarly, blank samples also show no significant difference (Figure 20.B). Cecum samples exhibit low contaminant proportions with no significant difference (Figure 20.C). Mucosal samples demonstrate minimal contaminant proportions. Despite the lack of statistically significant differences among the different kits for biopsies and blanks, there is a noticeable trend: Macherey-Nagel generally shows a higher proportion of contaminants compared to the other kits. Additionally, it is important to note the nearly 1 value observed in one Qiagen biopsy sample, which corresponds to a sample that was lost during the process.

The relative abundance of contaminants in the different colonic locations and kit batches was further explored (Figures 21 and 22)

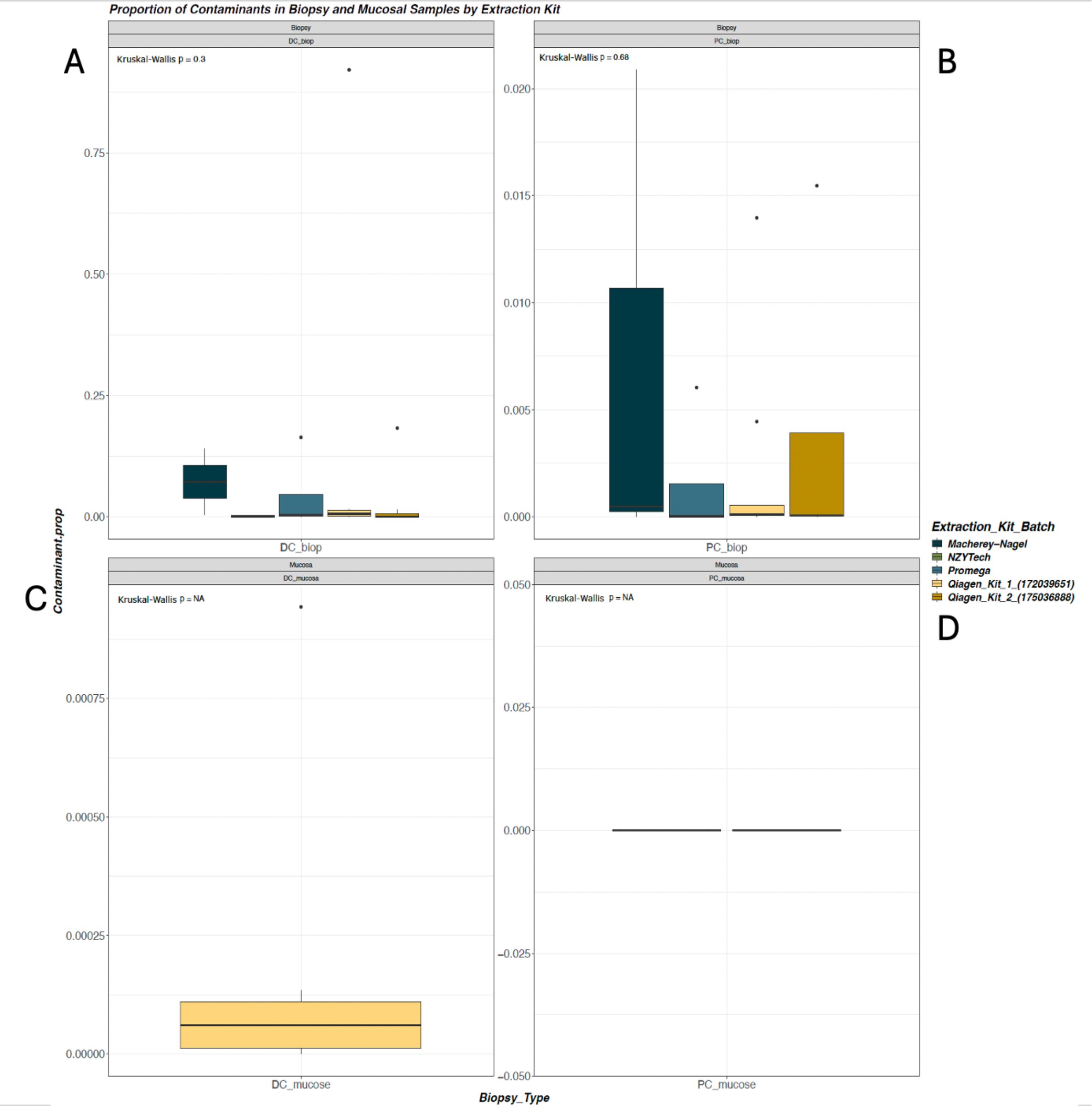


Figure 21: Proportion of contaminant bacterial Reads in each Biopsy Type filtered by Extraction Kit. The Figure illustrates the proportion of contaminants in various kits used for colon biopsies (Distal A, Proximal B) and mucosal scrapings (Distal C, Proximal D), categorized by sample type. A Kruskal-Wallis statistical test was conducted to determine the significance of the differences. Own elaboration.

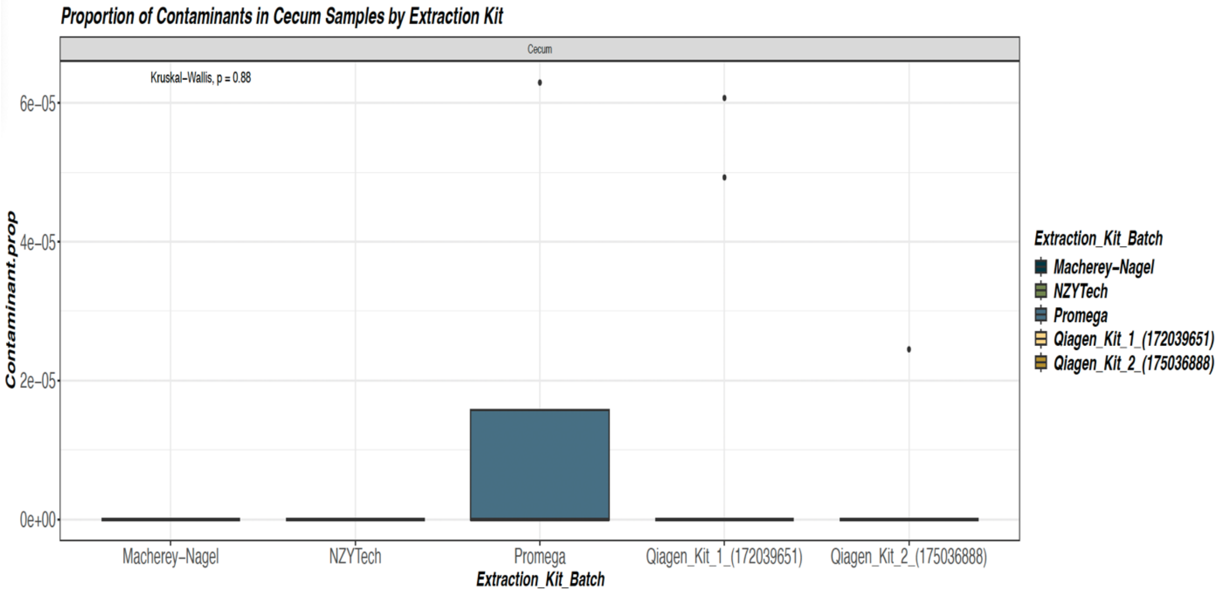


Figure 22: Proportion of contaminant bacterial Reads in Cecum Samples filtered by Extraction Kit. Kruskal-Wallis Statistical test was done to calculate significances. Own elaboration.

In these graphs, it can be observed that the proportion of contaminants in the samples varies depending on the sample type and the extraction kit used. While the differences are not statistically significant due to the limited number of samples per kit and colonic location, some trends could be observed. For example, in the case of distal and proximal colon biopsies (Figure 21.A and 21.B), the Macherey-Nagel and Promega kits show a slightly higher proportion of contaminants compared to the other extraction kits. For cecum samples, the Promega kit presents a higher proportion of contaminants. On the other hand, for mucosal scrapings (Figure 21.C and 21.C), no significant differences in the proportion of contaminants were found, regardless of the Qiagen batch used.

3.7 Microbiome composition in mouse colonic samples

The present study represents a pilot and aims at identifying the best methodologies for (low biomass) colonic sample processing for 16S and metatranscriptomic sequencing. Additionally, given the results of the 16S amplicon sequencing, a preliminary analysis of the microbiome composition in the sequenced samples was performed. Once the contaminants were identified, a quality control was carried out on the obtained results. In this quality control, samples with less than 5000 reads were discarded, including blanks and samples with extremely low biomass. The positive controls have also been removed from this analysis as they do not provide additional information in this case. Reads from bacteria identified as contaminants were also eliminated from all samples.

After filtering and data cleaning, a PCoA was conducted to determine any trend in sample distribution across the first principal components (Figure 23). The data were normalized using a CLR (Centered Log-Ratio) transformation, and Euclidean distance was used to calculate the PCoA. This approach helps in identifying patterns in the microbial composition across the different samples by reducing compositional bias and providing a clearer visualization of the variation in the dataset.

In this PCoA plot, a separation of samples into different groups was observed. This separation coincides with the research group from which the mice were originated (Figure 23.A).Notably, the samples from the mice in Marta Casado's group at IBV stand out, as these were the only ones on a high-fat diet, different from the diet of mice from other groups, while the rest of the samples from mice in other research groups are more similar to each other, likely because the diet fed to these mice was the same. In Figure 23.B, differences between the various batches of the extraction kits can be observed.

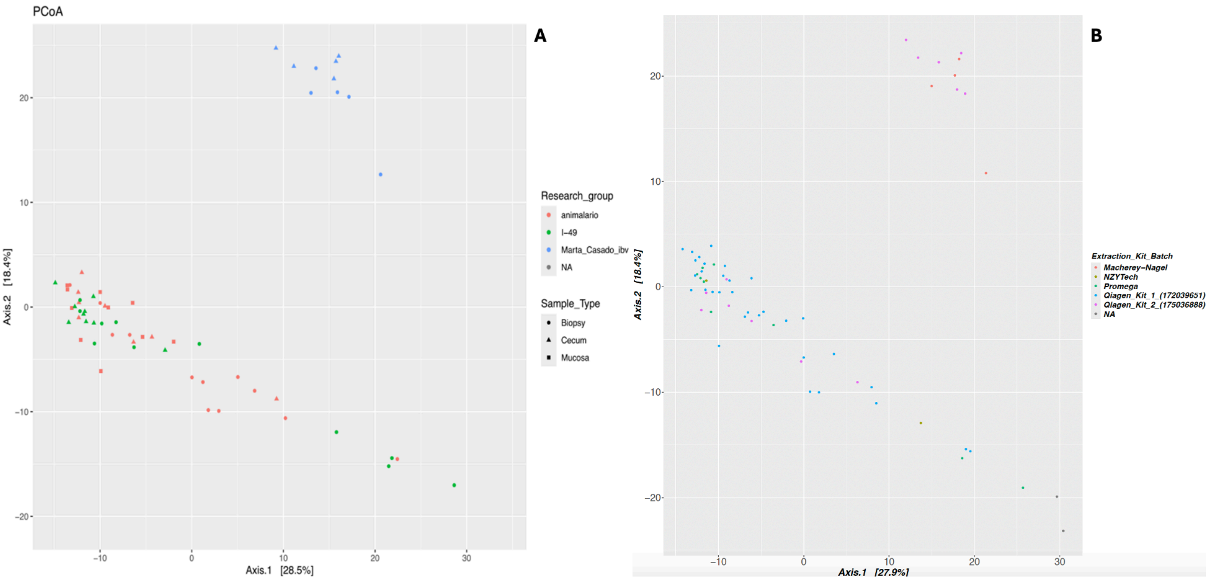


Figure 23: PCoA Plot of Sample Groupings by Research Group, Sample Type (A) and Extraction Kit Batch (B) after filtering the samples with less than 5000 reads. This Figure illustrates the groups formed among the samples, organized by Research Group, Sample Type and Extraction Kit Batch. NA in Research Group represents the Blanks and Positive Controls. Own elaboration

The PCoA suggests that there may be significant differences in bacterial composition between different types of samples and research groups. To observe the differences in bacterial composition among mice from different groups and sample types, a representation of the relative abundances of the 10 most abundant genera was conducted. The results of this analysis are presented in Figure 24.

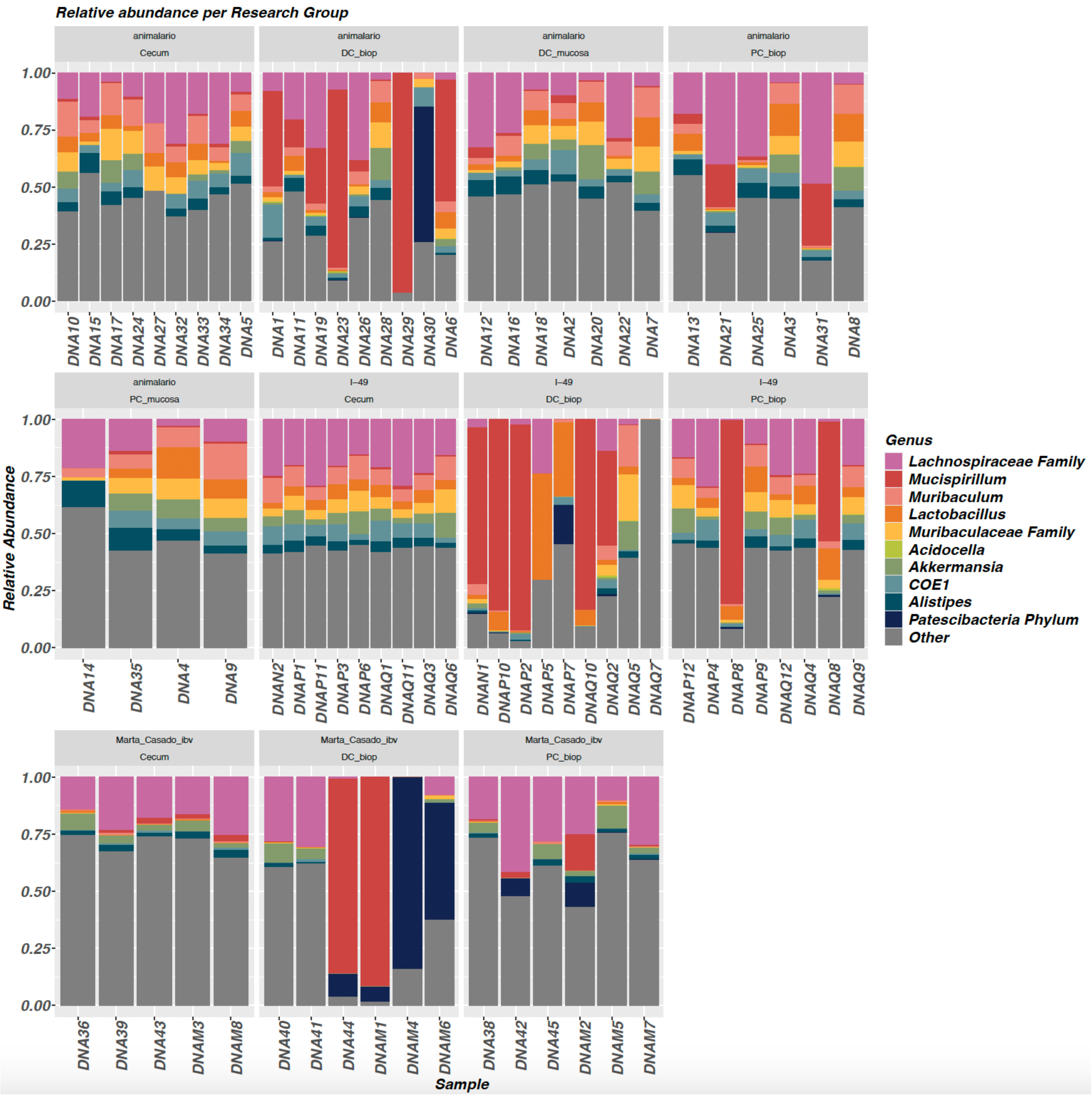


Figure 24: Relative abundance of Top 20 Genera in the samples, ordered by Research group and Sample Type. This bar plot illustrates the differences in the relative abundance of bacteria across each sample type and research group. The grey bars represent other genera not belonging to the 20 most abundant genera.

In Figure 24, differences are shown not only between mice from different research groups but also between sample types. For instance, *Mucispirillum* is the dominant bacterium in some samples, displacing almost all others, particularly in distal colon biopsies from both the animal facility,I-49 and IBV groups. The differences observed in mice from different research groups can be attributed to the varying conditions and diets they are subjected to. This explains why mice from Marta Casado's group exhibit a different bacterial composition than the other mice. For example, *Patescibacteria phylum* bacteria are more present in these mice than in others from the rest of groups.

There are likely many other significant differences, but without a statistical test, these observations remain speculative. Additionally, the bacterial composition varies across different sample types, with distal colon biopsies and mucosal scrapings differing in composition from proximal colon biopsies and mucosal scrapings, and both differing from cecal samples.

This study does not include a differential abundance analysis with statistical validation because it is outside the scope of this work, especially considering that this is a preliminary study. However, it can be suggested that diet and colon location likely influence the microbiota composition and that these factors might also affect bacterial function, something that could be further explored with metatranscriptomic analysis in future studies.

4.Discussion

The gut microbiota performs essential functions to maintain host homeostasis, and it is generally assumed that the microbiota associations with different human disorders are due to the impairment of these functions (such as short chain fatty acid production), in many cases concomitant to changes in community composition. Thus, to better comprehend the mechanisms underlying the associations between microbiota and human disease, it is imperative to improve our understanding of microbiota functions. Techniques such as metatranscriptomics allow to profile the active fraction of the microbiota and the functions performed by each member of the community. Metatranscriptomic analyses of the gut microbiota, especially those involving biopsies, present several challenges due to the low microbial biomass compared to that of the host, yet are essential to understand the functions being performed *in situ* at different intestinal sites, such as tumors or inflamed tissue. This, combined with the lack of a standardized protocol for processing and handling such samples, makes this task very complicated and difficult to perform (Sánchez-Rumí & Lloréns-Rico, 2024).

This study focused on optimizing the collection and processing of diverse mouse intestinal samples: cecal contents, mucosal samples and biopsies, and determining the best way to process them using different extraction kits, various RNA purification methods, and comparing the different types of samples. Additionally, efforts were made to identify potential contaminants specific to each extraction kit, with the intention of extending these findings to human biopsies.

In this study, 96 samples were processed using different DNA/RNA extraction kits and methodologies under comparison. The aspects of interest in this comparison include the quantity of RNA obtained, quality of RNA, proportion of bacterial RNA, and contamination levels. To evaluate these parameters, various analyses were performed, including quality measurements using Tape Bioanalyzer, qPCR, Qubit/Nanodrop measurements for RNA concentration, and 16S amplicon sequencing.

When comparing different types of samples, colon biopsies emerged as the most interesting option. Despite minor differences, colon biopsies generally offered better RNA integrity and higher concentrations than mucosal scrapings. This can be attributed to the higher concentration of host RNA in biopsies. Although the bacterial RNA proportions are lower in these samples, 16S amplicon sequencing demonstrates that it is possible to obtain relevant information regarding the microbiota, especially if a prior (RT)-qPCR step can be performed to determine bacterial relative abundances and select samples with a minimum bacterial proportion.

Thus, biopsies represent a suitable option for dual metatranscriptomic studies of both host and microbiota functions in these regions. Additionally, the ease and speed of obtaining colon biopsies compared to mucosal scrapings, which are more prone to RNA degradation due to the time required for their collection, further support the preference for biopsies. Furthermore, cecal samples are also advantageous due to their high bacterial biomass, making them a more cost-effective option for sequencing when determining host expression is not necessary.

The evaluation of various RNA extraction kits and sample types provided crucial information to optimize RNA quality in microbiota studies. Among the tested kits, the differences in performance were small, but there are some notable observations. For instance, the Promega kit resulted in a very low RNA concentration after purification, independent of the purification method employed. The Macherey-Nagel kit showed a slightly higher proportion of contaminants, although not significantly so. The NZYtech kit performed well according to most parameters, but it yielded fewer 16S reads from biopsies compared to Qiagen, and it should be noted that only a few samples were processed with this kit.

Regarding RNA purification methods, treatments involving on-column DNase proved effective in reducing DNA contamination. This is crucial for ensuring accurate and reliable metatranscriptomic results. The findings indicate that performing DNA removal treatments during RNA extraction is necessary to prevent genomic contamination and improve the overall quality of RNA samples. Although bead purification methods yielded slightly superior RNA quality, the cost and marginal improvement compared to on-column DNase purification suggest that the latter is a more practical and cost-effective alternative without significantly compromising RNA integrity.

Contaminant analysis performed on the 16S amplicon sequencing data highlights the importance to consider the nomenclature of bacteria, which can vary depending on the database used, as demonstrated by the data on the Zymo mock community, where *Pseudocitrobacter* genus was assigned to *Escherichia*. Study context is also relevant: in intestinal microbiota studies, for example, *Escherichia* might not be considered a contaminant due to its natural presence in the intestine. Similarly, although some studies have classified *Fusobacterium* as a contaminant (Lauder et al., 2016; Glassing et al., 2016), in the context of this intestinal microbiota study, this bacterium is also not considered a contaminant. This variability underscores the importance of contextualizing findings according to the type of study and the databases used to ensure an accurate interpretation of microbiological data.

Additionally, preliminary analysis of the 16S amplicon sequencing of the 96 samples, after filtering and quality control, revealed differences in microbiota composition based on laboratory, diet, and sample location, suggesting that functional differences may also be expected and identified via metatranscriptomics.

It is worth mentioning some limitations of the study: the number of samples is limited, and it has not been possible to test all possible combinations of kit, sample type, and purification method. As shown in Figure 4, the uneven distribution of samples among the different extraction kits is due to limitations in acquiring the kits. In many cases, free samples provided by commercial suppliers were used. For example, only two extractions could be performed with the NZYTech kit (one distal colon biopsy and one cecum sample) due to limited availability, which also resulted in a reduced or non-existent number of negative controls for certain kits.

Additionally, due to its high costs, especially in biopsies, metatranscriptomics has not been performed within the timeframe of this study, and therefore the presence of RNA contaminants in the kits cannot be ruled out. However, the limited stability of RNA, together with the reduced impact of contamination at the DNA level demonstrated by the 16S amplicon study analysis, make the impact of contamination in metatranscriptomics unlikely. The limited contamination is also the result of the application of very strict protocols to limit contamination in the laboratory.

In addition, based on the results obtained in this study, it will be possible to select the most suitable samples for future metatranscriptomic analyses using the identified criteria: a minimum RNA yield of 100 ng/µL and quality (RIN value of at least 7), as well as a minimum bacterial biomass. This selection can be achieved by applying a threshold on Cp values obtained via (RT-)qPCR, such as a Cp value below 20, and/or by applying a threshold on the number of bacterial reads obtained in 16S amplicon sequencing, such as 10,000 reads.Such sample selection will allow to optimize the sequencing costs, by sequencing only samples that are likely to yield bacterial transcriptional results.

Lastly, this pilot has been performed on mouse samples, due to their accessibility within our research institute and lack of ethical issues, as these samples are residuary material from other research projects. Therefore, it is plausible that our results do not fully translate to humans (for instance, in human studies a bowel prep is required to remove intestinal contents, potentially reducing bacterial biomass in biopsy sites). This underscores the importance of performing preliminary tests such as (RT-)qPCR and 16S sequencing to select samples for a cost-efficient study.

Overall, our observations underscore the importance of selecting appropriate extraction kits, sample types, and purification methods to ensure high-quality RNA, which is vital for reliable metatranscriptomic analyses. In conclusion, the Qiagen AllPrep DNA/RNA Mini extraction kit, along with on-column DNase purification, offers an optimal balance between cost, efficiency, and RNA quality. Colon biopsies are suitable if joint host and microbiota metatranscriptomic analysis is needed, or where a specific colonic location must be profiled, while cecal content samples provide a cost-effective alternative for mouse studies if these requirements are not needed.

These recommendations provide a solid foundation for future microbiota research, ensuring robust and reproducible results not only in mouse samples but also in human samples.

5. Conclusions

A total of 96 samples, including blanks and positive controls, were processed and subjected to various analyses. These analyses included total RNA quantity, quality, percentage of bacterial RNA, 16S rRNA, and contamination assessment. The results of this study demonstrate that the objectives set have been satisfactorily achieved:

1. A collection of mouse intestinal samples has been established, with a standardized protocol to collect and preserve the samples. This collection is currently expanding to include mouse from additional strains and experimental conditions. A total number of 168 samples were collected and further samples are being added to the collection routinely. The goal of this collection is to profile the metatranscriptome in a wide variety of conditions to understand the natural variability of gut microbial functions in mouse models.

2. **The Comparative Analysis of extraction Kits** according to criteria of RNA yield, quality, bacterial relative abundances and presence of contaminants revealed small differences among the kits compared, yet the Qiagen kit appeared to be suitable for these samples, while the NZYTech kit also showed promising results.

3. The results of the **Purification Methods Evaluation** indicate that DNA elimination during RNA extraction is essential. DNase I treatment in column was found to be a cost-effective method for this purpose.

4. A batch effect was observed in terms of contaminant detection between the two batches of the same Qiagen kit, underscoring the importance of including extraction blanks in all extraction experiments.

5. **The Sample Type comparative analysis** revealed that biopsy samples are suitable for gut microbiota studies where a specific location or host information is required, while cecal content samples can be used in mouse studies if the above requirements are not needed. Mucosal samples require additional processing steps in mouse samples which may affect the RNA integrity and increase the likelihood of environmental contamination.

6. **16S sequencing analysis preliminary results reveal differences in the microbiota composition** among mice from different research groups and variations in the bacterial composition across different sections of the colon, rendering it likely that such differences also exist at the functional level.

6. Bibliography

- Abellan-Schneyder, Isabel et al. “Primer, Pipelines, Parameters: Issues in 16S rRNA Gene Sequencing.” *mSphere* vol. 6,1 e01202-20. 24 Feb. 2021, doi:10.1128/mSphere.01202-20

- Ali, Nasir et al. “Current Nucleic Acid Extraction Methods and Their Implications to Point-of-Care Diagnostics.” BioMed research international vol. 2017 (2017): 9306564. doi:10.1155/2017/9306564

- Almeqdadi, Mohammad et al. “Gut organoids: mini-tissues in culture to study intestinal physiology and disease.” American journal of physiology. Cell physiology vol. 317,3 (2019): C405-C419. doi:10.1152/ajpcell.00300.2017

- Apprill, Amy & McNally, Sean & Parsons, Rachel & Weber, Laura. (2015). Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. Aquatic Microbial Ecology. 75. 10.3354/ame01753.

- Appunni, Sandeep et al. “Emerging Evidence on the Effects of Dietary Factors on the Gut Microbiome in Colorectal Cancer.” Frontiers in nutrition vol. 8 718389. 11 Oct. 2021, doi:10.3389/fnut.2021.718389

- Arumugam, Manimozhiyan et al. “Enterotypes of the human gut microbiome.” *Nature* vol. 473,7346 (2011): 174-80. doi:10.1038/nature09944

- Arumugam, Manimozhiyan et al. “Enterotypes of the human gut microbiome.” Nature vol. 473,7346 (2011): 174-80. doi:10.1038/nature09944

- Bäckhed, Fredrik et al. “Host-bacterial mutualism in the human intestine.” Science (New York, N.Y.) vol. 307,5717 (2005): 1915-20. doi:10.1126/science.1104816

- Barnett et al., (2021). microViz: an R package for microbiome data visualization and statistics. Journal of Open Source Software, 6(63), 3201, https://doi.org/10.21105/joss.03201

- Barton, H A et al. “DNA extraction from low-biomass carbonate rock: an improved method with reduced contamination and the low-biomass contaminant database.” *Journal of microbiological methods* vol. 66,1 (2006): 21-31. doi:10.1016/j.mimet.2005.10.005

- Bolger, Anthony M et al. “Trimmomatic: a flexible trimmer for Illumina sequence data.” *Bioinformatics (Oxford, England)* vol. 30,15 (2014): 2114-20. doi:10.1093/bioinformatics/btu170

- Bortolaia, Valeria et al. “ResFinder 4.0 for predictions of phenotypes from genotypes.” The Journal of antimicrobial chemotherapy vol. 75,12 (2020): 3491-3500. doi:10.1093/jac/dkaa345

- Callahan, Benjamin J et al. “DADA2: High-resolution sample inference from Illumina amplicon data.” *Nature methods* vol. 13,7 (2016): 581-3. doi:10.1038/nmeth.3869

- Chiu, Charles Y, and Steven A Miller. “Clinical metagenomics.” Nature reviews. Genetics vol. 20,6 (2019): 341-355. doi:10.1038/s41576-019-0113-7

- Choi, Seung-Chul et al. “Gut microbiota dysbiosis and altered tryptophan catabolism contribute to autoimmunity in lupus-susceptible mice.” *Science translational medicine* vol. 12,551 (2020): eaax2220. doi:10.1126/scitranslmed.aax2220

- Davis, Nicole M et al. “Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data.” *Microbiome* vol. 6,1 226. 17 Dec. 2018, doi:10.1186/s40168-018-0605-2

- de Martel, Catherine et al. “Global burden of cancers attributable to infections in 2008: a review and synthetic analysis.” *The Lancet. Oncology* vol. 13,6 (2012): 607-15. doi:10.1016/S1470-2045(12)70137-7

- Deutscher, Murray P. “Degradation of RNA in bacteria: comparison of mRNA and stable RNA.” Nucleic acids research vol. 34,2 659-66. 1 Feb. 2006, doi:10.1093/nar/gkj472

- Devriese, Sarah et al. “T84 monolayers are superior to Caco-2 as a model system of colonocytes.” Histochemistry and cell biology vol. 148,1 (2017): 85-93. doi:10.1007/s00418-017-1539-7

- Duvallet, Claire et al. “Meta-analysis of gut microbiome studies identifies disease-specific and shared responses.” Nature communications vol. 8,1 1784. 5 Dec. 2017, doi:10.1038/s41467-017-01973-8

- El-Sayed, Amr et al. “Microbiota's role in health and diseases.” *Environmental science and pollution research international* vol. 28,28 (2021): 36967-36983. doi:10.1007/s11356-021-14593-z

- Ferraretto, Anita et al. “Morphofunctional properties of a differentiated Caco2/HT-29 co-culture as an in vitro model of human intestinal epithelium.” Bioscience reports vol. 38,2 BSR20171497. 27 Apr. 2018, doi:10.1042/BSR20171497

- Filiatrault, Melanie J. “Progress in prokaryotic transcriptomics.” Current opinion in microbiology vol. 14,5 (2011): 579-86. doi:10.1016/j.mib.2011.07.023

- Fogh, J et al. “One hundred and twenty-seven cultured human tumor cell lines producing tumors in nude mice.” Journal of the National Cancer Institute vol. 59,1 (1977): 221-6. doi:10.1093/jnci/59.1.221

- Franzosa, Eric A et al. “Relating the metatranscriptome and metagenome of the human gut.” Proceedings of the National Academy of Sciences of the United States of America vol. 111,22 (2014): E2329-38. doi:10.1073/pnas.1319284111

- Fresia, Pablo et al. “Urban metagenomics uncover antibiotic resistance reservoirs in coastal beach and sewage waters.” Microbiome vol. 7,1 35. 28 Feb. 2019, doi:10.1186/s40168-019-0648-z

- Funabashi, Masanori et al. “A metabolic pathway for bile acid dehydroxylation by the gut microbiome.” *Nature* vol. 582,7813 (2020): 566-570. doi:10.1038/s41586-020-2396-4

- Gangadoo, Sheeana et al. “The Multiomics Analyses of Fecal Matrix and Its Significance to Coeliac Disease Gut Profiling.” International journal of molecular sciences vol. 22,4 1965. 17 Feb. 2021, doi:10.3390/ijms22041965

- Garrett, Wendy S. “Cancer and the microbiota.” *Science (New York, N.Y.)* vol. 348,6230 (2015): 80-6. doi:10.1126/science.aaa4972

- Geng, Jiafeng et al. “The links between gut microbiota and obesity and obesity related diseases.” *Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie* vol. 147 (2022): 112678. doi:10.1016/j.biopha.2022.112678

- Giannoukos, Georgia et al. “Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes.” Genome biology vol. 13,3 (2012): R23. doi:10.1186/gb-2012-13-3-r23

- Glassing, Angela et al. “Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples.” *Gut pathogens* vol. 8 24. 26 May. 2016, doi:10.1186/s13099-016-0103-7

- Glassing, Angela et al. “Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples.” *Gut pathogens* vol. 8 24. 26 May. 2016, doi:10.1186/s13099-016-0103-7

- Gomaa, Eman Zakaria. “Human gut microbiota/microbiome in health and diseases: a review.” *Antonie van Leeuwenhoek* vol. 113,12 (2020): 2019-2040. doi:10.1007/s10482-020-01474-7

- Grahn, Niclas et al. “Identification of mixed bacterial DNA contamination in broad-range PCR amplification of 16S rDNA V1 and V3 variable regions by pyrosequencing of cloned amplicons.” *FEMS microbiology letters* vol. 219,1 (2003): 87-91. doi:10.1016/S0378-1097(02)01190-4

- Granata, Ilaria et al. “Duodenal Metatranscriptomics to Define Human and Microbial Functional Alterations Associated with Severe Obesity: A Pilot Study.” *Microorganisms* vol. 8,11 1811. 17 Nov. 2020, doi:10.3390/microorganisms8111811

- Grant, Jennifer et al. “Simulating drug concentrations in PDMS microfluidic organ chips.” Lab on a chip vol. 21,18 3509-3519. 14 Sep. 2021, doi:10.1039/d1lc00348h

- Grünberger, Felix et al. “Nanopore sequencing of RNA and cDNA molecules in Escherichia coli.” RNA (New York, N.Y.) vol. 28,3 (2022): 400-417. doi:10.1261/rna.078937.121

- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

- Hamady, Micah, and Rob Knight. “Microbial community profiling for human microbiome projects: Tools, techniques, and challenges.” Genome research vol. 19,7 (2009): 1141-52. doi:10.1101/gr.085464.108

- Hugenholtz, Philip, and Gene W Tyson. “Microbiology: metagenomics.” Nature vol. 455,7212 (2008): 481-3. doi:10.1038/455481a

- Just, Sarah et al. “The gut microbiota drives the impact of bile acids and fat source in diet on mouse metabolism.” *Microbiome* vol. 6,1 134. 2 Aug. 2018, doi:10.1186/s40168-018-0510-8

- Kallmeyer, Jens et al. “Global distribution of microbial abundance and biomass in subseafloor sediment.” *Proceedings of the National Academy of Sciences of the United States of America* vol. 109,40 (2012): 16213-6. doi:10.1073/pnas.1203849109

- Kassambara A (2023). _ggpubr: 'ggplot2' Based Publication Ready Plots_. R package version 0.6.0, https://CRAN.R-project.org/package=ggpubr

- Kau, Andrew L et al. “Human nutrition, the gut microbiome and the immune system.” Nature vol. 474,7351 327-36. 15 Jun. 2011, doi:10.1038/nature10213

- Klindworth, Anna et al. “Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies.” Nucleic acids research vol. 41,1 (2013): e1. doi:10.1093/nar/gks808

- Lagier, Jean-Christophe et al. “The rebirth of culture in microbiology through the example of culturomics to study human gut microbiota.” Clinical microbiology reviews vol. 28,1 (2015): 237-64. doi:10.1128/CMR.00014-14

- Langille, Morgan G I et al. “Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences.” *Nature biotechnology* vol. 31,9 (2013): 814-21. doi:10.1038/nbt.2676

- Langmead, Ben, and Steven L Salzberg. “Fast gapped-read alignment with Bowtie 2.” Nature methods vol. 9,4 357-9. 4 Mar. 2012, doi:10.1038/nmeth.1923

- Lauder, Abigail P et al. “Comparison of placenta samples with contamination controls does not provide evidence for a distinct placenta microbiota.” *Microbiome* vol. 4,1 29. 23 Jun. 2016, doi:10.1186/s40168-016-0172-3

- Laurence, Martin et al. “Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes.” *PloS one* vol. 9,5 e97876. 16 May. 2014, doi:10.1371/journal.pone.0097876

- Ley, Ruth E et al. “Ecological and evolutionary forces shaping microbial diversity in the human intestine.” Cell vol. 124,4 (2006): 837-48. doi:10.1016/j.cell.2006.02.017

- Li, Siying. “Modulation of immunity by tryptophan microbial metabolites.” *Frontiers in nutrition* vol. 10 1209613. 21 Jun. 2023, doi:10.3389/fnut.2023.1209613

- Li D, Gai W, Zhang J, Cheng W, Cui N, Wang H. Metagenomic Next-Generation Sequencing for the Microbiological Diagnosis of Abdominal Sepsis Patients. Front Microbiol. 2022 Feb 2;13:816631. doi: 10.3389/fmicb.2022.816631. PMID: 35185847; PMCID: PMC8847725.

- Liu, Tiantian et al. “An empirical Bayes approach to normalization and differential abundance testing for microbiome data.” BMC bioinformatics vol. 21,1 225. 3 Jun. 2020, doi:10.1186/s12859-020-03552-z

- Liu, Yong-Xin et al. “A practical guide to amplicon and metagenomic analysis of microbiome data.” Protein & cell vol. 12,5 (2021): 315-330. doi:10.1007/s13238-020-00724-8

- Lloyd-Price, Jason et al. “Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases.” Nature vol. 569,7758 (2019): 655-662. doi:10.1038/s41586-019-1237-9 Manipulation_. R package version 1.1.4, https://CRAN.R-project.org/package=dplyr

- Mahmoudabadi, Gita et al. “Single Cell Transcriptomics Reveals the Hidden Microbiomes of Human Tissues” bioRxiv (2022): 10.11.511790

- Martínez-Maqueda, Daniel et al. “Food-derived peptides stimulate mucin secretion and gene expression in intestinal cells.” Journal of agricultural and food chemistry vol. 60,35 (2012): 8600-5. doi:10.1021/jf301279k

- Martínez-Porchas, Marcel, and Francisco Vargas-Albores. “An efficient strategy using k-mers to analyse 16S rRNA sequences.” Heliyon vol. 3,7 e00370. 27 Jul. 2017, doi:10.1016/j.heliyon.2017.e00370

- Mayer, Emeran A et al. “The Gut-Brain Axis.” *Annual review of medicine* vol. 73 (2022): 439-453. doi:10.1146/annurev-med-042320-014032

- McDonald, Daniel et al. “An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea.” The ISME journal vol. 6,3 (2012): 610-8. doi:10.1038/ismej.2011.139

- McMurdie, Paul J, and Susan Holmes. “phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data.” *PloS one* vol. 8,4 e61217. 22 Apr. 2013, doi:10.1371/journal.pone.0061217

- Nikolaev, Mikhail et al. “Homeostatic mini-intestines through scaffold-guided organoid morphogenesis.” Nature vol. 585,7826 (2020): 574-578. doi:10.1038/s41586-020-2724-8

- Ojala, Teija et al. “Current concepts, advances, and challenges in deciphering the human microbiota with metatranscriptomics.” Trends in genetics : TIG vol. 39,9 (2023): 686-702. doi:10.1016/j.tig.2023.05.004

- Olovo, Chinasa Valerie et al. “Faecal microbial biomarkers in early diagnosis of colorectal cancer.” Journal of cellular and molecular medicine vol. 25,23 (2021): 10783-10797. doi:10.1111/jcmm.17010

- Ottman, Noora et al. “The function of our microbiota: who is out there and what do they do?.” Frontiers in cellular and infection microbiology vol. 2 104. 9 Aug. 2012, doi:10.3389/fcimb.2012.00104

- Perler, Bryce K et al. “The Role of the Gut Microbiota in the Relationship Between Diet and Human Health.” *Annual review of physiology* vol. 85 (2023): 449-468. doi:10.1146/annurev-physiol-031522-092054

- Pickard, Joseph M et al. “Gut microbiota: Role in pathogen colonization, immune responses, and inflammatory disease.” *Immunological reviews* vol. 279,1 (2017): 70-89. doi:10.1111/imr.12567

- Puljiz, Zivana et al. “Obesity, Gut Microbiota, and Metabolome: From Pathophysiology to Nutritional Interventions.” *Nutrients* vol. 15,10 2236. 9 May. 2023, doi:10.3390/nu15102236

- Purushothaman, Srinithi et al. “Combination of Whole Genome Sequencing and Metagenomics for Microbiological Diagnostics.” International journal of molecular sciences vol. 23,17 9834. 30 Aug. 2022, doi:10.3390/ijms23179834

- Puschhof, Jens et al. “Intestinal organoid cocultures with microbes.” *Nature protocols* vol. 16,10 (2021): 4633-4649. doi:10.1038/s41596-021-00589-z

- Puschhof, Jens et al. “Organoids and organs-on-chips: Insights into human gut-microbe interactions.” *Cell host & microbe* vol. 29,6 (2021): 867-878. doi:10.1016/j.chom.2021.04.002

- Qi, Yuli et al. “In vitro models to study human gut-microbiota interactions: Applications, advances, and limitations.” Microbiological research vol. 270 (2023): 127336. doi:10.1016/j.micres.2023.127336

- Quaglio, Ana Elisa Valencise et al. “Gut microbiota, inflammatory bowel disease and colorectal cancer.” World journal of gastroenterology vol. 28,30 (2022): 4053-4060. doi:10.3748/wjg.v28.i30.4053

- Quince, Christopher et al. “Shotgun metagenomics, from sampling to analysis.” Nature biotechnology vol. 35,9 (2017): 833-844. doi:10.1038/nbt.3935

- Ratajczak, Weronika et al. “Immunomodulatory potential of gut microbiome-derived short-chain fatty acids (SCFAs).” *Acta biochimica Polonica* vol. 66,1 (2019): 1-12. doi:10.18388/abp.2018_2648

- Rauhut, R, and G Klug. “mRNA degradation in bacteria.” *FEMS microbiology reviews* vol. 23,3 (1999): 353-70. doi:10.1111/j.1574-6976.1999.tb00404.x

- R Core Team (2024). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- Reck, Michael et al. “Stool metatranscriptomics: A technical guideline for mRNA stabilisation and isolation.” BMC genomics vol. 16,1 494. 4 Jul. 2015, doi:10.1186/s12864-015-1694-y

- Reigstad, Christopher S, and Purna C Kashyap. “Beyond phylotyping: understanding the impact of gut microbiota on host biology.” *Neurogastroenterology and motility* vol. 25,5 (2013): 358-72. doi:10.1111/nmo.12134

- Rizzatti, G et al. “Proteobacteria: A Common Factor in Human Diseases.” BioMed research international vol. 2017 (2017): 9351507. doi:10.1155/2017/9351507

- Rubinstein, Mara Roxana et al. “Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/β-catenin signaling via its FadA adhesin.” *Cell host & microbe* vol. 14,2 (2013): 195-206. doi:10.1016/j.chom.2013.07.012

- Ruiz, Lorena et al. “Tackling probiotic and gut microbiota functionality through proteomics.” *Journal of proteomics* vol. 147 (2016): 28-39. doi:10.1016/j.jprot.2016.03.023

- Salter, Susannah J et al. “Reagent and laboratory contamination can critically impact sequence-based microbiome analyses.” *BMC biology* vol. 12 87. 12 Nov. 2014, doi:10.1186/s12915-014-0087-z

- Sánchez-Rumí MJ, Lloréns-Rico V. Estudiando las funciones de la microbiota intestinal: presente y futuro de la metatranscriptómica. SEBBM. 2024;220. doi:10.18567/sebbmrev_220.202406.dc003

- Schloss, Patrick D, and Sarah L Westcott. “Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis.” *Applied and environmental microbiology* vol. 77,10 (2011): 3219-26. doi:10.1128/AEM.02810-10

- Schloss, Patrick D et al. “Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies.” PloS one vol. 6,12 (2011): e27310. doi:10.1371/journal.pone.0027310

- Shakya, Migun et al. “Advances and Challenges in Metatranscriptomic Analysis.” Frontiers in genetics vol. 10 904. 25 Sep. 2019, doi:10.3389/fgene.2019.00904

- Shokralla, Shadi et al. “Next-generation sequencing technologies for environmental DNA research.” Molecular ecology vol. 21,8 (2012): 1794-805. doi:10.1111/j.1365-294X.2012.05538.x

- Tanner, M A et al. “Specific ribosomal DNA sequences from diverse environmental settings correlate with experimental contaminants.” *Applied and environmental microbiology* vol. 64,8 (1998): 3110-3. doi:10.1128/AEM.64.8.3110-3113.1998
- Tao, Yue et al. “Diagnostic Performance of Metagenomic Next-Generation Sequencing in Pediatric Patients: A Retrospective Study in a Large Children's Medical Center.” *Clinical chemistry* vol. 68,8 (2022): 1031-1041. doi:10.1093/clinchem/hvac067

- Truong, Duy Tin et al. “MetaPhlAn2 for enhanced metagenomic taxonomic profiling.” Nature methods vol. 12,10 (2015): 902-3. doi:10.1038/nmeth.3589

- Turnbaugh, Peter J et al. “Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome.” Cell host & microbe vol. 3,4 (2008): 213-23. doi:10.1016/j.chom.2008.02.015

- Turnbaugh, Peter J et al. “The human microbiome project.” Nature vol. 449,7164 (2007): 804-10. doi:10.1038/nature06244 version 2.1.5, https://CRAN.R-project.org/package=readr

- Walter, Jens, and Ruth Ley. “The human gut microbiome: ecology and recent evolutionary changes.” Annual review of microbiology vol. 65 (2011): 411-29. doi:10.1146/annurev-micro-090110-102830

- Walters, William A et al. “Meta-analyses of human gut microbes associated with obesity and IBD.” FEBS letters vol. 588,22 (2014): 4223-33. doi:10.1016/j.febslet.2014.09.039

- Walters, William A et al. “PrimerProspector: de novo design and taxonomic analysis of barcoded polymerase chain reaction primers.” Bioinformatics (Oxford, England) vol. 27,8 (2011): 1159-61. doi:10.1093/bioinformatics/btr087

- Walters, William et al. “Improved Bacterial 16S rRNA Gene (V4 and V4-5) and Fungal Internal Transcribed Spacer Marker Gene Primers for Microbial Community Surveys.” *mSystems* vol. 1,1 e00009-15. 22 Dec. 2015, doi:10.1128/mSystems.00009-15

- Wang, Juanjuan et al. “Gut-Microbiota-Derived Metabolites Maintain Gut and Systemic Immune Homeostasis.” *Cells* vol. 12,5 793. 2 Mar. 2023, doi:10.3390/cells12050793

- Weyrich, Laura S et al. “Laboratory contamination over time during low-biomass sample analysis.” *Molecular ecology resources* vol. 19,4 (2019): 982-996. doi:10.1111/1755-0998.13011

- Wickham et al., (2019). Welcome to the Tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

- Wickham H, François R, Henry L, Müller K, Vaughan D (2023). _dplyr: A Grammar of Data

- Wickham H, Hester J, Bryan J (2024). _readr: Read Rectangular Text Data_. R package

- Wickham H, Vaughan D, Girlich M (2024). *tidyr: Tidy Messy Data*. R package version 1.3.1, https://github.com/tidyverse/tidyr, https://tidyr.tidyverse.org.
- Xing, Zhikai et al. “RBUD: A New Functional Potential Analysis Approach for Whole Microbial Genome Shotgun Sequencing.” *Microorganisms* vol. 8,10 1563. 10 Oct. 2020, doi:10.3390/microorganisms8101563

- Yang, Bo et al. “Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis.” BMC bioinformatics vol. 17 135. 22 Mar. 2016, doi:10.1186/s12859-016-0992-y

- Zhang, Yancong et al. “Metatranscriptomics for the Human Microbiome and Microbial Community Functional Profiling.” Annual review of biomedical data science vol. 4 (2021): 279-311. doi:10.1146/annurev-biodatasci-031121-103035

Supplementary material (Annex)

Nº DNA	Nºmouse	Strain	Genotype	Diet	Research group	Euthanasia method
DNA38	31	C57BL/6	(KI)_No_expression_COX-2	High-fat diet_(A06071302)	Marta_Casado_ibv	Anesthesia_&_bleeding
DNA40	32	C57BL/6	(KI)_No_expression_COX-2	High-fat diet_(A06071302)	Marta_Casado_ibv	Anesthesia_&_bleeding
DNA41	35	C57BL/6	(KI)_expression_COX-2	High-fat diet_(A06071302)	Marta_Casado_ibv	Anesthesia_&_bleeding
DNA42	35	C57BL/6	(KI)_expression_COX-2	High-fat diet_(A06071302)	Marta_Casado_ibv	Anesthesia_&_bleeding
DNA44	34	C57BL/6	(KI)_No_expression_COX-2	High-fat diet_(A06071302)	Marta_Casado_ibv	Anesthesia_&_bleeding
DNA45	34	C57BL/6	(KI)_No_expression_COX-2	High-fat diet_(A06071302)	Marta_Casado_ibv	Anesthesia_&_bleeding
DNA6	7	C57BL/6	Wild _type	Normal	animalario	CO2
DNA8	7	C57BL/6	Wild _type	Normal	animalario	CO2
DNAM1	36	C57BL/6	(KI)_No_expression_COX-2	High-fat diet_(A06071302)	Marta_Casado_ibv	Anesthesia_&_bleeding
DNAM2	36	C57BL/6	(KI)_No_expression_COX-2	High-fat diet_(A06071302)	Marta_Casado_ibv	Anesthesia_&_bleeding
DNAM4	30	C57BL/6	(KI)_No_expression_COX-2	High-fat diet_(A06071302)	Marta_Casado_ibv	Anesthesia_&_bleeding
DNAM5	30	C57BL/6	(KI)_No_expression_COX-2	High-fat diet_(A06071302)	Marta_Casado_ibv	Anesthesia_&_bleeding
DNAM6	33	C57BL/6	(KI)_expression_COX-2	High-fat diet_(A06071302)	Marta_Casado_ibv	Anesthesia_&_bleeding
DNAM7	33	C57BL/6	(KI)_expression_COX-2	High-fat diet_(A06071302)	Marta_Casado_ibv	Anesthesia_&_bleeding
DNA37	31	C57BL/6	(KI)_No_expression_COX-2	High-fat diet_(A06071302)	Marta_Casado_ibv	Anesthesia_&_bleeding
DNABI2	NA	NA	NA	NA	NA	NA
DNABI3	NA	NA	NA	NA	NA	NA
DNABI4	NA	NA	NA	NA	NA	NA
DNABI5	NA	NA	NA	NA	NA	NA
DNABI6	NA	NA	NA	NA	NA	NA
DNABI7	NA	NA	NA	NA	NA	NA
DNABI8	NA	NA	NA	NA	NA	NA
DNABI9	NA	NA	NA	NA	NA	NA
DNABIM1	NA	NA	NA	NA	NA	NA
DNABIM2	NA	NA	NA	NA	NA	NA
DNABIP	NA	NA	NA	NA	NA	NA
DNABIP2	NA	NA	NA	NA	NA	NA
DNABIQ	NA	NA	NA	NA	NA	NA
DNABIQ2	NA	NA	NA	NA	NA	NA
DNABLANK	NA	NA	NA	NA	NA	NA
DNA10	7	C57BL/6	Wild _type	Normal	animalario	CO2
DNA36	30	C57BL/6	(KI)_No_expression_COX-2	High-fat diet_(A06071302)	Marta_Casado_ibv	Anesthesia_&_bleeding
DNA39	31	C57BL/6	(KI)_No_expression_COX-2	High-fat diet_(A06071302)	Marta_Casado_ibv	Anesthesia_&_bleeding
DNA43	35	C57BL/6	(KI)_expression_COX-2	High-fat diet_(A06071302)	Marta_Casado_ibv	Anesthesia_&_bleeding
DNA7	7	C57BL/6	Wild _type	Normal	animalario	CO2

DNA9	7	C57BL/6	Wild _type	Normal	animalario	CO2
LRHAMNOSUS	NA	NA	NA	NA	NA	NA
ZYMOMOCK	NA	NA	NA	NA	NA	NA
DNA5	8	C57BL/6	Wild _type	Normal	animalario	CO2
DNA15	5	CD1	Wild _type	Normal	animalario	CO2
DNAP5	23bis	C57BL/6	Irs2 _Luciferasa_(het)	Normal	I-49	CO2
DNA27	6	CD1	Wild _type	Normal	animalario	CO2
DNA33	14	C57BL/6	Wild _type	Normal	animalario	CO2
DNAQ10	22	C57BL/6	Irs2 _Luciferasa_(het)	Normal	I-49	CO2
DNA17	10	C57BL/6	Wild _type	Normal	animalario	CO2
DNA29	14	C57BL/6	Wild _type	Normal	animalario	CO2
DNA14	5	CD1	Wild _type	Normal	animalario	CO2
DNA13	5	CD1	Wild _type	Normal	animalario	CO2
DNAQ6	23	C57BL/6	Irs2 _Luciferasa_(het)	Normal	I-49	CO2
DNAQ3	21	C57BL/6	Irs2_Luciferasa_(KI)	Normal	I-49	CO2
DNA32	15	C57BL/6	Wild _type	Normal	animalario	CO2
DNA35	4	C57BL/6	Wild _type	Normal	animalario	CO2
DNA3	8	C57BL/6	Wild _type	Normal	animalario	CO2
DNA25	10	C57BL/6	Wild _type	Normal	animalario	CO2
DNA31	12	C57BL/6	Wild _type	Normal	animalario	CO2
DNAP12	23bis	C57BL/6	Irs2 _Luciferasa_(het)	Normal	I-49	CO2
DNAQ12	23	C57BL/6	Irs2 _Luciferasa_(het)	Normal	I-49	CO2
DNA34	16	C57BL/6	Wild _type	Normal	animalario	CO2
DNA22	6	CD1	Wild _type	Normal	animalario	CO2
DNAQ1	20	C57BL/6	Irs2 _Luciferasa_(het)	Normal	I-49	CO2
DNA11	5	CD1	Wild _type	Normal	animalario	CO2
DNAQ2	21	C57BL/6	Irs2_Luciferasa_(KI)	Normal	I-49	CO2
DNA21	11	C57BL/6	Wild _type	Normal	animalario	CO2
DNA26	12	C57BL/6	Wild _type	Normal	animalario	CO2
DNA28	3	C57BL/6	Wild _type	Normal	animalario	CO2
DNAP10	22bis	C57BL/6	Irs2 _Luciferasa_(het)	Normal	I-49	CO2
DNAP7	20bis	C57BL/6	Irs2 _Luciferasa_(het)	Normal	I-49	CO2
DNAP8	20bis	C57BL/6	Irs2 _Luciferasa_(het)	Normal	I-49	CO2
DNAQ7	20	C57BL/6	Irs2 _Luciferasa_(het)	Normal	I-49	CO2
DNAQ8	20	C57BL/6	Irs2 _Luciferasa_(het)	Normal	I-49	CO2
DNAP11	22bis	C57BL/6	Irs2 _Luciferasa_(het)	Normal	I-49	CO2
DNA18	1	C57BL/6	Wild _type	Normal	animalario	CO2
DNA20	3	C57BL/6	Wild _type	Normal	animalario	CO2
DNAQ4	22	C57BL/6	Irs2 _Luciferasa_(het)	Normal	I-49	CO2
DNAP6	23bis	C57BL/6	Irs2 _Luciferasa_(het)	Normal	I-49	CO2
DNAP1	20bis	C57BL/6	Irs2 _Luciferasa_(het)	Normal	I-49	CO2
DNAP3	21bis	C57BL/6	Irs2_Luciferasa_(KI)	Normal	I-49	CO2

DNAQ5	23	C57BL/6	Irs2 _Luciferasa_(het)	Normal	I-49	CO2
DNA12	5	CD1	Wild _type	Normal	animalario	CO2
DNA4	8	C57BL/6	Wild _type	Normal	animalario	CO2
DNA1	8	C57BL/6	Wild _type	Normal	animalario	CO2
DNA19	10	C57BL/6	Wild _type	Normal	animalario	CO2
DNA23	13	C57BL/6	Wild _type	Normal	animalario	CO2
DNAP9	21bis	C57BL/6	Irs2_Luciferasa_(KI)	Normal	I-49	CO2
DNAQ9	21	C57BL/6	Irs2_Luciferasa_(KI)	Normal	I-49	CO2
DNA16	2	C57BL/6	Wild _type	Normal	animalario	CO2
DNAP4	22bis	C57BL/6	Irs2 _Luciferasa_(het)	Normal	I-49	CO2
DNAP2	21 bis	C57BL/6	Irs2_Luciferasa_(KI)	Normal	I-49	CO2
DNAQ11	22	C57BL/6	Irs2 _Luciferasa_(het)	Normal	I-49	CO2
DNA24	2	C57BL/6	Wild _type	Normal	animalario	CO2
DNA2	8	C57BL/6	Wild _type	Normal	animalario	CO2
DNA30	16	C57BL/6	Wild _type	Normal	animalario	CO2
DNAM3	36	C57BL/6	(KI)_No_expression_COX-2	High-fat diet_(A06071302)	Marta_Casado_ibv	Anesthesia_&_bleeding
DNAM8	33	C57BL/6	(KI)_expression_COX-2	High-fat diet_(A06071302)	Marta_Casado_ibv	Anesthesia_&_bleeding
DNAN2	24	C57BL/6	Irs2 _Luciferasa_(het)	Normal	I-49	CO2
DNAN1	24	C57BL/6	Irs2 _Luciferasa_(het)	Normal	I-49	CO2

Table S1: Supplementary table summarizing information on the mice used for this study, including identification, strain, genotype, diet, research group of origin, and the method used for euthanasia.

Nº DNA	Sample Type	Biopsy Type	Extraction Kit Batch	RNA Purification Method
DNA38	Biopsy	PC_biop	Qiagen_Kit_2_(175036888)	Dnase
DNA40	Biopsy	DC_biop	Qiagen_Kit_2_(175036888)	Dnase
DNA41	Biopsy	DC_biop	Qiagen_Kit_2_(175036888)	Dnase
DNA42	Biopsy	PC_biop	Qiagen_Kit_2_(175036888)	Dnase
DNA44	Biopsy	DC_biop	Qiagen_Kit_2_(175036888)	Dnase
DNA45	Biopsy	PC_biop	Qiagen_Kit_2_(175036888)	Dnase
DNA6	Biopsy	DC_biop	Qiagen_Kit_1_(172039651)	Dnase
DNA8	Biopsy	PC_biop	Qiagen_Kit_1_(172039651)	Dnase
DNAM1	Biopsy	DC_biop	Macherey-Nagel	Dnase
DNAM2	Biopsy	PC_biop	Macherey-Nagel	Dnase
DNAM4	Biopsy	DC_biop	Macherey-Nagel	Dnase
DNAM5	Biopsy	PC_biop	Macherey-Nagel	Dnase
DNAM6	Biopsy	DC_biop	Macherey-Nagel	Dnase
DNAM7	Biopsy	PC_biop	Macherey-Nagel	Dnase
DNA37	Blank	Blank	Qiagen_Kit_2_(175036888)	Dnase
DNABI2	Blank	Blank	Qiagen_Kit_1_(172039651)	Dnase
DNABI3	Blank	Blank	Qiagen_Kit_1_(172039651)	Beads
DNABI4	Blank	Blank	Qiagen_Kit_1_(172039651)	Dnase
DNABI5	Blank	Blank	Qiagen_Kit_1_(172039651)	Dnase
DNABI6	Blank	Blank	Qiagen_Kit_2_(175036888)	Dnase
DNABI7	Blank	Blank	Qiagen_Kit_2_(175036888)	Dnase
DNABI8	Blank	Blank	Qiagen_Kit_2_(175036888)	Dnase
DNABI9	Blank	Blank	Qiagen_Kit_2_(175036888)	Dnase
DNABIM1	Blank	Blank	Macherey-Nagel	Dnase
DNABIM2	Blank	Blank	Macherey-Nagel	Dnase
DNABIP	Blank	Blank	Promega	Dnase_liquid
DNABIP2	Blank	Blank	Promega	Beads
DNABIQ	Blank	Blank	Qiagen_Kit_1_(172039651)	Dnase
DNABIQ2	Blank	Blank	Qiagen_Kit_1_(172039651)	Dnase

DNABLANK	Blank	Blank	Qiagen_Kit_1_(172039651)	No_RNA_Purification
DNA10	Cecum	Cecum	Qiagen_Kit_1_(172039651)	Dnase
DNA36	Cecum	Cecum	Qiagen_Kit_2_(175036888)	Dnase
DNA39	Cecum	Cecum	Qiagen_Kit_2_(175036888)	Dnase
DNA43	Cecum	Cecum	Qiagen_Kit_2_(175036888)	Dnase
DNA7	Mucosa	DC_mucosa	Qiagen_Kit_1_(172039651)	Dnase
DNA9	Mucosa	PC_mucosa	Qiagen_Kit_1_(172039651)	Dnase
LRHAMNOSUS	Positive_Control	Positive_Control	NA	Dnase
ZYMOMOCK	Positive_Control	Positive_Control	NA	Dnase
DNA5	Cecum	Cecum	Qiagen_Kit_1_(172039651)	No_RNA_Purification
DNA15	Cecum	Cecum	Qiagen_Kit_1_(172039651)	Beads
DNAP5	Biopsy	DC_biop	Promega	Dnase_liquid
DNA27	Cecum	Cecum	Qiagen_Kit_2_(175036888)	Dnase
DNA33	Cecum	Cecum	Qiagen_Kit_2_(175036888)	Dnase
DNAQ10	Biopsy	DC_biop	Qiagen_Kit_1_(172039651)	Dnase
DNA17	Cecum	Cecum	Qiagen_Kit_1_(172039651)	Dnase
DNA29	Biopsy	DC_biop	Qiagen_Kit_2_(175036888)	Dnase
DNA14	Mucosa	PC_mucosa	Qiagen_Kit_1_(172039651)	Beads
DNA13	Biopsy	PC_biop	Qiagen_Kit_1_(172039651)	Beads
DNAQ6	Cecum	Cecum	Qiagen_Kit_1_(172039651)	Dnase
DNAQ3	Cecum	Cecum	Qiagen_Kit_1_(172039651)	Dnase
DNA32	Cecum	Cecum	Qiagen_Kit_2_(175036888)	Dnase
DNA35	Mucosa	PC_mucosa	Qiagen_Kit_2_(175036888)	Dnase
DNA3	Biopsy	PC_biop	Qiagen_Kit_1_(172039651)	No_RNA_Purification
DNA25	Biopsy	PC_biop	Qiagen_Kit_1_(172039651)	Dnase
DNA31	Biopsy	PC_biop	Qiagen_Kit_2_(175036888)	Dnase
DNAP12	Biopsy	PC_biop	Promega	Beads
DNAQ12	Biopsy	PC_biop	Qiagen_Kit_1_(172039651)	Dnase
DNA34	Cecum	Cecum	Qiagen_Kit_2_(175036888)	Dnase
DNA22	Mucosa	DC_mucosa	Qiagen_Kit_1_(172039651)	Dnase
DNAQ1	Cecum	Cecum	Qiagen_Kit_1_(172039651)	Dnase
DNA11	Biopsy	DC_biop	Qiagen_Kit_1_(172039651)	Beads
DNAQ2	Biopsy	DC_biop	Qiagen_Kit_1_(172039651)	Dnase
DNA21	Biopsy	PC_biop	Qiagen_Kit_1_(172039651)	Dnase
DNA26	Biopsy	DC_biop	Qiagen_Kit_2_(175036888)	Dnase
DNA28	Biopsy	DC_biop	Qiagen_Kit_2_(175036888)	Dnase
DNAP10	Biopsy	DC_biop	Promega	Beads
DNAP7	Biopsy	DC_biop	Promega	Beads
DNAP8	Biopsy	PC_biop	Promega	Beads
DNAQ7	Biopsy	DC_biop	Qiagen_Kit_1_(172039651)	Dnase
DNAQ8	Biopsy	PC_biop	Qiagen_Kit_1_(172039651)	Dnase
DNAP11	Cecum	Cecum	Promega	Beads
DNA18	Mucosa	DC_mucosa	Qiagen_Kit_1_(172039651)	Dnase
DNA20	Mucosa	DC_mucosa	Qiagen_Kit_1_(172039651)	Dnase
DNAQ4	Biopsy	PC_biop	Qiagen_Kit_1_(172039651)	Dnase
DNAP6	Cecum	Cecum	Promega	Dnase_liquid
DNAP1	Cecum	Cecum	Promega	Dnase_liquid
DNAP3	Cecum	Cecum	Promega	Dnase_liquid
DNAQ5	Biopsy	DC_biop	Qiagen_Kit_1_(172039651)	Dnase
DNA12	Mucosa	DC_mucosa	Qiagen_Kit_1_(172039651)	Beads
DNA4	Mucosa	PC_mucosa	Qiagen_Kit_1_(172039651)	No_RNA_Purification
DNA1	Biopsy	DC_biop	Qiagen_Kit_1_(172039651)	No_RNA_Purification
DNA19	Biopsy	DC_biop	Qiagen_Kit_1_(172039651)	Dnase
DNA23	Biopsy	DC_biop	Qiagen_Kit_1_(172039651)	Dnase
DNAP9	Biopsy	PC_biop	Promega	Beads
DNAQ9	Biopsy	PC_biop	Qiagen_Kit_1_(172039651)	Dnase
DNA16	Mucosa	DC_mucosa	Qiagen_Kit_1_(172039651)	Dnase
DNAP4	Biopsy	PC_biop	Promega	Dnase_liquid
DNAP2	Biopsy	DC_biop	Promega	Dnase_liquid
DNAQ11	Cecum	Cecum	Qiagen_Kit_1_(172039651)	Dnase
DNA24	Cecum	Cecum	Qiagen_Kit_1_(172039651)	Dnase
DNA2	Mucosa	DC_mucosa	Qiagen_Kit_1_(172039651)	No_RNA_Purification
DNA30	Biopsy	DC_biop	Qiagen_Kit_2_(175036888)	Dnase
DNAM3	Cecum	Cecum	Macherey-Nagel	Dnase
DNAM8	Cecum	Cecum	Macherey-Nagel	Dnase
DNAN2	Cecum	Cecum	NZYTech	Dnase
DNAN1	Biopsy	DC_biop	NZYTech	Dnase

Table S2: Supplementary table summarizing information on sample and biopsy types, along with the extraction kit batches used for all extractions in this study.

Gene	Primer	Primer sequence
16S Universal	B_Uni16S-F	**CCATGAAGTCGGAATCGCTAG**
16S Universal	B_Uni16S-R	**GCTTGACGGGCGGTGT**
16S Universal	B_Uni16S.1	**CCATGAAGTCGGAATCGCTAGTAATCGTGGATCAGAATGCCACGGTGAATACGTTCCCGGGCCTTGTACACACCGCCCGTCACAC**
16S Universal	B_Uni16S.2	**GTGTGACGGGCGGTGTGTACAAGGCCCGGGAACGTATTCACCGTGGCATTCTGATCCACGATTACTAGCGATTCCGACTTCATGG**
ACTB	M_ACTB-F	**GCAAGCAGGAGTACGATGAGT**
ACTB	M_ACTB-R	**ACGCAGCTCAGTAACAGTCC**
ACTB	M_ACTB.1	**GCAAGCAGGAGTACGATGAGTCCGGCCCCTCCATCGTGCACCGCAAGTGCTTCTAGGCGGACTGTTACTGAGCTGCGT**
ACTB	M_ACTB.2	**ACGCAGCTCAGTAACAGTCCGCCTAGAAGCACTTGCGGTGCACGATGGAGGGGCCGGACTCATCGTACTCCTGCTTGC**

Table S3: Sequences of the primers for 16S and ACTB genes used in the qPCR assays

Genera	freq	prev	Total in Blanks	Total in True Samples	p,prev
Acidocella	0,022667536	29	15	14	4,03E-09
Acinetobacter	0,034222861	22	7	15	2,31E-02
Aureimonas_A	0,005860748	17	7	10	6,18E-05
Bradyrhizobium	0,000344819	2	2	0	0,013157895
Brevundimonas	0,000999123	6	5	1	0,003487464
Burkholderiaceae Family	0,002318709	2	2	0	0,013157895
Burkholderiales Order	0,000920046	4	3	1	0,007290876
Campylobacter_B	0,000440641	6	5	1	0,031404456
Cutibacterium	0,010762876	24	7	17	3,90E-02
Dyella	0,001105147	3	2	1	0,037513998
Fenollaria	0,000634068	3	2	1	0,037513998
Flectobacillus	0,003524256	4	3	1	0,000273935
Gordonia	0,000513724	2	2	0	0,013157895
Inhella	0,000160209	2	2	0	0,013157895
JACDEK01	0,000805859	3	2	1	0,037513998
Macrococcus	0,00253989	8	7	1	0,00155105
Massilia	0,004072807	9	8	1	0,000282413
Micrococcus	0,001595347	4	3	1	0,007290876
Moraxella_A	0,011422656	13	9	4	4,30E-05
Nitrospira_E	0,000283506	2	2	0	0,013157895
Obscuribacterales Order	0,000253092	4	3	1	0,007290876
Oceanispirochaeta	0,000520798	3	2	1	0,001959686
Paenirhodobacter	0,000712489	5	4	1	0,001262481
Paraburkholderia	0,006116663	7	6	1	0,000633727
Peptoniphilaceae Family	0,00084392	8	7	1	0,01375816
Pseudocitrobacter	0,018266245	36	16	20	3,85E-08
Pseudomonadaceae Family	0,000309911	2	2	0	0,013157895
Pseudomonas_F	0,001142298	3	2	1	3,75E-02
Psychrobacter	0,000911376	4	3	1	0,007290876
Ralstonia	0,004797889	8	7	1	0,000102418
Reyranella	0,000302835	2	2	0	0,013157895
Rhizobium	0,001089574	4	3	1	0,000273935
Rhodanobacter	0,000356607	2	2	0	0,013157895
Rhodobacteraceae Family	0,000326773	2	2	0	0,013157895
Rudaeicoccus	0,000696376	2	2	0	0,013157895
Sphingomonadaceae Family	0,001100179	3	2	1	0,001959686
Sphingomonas	0,000429983	2	2	0	0,013157895
Subtercola	0,001479452	3	2	1	0,037513998
TM7x	0,003051545	2	2	0	0,013157895

Table S4: Supplementary table summarizing the prevalence of contaminant bacteria (indicated by genera) in negative controls and true samples detected by decontam R package.