UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

DEPARTMENT OF INGEGNERIA DELL'INFORMAZIONE

COURSE OF COMPUTER ENGINEERING

Bioinformatics analysis of longitudinal fecal microbiota data from pig farms

Relatore
Prof. Giacomo Baruzzo

Correlatore
Massimo Bellato

Laureando
Piermarco Giustini

ANNO ACCADEMICO 2023-2024

Data di laurea 16/04/2024

Contents

1 Introduction

The past two decades have brought about a significant shift in our understanding of the role of the microbiome in piglet physiology. Early research has shown that, similar to humans, the prokaryotic inhabitants in piglets contribute substantially to their overall "holobiont" - a term that encompasses the host and its symbiotic microbial community. These microbial populations are not only comparable in cell number to the piglet's own cells but also provide a vastly greater number of genes. This complex microbial ecosystem plays a crucial role in various aspects of piglet health and development, including digestion, metabolism, immune system functionality, and potentially even influencing neurological development.

Furthermore, alterations in the microbiome have been increasingly associated with a variety of health issues, echoing trends observed in human microbiome studies. This parallel has sparked considerable interest and hope that the expanding knowledge of host-microbiome interactions could lead to novel strategies for managing and preventing diseases in these animals. As research progresses, there is an anticipation that insights gleaned from studying the microbiome will not only enhance our understanding of animal health and veterinary medicine but might also offer valuable parallels and insights applicable to human health.

Following a bit history, the first decade of microbiome research in this millennium focused on demonstrating the importance of the gut microbiome for physiological function, coupled with next-generation sequencing-based characterization of microbiome community structure in a variety of physiological and disease contexts. Such widespread mapping has led to the identification of associations, correlations, and predictions between the microbiome and various health outcomes, reported at ever-increasing scale and detail. In parallel, the field has extensively characterized the dynamic regulation of microbial signatures in response to a variety of factors, such as host immunity and genomics. With these substantial advances also came the recognition of major technical and conceptual obstacles challenging interpretation, generalization, and translation of microbiome findings to the clinical bedside.

These challenges include the intricate complexity of microbiome compositions, which are highly individualized and subject to numerous influencing factors. This variability has made it difficult to define a *healthy* microbiome or to establish standardized benchmarks for microbial health. Additionally, while correlations between microbiome compositions and certain health conditions have been identified, establishing causality remains a significant hurdle. Many studies have been observational in nature, and thus unable to conclusively determine whether changes in the microbiome are a cause or an effect of the associated conditions.

Moreover, the field has grappled with technological limitations, particularly in the early years. Although next-generation sequencing technologies have revolutionized our ability to analyze and understand the microbiome, they also brought challenges such as data interpretation, the need for advanced computational tools, and the difficulty in capturing the full diversity of microbial life. These issues were compounded by the vast amounts of data generated, necessitating sophisticated bioinformatics approaches for analysis.

Despite these challenges, the field of microbiome research has made significant strides, evolving from basic descriptive studies to more mechanistic investigations. Current research is increasingly focusing on understanding how microbial communities interact with their host at molecular and cellular levels, and how these interactions can be manipulated for therapeutic benefit. This includes exploring the role of the microbiome in modulating the immune system, influencing metabolic pathways, and even affecting neurological functions and behaviors.

2 Materials and Methods

This section outlines the methodology and data structure used in our analysis. The primary focus of this study was to develop and implement a streamlined pipeline for processing and analyzing microbiome data. The initial dataset was received in a compressed (.zip) format, containing several key files critical to our analysis pipeline. Essential files include 'piglets_metadata.tsv', 'paired-end-demux.qza', and 'classifier.qza'. These files form the foundation of our pipeline, enabling the effective handling and analysis of the microbiome data. Subsequent processes and analyses are built upon the data and structures derived from these core files.

2.1 Study Design and Population

Data for this study were obtained from a comprehensive metadata file encompassing information on 120 pig samples. Each sample represented a family unit consisting of a sow and her three piglets, categorized into distinct cells for analysis. The metadata provided intricate details, including the count of piglets per nest, mortality rates, instances of unweaned, and underweight individuals. Further elaboration on the metadata specifics is presented in the *Metadata Section*. Accompanying this, a raw count matrix of Operational Taxonomic Units (OTUs) was also provided, serving as a foundational element for our analysis.

2.2 Sample Collection

The study meticulously captured the temporal dynamics of the microbiome by collecting fecal samples from piglets at three strategically chosen timepoints. These timepoints were: T_0 (approximately 24 hours post-birth), T_1 (ranging from 24 to 96 hours post-birth), and T_2 (spanning from 96 to 250 hours post-birth). This strategic sampling schedule was designed to provide a comprehensive view of the microbiome's developmental trajectory in the early life stages of the piglets, thereby offering valuable insights into its evolution over these critical initial hours.

2.3 DNA Extraction and Sequencing

The methodology for DNA extraction centered around 16S rRNA gene sequencing. While detailed procedural information is not extensively documented, the process adhered to the standard methodologies prevalent in 16S rRNA sequencing, ensuring reliability and consistency in DNA extraction.

2.4 Development of a Customized Bioinformatics Pipeline

Our study necessitated the creation of a tailored bioinformatics pipeline, meticulously designed with a foundation in QIIME2's official guidelines, yet custom-modified to meet the specific needs of our research. This pipeline was structured into several pivotal phases:

1. **Data Import and Preprocessing:** Initiated with the import and preprocessing of FASTQ data, this stage was pivotal for extracting vital sequencing information, and the creation of the metadata file for qiime2.
2. **Feature Classification:** This step involved generating a detailed taxonomy file through the classification of various features.
3. **Demultiplexing of Paired-end Reads:** Included here was the selection of trimming values (left and right) based on quality plots.
4. **Demultiplexing and DADA2 Processing:** Utilizing the selected trimming values, this phase produced feature tables and sequence features. Subsequent steps included:

- Construction of a phylogenetic tree, yielding outputs such as rooted and unrooted trees, aligned sequences, and masked aligned sequences.
- Employment of an external classifier to annotate feature sequences for taxonomy classification.

5. **Distance Matrix Calculation:** Distance matrices, derived from the phylogenetic tree, were used for data imputation and to cleanse features lacking phylogenetic data.
6. **Data Imputation and Table Selection:** This involved calculating imputations based on study conditions, leading to the creation of three distinct imputed feature tables: original, log-transformed, and normalized. One of these tables was then selected for further analysis.
7. **Generation of Data Tables:** ASV, genus, and species tables were generated from the selected imputed features table.
8. **Data Normalization:** Techniques such as GMPR or CLR were applied for standardization, focusing on ASV, Genus, and Species data. This phase also included:
    - Creation of a taxa-filtered table, excluding features with zero values.
    - Collection of statistical data about samples and features (e.g., minimum frequency, quartiles, maximum frequency) to inform further analysis.
    - Selection of a statistical threshold for data trimming before conducting alpha and beta diversity analyses.
9. **Diversity Analysis:** Execution of alpha and beta diversity analyses on the trimmed data.
10. **Analytical Parameter Selection for ANCOM and MaAslin2:**
    - Retrieval of taxa-filtered data prior to normalization and statistical evaluation.
    - Selection of a statistical threshold for trimming data across the three tables: ASV, Genus, and Species.
11. **Advanced Analysis:** Performing ANCOM differential abundance analysis and MaAslin2 analysis on the prepared tables.

2.5 Computational Environment and Replication Technologies

The computational framework for this study was anchored in Visual Studio Code (VSCode), which offered an integrated development environment for scripting and programming in Bash, Python, and R. This versatile setup ensured an efficient and adaptable approach to handling the diverse computational needs of the project. Version control and data backup were meticulously managed using GitHub (Git Hub), enhancing the robustness and traceability of our computational work.

Emphasizing the replicability of the research, Docker and Conda environments were strategically utilized. These technologies provided a consistent and controlled environment for the analysis, ensuring that our computational processes could be reliably replicated in different settings. Qiime2, particularly its .qza and .qzv file formats, played a pivotal role in the visualization and interpretation of the study's results. The scripts developed for this study are fully automated, offering the flexibility to conduct various analyses based on the specific needs of the user.

In documenting this research journey, Overleaf was chosen as the primary tool for writing the thesis. Its LaTeX-based platform offered a streamlined and efficient environment for academic writing, facilitating the integration of complex computational and scientific content into a well-structured thesis document.

2.6 Automated Code

The core of our computational strategy for this study is encapsulated in a fully automated codebase. This automation is not just about efficiency; it's about adaptability and precision in handling diverse datasets and experimental conditions. Our scripts are designed to seamlessly adjust to different settings and parameters, a feature that has been indispensable in navigating the myriad of choices encountered in developing a robust bioinformatics pipeline.

This flexibility in our automated system allowed us to experiment with various configurations and parameters, enabling us to optimize our methodologies based on the specific needs of the study. The ability to rapidly iterate and test different scenarios has been a key factor in determining the most effective bioinformatics approaches and the arising of errors.

3 Metadata

The metadata table provides an overview of the collected samples and associated attributes. Each row represents a unique sample with details such as the sample ID, serial number, swab ID, animal identification, and information about the sow. It also includes data on the room where the sample was collected, the sex of the piglet, health indicators like the presence of diarrhea, and additional details about the sow's gestation, the nest, and the condition of the piglets (alive, dead, transferred, weaned). The last column provides information on treatments or medications administered to the sow.

The metadata are also adapted to be used in the qiime2 artifact, infarct the first row correspond to the the name of the metadata, the second one correspond to the the $q2:types$ that differes in 2 different type, *numerical* for numerical features and *categorical* for the categorical features, here a small example 1.

Sample ID	**Serial**	**Animal ID**	**...**	**Sow**	**Sex**	**Diarrhea**
#q2:types	numeric	categorical	...	numeric	categorical	categorical
1450087F1381048_S1_L001	1381048	S1_P0_T0	...	6247	f	n
1450088F1381049_S2_L001	1381049	S1_P1_T0	...	6247	m	e
1450089F1381050_S3_L001	1381050	S1_P2_T0	...	6247	f	e
1450090F1381051_S4_L001	1381051	S1_P3_T0	...	6247	m	e

Table 1: Sample metadata for piglet study, including data types

3.1 Manipulation of Metadata

In the dataset, significant modifications were made to the $ANIMAL_{ID}$ column to enhance the clarity and utility of the data for analysis. These modifications involved the addition of three new columns, each serving a specific purpose:

1. **Time Column**: A new column named $time$ was added to the dataset. This column is numerical, with values $\{0, 1, 2\}$, corresponding to different time points indicated in the $ANIMAL_{ID}$ column as $\{$T0, T1, T2$\}$. This modification aids in quantitatively representing the time aspect of the data.

2. **Sow Column**: Another column, *sow*, was introduced. This is also a numerical column, containing values ranging from 1 to 10. These values are directly mapped from the $ANIMAL_{ID}$ column, where they are represented as $\{$S0, S1, S2, ..., S10$\}$. This column is crucial for identifying and segregating data based on the sow identifier.

3. **Sow-Son Column**: The third column added is sow_{son}. It holds numerical values from 10 to 100, representing a combination of sow and piglet identifiers. This column is created by mapping combinations found in $ANIMAL_{ID}$ such as $\{$S1_P0, S1_P1, S2_P2, ..., S10_P0$\}$. The sow_son column is particularly useful for analyses that require understanding the relationship between sows and their offspring.

3.2 Percentage Analysis in the Study

The primary aim of this study is to explore the potential correlation between the number of piglets in a nest and the survival rate of these piglets. This investigation is crucial as it provides insights into factors affecting piglet mortality and survival in varying nest sizes.

3.2.1 Challenges with Numerical Values

The data at hand presents a challenge due to its purely numerical nature, particularly when comparing piglet counts across nests of different sizes. For example, comparing a nest with 10 piglets and 3 fatalities to a nest with 20 piglets and the same number of fatalities is not straightforward. The direct numerical comparison does not account for the proportional differences in nest sizes, leading to potential misinterpretations in the data.

3.2.2 Methodological Approach

To overcome this limitation and enable a more meaningful comparison, I have implemented a strategy that involves removing certain columns from the dataset. These columns are replaced with categorical ones, which are more suitable for the analysis. Specifically, the use of percentage calculations and mean values allows for a normalized comparison

across different nest sizes. This approach ensures that the data is evaluated in a context that accurately reflects the varying conditions of each nest, providing a clearer understanding of the relationship between nest size and piglet survival.

3.2.3 Proposed Solution

In order to enhance the interpretability and analytical utility of the dataset, a methodological refinement was implemented involving the creation of new columns based on percentage values rather than raw numerical data. This strategic transformation facilitates a more nuanced and insightful understanding of the data, particularly in the context of alpha and beta diversity analyses, as well as in the assessment of differential abundances. By shifting the focus to percentage-based metrics, it becomes possible to achieve a clearer and more meaningful overview of various clusters, thereby enriching the depth and quality of our analytical conclusions.

3.3 Correctness of Metadata: Addressing Inconsistencies in Data

The dataset includes five key columns that are intricately linked to the number of piglets in various states:

1. **Nest**: Represents the total number of piglets in the nest.
2. **Alive**: The count of piglets that are alive.
3. **Dead**: The number of piglets that are deceased.
4. **Uw_el**: The count of piglets that are underweight.
5. **Transferred**: Number of piglets transferred at time T0.

3.3.1 Study Over Time: Data Limitations

During the course of the study, several limitations regarding the temporal aspects of the data were identified:

- The exact time of piglet deaths is not known.
- The timing of piglet transfers is assumed to be at T0, as it is not explicitly provided in the metadata.
- The point in time when piglets become underweight is not recorded in the data.
- While the final count of alive piglets is known, the exact time corresponding to this count is unclear.

To address these challenges and ensure data accuracy, the following formulas were derived and applied:

$$alive_{T0} = nest_{T0} - dead_{T0} \tag{1}$$

$$weaned_{T2} = alive_{T0} - transferred_{T1T2} - underweight_{T1T2} \tag{2}$$

These formulas are crucial for reconciling the data with the actual events and timelines, thereby enhancing the reliability and interpretability of the metadata.

3.3.2 Identification of Missing Piglets in the Data

The equations previously delineated are largely consistent across the majority of samples in our metadata TSV file. To verify our hypotheses and ensure the integrity of the data, a Python script was executed to systematically validate these equations across the dataset. During this verification process, an intriguing pattern emerged: Equation (1) consistently holds true for all samples, indicating reliable data in those instances.

However, a deviation was observed in the case of Equation (2), where certain samples did not conform to the expected results. This discrepancy suggests the possibility of errors in the data, potentially stemming from inaccurate data entry or misaligned data recording processes. Such findings necessitate a closer examination of the dataset to identify and rectify these anomalies, thereby ensuring the accuracy and reliability of our analysis.

3.3.3 Proposed Solution

The resolution to the observed data discrepancies can be approached from different perspectives, each varying in simplicity. Firstly, an interactive Python script offers a straightforward solution: generating a new TSV file with the exclusion of the $transferred$ and uw_{el} columns. This approach effectively eliminates any misleading data that could arise from the problematic equations.

Alternatively, a more complex method involves manipulating the dataset to account for the disparities in piglet numbers. This could entail storing the differences in piglet counts in one of the two columns. However, this approach introduces a conceptual paradox, as it obscures the underlying reasons for the reduced number of weaned piglets and fails to clarify which factor is primarily responsible. Furthermore, altering the values might lead to more confusion rather than clarity, as it does not address the core issue of unknown variables affecting the $transferred$ and uw_{el} columns. Consequently, manipulating the data in this manner seems less feasible, and it might be more prudent to disregard the $transferred$ and uw_el columns in our analyses to avoid the introduction of potential biases and inaccuracies.

3.4 Metadata Analysis

The correlation matrix plotted 2 is a powerful tool used in statistical analysis to understand the strength and direction of relationships between multiple variables. Values close to +1 or -1 in the matrix indicate a strong positive or negative correlation, respectively, suggesting a direct or inverse relationship between the corresponding variables. Conversely, values near 0 imply a lack of significant correlation, indicating that the variables do not exhibit a noticeable positive or negative relationship. This interpretation of the correlation matrix is pivotal in identifying patterns and relationships in complex datasets [6]. I have plot the figures for the correlation data of my metadata

Instead the Principal Component Analysis (PCA) plotted 3 is a statistical procedure that utilizes an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. This technique is widely used to simplify the complexity in high-dimensional data while retaining trends and patterns. PCA is particularly powerful in identifying the dominant patterns in the data, reducing the dimensions of the data without much loss of information, and visualizing the structure of the data in a simple form [13].

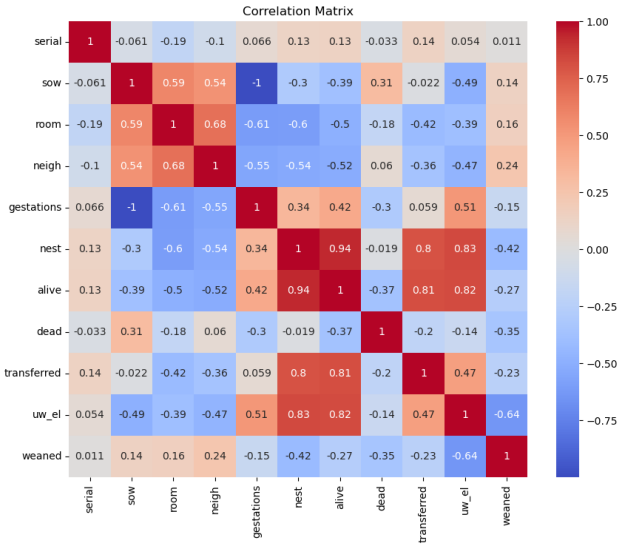


Figure 2: Correlation Matrix of Raw Metadata, illustrating the correlation coefficients between different data points. Values close to 1 indicate a strong positive correlation, exemplified along the diagonal where all values are exactly 1. Conversely, values approaching -1 signify a strong inverse correlation. Coefficients near 0 imply a lack of both positive and negative correlation, suggesting that these relationships may not be statistically significant or meaningful.

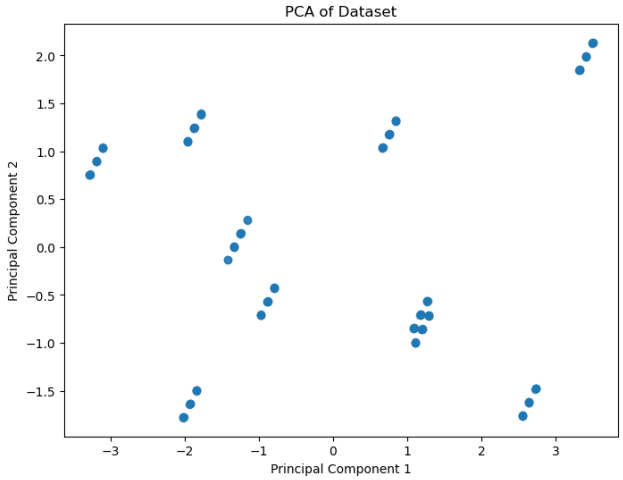


Figure 3: Principal Component Analysis (PCA) of Metadata: The figure shows the variance explained by each principal component, with values [0.46673002, 0.16316995], indicating the proportion of the dataset's variance that each component accounts for. The first component explains a significant portion of the variance, suggesting it captures the primary trend in the data.

3.5 Definitions of Fecal Sample Types in Piglets

The intestine of pigs harbors a mass of microorganisms which are essential for intestinal homeostasis and host health. Intestinal microbial disorders induce enteric inflammation and metabolic dysfunction, thereby causing adverse effects on the growth and health of pigs [8].

Hematic Diarrhea Sample: Characterized by the presence of blood in the diarrhea. It indicates gastrointestinal issues such as ulcers, infections, parasitic infestations, or diseases causing bleeding in the gastrointestinal tract. The feces are typically loose, watery, and contain fresh or digested bloo .

Non-Diarrhea Sample: Refers to normal, healthy feces. For piglets, this means well-formed feces, consistent in texture, and without abnormalities such as blood or excessive water content. Indicates good gastrointestinal health.

Diarrhea Sample: Characterized by loose, watery, and possibly more frequent bowel movements. Causes include dietary changes, infections, stress, or underlying health conditions. Unlike hematic diarrhea, standard diarrhea samples do not contain visible blood.

4 Qiime2

Quantitative Insights Into Microbial Ecology 2 (QIIME 2) is a powerful, extensible platform that enables users to analyze and interpret complex microbiome data. QIIME 2's versatility allows users to start their analysis at various stages of the data processing pipeline, depending on the nature and format of their raw data. This flexibility is crucial in microbiome research, where data types and processing stages can vary widely.

4.1 Data Import and Workflow Entry Points

Users typically begin their QIIME 2 journey with raw sequencing data, such as FASTQ or FASTA files. These files contain DNA sequences along with quality scores for each base. The first critical step is importing this raw data into QIIME 2, which converts it into a QIIME 2 artifact (.qza). The .qza format is a fundamental component of the QIIME 2 framework, encapsulating data in a structured manner that integrates relevant metadata and provenance information. This contrasts with the .qzv format, which represents visualization artifacts. While .qza files are for data storage and analysis, .qzv files are essentially reports or visual summaries of the data or analysis results.

4.2 Key Processing Steps in QIIME 2

- Demultiplexing: This step involves sorting the raw sequence data to identify the sample source of each read.
- Denoising or Clustering: Sequences are either denoised into Amplicon Sequence Variants (ASVs) or clustered into Operational Taxonomic Units (OTUs). This process aims to reduce sequence errors and dereplicate sequences, creating a cleaner, more accurate dataset.
- Creation of Feature Table and Representative Sequences: The outcome of denoising or clustering is a feature table and a set of representative sequences. The feature table is a crucial element in QIIME 2; it's a matrix that correlates samples with observed features (like OTUs or ASVs), quantifying the presence of these features in each sample.

4.3 Analytical Capabilities in QIIME 2

With the processed data in a feature table, QIIME 2 enables a wide array of analyses:

- Taxonomic Classification: This involves identifying the species or taxonomic groups present in the samples.
- Diversity Analysis (Alpha and Beta Diversity): These analyses assess the diversity within individual samples (alpha diversity) and the comparative diversity between samples (beta diversity).
- Phylogenetic Analysis: For sequencing phylogenetic markers (e.g., 16S rRNA genes), QIIME 2 can align sequences to explore phylogenetic relationships among features.
- Differential Abundance Analysis: This identifies features that significantly vary in abundance across different experimental conditions or groups.

4.4 Understanding Terminology and Data in Qiime2

Understanding the data and terminology used in microbiome studies, particularly when working with tools like QIIME2, requires a nuanced approach. This section aims to clarify the terminology and data representation used throughout this report, focusing on the context of a study involving piglets.

Firstly, the term *sample* in this context refers specifically to an individual piglet. In our dataset, the total number of piglets, or samples, is 120. Each sample represents a unique set of data points collected from one piglet, which includes various measurements and observations pertinent to the study.

On the other hand, when we mention *feature* is typically represented by an alphanumeric code in the dataset. To the uninitiated, this representation might seem obscure. However, it's important to understand that each of these alphanumeric codes corresponds to a distinct *taxonomic classification*.

For each feature, a confidence level is assigned, which indicates the reliability or accuracy of the identification of the microbiota environment. A higher confidence score suggests greater certainty about the feature's characterization. This score is crucial for ensuring that the analyses and conclusions drawn from the data are based on accurate and reliable identifications of the microbiota environments.

Also the features sequences are present where we have the same alpha code number but in this case we directly refer to *AGTC* read, In that table 3 we can see the statistics for our read from our raw FASTA file.

In summary, the dataset comprises 120 samples, each representing a piglet, and numerous features precisely 3020, each corresponding to a unique taxonomic classification within these samples. The confidence scores associated with each feature add an additional layer of reliability to our understanding of the microbiota composition in each piglet. There is a small sample of our taxonomic classification file 2.

Feature ID	Taxon	Confidence
47f3d645d9603837 1757074de1d8fb8d	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__fragilis	0.95
875f9ef9a121a300 2ea097444690a20a	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Clostridiaceae; g__Clostridium; s__perfringens	0.92
ffc36e27c8204266 4a16bcd4d380b286	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacteriales; f__Enterobacteriaceae	0.89

Table 2: Examples of features and their corresponding taxonomic classifications with confidence levels

Descriptive Statistics		Seven Number Summary	
Statistic	**Value**	**Quantile**	**Value**
Count	2786	2%	386
Minimum	282	9%	402
Maximum	452	25%	403
Mean	410.803	50% (Median)	405
Range	170	75%	422
Standard Deviation	20.4665	91%	427
		98%	428

Table 3: Combined Descriptive Statistics and Seven Number Summary of Feature Sequences

5 Demultiplexing Sequences

Demultiplexing in the context of sequencing data refers to the process of identifying and separating sequences that originate from different samples. When multiple samples are sequenced together, a unique barcode sequence is added to each sample's sequences. Demultiplexing involves using these barcode sequences to sort the combined sequences back into individual samples. In QIIME 2, the demultiplexing process requires knowledge of the barcode sequence associated with each sample. In our case, we have the input sequence of forward and reverse reads shown here.

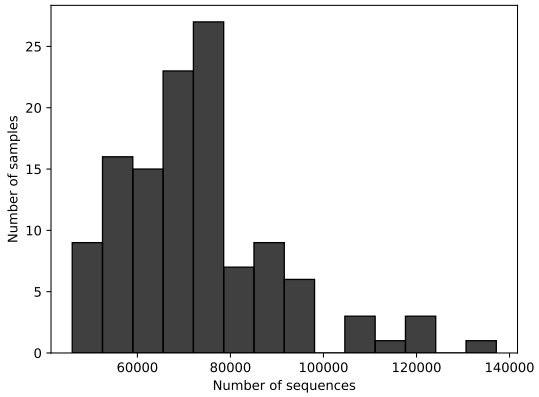


Figure 4: **Forward Reads** Histogram of Sequencing Read Counts in Samples: The x-axis of the plot represents the number of sequencing reads, indicating the quantity of DNA sequences obtained for each sample. The y-axis shows the frequency, denoting the number of samples that correspond to each specific read count range. This visualization helps in understanding the distribution of sequencing reads among the sampled population.

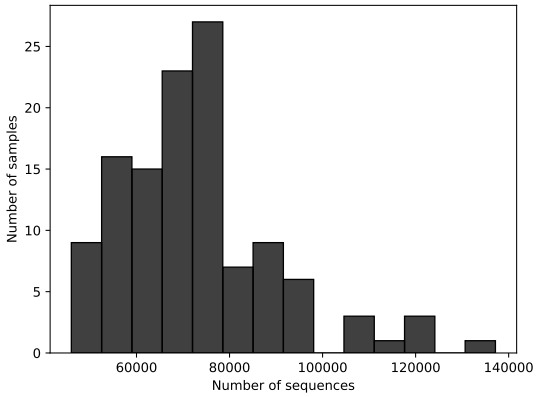


Figure 5: **Backward Reads** Histogram of Sequencing Read Counts in Samples: The x-axis of the plot represents the number of sequencing reads, indicating the quantity of DNA sequences obtained for each sample. The y-axis shows the frequency, denoting the number of samples that correspond to each specific read count range. This visualization helps in understanding the distribution of sequencing reads among the sampled population.

Table 4: Statistics for **Forward and Reverse Reads**

Statistic	**Forward Reads**	**Reverse Reads**
Minimum	45,991	45,991
Median	71,013.5	71,013.5
Mean	72,808.96	72,808.96
Maximum	137,157	137,157
Total	8,737,075	8,737,075

After completing the initial demultiplexing step, a vital preprocessing task is the trimming of sequences. This process entails the removal of the **primer sequences** from the left (forward) and right (reverse) ends of each sequence. Trimming is essential for ensuring the accuracy of subsequent analyses by eliminating non-biological sequences that could otherwise distort the results. Once trimming is accomplished, the sequences are prepared for further steps, such as denoising and taxonomic classification. The starting count of our sequences, as indicated in the trimmed demultiplexing quality plot, stands at *7,482,619*. However, a substantial reduction in this number is anticipated due to the sequence overlap analysis inherent in the DADA2 method [3].

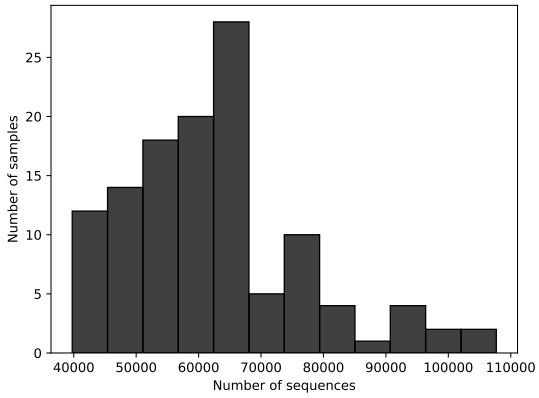


Figure 6: **Trimmed Forward Reads** Histogram of Sequencing Read Counts in Samples: The x-axis of the plot represents the number of sequencing reads, indicating the quantity of DNA sequences obtained for each sample. The y-axis shows the frequency, denoting the number of samples that correspond to each specific read count range. This visualization helps in understanding the distribution of sequencing reads among the sampled population.

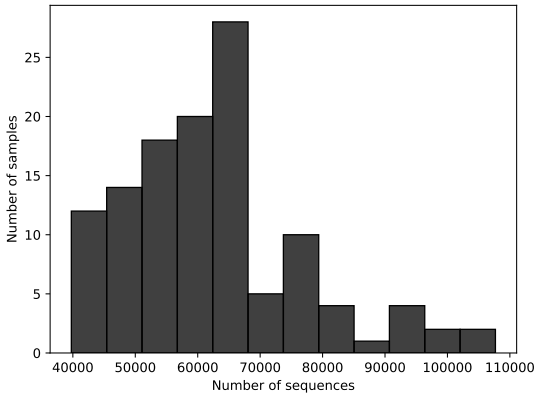


Figure 7: **Trimmed Backward Reads** Histogram of Sequencing Read Counts in Samples: The x-axis of the plot represents the number of sequencing reads, indicating the quantity of DNA sequences obtained for each sample. The y-axis shows the frequency, denoting the number of samples that correspond to each specific read count range. This visualization helps in understanding the distribution of sequencing reads among the sampled population.

Table 5: Statistics for **Trimmed Forward and Reverse Reads**

Statistic	**Forward Reads**	**Reverse Reads**
Minimum	39,756	39,756
Median	61,078.5	61,078.5
Mean	62,355.15	62,355.15
Maximum	107,687	107,687
Total	7,482,619	7,482,619

5.1 Quality Plots

After determining the quantity of sequences obtained per sample and summarizing their distribution, it is crucial to evaluate the quality of these sequencing results. Not all reads generated during the sequencing step are of sufficient quality to be utilized in the subsequent denoising process. To assess this, we employ quality plots for both forward and reverse reads.

Quality plots are essential tools in next-generation sequencing (NGS) data analysis. They visually represent the quality scores of nucleotides across all reads, offering a snapshot of the data quality. These plots typically display the **Phred quality score**, which indicates the probability of an error in base calling as a **box plot**. The higher the score, the lower the probability of an error.

5.1.1 Phred in Quality Score

The Phred quality score is an integral metric in Next-Generation Sequencing (NGS) for assessing the accuracy of base calls. It quantifies the confidence level of each nucleotide identification in the sequence data.

The Phred score (denoted as Q) is calculated using the probability P that a base call A, G, T, C is incorrect. The relationship is given by the formula:

$$Q = -10\log_{10}(P) \tag{3}$$

where:

- Q is the Phred quality score.
- P is the probability of an incorrect base call.
- A higher Q value indicates a lower probability of error, and vice versa.

5.1.2 Interpretation of Scores

- A Phred score of 20, which corresponds to $P = 0.01$, indicates a 1 in 100 chance of an error in the base call.
- A Phred score of 30, equating to $P = 0.001$, implies a 1 in 1000 chance of error.

Quality scores across all reads in NGS data are often visualized using box plots for each base position in the reads.

5.1.3 Box Plot Structure

- **X-Axis:** Represents the position in the sequencing read.
- **Y-Axis:** Shows the Phred quality score.
- **Box:** The box indicates the interquartile range (IQR) of the quality scores.
- **Median Line:** A line within the box shows the median quality score.
- **Whiskers:** Extend from the box to indicate the range of the data, with points outside representing outliers, in this case are not present, due to the conformation of our data.

5.1.4 Interpreting Box Plots

- A high median score indicates high confidence in base calls at that position across all reads.
- Consistent quality scores across reads are shown by a narrow IQR (box).
- Wide variations or low scores, especially towards the end of reads, suggest lower confidence in base calls, guiding decisions on sequence trimming or filtering.

5.2 Quality Plots calculation and treshold

We visualize multiple quantiles of quality scores at each position, such as the 2%, 9%, 25%, and 50% quantiles. These quantiles help in understanding the distribution of quality scores at each position in the read.

For practical quality assessment, certain threshold values are commonly used. These include quality scores of 15, 20, and 25, which correspond to error probabilities of 0.03%, 0.01%, and 0.003%, respectively. In quality plots, these thresholds are represented as horizontal lines. Reads with quality scores falling below these thresholds are generally considered unreliable and are often filtered out or trimmed in the preprocessing steps.

5.2.1 Trimming position Calculation

In this section we see how it is possible to calculate a good trimming position based on the quality plots.

$$Q_p(x) = \text{Percentile Quality Score at position } x \text{ for percentile } p \tag{4}$$

$Q_p(x)$ represents the quality score at a specific position x in the read for a given percentile $p = [2\%, 9\%, 25\%, 50\%]$.

$$T_q = [15, 20, 25] \tag{5}$$

T_q is the threshold line at a constant quality score for q.

$$\text{Right}_{trim} = \min\{i \mid Q_p(i) < T_q\} \tag{6}$$

This formula finds the smallest index i where the median quality score falls below the specified quality threshold.

$$\text{Left}_{trim} = \max\{i \mid Q_p(i) > T_q\} \tag{7}$$

This calculates the largest index i where the median quality score exceeds the specified quality threshold.

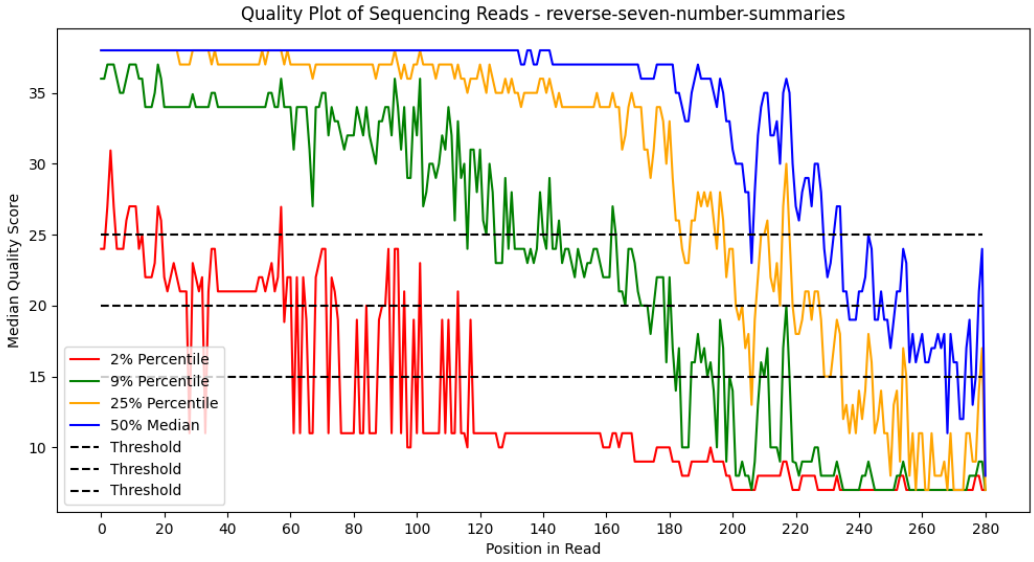


Figure 8: **Reverse Quality Plot:** This plot provides an insightful visualization of sequencing quality across nucleotide positions. The x-axis represents the position of nucleotides in each read. On the y-axis, we display the quality scores corresponding to various percentiles for each nucleotide position. Specifically, four key percentiles are represented through line plots, showing the 2%, 9%, 25%, and 50% (median) quantiles. Additionally, three horizontal lines are included to indicate predetermined error thresholds. These thresholds are critical in determining the appropriate point for initiating trimming. The intersection of the median line (in blue) with the first error threshold line (in black) suggests the optimal starting point for trimming the sequences.

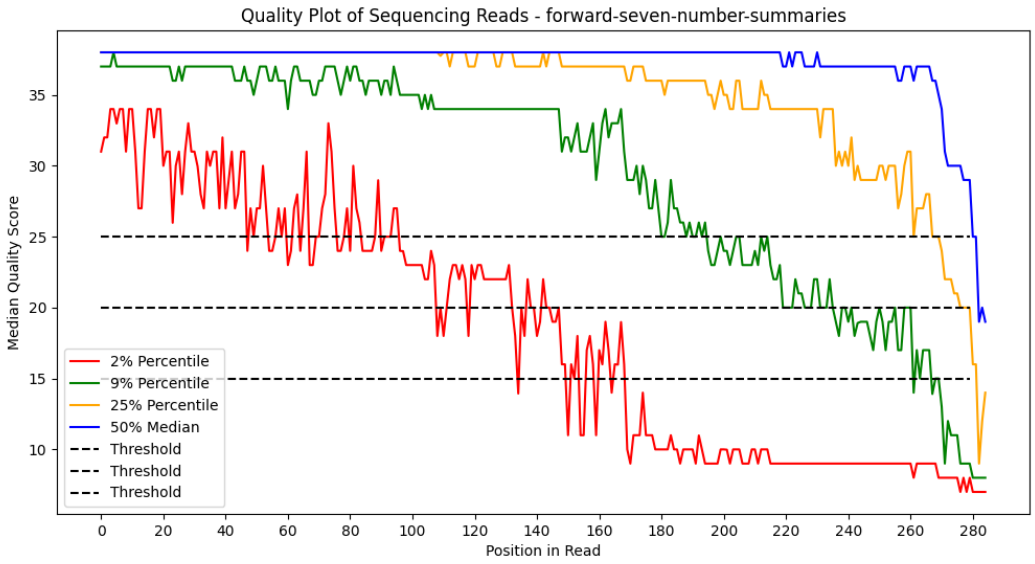


Figure 9: **Forward Quality Plot:** This plot provides an insightful visualization of sequencing quality across nucleotide positions. The x-axis represents the position of nucleotides in each read. On the y-axis, we display the quality scores corresponding to various percentiles for each nucleotide position. Specifically, four key percentiles are represented through line plots, showing the 2%, 9%, 25%, and 50% (median) quantiles. Additionally, three horizontal lines are included to indicate predetermined error thresholds. These thresholds are critical in determining the appropriate point for initiating trimming. The intersection of the median line (in blue) with the first error threshold line (in black) suggests the optimal starting point for trimming the sequences.

In these plots 8 9, we can observe the distribution of quality scores across the length of the reads. The highlighted quantile lines (2%, 9%, 25%, 50%) provide insight into the variability of quality at each position. Additionally, the marked thresholds (quality scores of 15, 20, 25) serve as benchmarks to gauge the proportion of reads meeting the desired quality standards. Below also the trimming value based on the quality score choosed, to note is the left trim that is always 0 because the quality of the read in the left site for both are pretty high so we does not need to trim anything there.

Table 6: Trimming positions at different quality thresholds (50% quantile)

File	Threshold	Left Trim	Right Trim
Forward-Seven-Number-Summaries	15	0	284
	20	0	282
	25	0	282
Reverse-Seven-Number-Summaries	15	0	268
	20	0	237
	25	0	206

In our analysis, we have opted for a more conservative approach by setting the threshold quality score at **25**. This decision aims to maintain a higher standard of data integrity and reliability.

6 Denoising DADA2

In the denoising step we are going to use the DADA2 algorithm, implemented in QIIME 2, is a widely used method for this purpose. DADA2 stands out as a pivotal software package in the realm of microbial ecology, specifically tailored for processing Illumina-sequenced amplicon data. DADA2 builds on its predecessor, DADA, a model-based approach that eschewed the construction of OTUs for error correction.

The new DADA2 software implements a quality-aware model specifically for Illumina amplicon errors. It operates reference-free, applicable to any genetic locus, and manages the complete amplicon workflow including filtering, dereplication, sample inference, chimera identification, and merging of paired-end reads. The main step on how DADA2 works can be found in this paper [4].

6.1 Qiim2 DADA2 key points

- **Error Model**: DADA2 learns the error rates of the sequencing process. It models the probability $P_{err}(r|a)$ that a nucleotide a is read as r due to an error.
- **Inference Algorithm**: DADA2 uses a sample-specific error model and a variant of the expectation-maximization algorithm to infer the most likely set of true ASVs present in the sample.
- **Quality Score-Based Filtering**: Based on user-defined thresholds, DADA2 trims and filters the reads. Trimming is done to remove low-quality tails of reads, which are more prone to errors. The parameters '–p-trim-left-f', '–p-trim-left-r', '–p-trunc-len-f', and '–p-trunc-len-r' control these aspects, this value are directly taken from the table created with quality plot 6.

6.2 Qiime2 DADA2 Outputs

The outputs provided by the denoising step using the value of 25 **as threshold** hence $[0, 282]$ for forward and $[0, 208]$ for the reversi in Qiime2 are:

- **Feature Table**: This is provided as a .qza artifact and can be conceptualized as a matrix where the rows represent samples and the columns represent features.

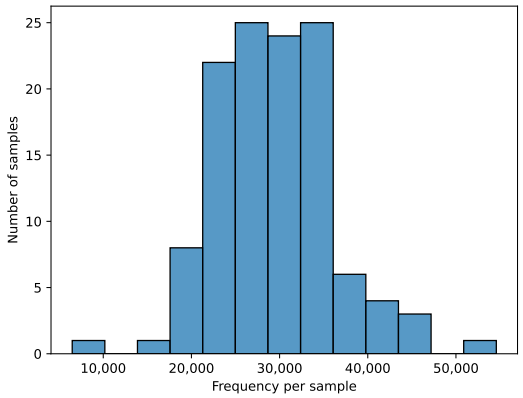


Figure 10: **Sample Frequency:** This graph provides a detailed representation of the frequency distribution within the samples. The x-axis displays the frequency count, indicating the number of occurrences or counts recorded in the dataset. On the y-axis, we have the number of samples that correspond to each frequency count.

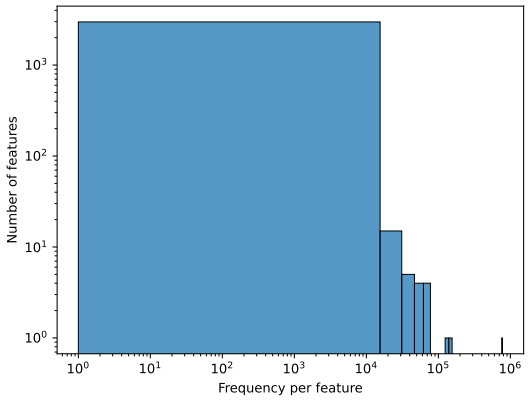


Figure 11: **Feature Frequency:** The graph presents an insightful view into the frequency distribution across features. The x-axis illustrates the frequency per feature, showing the count or number of occurrences for each feature within the dataset. Correspondingly, the y-axis indicates the number of features at each frequency level.

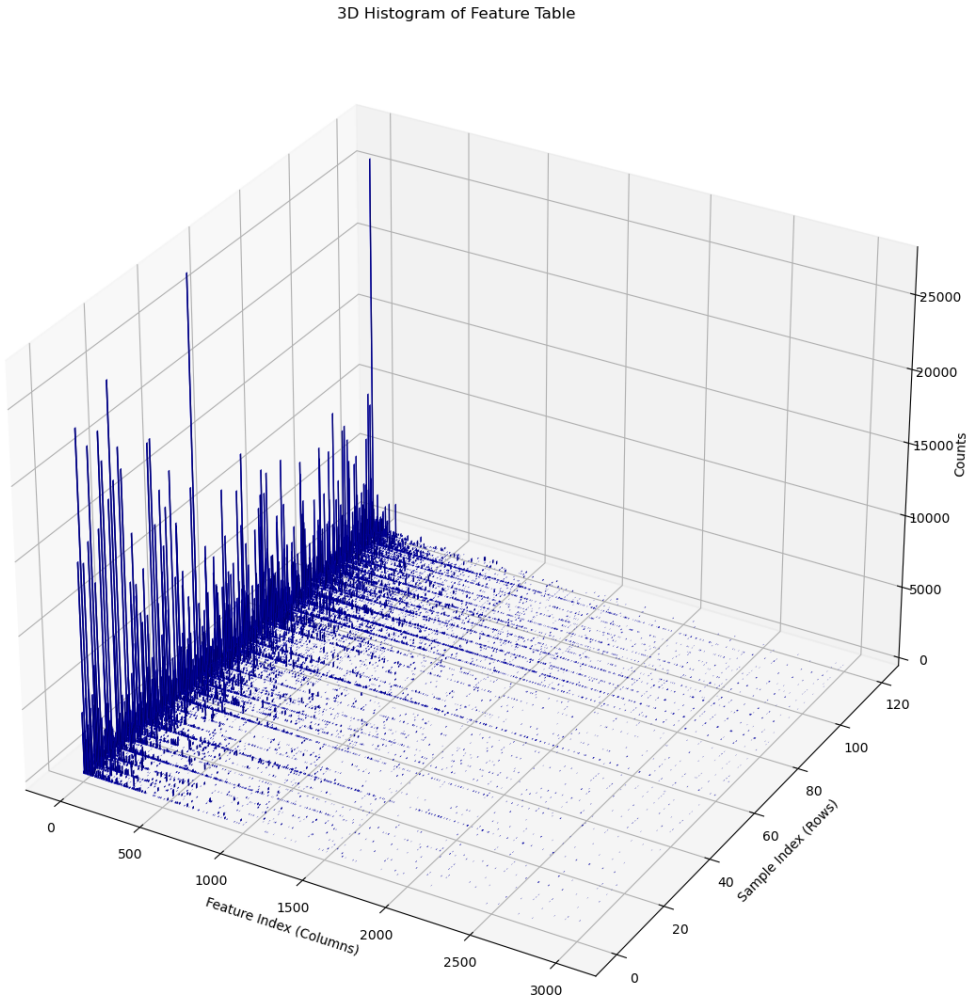


Figure 12: **3D Histogram of the Non-Imputed Feature Table:** This figure presents a comprehensive 3D histogram. On the x-axis, we display the feature index, representing the total number of features in our count table. The z-axis corresponds to the number of samples. Finally, the y-axis quantifies the count of features for each bin, providing a detailed view of the distribution across the dataset.

Table 7: Overview of Table Summary for **Non-Imputed Feature Table**

Metric	**Value**
Number of Samples	120
Number of Features	3,011
Total Frequency	3,532,911

Table 8: Detailed Frequency Metrics for **Non-Imputed Feature Table**

	Frequency Per Sample	Frequency Per Feature
Minimum Frequency	6,481.0	1.0
1st Quartile	24,772.75	13.0
Median Frequency	29,157.5	48.0
3rd Quartile	33,583.5	243.0
Maximum Frequency	54,579.0	775,782.0
Mean Frequency	29,440.925	1,173.33

- **Feature Sequences**: Also provided as a .qza artifact, this can be visualized as a table with two columns: Feature ID and Sequence. Each row corresponds to an alphanumerical ID representing features and the associated sequences in nucleotide format (e.g., AGTC).
- **Denoising Stats**: This is a .qza artifact that provides a summary of the actions performed by DADA2 for each sample, including information on the number of sequences processed, filtered, denoised, merged, and non-chimeric.

7 Taxonomic Classification

We explore the hierarchical taxonomy present within our dataset, encompassing up to seven distinct levels of classification. These taxonomic levels delineate various stages of granularity in characterizing the microbial taxa present in our dataset. The following figures depict the taxonomic hierarchy and the corresponding legends associated with each bacterial entity, commencing from the initial level (Level 1), where we encounter the broad classification of two major bacterial families, and extending through the succeeding six levels of taxonomy (Levels 2 through 7).

7.1 Taxon Level and Their Meaning

In biological classification, the acronym KPCOFGS represents a hierarchy of levels, each of which is significant for the classification of organisms. This system is instrumental in understanding the relationships and characteristics of various organisms, particularly within the fecal microbiome environment. The fecal microbiome is a complex and dynamic community of microorganisms present in the gastrointestinal tract, playing a crucial role in health and disease. Here, we will briefly describe each taxonomic level with respect to its relevance in fecal microbiome studies:

- **Kingdom:** The highest level of classification that groups together all forms of life with fundamental similarities. In the context of fecal microbiome, the Kingdom predominantly includes Bacteria and Archaea.
- **Phylum:** This level under the kingdom makes a more distinct group of organisms. For example, in fecal microbiome studies, common phyla include Firmicutes and Bacteroidetes.
- **Class:** This category groups organisms that share even more specific characteristics. For example, the Clostridia class is a significant component within the Firmicutes phylum in the gut.
- **Order:** A subdivision that groups organisms even more closely. For instance, the order Bacteroidales is predominant within the Bacteroidetes phylum in the gut microbiome.
- **Family:** This category brings together a group of related genera. An example is the Lachnospiraceae family within the Clostridia class, often found in the human gut.
- **Genus:** A genus comprises one or more species that are closely related. For instance, the genus Bacteroides is commonly studied in fecal microbiome analyses.
- **Species:** The most specific level of classification, a species is a group of organisms that can interbreed and produce fertile offspring. A familiar species in gut microbiome studies is Escherichia coli.

Each of these taxonomic levels provides crucial insights into the composition and function of the fecal microbiome all the information can be found on [1].

7.2 Build Taxonomy Classification in Qiime2

Taxonomy classification in Qiime2 was executed using a specialized script that utilizes the **classifier.qza**. This classifier, integral to the Qiime2 platform, employs machine learning algorithms to accurately categorize microbial sequences into taxonomic groups. The classifier.qza is a pre-trained model informed by reference taxonomic data, which enables the precise identification of taxa in the microbiome samples based on their sequence data. This step creates the **taxonomic.qza** 2.

7.3 Build Phylogenetic Tree in Qiime2

The construction of a phylogenetic tree is a critical step in many bioinformatics analyses, especially in studies of microbial communities. This process involves the generation of a tree that represents the evolutionary relationships among various taxa (such as species or operational taxonomic units) present in the dataset.

In QIIME 2, the creation of a phylogenetic tree is facilitated by the command `align-to-tree-mafft-fasttree`, which is part of the QIIME 2 pipeline. This command performs two key functions:

1. **Sequence Alignment**: The first step is to align the sequences present in the **feature_sequences.qza** file. This file, generated in the previous step of the workflow, contains the representative sequences for each feature identified in the samples. The alignment process arranges these sequences in a way that positions homologous (evolutionarily related) characters in the same column.

2. **Tree Construction**: Once the sequences are aligned, the command then utilizes this alignment to construct a phylogenetic tree. The tree-building algorithm used in QIIME 2 is **FastTree**, which efficiently generates an approximate maximum-likelihood tree. The resulting tree illustrates the inferred evolutionary relationships among the sampled taxa.

The phylogenetic tree thus produced is a fundamental component for subsequent analyses like diversity calculations and phylogenetic-aware metrics. It provides a context for interpreting the ecological and evolutionary dynamics of the microbial community under study.

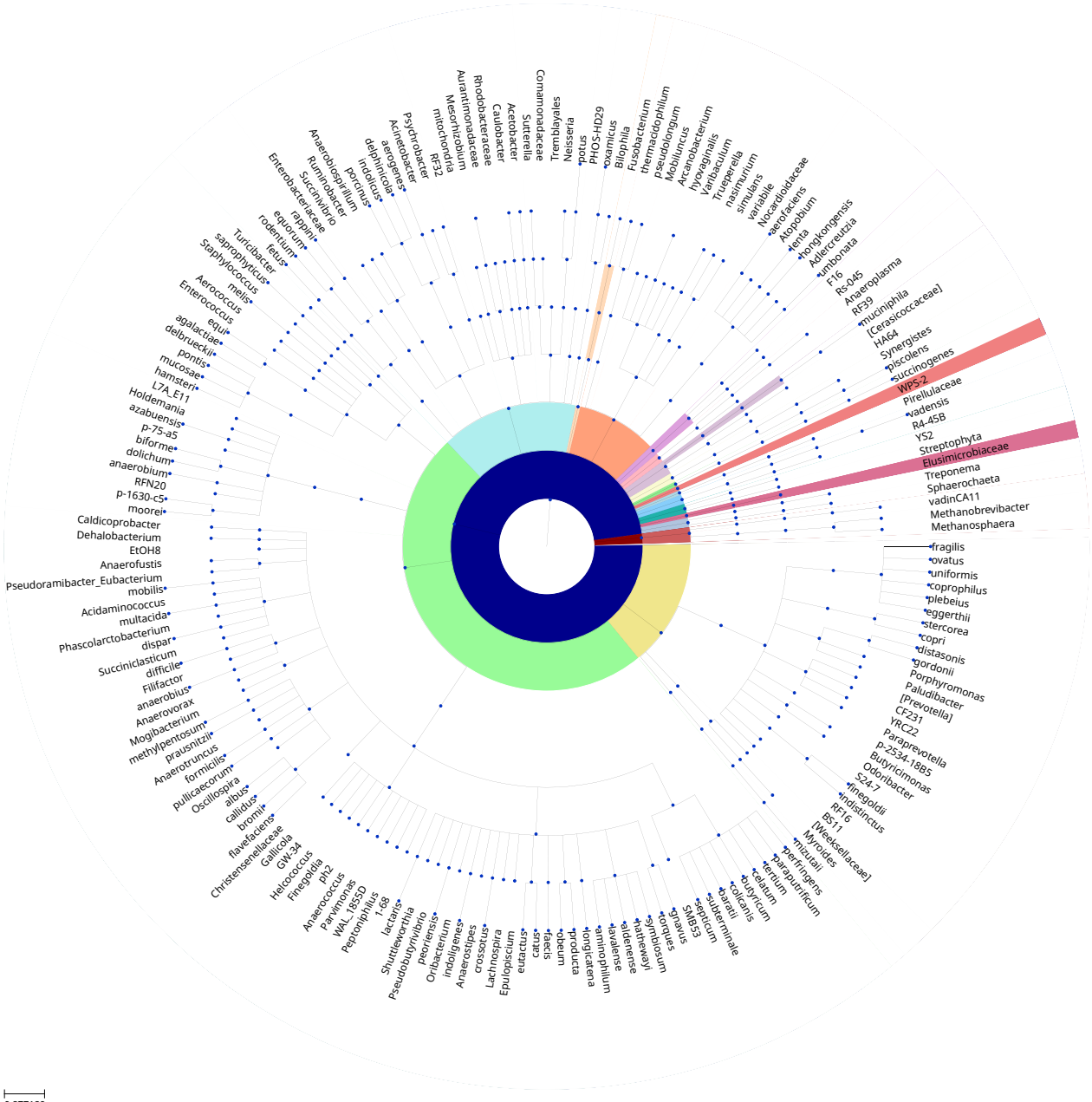
Note: It is essential to ensure that the input data (i.e., the **feature_sequences.qza** file) is of high quality and accurately represents the diversity of the microbial community, as this directly impacts the reliability of the phylogenetic tree.

7.4 Plot the Phylogenetic Tree for Taxonomic Classification

The phylogenetic tree 13 for taxonomy classification, along with its labels 14, is constructed using a Python algorithm, utilizing the ete3 library. The following pseudocode outlines the main steps of this algorithm:

```
Algorithm: Build Phylogenetic Tree from Taxonomic Data
Input: CSV file with taxonomic data
Output: Visual phylogenetic tree of the taxonomy

1: Load CSV Data
   Read CSV file containing taxonomic lineages
   Each row represents a taxonomic lineage

2: Initialize Tree
   Create an empty tree structure
   Set root of the tree

3: Process Lineages
   for each lineage in CSV Data:
       Split lineage into taxonomic levels
       for each taxonomic level in lineage:
           Remove rank prefixes (e.g., 'k__', 'p__')
           if level not in current tree path:
```

```
        Add new node for the level
    Move to the node corresponding to the current level

6: Render Tree
   Save the tree as a graphical representation
```



0.877189

Figure 13: This Taxonomic Chart visually represents the biological classification system using a phylogenetic tree. Each branch of the tree corresponds to a taxonomic level, labeled KPCOFGS, which stands for Kingdom, Phylum, Class, Order, Family, Genus, and Species. The tree illustrates the evolutionary relationships among the taxa.

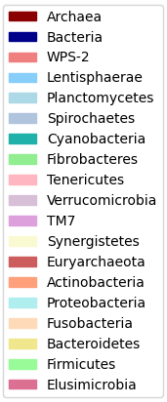


Figure 14: Legend for the phylogenetic tree

8 Imputation

In bioinformatics, imputation refers to the process of predicting and filling in missing data in a dataset. This is particularly important in bioinformatics pipelines, where datasets often have gaps due to various reasons such as experimental limitations, technical errors, or insufficient coverage. Imputation techniques use available data to estimate these missing values accurately[9].

Those are the key feature of the imputation methods:

- **Enhancing Data Completeness**: Filling in missing values to create a complete dataset, which is crucial for many downstream analyses.
- **Improving Statistical Power**: By addressing missing data, imputation helps in retaining the integrity of statistical analyses, ensuring more reliable and robust results.
- **Enabling Comprehensive Analysis**: Imputed datasets allow for the application of various computational methods that require complete data, thus facilitating a more thorough exploration of the biological questions at hand.

8.1 Data Pre-processing

Normalization of taxon counts across samples is a critical step to ensure comparable scales. In cases where normalization criteria are not met, mbImpute performs a default normalization by library size and applies a logarithmic transformation to these normalized counts, thus mitigating the impact of large count disparities.

Note: There is a bug in this parameter in R packages the **normalization** don't work correctly, it is better to not include the parameter mbImpute will arrange itself alone.

8.2 Identification of Taxon Abundances for Imputation

mbImpute employs a mixture model to approximate taxon abundances, with the model consisting of two components:

- A Gamma distribution representing non-biological zeros and low abundances.
- A normal distribution accounting for the actual abundances of the taxon.

The model is formulated as follows:

$$d_{ij} = \frac{\hat{p}_j \cdot f(Y_{ij}; \hat{\alpha}_j, \hat{\beta}_j)}{\hat{p}_j \cdot f(Y_{ij}; \hat{\alpha}_j, \hat{\beta}_j) + (1 - \hat{p}_j) \cdot f_N(Y_{ij}; X_i^T \cdot \hat{\gamma}_j, \hat{\sigma}_j^2)} \tag{8}$$

where:

- d_{ij} is the decision metric for imputation necessity

- f and f_N are the probability density functions for the Gamma and normal distributions, respectively
- $\hat{p}_j, \hat{\alpha}_j, \hat{\beta}_j, \hat{\gamma}_j, \hat{\sigma}_j^2$ are the estimated model parameters.

8.3 Imputation Process

The imputation process involves utilizing the mixture distribution to assess the likelihood of an observed abundance being a missing value. This is influenced by the linear function of sample covariates that models the normal mean parameter for each taxon.

8.4 Leveraging using Metadata and Phylogenetic Tree

The mbImpute method's effectiveness is significantly augmented by incorporating a metadata file and focusing on specific study conditions, such as the **diarrhea** column in our dataset. Additionally, a phylogenetic tree serves as a foundational element for creating a distance matrix, further refining our analytical approach.

Initially, we import the metadata file provided and determine our study condition. To facilitate this, we divide the metadata into two distinct tables: one containing the study condition **diarrhea** and the other excluding it. This separation is vital for focused analysis on our condition of interest.

Subsequently, we select a phylogenetic tree, either rooted or unrooted, previously constructed. The choice depends on the specific requirements of our study in our case we have proceed by using the **rooted-tree**.

8.5 Calculation of Distance Matrix

Upon selecting the appropriate tree, we proceed to calculate the distance matrix. The distance matrix, D, from a phylogenetic tree is computed using the cophenetic function, where D_{ij} represents the distance between taxa i and j in the tree:

$$D_{ij} = \text{Cophenetic Distance between nodes } i \text{ and } j$$

With the distance matrix calculated, the next step involves preparing it for the mbImpute function. However, prior to this, it's crucial to align the features of the distance matrix with the feature table. Given the feature table's quality trimming, some features might be absent. Therefore, we first identify features not present in the feature table and eliminate them from the distance matrix. This process results in a reshaped distance matrix, ensuring that only high-quality, relevant features are included.

Consider the scenario where the original shape of the distance matrix is 3011×3011 and the feature table has dimensions 120×3011. In this case, the distance matrix needs to be reshaped to 3011×3011 to align with the feature table. It is crucial to ensure that features trimmed from the distance matrix correspond accurately to those in the feature table. This alignment is essential because the **taxonomy.tsv** file, which often guides the feature selection, might contain a different number of features. Although typically, the **taxonomy.tsv** is derived from the **seq-aligned.qza** file and should have a matching number of features, discrepancies can arise due to data manipulation and rearrangement in preprocessing stages. Therefore, verifying the consistency between the distance matrix and the feature table is a critical step before proceeding with further analysis

This meticulous process of leveraging metadata, particularly the study condition, and the strategic use of a phylogenetic tree as a distance matrix, significantly enhances the accuracy and relevance of the mbImpute function in our analysis.

In this following study, we explored three different imputation methods using mbImpute: non-normalized, logarithmically normalized (Lng-Normalized), and normalized. Each method offers distinct insights into the frequency distribution of samples and features within our microbiome data.

8.6 Imputed Feature Table

In the imputation process without normalization, we address the zeros in the data without attempting to accurately scale or refine the data measurements.

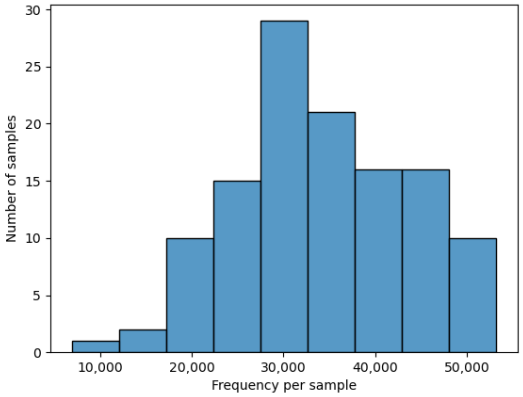


Figure 15: **Sample Frequency Imputed:** This graph provides a detailed representation of the frequency distribution within the samples. The x-axis displays the frequency count, indicating the number of occurrences or counts recorded in the dataset. On the y-axis, we have the number of samples that correspond to each frequency count.

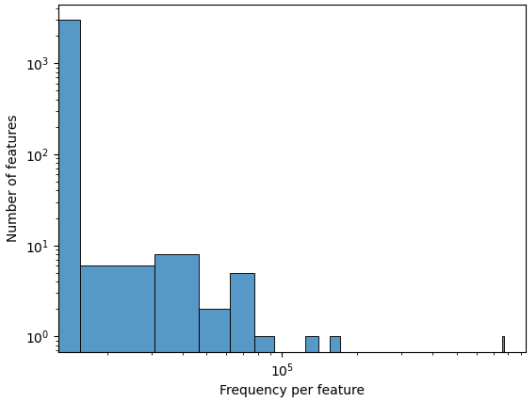


Figure 16: **Feature Frequency Imputed:** The graph presents an insightful view into the frequency distribution across features. The x-axis illustrates the frequency per feature, showing the count or number of occurrences for each feature within the dataset. Correspondingly, the y-axis indicates the number of features at each frequency level.

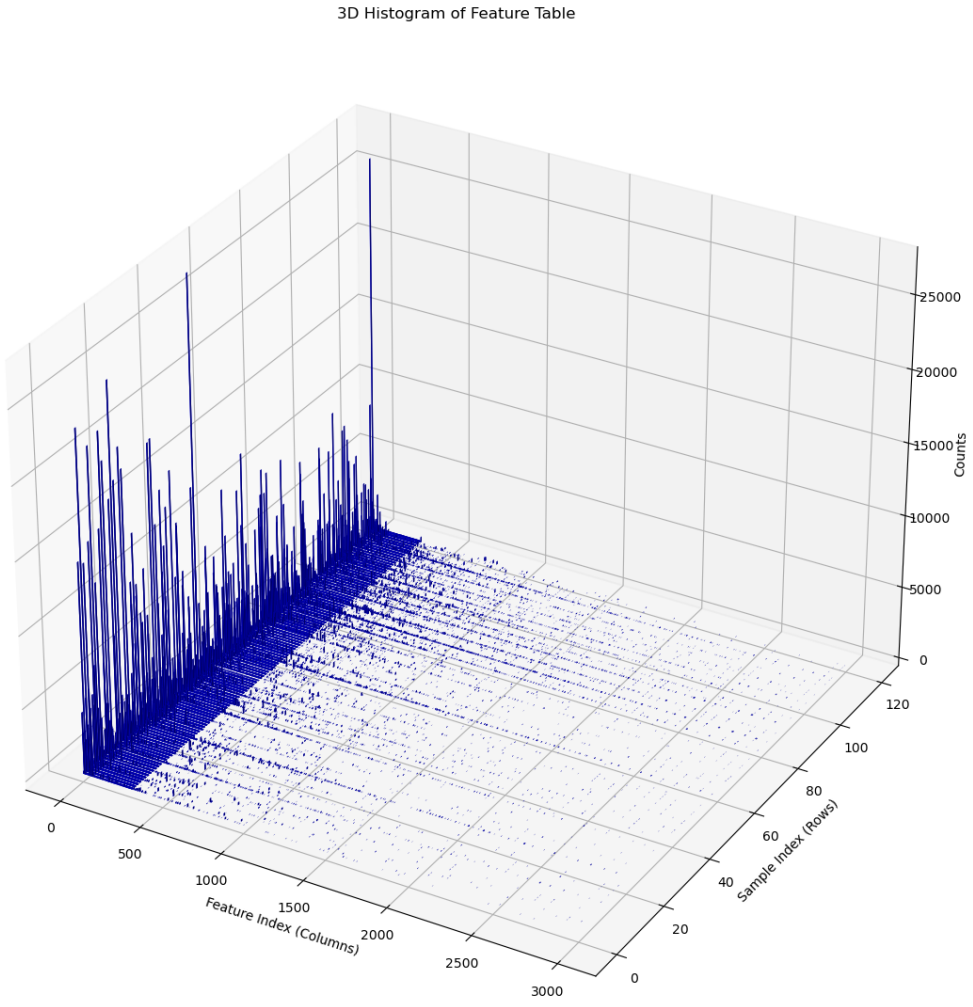


Figure 17: **3D Histogram of the Imputed Feature Table:** This figure presents a comprehensive 3D histogram. On the x-axis, we display the feature index, representing the total number of features in our count table. The z-axis corresponds to the number of samples. Finally, the y-axis quantifies the count of features for each bin, providing a detailed view of the distribution across the dataset.

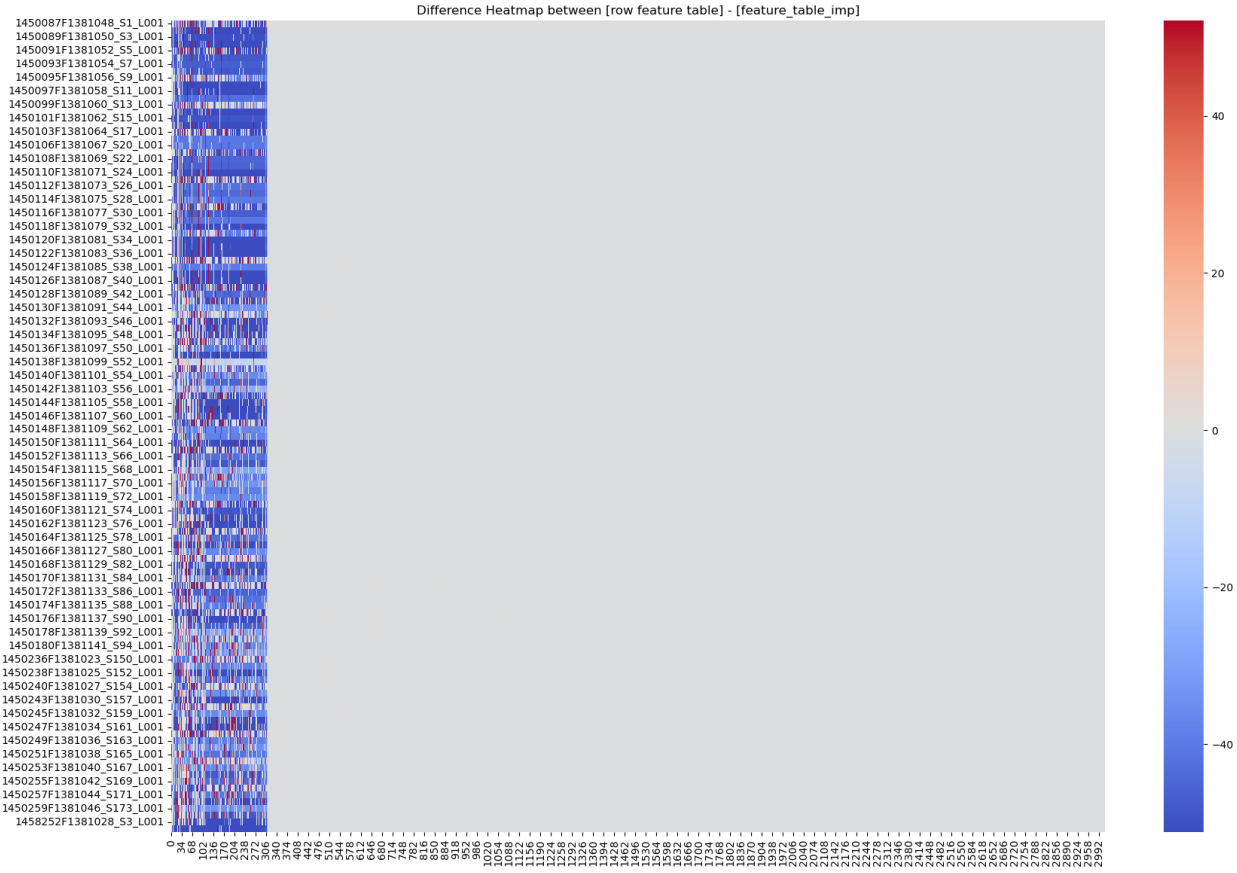


Figure 18: This heatmap displays a comparison between **the original feature table minus the imputed feature table** . On the right side of the heatmap, a color bar is provided to illustrate the differences, showing the original non-imputed data subtracted from the new imputed data. The features are aligned along the x-axis, while the samples are organized on the y-axis.

Table 9: Overview of Table Summary for **Imputed Feature Table**

Metric	**Sample**
Number of Samples	120
Number of Features	3,011
Total Frequency	4,084,753

Table 10: Detailed Frequency Metrics for **Imputed Feature Table**

	Frequency Per Sample	**Frequency Per Feature**
Minimum Frequency	6,902.0	0.0
1st Quartile	28,305.5	12.0
Median Frequency	33,046.0	46.0
3rd Quartile	40,949.25	237.0
Maximum Frequency	53,245.0	775,662.0
Mean Frequency	34,039.60	1,356.61

8.7 Imputed Feature Table Log Transform

The Lng-Normalized imputation adjusts for these variations by applying a logarithmic scale, which helps balance the representation of abundant and less abundant features. This approach reduces the impact of highly abundant features, allowing for a more nuanced analysis of the data.

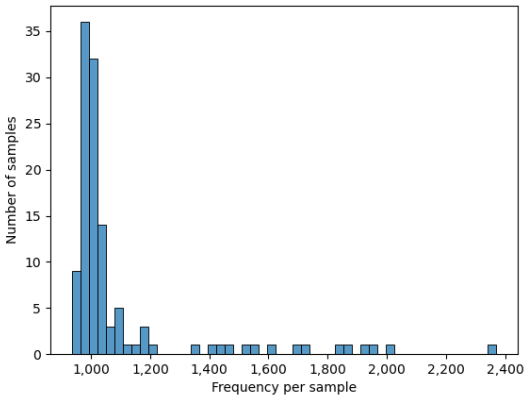


Figure 19: **Sample Frequency Imputed (LNG):** This graph provides a detailed representation of the frequency distribution within the samples. The x-axis displays the frequency count, indicating the number of occurrences or counts recorded in the dataset. On the y-axis, we have the number of samples that correspond to each frequency count.

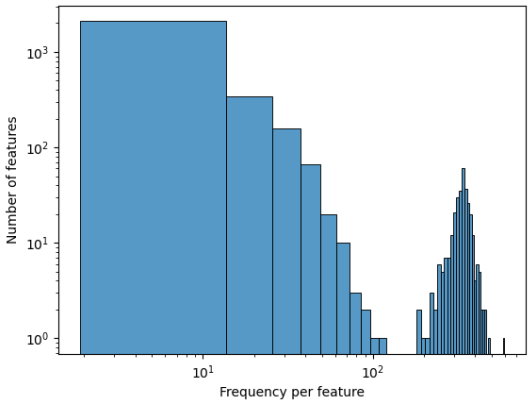


Figure 20: **Feature Frequency Imputed (LNG):** The graph presents an insightful view into the frequency distribution across features. The x-axis illustrates the frequency per feature, showing the count or number of occurrences for each feature within the dataset. Correspondingly, the y-axis indicates the number of features at each frequency level.

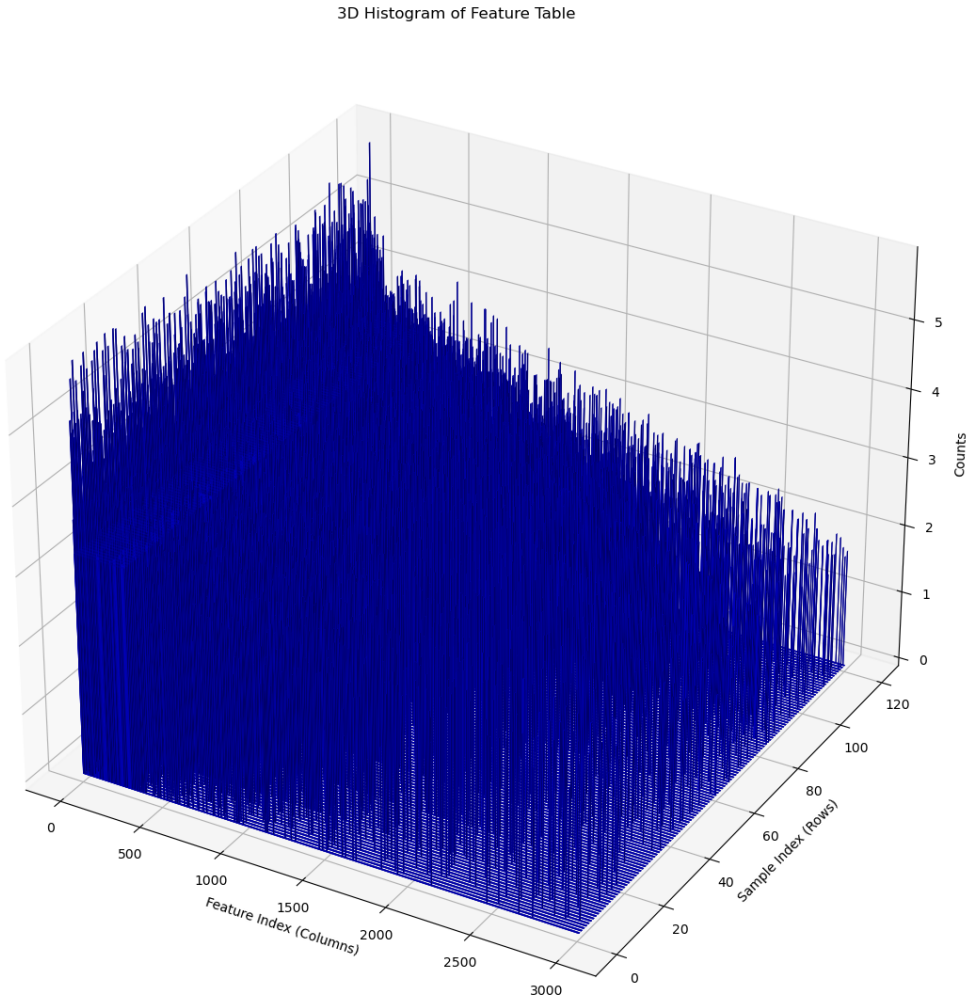


Figure 21: **3D Histogram of the Imputed Feature Table (LNG):** This figure presents a comprehensive 3D histogram. On the x-axis, we display the feature index, representing the total number of features in our count table. The z-axis corresponds to the number of samples. Finally, the y-axis quantifies the count of features for each bin, providing a detailed view of the distribution across the dataset.

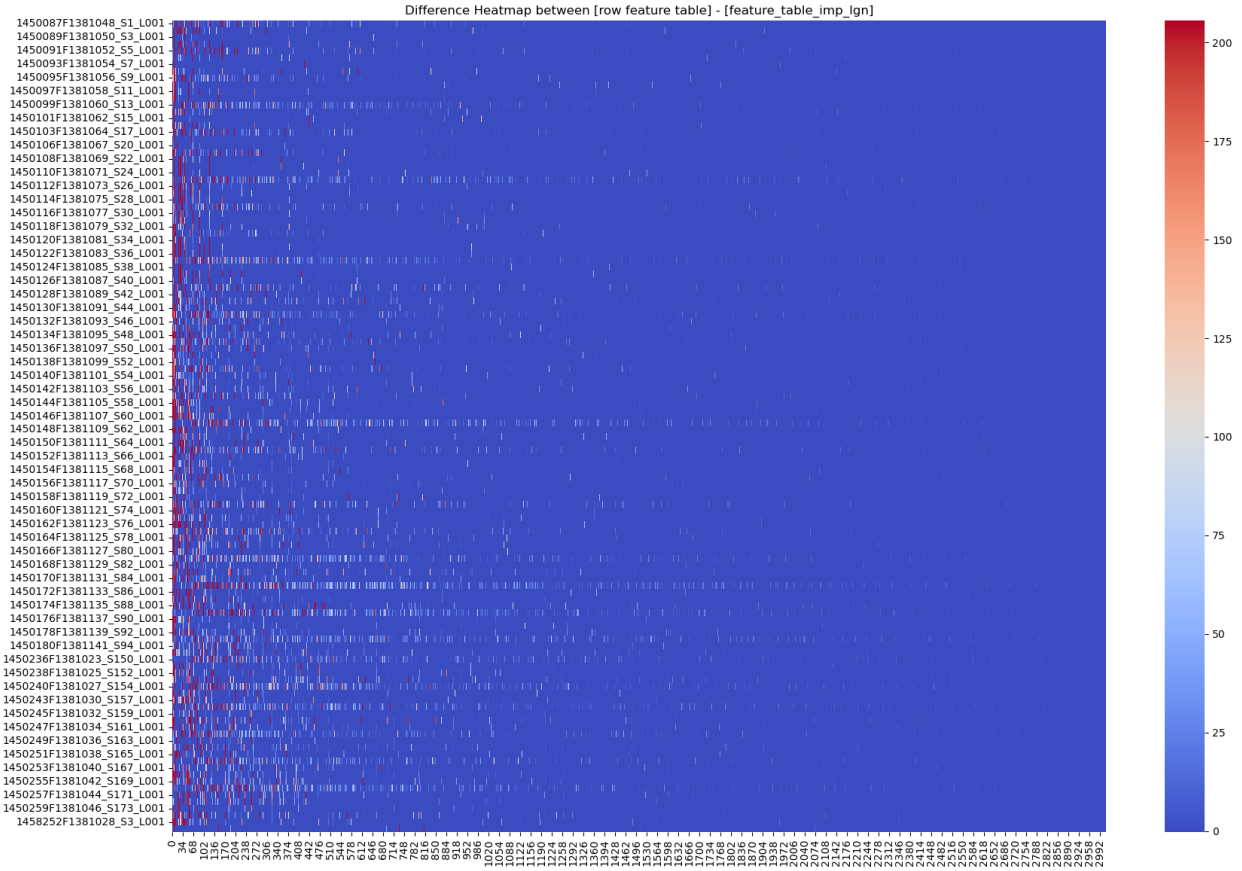


Figure 22: This heatmap displays a comparison between **the original feature table minus the imputed feature table (LNG)**. On the right side of the heatmap, a color bar is provided to illustrate the differences, showing the original non-imputed data subtracted from the new imputed data. The features are aligned along the x-axis, while the samples are organized on the y-axis.

Table 11: Overview of Table Summary for **Imputed Feature Table (LNG)**

Metric	Value
Number of Samples	120
Number of Features	3,011
Total Frequency	131,731

Table 12: Detailed Frequency Metrics for **Imputed Feature Table (LNG)**

	Frequency Per Sample	Frequency Per Feature
Minimum Frequency	934.45	1.88
1st Quartile	980.94	3.23
Median Frequency	1,001.92	6.13
3rd Quartile	1,045.57	18.45
Maximum Frequency	2,370.80	593.90
Mean Frequency	1,097.77	43.75

8.8 Imputed Feature Table Normalized

Finally, the normalized imputation presents a comprehensive view by adjusting all sample and feature frequencies to a common scale. This normalization aids in identifying patterns and trends that might be obscured in the raw data due to disparate abundance scales.

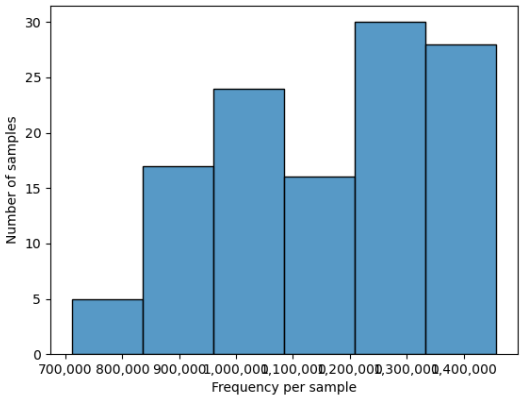


Figure 23: **Sample Frequency Imputed (NRM):** This graph provides a detailed representation of the frequency distribution within the samples. The x-axis displays the frequency count, indicating the number of occurrences or counts recorded in the dataset. On the y-axis, we have the number of samples that correspond to each frequency count.

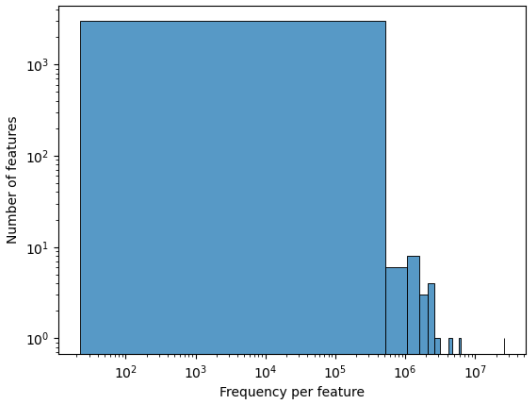


Figure 24: **Feature Frequency Imputed (NRM):** The graph presents an insightful view into the frequency distribution across features. The x-axis illustrates the frequency per feature, showing the count or number of occurrences for each feature within the dataset. Correspondingly, the y-axis indicates the number of features at each frequency level.

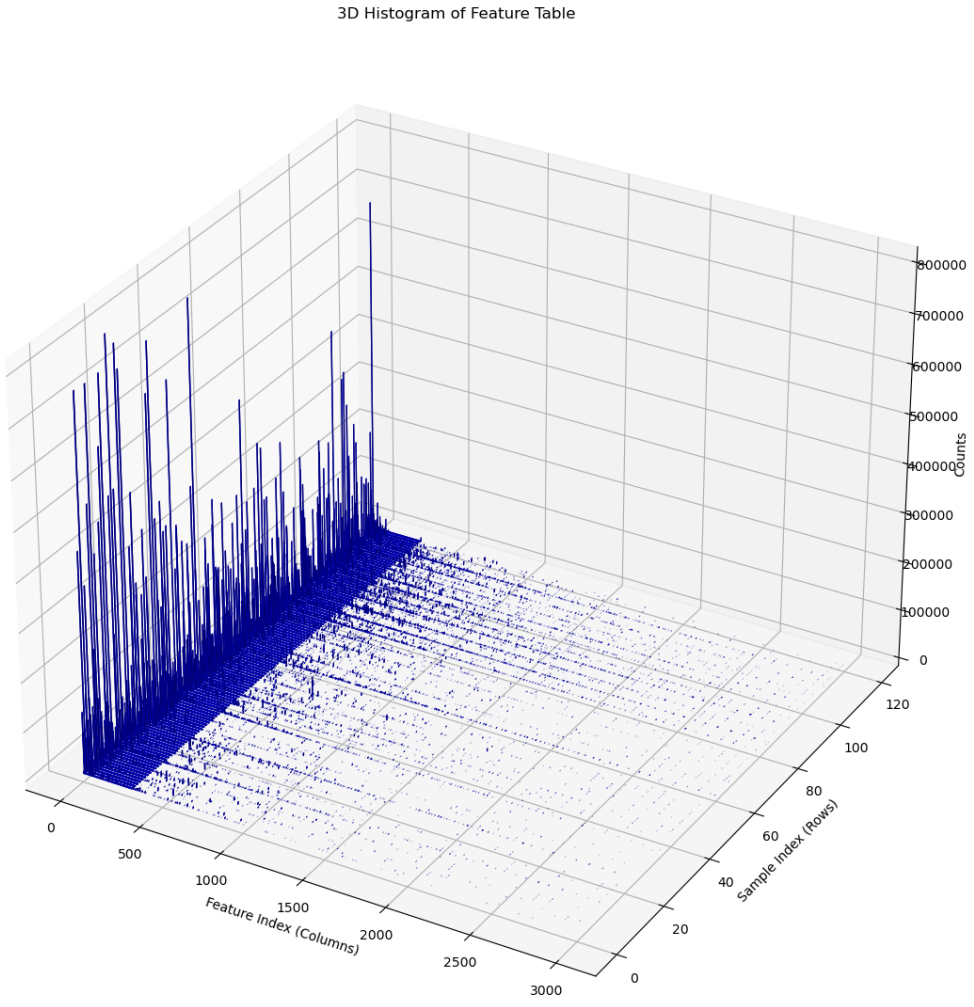


Figure 25: **3D Histogram of the Imputed Feature Table (NRM):** This figure presents a comprehensive 3D histogram. On the x-axis, we display the feature index, representing the total number of features in our count table. The z-axis corresponds to the number of samples. Finally, the y-axis quantifies the count of features for each bin, providing a detailed view of the distribution across the dataset.

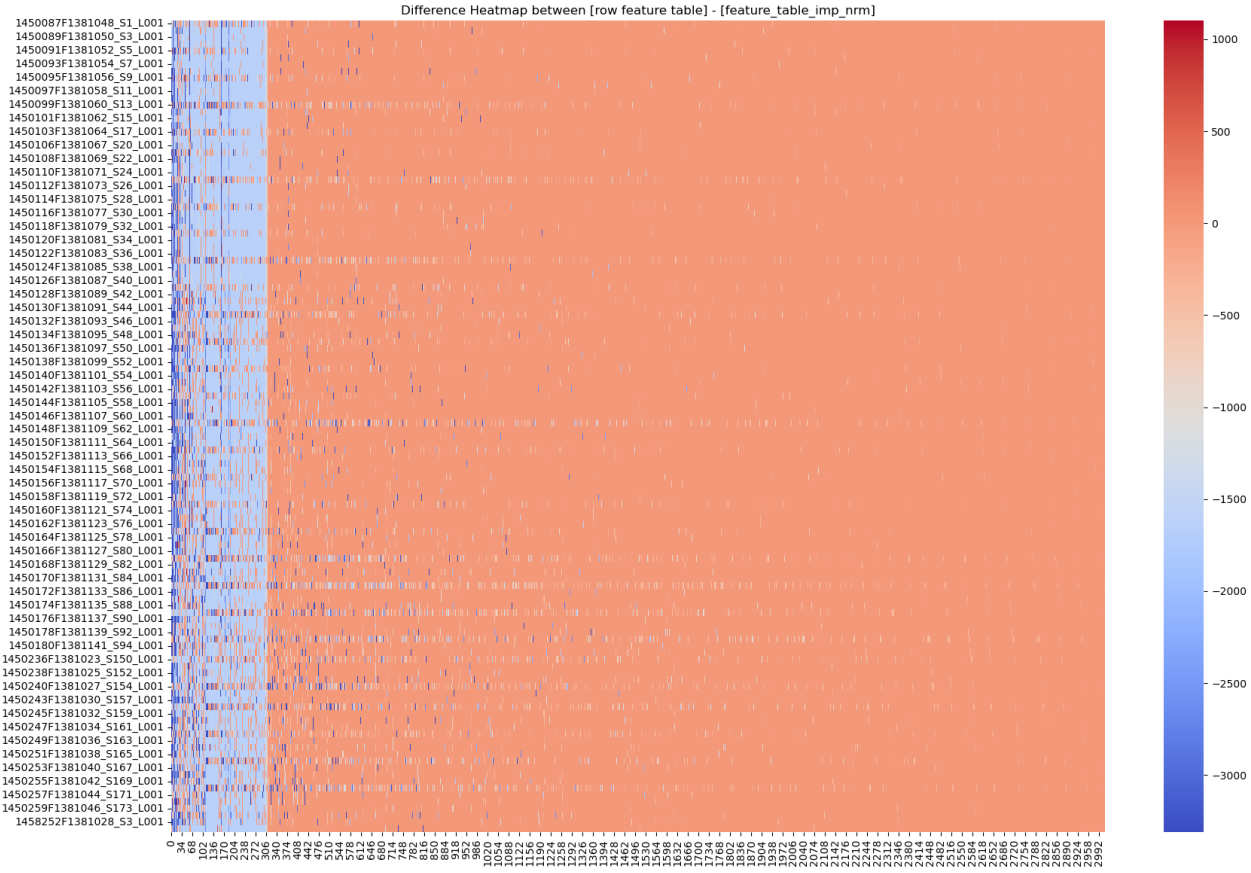


Figure 26: This heatmap displays a comparison between **the original feature table minus the imputed feature table (NRM)**. On the right side of the heatmap, a color bar is provided to illustrate the differences, showing the original non-imputed data subtracted from the new imputed data. The features are aligned along the x-axis, while the samples are organized on the y-axis.

Table 13: Overview of Table Summary for **Imputed Feature Table (NRM)**

Metric	Value
Number of Samples	120
Number of Features	3,011
Total Frequency	139,454,110

Table 14: Detailed Frequency Metrics for **Imputed Feature Table (NRM)**

	Frequency Per Sample	Frequency Per Feature
Minimum Frequency	710,930.0	22.0
1st Quartile	998,982.75	443.0
Median Frequency	1,192,127.0	1,619.0
3rd Quartile	1,323,122.25	8,196.0
Maximum Frequency	1,457,812.0	26,366,963.0
Mean Frequency	1,162,117.58	46,314.88

8.9 Conclusion on Data Imputation

In this section, we synthesize our findings from examining three distinct datasets: the imputed dataset, the normalized dataset, and the log-scaled dataset. Our goal was to determine the most suitable dataset for further analysis.

Initially, the **normalized** dataset seemed promising as it presumably equalizes the data distribution. However, a significant issue emerged with this approach. The feature frequency in the normalized dataset exhibited an exponential increase, soaring from approximately $4,000$ to an overwhelming $140,000$.
Note: This create memory issue due to low resources.

Regarding the **log-scaled** dataset, its inherent flaw was evident from the outset. The dataset is centered around zero, which poses a major limitation for its application in software like Qiime2, as this tool requires data to be positively skewed. An adjustment to shift the entire dataset to the positive x-axis was considered, but this approach would introduce a significant bias, making it an unsuitable solution.

Consequently, after thorough consideration, we opted to use the **imputed** dataset for all subsequent analyses. This dataset maintains a balance by modestly increasing the feature frequency, unlike the exponential surge observed in the normalized set, thereby preserving the integrity of the original data without the zero-inflation issue.

9 Normalization

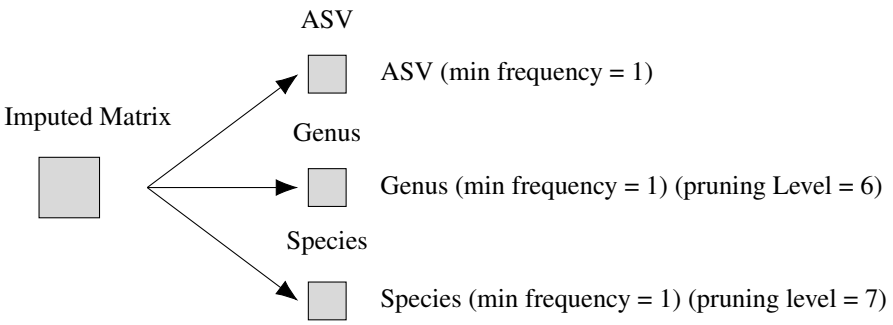
Having reached this stage in our research, we are equipped with **imputed data** and a clear taxonomic, which together pave the way for a comprehensive understanding of our analytical objectives. The subsequent phase of our analysis involves a series of methodical steps to normalize our data.

9.1 Construction of Taxonomic Tables (Pruning)

Our primary task is to construct three distinct tables: the ASV (Amplicon Sequence Variant) table, the Genus table, and the Species table.

- **ASV Table:** This table represents a raw dataset selected from three different types of imputation methods. It is further refined by applying a threshold based on minimum frequency levels in my case I have decided to use **minimum frequency** $= 1$, this will exclude all feature with less then one frequency present.
- **Genus and Species Tables:** The construction of these tables is more nuanced. As delineated in the preceding chapter, the biological classification system, denoted as KPCOFGS (Kingdom, Phylum, Class, Order, Family, Genus, Species), serves as a foundational guide. For the Genus and Species tables, we focus on their respective taxonomic levels, incorporating a 'pruning factor' in our filtering process.
  Specifically, a pruning factor of 6 is applied for the Genus table, positioning it at the sixth level and thereby excluding all species-level data. Conversely, for the Species table, the pruning factor is set to 7. Further more here a filtering is applied as ASV table **minimum frequency** $= 1$, for genus and species table.

This structured approach allows for a precise extraction of taxonomic data at different levels, facilitating a more targeted and effective analysis. The distinction and separation of data into the ASV, Genus, and Species tables are crucial for the subsequent phases of our study, where each table will offer unique insights pertinent to our research objectives. Also a filtering of feature is performed with value 1 the intent is to not exclude any kind of relevant feature in this step.



9.1.1 ASV Feature Table Taxa Filtered

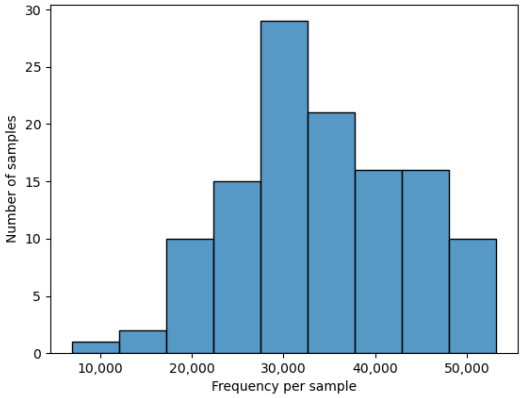


Figure 27: **Sample Frequency ASV Taxa Filtered:** This graph provides a detailed representation of the frequency distribution within the samples. The x-axis displays the frequency count, indicating the number of occurrences or counts recorded in the dataset. On the y-axis, we have the number of samples that correspond to each frequency count.

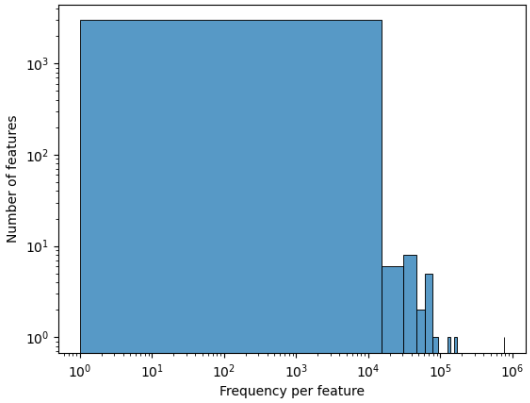


Figure 28: **Feature Frequency Imputed ASV Taxa Filtered:** The graph presents an insightful view into the frequency distribution across features. The x-axis illustrates the frequency per feature, showing the count or number of occurrences for each feature within the dataset. Correspondingly, the y-axis indicates the number of features at each frequency level.

Table 15: Overview of Table Summary for **ASV Taxa Filtered**

Metric	Value
Number of Samples	120
Number of Features	3,001
Total Frequency	4,084,753

Table 16: Detailed Frequency Metrics for **ASV Taxa Filtered**

Metric	Frequency Per Sample	Frequency Per Feature
Minimum Frequency	6,902.0	1.0
1st Quartile	28,305.5	12.0
Median Frequency	33,046.0	46.0
3rd Quartile	40,949.25	238.0
Maximum Frequency	53,245.0	775,662.0
Mean Frequency	34,039.61	1,361.13

9.1.2 Genus Feature Table Taxa Filtered

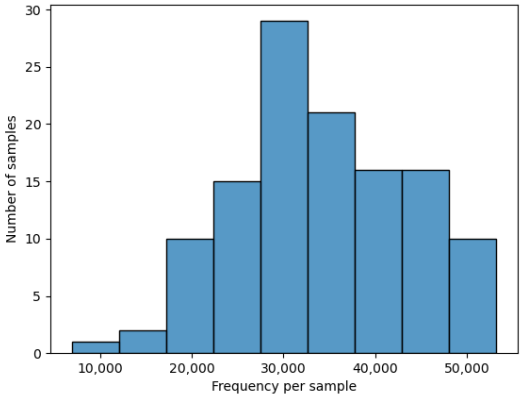


Figure 29: **Sample Frequency Genus Taxa Filtered:** This graph provides a detailed representation of the frequency distribution within the samples. The x-axis displays the frequency count, indicating the number of occurrences or counts recorded in the dataset. On the y-axis, we have the number of samples that correspond to each frequency count.

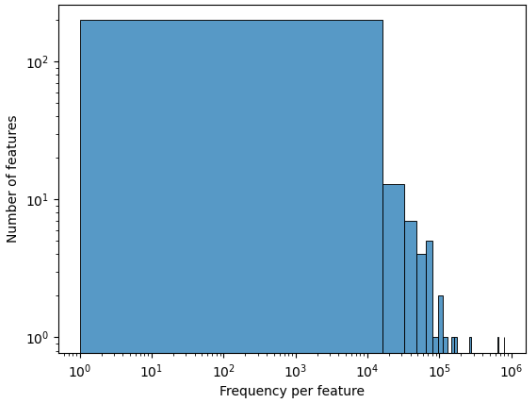


Figure 30: **Feature Frequency Imputed Genus Taxa Filtered:** The graph presents an insightful view into the frequency distribution across features. The x-axis illustrates the frequency per feature, showing the count or number of occurrences for each feature within the dataset. Correspondingly, the y-axis indicates the number of features at each frequency level.

Table 17: Overview of Table Summary for **Genus Taxa Filtered**

Metric	Value
Number of Samples	120
Number of Features	237
Total Frequency	4,084,753

Table 18: Detailed Frequency Metrics for **Genus Taxa Filtered**

Metric	Frequency Per Sample	Frequency Per Feature
Minimum Frequency	6,902.0	1.0
1st Quartile	28,305.5	50.0
Median Frequency	33,046.0	576.0
3rd Quartile	40,949.25	7,165.0
Maximum Frequency	53,245.0	808,241.0
Mean Frequency	34,039.61	17,235.24

9.1.3 Species Feature Table Taxa Filtered

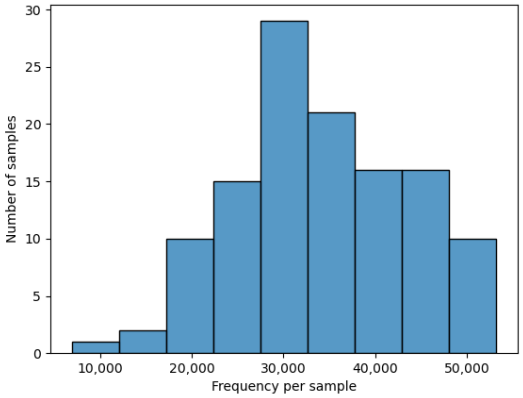


Figure 31: **Sample Frequency Species Taxa Filtered:** This graph provides a detailed representation of the frequency distribution within the samples. The x-axis displays the frequency count, indicating the number of occurrences or counts recorded in the dataset. On the y-axis, we have the number of samples that correspond to each frequency count.

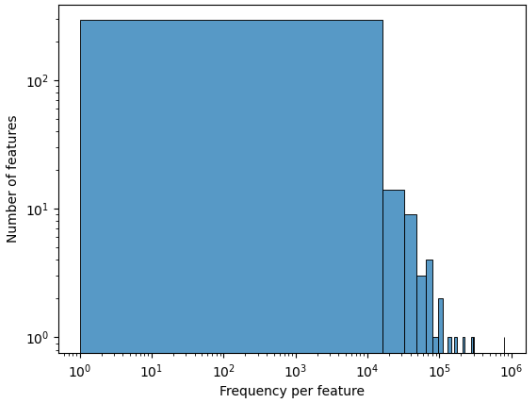


Figure 32: **Feature Frequency Imputed Species Taxa Filtered:** The graph presents an insightful view into the frequency distribution across features. The x-axis illustrates the frequency per feature, showing the count or number of occurrences for each feature within the dataset. Correspondingly, the y-axis indicates the number of features at each frequency level.

Table 19: Overview of Table Summary for **Species Taxa Filtered**

Metric	Value
Number of Samples	120
Number of Features	332
Total Frequency	4,084,753

Table 20: Detailed Frequency Metrics for **Species Taxa Filtered**

Metric	Frequency Per Sample	Frequency Per Feature
Minimum Frequency	6,902.0	1.0
1st Quartile	28,305.5	50.0
Median Frequency	33,046.0	438.5
3rd Quartile	40,949.25	6,493.0
Maximum Frequency	53,245.0	808,241.0
Mean Frequency	34,039.61	12,303.47

9.1.4 Conclusion on Pruning and Taxa Filtering

In this conclusio, we want to expose the difference between the **Imputed Feature Table** against three others derived from various levels of pruning and taxa filtering: ASV, Genus, and Species. Each table yields unique perspectives on the microbial community's structure and diversity in our samples, showcasing the outcomes of distinct data processing methods.

Consistency Across Samples: A notable consistency is observed in the number of samples (120) and total frequency (4,084,753) across all tables. This uniformity is vital for comparative purposes, as it ensures that the changes in data are due to the processing methods rather than alterations in sample size or overall count.

Variation in Number of Features: A significant disparity exists in the number of features among these tables. The

Imputed Feature Table is the most comprehensive, containing 3,011 features. The ASV Taxa Filtered table shows a slight decrease to 3,001 features. Notably, the Genus and Species tables exhibit substantial reductions, with only 237 and 332 features, respectively. This suggests a more aggregated data approach at the Genus and Species levels, potentially leading to a loss of finer details.

Frequency Metrics: The Minimum Frequency per feature in the Imputed table is 0.0, implying the presence of features not found in all samples. In contrast, the taxa-filtered tables (ASV, Genus, Species) display a minimum frequency of 1.0, indicating the exclusion of such features. The Mean Frequency per feature shows a varied pattern, with the Genus Taxa Filtered table at the highest (17,235.24) and the Imputed table at the lowest (1,356.61). This variation suggests a denser concentration of counts in fewer features in the Genus and Species tables, as opposed to the more distributed pattern in the Imputed and ASV tables. Furthermore, the Maximum Frequency per feature in both the Genus and Species tables is significantly higher (808,241.0) compared to that in the Imputed (775,662.0) and ASV (775,662.0) tables, highlighting a similar trend of count concentration.

9.2 Normalization of ASV, Genus, Species Feature Table

The processing script is designed to automatically explore various normalization possibilities, allowing us to select the most suitable method for our dataset. As illustrated in the following section, two prominent normalization techniques can be applied to each of the matrices created (ASV, Genus, Species). The normalization method are:
Geometric Mean of Pairwise Ratios (GMPR) and Centered Log-Ratio (CLR) normalization. In this section we are going to see this method and the result that they achieve.

9.2.1 GMPR Normalization

The GMPR normalization, proposed by Chen et al. [5], is a method designed to adjust for varying sequencing depths in compositional data, such as those encountered in microbiome studies. It normalizes the data using the geometric mean of pairwise ratios (GMPR) between samples. The GMPR normalization for the i-th feature in a sample is mathematically defined as:

$$\text{GMPR}(x_i) = \frac{x_i}{g_m}$$

Where:
- x_i is the count of the i-th feature in the sample
- g_m is the geometric mean9.2.3 of the counts of all features in that sample

This technique is particularly effective in handling sparse and compositional data, providing a way to mitigate the impact of varying sequencing depths. It differs from the CLR (Centered Log-Ratio) transformation, which normalizes the data row by row, focusing on the relative abundance of features within each sample.

Example: Consider a sample with feature counts $[10, 30, 60]$. The geometric mean g_m of these counts is calculated as:

$$((10 \times 30 \times 60)^{\frac{1}{3}} \approx 26$$

The GMPR normalized value for the first feature (with count 10) is then calculated as

$$\frac{10}{26}$$

This example illustrates how GMPR normalization adjusts each feature count relative to the overall composition of the sample, thus allowing for more meaningful comparisons across samples with different sequencing depths.

9.2.2 CLR Normalization

The Centered Log-Ratio (CLR) normalization, introduced by Aitchison [2], is a foundational technique in compositional data analysis. The CLR transformation is applied to each component of a compositional vector, normalizing the data to account for the relative proportions of each part. The mathematical formula for CLR normalization of the j-th component in the i-th sample is given by:

$$\text{CLR}(x_i) = \ln\left(\frac{x_i}{g_m}\right)$$

Where:
- x_i is the count of the i-th feature
- g_m is the geometric mean 9.2.3 of the counts of all features in that sample

This normalization approach transforms the compositional data into log-ratio coordinates, enabling the application of standard statistical techniques. CLR is particularly useful for analyzing data where the relative abundance of components (such as microbial species in a microbiome study) is more important than their absolute counts.

Example: Consider a sample with feature counts $[2, 8, 32]$. The geometric mean of these counts is:

$$((2 \times 8 \times 32)^{\frac{1}{3}} \approx 8$$

The CLR normalized value for the first feature (with count 2) is calculated as:

$$\ln\left(\frac{2}{8}\right) = \ln(0.25) \approx -1.39$$

This example demonstrates how CLR normalization adjusts each feature count relative to the geometric mean of the sample, facilitating comparisons of relative abundances across different samples.

9.2.3 Geometric Mean

The geometric mean of a set of positive numbers is calculated by multiplying all the numbers together and then taking the n-th root of the product, where n is the total number of values. The formula for the geometric mean of a set of numbers $x_1, x_2, ..., x_n$ is:

$$\text{Geometric Mean} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}$$

Where:
- $\prod$ denotes the product of the series
- x_i is each individual number in the set
- n is the total number of values in the set

For example, the geometric mean of three numbers 2, 8, and 32 is calculated as:

$$\text{Geometric Mean} = (2 \times 8 \times 32)^{\frac{1}{3}} = (512)^{\frac{1}{3}} \approx 8$$

This formula is particularly useful in situations where the numbers are of different orders of magnitude or when they represent rates of change (like growth rates).

9.3 ASV Feature Table Imputed GMPR Normalized

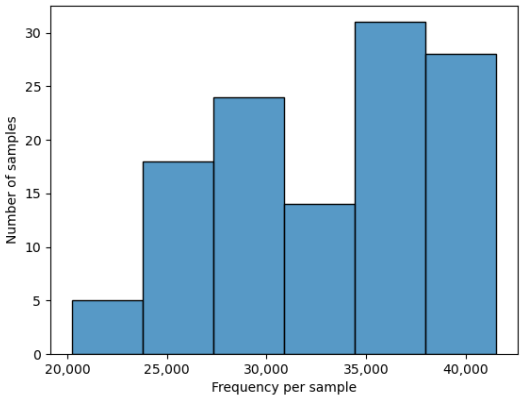


Figure 33: **Sample Frequency ASV GMPR Normalized:** The graph presents an insightful view into the frequency distribution across features. The x-axis illustrates the frequency per feature, showing the count or number of occurrences for each feature within the dataset. Correspondingly, the y-axis indicates the number of features at each frequency level.

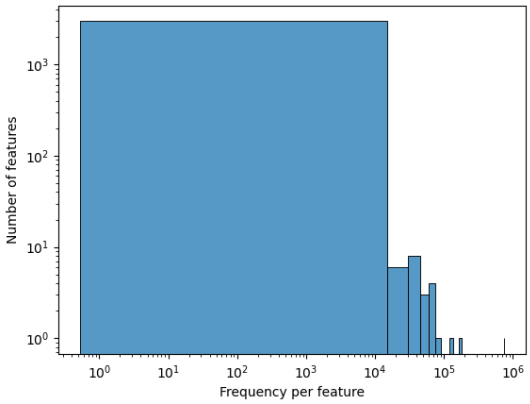


Figure 34: **Feature Frequency ASV GMPR Normalized:** The graph presents an insightful view into the frequency distribution across features. The x-axis illustrates the frequency per feature, showing the count or number of occurrences for each feature within the dataset. Correspondingly, the y-axis indicates the number of features at each frequency level.

Table 21: Overview of Table Summary for **ASV GPMR Normalized**

Metric	Value
Number of Samples	120
Number of Features	3,001
Total Frequency	3,963,204

Table 22: Detailed Frequency Metrics for **ASV GPMR Normalized**

Metric	Frequency Per Sample	Frequency Per Feature
Minimum Frequency	20,206.28	0.5192
1st Quartile	28,129.80	11.4611
Median Frequency	33,896.12	44.4467
3rd Quartile	37,650.39	230.232
Maximum Frequency	41,553.78	754,013.7675
Mean Frequency	33,026.70	1,320.63

9.4 Genus Feature Table Imputed GMPR Normalized

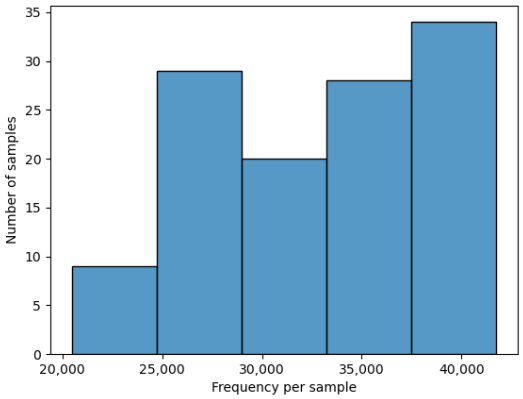


Figure 35: **Sample Frequency Genus GMPR Normalized:** The graph presents an insightful view into the frequency distribution across features. The x-axis illustrates the frequency per feature, showing the count or number of occurrences for each feature within the dataset. Correspondingly, the y-axis indicates the number of features at each frequency level.

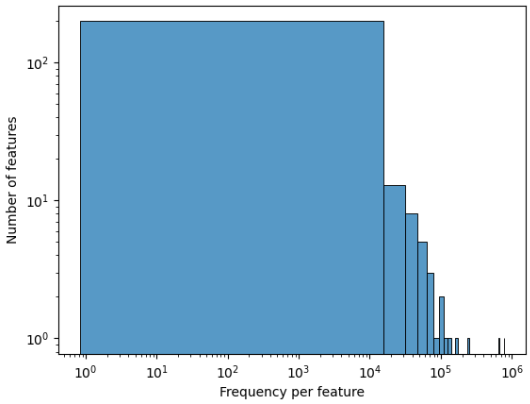


Figure 36: **Feature Frequency Genus GMPR Normalized:** The graph presents an insightful view into the frequency distribution across features. The x-axis illustrates the frequency per feature, showing the count or number of occurrences for each feature within the dataset. Correspondingly, the y-axis indicates the number of features at each frequency level.

Table 23: Overview of Table Summary for **Genus GMPR Normalized**

Metric	**Value**
Number of Samples	120
Number of Features	237
Total Frequency	3,966,781

Table 24: Detailed Frequency Metrics for **Genus GMPR Normalized**

Metric	**Frequency Per Sample**	**Frequency Per Feature**
Minimum Frequency	20,463.39	0.8221
1st Quartile	27,683.85	45.5994
Median Frequency	33,862.56	510.6687
3rd Quartile	37,924.90	7,166.15
Maximum Frequency	41,749.80	789,341.71
Mean Frequency	33,056.51	16,737.48

9.5 Species Feature Table Imputed GMPR Normalized

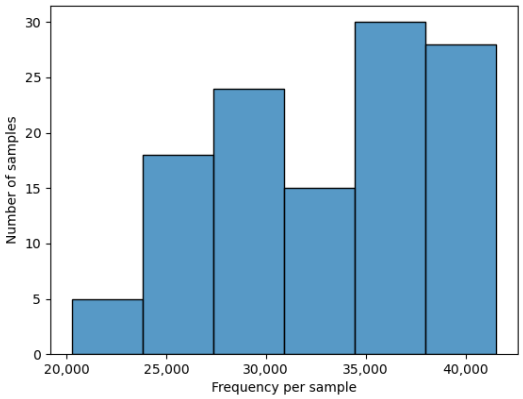


Figure 37: **Sample Frequency Species GMPR Normalized:** The graph presents an insightful view into the frequency distribution across features. The x-axis illustrates the frequency per feature, showing the count or number of occurrences for each feature within the dataset. Correspondingly, the y-axis indicates the number of features at each frequency level.

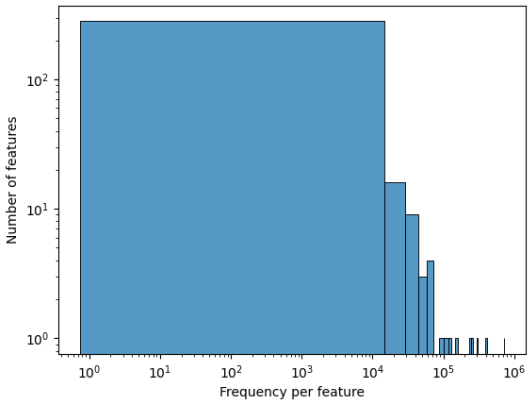


Figure 38: **Feature Frequency Species GMPR Normalized:** The graph presents an insightful view into the frequency distribution across features. The x-axis illustrates the frequency per feature, showing the count or number of occurrences for each feature within the dataset. Correspondingly, the y-axis indicates the number of features at each frequency level.

Table 25: Overview of Table Summary for **Species GMPR Normalized**

Metric	Value
Number of Samples	120
Number of Features	332
Total Frequency	3,963,139

Table 26: Detailed Frequency Metrics for **Species GMPR Normalized**

Metric	Frequency Per Sample	Frequency Per Feature
Minimum Frequency	20,261.01	0.6581
1st Quartile	28,161.97	46.6806
Median Frequency	33,721.55	422.7581
3rd Quartile	37,655.46	6,374.32
Maximum Frequency	41,553.77	785,117.77
Mean Frequency	33,026.17	11,937.17

9.6 Conclusion on the GMPR Normalization

In our comprehensive analysis of the GMPR Normalized datasets across various taxonomic levels (ASV, Genus, Species), several key insights have been observed. These findings are pivotal in understanding the impacts of GMPR normalization and the representational differences across taxonomic resolutions.

The GMPR (Geometric Mean of Pairwise Ratios) normalization, known for adjusting raw counts to address **compositional data constraints**, inherently leads to some loss of frequencies. This normalization makes the data more comparable across samples by adjusting the original frequency counts, resulting in lower total frequencies in the normalized data compared to the raw data.

A remarkable **anomaly** emerges in the Genus table, where the total frequency is unexpectedly higher (3,966,781)

compared to the ASV table (3,963,204). This unusual increase, despite the Genus table having fewer features, might be attributed to the normalization process disproportionately amplifying certain dominant features at the Genus level. While this results in an overall higher frequency count, it's important to note that such anomalies, although uncommon, are possible and do not substantially alter the overall feature count.

10 Alpha and Beta Diversity Analysis

The analysis of alpha and beta diversity is rooted in the study of the 16S rRNA gene, a component universally conserved across bacterial species. This gene's structure is crucial in phylogenetic research due to the variability found in specific regions, allowing for the precise identification and classification of bacteria at different taxonomic levels. Using data derived from 16S sequencing, two fundamental measures of microbial diversity are employed:

- **Alpha Diversity** encapsulates the diversity within a single microbial community, offering insights into the richness and evenness of species in that specific environment.
- **Beta Diversity**, conversely, focuses on the diversity between distinct communities. This measure is pivotal in understanding the differences or similarities in species composition across various samples.

In the ensuing section, we will delve into the principal methodologies employed for analyzing alpha and beta diversity, as highlighted in [7].

10.1 Alpha Diversity Bar Plot

Alpha diversity bar plots are a fundamental tool in microbiota studies, providing a visual representation of the diversity within individual microbial communities across various samples.These plots are instrumental in understanding and comparing the richness and evenness of species within each sample.

The primary purpose of an alpha diversity bar plot is to compare the microbial diversity across different samples or groups. Each bar in the plot represents a categorically defined group of samples, with the height of the bar indicating the level of alpha diversity. This visualization facilitates the identification of patterns and differences in microbial diversity under different conditions.

Interpreting an alpha diversity bar plot involves assessing the height of the bars to gauge the diversity within each sample or group. Higher bars indicate a greater variety of species, suggesting a richer microbial ecosystem. Conversely, lower bars suggest reduced diversity, potentially pointing to ecological imbalances or specific conditions affecting the microbial community.

To accurately assess our metrics, we will employ GMPR-normalized data at both the ASV Genus and Species levels. It's important to note that calculations of alpha diversity will be confined to categorical data. In contrast, for non-categorical data, our approach will involve a comprehensive 'one versus all' comparison across all samples.

Important Note: The use of Centered Log-Ratio (CLR) normalized data is precluded in our analysis for a specific reason. Post-normalization, the resulting data are centered around zero, encompassing both positive and negative values for frequencies. This presents a significant challenge in the context of Qiime2 software, which does not accommodate negative values in the count table for Alpha and Beta diversity analyses. While a potential workaround could involve shifting CLR data to ensure all values are positive, this approach is not ideal. Such a modification constitutes a substantial manipulation of the data, potentially compromising its integrity and skewing the results of our analyses. Therefore, we have opted against using CLR-normalized data to maintain the accuracy and reliability of our findings.

In the subsequent subsection, we will delve into an in-depth discussion of four pivotal metrics that have been integral to our research. These metrics are meticulously selected to encapsulate a comprehensive spectrum of information, each focusing on a distinct aspect of alpha diversity analysis:

- **Shannon Diversity (Diversity Class)**: This metric is a cornerstone in our analysis, offering insights into the ecological diversity of a community by considering both the abundance and evenness of species present.
- **Pielou's Evenness (Evenness Class)**: To gauge the uniformity of species distribution, we employ Pielou's evenness. This metric is critical in understanding how evenly individuals are distributed across the different species in our study.

- **Observed OTUs (Richness Class)**: Representing the 'Richness' category, the count of Observed OTUs (Operational Taxonomic Units) serves as a straightforward yet powerful indicator of species richness within our samples.
- **Faith's Phylogenetic Diversity (Phylogenetic Diversity Class)**: Unique to our study is the use of Faith's PD, which quantifies biodiversity by considering the branch lengths of a phylogenetic tree. This metric is particularly insightful, as it incorporates the evolutionary history of the species. However, it's pertinent to note that Faith's PD is applicable exclusively to ASV (Amplicon Sequence Variant) tables where a comprehensive phylogenetic tree is available.

Each of these metrics has been carefully chosen for their ability to provide a nuanced understanding of ecological diversity. By employing these diverse yet complementary measures, our analysis achieves a holistic view of the alpha diversity present in our study, ensuring a robust and comprehensive ecological assessment.

10.1.1 Diversity Class (Shannon Diversity)

The Diversity class encompasses metrics that measure both species richness and evenness. A prominent metric in this class is the Shannon Diversity Index.

- **Metric:**

$$H = -\sum_{i=1}^{S} p_i \log(p_i)$$

  where:

  - H is the Shannon index
  - S is the number of species (features)
  - p_i is the proportion of individuals belonging to species i

- **Interpretation:** Higher values of H indicate a community with diverse species that are also evenly distributed.

10.1.2 Evenness Class (Pielou's Evenness)

This class focuses on the distribution of individuals across different species, assessing how evenly the community's individuals are spread.

- **Metric:**

$$J' = \frac{H}{\log(S)}$$

  - J' represents Pielou's evenness
  - H the Shannon index
  - S the number of species

- **Interpretation:** Values near 1 indicate uniform distribution among species.

10.1.3 Richness Class (Observed OTUs)

The Richness class quantifies the number of different species present in a sample, offering a straightforward diversity assessment.

- **Metric:**

$$S^{obs}$$

  - S^{obs} The count of distinct species or OTUs

- **Interpretation:** A higher count indicates a richer variety of species.

10.1.4 Phylogenetic Diversity Class (Faith's PD)

Phylogenetic diversity measures incorporate the evolutionary relationships between species, providing a deeper understanding of community structure.

- **Metric:**

$$PD = \sum_{\text{all branches}} \text{length}$$

  - PD is the sum of the lengths of phylogenetic tree branches connecting the species in a sample

- **Interpretation:** Higher PD values indicate a broad evolutionary diversity.

10.1.5 Kruskal-Wallis Test

The Kruskal-Wallis test is a non-parametric statistical test used to determine if there are statistically significant differences between two or more groups of an independent variable on a continuous or ordinal dependent variable. In our case is used with the display of the alpha bar plot to have a better understanding of the data. Below we can brifely see how it works:

1. **Ranking the Data:**
   - Combine all data from different groups into a single dataset.
   - Rank all the observations, from the smallest to the largest. Assign average ranks in case of ties.
2. **Calculating the Test Statistic (H):** The formula for the Kruskal-Wallis test statistic, H, is given by:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{g} \frac{R_i^2}{n_i} - 3(N+1)$$

   where:
   - N is the total number of observations across all groups.
   - g is the number of groups.
   - R_i is the sum of the ranks for the i-th group.
   - n_i is the number of observations in the i-th group.
3. **Determining Statistical Significance:**
   - The test statistic H is approximately chi-squared distributed with $g-1$ degrees of freedom.
   - Compare the computed H value with the chi-squared distribution to determine the p-value.

10.2 Beta Diversity Emperor Plot

Emperor plots are a type of visualization tool used extensively in bioinformatics for representing complex, high-dimensional data in an interpretable three-dimensional space. These plots are particularly useful in microbiome studies, where they aid in visualizing and interpreting patterns within multivariate data sets.

10.2.1 Mathematical Foundations

The construction of an Emperor plot is based on the results of Principal Coordinates Analysis (PCoA) or other similar dimensionality reduction techniques. The key steps involved in PCoA include:

1. **Distance Matrix Computation:** The first step involves calculating a distance matrix D, where D_{ij} represents the distance between sample i and sample j.
2. **Double Centering and Matrix Transformation:** The distance matrix D is then transformed into a new matrix B using double centering:

$$B_{ij} = -\frac{1}{2}(D_{ij}^2 - \overline{D}_{i.}^2 - \overline{D}_{.j}^2 + \overline{D}^2)$$

   where $\overline{D}_{i.}$, $\overline{D}_{.j}$, and $\overline{D}$ are the mean distances for rows, columns, and the overall mean, respectively.

3. **Eigenvalue Decomposition:** The matrix B is subjected to eigenvalue decomposition to extract eigenvalues and eigenvectors. The eigenvectors correspond to the principal coordinates.

4. **Plotting:** The principal coordinates (usually the first three for a 3D plot) are used to plot the samples in the Emperor plot, providing a visual representation of the data's multivariate structure.

Several metrics can be used to calculate distances matrix D between samples in Emperor plots, each providing different insights into the data. In my analisys we are going to use.

10.2.2 Dissimilarity class (Bray-Curtis)

The Bray-Curtis dissimilarity is a widely used metric in ecological studies to quantify the difference between two samples based on species composition. It is defined as:

$$BC_{ij} = \frac{\sum_{k=1}^{S} |x_{ik} - x_{jk}|}{\sum_{k=1}^{S} (x_{ik} + x_{jk})}$$

where:

- x_{ik} and x_{jk} represent the quantities of species k in samples i and j, respectively.
- S is the total number of species.

This metric is sensitive to changes in species abundance and effectively measures the dissimilarity between two samples. The Bray-Curtis dissimilarity ranges from 0 (identical composition) to 1 (completely different composition) [12].

10.2.3 Dissimilarity class (Jaccard Index)

The Jaccard Index is another dissimilarity metric, primarily focused on species presence or absence rather than abundance. It is calculated as:

$$J_{ij} = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$

where:

- M_{11} is the number of species present in both samples.
- M_{01} and M_{10} are the number of species exclusive to each sample.

The Jaccard Index is particularly useful for comparing community composition when abundance data is not available or is less reliable.

10.2.4 Phylogenetic Dissimilarity Class(UniFrac Metrics)

UniFrac metrics, including both weighted and unweighted versions, offer a phylogenetic approach to measuring community dissimilarity, and they can be only used on ASV tables. These metrics incorporate phylogenetic information by comparing the genetic distances between sets of taxa in the samples. The weighted version considers both the presence/absence and the abundance of taxa, while the unweighted version focuses solely on the presence/absence. This makes UniFrac particularly insightful in studies where evolutionary relationships between species are of interest.

11 Alpha Diversity Bar Plots and Tables

In this section, we delve into the exploration of alpha diversity at various taxonomic levels - ASV (Amplicon Sequence Variant), genus, and species - in the context of specific groups and conditions such as **Diarrhea** and **Is Sow**. Each subsection presents a series of bar plots, visually comparing alpha diversity metrics across different groups, followed by corresponding tables displaying the results of **Kruskal-Wallis tests**. *Note*: The bar plot are of **time** and **sex** are not showed because we focused more due to the result on the other two.

11.1 ASV Alpha Bar Plot

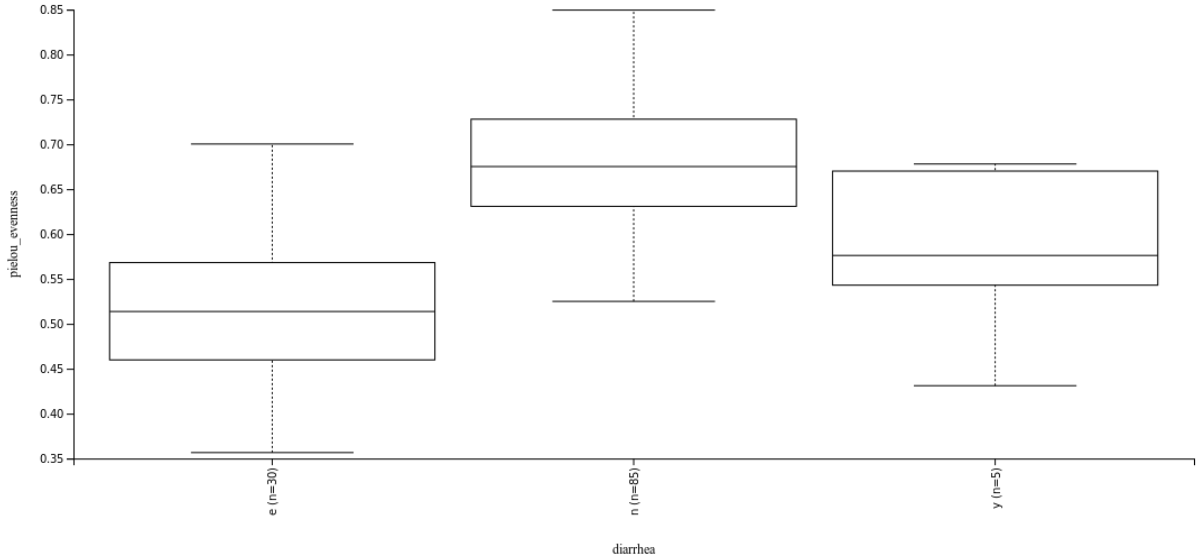


Figure 39: **Evenness Vector for Diarrhea at ASV Level**: On the x-axis, the graph displays the three distinct labels: e, n, and y, representing different categories within the dataset. The y-axis illustrates the values of the alpha evenness vector at the ASV level, quantifying the diversity within each category.

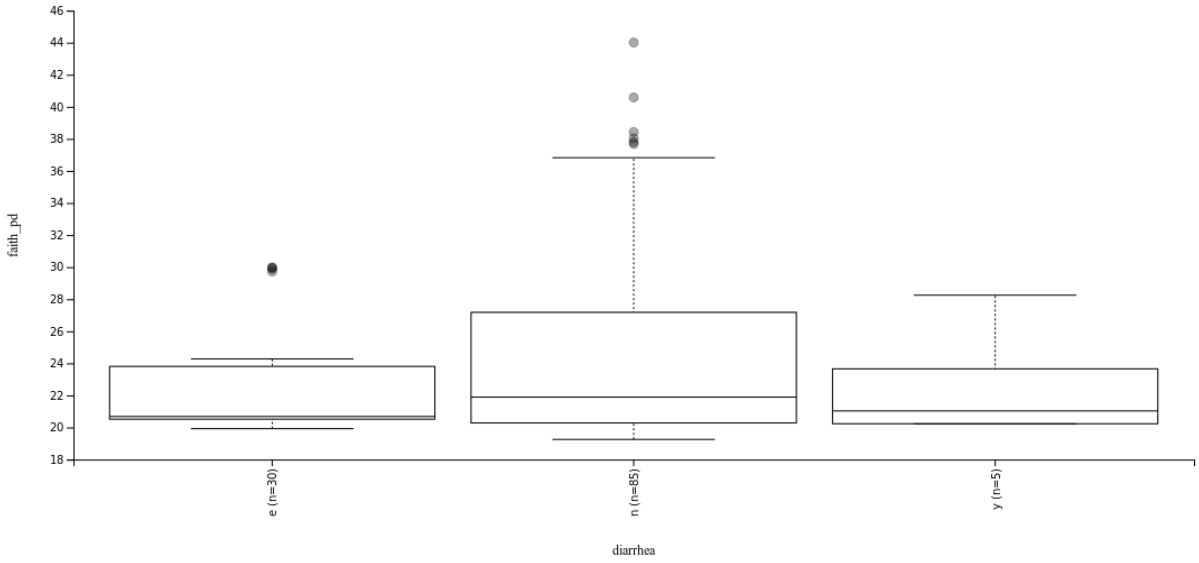


Figure 40: **Faith's PD Vector for Diarrhea at ASV Level**: On the x-axis, the graph displays the three distinct labels: e, n, and y, representing different categories within the dataset. The y-axis illustrates the values of Faith's PD vector at the ASV level, reflecting the phylogenetic diversity within each category.

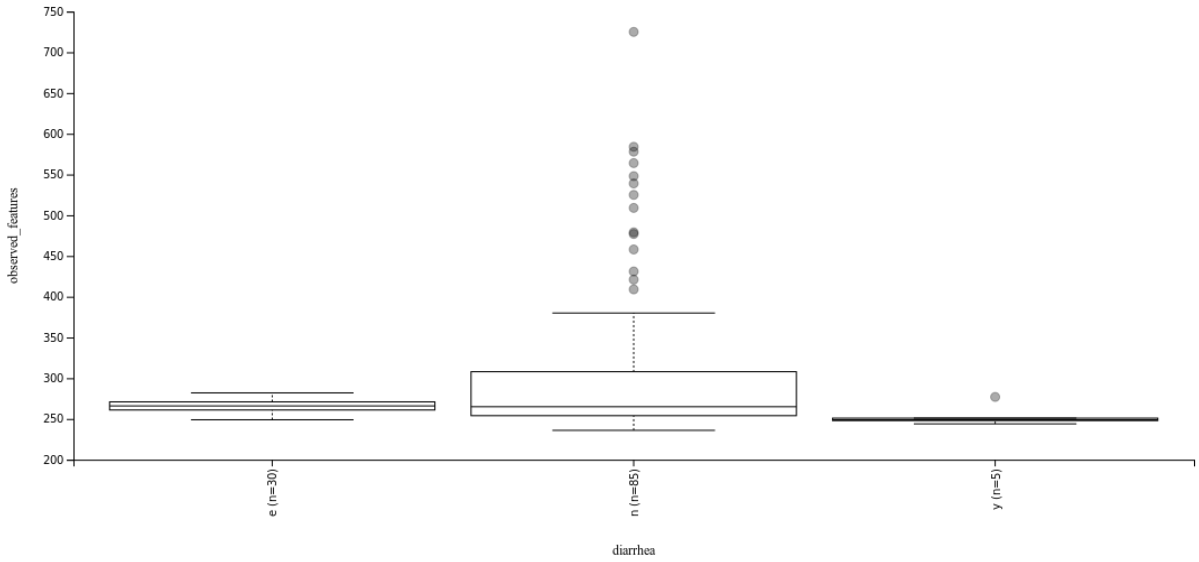


Figure 41: **Observed Feature for Diarrhea at ASV Level**: The x-axis of the graph shows the three distinct labels: e, n, and y, indicative of different categories. The y-axis represents the count of observed features at the ASV level, indicating the variety of taxa present in each category.

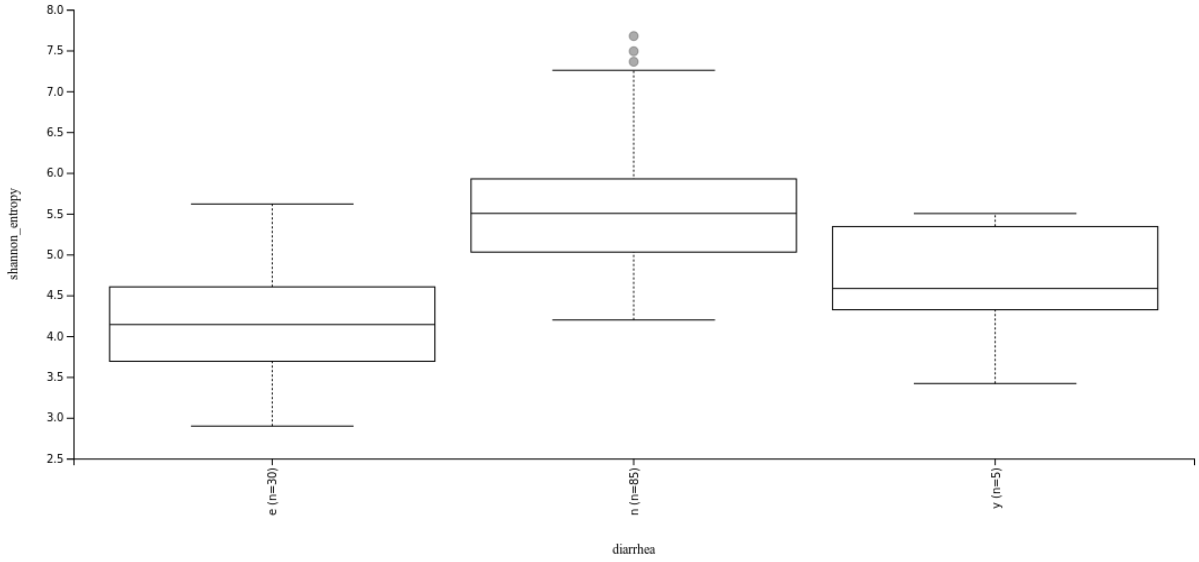


Figure 42: **Shannon Vector for Diarrhea at ASV Level**: This graph categorizes the data along the x-axis using labels e, n, and y. The y-axis shows the Shannon diversity index at the ASV level, illustrating the ecological richness and evenness of each category.

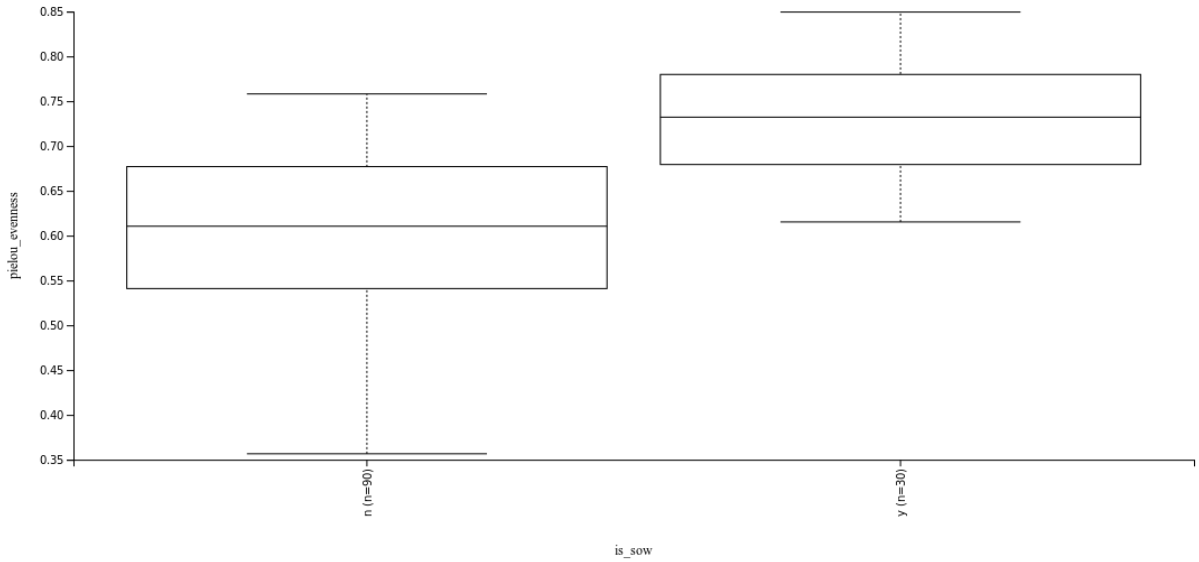


Figure 43: **Evenness Vector for Is Sow at ASV Level**: Displayed on the x-axis are the labels n and y, representing different categories. The y-axis quantifies the alpha evenness at the ASV level, providing insights into the diversity within each category.

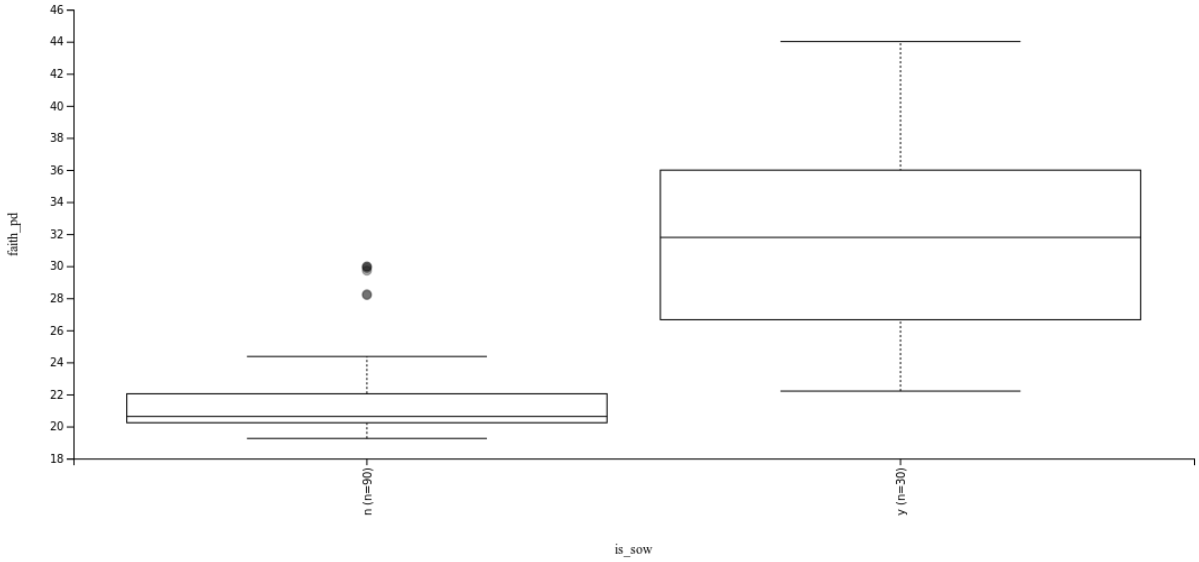


Figure 44: **Faith's PD Vector for Is Sow at ASV Level**: The graph shows labels n and y on the x-axis, denoting distinct categories. The y-axis reveals the Faith's PD vector values at the ASV level, reflecting the phylogenetic diversity in each category.

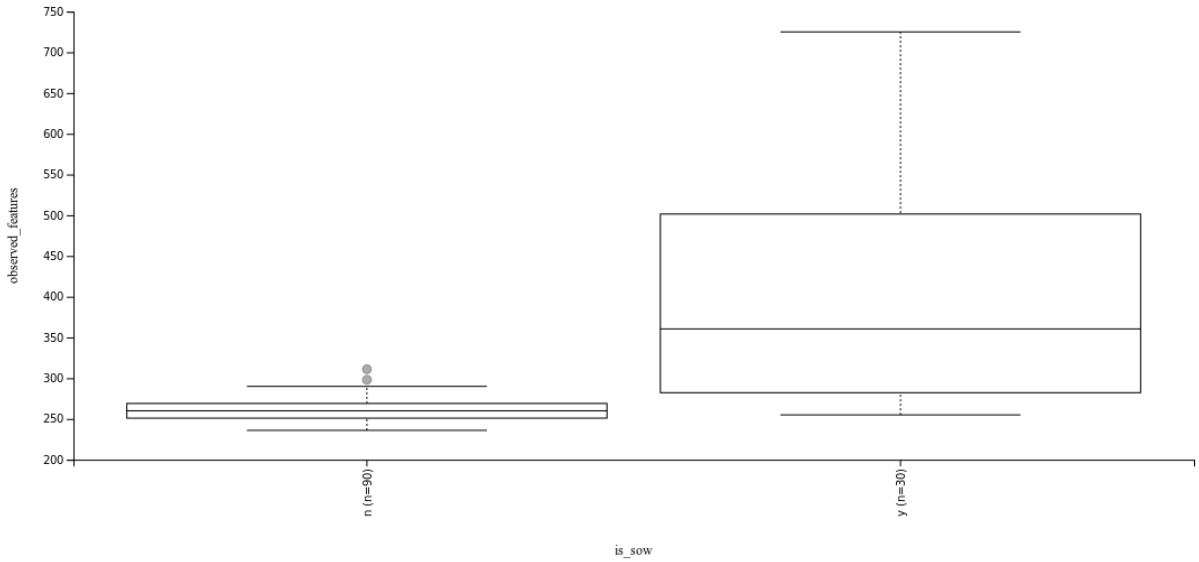


Figure 45: **Observed Feature for Is Sow at ASV Level**: In this graph, the x-axis presents the labels n and y for different categories. The y-axis indicates the number of observed features at the ASV level, shedding light on the taxa variety in each category.

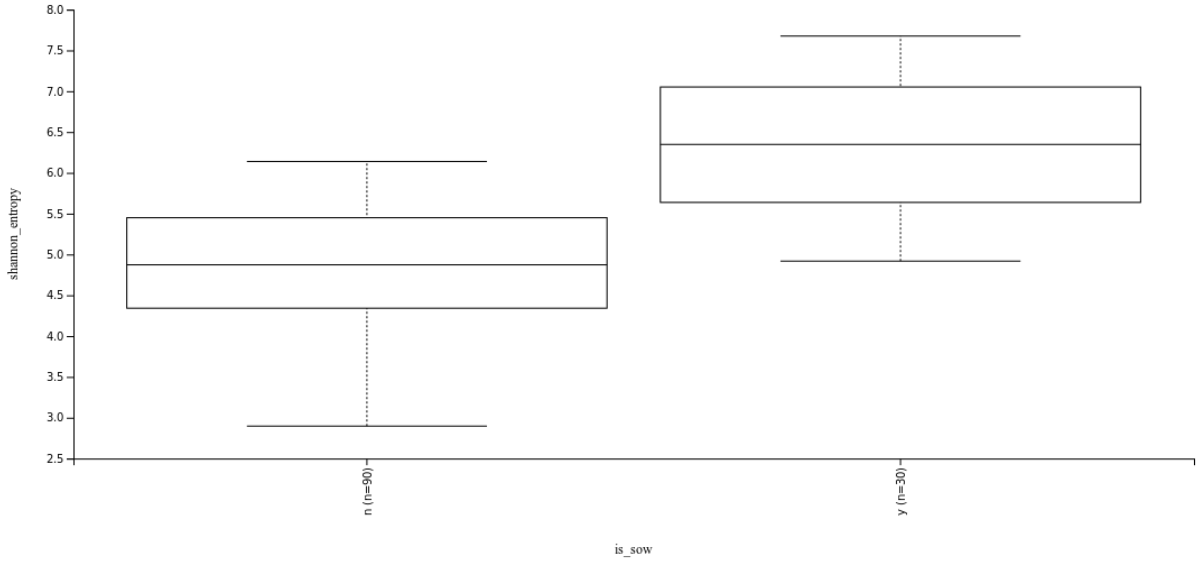


Figure 46: **Shannon Vector for Is Sow at ASV Level**: This graph illustrates the Shannon diversity index at the ASV level on the y-axis, with the x-axis categorizing the data under labels n and y, signifying different categories.

11.2 ASV Kruskal-Wallis Results

Table 27: Overall Test Results for Alpha Diversity Metrics at the **ASV** level

Metric	Sex Label		Diarrhea Label	
	H-value	**P-value**	**H-value**	**P-value**
Evenness	9.99	0.00157	57.75	2.88×10^{-13}
Observed Features	12.47	0.000413	3.84	0.1466
Shannon	10.92	0.000949	57.33	3.55×10^{-13}
Faith PD	9.29	0.00230	11.77	0.00279
	Is Sow Label		**Time Label**	
	H-value	**P-value**	**H-value**	**P-value**
Evenness	39.80	2.81×10^{-10}	33.87	4.41×10^{-8}
Observed Features	41.54	1.16×10^{-10}	18.13	0.000116
Shannon	45.34	1.66×10^{-11}	31.37	1.54×10^{-7}
Faith PD	56.48	5.68×10^{-14}	1.70	0.4269

Table 28: Kruskal-Wallis Pairwise Comparisons for Alpha Diversity Metrics at ASV Level for **Sex Label**

Metric	Comparison	H-value	P-value	Q-value
Evenness	f vs m	9.99	0.001575	0.001575
Observed Features	f vs m	12.47	0.000413	0.000413
Shannon	f vs m	10.92	0.000949	0.000949
Faith PD	f vs m	9.29	0.002303	0.002303

Table 29: Kruskal-Wallis Pairwise Comparisons for Alpha Diversity Metrics at ASV Level for **Diarrhea Label**

Metric	Comparison	H-value	P-value	Q-value
Evenness	e vs n	53.185598	$3.034756e-13$	$9.104268e-13$
	y vs e	0.980000	0.3221988	0.3221988
	n vs y	8.705236	0.0031729	0.0047595
Observed Features	e vs n	0.328875	0.566322	0.566322
	y vs e	3.564041	0.059044	0.102245
	n vs y	3.326705	0.068163	0.102245
Shannon	e vs n	52.907262	$3.496744e-13$	$1.049023e-12$
	y vs e	0.980000	0.3221988	0.3221988
	n vs y	8.498591	0.0035542	0.0053313
Faith PD	e vs n	11.568195	0.000671	0.002013
	y vs e	0.500000	0.479500	0.479500
	n vs y	0.700065	0.402762	0.479500

Table 30: Kruskal-Wallis Pairwise Comparisons for Alpha Diversity Metrics at ASV Level for **Is Sow Label**

Metric	Comparison	H-value	P-value	Q-value
Evenness	n vs y	39.804628	$2.806792e-10$	$2.806792e-10$
Observed Features	n vs y	41.536995	$1.156614e-10$	$1.156614e-10$
Shannon	n vs y	45.337778	$1.658196e-11$	$1.658196e-11$
Faith PD	n vs y	56.477502	$5.684506e-14$	$5.684506e-14$

Table 31: Kruskal-Wallis Pairwise Comparisons for Alpha Diversity Metrics at ASV Level for **Time Label**

Metric	Comparison	H-value	P-value	Q-value
Evenness	T0 vs T1	13.724537	$2.116709e-04$	$3.175064e-04$
	T0 vs T2	27.805926	$1.341147e-07$	$4.023441e-07$
	T1 vs T2	9.900833	0.0016520	0.0016520
Observed Features	T0 vs T1	9.313160	0.002275	0.003413
	T0 vs T2	2.600819	0.106809	0.106809
	T1 vs T2	15.354427	0.000089	0.000267
Shannon	T0 vs T1	12.813333	$3.441579e-04$	$5.166868e-04$
	T0 vs T2	25.618148	$4.161075e-07$	$1.24832e-06$
	T1 vs T2	9.245926	0.0023602	0.0023602
Faith PD	T0 vs T1	0.088981	0.765476	0.765476
	T0 vs T2	1.893426	0.168816	0.506448
	T1 vs T2	0.592593	0.441418	0.662127

11.2.1 ASV Alpha Metrics conclusion

Examining the Kruskal-Wallis test results across various labels (Diarrhea, Time, Sow, and Sex) for alpha diversity metrics at the ASV level provides valuable insights:

- **Significant Variation in Evenness and Shannon Diversity**: Both the Evenness and Shannon metrics show significant differences across most labels, with particularly low p-values (e.g., 2.88×10^{-13} for Diarrhea and 1.66×10^{-11} for Sow in Shannon diversity). This indicates considerable variability in microbial community composition related to these labels.
- **Observed Features Variation**: For the Observed Features metric, notable differences are evident in some labels (e.g., Time and Sow labels) with significant p-values (e.g., 1.16×10^{-10} for Sow). However, this metric shows less variability in the Diarrhea and Sex labels.
- **Faith PD Pielou**: This metric consistently shows significant differences across all labels, indicating variations in phylogenetic diversity linked to each categorical label.

11.3 Genus Alpha Bar Plot

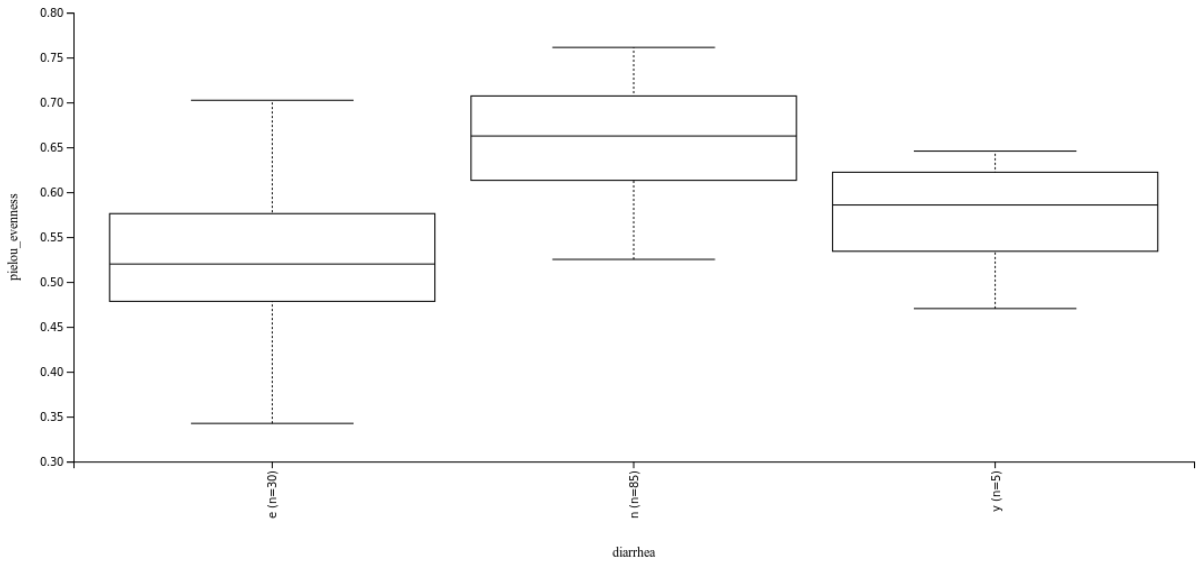


Figure 47: **Evenness Vector for Diarrhea at Genus Level**: On the x-axis, the graph displays the distinct labels associated with diarrhea, representing different categories within the dataset. The y-axis illustrates the values of the alpha evenness vector at the genus level, quantifying the diversity within each category.

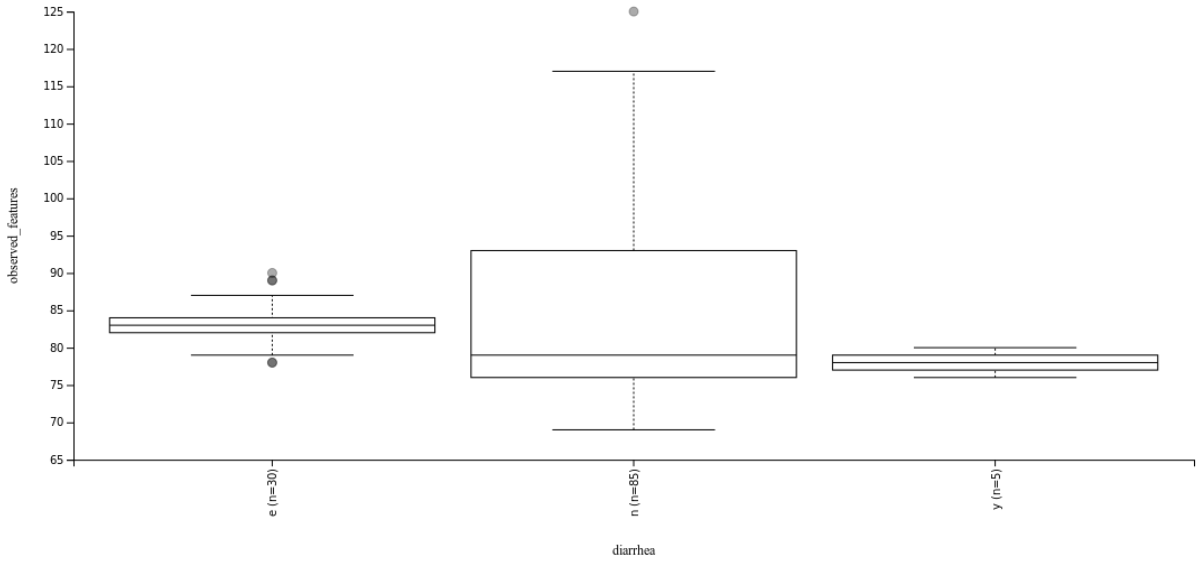


Figure 48: **Observed Feature for Diarrhea at Genus Level**: This graph shows the different categories related to diarrhea on the x-axis. The y-axis indicates the count of observed features at the genus level, providing insight into the variety of taxa within each category.

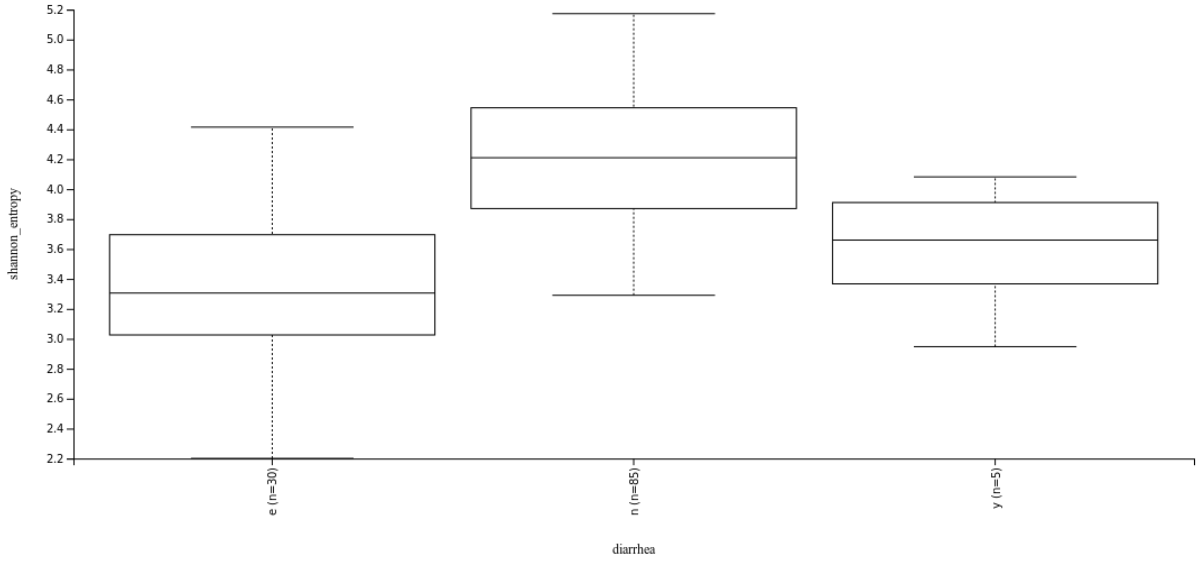


Figure 49: **Shannon Vector for Diarrhea at Genus Level**: Here, the x-axis represents the various categories linked to diarrhea. The y-axis shows the Shannon diversity values, reflecting the ecological richness and evenness at the genus level for each category.

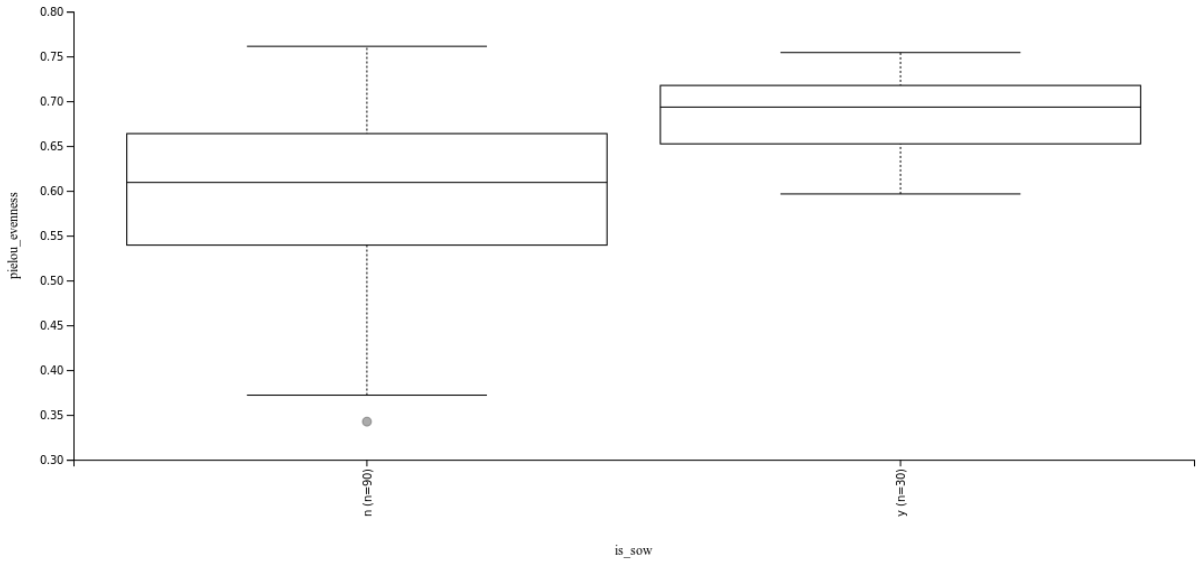


Figure 50: **Evenness Vector for Is Sow at Genus Level**: On the x-axis, the graph categorizes the data based on the 'Is Sow' label. The y-axis quantifies the evenness of the microbial community at the genus level within each category.

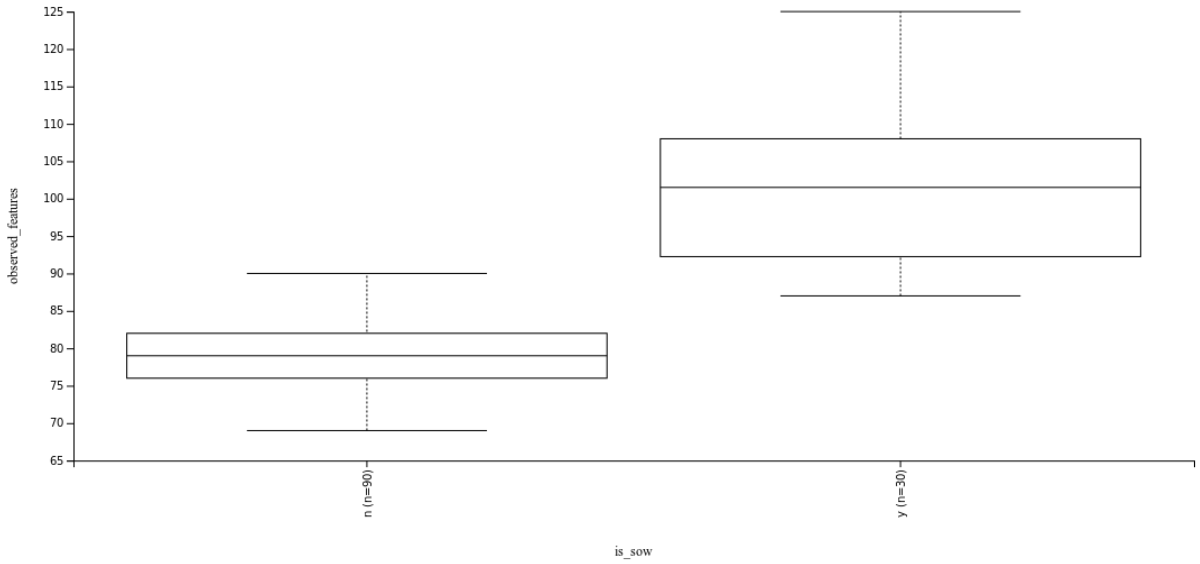


Figure 51: **Observed Feature for Is Sow at Genus Level**: This graph highlights the observed features at the genus level with the 'Is Sow' categorization on the x-axis. It reveals the variation in the number of distinct features observed in each category.

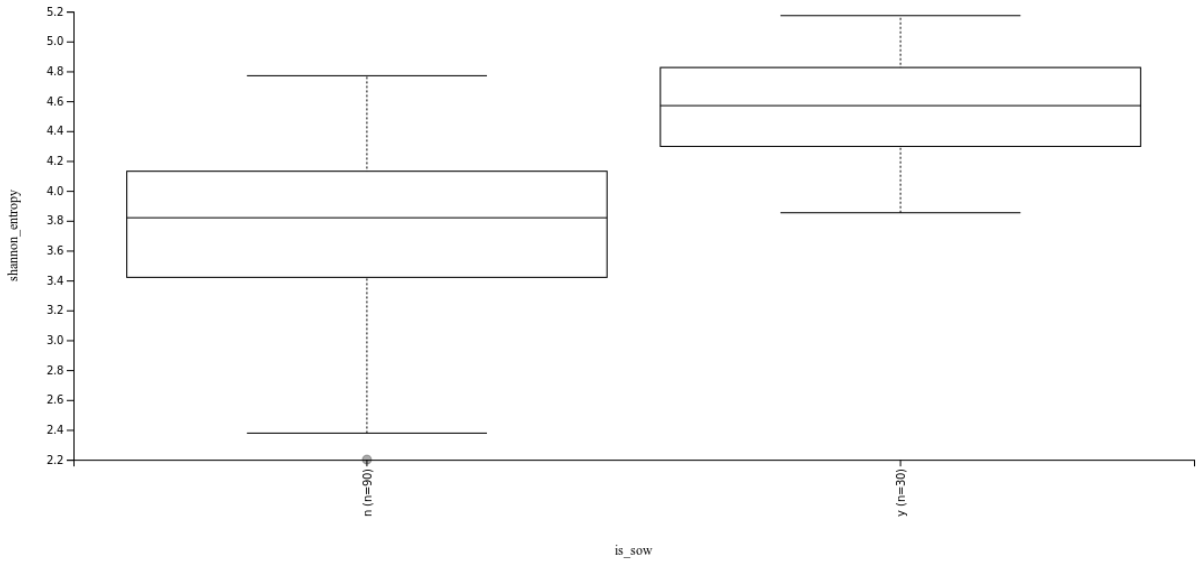


Figure 52: **Shannon Vector for Is Sow at Genus Level**: The graph presents the Shannon diversity index for categories based on the 'Is Sow' label on the x-axis. The y-axis illustrates the Shannon diversity measure at the genus level, indicating the ecological diversity within each category.

Table 32: Overall Test Results for Alpha Diversity Metrics at the **Genus** level

Metric	**Sex Label**		**Diarrhea Label**	
	H-value	**P-value**	**H-value**	**P-value**
Evenness	10.40	0.00126	53.04	3.04×10^{-12}
Observed Features	12.63	0.000379	5.16	0.0759
Shannon	12.05	0.000518	52.58	3.82×10^{-12}
	Is Sow Label		**Time Label**	
	H-value	**P-value**	**H-value**	**P-value**
Evenness	45.91	1.24×10^{-11}	24.43	4.96×10^{-6}
Observed Features	64.54	9.47×10^{-16}	2.82	0.2438
Shannon	54.22	1.79×10^{-13}	22.32	1.42×10^{-5}

Table 33: Kruskal-Wallis Pairwise Comparisons for Alpha Diversity Metrics at Genus Level for **Sex Label**

Metric	**Comparison**	**H-value**	**P-value**	**Q-value**
Evenness	f vs m	10.40	0.00126	0.00126
Observed Features	f vs m	12.63	0.000379	0.000379
Shannon	f vs m	12.05	0.000518	0.000518

Table 34: Kruskal-Wallis Pairwise Comparisons for Alpha Diversity Metrics at Genus Level for **Diarrhea Label**

Metric	Comparison	H	P-value	Q-value
Evenness	e vs n	48.38	3.52×10^{-12}	1.06×10^{-11}
	y vs e	0.57	0.4507	0.4507
	n vs y	8.71	3.17×10^{-3}	4.76×10^{-3}
Observed Features	e vs n	2.54	0.1110	0.1664
	y vs e	1.49	0.2226	0.2226
	n vs y	3.05	0.0805	0.1664
Shannon	e vs n	47.93	4.41×10^{-12}	1.32×10^{-11}
	y vs e	0.38	0.5400	0.5400
	n vs y	8.71	3.17×10^{-3}	4.76×10^{-3}

Table 35: Kruskal-Wallis Pairwise Comparisons for Alpha Diversity Metrics at Genus Level for **Is Sow Label**

Metric	Comparison	H	P-value	Q-value
Evenness	n vs y	45.91	1.24×10^{-11}	1.24×10^{-11}
Observed Features	n vs y	64.54	9.47×10^{-16}	9.47×10^{-16}
Shannon	n vs y	54.22	1.79×10^{-13}	1.79×10^{-13}

Table 36: Kruskal-Wallis Pairwise Comparisons for Alpha Diversity Metrics at Genus Level for **Time Label**

Metric	Comparison	H	P-value	Q-value
Evenness	T0 vs T1	10.83	9.99×10^{-4}	1.50×10^{-3}
	T0 vs T2	20.02	8.00×10^{-6}	2.30×10^{-5}
	T1 vs T2	6.21	0.0127	0.0127
Observed Feature	T0 vs T1	2.52	0.1122	0.2963
	T0 vs T2	0.04	0.8430	0.8430
	T1 vs T2	1.66	0.1975	0.2963
Shannon	T0 vs T1	9.60	1.95×10^{-3}	2.92×10^{-3}
	T0 vs T2	18.34	1.90×10^{-5}	5.60×10^{-5}
	T1 vs T2	5.93	0.0149	0.0149

11.3.1 Genus Alpha Metrics Conclusion

The analysis of Kruskal-Wallis test results for alpha diversity metrics at the Genus level, considering various labels (Sex, Diarrhea, Is Sow, and Time), yields key insights:

- **Significant Variation in Evenness and Shannon Diversity**: A significant variance is observed in both Evenness and Shannon metrics across most labels. Notably, the Shannon diversity shows pronounced variability with substantial p-values, such as 3.04×10^{-12} for Diarrhea and 1.79×10^{-13} for Is Sow. This underscores the significant influence of these labels on microbial community structure.
- **Observed Features Variation**: The Observed Features metric presents significant differences under certain labels. For instance, it shows high significance with a p-value of 9.47×10^{-16} for the Is Sow label, while exhibiting less variability with the Sex and Diarrhea labels, suggesting that specific factors more profoundly affect the observable variety in microbial features at the Genus level.
- **Consistency across Labels**: The consistent pattern of significance across different labels for all metrics (Evenness, Observed Features, and Shannon) highlights the robust impact of these categorical factors on the diversity within microbial communities at the Genus level.

11.4 Species Alpha Bar Plot

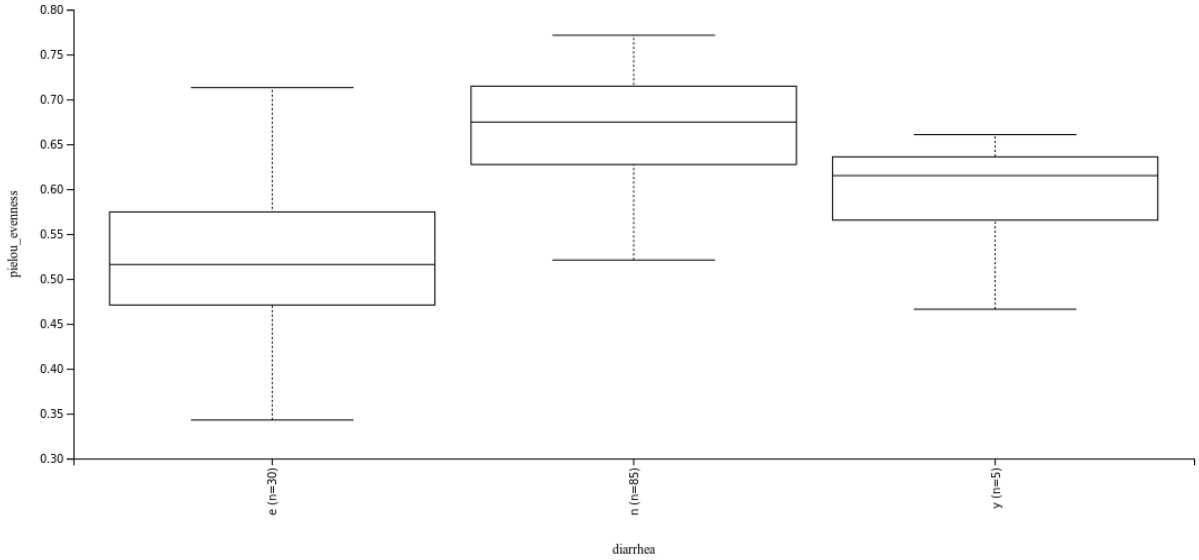


Figure 53: **Evenness Vector for Diarrhea at Species Level**: The x-axis of the graph categorizes data based on different conditions related to diarrhea, while the y-axis quantifies the evenness of the microbial community at the species level within each category.

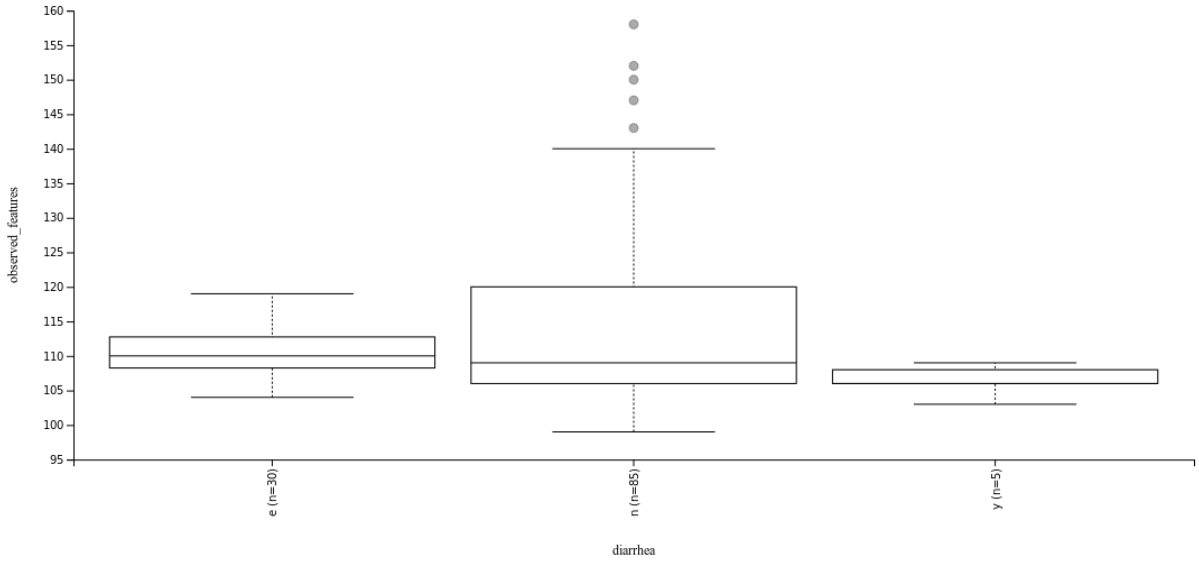


Figure 54: **Observed Feature for Diarrhea at Species Level**: This graph presents the count of observed features at the species level for different diarrhea-related conditions (x-axis). The y-axis indicates the diversity in the number of distinct features observed within each category.

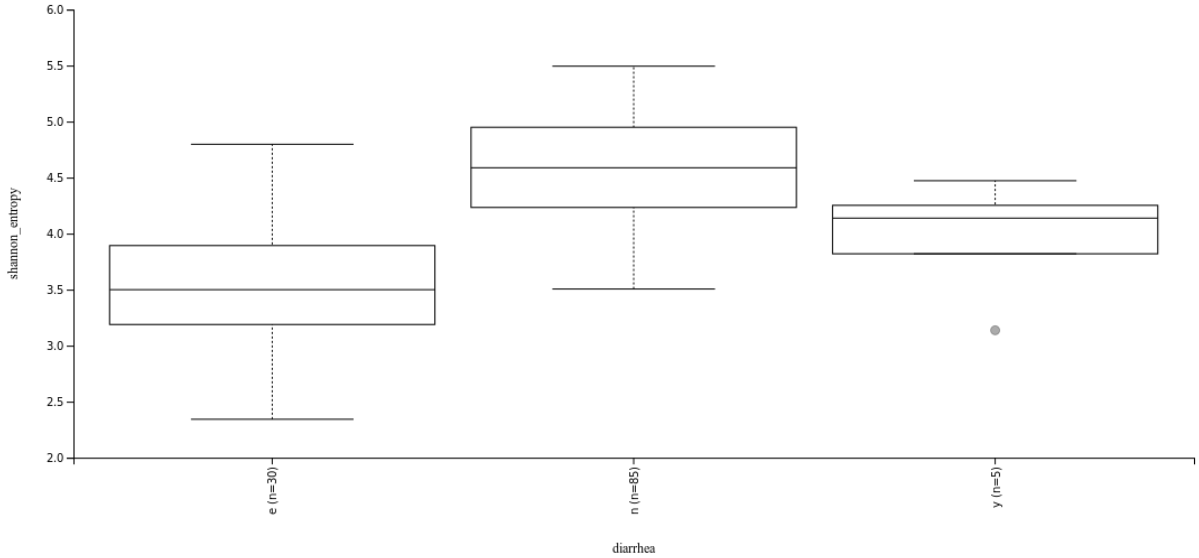


Figure 55: **Shannon Vector for Diarrhea at Species Level**: Here, the x-axis displays various categories associated with diarrhea. The y-axis shows the Shannon diversity values, reflecting the ecological richness and evenness at the species level for each category.

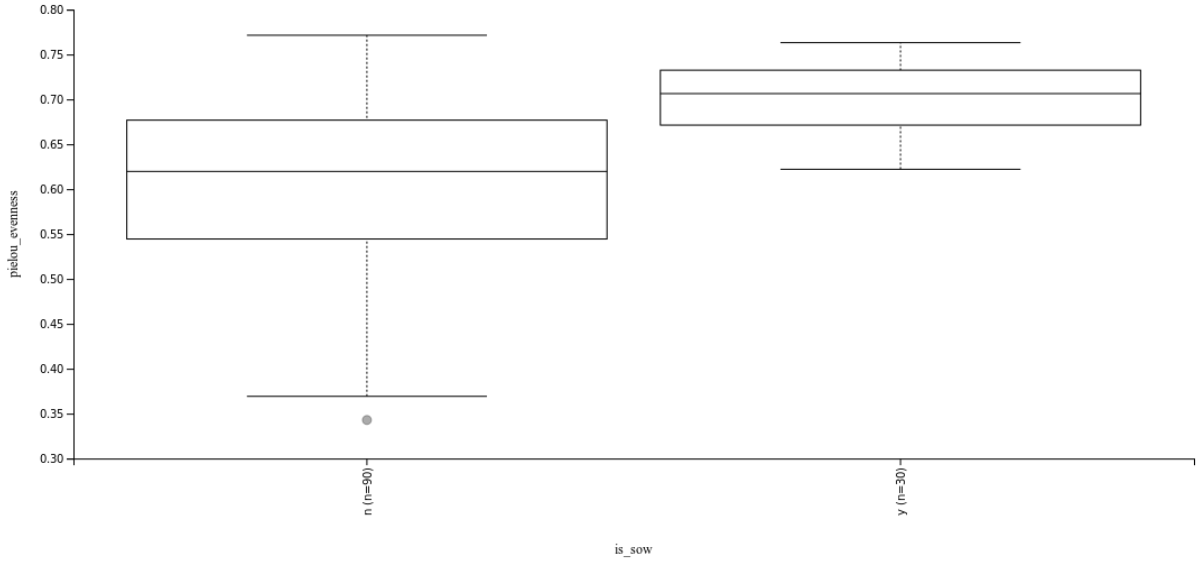


Figure 56: **Evenness Vector for Is Sow at Species Level**: On the x-axis, the graph categorizes data based on the 'Is Sow' label. The y-axis shows the evenness of the microbial community at the species level within each category, illustrating the balance of species presence.

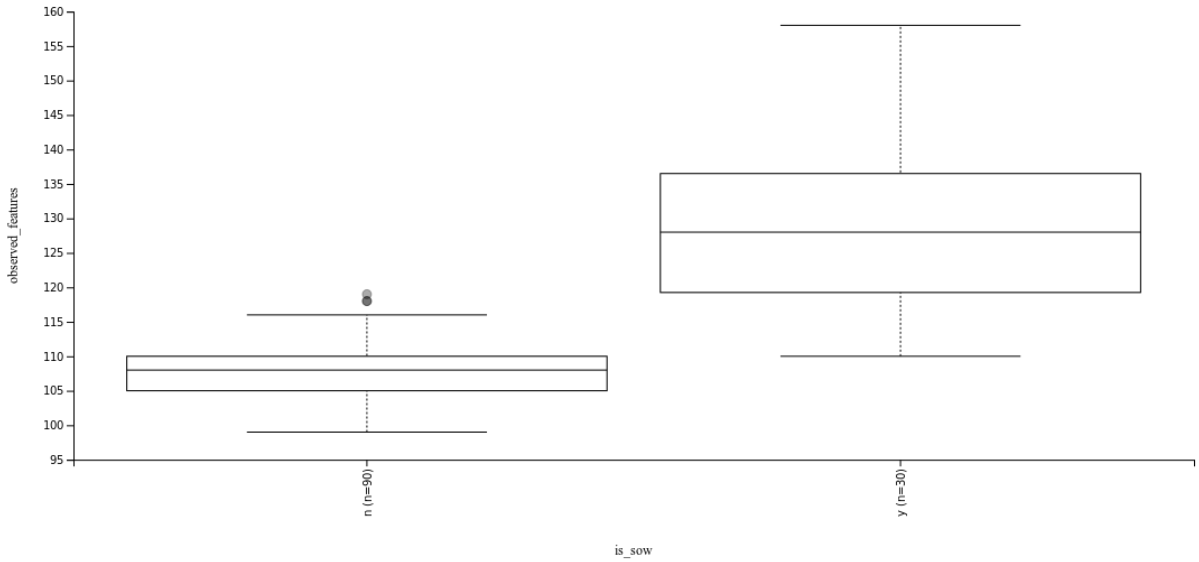


Figure 57: **Observed Feature for Is Sow at Species Level**: This graph illustrates the variety of observed features at the species level, with categories based on the 'Is Sow' label on the x-axis. The y-axis reveals the variation in the number of distinct species observed in each category.

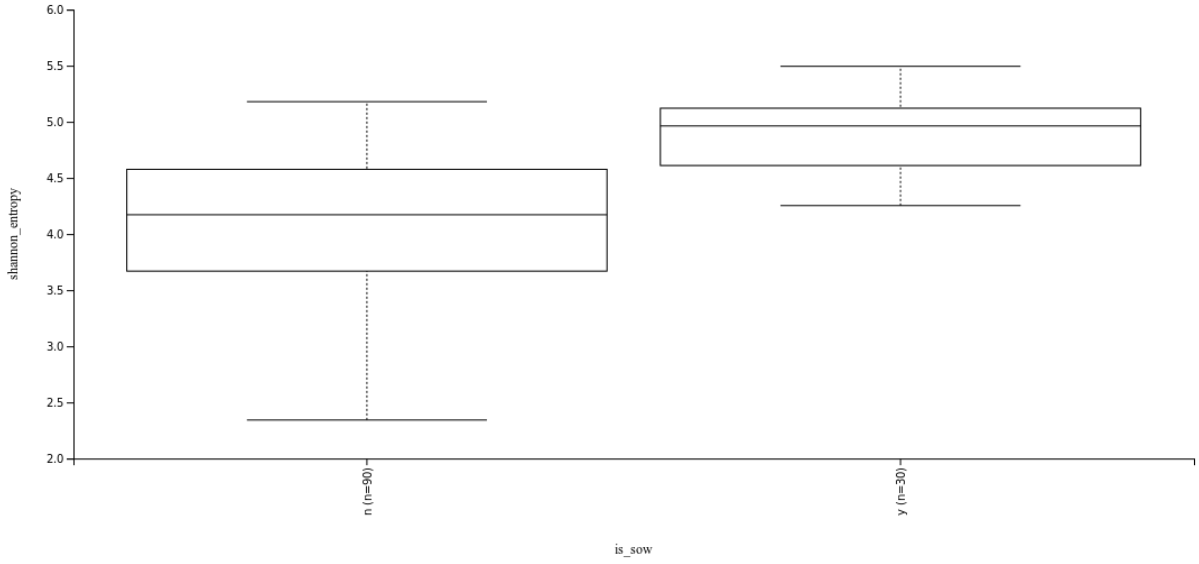


Figure 58: **Shannon Vector for Is Sow at Species Level**: Displaying the Shannon diversity index for categories based on the 'Is Sow' label (x-axis), this graph quantifies the ecological diversity at the species level for each category, indicated on the y-axis.

Table 37: Overall Test Results for Alpha Diversity Metrics at the **Species** level

Metric	Sex Label		Diarrhea Label	
	H-value	**P-value**	**H-value**	**P-value**
Evenness	9.85	0.00169	57.47	3.31×10^{-13}
Observed Features	11.91	0.000557	6.47	0.0393
Shannon	11.86	0.000572	57.57	3.15×10^{-13}
	Is Sow Label		**Time Label**	
	H-value	**P-value**	**H-value**	**P-value**
Evenness	33.99	5.53×10^{-9}	32.74	7.77×10^{-8}
Observed Features	51.22	8.24×10^{-13}	4.31	0.1160
Shannon	41.74	1.04×10^{-10}	31.56	1.40×10^{-7}

Table 38: Kruskal-Wallis Pairwise Comparisons for Alpha Diversity Metrics at Species Level for **Sex Label**

Metric	Comparison	H-value	P-value	Q-value
Evenness	f vs m	9.85	0.001695	0.001695
Observed Features	f vs m	11.91	0.000557	0.000557
Shannon	f vs m	11.86	0.000572	0.000572

Table 39: Kruskal-Wallis Pairwise Comparisons for Alpha Diversity Metrics at Species Level for **Diarrhea Label**

Metric	Comparison	H-value	P-value	Q-value
Evenness	e vs n	52.999959	$3.335553e-13$	$1.000666e-12$
	y vs e	1.280000	0.2578990	0.2578990
	n vs y	8.498591	0.0035542	0.0053313
Observed Features	e vs n	3.128228	0.076947	0.09375
	y vs e	2.808767	0.09375	0.09375
	n vs y	3.634219	0.056603	0.09375
Shannon	e vs n	52.814645	$3.665579e-13$	$1.099674e-12$
	y vs e	1.175556	0.2782627	0.2782627
	n vs y	8.914363	0.0028294	0.0042440

Table 40: Kruskal-Wallis Pairwise Comparisons for Alpha Diversity Metrics at Species Level for **Is Sow Label**

Metric	Comparison	H-value	P-value	Q-value
Evenness	n vs y	33.992433	$5.532681e-09$	$5.532681e-09$
Observed Features	n vs y	51.223129	$8.244155e-13$	$8.244155e-13$
Shannon	n vs y	41.739431	$1.042845e-10$	$1.042845e-10$

Table 41: Kruskal-Wallis Pairwise Comparisons for Alpha Diversity Metrics at Species Level for **Time Label**

Metric	Comparison	H-value	P-value	Q-value
Evenness	T0 vs T1	14.373704	$1.498809e-04$	$2.248213e-04$
	T0 vs T2	26.601481	$2.500519e-07$	$7.501558e-07$
	T1 vs T2	8.783704	0.0030393	0.0030393
Observed Features	T0 vs T1	2.870785	0.090201	0.135301
	T0 vs T2	0.223699	0.636236	0.636236
	T1 vs T2	3.379180	0.066025	0.135301
Shannon	T0 vs T1	13.300093	$2.653930e-04$	$3.98390e-04$
	T0 vs T2	25.230000	$5.088450e-07$	$1.52653e-06$
	T1 vs T2	9.600370	0.0019454	0.0019454

11.4.1 Species Alpha Metrics Conclusion

The comprehensive analysis of alpha diversity metrics at the Species level, considering labels such as Sex, Diarrhea, Is Sow, and Time, offers insightful conclusions:

- **Notable Variations in Evenness and Shannon Diversity**: There are significant variations in both Evenness and Shannon diversity metrics across all labels. Particularly striking are the results for the Diarrhea label, where both metrics show extremely low p-values (e.g., Evenness at 3.31×10^{-13} and Shannon at 3.15×10^{-13}), indicating a strong influence of diarrhea on species diversity.
- **Observed Features Metrics**: The Observed Features metric demonstrates significant variations under the Is Sow label (p-value of 8.24×10^{-13}), suggesting a substantial influence of this label on species variety. However, it shows more modest differences for the Diarrhea and Sex labels, indicating a varied impact across labels.
- **Consistency Across Labels**: The metrics show a consistent pattern of significance across different labels, underscoring the impact of factors like Sex, Diarrhea status, Is Sow status, and time on species diversity. Particularly, the Time label shows significant temporal variations, as seen in the Evenness and Shannon metrics.
- **Pairwise Comparisons Reinforcing Trends**: The pairwise comparisons for each label corroborate these findings. Notably, in the Sex label, both females and males exhibit significant differences in all metrics. The Diarrhea label shows marked differences, especially between 'e' vs 'n' groups in Evenness and Shannon, while the Is Sow and Time labels demonstrate significant variability across all metrics, emphasizing the influence of these factors on species diversity.

11.5 Alpha Analysis overall conclusion

This comprehensive analysis of Kruskal-Wallis test results across various alpha diversity metrics at different taxonomic levels elucidates the intricate influences of *Is Sow, Time, Diarrhea*, and *Sex* on microbial community composition.

Dominant Influence of Is Sow Category: Notably, the 'Is Sow' category consistently shows the highest Kruskal-Wallis H and Q values, especially in metrics like Observed Features and Shannon Entropy, underlining its significant role in microbial diversity.

Variable Impact of Time and Diarrhea:

- **Time**: This category significantly affects the number and distribution of microbial entities. For instance, the Observed Features metric shows notable changes over time, reflecting dynamic shifts in the microbial community composition. This could be attributed to environmental influences or host-related factors affecting microbial balance.
- **Diarrhea**: The influence of diarrhea is particularly evident in diversity and evenness metrics like Shannon Vector and Evenness. These findings suggest that diarrhea primarily impacts the distribution and relative abundance of the microbial species rather than altering the unique variety or the phylogenetic structure of the microbial community.

Limited Role of Sex: The analysis reveals that 'Sex' does not significantly differentiate alpha diversity, as indicated by the consistently lower H values across all metrics. This suggests a minimal impact of gender on the microbial diversity within the scope of this study.

12 Beta Diversity Emperor Plots

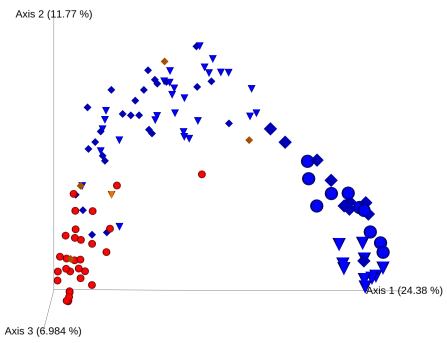


Figure 59: Bray Curtis - ASV View 1

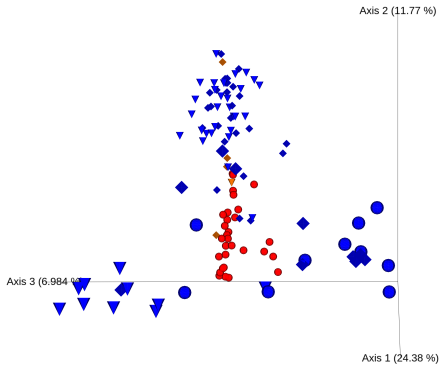


Figure 60: Bray Curtis - ASV View 2

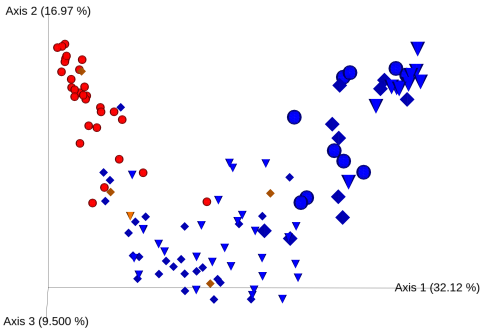


Figure 61: Bray Curtis - Genus View 1

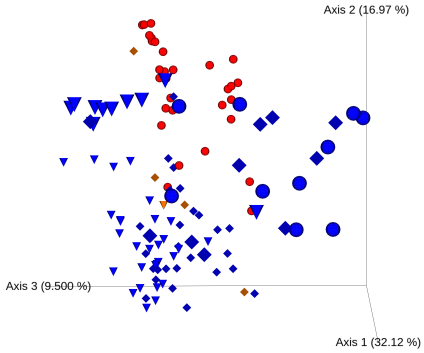


Figure 62: Bray Curtis - Genus View 2

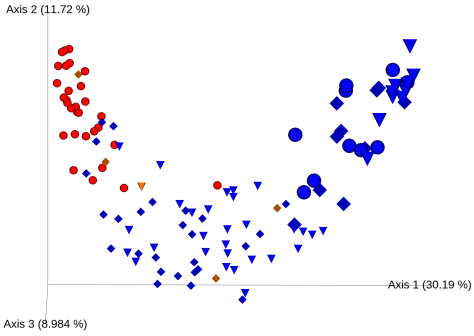


Figure 63: Bray Curtis - Species View 1

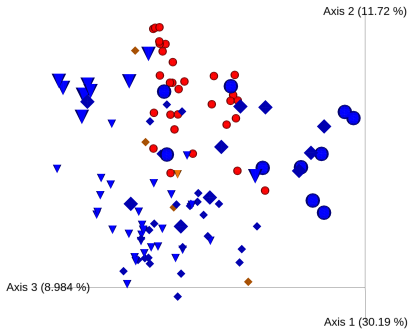


Figure 64: Bray Curtis - Species View 2

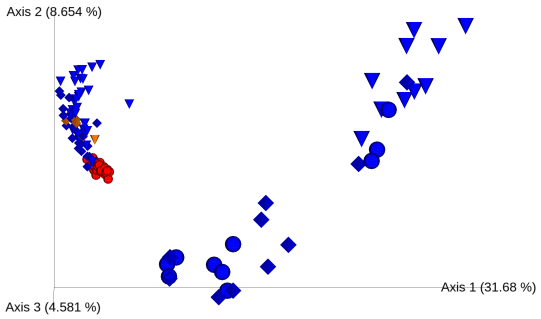


Figure 65: Jaccard - ASV View 1

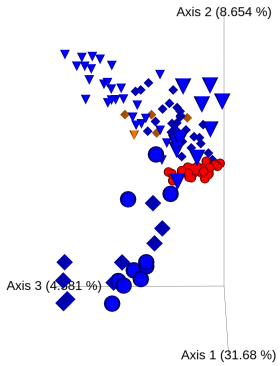


Figure 66: Jaccard - ASV View 2

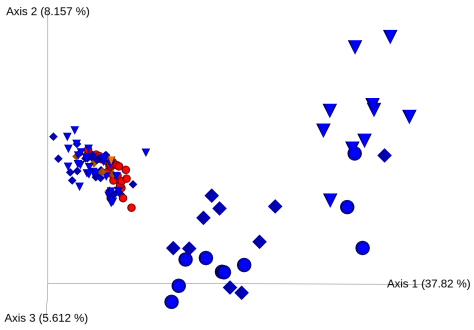


Figure 67: Jaccard - Genus View 1

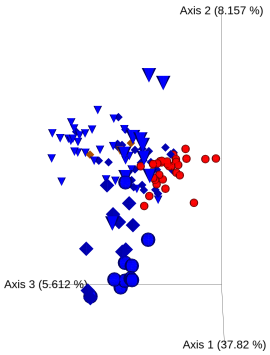


Figure 68: Jaccard - Genus View 2

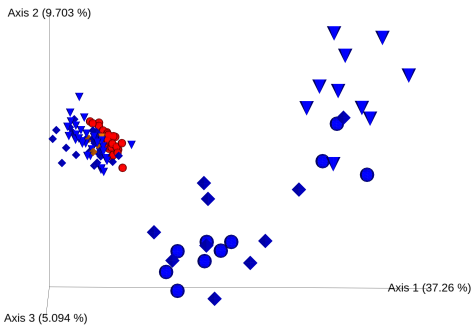


Figure 69: Jaccard - Species View 1

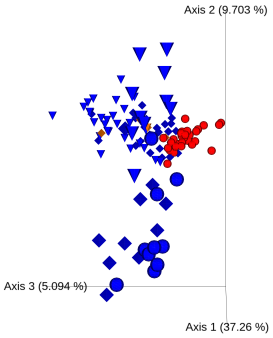


Figure 70: Jaccard - Species View 2

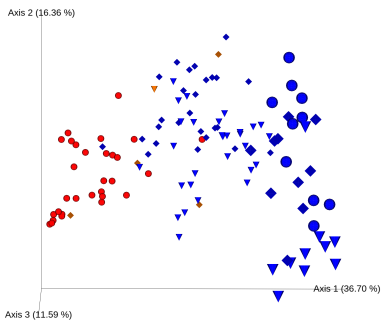


Figure 71: Weighted UniFrac - ASV View 1

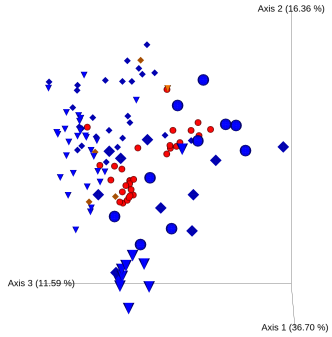


Figure 72: Weighted UniFrac - ASV View 2

12.1 Beta Analysis Conclusion

In our investigation, Emperor plots were utilized to visualize the beta diversity of microbial communities. Metrics such as Bray-Curtis, Jaccard, and both weighted and unweighted UniFrac were employed at the ASV level. This approach was in accordance with our alpha diversity analysis expectations, predicting a clear distinction between piglets associated with the sow (**is sow**) and those that were not. Additionally, we hypothesized the convergence of the microbiome of adult piglets over time, an aspect we sought to explore in our Emperor plots.

To aid our analysis, a color-coded legend was designed: yellow for **diarrhea**, blue for **no diarrhea**, and red for **ematic** conditions. Gender differentiation was also considered, representing male samples with cones and female samples with dots. The **is sow** category was further emphasized with larger labels.

The Bray-Curtis results aligned closely with our hypotheses, indicating a clustered grouping of initial **ematic** samples, followed by a dispersal of **diarrhea** and **no diarrhea** samples. Detailed examination of the Bray-Curtis plots revealed a distinct trajectory from an **ematic** cluster to **no diarrhea** samples, with **is sow** samples forming a distinctly separate cluster towards the trajectorys end.

A temporal differentiation in the microbiota was observed, particularly with **ematic** samples exclusive to time T0, suggesting a gradual alignment with the **is sow** microbiome cluster. This supports the predicted microbiota evolution over time towards a mature **is sow** state.

In the Jaccard Emperor plots, which highlight dissimilarities between groups, a clear bifurcation was observed between the **is sow** and **non-sow** groups. Additionally, within the **non-sow** group, the distinction between **ematic** and non-ematic samples was evident, while **diarrhea** samples remained interspersed within the **no diarrhea** group.

However, for the ASV level unweighted UniFrac metric, the analysis did not yield clearly defined clusters, resulting in rather sparse groupings. Conversely, the weighted UniFrac metric demonstrated trends similar to those observed in the Bray-Curtis plots, reinforcing our observations regarding the microbial community structures and their distinguishing features.

13 ANCOM for Differential Abundance Analysis

The Analysis of Composition of Microbiomes (ANCOM) is a statistical approach tailored for differential abundance analysis in microbiome studies. This method is particularly adept at handling the inherent complexities of compositional data, a common characteristic of microbiome datasets, the following information are taken from the original ANCOM paper suggested by Qiime2 documentation [11].

13.1 Compositional Data Challenges

Microbiome datasets are typically compositional, meaning that they represent relative abundances of microbial taxa in a fixed total, all this idea can be found at [14]. This attribute leads to two primary challenges:

1. **Relative Scale:** The data only provide information about the relative abundances of taxa, not their absolute quantities. Therefore, the increase or decrease in a taxon's abundance is interpretable only in the context of other taxa within the same sample.

2. **Sum Constraint:** The total sum of taxa abundances in each sample is constant. This constraint induces a negative correlation among taxa, complicating standard statistical analyses which assume independence of features.

To check if the data is compositional, we can use a simple mathematical approach based on the fundamental properties of compositional data. Compositional data, by definition, consists of parts of a whole where the information is contained in the ratios of the parts, not in the absolute values.

Given a dataset with n samples and m taxa (or features), where each sample i has a vector of taxa abundances $X_i = [x_1, x_2, ..., x_m]$, you can check for the compositional nature as follows:

- **Sum of Feature**: Calculate the sum of all features in each sample:

$$S_i = \sum_{j=1}^{m} x_{ij}$$

  where:
  x_{ij} is the abundance of taxon j in sample i.

- **Check for Constant Sum**: Determine if this sum S_i is constant (or nearly constant) across all samples, citing Qiime2 the standatd deviation should be under the 25% to let ANCOM statistical analysis work correctly.

Consider a simple example with three taxa (A, B, C) in a single sample. Let their absolute abundances be denoted as follows:

$$A = 20, \tag{9}$$
$$B = 30, \tag{10}$$
$$C = 50. \tag{11}$$

The total abundance T in the sample is the sum of the abundances of A, B, and C:

$$T = A + B + C \tag{12}$$
$$= 100. \tag{13}$$

We can then compute the relative abundances of each taxon as a fraction of the total abundance:

$$A_r = \frac{A}{T} = \frac{20}{100} = 0.2, \tag{14}$$
$$B_r = \frac{B}{T} = \frac{30}{100} = 0.3, \tag{15}$$
$$C_r = \frac{C}{T} = \frac{50}{100} = 0.5. \tag{16}$$

In this scenario, the relative nature of A_r, B_r, and C_r implies that an increase in the abundance of one taxon results in a proportional decrease in the relative abundance of the others, highlighting the compositional nature of the data.

The table below presents the standard deviation and its percentage of the mean for various categories and tables. Notably, both the **Taxa Filt** and **Norm GMPR** tables for ASVs meet the criterion of having a standard deviation less than 25% of the mean. This observation is particularly significant for the application of ANCOM. In this specific case, the **Norm GMPR** table was selected for its compliance with this parameter. The detailed data are as follows:

Category	Table	Standard Deviation	Std Dev as % of Mean
ASV	Taxa Filt	8647.90	26.52%
ASV	Norm GMPR	8115.63	16.18%
ASV	Norm CLR	253.41	-1994.00%
Genus	Taxa Filt	8647.90	26.52%
Genus	Norm GMPR	8683.07	16.10%
Genus	Norm CLR	46.26	-2636.07%
Species	Taxa Filt	8647.90	26.52%
Species	Norm GMPR	8567.43	16.75%
Species	Norm CLR	51.32	-2477.81%

Table 42: Standard Deviation and Percentage of Mean for Various Categories and Tables

13.2 W Statistics in ANCOM for Multi-Label Metadata

One of the key outputs of ANCOM is the single W statistic for each feature (taxon), which is instrumental in identifying differentially abundant taxa across various groups. In datasets with multiple conditions, such as samples classified under different health states or environmental conditions, ANCOM provides a single W statistic for each taxon. This statistic is not a measure of absolute or even relative abundance, but rather focuses on the ratios of abundances of taxa and how these ratios vary significantly across different conditions.

13.2.1 Calculation of the W Statistic

The W statistic for a taxon is determined through the following steps:

1. **Ratio Formation**: For each taxon, calculate the ratio of its abundance to that of every other taxon across all samples.
2. **Statistical Testing**: Conduct statistical tests to determine whether these ratios are significantly different across the various conditions or labels in the study.
3. **Counting Significant Differences**: The W statistic is the count of the number of times the abundance ratios of a particular taxon are significantly different across groups.

13.2.2 Example In diarrhea column

In our dataset, we have samples classified under three different conditions related to diarrhea metadata: *yes*, *no*, and *ematic*.

- For a specific taxon, say Taxon A, we calculate its abundance ratio with every other taxon (e.g., Taxon B, Taxon C) for each of the three conditions.
- We then perform statistical tests to determine if the ratios of Taxon A to Taxon B and Taxon A to Taxon C are significantly different across the y, n, and e conditions.
- If, for example, the ratio of Taxon A to Taxon B is significantly different in 4 out of 10 comparisons across conditions, and Taxon A to Taxon C is significantly different in 6 out of 10 comparisons, the W statistic for Taxon A would be:

$$4+6=10$$

.

13.3 Comparative Analysis of multi label and pairwise W Statistics

The single W statistic used in ANCOM offers a comprehensive measure of a taxon's differential abundance across multiple conditions. This approach is notably advantageous for identifying taxa that exhibit consistent patterns of change across diverse conditions, extending beyond the limitations of simple pairwise comparisons.

To evaluate this hypothesis, a pairwise test using the ANCOM tool in QIIME2 was conducted. The goal was to

determine if the pairwise approach identifies different taxa compared to the multilabel approach. Our trials indicated that the difference between the pairwise and multilabel analyses was minimal. The pairwise method identified slightly more differentially abundant taxa (1 to 3 maximum) than the multilabel approach. However, the taxa identified through the pairwise method generally had significantly lower W statistics than those identified in the multilabel analysis.

These observations lead to the conclusion that employing both multilabel and pairwise approaches does not result in significant changes in the identification of differentially abundant taxa. Our subsequent step was to interpret the ANCOM results for the multilabel approach. In the following chapter, we illustrate how we assessed the magnitude of differential abundance using the log-ratio method in the multivariate context.

13.3.1 Calculation of Significant Log Ratios for multi label approach

We previously analyzed the ANCOM results for diarrhea, particularly focusing on the W statistics rank. However, it remained unclear which among the three features (n, e, y) was most significant between each others, hence we have to find a way to spot order of difference between each category.

- The data for each category (n, e, y) are presented in percentiles as follows:

$$Percentiles = [0.0, 0.25, 0.5, 0.75, 1.0]$$

- For each category (e,n,y), we calculate the sum as follows:

$$x_{label} = \sum_{i=1}^{N} x_i$$

  where x_i represents the quantile value, i the quantile and $N = 5$ is the total number of quantiles. We denote x_{label} as the sum for each label in our case (e,n,y).

- Following the computation of these sums, we calculate the ratio using the formula:

$$ratio = \log\left(\frac{\text{label}_i}{\text{label}_j}\right)$$

  In this formula, i and j represent two different labels. This method allows us to quantitatively determine the significance of the differences between the categories.

- In the following example n,e,y represent the sum of value for the percentile:

$$ratio\ \frac{n}{e} = -5.54$$

$$ratio\ \frac{n}{y} = +0.45$$

  In the that case we know that we have to focus on the difference between n and e, because the absolute value of their ration is higher compared to the other.

13.4 Differential Abbundances ANCOM Results

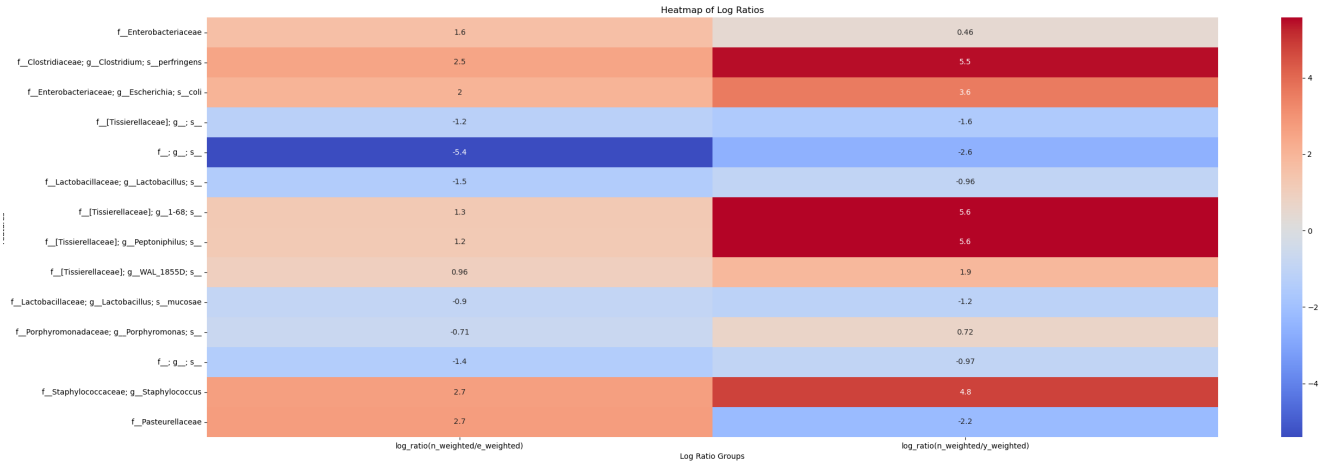


Figure 73: **ASV ANCOM Log Ratio Analysis:** This figure presents a detailed analysis of the log ratios, with the x-axis showing two calculated log ratios: the log ratio of 'n' over 'e' and the log ratio of 'n' over 'y'. On the left side, taxa are ranked according to the W statistics derived from the ANCOM analysis. The right side features a color bar, which aids in interpreting the data. The color spectrum ranges from intense red, indicating highly positive values, to intense blue, denoting highly negative values. This color coding provides an intuitive understanding of the log ratio magnitudes and their respective directions.

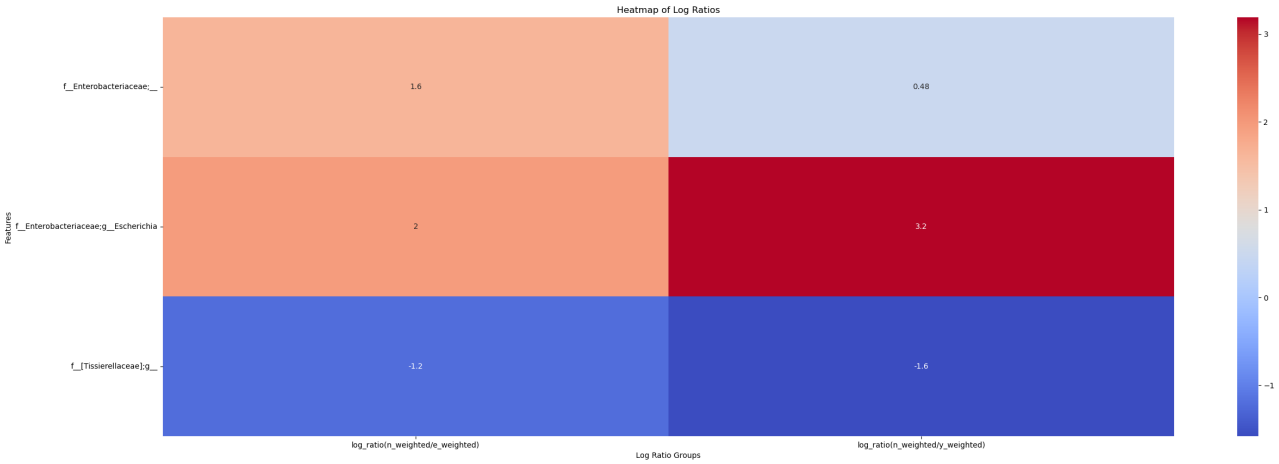


Figure 74: **Genus ANCOM Log Ratio Analysis:** This figure presents a detailed analysis of the log ratios, with the x-axis showing two calculated log ratios: the log ratio of 'n' over 'e' and the log ratio of 'n' over 'y'. On the left side, taxa are ranked according to the W statistics derived from the ANCOM analysis. The right side features a color bar, which aids in interpreting the data. The color spectrum ranges from intense red, indicating highly positive values, to intense blue, denoting highly negative values. This color coding provides an intuitive understanding of the log ratio magnitudes and their respective directions.

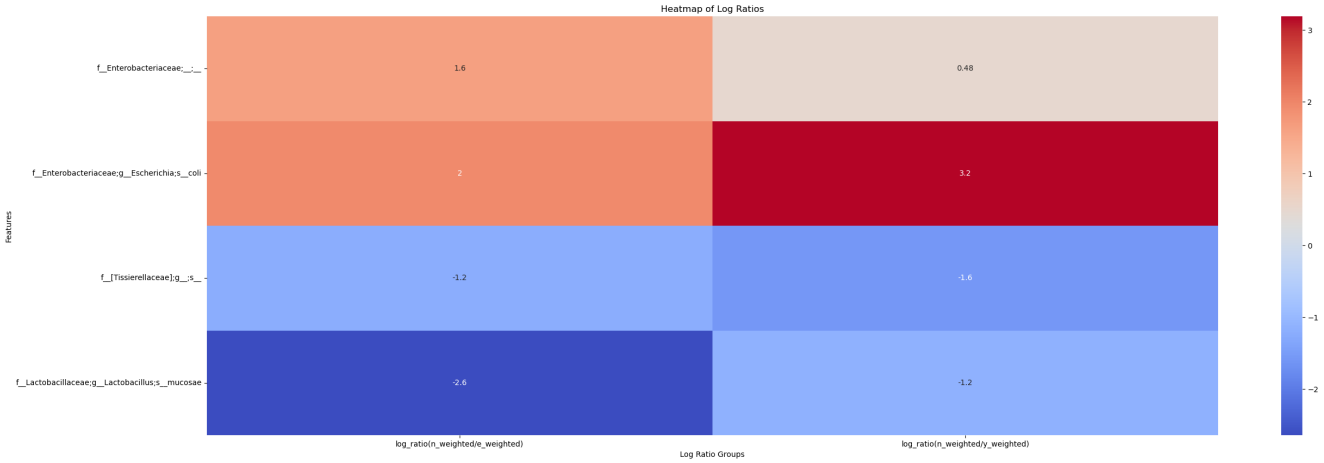


Figure 75: **Species ANCOM Log Ratio Analysis:** This figure presents a detailed analysis of the log ratios, with the x-axis showing two calculated log ratios: the log ratio of 'n' over 'e' and the log ratio of 'n' over 'y'. On the left side, taxa are ranked according to the W statistics derived from the ANCOM analysis. The right side features a color bar, which aids in interpreting the data. The color spectrum ranges from intense red, indicating highly positive values, to intense blue, denoting highly negative values. This color coding provides an intuitive understanding of the log ratio magnitudes and their respective directions.

The images are generated using the ANCOM **significant-log-ratio.tsv** file, which organizes different taxa based on their W statistics (13.2). In this scheme, taxa are arranged in descending order according to their W statistics, with the first taxon displaying the highest statistic. This ordering provides essential insights into the results of Differential Abundance (DA) analysis. Another sophisticated aspect, elaborated in this ANCOM chapter, involves the assessment of the magnitude of differences among the three designated categories: **y, e, n**. These disparities are visually represented in the heatmap, highlighting the log-ratio values (13.3.1).

Interpreting these heatmaps is intuitive: a larger absolute value in a specific row for a taxon indicates a notable difference between two categories. The sign of this value, be it positive or negative, signifies the direction of the disparity. Negative values, depicted in blue on the heatmap, denote a ratio between 0 and 1, suggesting that the denominator is more predominant than the numerator. Conversely, positive values indicate that the numerator is more prevalent.

Our analysis is not driven by a biological objective but rather focuses on the results, aiming to validate our analytical methodology. At this point, we refrain from delving into detailed interpretations of these data, concentrating instead on affirming the robustness and validity of our analytical approach.

13.5 Conclusion on ANCOM

ANCOM's approach to handling compositional data and its provision of a single W statistic per feature makes it a robust tool for differential abundance analysis in microbiome studies. However, the suitability of ANCOM should be carefully considered in the context of the dataset's characteristics, particularly the number of labels in the metadata and the expected proportion of differentially abundant features.

To address the challenges associated with compositional data and enhance the efficacy of ANCOM, especially when using GMPR-normalized count tables, it is advisable to implement a *more stringent filtering process prior to analysis*. This step involves the careful removal of taxa with low counts or high variability across samples, which can reduce noise and improve the reliability of differential abundance detection. By refining the dataset through such preprocessing, researchers can better align the data with the assumptions underlying ANCOM, thereby ensuring more accurate and interpretable results.

14 MaAsLin2 Analysis

MaAsLin2, short for Multivariable Association in Population-scale Meta-omics Studies, is a tool designed for multivariable association testing in microbiome profiles, including taxonomic, functional, or metabolomic data. It offers a comprehensive system that integrates preprocessing, normalization, transformation, and statistical modeling tailored for the complexities of microbial multi-omics data, like compositionality, overdispersion, and high-dimensionality the datailed explained above are taken from the original MaAsLin2 paper [10].

Furthermore MaAsLin2 as ANCOM require the data to be compositional to have better results on the final analysis more detail on compositional data can be found in ANCOM section.

14.1 Mathematical Framework

The core of MaAsLin2's methodology lies in its use of generalized linear and mixed models. These models are adaptable for a wide range of epidemiological study designs, including cross-sectional and longitudinal studies, and for various data types such as counts and relative abundances. MaAsLin2 uniquely manages repeated measurements and multiple covariates, crucial for maintaining statistical power and accuracy in complex microbiome data analysis

MaAsLin2 contrasts with traditional methods that focused on overall associations between microbiome structures. Instead, it provides feature-level inference, crucial for high-resolution characterization of microbial associations. This involves a detailed adjustment process for each feature-metadata pair, accommodating complex study designs and overcoming limitations of early methods that did not fully account for the intricacies of microbiome data.

14.2 Balancing Fixed and Random Effects

The analysis was structured to accommodate both fixed and random effects, addressing the inherent variability characteristic of microbiome studies. The **fixed effects** in our model corresponded to known covariates of interest, including sex, sow status, and temporal conditions. In contrast, **random effects** were incorporated to account for unexplained variability within the data. This approach facilitated a more detailed understanding of the associations of microbial characteristics. Specifically, we chose to focus on the case of diarrhea.

NOTE: A significant issue encountered during the use of the MaAsLin2 package was related to memory usage. Applying the method with more than two fixed effects resulted in substantial memory consumption, leading to memory errors in the docker container. This problem likely stemmed from the limited resources available on my computer and was particularly noticeable when processing the amplicon sequence variant (ASV) table.

Our analysis also underscored the importance of choosing the appropriate *reference level* in MaAsLin2. Different reference levels, such as:

$$\text{reference} = (\text{diarrhea, n}) \tag{17}$$

compared to:

$$\text{reference} = (\text{diarrhea, y}) \tag{18}$$

led to distinct results. Modifying the reference level influenced the **q-values** and consequently altered the ranking of the significance of the features. This observation highlighted the tool's **sensitivity** to reference level specifications as well as to the configuration of all other chosen random and fixed effects.

These findings emphasize the need for careful and accurate preliminary studies before selecting these values. The choice of reference level, in conjunction with the configuration of fixed and random effects, plays a pivotal role in the interpretation and reliability of MaAsLin2 analyses. In our case, we used the most common reference level, specifically the **no diarrhea** label.

14.3 Differential Abbundances MaAsLin2 Results

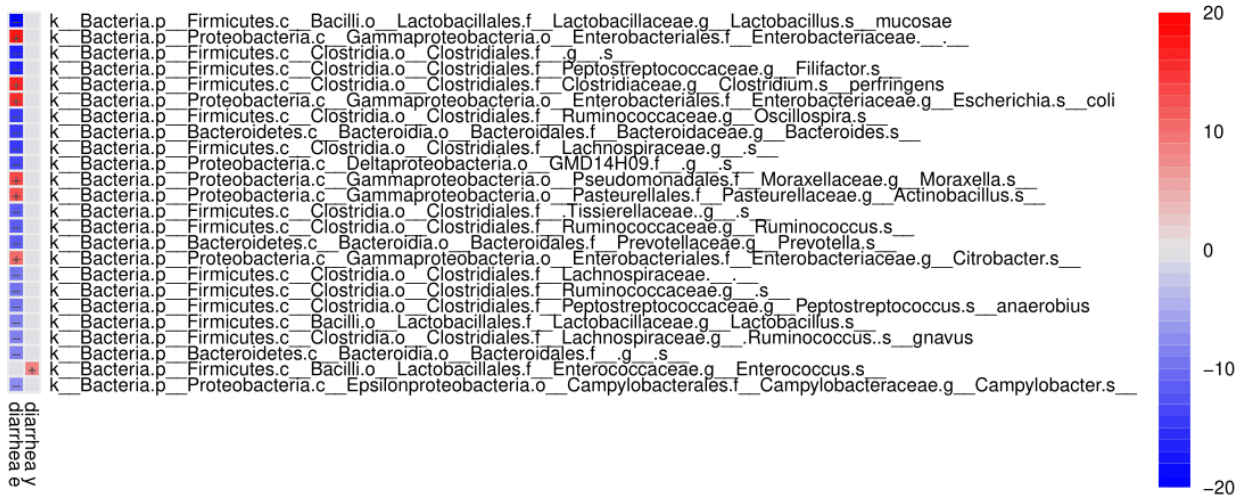


Figure 76: **Correlation Heatmap of ASVs** This figure presents a heatmap illustrating the correlations based on the reference value of *no diarrhea*. On the left side, the correlation is visually represented through a color gradient ranging from red to blue, with white indicating neutral correlation. Red signifies a positive correlation, blue denotes a negative correlation, and white indicates the absence of correlation for the other two label, e and y. In the middle of the heatmap are the names of the taxa identified as differentially abundant. On the right side, a color bar is provided to interpret the varying levels of correlation.

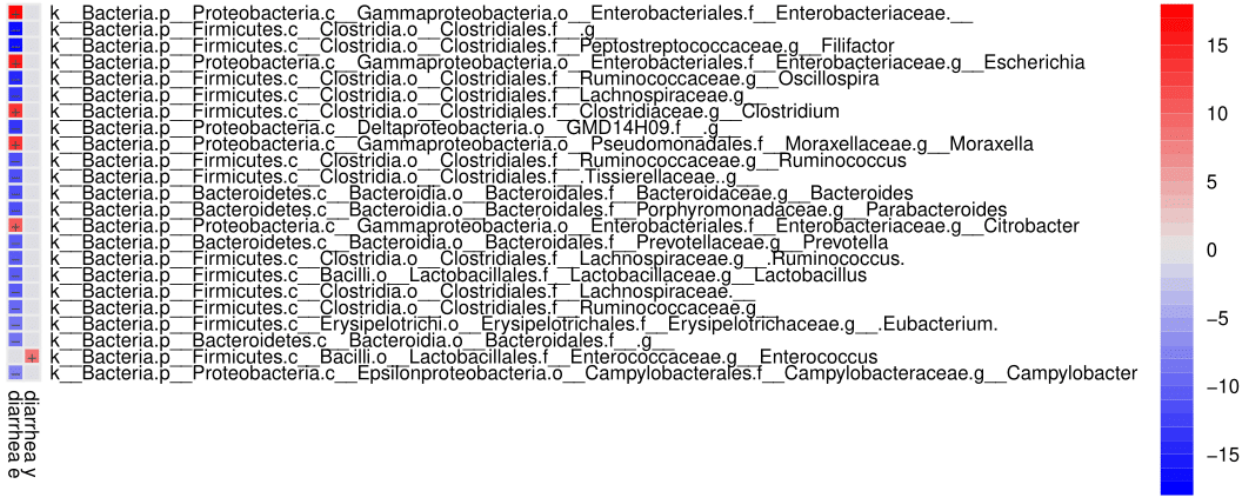


Figure 77: **Correlation Heatmap of Genus** This figure presents a heatmap illustrating the correlations based on the reference value of *no diarrhea*. On the left side, the correlation is visually represented through a color gradient ranging from red to blue, with white indicating neutral correlation. Red signifies a positive correlation, blue denotes a negative correlation, and white indicates the absence of correlation for the other two label, e and y. In the middle of the heatmap are the names of the taxa identified as differentially abundant. On the right side, a color bar is provided to interpret the varying levels of correlation.

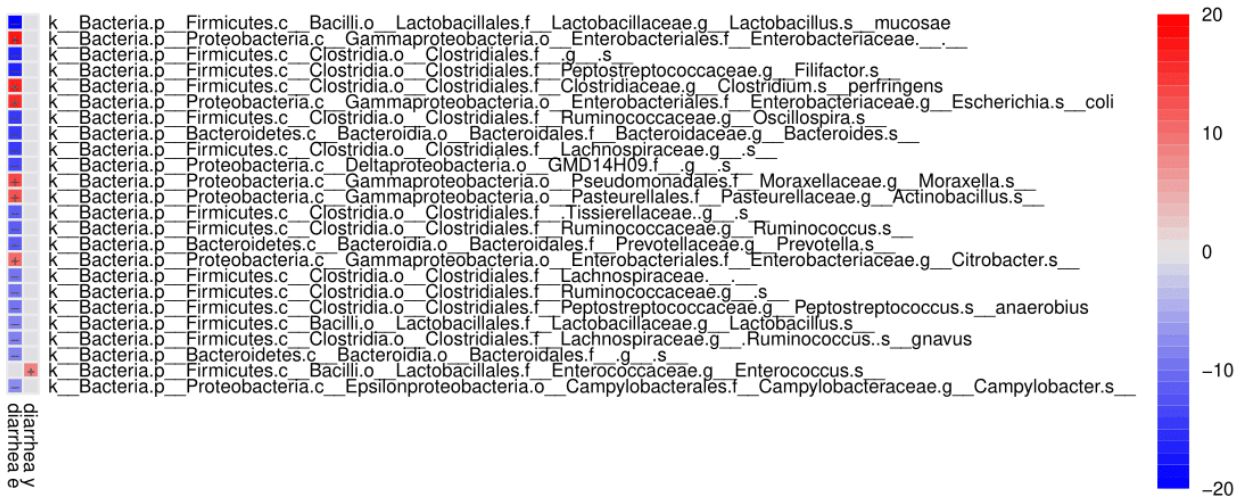


Figure 78: **Correlation Heatmap of Species** This figure presents a heatmap illustrating the correlations based on the reference value of *no diarrhea*. On the left side, the correlation is visually represented through a color gradient ranging from red to blue, with white indicating neutral correlation. Red signifies a positive correlation, blue denotes a negative correlation, and white indicates the absence of correlation for the other two label, e and y. In the middle of the heatmap are the names of the taxa identified as differentially abundant. On the right side, a color bar is provided to interpret the varying levels of correlation.

Before proceeding, it is important to understand how the correlation is plotted. The formula is:

$$\text{Metric} = \text{sign}(\text{Coefficient}) \times -\log_{10}(\text{q-value}) \tag{19}$$

Where: The *Coefficients* are derived from the results of the MaAsLin2 table, and the *q-values* are obtained from the results of multilabel analyses. It is important to note that in the case of pairwise comparisons, we are unable to plot this correlation using the same approach.

The table provides a broad range of Differential Abundances (DA) in the case of n vs e. Consequently, most of the colored output is concentrated in the e column, whereas there are fewer values indicating correlation between e and y.

The analysis of the three correlation matrices ASV, Genus, and Species reveals a consistent pattern: taxa identified as Differentially Abundant are remarkably similar across these different taxonomic levels. This consistency underscores the stability of MaAsLin2 across various analyses. A notable example illustrating this stability is the consistent correlation of *Lactobacillus* across all three tables, where it is uniquely correlated with both the n and y categories. This recurrent observation across different taxonomic resolutions not only confirms the robustness of MaAsLin2 but also highlights the potential biological relevance of *Lactobacillus* in the study context. This finding aligns with the underlying biological assumption that significant microbial features should manifest consistently across various levels of taxonomic classification, further reinforcing the validity of our analytical approach.

This is not the only way to view the data; in fact, MaAsLin also generates bar plots for the most significant taxa and produces a file containing all the taxon results. These files can be found in the folder where our script is located, allowing for a more in-depth analysis.

14.4 Conclusion

MaAsLin2 represents a significant advancement in the field of microbiome epidemiology, with its sophisticated modeling capabilities and the ability to handle complex study designs and data types. This makes it an invaluable tool for uncovering intricate microbial associations. The insights gained from our study using MaAsLin2 are substantial. However, a notable aspect of MaAsLin2's application is its **sensitivity** due to the variability in the choice of analysis parameters.

Additionally, the implementation of random effects in MaAsLin2 can lead to different conclusions than its fixed effects

counterpart, underscoring the importance of careful model selection in analysis. The tool's sensitivity to parameter choices emphasizes the critical need for thorough preliminary analysis and judicious selection of model parameters to ensure robust and meaningful results from MaAsLin2 analyses.

15 Intersecting Findings from MaAsLin2 and ANCOM in Differential Abundance Analysis

In our study, we sought to validate the consistency of differential abundance (DA) findings by comparing results obtained from two statistical methods: MaAsLin2 and ANCOM. This comparison is essential to reinforce the robustness of our observations about DA features across three different levels: ASV, genus, and species.

To facilitate this comparison, we developed a script that assigns a ranking to DA features. These rankings are based on the W-statistics from ANCOM and the q-values from MaAsLin2. Implementing this ranking process was a key step in assessing the level of concordance between the two methods.

NOTE: Our approach does not constitute a benchmarking exercise comparing the performance of the two methods. Rather, it is a comparison based on the biological results. Specifically, we anticipate that a significant number of DAs should be consistent between the two methods, underscoring the reliability of our findings.

Statistic	**Value**
Total Entries	27
Number of Entries Identified by ANCOM	14
Number of Entries Identified by MaASLin2	25
Entries with Intersection	11
Percentage of Intersection	48.1%
Percentage of No Intersection	51.9%
Intersection Percentage in ANCOM Entries	78.57%
Intersection Percentage in MaASLin2 Entries	44%

Table 43: **Summary of Biological Intersections for ASV**: Identified by ANCOM and MaASLin2 Tools. This table displays the total number of taxa identified, the number identified exclusively by each tool, and the percentages of intersections between them.

Statistic	**Value**
Total Entries	23
Number of Entries Identified by ANCOM	3
Number of Entries Identified by MaASLin2	23
Entries with Intersection	3
Percentage of Intersection	13.04%
Percentage of No Intersection	86.96%
Intersection Percentage in ANCOM Entries	100%
Intersection Percentage in MaASLin2 Entries	13.04%

Table 44: **Summary of Biological Intersections for Genus**: Identified by ANCOM and MaASLin2 Tools. The table displays the total number of taxa identified, the number identified exclusively by each tool, and the percentages of intersections between them.

Statistic	Value
Total Entries	24
Number of Entries Identified by ANCOM	4
Number of Entries Identified by MaASLin2	24
Entries with Intersection	4
Percentage of Intersection	16.67%
Percentage of No Intersection	83.33%
Intersection Percentage in ANCOM Entries	100%
Intersection Percentage in MaASLin2 Entries	16.67%

Table 45: **Summary of Biological Intersections for Species**: Identified by ANCOM and MaASLin2 Tools. The table displays the total number of taxa identified, the number identified exclusively by each tool, and the percentages of intersections between them.

Based on the data presented in the tables and the analysis of the intersections between ANCOM and MaASLin2 results, a clear conclusion can be drawn regarding the comparative effectiveness and insight provided by these tools in microbial abundance studies.

The tables reveal that MaASLin2 consistently identifies a greater number of entries compared to ANCOM across various taxonomic levels, indicating its enhanced capability in detecting abundant taxa. Notably, a substantial portion of the taxa identified by ANCOM also appear in the MaASLin2 results, as evidenced by the high percentage of intersections. This overlapping demonstrates that both methods are valid and reliable for analysis, yet it is MaASLin2 that exhibits a broader scope in identifying microbial abundances.

Furthermore, the fact that stricter q-value criteria for ANCOM do not lead to significant changes in the intersection percentages suggests that the inherent ranking of taxa is mostly preserved across both methods (This conclusion was lead to different try of intersection using different range limit for q-value in this case the q-value is ≤ 0.005). However, the superior ability of MaASLin2 to capture a more extensive range of taxa, potentially leading to a more comprehensive and nuanced analysis, is evident.

In conclusion, while ANCOM proves to be a robust tool for microbial abundance analysis, MaASLin2 demonstrates a heightened capacity to uncover a wider array of insights within the data. This characteristic makes MaASLin2 particularly valuable for in-depth studies where capturing the full spectrum of microbial diversity is crucial.

16 Longitudinal Analysis

Longitudinal analysis provides critical insights into how variables change over time. Unlike cross-sectional studies that offer a snapshot at a single time point, longitudinal studies track the same subjects across multiple time points. This approach allows us to observe the temporal dynamics and causal relationships within our data, offering a deeper understanding of the processes under investigation.

In our experiment, we meticulously prepared our metadata to support robust longitudinal analysis. We transformed time-related categorical data into numerical form, converting $T0, T1, T2 \longrightarrow 0, 1, 2$ for more effective temporal analysis. Additionally, we restructured the *serial* column to assign unique IDs to each piglet, enabling us to accurately track individual changes over time. This meticulous preparation extended to incorporating alpha diversity metrics (*shannon*, *observed features*, *evenness vector pielou*, *faith pd*) from prior analyses into our metadata, thus enriching our dataset for more detailed analysis.

16.1 Linear Mixed Effects Models

To delve into the complex, dependent nature of our longitudinal data, we implemented Linear Mixed Effects (LME) Models. This advanced statistical tool was crucial for understanding the impact of both fixed effects (our experimental conditions) and random effects (inherent subject variability). Particularly well-suited for our repeated-measures data, LME models accounted for correlations between these measures, allowing us to meticulously dissect the influence of predictor variables on the response variable and understand the evolution of our experimental data over time.

The focus of our longitudinal analysis was on ASV, genus, and species. We specifically analyzed how variables changed over time between two distinct groups: is_sow and non-is_sow. We excluded diarrhea from our current analysis for two reasons. Firstly, ematic samples are exclusively present at time T0, limiting the scope for temporal comparison. Secondly, insights from previous alpha and beta diversity analyses indicated that a temporal difference analysis between is_sow and non-is_sow would be more meaningful. We hypothesized that the is_sow subjects would exhibit fewer changes in features over time compared to the non-is_sow subjects.

For this analysis, we used the formula $alpha \sim time \times is\_sow$. While it was possible to add more correlations, we were limited by memory usage constraints. The results of the Linear Mixed Effects Model for alpha diversity metrics, analyzing the impact of time and the 'Is Sow' condition, are summarized in the following tables. Asterisks (*) indicate statistically significant values. The 'Coeff' column represents the estimated effect size of each parameter. 'StdE' stands for Standard Error of the estimate, indicating the variability. 'Z-Val' is the Z-score used to determine statistical significance. 'P-Val' is the probability of observing the effect size if the null hypothesis is true. The '95% Confidence' column provides the interval within which the true effect size is likely to fall.

16.1.1 ASV LME results

The model results showed significant interactions for several metrics. Notably, in the case of Observed Features, there is a pronounced interaction effect between time and the is_sow condition. This finding supports our hypothesis that is_sow subjects would display distinct temporal patterns compared to non-is_sow subjects. In particular, the is_sow group showed more significant changes over time, contrasting with our initial expectation of fewer changes in this group, this is quite strange and need further investigation.

Alpha Metric	Parameter	Coeff	StdE	Z-Val	P-Val	95% Conf-Int
Observed Features	Baseline	301.739	8.650	34.883	<0.001*	[284.785, 318.692]
	Is Sow (Yes)	32.261	17.300	1.865	0.062	[-1.646, 66.168]
	Time	2.817	6.700	0.420	0.674	[-10.316, 15.949]
	Time × Is Sow (Yes)	113.183	13.400	8.446	<0.001*	[86.919, 139.448]
Faith PD	Baseline	23.371	0.498	46.912	<0.001*	[22.395, 24.348]
	Is Sow (Yes)	4.786	0.996	4.804	<0.001*	[2.834, 6.739]
	Time	-0.411	0.386	-1.064	0.287	[-1.167, 0.346]
	Time × Is Sow (Yes)	4.224	0.772	5.473	<0.001*	[2.711, 5.737]
Pielou Evenness	Baseline	0.624	0.011	59.155	<0.001*	[0.603, 0.645]
	Is Sow (Yes)	0.158	0.021	7.489	<0.001*	[0.117, 0.199]
	Time	0.073	0.008	8.994	<0.001*	[0.057, 0.089]
	Time × Is Sow (Yes)	-0.020	0.016	-1.203	0.229	[-0.052, 0.012]
Shannon Entropy	Baseline	5.136	0.096	53.398	<0.001*	[4.947, 5.324]
	Is Sow (Yes)	1.416	0.192	7.361	<0.001*	[1.039, 1.793]
	Time	0.617	0.075	8.288	<0.001*	[0.471, 0.763]
	Time × Is Sow (Yes)	0.164	0.149	1.101	0.271	[-0.128, 0.456]

Table 46: ASV Level Linear Mixed Effects Model results for different diversity metrics, analyzing the impact of time and the 'Is Sow' condition. Asterisks (*) indicate statistically significant values.

16.1.2 Genus LME Results

The genus level analysis using Linear Mixed Effects Models yielded notable findings. Similar to the ASV results, significant interactions were observed, particularly in the case of Observed Features. A significant interaction between time and the is_sow condition was found, aligning with our hypothesis that the is_sow group would display distinct temporal patterns compared to the non-is_sow group.

In terms of Pielou Evenness, we observed a consistent increase over time in the is_sow group, indicated by the positive time coefficient. This suggests a dynamic shift in species distribution within this group. For Shannon Entropy, the is_sow group showed a notable increase in microbial diversity, reflecting a more complex microbial community over time.

Alpha Metric	Parameter	Coeff	StdE	Z-Val	P-Val	95% Conf-Int
Observed Features	Baseline	90.000	0.798	112.818	<0.001*	[88.436, 91.564]
	Is Sow (Yes)	11.750	1.595	7.365	<0.001*	[8.623, 14.877]
	Time	-0.367	0.618	-0.593	0.553	[-1.578, 0.844]
	Time × Is Sow (Yes)	7.017	1.236	5.678	<0.001*	[4.594, 9.439]
Pielou Evenness	Baseline	0.590	0.009	65.881	<0.001*	[0.572, 0.608]
	Is Sow (Yes)	0.145	0.018	8.106	<0.001*	[0.110, 0.180]
	Time	0.054	0.007	7.881	<0.001*	[0.041, 0.068]
	Time × Is Sow (Yes)	-0.032	0.014	-2.338	0.019*	[-0.059, -0.005]
Shannon Entropy	Baseline	3.828	0.059	65.335	<0.001*	[3.713, 3.943]
	Is Sow (Yes)	1.073	0.117	9.158	<0.001*	[0.844, 1.303]
	Time	0.350	0.045	7.720	<0.001*	[0.261, 0.439]
	Time × Is Sow (Yes)	-0.134	0.091	-1.478	0.139	[-0.312, 0.044]

Table 47: Genus Level Linear Mixed Effects Model results for different diversity metrics, analyzing the impact of time and the 'Is Sow' condition. Asterisks (*) indicate statistically significant values.

16.1.3 Species LME Results

At the species level, the LME model results indicate notable patterns, particularly in the context of the "is sow" and time interaction. The observed features show a statistically significant interaction between time and the "is sow" condition, with a marked increase in the number of features over time for the "is sow" group. This suggests that the microbial diversity within this group is not only distinct but also evolves more noticeably over time compared to the non-"is sow" group.

In terms of Pielou Evenness, a significant and positive time coefficient is observed in the "is sow" group, indicating an increase in evenness of species distribution over time. This could imply a more balanced and stable microbial community as time progresses.

For Shannon Entropy, the data reveal a significant increase in microbial diversity within the "is sow" group over time. This increase in entropy suggests a more complex and varied microbial ecosystem developing in this group, highlighting the dynamic nature of microbial communities and their susceptibility to temporal changes.

Diversity Metric	Parameter	Coeff	StdE	Z-Val	P-Val	95% Confidence
Observed Features	Baseline	120.300	0.945	127.285	<0.001*	[118.448, 122.152]
	Is Sow (Yes)	8.100	1.890	4.285	<0.001*	[4.395, 11.805]
	Time	-0.033	0.732	-0.046	0.964	[-1.467, 1.400]
	Time × Is Sow (Yes)	8.833	1.463	6.038	<0.001*	[5.966, 11.701]
Pielou Evenness	Baseline	0.600	0.009	64.153	<0.001*	[0.582, 0.618]
	Is Sow (Yes)	0.146	0.019	7.814	<0.001*	[0.110, 0.183]
	Time	0.067	0.007	9.204	<0.001*	[0.052, 0.081]
	Time × Is Sow (Yes)	-0.047	0.014	-3.243	0.001*	[-0.075, -0.019]
Shannon Entropy	Baseline	4.145	0.065	63.560	<0.001*	[4.017, 4.273]
	Is Sow (Yes)	1.081	0.130	8.289	<0.001*	[0.826, 1.337]
	Time	0.462	0.051	9.142	<0.001*	[0.363, 0.561]
	Time × Is Sow (Yes)	-0.252	0.101	-2.492	0.013	[-0.450, -0.054]

Table 48: Species Level Linear Mixed Effects Model results for different diversity metrics, analyzing the impact of time and the 'Is Sow' condition. Asterisks (*) indicate statistically significant values.

16.1.4 Conclusion

Our analysis using Linear Mixed Effects (LME) models revealed an unexpected trend: the "is sow" group exhibited more variation over time compared to the "non-sow" group. This finding prompts a multi-faceted discussion to explore potential underlying reasons.

A primary consideration is the baseline microbial diversity in the "is sow" group. A higher diversity at the outset

may lead to a microbiome that is more responsive or susceptible to environmental or internal factors, resulting in pronounced shifts over time. Furthermore, the variability observed might also stem from intrinsic biological differences between the groups or external factors like sampling methods, which can influence the observed changes.

The unique biological factors within the "is sow" group could also play a significant role. Differences in environmental exposures, dietary patterns, health statuses, or physiological changes could contribute to a more dynamic microbiome in this group. Additionally, the initial microbial composition in "is sow" samples might inherently be more unstable or prone to changes, leading to more observable shifts in the microbiome.

Statistical considerations must also be acknowledged. The modeling process itself, including the choice of the model, the specification of interactions, or the handling of random effects, could influence the outcomes and potentially present a skewed view of the microbial dynamics.

Unmeasured variables or factors that disproportionately affect the "is sow" group might also drive the higher variability observed in this group. These could include environmental, physiological, or other biological factors not accounted for in the study.

Another critical aspect is the role of emetic samples in the "non-sow" group, particularly at time T0. If these samples have a distinct and stable microbial profile, it could artificially reduce the observed variability in this group. This aspect might mask the true temporal dynamics of the microbiome in the "non-sow" samples.

In summary, while our study presents clear evidence of greater variability in the "is sow" group compared to the "non-sow" group, the reasons behind this pattern are complex and likely multifactorial. A deeper exploration into specific microbial taxa, environmental, and physiological factors would be essential to fully understand the microbial dynamics in these groups.

16.2 Volatility Analysis

In our study, Volatility Analysis played a crucial role in monitoring and visualizing the temporal fluctuations of key dependent variables, particularly alpha diversity metrics. Utilizing this method, we generated interactive line plots that vividly illustrated trends and patterns within our longitudinal dataset. These visualizations were invaluable for uncovering insights into the stability or variability of essential measures, such as microbial diversity, across different groups over the course of the experiment.

Notably, the focus of our analysis shifted to the hypothesis surrounding the is_sow column, examining its behavior and impact on microbial community dynamics. The interactive nature of Volatility Analysis allowed us to delve deeply into how these communities evolved over time under different conditions, including the is_sow factor.

This approach provided a nuanced understanding that challenged and refined our initial interpretations from the Linear Mixed Effects (LME) models. In light of these findings, we gained a more comprehensive perspective on the complex interplay between temporal factors and microbial diversity. The insights gleaned from this analysis did not merely confirm our previous results but offered a more intricate picture, emphasizing the dynamic nature of microbial ecosystems and their sensitivity to specific conditions like the is_sow status.

16.2.1 PLOT

16.3 Feature Volatility Analysis

We employed Feature Volatility Analysis from the q2-longitudinal plugin to identify key microbial features predictive of time-related changes. This allowed us to dynamically visualize and discern which features significantly contribute to variations over time or due to treatment. It enabled us to identify specific features that play influential roles in the state changes observed during our experiment. The data are plotted in **.qzv** file and able to been seen in the analisys for further details.

References

[1] Taxonomy - definition, classification & example, 2023.

[2] J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, January 1982.

[3] Evan Bolyen, J.R. Rideout, Matthew R. Dillon, et al. Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. *Nature Biotechnology*, 37:852–857, 2019.

[4] Benjamin J Callahan, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. Dada2: High-resolution sample inference from illumina amplicon data. *Nature Methods*, 13(7):581–583, 2016.

[5] Li Chen, J. Reeve, Lu Zhang, Shengbing Huang, Xuefeng Wang, and Jun Chen. A robust normalization method for zero-inflated count data in microbiome sequencing. *PeerJ*, 6:e4600, 2018.

[6] Davide Chicco, Niklas Tötsch, and Giuseppe Jurman. The matthews correlation coefficient (mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining*, 14(1), 2021.

[7] F. Finotello, E. Mastrorilli, and B. Di Camillo. Measuring the diversity of the human microbiota with targeted next-generation sequencing. *Briefings in Bioinformatics*, 18(4):723–735, 2016.

[8] Jun Hu, Lingli Chen, Yimei Tang, Chunlin Xie, Baoyang Xu, Min Shi, Wenyong Zheng, Shuyi Zhou, Xinkai Wang, Liu Liu, Yiqin Yan, Tao Yang, Yaorong Niu, Qiliang Hou, Xiaofan Xu, and Xianghua Yan. Standardized preparation for fecal microbiota transplantation in pigs. *Frontiers in Microbiology*, 9:1328, 2018.

[9] Rui Jiang, Weizhong Li, and Jing Li. mbimpute: an accurate and robust imputation method for microbiome data. *Genome Biology*, 22(192), 2021.

[10] H. Mallick, A. Rahnavard, L. J. McIver, S. Ma, Y. Zhang, L. H. Nguyen, T. L. Tickle, G. Weingart, B. Ren, E. H. Schwager, S. Chatterjee, K. N. Thompson, J. E. Wilkinson, A. Subramanian, Y. Lu, L. Waldron, J. N. Paulson, E. A. Franzosa, H. C. Bravo, and C. Huttenhower. Multivariable association discovery in population-scale meta-omics studies. *PLoS Computational Biology*, 17(11):e1009442, 2021.

[11] Siddhartha Mandal, William Van Treuren, Rob Knight, Merete Eggesbø, R. White, and Shyamal D. Peddada. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health and Disease*, 26:27663, 2015.

[12] László Orlóci. Revisions for the bray and curtis ordination. *Canadian Journal of Botany*, 52(7):1773–1776, 1974.

[13] Daniel Rojas-Valverde, J. Pino-Ortega, C. Gómez-Carmona, and Markel Rico-González. A systematic review of methods and criteria standard proposal for the use of principal component analysis in team's sports science. *International Journal of Environmental Research and Public Health*, 17(23):8712, 2020.

[14] Antoni Susin, Yiwen Wang, Kim-Anh Lê Cao, and M Luz Calle. Variable selection in microbiome compositional data analysis. *NAR Genomics and Bioinformatics*, 2(2):lqaa029, 05 2020.