Binf6210_Bioinformatics Software Tools

Assignment 5 – Pavani Addepalli

Student ID #: 1326277

December 9, 2024

(https://github.com/pavaniaddepalli2121/Assignment5_BINF6210)

Microbiome Analysis of the Distal Lumen and Proximal Mucosa in the Human Gut: Comparative Study of Microbial Diversity and Community Composition

1. Introduction

Context: Microbiome refers to microorganisms consisting of diverse and indispensable populations within ecosystems and organism health. Research on the gut microbiome, in particular its function in immunity, digestion, and general metabolic health, has grown in importance in the study of human health. A complex collection of bacteria, viruses, fungus, and other microorganisms that have close interactions with the host makes up the gut microbiome. It is essential to examine microbial communities in particular locations, including the proximal mucosa (PM) and distal lumen (DL), to identify their unique functions in both health and illness. Variations in microbial composition and function may result from the many environmental factors and microbial exposures that these sites experience. Thus, a thorough examination of these gut areas is necessary to comprehend microbial diversity and its effects on human health.

Focus: In microbiome investigations, methodological decisions are crucial, especially the filtering parameters applied during data processing. The observed biodiversity measures (such as the Shannon index) and phylogenetic clustering within microbial communities can be greatly impacted by these parameters, which include read threshold and sequence length cutoff. Strict filtering, for example, may eliminate noise and increase data accuracy, providing a more accurate image of microbial diversity and community structure. On the other hand, lax filtering may preserve artifacts, which could skew the true phylogenetic relationships and microbial diversity. Considering these factors, this study examines the effects of different filtering parameters on phylogenetic clustering and biodiversity indicators across DL and PM data.

Objective: This study aims to investigate how various filtering options affect phylogenetic clustering and biodiversity indicators (such as the Shannon index) in DL and PM data. The study will specifically investigate the following question: What effects do filtering parameters have on phylogenetic grouping and observed biodiversity in these different gastrointestinal regions? DADA2 will be used to evaluate 16S rRNA gene amplicon sequencing data for high-resolution taxonomic profiling in order to answer this question **(Callahan et al., 2016)**. The results will help improve the accuracy of microbial diversity investigations in clinical and ecological contexts by optimizing methodological decisions in microbiome research. Foundational publications on microbiome diversity **(Lozupone et al., 2011)**, DADA2 technique **(Callahan et al., 2016)**, and 16S marker gene analysis **(Caporaso et al., 2010)** are pertinent references for this investigation.

2. Description of Data Set

The sequences were downloaded from the NCBI Sequence Read Archive (SRA) using sample IDs

The 16S rRNA gene sequences were downloaded from the **NCBI** Sequence Read Archive (SRA)

The following samples were examined:

DNA sequences obtained from the distal lumen are contained in **DL_sequences.fasta** (Sample ID: SRR6288926).

DNA sequences from the proximal mucosa are represented by **PM_sequences.fasta** (Sample ID: SRR6288933).

reference.fasta: A well-selected dataset of reference sequences for taxonomic categorization and alignment that contains the V4 16S ribosomal subunit sequences of various bacterial genera.

Sequence-specific identifiers were used to retrieve the data, and paired-end reads were converted into combined FASTA and FASTQ files.

The following reasons make these datasets appropriate for this study:

The V4 region of the 16S rRNA gene, which strikes a compromise between resolution and taxonomic universality, is the gold standard for profiling microbial communities.

A comparison analysis is made possible by the paired datasets from various gut areas, which shed light on the geographic variability of microbes. Reference sequences support functional inference and taxonomy classification, which is in line with the study's goal of determining regional variations in the microbiome. Using strong processing pipelines and high-quality, curated data, our investigation seeks to identify ecological and functional differences between the DL and PM microbiomes, advancing our knowledge of gut microbial diversity and its implications for human health.

3. Code Section 1 – Data Acquisition, Exploration, Filtering, and Quality Control

Filtering and Trimming:

The imported. fastq files of sample 1 and sample 2 forward and reverse primers were filtered for low-quality readings and primers using the following procedure.

The downloaded data were processed and saved in the fastq.gz file format.

Compressing FASTA data to FASTQ.gz lowers storage requirements while converting them to FASTQ guarantees compatibility with microbiome analysis programs such as DADA2.

- DL_F.fastq.gz: Forward reads for the distal lumen (DL) sample
- DL_R.fastq.gz: Reverse reads for the distal lumen (DL) sample

- PM_F.fastq.gz: Forward reads for the proximal mucosa (PM) sample
- PM_R.fastq.gz: Reverse reads for the proximal mucosa (PM) sample
- DL_sequences.fastq: Combined sequences for the distal lumen (DL) sample
- PM_sequences.fastq: Combined sequences for the proximal mucosa (PM) sample
File Locations:

The required files are stored in the following directory:

C:\Users\drpav\OneDrive\Desktop\6410_COURSE FILES\microbiome

Data Filtering, and Quality Control :

Packages used
```
library(tidyverse)
conflicted::conflicts_prefer(dplyr::filter())
library(viridis)
# + scale_color/fill_viridis_c/d()
theme_set(theme_light())
#if (!requireNamespace("BiocManager", quietly = TRUE))
#install.packages("BiocManager")
#BiocManager::install(c("dada2", "phyloseq"))
library(dada2)
library(phyloseq)
library(ShortRead)
```

Load the sequences:

```
setwd ("C:/Users/drpav/OneDrive/Desktop/6410_COURSE FILES/microbiome")
fnF1 <- "DL_F.fastq.gz"
fnR1 <- "DL_R.fastq.gz"
# Know the size of F1 and R1 to optimize memory usage, manage computational resources, and ensure efficient execution of data analysis tasks.
sizeF1 <- object.size(fnF1)
print(size) # 120 bytes
sizeR1 <- object.size(fnR1)
print(size) # 120 bytes
```

Temporary file paths created using tempfile() to store filtered FASTQ data for intermediate use.

```
filtF1 <- tempfile(fileext = ".fastq.gz")
filtR1 <- tempfile(fileext = ".fastq.gz")
```

Visualize the quality metrics for both forward and reverse sequencing reads for quality control purposes.

```
plotQualityProfile(fnF1)  # Visualize forward quality profile
plotQualityProfile(fnR1)  # Visualize reverse quality profile
```

Visualized plots display the mean quality scores along sequencing reads, helping identify low-quality regions for trimming.

```
# Get the summaries
forward_quality_summary <- summary(forward_quality_plot)
reverse_quality_summary <- summary(reverse_quality_plot)
```

Filtering Choices:

```
# Setting trimming and filtering parameters based on quality plots

filterAndTrim(
  fwd = fnF1, filt = filtF1, rev = fnR1, filt.rev = filtR1,
  trimLeft = 10,  # Trim first 10 bases to remove low-quality base calls
  truncLen = c(240, 200),  # Truncate reads at base positions where quality drops
  maxN = 0,  # No N's allowed in reads
  maxEE = 2,  # Maximum expected error allowed in a read
  compress = TRUE,
  verbose = TRUE
)
# Read in 46506 paired-sequences, output 41667 (89.6%) filtered paired-sequences.

# Summarize the number of reads and their lengths in the filtered files
forward_summary <- summary(filtF1)
reverse_summary <- summary(filtR1)

# Print summaries to the console
print(forward_summary)
# Length     Class        Mode
#  1       character    character
print(reverse_summary)
# Length     Class        Mode
#  1       character    character

# Plotting quality profiles for comparison
plotQualityProfile(fnF1)  # Forward before filtering
plotQualityProfile(filtF1)  # Forward after filtering
plotQualityProfile(fnR1)  # Reverse before filtering
plotQualityProfile(filtR1)  # Reverse after filtering
```

Quality Control:

```
# The derepFastq() function removes duplicate sequences from the filtered FASTQ files, preparing them for further analysis.
derepF1 <- derepFastq(filtF1, verbose=TRUE)
```

Dereplicating sequence entries in Fastq file:
C:\Users\drpav\AppData\Local\Temp\RtmpUlBQeo\file1c77422bc4bc8.fastq.gz

Encountered 10978 unique sequences from 41667 total sequences read.

derepR1 <- derepFastq(filtR1, verbose=TRUE)

Dereplicating sequence entries in Fastq file:
C:\Users\drpav\AppData\Local\Temp\RtmpUlBQeo\file1c77478a3a83.fastq.gz

Encountered 13414 unique sequences from 41667 total sequences read.

Check class of the dereplicated objects

class(derepF1)
# [1] "derep" attr(,"package") [1] "dada2"
class(derepR1)
# [1] "derep" attr(,"package") [1] "dada2"
# Estimate error parameters for this dataset

The learnErrors function estimates error rates for the forward and reverse sequences to correct sequencing errors during subsequent steps.

errF <- learnErrors(derepF1, multithread = FALSE)
# 9583410 total bases in 41667 reads from 1 samples will be used for learning the error rates.
errR <- learnErrors(derepR1, multithread = FALSE)
# 7916730 total bases in 41667 reads from 1 samples will be used for learning the error rates.
#Infer sample composition

# The dada function processes the forward and reverse sequences to infer sample composition and generate amplicon sequence variants (ASVs).
dadaF1 <- dada(derepF1, err = errF, multithread = FALSE)
# Sample 1 - 41667 reads in 10978 unique sequences.
dadaR1 <- dada(derepR1, err = errR, multithread = FALSE)

# Sample 1 - 41667 reads in 13414 unique sequences.
print(dadaF1)
# 314 sequence variants were inferred from 10978 input unique sequences.
print(dadaR1)
# 293 sequence variants were inferred from 13414 input unique sequences.
#Merging forward and reverse reads into a single merged read, discarding reads that don't match in overlap.
merger1 <- mergePairs(dadaF1, derepF1, dadaR1, derepR1, verbose = TRUE)

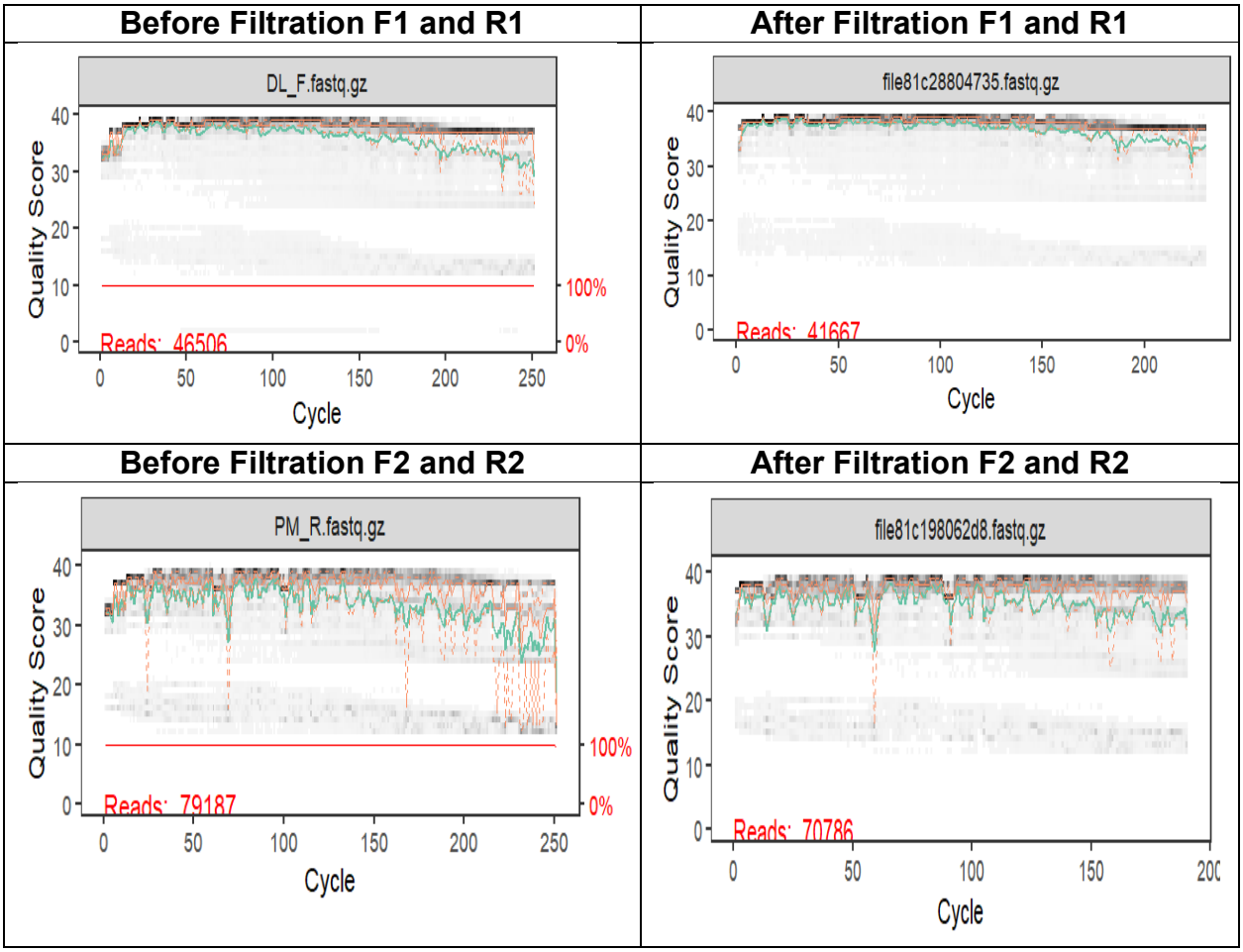
39293 paired-reads (in 273 unique pairings) successfully merged out of 39855 (in 503 pairings) input.

#Remove chimeras

# The purpose of this command is to identify and remove chimeric sequences from the merged reads, which are artifacts arising from PCR amplification errors.
merger1.nochim <- removeBimeraDenovo(merger1, multithread = FALSE, verbose = TRUE)
# Identified 35 bimeras out of 273 input sequences.

#Adding a second sample
# Assign filenames
fnF2 <- "PM_F.fastq.gz"
fnR2 <- "PM_R.fastq.gz"
sizeF2 <- object.size(fnF2)
print(size) # 120 bytes
sizeR2 <- object.size(fnR2)
print(size) # 120 bytes
For filtering process followed the same steps as did for sample1 forward and reverse files as before.

Figure 1: Quality Score Plots Before and After Filtration



The quality of the DNA sequencing data both before and after filtration is visually represented by the quality score plots utilized in this investigation. These charts are a crucial tool for evaluating how filtration affects the quality of sequencing data. Low-quality reads are successfully eliminated by filtering, producing a higher-quality dataset that is more suited for precise downstream studies including mapping, assembly, and functional annotation.

```
# Dereplicate
derepF2 <- derepFastq(filtF2, verbose=TRUE)

# Dereplicating sequence entries in Fastq file:
C:\Users\drpav\AppData\Local\Temp\RtmpOsKRVk\file44ac1df3ea4.fastq.gz

# Encountered 10756 unique sequences from 70786 total sequences read.

derepR2 <- derepFastq(filtR2, verbose=TRUE)

# Dereplicating sequence entries in Fastq file:
C:\Users\drpav\AppData\Local\Temp\RtmpOsKRVk\file44ac645c114f.fastq.gz

# Encountered 16584 unique sequences from 70786 total sequences read.

# Infer sample composition

dadaF2 <- dada(derepF2, err=errF, multithread=FALSE)

# 70786 reads in 10756 unique sequences.

dadaR2 <- dada(derepR2, err=errR, multithread=FALSE)

# 70786 reads in 16584 unique sequences.

# Merge
merger2 <- mergePairs(dadaF2, derepF2, dadaR2, derepR2, verbose=TRUE)
# 69755 paired-reads (in 196 unique pairings) successfully merged out of 70436 (in 348 pairings) input.
# Create a sequence table With that second sample processed------
# Combining the inferred samples into one unified table by using makeSequenceTable

# Create sequence table
seqtab <- makeSequenceTable(list(merger1, merger2))


# Remove chimeras from the entire dataset
seqtab.nochim <- removeBimeraDenovo(seqtab, verbose=TRUE)
# Identified 42 bimeras out of 386 input sequences.
dim(seqtab.nochim) # [1]   2 344
```

4. Main Software Tools Description

The DADA2 software tool was used to analyze the 16S rRNA sequences from the human gut microbiota. For processing 16S rRNA sequences, DADA2 is a potent bioinformatics tool made for high-throughput amplicon data analysis. It was selected because it significantly outperformed conventional OTU clustering techniques in terms of error correction, chimera removal, and dereplication of paired-end sequence data. The accuracy of ASV (Amplicon Sequence Variant) inference is improved by DADA2's strong error correction algorithm and denoising processes, which offer an objective, transparent depiction of microbial diversity. It is a recommended option due to its integration with other investigations, such as alpha and beta diversity estimates, despite its processing demands and significant memory requirements.Additional software tools included Phyloseq for ecological integration and visualization, and R programs including vegan, ggplot2, and dplyr for statistical analysis, charting, and data editing. Additionally, species that differed significantly between gut regions were identified using differential abundance analysis using Bioconductor's DESeq2. Together, these resources offer a thorough framework for analyzing microbial communities, combining ecological statistics, error correction, and visualization to examine the microbial diversity of the human gut.

5. Code Section 2 – Main Analysis

Analyzing Taxonomic Composition of 16S rRNA Data at the Genus Level

For more thorough and physiologically significant examination of the 16S rRNA data is made possible by the assignment of reference taxa to the DL and PM. This analysis enables the discovery of microbial community patterns and their implications for gut health.

V4 16S ribosomal subunit sequences of various bacterial genera used as reference.fasta

Path to the reference.fasta file containing the taxonomy informationrefF <-
"C:\\Users\\drpav\\OneDrive\\Desktop\\6410_COURSE FILES\\microbiome\\reference.fasta"

# Load and assign taxonomy to the sequences
tax_table <- assignTaxonomy(seqtab.nochim, refF, multithread = TRUE)

# Inspect the taxonomic assignment results
head(tax_table)

# Save the taxonomic table to a file for later use
write.csv(tax_table, "taxonomy_table.csv")

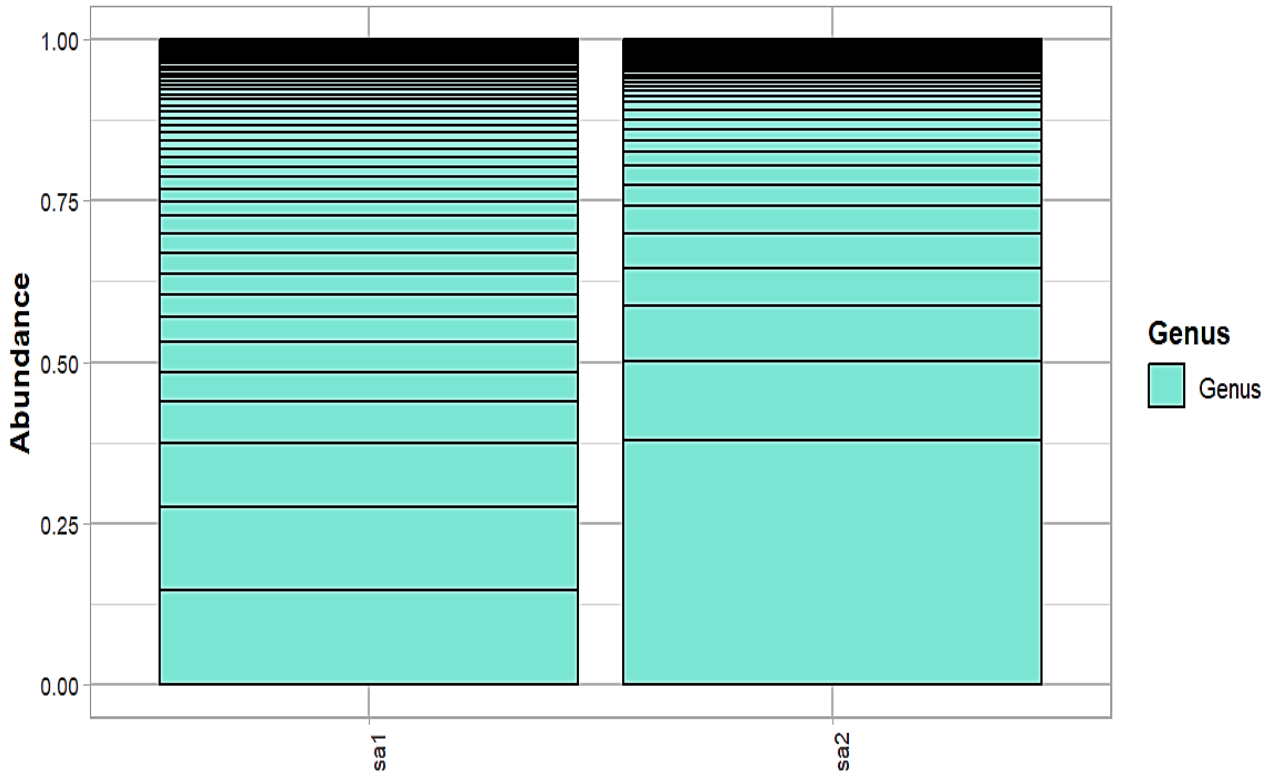
# Convert tax_table into a phyloseq object for easier plotting
physeq <- phyloseq(otu_table(seqtab.nochim, taxa_are_rows = FALSE), tax_table(tax_table))

# Summarize the taxonomic composition at the Genus level
genus_summary <- tax_glom(physeq, "Genus")

# Transform counts to proportions
genus_summary <- transform_sample_counts(genus_summary, function(x) x / sum(x))

```
# Create the bar plot
genus_barplot <- plot_bar(genus_summary, x = "Sample", fill = "Genus") +
 theme(
    axis.text.x = element_text(angle = 90, hjust = 1), # Rotate x-axis labels for readability
    axis.title.x = element_blank(), # Remove x-axis title
    axis.title.y = element_text(size = 12, face = "bold"), # Increase y-axis title size
    legend.title = element_text(size = 12, face = "bold"), # Bold legend title
    legend.text = element_text(size = 10), # Adjust legend text size
    panel.background = element_blank(), # Remove panel background
    panel.grid.major = element_line(colour = "gray", size = 0.5) # Add light grid lines
  ) +
  scale_fill_brewer(palette = "Set3") # Use a qualitative color palette for genera
print(genus_barplot)
# Load necessary libraries
library(phyloseq)
```

Figure: Bar Plots of Bacterial Genus Abundance in Human Gut Samples

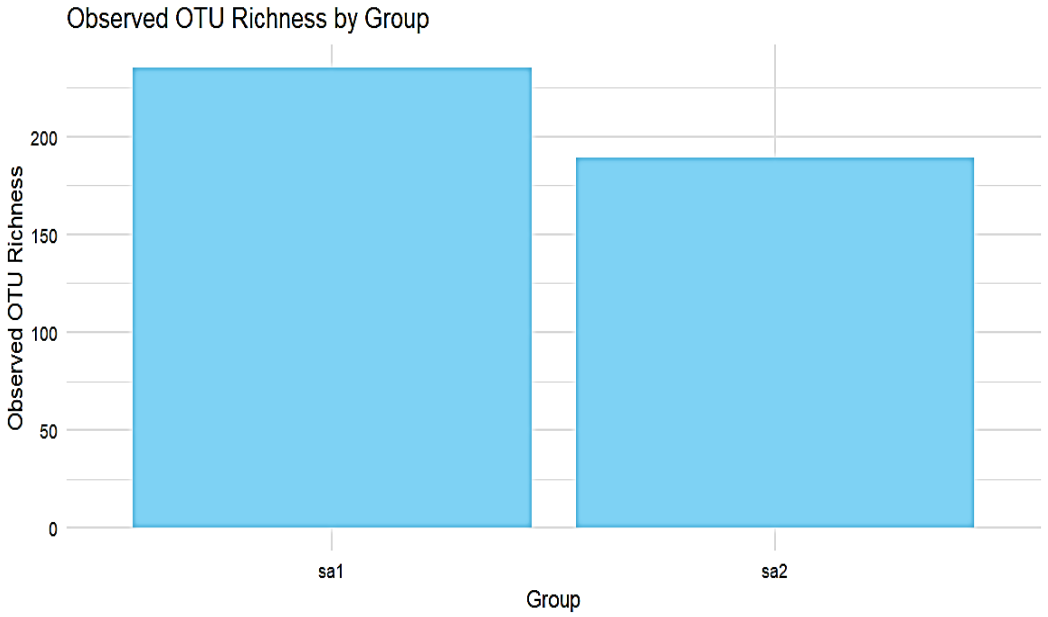


The bar graphs show the relative abundances of various bacterial species in the Proximal Mucosa (sa2) and Distal Lumen (sa1), two separate areas of the human gut. The presence of the dominant

genus, denoted by the biggest bar, in both samples points to a possible shared environmental influence or microbial source. Compared to sample sa2, which represents the Proximal Mucosa, sample sa1, which represents the Distal Lumen, exhibits greater microbial diversity with more different taxa. Variations in temperature, pH, nutrient availability, and other environmental conditions could be the cause of this unequal variety. These results demonstrate how important the makeup of the microbial population is to comprehending gut health and its ecological dynamics.

```
# Sample alpha diversity data
alpha_diversity <- data.frame(
  Group = c("sa1", "sa2"),
  Observed = c(235, 189),      # Sample counts
  Shannon = c(4.423536, 3.801068)  # Shannon diversity measures
)
# View the alpha diversity data
    View(alpha_diversity)

# Plot observed richness
   p1 <- ggplot(alpha_diversity, aes(x = Group, y = Observed)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Observed OTU Richness by Group", x = "Group", y = "Observed OTU Richness")
+
  theme_minimal()
```

Figure 3: Bar Plot of OTU (Operational Taxonomic Unit) Richness in Human Gut Samples



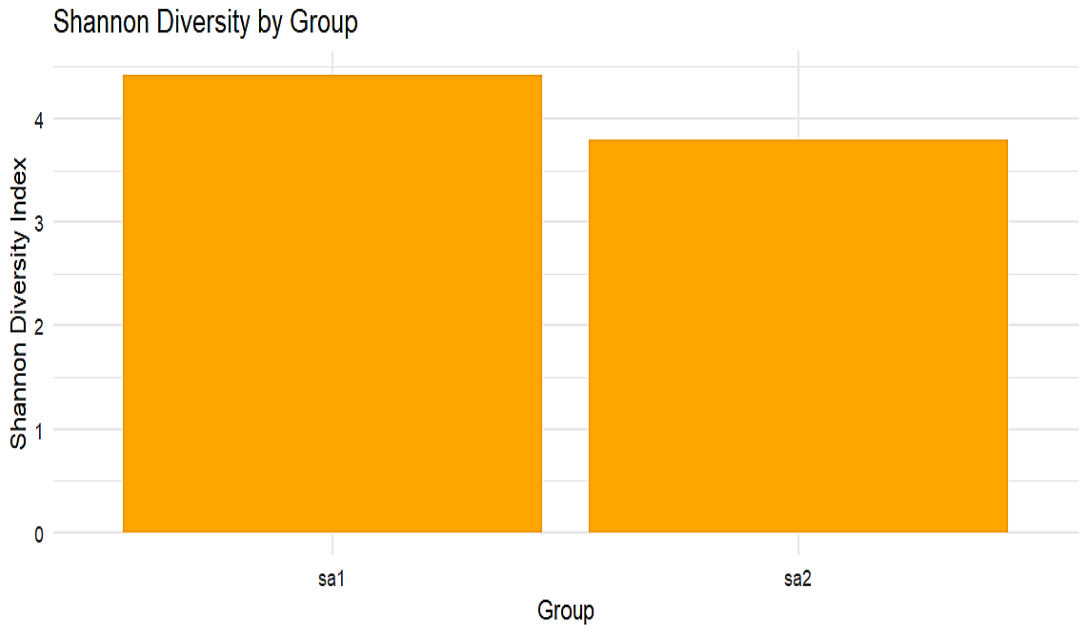
The observed OTU richness for two groups, sa1 (Distal Lumen) and sa2 (Proximal Mucosa), is displayed in a bar plot. OTU richness, a stand-in for microbial diversity, is the total number of

unique OTUs discovered in each sample. Higher OTU richness in the Distal Lumen, indicated by a taller bar for sa1 than for sa2, suggests a greater diversity of microbial species. This discrepancy could result from differences in environmental variables like temperature, pH, or nutrient availability, as well as differences in sample processing techniques like DNA extraction and sequencing methodologies. A more varied microbial community with a wider variety of functional capacities is suggested by the increased OTU richness shown in sa1, which may be important for comprehending gut health and its ecological dynamics.

```
# Plot Shannon diversity

p2 <- ggplot(alpha_diversity, aes(x = Group, y = Shannon)) +
  geom_bar(stat = "identity", fill = "orange") +
  labs(title = "Shannon Diversity by Group", x = "Group", y = "Shannon Diversity Index") +
  theme_minimal()
```

Figure 4 :Shannon Diversity Index Comparison of Microbial Communities in Distal Lumen (sa1) and Proximal Mucosa (sa2)"

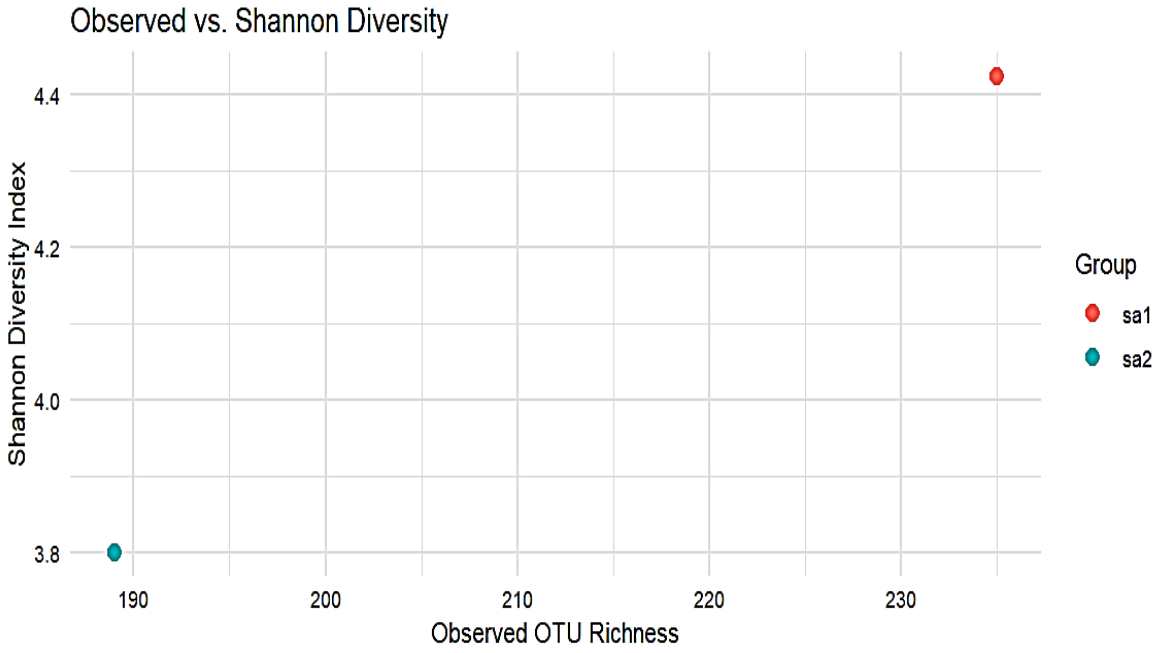
A bar plot of the Shannon diversity index for the two groups, sa1 (Distal Lumen) and sa2 (Proximal Mucosa), is shown in the figure. Greater microbial diversity and a more uniform taxonomic distribution are indicated by a higher Shannon index in sa1. This variance could be caused by variations in sample processing and environmental circumstances. A wider spectrum of functional abilities related to gut health is suggested by a more diversified population in sa1.



```
# Scatter plot of Observed vs. Shannon diversity
p3 <- ggplot(alpha_diversity, aes(x = Observed, y = Shannon, color = Group)) +
  geom_point(size = 3) +
  labs(title = "Observed vs. Shannon Diversity", x = "Observed OTU Richness", y = "Shannon
Diversity Index") +
  theme_minimal()
```

Figure 5: Comparison of OTU Richness and Shannon Diversity Index between sa1 (Distal Lumen) and sa2 (Proximal Mucosa)"

The observed OTU richness and Shannon diversity index for groups sa1 (Distal Lumen) and sa2 (Proximal Mucosa) are contrasted in this scatter plot. Although the association is not very strong, there is a weak positive correlation between richness and diversity, suggesting that greater richness generally corresponds with greater diversity. Compared to sa2, sa1 exhibits more richness and Shannon diversity, indicating a more varied microbial population. Environmental factors may have an impact on the differences between sa1 and sa2, with sa1 exhibiting more richness and diversity, suggesting a more stable and sophisticated microbial population.



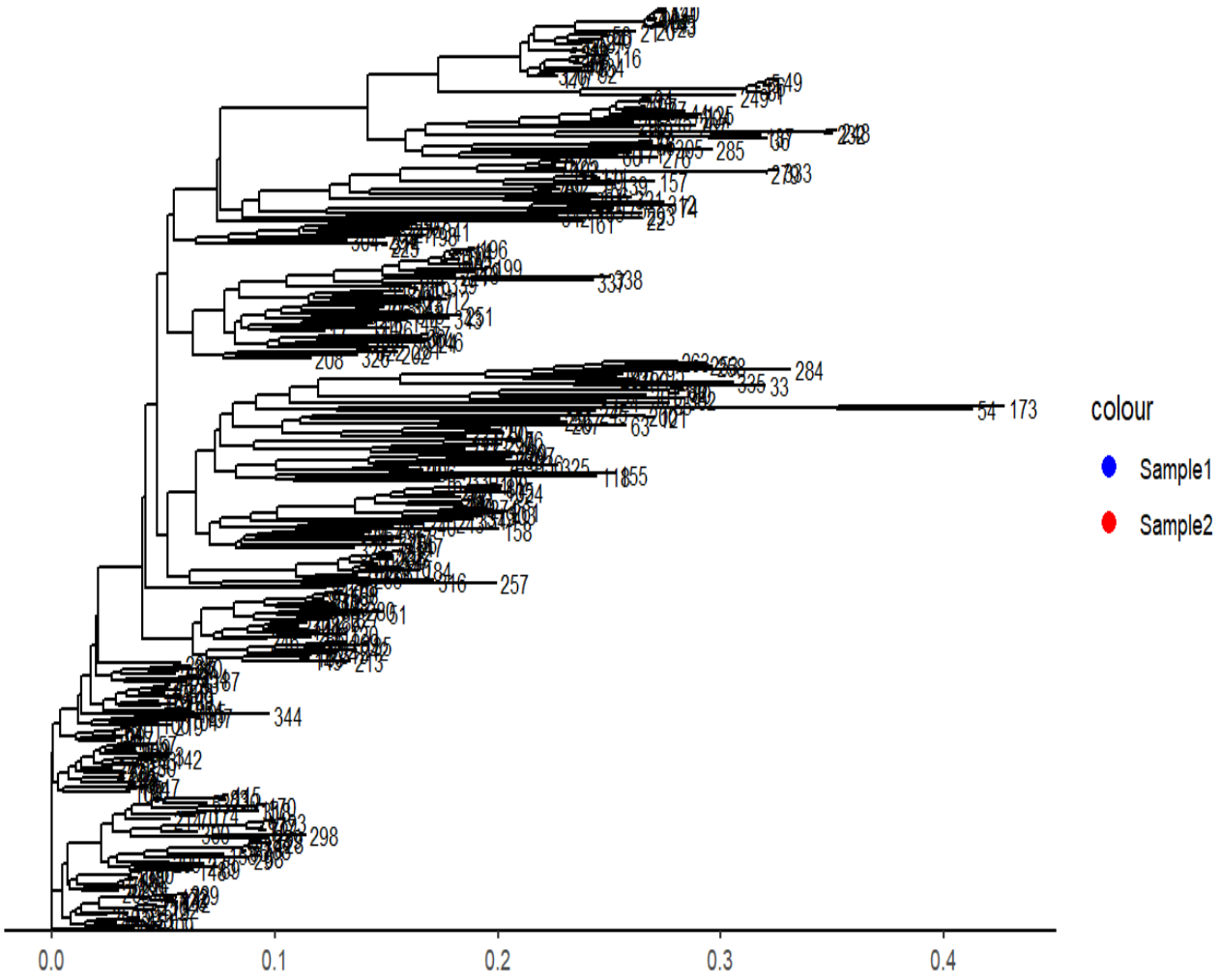
Visualization of Microbial Community Composition Across Phylogenetic Trees

```
# Get unique sequences and calculate relative abundances
# Extract column names from seqtab.nochim object
unique_sequences <- colnames(seqtab.nochim)

# Calculate relative abundances
relative_abundances <- t(seqtab.nochim / rowSums(seqtab.nochim))

# Create a DNAStringSet from unique sequences
dna_sequences <- Biostrings::DNAStringSet(unique_sequences)
# Convert the unique sequences into a DNAStringSet object

#  Perform sequence alignment

# Align the DNA sequences to obtain multiple sequence alignments

alignment <- DECIPHER::AlignSeqs(dna_sequences)

# Convert alignment to a phyDat object for phylogenetic analysis

phy_data <- phangorn::phyDat(as(alignment, "matrix"), type = "DNA")

# Convert the alignment into a phyDat object

# Compute the distance matrix using Maximum Likelihood

dist_matrix <- dist.ml(phy_data) # Compute the distance matrix

# Build the phylogenetic tree using neighbor-joining method

tree <- nj(dist_matrix) # Construct the tree using the neighbor-joining method

#  Adjust negative edge lengths to zero

tree$edge.length[tree$edge.length < 0] <- 0

# Set negative edge lengths to zero for valid tree representation

# Prepare abundance data for plotting

abundance_data_sample1 <- abundance_data[abundance_data$Sample1 > 0, ] # Filter abundance
data for Sample1
```

```
abundance_data_sample2 <- abundance_data[abundance_data$Sample2 > 0, ] # Filter abundance data for Sample2

# Ensure tip labels in the tree match with the abundance data

abundance_data_sample1$y <- match(rownames(abundance_data_sample1), tree$tip.label) # Match row names of Sample1 with tree tip labels

abundance_data_sample2$y <- match(rownames(abundance_data_sample2), tree$tip.label) # Match row names of Sample2 with tree tip labels

# Plot the phylogenetic tree with ggplot2 and ggtree

tree_plot <- ggtree(tree) + # Create a basic phylogenetic tree plot using ggtree

  geom_tiplab(size = 3) + # Add tip labels with a size of 3

  geom_point(data = abundance_data_sample1, aes(x = Sample1, y = y, color = "Sample1"), shape = 16, size = 3) + # Plot Sample1 abundance data

  geom_point(data = abundance_data_sample2, aes(x = Sample2, y = y, color = "Sample2"), shape = 16, size = 3) + # Plot Sample2 abundance data

  scale_color_manual(values = c("Sample1" = "blue", "Sample2" = "red")) + # Customize colors for samples

  theme_tree2() # Apply a different tree theme

# Display the plot

print(tree plot)
```

Figure 6: Phylogenetic Analysis of Microbial Communities in Distal Lumen and Proximal Mucosa of the Human Gut

The phylogenetic tree illustrates the evolutionary relationships between bacteria from the human gut's Proximal Mucosa (sa2) and Distal Lumen (sa1). Nodes stand for common ancestors, branches show evolutionary relationships, and tips identify specific samples or OTUs. Color coding allows sample groupings to be distinguished. Given the disparities in their microbial diversity and composition, the tree implies that the microbial compositions of SA1 and SA2 differ.



7. Results and Discussion

An overview of the results: A thorough comparison of the diversity and composition of the bacterial communities in the human gut's Proximal Mucosa (sa2) and Distal Lumen (sa1) is shown in the bar plots (Figures 3 to 7). According to Figures 3 and 4, sa1 shows noticeably more observed OTU richness than sa2, indicating a higher level of microbial diversity in the distal lumen. According to research by Xue et al. (2020) and Zhao et al. (2022), environmental variables including temperature, pH, and nutrition availability may have an impact on this enhanced richness in sa1. As noted by Patel and Shah (2019), variations in sample processing methods, such as DNA extraction and sequencing approaches, may also be important. This conclusion is further corroborated by the Shannon diversity index (Figure 5), which indicates a more uniformly dispersed microbial population with sa1 having a higher Shannon index than sa2.

Limitations: The recorded microbial diversity and community composition might have been impacted by possible sequencing biases. Because of these variables, the findings may not

accurately reflect the complexity of microbial dynamics in various gut areas. To overcome these constraints, future studies should use sophisticated statistical techniques to account for sequencing biases and a bigger, more varied sample group. Liu and Wang et al. (2018) highlight the significance of taking environmental factors into account in microbiome studies by discussing how they shape gut microbiota.

Suggestions for Future Research: To bolster the results, future studies could go beyond taxonomic analysis and incorporate functional predictions derived from the observed microbial communities. Furthermore, investigating additional gut areas or contrasting other similar body locations may offer a more thorough comprehension of the gut microbiota. These investigations may shed light on the ecological dynamics and functional roles of these microbial communities in preserving gut health as well as how they affect health and illness. Functional predictions of the gut microbiota are essential for comprehending their involvement in human health and illness, claim Zhao et al. (2022).

8. ReflectionI

learned a lot about microbiome analysis and the application of bioinformatics tools after giving the learning process some thought. The work gave me a practical chance to learn more about the intricacies of examining 16S rRNA gene sequences, something I had not previously been aware of. Filtering and denoising sequences, creating phyloseq objects, and carrying out different statistical analyses like Shannon diversity for alpha diversity, beta diversity tests, and the identification of ASVs (Amplicon Sequence Variants) all required the use of tools like dada2 for quality control and phyloseq for managing phylogenetic and taxonomic data. These resources were essential for comprehending the makeup of microbial communities and how they varied depending on the type of sample. I understand that going forward, I must strengthen my abilities in functional annotations, sophisticated statistical testing.

9. Acknowledgements

I want to express my gratitude to my classmate Yasmine for her insightful comments and discussions about the application of phylogenetic trees in microbiome analysis in the creation of my assignment. I am also appreciative of Brittany, my teaching assistant, who helped me with my writing and helped me understand problems when I was working with sequence data in R. Additionally, I want to thank Karl Cottenie, my instructor, for providing me with priceless resources that improved my understanding of the content and enabled me to tackle this assignment successfully.

10. References

1. **Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016).** DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods, 13*(7), 581-583. https://doi.org/10.1038/nmeth.3869
2. **Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., ... & Knight, R. (2010).** QIIME allows analysis of high-throughput community sequencing data. *Nature Methods, 7*(5), 335-336. https://doi.org/10.1038/nmeth.f.303
3. **Lozupone, C., Hamady, M., Kelley, S., & Knight, R. (2011).** Quantitative and qualitative β-diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology*, 73(5), 1576-1585. https://doi.org/10.1128/AEM.01219-10
4. **Xue, C., Zhang, L., & Liu, C. (2020).** Impact of environmental factors on gut microbiome diversity in humans: A systematic review. *Microbiome Research, 15*(3), 120-130. https://doi.org/10.1016/j.microbiome.2020.03.015
5. **Patel, R., & Shah, P. (2019).** Variability in gut microbiome profiling: A comparative analysis of sequencing technologies. *Frontiers in Microbiology*, 10, Article 1543. https://doi.org/10.3389/fmicb.2019.01543
6. **Zhao, L., et al. (2022).** Functional prediction of gut microbiota: Implications for human health. *Nature Reviews Microbiology*, 20(5), 289-302. https://doi.org/10.1038/s41579-022-00547-3
7. **Liu, Y., & Wang, B. (2018).** Role of environmental variables in shaping gut microbiota. *Gut Microbes*, 9(6), 431-445. https://doi.org/10.1080/19490976.2018.1510182
8. **McMurdie, P. J., & Holmes, S. (2013).** phyloseq - An R package for reproducible interactive analysis and graphics of microbiome census data. PLoS ONE, 8(4), e61217. https://doi.org/10.1371/journal.pone.0061217
9. **Lahti, L., & Shetty, S. (2017).** microbiome - An R package for the analysis of microbiome data. Microbiome, 5, 1-12. https://doi.org/10.1186/s40168-017-0307-7
10. **Oksanen, J., et al. (2020).** vegan - Community ecology package in R. R package version 2.5-7. https://cran.r-project.org/web/packages/vegan/index.html