Human Papilloma Virus Infection impact on the Intestinal Microbiome in Colorectal Cancer Patients

Elliot Dean

2024-04-30

Contents

1 R Environment Setup

R Packages are installed (if needed):

```r
install.packages("tidyverse")               # Compilation of packages for data management
install.packages("here")                    # Provides a relative path to files
install.packages("gt")                      # Create Presentation-Ready Display Tables
install.packages("devtools")                # Collection of package development tools
devtools::install_github("jbisanz/qiime2R") # Import QIIME2 artifacts to R
install.packages("vegan")                   # Functions for community ecologists
install.packages("BiocManager")             # Open source software for Bioinformatics
BiocManager::install("phyloseq")            # Explore microbiome profiles using R
BiocManager::install("microbiome")          # Utilities for microbiome analysis
```

Libraries are activated:

```r
library(tidyverse)
library(readxl)
library(here)
library(gt)
library(qiime2R)
library(vegan)
library(phyloseq)
library(microbiome)
```

A standardized set of colors was chosen to represent metadata properties of the samples:

```r
# Specify specific colors for important metadata
color_palette <- c(`CRC+HPV` = "deeppink", CRC = "deepskyblue")
```

Custom Functions:

```r
# Used to create Beta Diversity Ordination Plots
# phyobj = Phyloseq Object  |   ordobj = Ordination Object

ordination_plotting <- function(phyobj, ordobj) {
  plot_ordination(phyobj, ordobj, color = "condition") +
  geom_point(size = 5, alpha = 0.5) +
  scale_color_manual(values = color_palette) +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(color = "Condition") }

# Function to create a boxplot using ggplot2
# data = Data Frame

plotbox <- function(data) {
  boxq1 <- quantile(data$Abundance, 0.1)
  boxq9 <- quantile(data$Abundance, 0.9)
  box_data <- data %>% filter(data$Abundance >= boxq1 & data$Abundance <= boxq9)

  ggplot(data, aes(x = condition, y = Abundance, fill = condition)) +
    geom_boxplot(outliers = FALSE) +
    geom_point(data = box_data, mapping = aes(x = condition, y = Abundance),
               color = "black", position = "jitter") +
    labs(x = "Condition") +
    scale_fill_manual(values = color_palette) +
    theme(legend.position = "Hide", plot.title = element_text(hjust = 0.5)) }
```

```
# Calculate averages & perform t-test
# group1 = Data Frame 1 | group2 = Data Frame 2

groupavg <- function(group1, group2) {
  mean1 <- mean(group1$Abundance)
  mean2 <- mean(group2$Abundance)
  ttest <- t.test(group1$Abundance, group2$Abundance)$p.value
  result <-
    paste("CRC Average =", mean1," ", "CRC + HPV Average =", mean2,'\n',"T-Test, p =", ttest)
  cat(result, "\n") }
```

Note on plotbox(): To keep boxplots organized, outliers (as calculated by **geom_boxplot**) have been excluded from plots. In order to include more sample points, the dataset was slightly constrained using **quantile**, and this new data frame was used to plot individual values via **geom_point**.

Load .RData Objects if necessary:

```
load("rdata_files/sample_info.RData")
load("rdata_files/gse_physeq.RData")
```

---

2 Introduction

16S RNA Sequencing data from colorectal cancer (CRC) patient stool samples with or without concomitant Human Papilloma Virus (HPV) infection was received via a shared **box** storage folder. The sequencing was performed on an **Illumina** platform, using primers for the V3 & V4 hypervariable regions. The folder contained one subfolder - **fastq_raw**, and one file - **2024_04_11_gse216589_metadata_v1.xlsx**. The data files were in fastq format, 40 files total (this was confirmed after downloading using **Finder** in MacOS). The metadata associated with the samples listed 11 groups - **Condition** (either CRC or CRC + HPV), **crc_sex** (type of cancer), **crc_site** (location of cancer), **Organism** (*Homo sapiens*), **Patient ID**, **sex** (male or female), **Source Name**, **Time**, **tumor_grading**, and **tumor_stage**.

2.1 Preparing Sample Metadata

The original sample metadata file had a Microsoft Excel extension, and needed to be converted into a tab separated value document. To do this, the xlsx was imported into **R** using the **read_excel** method. Subsequently, **2024_04_11_gse216589_metadata.tsv** was created by running the **write.table** command, taking the xlsx as input, and specifying the separator to be used.

```
xcel_data <- read_excel(here("set02/2024_04_11_gse216589_metadata_v1.xlsx")) %>%
  rename(`sample-id` = `sample_id`)

write.table(xcel_data, file = here("2024_04_11_gse216589_metadata.tsv"),
  sep = "\t", quote = FALSE, row.names = FALSE)
```

2.2 Preparing Manifest File

To facilitate import of sequence data into **QIIME2**, a manifest file was created (**2024_04_11_gse216589_manifest.tsv**). The file contains sample ID, along with full path names to the sequencing files (for each sample). To standardize path names, the **here** method is used to store the character string of the RStudio project working directory.

```
# Use xcel_data as a table foundation
# Create new columns using mutate, manipulate names (strings) with paste0

there <- here()

gse_manifest <- xcel_data %>%
  mutate(`forward-absolute-filepath` =
           paste0(there, "/set02/fastq_raw/", `sample-id`, "_1.fastq")) %>%
  mutate(`reverse-absolute-filepath` =
           paste0(there, "/set02/fastq_raw/", `sample-id`, "_2.fastq")) %>%
  select(`sample-id`, `forward-absolute-filepath`, `reverse-absolute-filepath`)

write.table(gse_manifest, here("2024_04_11_gse216589_manifest.tsv"),
            sep = "\t", quote = FALSE, row.names = FALSE)
```

2.3 Objectives

The link between HPV infection and certain types of cancer (depending on viral subtype) has been demonstrated thoroughly. Far less is known about the impact of HPV infection on the composition of the intestinal microbiome, especially in patients with CRC. In this analysis, two major groups (CRC+/HPV- and CRC+/HPV+) were compared to determine if the presence, and/or abundance, of certain bacterial taxa found in stool samples were significantly different.

---

3 From FASTQ to QZA

To run QIIME via a command line interface, a system terminal was used. Activating the conda environment containing QIIME2 is necessary prior to starting.

Importing the sequences into a QIIME archive was completed using the following command:

```
qiime tools import \
  --type SampleData[PairedEndSequencesWithQuality] \
  --input-path 2024_04_11_gse216589_manifest.tsv \
  --output-path qiime/2024_04_11_gse216589_demux.qza \
  --input-format PairedEndFastqManifestPhred33V2
```

import is one of the fundamental procedures in the QIIME **tools** library.

The import method reads the manifest file provided to it, and collects all of the sequence files denoted (along with their respective sample id) into an archive.

The qza file was used to produce a qzv file:

```
qiime demux summarize \
  --i-data qiime/2024_04_11_gse216589_demux.qza \
  --o-visualization qiime/2024_04_11_gse216589_demux.qzv
```

The sequences, which have already been demultiplexed, were reviewed using QIIME2 View.

Demultiplexing is the method by which sequencing reads are assigned to their sample of origin based on the sequence of their corresponding DNA barcode (identifier).

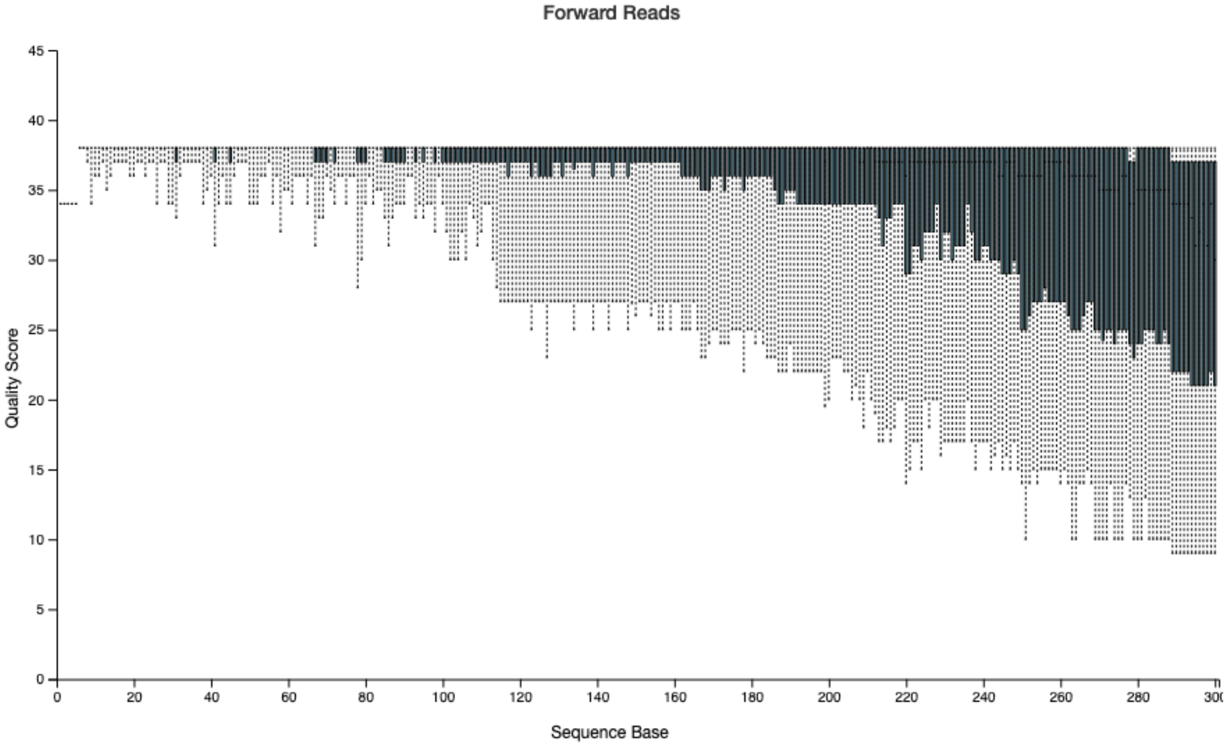


Figure 1. Quality Plot, Forward Reads

The Quality Score reflects the confidence of the sequence data.

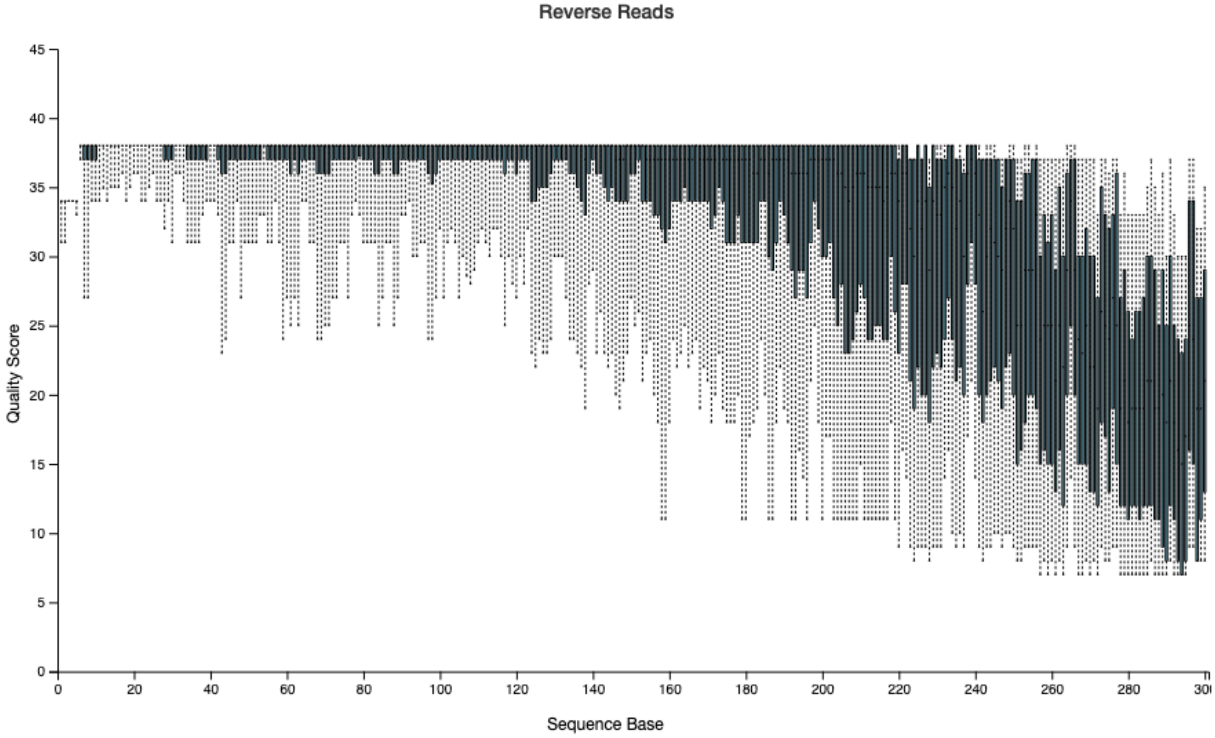


Figure 2. Quality Plot, Reverse Reads

4 Sequence Quality Control

Once the sequence files were imported, the demux summary was created and viewed. For this analysis, the determination of sequence cutoff location based on quality score was a value of 10. More explicitly, the parameter used for quality score value was the Lower Whisker value of the bar plot for each base. As the early reads for both directions are always above a quality score of 20, no sequence was trimmed from left side. The quality scores of the forward reads remained high (at or above 14) until position 250. It gradually decreased until a score of 10 for position 288, dropping to 9 after that location. As such, 288 was chosen as the truncation length for the forward reads. The quality scores of the reverse reads remained high (at or above 18) until position 157. It gradually decreased until a score of 11 for position 219, dropping to 9 after that location. As such, 219 was chosen as the truncation length for the reverse reads.

The DADA2 filtering technique utilized is paired end sequence denoising.

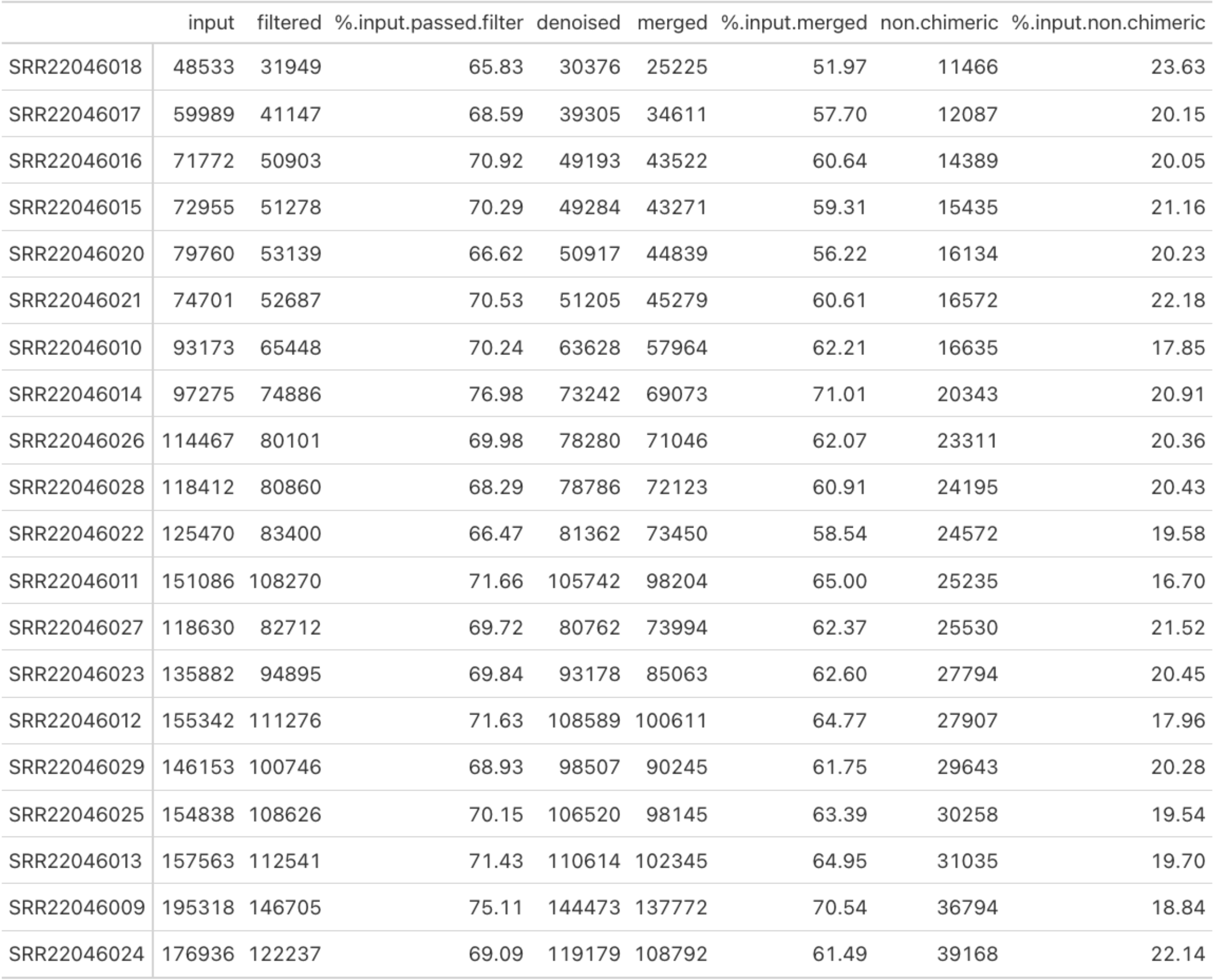
```
qiime dada2 denoise-paired \
  --i-demultiplexed-seqs qiime/2024_04_11_gse216589_demux.qza \
  --p-trim-left-f 0 \
  --p-trim-left-r 0 \
  --p-trunc-len-f 288 \
  --p-trunc-len-r 219 \
  --o-table qiime/gse_table.qza \
  --o-representative-sequences qiime/gse_rep_seqs.qza \
  --o-denoising-stats qiime/gse_stats.qza
```

Denoising methods filter out noisy sequences, correct errors in marginal sequences, remove chimeric sequences, remove singletons, join denoised paired-end reads, and dereplicate those sequences.

To review the denoising statistics, **gse_stats.qza** was imported into R using the **read_qza** method from the **qiime2R** library.

```
gse_dada2 <- read_qza(here("qiime/gse_stats.qza"))

gse_dada2$data %>% rownames_to_column() %>% arrange(non.chimeric) %>% gt() %>%
  cols_label(`percentage.of.input.passed.filter` = "%.input.passed.filter",
             `percentage.of.input.merged` = "%.input.merged",
             `percentage.of.input.non.chimeric` = "%.input.non.chimeric")
```



	input	filtered	%.input.passed.filter	denoised	merged	%.input.merged	non.chimeric	%.input.non.chimeric
SRR22046018	48533	31949	65.83	30376	25225	51.97	11466	23.63
SRR22046017	59989	41147	68.59	39305	34611	57.70	12087	20.15
SRR22046016	71772	50903	70.92	49193	43522	60.64	14389	20.05
SRR22046015	72955	51278	70.29	49284	43271	59.31	15435	21.16
SRR22046020	79760	53139	66.62	50917	44839	56.22	16134	20.23
SRR22046021	74701	52687	70.53	51205	45279	60.61	16572	22.18
SRR22046010	93173	65448	70.24	63628	57964	62.21	16635	17.85
SRR22046014	97275	74886	76.98	73242	69073	71.01	20343	20.91
SRR22046026	114467	80101	69.98	78280	71046	62.07	23311	20.36
SRR22046028	118412	80860	68.29	78786	72123	60.91	24195	20.43
SRR22046022	125470	83400	66.47	81362	73450	58.54	24572	19.58
SRR22046011	151086	108270	71.66	105742	98204	65.00	25235	16.70
SRR22046027	118630	82712	69.72	80762	73994	62.37	25530	21.52
SRR22046023	135882	94895	69.84	93178	85063	62.60	27794	20.45
SRR22046012	155342	111276	71.63	108589	100611	64.77	27907	17.96
SRR22046029	146153	100746	68.93	98507	90245	61.75	29643	20.28
SRR22046025	154838	108626	70.15	106520	98145	63.39	30258	19.54
SRR22046013	157563	112541	71.43	110614	102345	64.95	31035	19.70
SRR22046009	195318	146705	75.11	144473	137772	70.54	36794	18.84
SRR22046024	176936	122237	69.09	119179	108792	61.49	39168	22.14

Table 1. DADA2 Denoising Statistics

5 Phylogenetic Tree Generation

Using the representative sequences, the **fasttree** method from QIIME **phylogeny** was implemented:

```
qiime phylogeny align-to-tree-mafft-fasttree \
  --i-sequences qiime/gse_rep_seqs.qza \
  --o-alignment qiime/aligned_gse_rep_seqs.qza \
  --o-masked-alignment qiime/masked_aligned_gse_rep_seqs.qza \
  --o-tree qiime/unrooted_tree_gse_rep_seqs.qza \
  --o-rooted-tree qiime/rooted_tree_gse_rep_seqs.qza
```

6 Taxonomic Analysis

In order to assign taxonomy to species present in the samples, a classifier with sequence / feature information at every taxonomic level is used. This analysis utilized the **Silva 138 99% OTUs from 515F/806R region of sequences** library.

The selection of the Silva 138 classifier was based on the results of 2 studies (Odom *et al.*, Balvočiūtė *et al.*). Silva has historically been updated more frequently than one of the other major classifiers, Greengenes2. In addition, Silva 138 has a greater number of features (OTU's) to classify against compared to Greengenes, allowing more precision in taxonomic assignment, even to the species level. However, as these classifiers have been trained by QIIME2's **feature-classifier** plugin, there are specific properties of the library that are tied to the QIIME workflow.

```
qiime feature-classifier classify-sklearn \
  --i-classifier qiime/silva-138-99-515-806-nb-classifier.qza \
  --i-reads qiime/gse_rep_seqs.qza \
  --o-classification qiime/gse_taxonomy.qza
```

7 Creating a Phyloseq Object

qiime2R has a method (**qza_to_phyloseq**) to create a **Phyloseq** object from 3 QIIME archive files & the metadata tsv.

```
gse_physeq <-
  qza_to_phyloseq(
    features = here("qiime/gse_table.qza"),                   # Feature Table
    tree = here("qiime/rooted_tree_gse_rep_seqs.qza"),        # Rooted Phylogenetic Tree
    taxonomy = here("qiime/gse_taxonomy.qza"),                # Taxonomy
    metadata = here("2024_04_11_gse216589_metadata.tsv"))     # Sample Metadata
```

7.1 Accessing Metadata

The metadata associated with the Phyloseq object can be obtained using the **sample_data** method. However, this is a unique class of the same name, that may not be compatible with all non-ecology libraries. To make the information accessible in multiple forms, the sample data is stored in the **sample_info** object, and the rownames (SampleID) are added as a new column.

In addition, the **as_tibble** modifier is used to create a "tibble" (data frame) of the information.

```
sample_info <- sample_data(gse_physeq)       # Create a data object of sample metadata

sample_info$SampleID <- rownames(sample_info)        # Add Sample ID to metadata table

sample_information <- as_tibble(sample_data(gse_physeq))            # Data frame format
```

7.2 Extracting Data from an Object

As a Phyloseq object has multiple data types in different formats, it can be difficult to access information quickly. To alleviate this issue, a method called **psmelt** creates a large data frame of core data, including: OTU, Sample ID, Abundance Values, Metadata Categories, and every taxonomic level from Domain to Species (when available).

```
# Phyloseq object is "melted" into a large data frame of components
gse_phydata <- psmelt(gse_physeq)
```

8 Saving Items

All of the objects created thus far may be saved as RData files for easy retrieval the next time the environment is activated. Two collections of RData will be created - one set for the Physeq object, and another for the sample information.

```
# Save information from the phyloseq object
save(sample_info, sample_information, file = here("rdata_files/sample_info.RData"))

# Save phyloseq object items
save(gse_physeq, gse_phydata, file = here("rdata_files/gse_physeq.RData"))
```

---

9 Alpha Diversity

Species richness can be visualized with the **plot_richness** method in Phyloseq. Here, 3 metrics are displayed side by side in facet grids. The samples have been grouped by Condition.

```
plot_richness(physeq = gse_physeq, x = "condition", color = "condition",
  measures = c("Simpson", "Shannon", "InvSimpson")) +
geom_boxplot() +
  scale_color_manual(values = color_palette) +
  labs(title = "Alpha Diversity Metrics", x = NULL, y = "Alpha Diversity Measure") +
  theme(legend.position = "Hide", plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 0, hjust = 0.5))
```

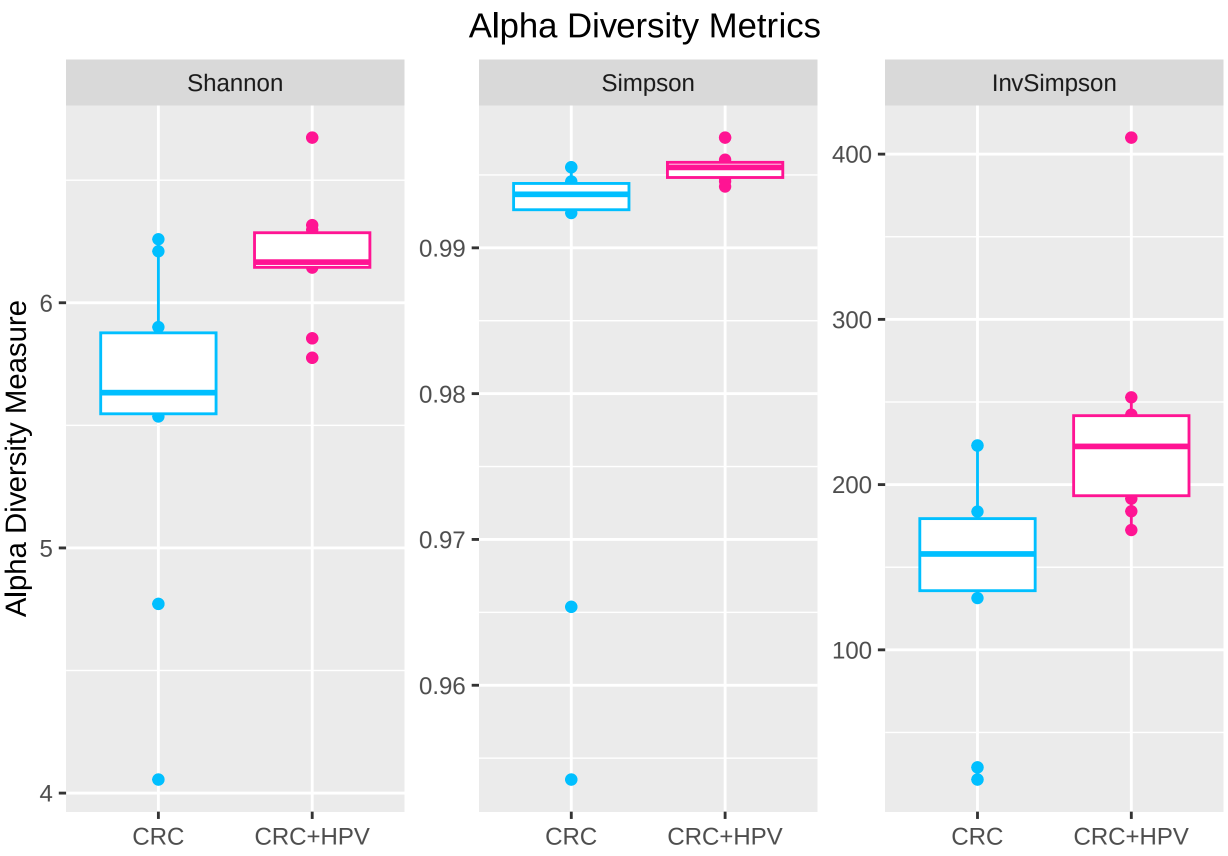


Figure 3. Shannon, Simpson, and Inverse Simpson Diversity

In addition to the 3 metrics presented so far, other Alpha Diversity measures can be calculated with the **estimate_richness** method in Phyloseq. To assess the evenness of populations in a sample, the **evenness** method from the **microbiome** library is utilized.

By creating a new column in both tables with the sample id, they can be joined together to access the complete dataset.

```
gse_ads <- estimate_richness(gse_physeq) %>%     # Calculate Alpha Diversity Values
  rownames_to_column(., "SampleID") %>%          # Add row names to column
  inner_join(sample_info)                        # Merge diversity values with metadata

gse_evenness <- evenness(gse_physeq) %>% rename(simpson_evenness = simpson) %>%
  rownames_to_column(., "SampleID")

gse_ads <- gse_ads %>% inner_join(gse_evenness)
```

9.1 Mann-Whitney U Test

To determine significance of the Alpha Diversity results, a Wilcoxon–Mann–Whitney Test may be carried out. This non-parametric statistical test is used to compare distributions between two groups, along a single variable.

By separating the table created earlier (**gse_ads**) into specific groups (by a single variable), multiple metrics can be easily referenced for testing.

The first set of groups created for analysis are split by condition.

```
gse_ads_crc <- gse_ads %>% filter(condition == "CRC")
gse_ads_crc_hpv <- gse_ads %>% filter(condition == "CRC+HPV")

# Shannon Diversity significance
wilcox.test(gse_ads_crc$Shannon, gse_ads_crc_hpv$Shannon)$p.value
```

```
## [1] 0.008930698
```

```
# Simpson Diversity significance
wilcox.test(gse_ads_crc$Simpson, gse_ads_crc_hpv$Simpson)$p.value
```

```
## [1] 0.002089242
```

```
# Pielou Evenness significance
wilcox.test(gse_ads_crc$pielou, gse_ads_crc_hpv$pielou)$p.value
```

```
## [1] 0.003886207
```

All 3 tests determined that species richness and evenness were superior in the CRC+HPV group compared to CRC alone.

Next, 2 groups were created based on sex.

```
gse_ads_male <- gse_ads %>% filter(sex == "male")
gse_ads_female <- gse_ads %>% filter(sex == "female")

# Shannon Diversity significance
wilcox.test(gse_ads_male$Shannon, gse_ads_female$Shannon)$p.value
```

```
## [1] 0.4377967
```

```
# Simpson Diversity significance
wilcox.test(gse_ads_male$Simpson, gse_ads_female$Simpson)$p.value
```

```
## [1] 0.3506966
```

```
# Pielou Evenness significance
wilcox.test(gse_ads_male$pielou, gse_ads_female$pielou)$p.value
```

```
## [1] 0.3113777
```

All 3 tests clearly demonstrated sex did not influence alpha diversity.

10 Beta Diversity

In addition to distance metrics, Phyloseq has a method to perform Principal Coordinate Analyses (and other ordinance calculations), **ordinate**. There are dozens of different combinations of ordination methods, distance metrics, and other parameters to control the analysis of samples.

A custom function was designed to implement the **plot_ordination** method from Phyloseq (**ordination_plotting**). This function takes a phyloseq object and an ordination object, the output being the Beta Diversity metric of interest.

3 Beta Diversity metrics (Weighed Unifrac, Unweighed Unifrac, and Bray-Curtis Dissimilarity) were used:

```
# Ordinate requires a Phyloseq object, method, & distance metric
gse_pcoa_wunifrac <- ordinate(gse_physeq, "PCoA", "wunifrac")

ordination_plotting(gse_physeq, gse_pcoa_wunifrac) +
  labs(title = "Weighted Unifrac Distance")
```

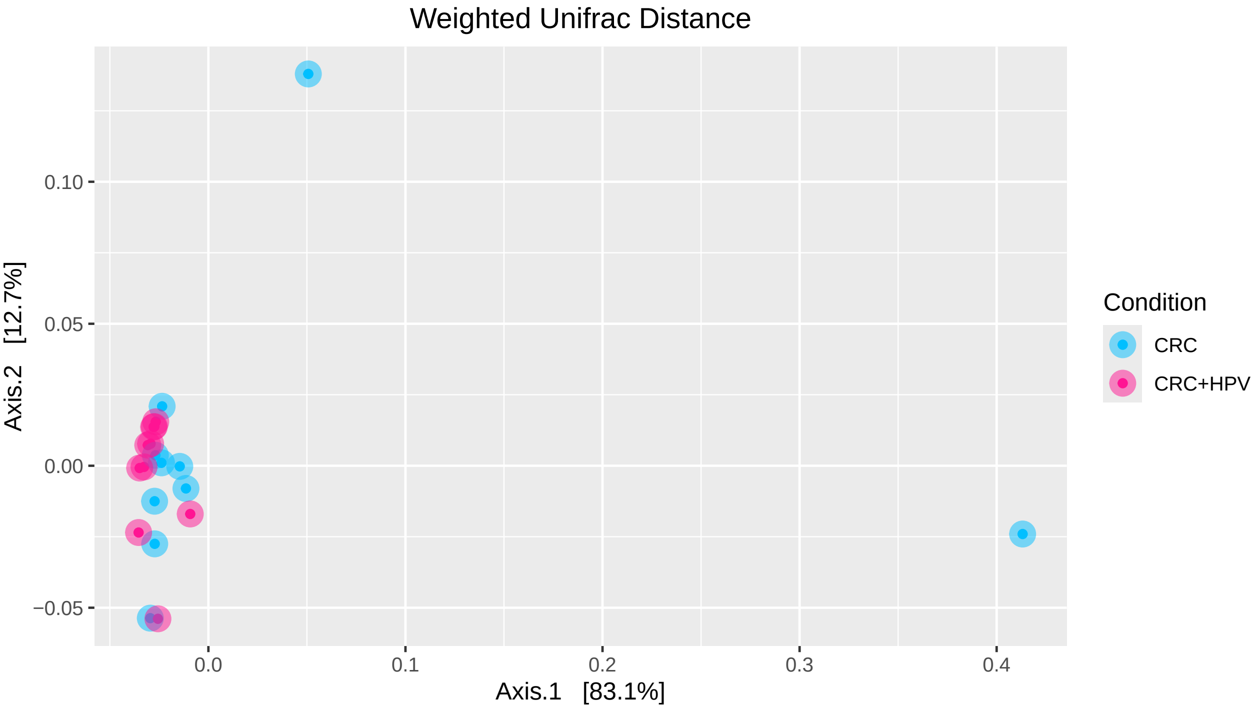


Figure 4. Unifrac Distance, Weighted

```
gse_pcoa_unifrac <- ordinate(gse_physeq, "PCoA", "unifrac")
ordination_plotting(gse_physeq, gse_pcoa_unifrac) +
  labs(title = "Unweighted Unifrac Distance")
```

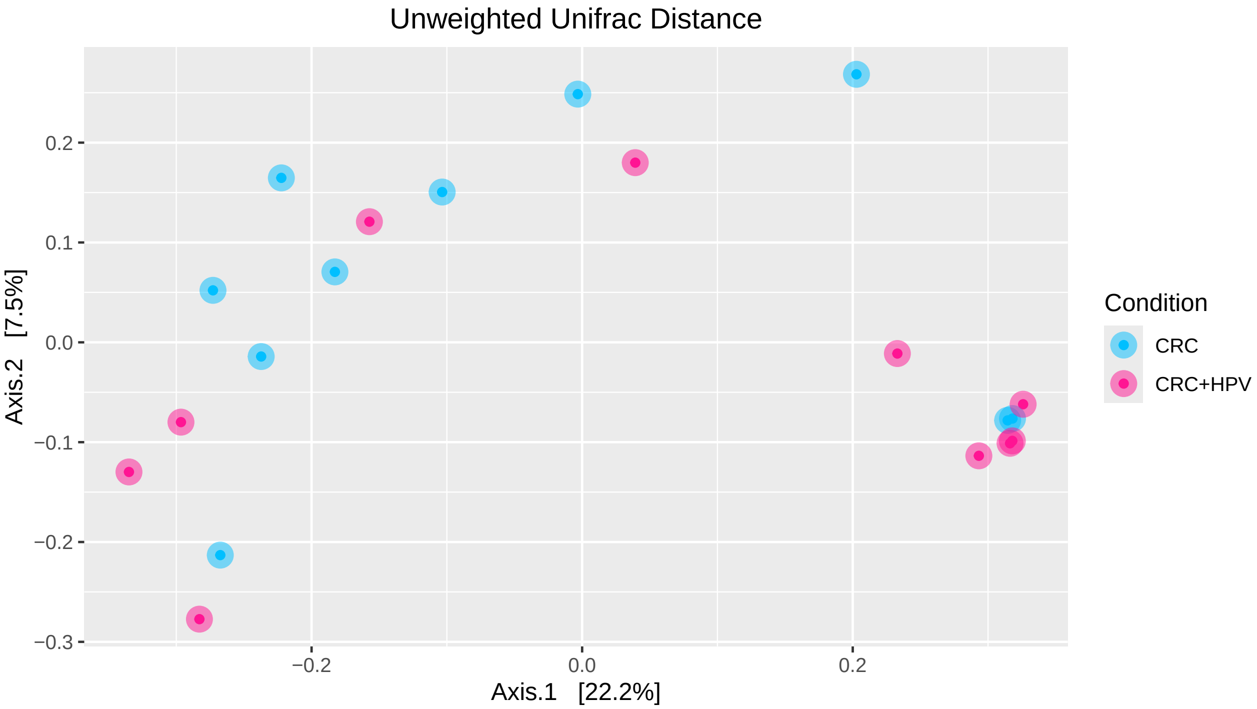


Figure 5. Unifrac Distance, Unweighted

```r
gse_pcoa_bray <- ordinate(gse_physeq, "PCoA", "bray")
ordination_plotting(gse_physeq, gse_pcoa_bray) +
  labs(title = "Bray-Curtis Distance")
```

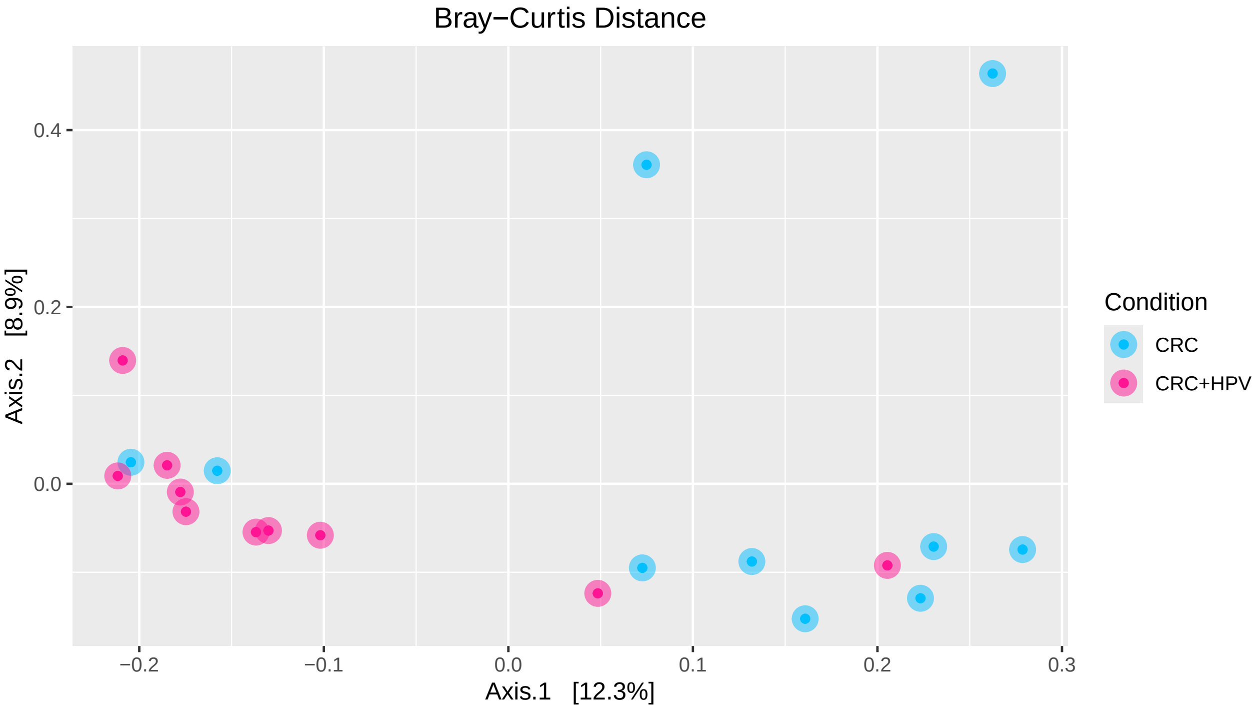


Figure 6. Bray-Curtis Distance

10.1 ANOSIM

ANOSIM (Analysis of Similarities) is a non-parametric statistical test used in ecology to compare the similarity of groups of samples. It assesses whether the difference between groups is significantly greater than the difference within groups. The R package **vegan** has a function to perform ANOSIM, aptly named **anosim**. The analysis can be custom tailored in a number of ways, including a specific distance matrix to be used (the **distance** method used is from the Phyloseq library, however). Here, ANOSIM is performed using 4 different distance measures (Weighted Unifrac, Unweighted Unifrac, Bray-Curtis, and Jaccard).

```
gse_wu <- distance(gse_physeq, "wunifrac")
gse_uu <- distance(gse_physeq, "uunifrac")
gse_bray <- distance(gse_physeq, "bray")
gse_jaccard <- distance(gse_physeq, "jaccard")

anosim(gse_wu, sample_information$condition)$signif
```

```
## [1] 0.136
```

```
anosim(gse_uu, sample_information$condition)$signif
```

```
## [1] 0.265
```

```
anosim(gse_bray, sample_information$condition)$signif
```

```
## [1] 0.005
```

```
anosim(gse_jaccard, sample_information$condition)$signif
```

```
## [1] 0.004
```

Both UniFrac distances were not significantly different between groups. Bray-Curtis Dissimilarity and Jaccard Distance both had $p < 0.01$, however. This indicates that the latter 2 tests determined the population differences between the two groups were higher than within them.

10.2 PERMANOVA

PERMANOVA (Permutational Multivariate Analysis of Variance) is also a non-parametric test which determines the degree to which groups of samples share certain characteristics. However, PERMANOVA is more focused on comparing group values to the overall means of the data. PERMANOVA tests the null hypothesis that there are no differences in multivariate dispersion and location among the groups. It can be implemented by vegan using the **adonis2** method.

A distance matrix (such as those calculated in the previous step) is used as the left hand side of the formula for the function. The right hand side is the condition or variable type that is to be compared. Lastly, metadata for the samples being studied is necessary.

```
adonis2(gse_wu ~ condition, data = sample_information)
```

```
## Permutation test for adonis under reduced model
## Terms added sequentially (first to last)
## Permutation: free
## Number of permutations: 999
##
## adonis2(formula = gse_wu ~ condition, data = sample_information)
##           Df SumOfSqs      R2      F Pr(>F)
## condition  1 0.016515 0.07369 1.4319  0.098 .
## Residual  18 0.207603 0.92631
## Total     19 0.224118 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
adonis2(gse_uu ~ condition, data = sample_information)
```

```
## Permutation test for adonis under reduced model
## Terms added sequentially (first to last)
## Permutation: free
## Number of permutations: 999
##
## adonis2(formula = gse_uu ~ condition, data = sample_information)
##           Df SumOfSqs      R2      F Pr(>F)
## condition  1   0.3277 0.05643 1.0764  0.246
## Residual  18   5.4793 0.94357
## Total     19   5.8070 1.00000
```

```r
adonis2(gse_bray ~ condition, data = sample_information)
```

```
## Permutation test for adonis under reduced model
## Terms added sequentially (first to last)
## Permutation: free
## Number of permutations: 999
##
## adonis2(formula = gse_bray ~ condition, data = sample_information)
##           Df SumOfSqs      R2     F Pr(>F)
## condition  1   0.4122 0.07843 1.532  0.004 **
## Residual  18   4.8427 0.92157
## Total     19   5.2548 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
adonis2(gse_jaccard ~ condition, data = sample_information)
```

```
## Permutation test for adonis under reduced model
## Terms added sequentially (first to last)
## Permutation: free
## Number of permutations: 999
##
## adonis2(formula = gse_jaccard ~ condition, data = sample_information)
##           Df SumOfSqs      R2      F Pr(>F)
## condition  1   0.4688 0.06817 1.3168  0.007 **
## Residual  18   6.4076 0.93183
## Total     19   6.8764 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The PERMANOVA results follow the same trend that is seen in ANOSIM - both UniFrac measures returned non-significant differences between the groups, while Bray-Curtis & Jaccard had $p < 0.01$.

11 *Lactobacillus* Analysis

```r
# Filter physeq object by microbes belonging to Lactobacillus
gse_lactobacillus <- subset_taxa(gse_physeq, Genus == "Lactobacillus")

# Create data frame & filter items with zero values
gse_lactobacillus_data <- psmelt(gse_lactobacillus) %>% filter(Abundance > 0)
```

```r
plotbox(gse_lactobacillus_data) +
  labs(title = "Abundance of Lactobacillus in Patient Samples",
       y = "Frequency of Lactobacillus")
```

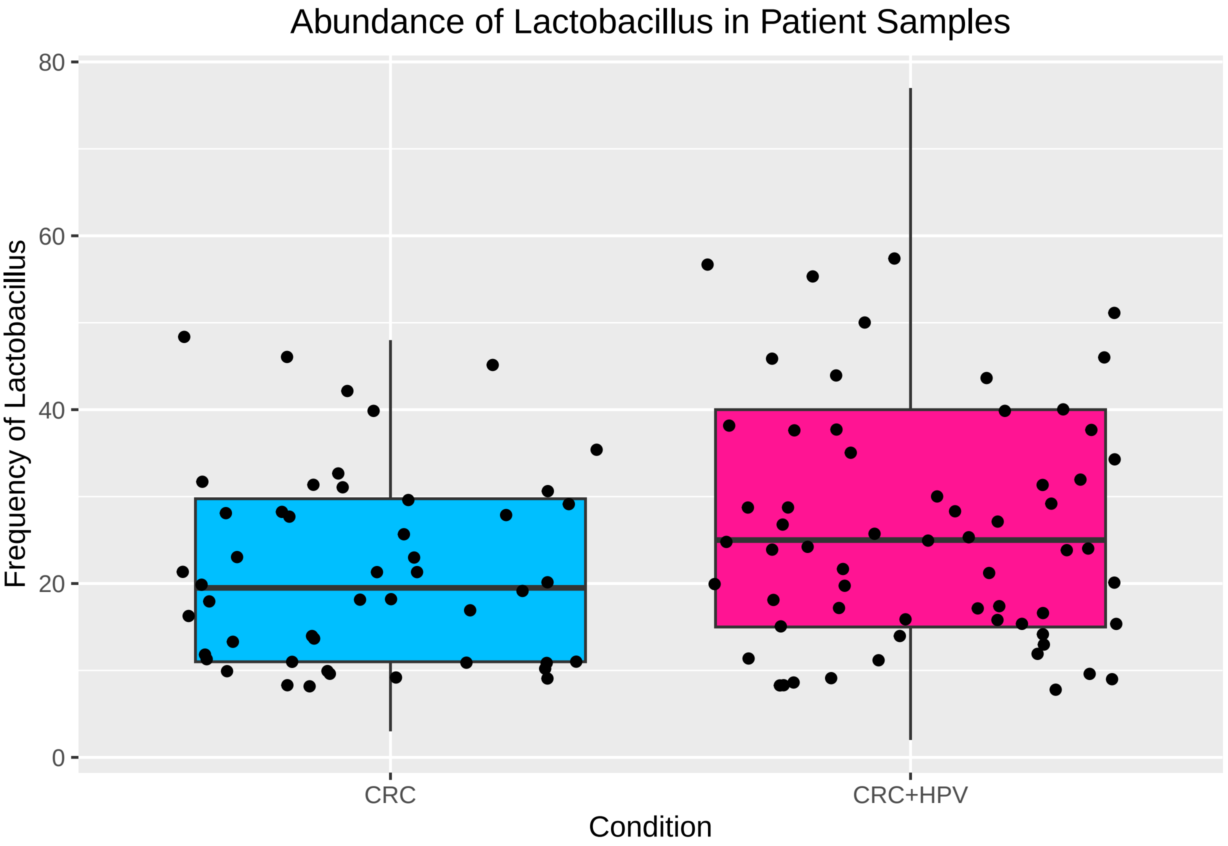


Figure 7. Comparison of Lactobacillus between groups

```r
# Create groups from original data frame
lac_crc <- gse_lactobacillus_data %>% filter(condition == "CRC")
lac_crc_hpv <- gse_lactobacillus_data %>% filter(condition == "CRC+HPV")

# Use function to calculate values
groupavg(lac_crc, lac_crc_hpv)
```

```
## CRC Average = 22.5555555555556   CRC + HPV Average = 31.9350649350649
##  T-Test, p = 0.00883644761675433
```

The results indicate a significantly higher number of *Lactobacillus* in the CRC+HPV group.

12 *Anaerococcus* Analysis

```
# Filter physeq object by microbes belonging to Anaerococcus
gse_anaerococcus <- subset_taxa(gse_physeq, Genus == "Anaerococcus")

# Create data frame & filter items with zero values
gse_anaerococcus_data <- psmelt(gse_anaerococcus) %>% filter(Abundance > 0)
```

```
plotbox(gse_anaerococcus_data) +
  labs(title = "Abundance of Anaerococcus in Patient Samples",
       y = "Frequency of Anaerococcus")
```

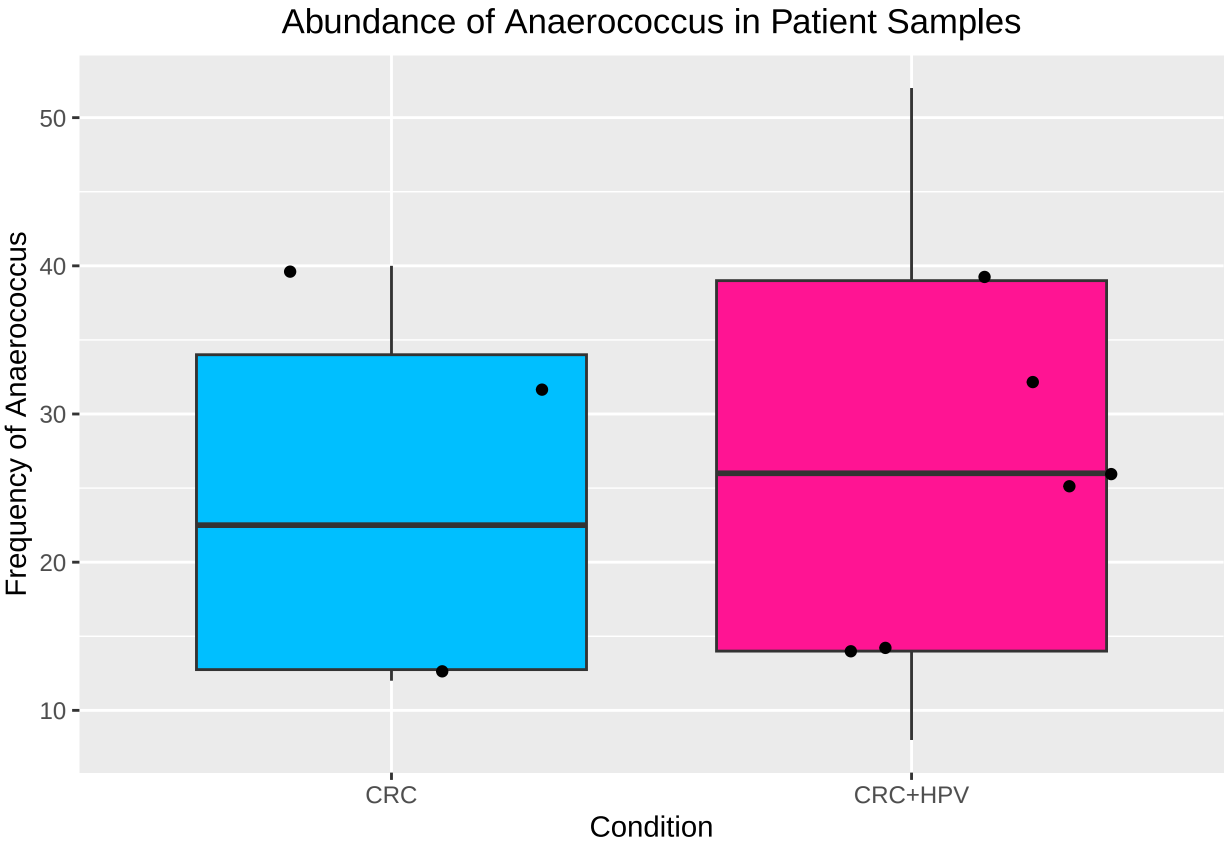


Figure 8. Comparison of Anaerococcus between groups

```
# Create groups from original data frame
ana_crc <- gse_anaerococcus_data %>% filter(condition == "CRC")
ana_crc_hpv <- gse_anaerococcus_data %>% filter(condition == "CRC+HPV")

groupavg(ana_crc, ana_crc_hpv)
```

```
## CRC Average = 24.25   CRC + HPV Average = 29
##  T-Test, p = 0.606006655234261
```

The average number of *Anaerococcus* in the CRC+HPV group is higher.

However, by t-test the difference is insignificant.

13 *Turicibacter* Analysis

```
# Filter physeq object by microbes belonging to Turicibacter
gse_turicibacter <- subset_taxa(gse_physeq, Genus == "Turicibacter")

# Create data frame & filter items with zero values
gse_turicibacter_data <- psmelt(gse_turicibacter) %>% filter(Abundance > 0)
```

```
plotbox(gse_turicibacter_data) +
  labs(title = "Abundance of Turicibacter in Patient Samples",
       y = "Frequency of Turicibacter")
```

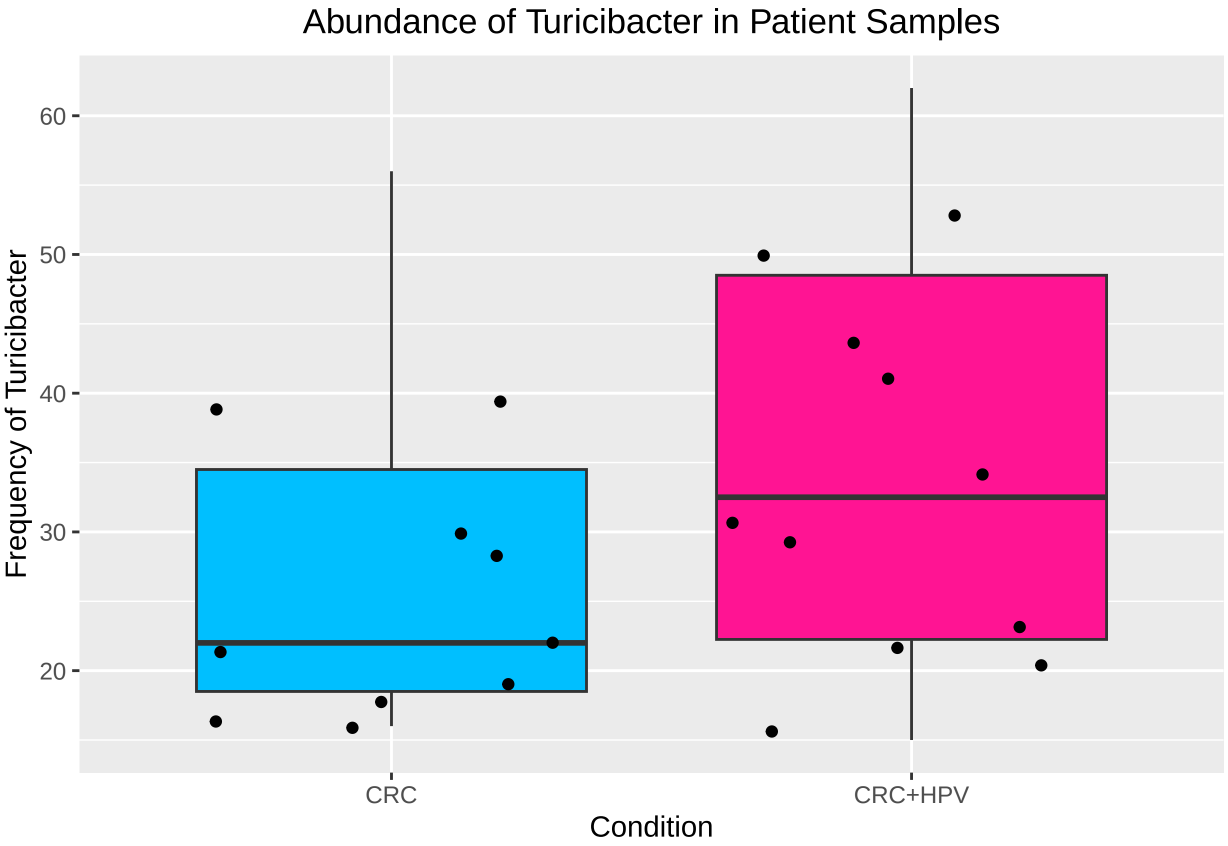


Figure 9. Comparison of Turicibacter between groups

```
# Create groups from original data frame
tub_crc <- gse_turicibacter_data %>% filter(condition == "CRC")
tub_crc_hpv <- gse_turicibacter_data %>% filter(condition == "CRC+HPV")

groupavg(tub_crc, tub_crc_hpv)
```

```
## CRC Average = 27.6363636363636   CRC + HPV Average = 38.6428571428571
##  T-Test, p = 0.14340286917177
```

The average number of *Turicibacter* in the CRC+HPV group is higher.

However, by t-test the difference is insignificant.

14 *Peptostreptococcus* Analysis

```
# Filter physeq object by microbes belonging to Peptostreptococcus
gse_peptostreptococcus <- subset_taxa(gse_physeq, Genus == "Peptostreptococcus")

# Create data frame & filter items with zero values
gse_peptostreptococcus_data <- psmelt(gse_peptostreptococcus) %>% filter(Abundance > 0)
```

```
plotbox(gse_peptostreptococcus_data) +
  labs(title = "Abundance of Peptostreptococcus in Patient Samples",
       y = "Frequency of Peptostreptococcus")
```

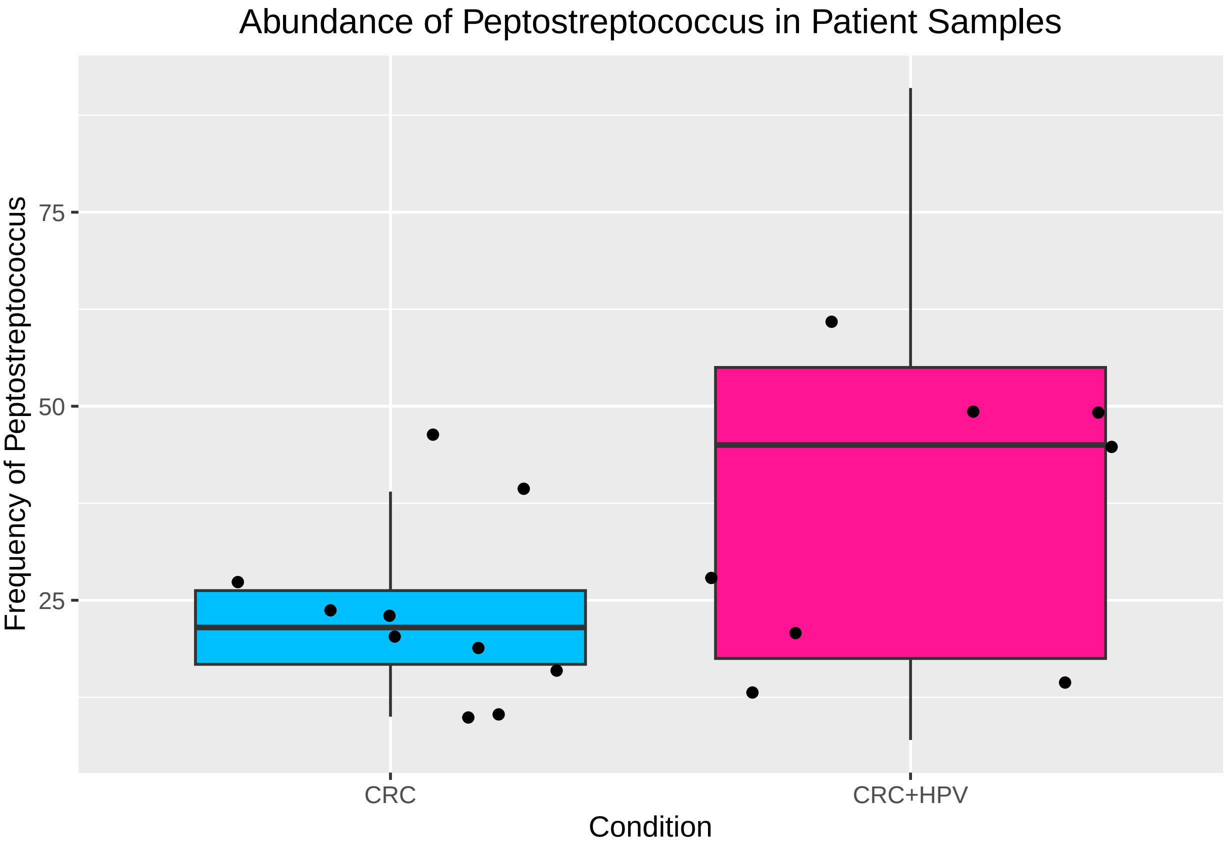


Figure 10. Comparison of Peptostreptococcus between groups

```
# Create groups from original data frame
pep_crc <- gse_peptostreptococcus_data %>% filter(condition == "CRC")
pep_crc_hpv <- gse_peptostreptococcus_data %>% filter(condition == "CRC+HPV")

groupavg(pep_crc, pep_crc_hpv)
```

```
## CRC Average = 23.4   CRC + HPV Average = 41.9090909090909
##  T-Test, p = 0.0672461043334317
```

The average number of *Peptostreptococcus* in the CRC+HPV group is higher.

However, by t-test the difference is, just slightly, insignificant.

15 *Ruminococcus* Analysis

```
# Filter physeq object by microbes belonging to Ruminococcus
gse_ruminococcus <- subset_taxa(gse_physeq, Genus == "Ruminococcus")

# Create data frame & filter items with zero values
gse_ruminococcus_data <- psmelt(gse_ruminococcus) %>% filter(Abundance > 0)
```

```
plotbox(gse_ruminococcus_data) +
  labs(title = "Abundance of Ruminococcus in Patient Samples",
       y = "Frequency of Ruminococcus")
```

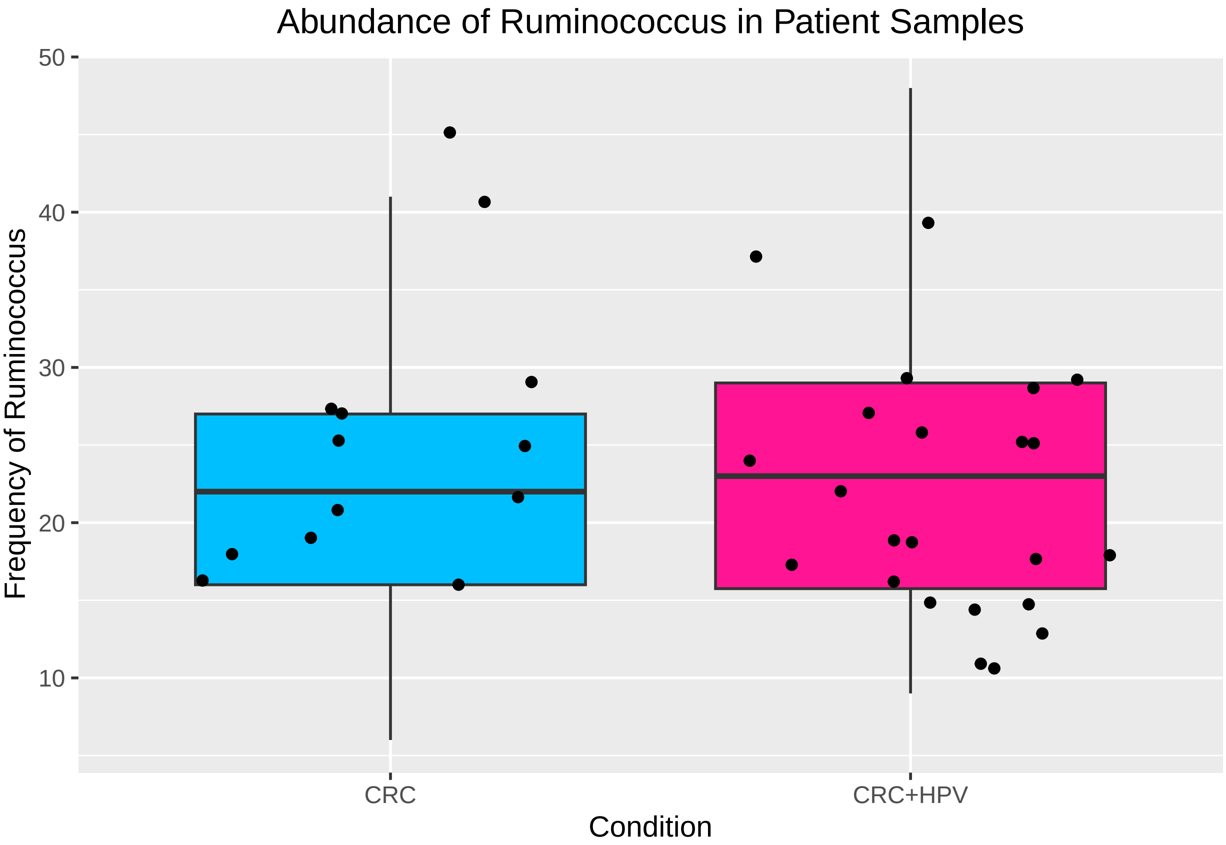


Figure 11. Comparison of Ruminococcus between groups

```
# Create groups from original data frame
rum_crc <- gse_ruminococcus_data %>% filter(condition == "CRC")
rum_crc_hpv <- gse_ruminococcus_data %>% filter(condition == "CRC+HPV")

groupavg(rum_crc, rum_crc_hpv)
```

```
## CRC Average = 24.4117647058824   CRC + HPV Average = 26.0714285714286
##  T-Test, p = 0.710846598304479
```

The average number of *Ruminococcus* in the CRC+HPV group is higher.

However, by t-test the difference is insignificant.

16 Total Microbiome Difference

```
data_gse <- gse_phydata %>% filter(Abundance > 100 & Phylum != "NA" & Class != "NA")

ggplot(data_gse, aes(x = condition, y = Abundance, fill = Class)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~Phylum, scales = "free_x") +
  labs(title = "Prevalence of Microbes in Subjects by Phylum & Class", x = NULL,
       y = "Abundance") +
  theme(legend.position = "bottom", legend.key.size = unit(0.5, "cm"),
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(size = 9, angle = 0, hjust = 0.5))
```

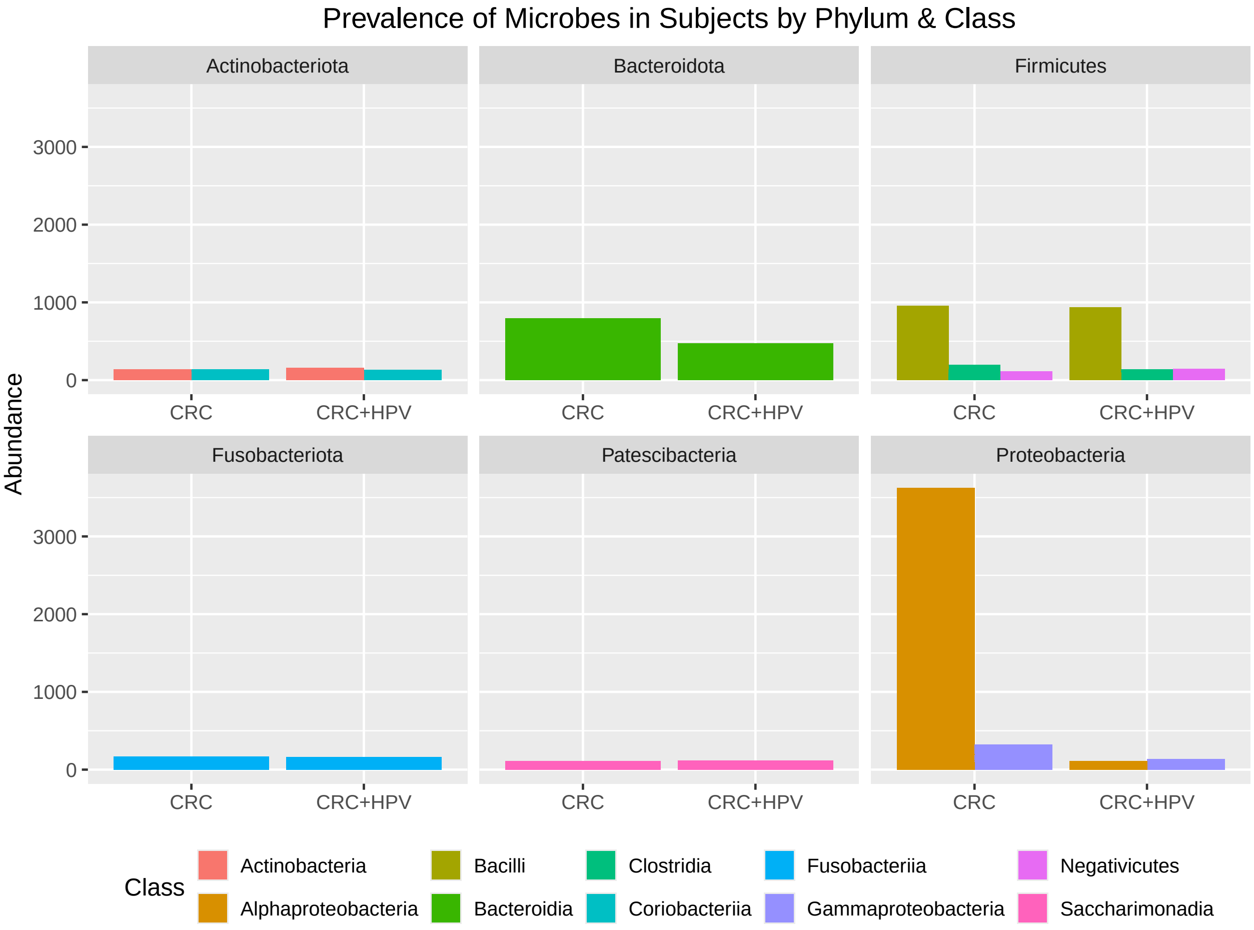


Figure 12. Taxonomic Overview of Experiment Data

To get a basic overview of the types of microbes present in the samples, the physeq data was filtered to remove rare species (only keeping entries with Abundance > 100) and unclassified taxa.

If the total microbiome is plotted by condition, Phylum, and Class, it can be seen immediately that HPV Negative patients have an enormous abundance of *Proteobacteria* (especially *Alphaproteobacteria*), while HPV Positive patients have a significantly lower number.

17 Conclusion

The intestinal microbiome is a complex community with immense numbers of bacterial populations belonging to a wide variety of taxa. These populations can be affected by a number of factors, including infection and cancer. This analysis endeavored to compare the intestinal microbiota of patients with CRC, and patients with concomitant HPV infection. For all three Alpha Diversity metrics tested, the CRC+HPV group demonstrated higher community richness and evenness. The Beta Diversity metrics tested returned inconsistent results. While UniFrac measures did not identify significant differences between the groups, both Bray-Curtis and Jaccard did.

Five Genus level taxa were compared between the two groups, for all of these, abundance levels were higher in the CRC+HPV patients. Upon reviewing the Phylum level frequencies, it was noted that the CRC group contained far more microbes belonging to a single class (*Alphaproteobacteria*). CRC patients also had higher numbers of Class *Bacteroidia*.

While the data does not ubiquitously support the claim that CRC patients with HPV have a more diverse and balanced intestinal microbiota than those without HPV, many metrics calculated establish strong evidence. In addition, an article by Ambrosio *et al.* reports that Alpha Diversity measures were indeed higher in patient samples with CRC+HPV compared to CRC alone. The study also identified decreased *Bacteroides* levels in patients with CRC+HPV.

18 Session Info

This command will output details about the R environment used to produce this document:

```r
sessionInfo()
```

```
## R version 4.3.3 (2024-02-29)
## Platform: x86_64-apple-darwin20 (64-bit)
## Running under: macOS Monterey 12.7.4
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.3-x86_64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.3-x86_64/Resources/lib/libRlapack.dylib;  LAPACK
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: America/New_York
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] microbiome_1.24.0 phyloseq_1.46.0   vegan_2.6-4       lattice_0.22-6
##  [5] permute_0.9-7     qiime2R_0.99.6    gt_0.10.1         here_1.0.1
##  [9] readxl_1.4.3      lubridate_1.9.3   forcats_1.0.0     stringr_1.5.1
## [13] dplyr_1.1.4       purrr_1.0.2       readr_2.1.5       tidyr_1.3.1
## [17] tibble_3.2.1      ggplot2_3.5.0     tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] bitops_1.0-7            gridExtra_2.3           rlang_1.1.3
##  [4] magrittr_2.0.3          ade4_1.7-22             compiler_4.3.3
##  [7] mgcv_1.9-1              vctrs_0.6.5             reshape2_1.4.4
## [10] pkgconfig_2.0.3         crayon_1.5.2            fastmap_1.1.1
## [13] backports_1.4.1         XVector_0.42.0          labeling_0.4.3
## [16] utf8_1.2.4              rmarkdown_2.26          tzdb_0.4.0
## [19] xfun_0.43               zlibbioc_1.48.2         GenomeInfoDb_1.38.8
## [22] jsonlite_1.8.8          biomformat_1.30.0       highr_0.10
## [25] rhdf5filters_1.14.1     Rhdf5lib_1.24.2         parallel_4.3.3
## [28] cluster_2.1.6           R6_2.5.1                stringi_1.8.3
## [31] zCompositions_1.5.0-3   rpart_4.1.23            cellranger_1.1.0
## [34] Rcpp_1.0.12             iterators_1.0.14        knitr_1.45
## [37] base64enc_0.1-3         IRanges_2.36.0          Matrix_1.6-5
## [40] splines_4.3.3           nnet_7.3-19             igraph_2.0.3
## [43] timechange_0.3.0        tidyselect_1.2.1        rstudioapi_0.16.0
## [46] yaml_2.3.8              codetools_0.2-20        plyr_1.8.9
## [49] Biobase_2.62.0          withr_3.0.0             Rtsne_0.17
## [52] evaluate_0.23           foreign_0.8-86          survival_3.5-8
## [55] xml2_1.3.6              Biostrings_2.70.3       pillar_1.9.0
## [58] DT_0.33                 checkmate_2.3.1         foreach_1.5.2
## [61] stats4_4.3.3            NADA_1.6-1.1            generics_0.1.3
## [64] rprojroot_2.0.4         RCurl_1.98-1.14         truncnorm_1.0-9
## [67] S4Vectors_0.40.2        hms_1.1.3               munsell_0.5.1
## [70] scales_1.3.0            glue_1.7.0              Hmisc_5.1-2
## [73] tools_4.3.3             data.table_1.15.4       rhdf5_2.46.1
## [76] grid_4.3.3              ape_5.7-1               colorspace_2.1-0
## [79] nlme_3.1-164            GenomeInfoDbData_1.2.11 htmlTable_2.4.2
## [82] Formula_1.2-5           cli_3.6.2               fansi_1.0.6
## [85] gtable_0.3.4            digest_0.6.35           BiocGenerics_0.48.1
## [88] farver_2.1.1            htmlwidgets_1.6.4       htmltools_0.5.8.1
## [91] multtest_2.58.0         lifecycle_1.0.4         MASS_7.3-60.0.1
```

19 References

QIIME2

Bolyen, E., Rideout, J.R., Dillon, M.R. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* **37**, 852–857 (2019). https://doi.org/10.1038/s41587-019-0209-9

Silva 138 Classifier - Download

MD5: e05afad0fe87542704be96ff483824d4

Bokulich, N.A., Kaehler, B.D., Rideout, J.R. *et al.* Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* **6**, 90 (2018). https://doi.org/10.1186/s40168-018-0470-z

Balvočiūtė, M., Huson, D.H. SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare?. *BMC Genomics* **18** (Suppl 2), 114 (2017). https://doi.org/10.1186/s12864-017-3501-4

Odom, A.R., Faits, T., Castro-Nallar, E. *et al.* Metagenomic profiling pipelines improve taxonomic classification for 16S amplicon sequencing data. *Sci Rep* **13**, 13957 (2023). https://doi.org/10.1038/s41598-023-40799-x

Velda J. Gonzalez-Mercado, Jean Lim, Leorey N. Saligan, Nicole Perez, Carmen Rodriguez, Raul Bernabe, Samia Ozorio, Elsa Pedro, Farrah Sepehri, Brad Aouizerat, "Gut Microbiota and Depressive Symptoms at the End of CRT for Rectal Cancer: A Cross-Sectional Pilot Study”, *Depression Research and Treatment*, vol. 2021, Article ID 7967552, 10 pages, 2021. https://doi.org/10.1155/2021/7967552

Ambrosio, M.R., Niccolai, E., Petrelli, F. *et al.* Immune landscape and oncobiota in HPV-Associated Colorectal Cancer: an explorative study. *Clin Exp Med* 23, 5101–5112 (2023). https://doi.org/10.1007/s10238-023-01165-3

qiime2R

Phyloseq

microbiome

vegan