Ecology and Comparative Analysis of the Gut Microbiome of Cancer Patients Against Healthy Controls

The gut microbiome plays a crucial role in maintaining homeostasis by degrading complex fibers, synthesizing vitamins, modulating the immune system, and preserving gut integrity through mutualistic interactions with the host, as well as the production of metabolites.[1] Patients with acute inflammatory diseases such as IBD, UC, and Crohn's disease exhibit significant differences in the composition and diversity of their gut microbiome compared to healthy individuals.[2] These variations in microbial population and diversity between healthy and diseased individuals are collectively referred to as gut dysbiosis.[3,4] A strong link exists between gut dysbiosis and overall health.[2,5–7] However, the mechanisms by which gut microbiome interacts with the host to contribute to disease phenotypes remain poorly understood. This critical knowledge gap hinders the development of effective treatments for inflammatory bowel disease, highlighting the urgent need to investigate the functional roles of specific bacterial taxa and their interactions with the host.

Gut dysbiosis is a common phenotype observed in IBD, neurodegenerative diseases, and colorectal cancer.[4,8] For instance, the gut microbiome of IBD patients undergoes significant shifts, with an increased presence of aerotolerant bacterial pathogens.[9] The Human Microbiome Project (HMP) established a genetic reference database for gut bacterial taxa, enabling comparative analyses of the gut microbiome in healthy and diseased individuals. Given the rising incidence of colorectal cancer and pancreatic cancer and its association with changes in the gut microbiome, there have been increased efforts to study the relationship between cancer and gut dysbiosis and possibly identify and characterize the effects of bacterial pathobionts in cancer progression.[6]

The overall goal of this project is to investigate the differences in the gut microbiome of pancreatic cancer patients versus healthy individuals. *Our central hypothesis* is that the gut microbiome of pancreatic cancer patients is different in composition and abundance than a healthy individual. We will carry out a meta-analysis of already existing clinical data to do a comparative taxonomic analysis of the gut microbiome of healthy and cancer patients. The rationale for this project is that while lots of studies have shown significant differences in the gut microbiome of patients suffering from colorectal cancer versus healthy individuals, little is known about the pancreatic cancer gut microbiome. Thus, this study aims to identify possible bacterial pathobionts that are associated with pancreatic cancer and possibly use them as biomarkers for early diagnosis of pancreatic cancer.

Aim 1: To investigate the comparative differences in the gut microbiome of pancreatic cancer patients versus healthy individuals. This aims to give us an idea t the phylum level, the differences in the bacterial composition of cancer and healthy individuals. We will use data from the European Nucleotide Archive (project ID: 994901, Accession no..: PRJNA994901, Project title: Healthy human gut microbiome subjected to B9 and B12 treatment, and NCBI with study no: PRJNA542319).

Aim 2: Identify bacterial pathobionts highly abundant in pancreatic cancer patients. We will profile the top species of bacteria that are found in pancreatic cancer patients.

We will utilize R Data analytic tool (version 4.2) and my conversation with Chat GPT 4 to generate codes in R and data.

Results

R-Scripts used for Data Analysis.

1. **Comparative Gut microbial abundance in healthy and pancreatic cancer patients**

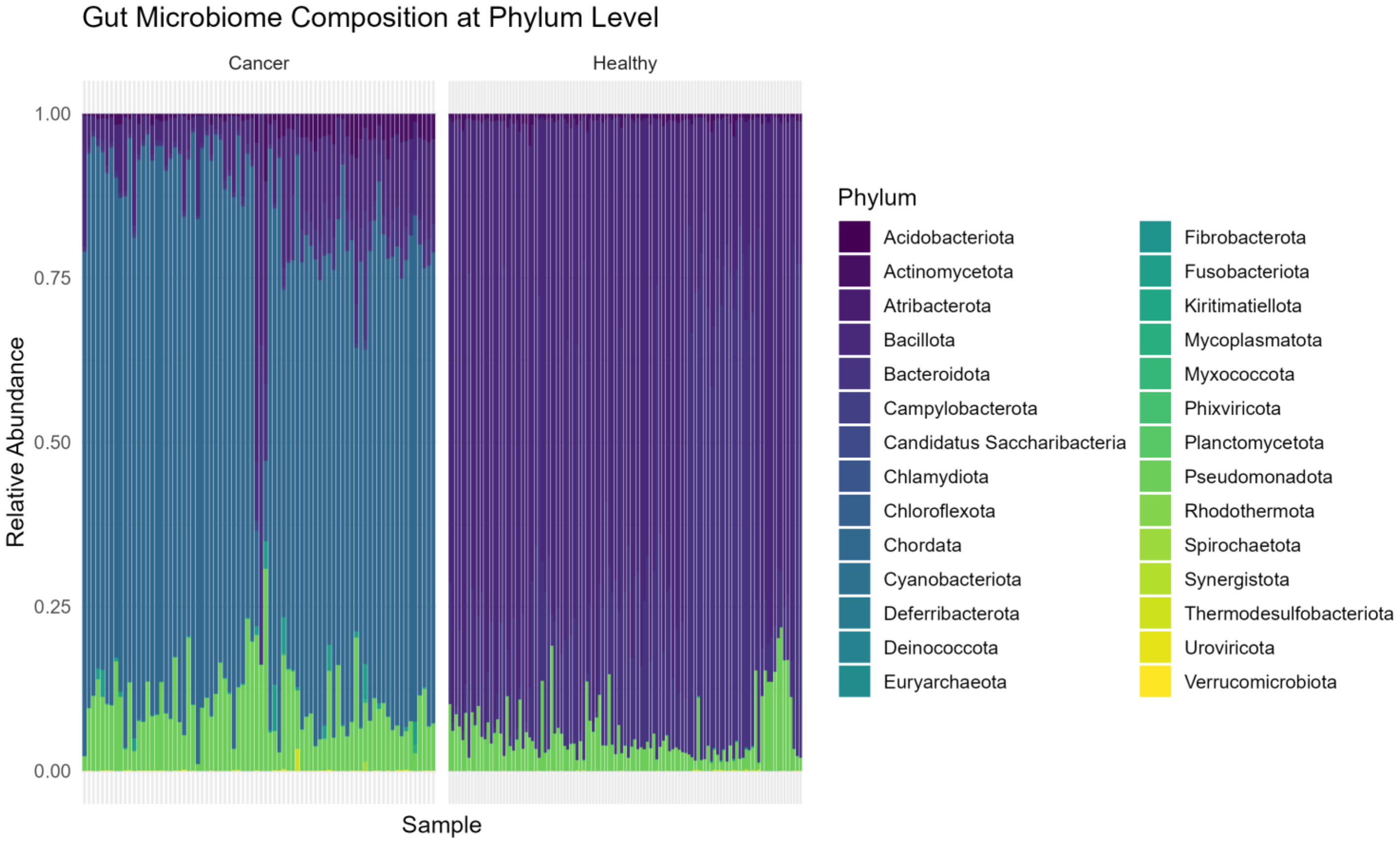

Fig.1: Comparative abundance of the gut microbiome of healthy VS Cancer patients

R Scripts

```
# Load libraries
library(readr)
library(dplyr)
library(tidyr)
library(ggplot2)
library(viridis)
```

```
# Step 1: Read data (phylum in rows, samples in columns)
hh_url<-
"https://raw.githubusercontent.com/okpecallistus/Uchenna5202stuff/refs/heads/main/HH_combined_bracken_phylum_fraction.csv"
pc_url<-
"https://raw.githubusercontent.com/okpecallistus/Uchenna5202stuff/refs/heads/main/PC_combined_bracken_phylum_fraction.csv"
hh_phylum <- read.csv(hh_url, row.names = 1, check.names = FALSE)
pC_phylum <- read.csv(pc_url, row.names = 1, check.names = FALSE)
```

```
# Step 2: Transpose and reshape to long format

hh_long <- hh_phylum %>%
  t() %>%
  as.data.frame() %>%
  mutate(Sample = rownames(.), Group = "Healthy") %>%
  pivot_longer(-c(Sample, Group), names_to = "Phylum", values_to = "Abundance")
pc_long <- pC_phylum %>%
  t() %>%
  as.data.frame() %>%
  mutate(Sample = rownames(.), Group = "Cancer") %>%
  pivot_longer(-c(Sample, Group), names_to = "Phylum", values_to = "Abundance")
```

```
# Step 3: Combine datasets

long_data <- bind_rows(hh_long, pc_long)
 group_by(Sample) %>%
  mutate(Abundance = Abundance / sum(Abundance, na.rm = TRUE)) %>%
  ungroup()
#Plot the bar graph
bar_plot<-ggplot(long_data_norm, aes(x = Sample, y = Abundance, fill = Phylum))
+
  geom_bar(stat = "identity") +
  facet_wrap(~ Group, scales = "free_x") +
  theme_minimal(base_size = 12) +
  labs(title = "Gut Microbiome Composition at Phylum Level",
       y = "Relative Abundance", x = "Sample") +
  theme(axis.text.x = element_blank(), axis.ticks.x = element_blank()) +
  scale_fill_viridis_d(option = "D")
print(bar_plot)
ggsave("phylum_barplot.png", bar_plot, width = 10, height = 6, dpi = 300)
```

2. Ecological properties of the Gut microbiome in Healthy and Cancer patients(α-diversity)

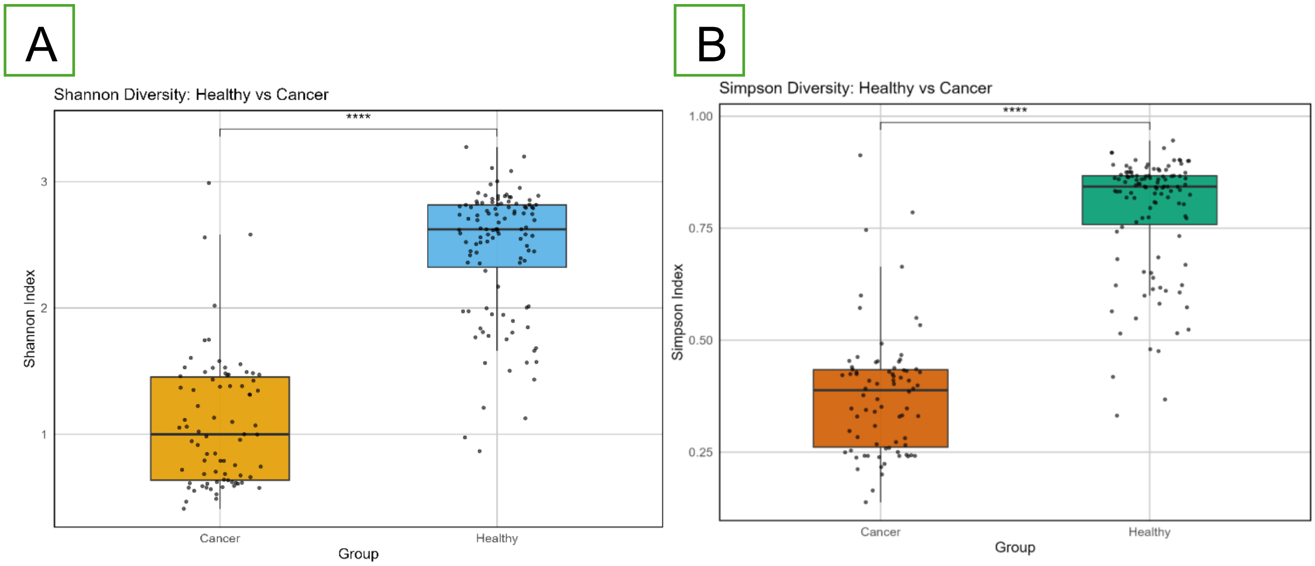

Fig. 2: Alpha diversity of gut microbiome for healthy and cancer subjects. A) Shannon diversity index. B) Simpson diversity index.

R Scripts:

```
#install.packages(c("readxl", "vegan", "ggplot2", "dplyr",
"ggpubr"))
library(readxl)
library(vegan)
library(ggplot2)
library(dplyr)
library(ggpubr)
view(diversity_df)
view(pC_diversity_df)D
```

```
#Read my data and assign them names

hh_species<- (Healthy individuals)
pC_species<- (cancer patients)
#Transpose data
pC_species_t<-t(pC_species)
hh_species_t<-t(hh_species)
```

```
#Convert data to numeric
Creat a new data frame, retaining column and row names
hh_species_t_numeric<-as.data.frame(hh_species_t)
#use lapply to convert each column to numeric while preserving names
```

```
#Compute Diversity Indices
library(vegan)
pC_shannon_index_sp<-diversity(pC_species_t_numeric, index = "shannon")
pC_simpson_index_sp<-diversity(pC_species_t_numeric, index = "simpson")
shannon_index_species<-diversity(hh_species_t_numeric,index = "shannon")
simpson_index_species<-diversity(hh_species_t_numeric, index = "simpson")
#Prepare Data for Plotting
pC_diversity_df<-data.frame(SampleID = rownames(pC_species_t_numeric),Shannon = 
pC_shannon_index_sp, Simpson = pC_simpson_index_sp)
view(pC_diversity_df)
diversity_df <- data.frame(
  SampleID = rownames(hh_species_t_numeric),
  Shannon = shannon_index_species,
  Simpson = simpson_index_species)
```

```
#Group Data

pC_diversity_df$Group<-"Cancer"

diversity_df$Group<-"Healthy"

#Combine Both Datasets

combined_div<-rbind(diversity_df, pC_diversity_df)

view(combined_div)

#Plot: Simpson BOXplot with Statistical Test

library(ggplot2)

library(ggpubr)
```

```
#Plot Simpson Graph
p_simpson <- ggplot(combined_div, aes(x = Group, y = Simpson, fill = Group)) +
  geom_boxplot(width = 0.5, alpha = 0.9, outlier.shape = NA) +
  geom_jitter(width = 0.15, alpha = 0.6, size = 1, color = "black") +
# Statistical test with bar and asterisks
  stat_compare_means(method = "wilcox.test", label = "p.signif",
                     comparisons = list(c("Healthy", "Cancer")),
                     tip.length = 0.02, size = 5) +
  # Apply a clean theme and add gridlines + border
  theme_minimal(base_size = 12) +
  theme( panel.grid.major = element_line(color = "grey80"),
    panel.grid.minor = element_blank(),
    panel.border = element_rect(color = "black", fill = NA, size = 1),
    axis.text = element_text(size = 11), axis.title = element_text(size = 13), legend.position = "none") +
  labs(title = "Simpson Diversity: Healthy vs Cancer", x = "Group",
    y = "Simpson Index" ) +
  # Color-blind friendly & elegant colors (Okabe-Ito palette)
  scale_fill_manual(values = c("Healthy" = "#009E73", "Cancer" = "#D55E00"))
print(p_simpson)ggsave("combined_Simpson_Diversity_Boxplot.png", plot = p_simpson, width = 8, height
= 6, dpi = 300)
# Wilcoxon rank-sum test (non-parametric)
wilcox.test(Simpson ~ Group, data = combined_div)
```

```
#Copute Shanono index:
p_shannon <- ggplot(combined_div, aes(x = Group, y = Shannon, fill = Group)) +
  geom_boxplot(width = 0.5, alpha = 0.9, outlier.shape = NA) +
  geom_jitter(width = 0.15, alpha = 0.6, size = 1, color = "black") +
  # Statistical test with bar and asterisks
  stat_compare_means(method = "wilcox.test", label = "p.signif",
  comparisons = list(c("Healthy", "Cancer")),  tip.length = 0.02, size = 5) +
  # Apply a clean theme and add gridlines + border
 theme_minimal(base_size = 12) + theme(panel.grid.major = element_line(color =
"grey80"),panel.grid.minor = element_blank(), panel.border = element_rect(color = "black", fill = NA, size =
1),axis.text = element_text(size = 11), axis.title = element_text(size = 13),
    legend.position = "none" ) +
  labs( title = "Shannon Diversity: Healthy vs Cancer",
    x = "Group",  y = "Shannon Index") +
    # Color-blind friendly & elegant colors (Okabe-Ito palette)
  scale_fill_manual(values = c("Healthy" = "#56B4E9", "Cancer" = "#E69F00"))
print(p_shannon)
ggsave("combined_Simpson_Diversity_Boxplot.png", plot = p_shannon, width = 8, height = 6, dpi = 300)
# Wilcoxon rank-sum test (non-parametric)
wilcox.test(Shannon ~ Group, data = combined_div)
```

3. Ecological properties of the Gut microbiome in Healthy and Cancer patients (β-diversity)

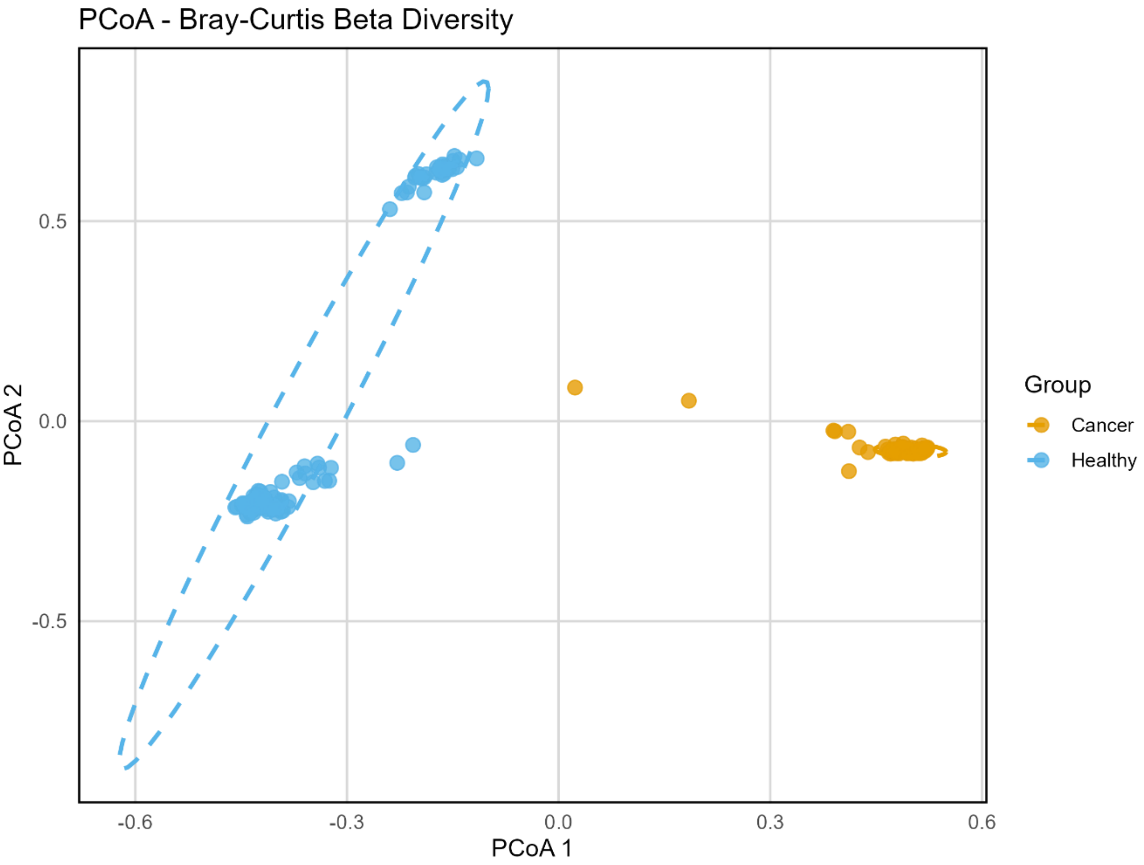

Fig. 3: β-diversity of gut microbiome for healthy and cancer subjects.

R Script:

```
#LOAD NECESSARY PACKAGES

library(readxl)

library(vegan)

library(ggplot2)

library(dplyr)

hh_df_sp<-as.data.frame(hh_species_t)

pC_df_sp<-as.data.frame(pC_species_t)

hh_df_sp$Group<-"Healthy"

view(hh_df_sp)

view(pC_diversity_df)
```

```
pC_df_sp$Group<-"Cancer"

#Combine and Clean Data

#First, since the species are different in the different datset, (Force matching is used)

#1: Get the union of all species names (columns)

all_species_beta<-union(colnames(hh_df_sp), colnames(pC_df_sp))

#Add missing Columns (species) with zeros to cancer data

missing_in_pC <- setdiff(all_species_beta, colnames(pC_df_sp))

pC_df_sp[missing_in_pC] <- 0

#Re-order columns to match correctly

hh_df_sp <- hh_df_sp[, all_species_beta]

pC_df_sp <- pC_df_sp[, all_species_beta]
```

```
# Fill missing species with zeros (not yet reordered)

missing_in_hh <- setdiff(all_species_beta, colnames(hh_df_sp))

missing_in_pc <- setdiff(all_species_beta, colnames(pC_df_sp))

hh_df_sp[missing_in_hh] <- 0

pC_df_sp[missing_in_pc] <- 0

#Order to Match

hh_df_sp <- hh_df_sp[, all_species_beta]

pC_df_sp <- pC_df_sp[, all_species_beta]

view(hh_df_sp)

#Group

pC_df_sp$Group<-"Cancer"

hh_df_sp$Group<-"Healthy"

#Combine

combined_betadiv_df <- rbind(hh_df_sp, pC_df_sp)

view(combined_betadiv_df)

hh_df_sp$Group <- "Healthy"

pC_df_sp$Group <- "Cancer"

view(hh_df_sp)
```

```
#After combining
group_labels <- combined_betadiv_df$Group
combined_betadiv_df$Group <- NULL  # remove Group column before calculating distances
view(group_labels)
#Calculate Bray-Cutis distance
# Make sure group labels are stored
group_labels <- c(rep("Healthy", nrow(hh_df_sp)), rep("Cancer", nrow(pC_df_sp)))
# Combine numeric species tables (already fixed earlier)
combined_betadiv_df <- rbind(hh_df_sp, pC_df_sp)
# Calculate Bray-Curtis distance matrix
bray_dist <- vegdist(combined_betadiv_df, method = "bray")
#Check data frame:
str(combined_betadiv_df)
#STore Groups
group_labels <- combined_betadiv_df$Group
```

```
#Remove group columns before computing distances

# Keep only numeric species data

combined_numeric_df <- combined_betadiv_df[, sapply(combined_betadiv_df, is.numeric)]

#Compute Bryacutis

library(vegan)

bray_dist <- vegdist(combined_numeric_df, method = "bray")

# Run PCoA (Principal Coordinates Analysis)

pcoa_result <- cmdscale(bray_dist, eig = TRUE, k = 2)
```

```
# Create a data frame for plotting

pcoa_df <- data.frame(

  SampleID = rownames(combined_betadiv_df),

  Dim1 = pcoa_result$points[, 1],

  Dim2 = pcoa_result$points[, 2],

  Group = group_labels)

library(ggplot2)
```

OK

```
# Create the plot and assign it to an object

pcoa_plot <- ggplot(pcoa_df, aes(x = Dim1, y = Dim2, color = Group)) +

  geom_point(size = 3, alpha = 0.8) +

  stat_ellipse(level = 0.95, linetype = "dashed", size = 1) +  # 95% CI ellipse
```

4. Comparative Taxonomic Analysis of Gut Microbiome in healthy VS Cancer subjects

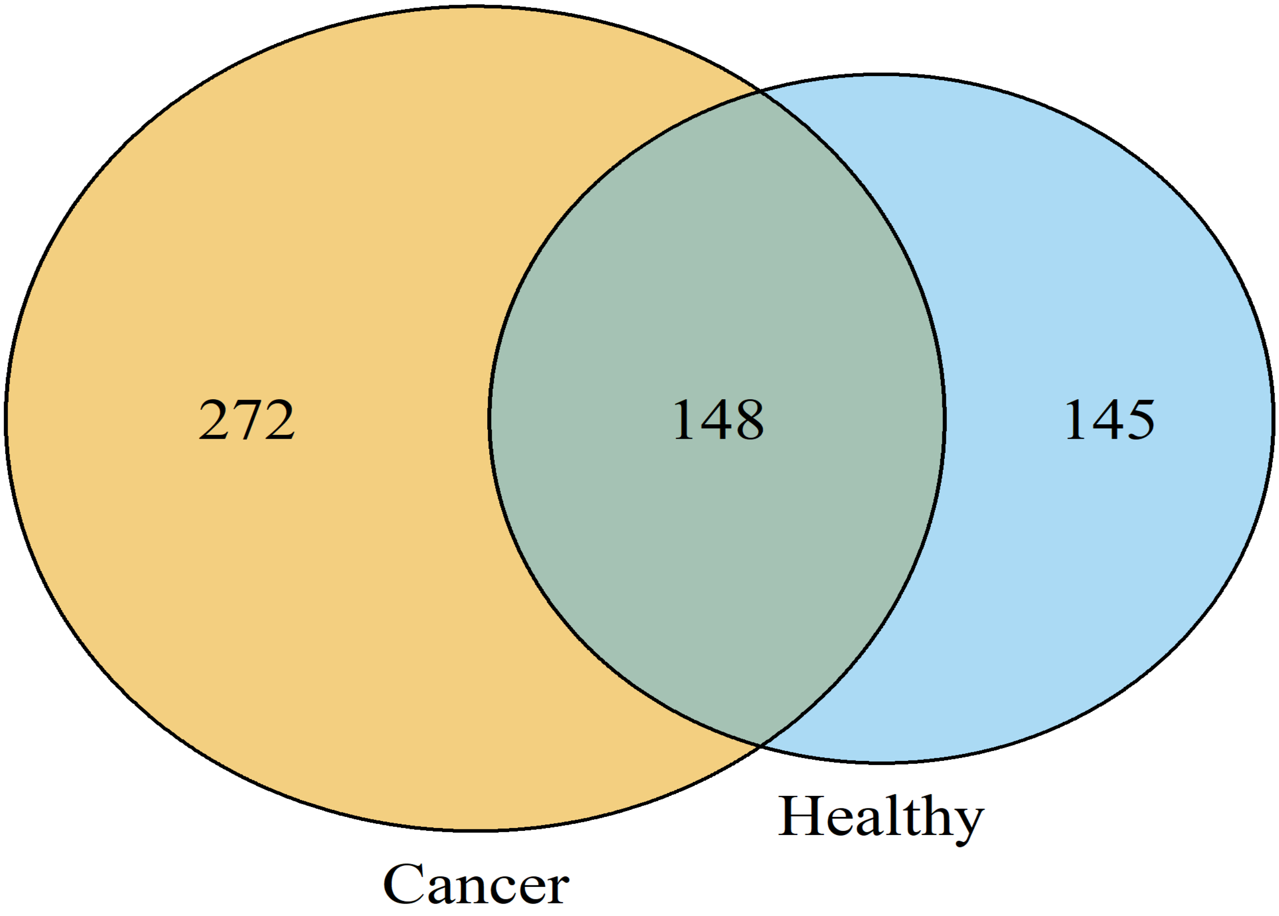

Fig. 4: Differential abundance of microbial species in health and disease state

R Scripts

Codes:

```
library(VennDiagram)

# Step 1: Identify numeric columns only

hh_numeric_cols <- sapply(h_t_clean, is.numeric)

pc_numeric_cols <- sapply(p_t_clean, is.numeric)

# Step 2: Get species (columns) with non-zero abundance

hh_species <- colnames(h_t_clean)[hh_numeric_cols][colSums(h_t_clean[, hh_numeric_cols]) > 0]

pc_species <- colnames(p_t_clean)[pc_numeric_cols][colSums(p_t_clean[, pc_numeric_cols]) > 0]

# Step 3: Generate and save Venn diagram

venn.plot <- venn.diagram(

  x = list(Healthy = hh_species, Cancer = pc_species),

  category.names = c("Healthy", "Cancer"),

  filename = "venn_species_overlap.png",  # Saves to current working directory

  output = TRUE,

  imagetype = "png",

  height = 2000,

  width = 2000,

  resolution = 300,

  col = "black",

  fill = c("#56B4E9", "#E69F00"),

  alpha = 0.5,

  cex = 2,

  cat.cex = 2,

  cat.pos = 0

)
```

5. Comparative Taxonomic Analysis of Gut Microbiome in healthy VS Cancer subjects

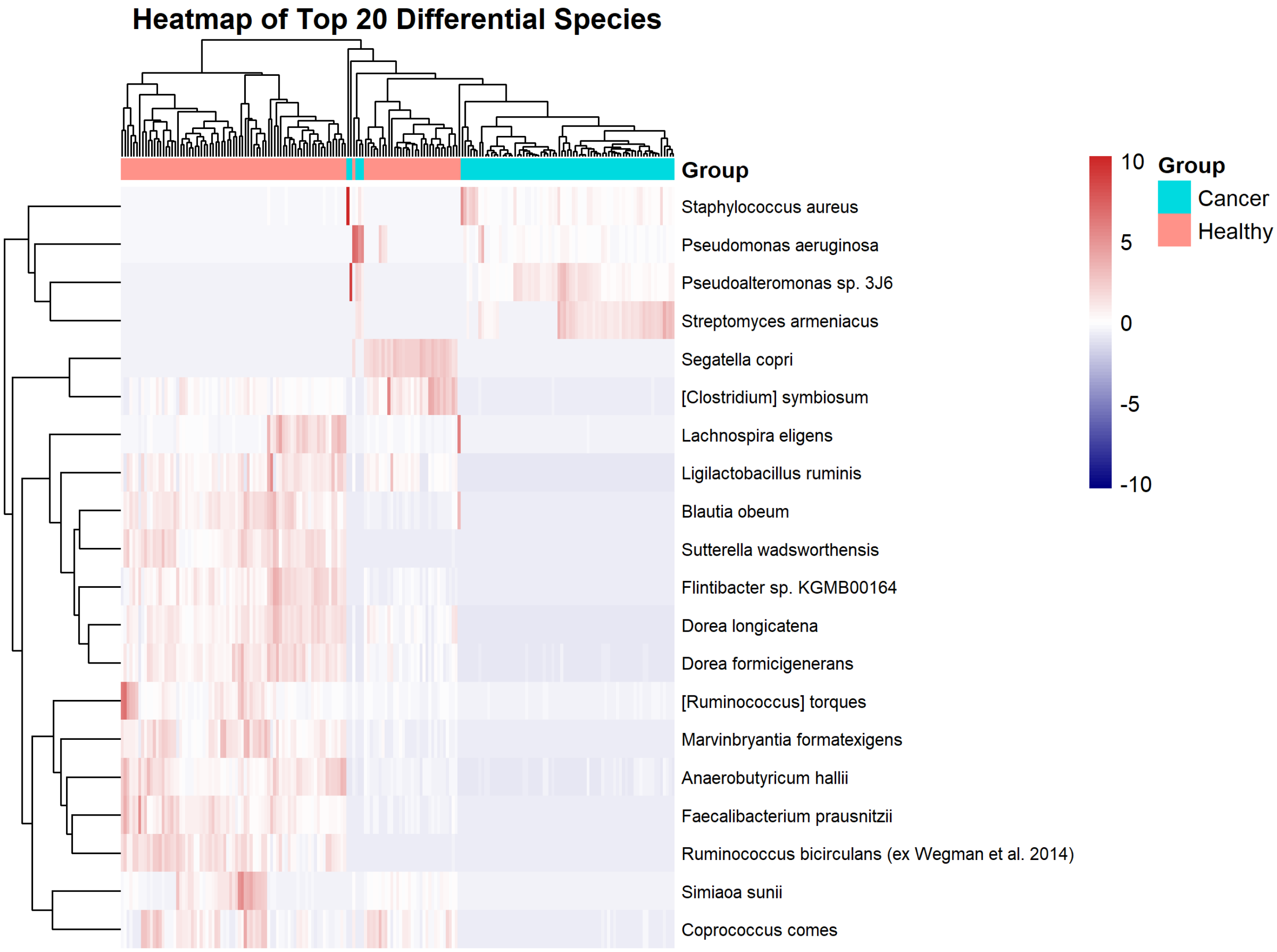

Fig. 5: Top 20 differentially abundant species

R Scripts:

Codes:

```
library(pheatmap)

# Choose top 20 species based on lowest adjusted p-values

top_heatmap_species <- results_df$Species[order(results_df$p_adj)][1:20]
```

6. Comparative Taxonomic Analysis of Gut Microbiome in healthy VS Cancer subjects

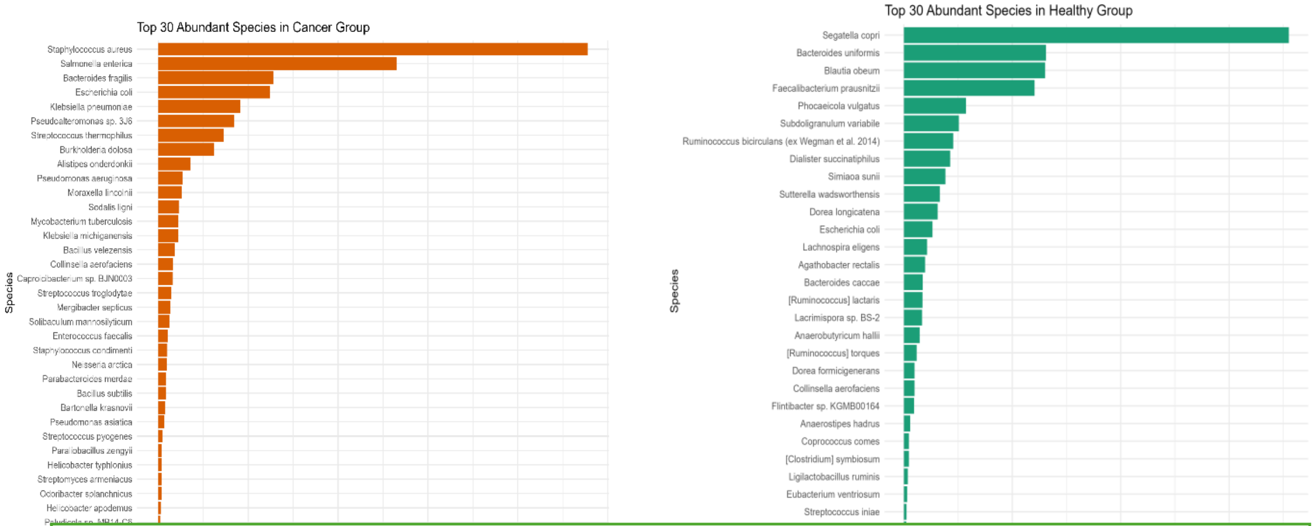

Fig. 6 Differential species abundance of gut microbiomes in Healthy and cancer subjects

R Scripts

```
#Top 30 Gut microbiome

# Load libraries

library(dplyr)

library(ggplot2)

# Step 1: Prepare the abundance matrix and group info

group_factor <- combined_data$Group
```

```
# Step 7: Extract abundance values for top species
healthy_abundances <- healthy_data[, top_healthy_species]
cancer_abundances  <- cancer_data[, top_cancer_species]
# Step 8: Compute mean abundance per species
healthy_means <- colMeans(healthy_abundances, na.rm = TRUE)
cancer_means  <- colMeans(cancer_abundances, na.rm = TRUE)
df_healthy <- data.frame(Species = names(healthy_means), Abundance = healthy_means, Group =
```

Conclusion

The data suggests a difference in the richness and diversity of the gut microbiome in cancer patients versus healthy individuals.

There is an increase in proteobacteria in cancer patients compared to healthy individuals, suggesting gut dysbiosis.

Strikingly, unique taxa are common in the cancer gut microbiome, including *Fusobacterium sp., Staphylococcus sp. Fusobacterium* is known to be highly associated with colorectal cancer.

Supplementary Data and R Scripts:

https://github.com/okpecallistus/Uchenna5202stuff/blob/main/Codes%20for%20My%20analysis%20Tutorial.pdf

References:

1. Fricker, A. D., Yao, T., Lindemann, S. R. & Flores, G. E. Enrichment and characterization of human-associated mucin-degrading microbial consortia by sequential passage. *FEMS Microbiol Ecol* 100, (2024).

2. Chamorro, N. *et al.* Landscapes and bacterial signatures of mucosa-associated intestinal microbiota in Chilean and Spanish patients with inflammatory bowel disease. *Microbial Cell* 8, 223–238 (2021).

3. Zitomersky, N. L. *et al.* Characterization of Adherent Bacteroidales from Intestinal Biopsies of Children and Young Adults with Inflammatory Bowel Disease. *PLoS One* 8, (2013).

4. Hurych, J. *et al.* Faecal Bacteriome and Metabolome Profiles Associated with Decreased Mucosal Inflammatory Activity Upon Anti-TNF Therapy in Paediatric Crohn's Disease. *J Crohns Colitis* 18, 106–120 (2024).

5. Lima, I. S. *et al.* Gut Dysbiosis: A Target for Protective Interventions against Parkinson's Disease. *Microorganisms* 11, (2023).

6. Singh, S. *et al.* Implication of Obesity and Gut Microbiome Dysbiosis in the Etiology of Colorectal Cancer. *Cancers* vol. 15 Preprint at https://doi.org/10.3390/cancers15061913 (2023).

7. Fassarella, M. *et al.* Gut microbiome stability and resilience: Elucidating the response to perturbations in order to modulate gut health. *Gut* vol. 70 595–605 Preprint at https://doi.org/10.1136/gutjnl-2020-321747 (2021).

8. Lima, I. S. *et al.* Gut Dysbiosis: A Target for Protective Interventions against Parkinson's Disease. *Microorganisms* 11, (2023).

9. Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* 569, 655–662 (2019).

10. Some of the data analysis code and figure generation were assisted by OpenAI's ChatGPT (version GPT-4, 2025), an AI language model.“

11. Hadley Wickham, Mine Cetinkaya-Rundel, and Garret Grolemund, R for Data Science (2e)https://r4ds.hadley.nz/preface-2e.html