# Shannon Diversity Plot

#Load & install packages

Install.packages(“tidyverse”)

Install.packages(“vegan”)

library(tidyverse)

library(vegan)

#STEP 1: Load & clean abundance data

relabund <- read.delim("bacteria.WGS.solid.case.relabund.txt", row.names = 1, check.names = FALSE)

relabund\_t <- as.data.frame(t(relabund)) %>%

rownames\_to\_column(var = "SampleID")

# STEP 2: Force everything to numeric, replace NA with 0

abund\_numeric <- relabund\_t %>%

mutate(across(-SampleID, ~ as.numeric(as.character(.)))) %>%

replace(is.na(.), 0)

# STEP 3: Filter out samples with 0 abundance

abund\_numeric <- abund\_numeric %>%

filter(rowSums(select(., -SampleID)) > 0)

# STEP 4: Calculate Shannon Diversity

abund\_matrix <- as.matrix(select(abund\_numeric, -SampleID))

storage.mode(abund\_matrix) <- "double"

shannon <- diversity(abund\_matrix, index = "shannon")

# STEP 5: Create dataframe with SampleID and Shannon Diversity

diversity\_df <- data.frame(

SampleID = abund\_numeric$SampleID,

Shannon = shannon)

# STEP 6: Load and clean metadata

metadata <- read.delim("metadata.WGS.solid.case.txt", check.names = FALSE)

metadata <- metadata %>%

# fix SampleID column

mutate(SampleID = toupper(trimws(bcr\_patient\_barcode)))

# STEP 7: Standardize SampleID for join

diversity\_df$SampleID <- toupper(trimws(diversity\_df$SampleID))

# STEP 8: Merge data and filter for complete cases

final\_data <- left\_join(diversity\_df, metadata, by = "SampleID") %>%

filter(!is.na(pathologic\_stage), !is.na(Shannon))

#remove empty reads

final\_data <- final\_data %>%

filter(!is.na(pathologic\_stage) & pathologic\_stage != "")

# STEP 9: Plot Shannon Diversity

ggplot(final\_data, aes(x = pathologic\_stage, y = Shannon, fill = pathologic\_stage)) +

geom\_boxplot() +

theme\_minimal() +

labs(title = "Shannon Diversity by Pathologic Stage in CRC Patients",

x = "Pathologic Stage", y = "Shannon Diversity Index") +

theme(axis.text.x = element\_text(angle = 45, hjust = 1))

#Run Statistics

kruskal.test(Shannon ~ pathologic\_stage, data = final\_data)

output: Kruskal-Wallis chi-squared = 39.78, df = 13, p-value = 0.00015

A Kruskal-Wallis test revealed a statistically significant difference in microbial Shannon diversity across colorectal cancer stages (χ² = 39.78, df = 13, *p* = 0.00015), indicating that tumor stage is associated with changes in alpha diversity

#PCA Plot

# Load libraries

library(tidyverse)

# STEP 1: Prepare abundance matrix

abund\_matrix <- relabund\_t %>%

column\_to\_rownames(var = "SampleID") %>%

mutate\_all(~ as.numeric(as.character(.))) %>%

replace(is.na(.), 0)

# Remove taxa with zero variance (required for PCA)

abund\_matrix <- abund\_matrix[, apply(abund\_matrix, 2, var) > 0]

# STEP 2: Run PCA

pca <- prcomp(abund\_matrix, center = TRUE, scale. = TRUE)

# Get scores for each sample

pca\_scores <- as.data.frame(pca$x)

pca\_scores$SampleID <- rownames(pca\_scores)

# STEP 3: Merge with metadata

metadata <- metadata %>% mutate(SampleID = toupper(trimws(bcr\_patient\_barcode)))

pca\_scores$SampleID <- toupper(trimws(pca\_scores$SampleID))

pca\_data <- left\_join(pca\_scores, metadata, by = "SampleID")

# STEP 4: Plot PCA, colored by pathologic stage

ggplot(pca\_data, aes(x = PC1, y = PC2, color = pathologic\_stage)) +

geom\_point(size = 2, alpha = 0.8) +

theme\_minimal() +

labs(

title = "PCA of Microbial Composition in CRC Patients",

x = paste0("PC1 (", round(summary(pca)$importance[2, 1] \* 100, 1), "%)"),

y = paste0("PC2 (", round(summary(pca)$importance[2, 2] \* 100, 1), "%)")

) +

theme(legend.position = "bottom")