Final Microbiome Project

Muhtarin Khayer Brohee

2025-04-24

Table of Contents

## Loading Library

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.4 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.4   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

## Introduction

This R Markdown document presents a reproducible workflow for analyzing genus- and species-level microbial profiles derived from Kraken2 reports of gut microbiome samples from male and female American cockroach. The analysis includes presence-absence matrices, richness estimates, statistical comparisons, and genus-level heatmaps.

## Load and Process Kraken Reports

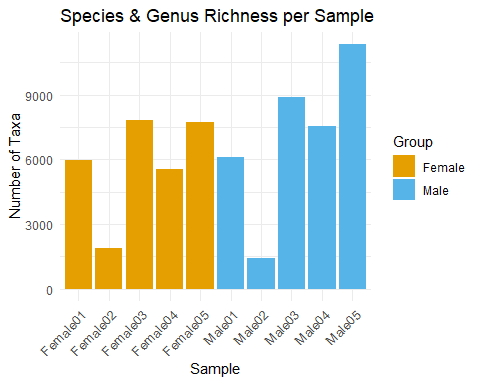
# Set folder with Kraken report files  
report\_dir <- "C:/Users/muhta/OneDrive/Documents/GitHub/PLPA-6820/Finalproject"  
files <- list.files(report\_dir, pattern = "\_kraken\_report.txt$", full.names = TRUE)  
  
# Function to extract presence from each report  
extract\_species\_presence <- function(filepath) {  
 # Get sample ID like F\_01 or M\_01 from filename  
 filename <- tools::file\_path\_sans\_ext(basename(filepath))  
 sample\_raw <- str\_extract(filename, "^[FM]\_\\d{2}")  
 group <- ifelse(str\_starts(sample\_raw, "F"), "Female", "Male")  
 sample\_name <- paste0(group, str\_extract(sample\_raw, "\\d+"))  
   
 # Read Kraken report  
 df <- read\_tsv(filepath, col\_names = FALSE, col\_types = cols(.default = "c"), quote = "", trim\_ws = FALSE)  
 colnames(df) <- c("percent", "reads", "direct", "rank", "taxid", "name")  
   
 df <- df %>%  
 mutate(reads = as.numeric(reads),  
 name = str\_trim(name))  
   
 # Include both genus and species  
 taxa\_df <- df %>%  
 filter(rank %in% c("G", "S"), reads > 0) %>%  
 mutate(present = 1) %>%  
 select(name, present)  
   
 # Convert to single-row tibble  
 presence\_vector <- deframe(taxa\_df)  
 tibble(Sample = sample\_name, !!!set\_names(presence\_vector, names(presence\_vector)))  
}  
  
# Apply function to all files  
presence\_list <- lapply(files, extract\_species\_presence)  
presence\_matrix <- reduce(presence\_list, full\_join, by = "Sample") %>%  
 replace(is.na(.), 0)  
  
# Save raw matrix  
write\_csv(presence\_matrix, file.path(report\_dir, "presence\_absence\_matrix.csv"))  
cat("Saved full presence/absence matrix.\n")

## Saved full presence/absence matrix.

# Aggregate runs by sample (e.g., Female01)  
presence\_aggregated <- presence\_matrix %>%  
 group\_by(Sample) %>%  
 summarise(across(where(is.numeric), ~ as.numeric(any(. == 1)))) %>%  
 ungroup()  
  
# Add sex group and richness  
presence\_aggregated <- presence\_aggregated %>%  
 mutate(Group = str\_extract(Sample, "Female|Male"),  
 Richness = rowSums(across(where(is.numeric))))  
  
# Save aggregated matrix  
write\_csv(presence\_aggregated, file.path(report\_dir, "presence\_absence\_matrix\_aggregated.csv"))

## Species & Genus Richness per Sample

# Colorblind-friendly palette  
cb\_palette <- c("Female" = "#E69F00", "Male" = "#56B4E9")  
  
# Updated plot  
richness\_plot <- ggplot(presence\_aggregated, aes(x = Sample, y = Richness, fill = Group)) +  
 geom\_col() +  
 scale\_fill\_manual(values = cb\_palette) +  
 labs(title = "Species & Genus Richness per Sample",  
 x = "Sample", y = "Number of Taxa") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1, size = 10))  
  
# Save plot  
ggsave("C:/Users/muhta/OneDrive/Documents/GitHub/PLPA-6820/Finalproject/figures/richness\_plot\_aggregated.png",  
 plot = richness\_plot, width = 8, height = 4, dpi = 300)  
# Draw the plot  
print(richness\_plot)



# Set path to save the PDF  
pdf("C:/Users/muhta/OneDrive/Documents/GitHub/PLPA-6820/Finalproject/figures/richness\_plot\_aggregated.pdf",  
 width = 8, height = 4) # you can adjust size  
# Close the PDF device  
dev.off()

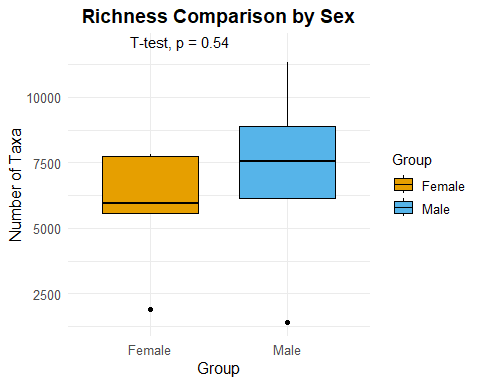
## png   
## 2

cat("PDF saved successfully.\n")

## PDF saved successfully.

## Richness Comparison by Sex (T-test)

# Build the richness boxplot with t-test  
richness\_boxplot <- ggboxplot(  
 presence\_aggregated,  
 x = "Group",  
 y = "Richness",  
 fill = "Group",  
 color = "black",  
 palette = c("Female" = "#E69F00", "Male" = "#56B4E9") # Okabe-Ito palette  
) +  
 stat\_compare\_means(  
 method = "t.test",  
 label.y = max(presence\_aggregated$Richness) \* 1.05 # position above max value  
 ) +  
 labs(  
 title = "Richness Comparison by Sex",  
 x = "Group",  
 y = "Number of Taxa"  
 ) +  
 theme\_minimal() +  
 theme(  
 plot.title = element\_text(size = 14, face = "bold", hjust = 0.5),  
 axis.text = element\_text(size = 10),  
 axis.title = element\_text(size = 12),  
 legend.title = element\_text(size = 11),  
 legend.text = element\_text(size = 10)  
 )  
  
print(richness\_boxplot)



pdf("C:/Users/muhta/OneDrive/Documents/GitHub/PLPA-6820/Finalproject/figures/richness\_boxplot\_final.pdf",  
 width = 6, height = 4)  
print(richness\_boxplot)  
dev.off()

## png   
## 2

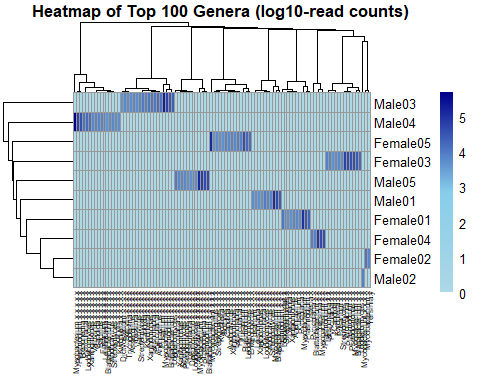
ggsave("C:/Users/muhta/OneDrive/Documents/GitHub/PLPA-6820/Finalproject/figures/richness\_boxplot\_final.png",  
 plot = richness\_boxplot, width = 6, height = 4, dpi = 300)

## Genus-Level Heatmap

# Load required libraries  
library(tidyverse)  
library(pheatmap)  
library(tibble)  
  
# Set report directory  
report\_dir <- "C:/Users/muhta/OneDrive/Documents/GitHub/PLPA-6820/Finalproject"  
files <- list.files(report\_dir, pattern = "\_kraken\_report.txt$", full.names = TRUE)  
  
# Function to extract genus-level read counts with renamed samples  
extract\_genus\_counts <- function(filepath) {  
 full\_name <- tools::file\_path\_sans\_ext(basename(filepath)) # Full filename  
 short\_id <- str\_extract(full\_name, "F\_\\d+|M\_\\d+") # e.g., F\_01  
 sample\_name <- short\_id %>%  
 str\_replace("F\_", "Female") %>%  
 str\_replace("M\_", "Male")  
   
 df <- read\_tsv(filepath, col\_names = FALSE, col\_types = cols(.default = "c"), quote = "", trim\_ws = FALSE)  
 colnames(df) <- c("percent", "reads", "direct", "rank", "taxid", "name")  
   
 df %>%  
 filter(rank == "G") %>%  
 mutate(reads = as.numeric(reads),  
 name = str\_trim(name)) %>%  
 group\_by(name) %>%  
 summarise(reads = sum(reads, na.rm = TRUE), .groups = "drop") %>%  
 pivot\_wider(names\_from = name, values\_from = reads, values\_fill = 0) %>%  
 mutate(Sample = sample\_name)  
}  
  
# Process all Kraken reports  
genus\_read\_list <- lapply(files, extract\_genus\_counts)  
  
# Merge into full matrix  
genus\_read\_matrix <- reduce(genus\_read\_list, full\_join, by = "Sample") %>%  
 replace(is.na(.), 0)  
  
# Save full matrix as CSV  
write\_csv(genus\_read\_matrix, file.path(report\_dir, "genus\_read\_counts\_matrix.csv"))  
  
# Load matrix, convert to log10, rename rows  
genus\_counts <- read\_csv(file.path(report\_dir, "genus\_read\_counts\_matrix.csv")) %>%  
 column\_to\_rownames("Sample")

## Rows: 10 Columns: 15513  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): Sample  
## dbl (15512): Abyssalbus.x, Abyssibius.x, Abyssicoccus.x, Acaryochloris.x, Ac...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

genus\_log <- log10(genus\_counts + 1)  
  
# Top 100 most abundant genera  
top\_100 <- colSums(genus\_log) %>%  
 sort(decreasing = TRUE) %>%  
 head(100) %>%  
 names()  
heatmap\_data <- genus\_log[, top\_100]  
  
# Plot heatmap  
pdf(file.path(report\_dir, "figures/genus\_readcount\_heatmap\_top100.pdf"), width = 18, height = 8)  
pheatmap(  
 heatmap\_data,  
 cluster\_rows = TRUE,  
 cluster\_cols = TRUE,  
 color = colorRampPalette(c("lightblue", "skyblue", "darkblue"))(100),  
 fontsize\_row = 10,  
 fontsize\_col = 6,  
 angle\_col = 90,  
 main = "Heatmap of Top 100 Genera (log10-read counts)"  
)



dev.off()

## pdf   
## 3

cat("Heatmap saved to: figures/genus\_readcount\_heatmap\_top100.pdf\n")

## Heatmap saved to: figures/genus\_readcount\_heatmap\_top100.pdf

## Conclusion

This project provides a full, reproducible pipeline for summarizing, visualizing, and comparing microbial diversity across sexes from Kraken2 output files in adult American Cockroahces(Periplaneta americana). The workflow includes presence/absence matrix generation, genus-level quantification, richness analysis, statistical testing, and a top-100 heatmap.