Contamination_decontamination_experiment-FS

- 2 AMR++ pipeline outputs were used.

contam.csv (before decontamination)
decontam.csv (after decontamination)

- R script was used to filter genes under 100 reads.

```
# Load necessary libraries
library(tidyverse)

# Read the CSV file
data <- read.csv("contam.csv")

# Convert data to long format and filter
filtered_data <- data %>%
  pivot_longer(cols = starts_with("ERR"),
               names_to = "Sample",
               values_to = "Read_Count") %>%
  filter(grepl("^MEG", gene_accession), Read_Count >= 100)

# Save filtered data to CSV
write.csv(filtered_data, "contamFiltered.csv", row.names = FALSE)

# Print message
cat("Filtered data saved to filtered_data.csv\n")
```

- New filtered outputs:

contamFiltered.csv (before decontamination)
decontamFiltered.csv (after decontamination)

- An R script was used to:
  - Perform ANOVA and Tukey test to show statistical significance in MEG number counts (**`p-value < 2e-16`**).
  - Identify unique MEG numbers in each group.
  - Create a table showing how many times each MEG number (gene type) was counted in both groups.

```
# Load required libraries
library(dplyr)
library(tidyr)
library(multcomp)

# Read the CSV files
decontam <- read.csv("decontamFiltered.csv")
contam <- read.csv("contamFiltered.csv")

# Create a list of data frames and add a 'Group' column to each
df_list <- list(decontam = decontam, contam = contam)
df_list <- lapply(names(df_list), function(name) {
  df <- df_list[[name]]
  df$Group <- name
  return(df)
})

# Combine data frames into one
combined_df <- bind_rows(df_list)

# Count the frequency of each 'gene_accession' in each group
```

```
frequency_table <- combined_df %>%
  group_by(Group, gene_accession) %>%
  summarise(Frequency = n(), .groups = 'drop')

# Perform ANOVA
anova_result <- aov(Frequency ~ Group, data = frequency_table)

# Print ANOVA Summary
anova_summary <- summary(anova_result)
print("ANOVA Summary:")
print(anova_summary)

# Extract the p-value from the ANOVA result
p_value <- anova_summary[[1]]$`Pr(>F)`[1]

# Perform Tukey HSD test regardless of ANOVA result
tukey_result <- TukeyHSD(anova_result)

# Print Tukey HSD results
print("Tukey HSD Test Results:")
print(tukey_result)

# Identify unique MEG numbers in each group
unique_meg <- combined_df %>%
  group_by(gene_accession) %>%
  summarise(Unique_in = toString(unique(Group)), .groups = 'drop') %>%
  filter(nchar(gsub("[^,]", "", Unique_in)) == 0)

print("Unique MEG numbers in each group:")
print(unique_meg)

# Write unique MEG numbers to a CSV file
write.csv(unique_meg, "unique_meg_numbers.csv", row.names = FALSE)

# Identify and output all differences in MEG numbers between groups
differences <- frequency_table %>%
  pivot_wider(names_from = Group, values_from = Frequency, values_fill = list(Frequency = 0))
%>%
  rowwise() %>%
  mutate(Difference = abs(decontam - contam))

print("All differences in MEG numbers between files:")
print(differences)

# Write all differences to a CSV file
write.csv(differences, "all_meg_number_differences.csv", row.names = FALSE)
```

- The 'all_meg_number_differences' data frame was copied and pasted into an Excel table 'TPFPTNFN_decon_contam'. False positives and false negatives and true positives and true negatives were calculated (Excel formula) by:
  - Formulating the difference between the two groups for each MEG number, e.g. **=IF(C2=B2, C2,ABS(C2-B2))**
  - Using the difference, false positives and negatives and true positives and negatives were calculated in each group, e.g. **=IF(AND(C2=0,B2=0),"TN", IF(C2=B2, "TP", IF(C2<B2, "FP",IF(C2>B2, "FN"))))**

- False positives, false negatives, true positives, and true negatives were counted using the **Cmd + F** function.

- Drug-resistance groups were counted across threshold groups:
  - On Excel 'Decon_Contam_ResGroup', an empty table was manually created listing drug-resistance groups down one column and the two groups across the top.

  - An R script was used to count the number of times a drug-resistant group type was identified in decontam and contam, using a key word (e.g. group name 'Multi-drug'). (Note: in this R script, unlike for the threshold group analysis, decontam must be replaced with contam when identifying genes in contaminated data).

```
# Load necessary package
library(dplyr)

# Read the CSV file
data <- read.csv("all_meg_number_differences.csv", stringsAsFactors = FALSE)

# Specify the keyword
keyword <- "Multi-drug"

# Filter rows that contain the keyword and select the contam column
filtered_numbers <- data %>%
  filter(apply(data, 1, function(row) any(grepl(keyword, row, ignore.case = TRUE)))) %>%
  select(decontam)

# Convert the contam column to numeric and sum the values
total_sum <- sum(as.numeric(filtered_numbers$decontam), na.rm = TRUE)

# Print the total sum
cat("The total sum of values in contam associated with the gene type is:", total_sum, "\n")
```

- The output would look something like this:

```
The total sum of values in contam associated with the gene type is: 20
```

And the number was manually (copied and pasted) in the empty table on excel.

- The drug-resistance group type table for decontam and contam was used to calculate percentages with excel formula which was converted into a bar chart (Figure 3).