MLST ST Analysis Documentation

This document includes the full MLST ST analysis script and expected output files.

Expected Output Files:

```
ST_distribution_barplot.png
ST_distribution.csv
Allele_variation_summary.csv
```

Script:

```
library(dplyr)
library(ggplot2)
library(readr)
library(tidyr)

# ===== INPUT PATH =====
input_file <- "/data/internship_data/nidhi/aba/output/mlst_output/mlst_results.csv"

# Read file with NO headers
mlst <- read_csv(input_file, col_names = FALSE)

# Assign proper column names
colnames(mlst) <- c(
 "file", "species", "ST",
 "cpn60", "fusA", "gltA", "pyrG", "recA", "rplB", "rpoB"
)

# ------------------------------
# 1. ST DISTRIBUTION
# ------------------------------
st_table <- mlst %>%
 count(ST, name = "Count") %>%
 mutate(Percentage = round((Count / sum(Count)) * 100, 2))

write.csv(st_table, "ST_distribution.csv", row.names = FALSE)

# Barplot of ST distribution
png("ST_distribution_barplot.png", width = 900, height = 600)
ggplot(st_table, aes(x = reorder(ST, -Count), y = Count)) +
 geom_bar(stat = "identity") +
 theme_minimal() +
 labs(title = "MLST Sequence Type Distribution",
 x = "Sequence Type (ST)", y = "Count") +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
dev.off()

# ------------------------------
# 2. ALLELE VARIATION ANALYSIS
# ------------------------------
allele_cols <- c("cpn60", "fusA", "gltA", "pyrG", "recA", "rplB", "rpoB")

allele_variation <- mlst %>%
 select(all_of(allele_cols)) %>%
 summarise(across(everything(), ~n_distinct(.))) %>%
 pivot_longer(cols = everything(),
 names_to = "Locus",
 values_to = "Unique_Alleles") %>%
 arrange(desc(Unique_Alleles))

write.csv(allele_variation, "Allele_variation_summary.csv", row.names = FALSE)

cat("
Analysis complete.
")
cat("Generated files:
")
cat(" ST_distribution.csv
")
cat(" ST_distribution_barplot.png
```

```
")
cat(" Allele_variation_summary.csv
")
```