Analysis of GC Content in Bacterial Genomes

CJA

2025-12-16

1. Introduction This report presents an analysis of the GC content (%) in bacterial genomes across different genera. The data were retrieved from the NCBI database and processed automatically.

2. Data Loading

```
df <- read.csv("D:/Programation/Nouveau dossier/gc_genomes.csv")
df$genus <- as.factor(df$genus)
```

3. Descriptive Statistics

```
desc <- df %>%
group_by(genus) %>%
summarise(
mean_gc = mean(gc_content_percent),
sd_gc = sd(gc_content_percent),
n = n()
)
desc
```

```
## # A tibble: 3 x 4
##   genus        mean_gc sd_gc     n
##   <fct>          <dbl> <dbl> <int>
## 1 Bacillus        42.5 2.61     10
## 2 Escherichia     50.5 0.368    10
## 3 Pseudomonas     62.4 2.84     10
```

4. Normality Test

```
norm_test <- df %>%
group_by(genus) %>%
summarise(shapiro_p = shapiro.test(gc_content_percent)$p.value)
norm_test
```

```
## # A tibble: 3 x 2
##   genus        shapiro_p
##   <fct>            <dbl>
## 1 Bacillus         0.161
## 2 Escherichia      0.0133
## 3 Pseudomonas      0.537
```

5. Global Statistical Test

```
if(all(norm_test$shapiro_p > 0.05)){
test <- aov(gc_content_percent ~ genus, data=df)
summary(test)
posthoc <- TukeyHSD(test)
posthoc
} else {
test <- kruskal_test(gc_content_percent ~ genus, data=df)
test
posthoc <- df %>%
dunn_test(gc_content_percent ~ genus, p.adjust.method = "bonferroni")
posthoc
}
```

```
## # A tibble: 3 x 9
##   .y.                 group1      group2         n1    n2 statistic           p      p.adj p.adj.sign:
## * <chr>               <chr>       <chr>       <int> <int>     <dbl>       <dbl>      <dbl> <chr>
## 1 gc_content_percent Bacillus    Escherichia    10    10      2.54 0.0110         0.0329  *
## 2 gc_content_percent Bacillus    Pseudomonas    10    10      5.09 0.000000364    0.00000109 ****
## 3 gc_content_percent Escherichia Pseudomonas    10    10      2.54 0.0110         0.0329  *
```
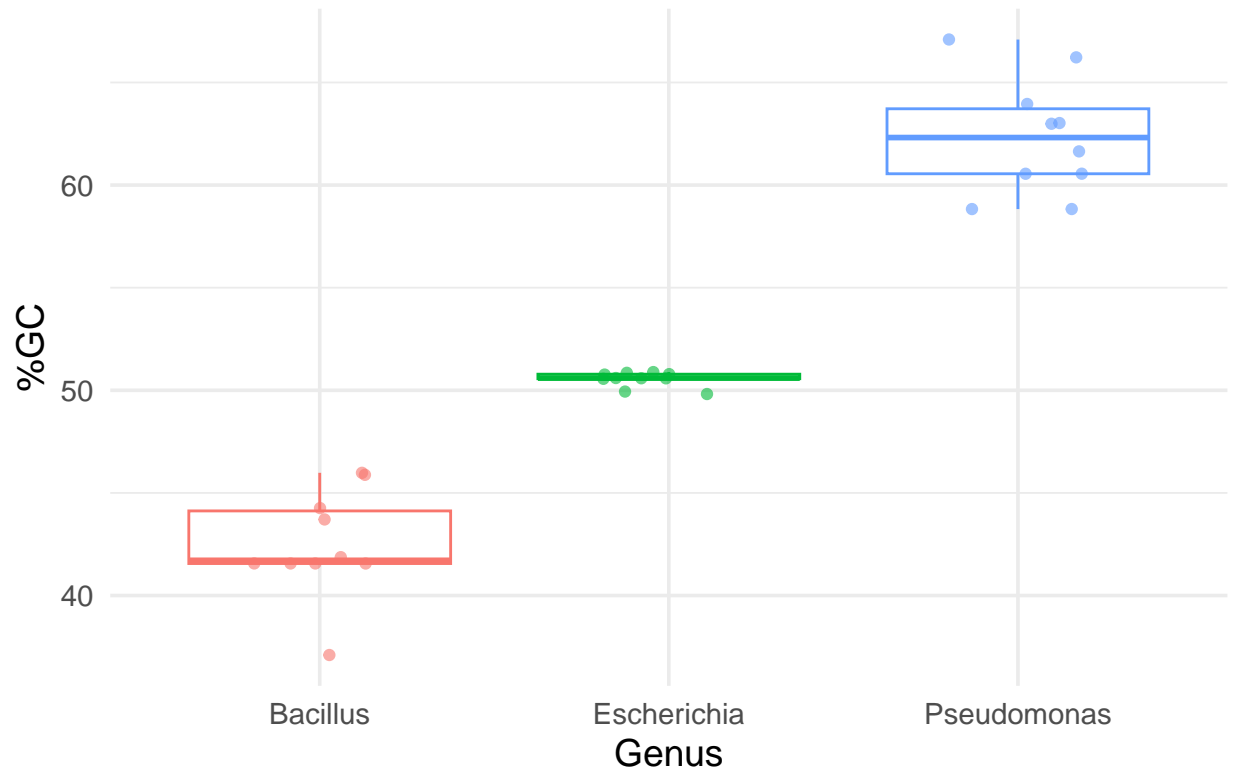
6. Visualization

```
p <- ggplot(df, aes(x=genus, y=gc_content_percent, color=genus)) +
geom_boxplot(outlier.shape = NA) +
geom_jitter(width=0.2, alpha=0.6) +
theme_minimal(base_size = 14) +
labs(
title = "GC Content (%) by Bacterial Genus",
x = "Genus",
y = "%GC"
) +
theme(legend.position = "none")
p
```

GC Content (%) by Bacterial Genus

7. Interpretation The results highlight differences in GC content between bacterial genera. Statistical tests and visualizations provide a clear quantitative view of these differences. These insights can be used for comparative genomics studies or publications.