Comparative Genomics Analysis of Escherichia coli

CJA

2025-12-16

1. Introduction

This report presents a comparative genomics analysis based on complete bacterial genomes. The dataset was automatically retrieved from public databases and processed through a reproducible pipeline. The focus is on GC content, genome size, and exploratory clustering.

2. Data Loading

```
df <- read.csv("D:/Programation/phylo E.Coli/genome_features.csv")

df$genus <- as.factor(df$genus)

str(df)
```

```
## 'data.frame':    50 obs. of  5 variables:
##  $ genome_id         : chr  "AP025945.1" "AP045036.1" "AP045043.1" "AP045049.1" ...
##  $ genus             : Factor w/ 1 level "Escherichia": 1 1 1 1 1 1 1 1 1 1 ...
##  $ genome_size_bp    : int  4857907 4873051 4772753 4924793 4889516 4979719 4941033 4866693 4762631 4
##  $ gc_content_percent: num  49.8 50.9 50.9 50.6 50.6 ...
##  $ n_contigs         : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
summary(df)
```

```
##   genome_id                genus    genome_size_bp    gc_content_percent   n_contigs
##  Length:50          Escherichia:50   Min.   :4392175   Min.   :49.76      Min.   :1
##  Class :character                    1st Qu.:4777857   1st Qu.:50.55      1st Qu.:1
##  Mode  :character                    Median :4869872   Median :50.59      Median :1
##                                      Mean   :4860591   Mean   :50.57      Mean   :1
##                                      3rd Qu.:4938234   3rd Qu.:50.77      3rd Qu.:1
##                                      Max.   :5336054   Max.   :50.89      Max.   :1
```

3. Descriptive Statistics

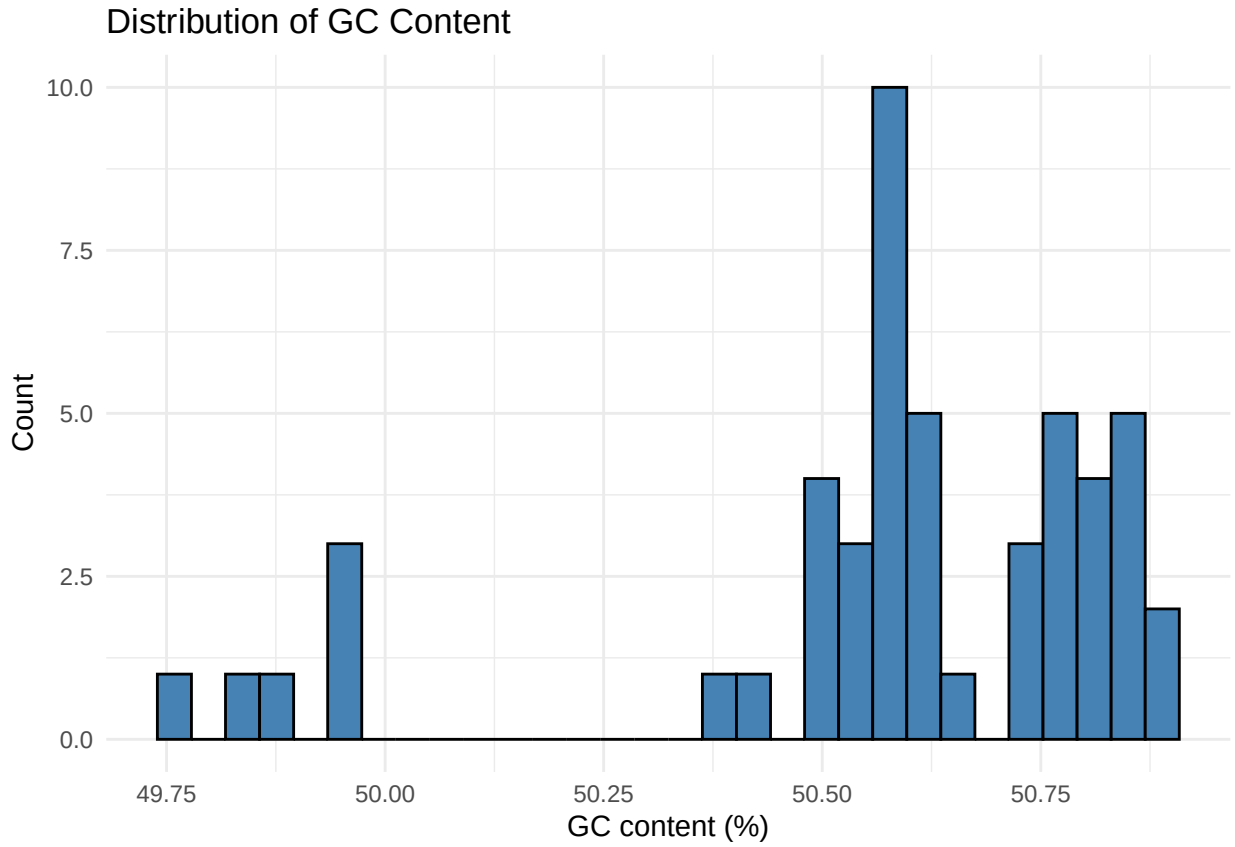
```
desc_stats <- df %>%
group_by(genus) %>%
summarise(
n_genomes = n(),
mean_gc = mean(gc_content_percent),
sd_gc = sd(gc_content_percent),
mean_size_mb = mean(genome_size_bp) / 1e6
)

desc_stats
```

```
## # A tibble: 1 x 5
##   genus       n_genomes mean_gc sd_gc mean_size_mb
##   <fct>           <int>   <dbl> <dbl>        <dbl>
## 1 Escherichia        50    50.6 0.289         4.86
```

4. Distribution of GC Content

```
ggplot(df, aes(gc_content_percent)) +
geom_histogram(bins = 30, fill = "steelblue", color = "black") +
theme_minimal() +
labs(
title = "Distribution of GC Content",
x = "GC content (%)",
y = "Count"
)
```



5. Normality Assessment

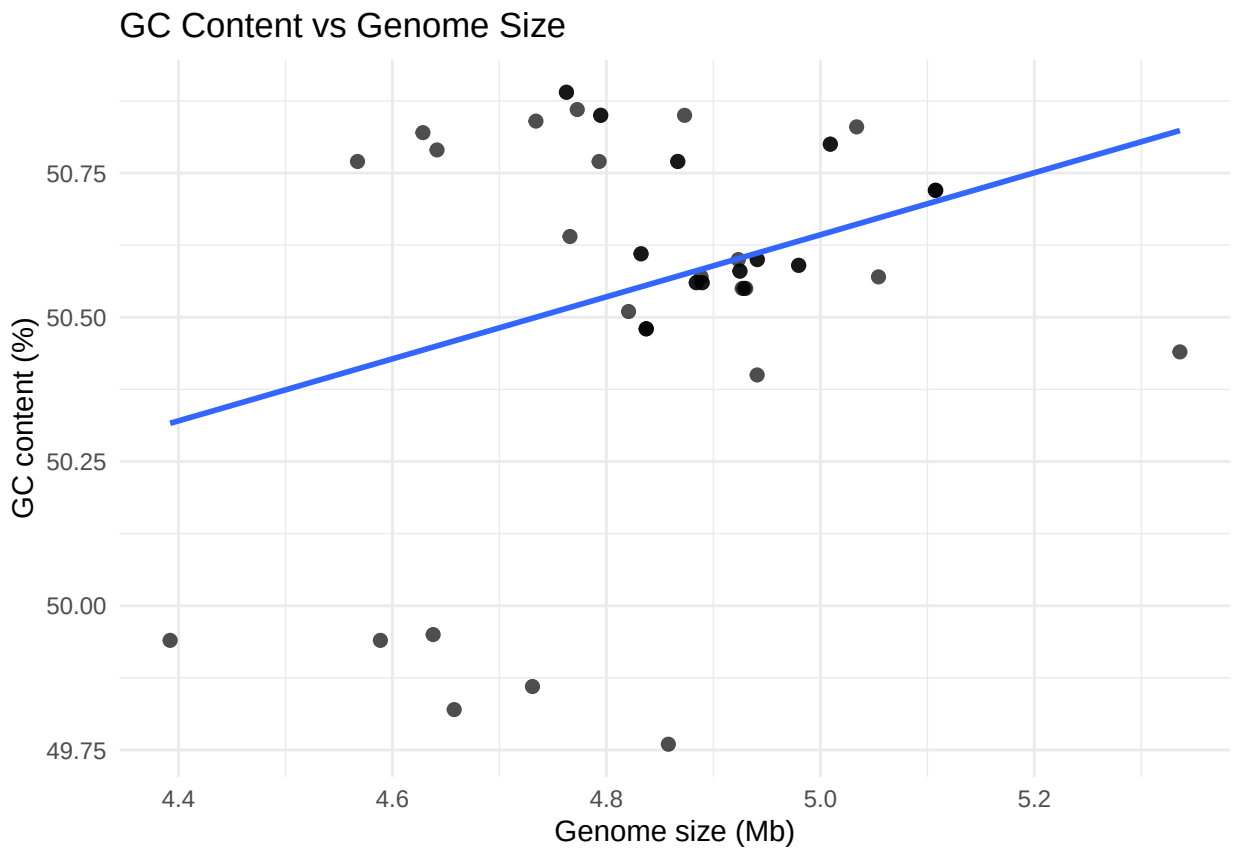
```
normality <- df %>%
group_by(genus) %>%
shapiro_test(gc_content_percent)

normality
```

```
## # A tibble: 1 x 4
##   genus       variable           statistic           p
##   <fct>       <chr>                  <dbl>       <dbl>
## 1 Escherichia gc_content_percent     0.799 0.000000848
```

6. Relationship Between GC Content and Genome Size

```
ggplot(df, aes(genome_size_bp / 1e6, gc_content_percent)) +
geom_point(size = 2, alpha = 0.7) +
geom_smooth(method = "lm", se = FALSE) +
theme_minimal() +
labs(
title = "GC Content vs Genome Size",
x = "Genome size (Mb)",
y = "GC content (%)"
)
```



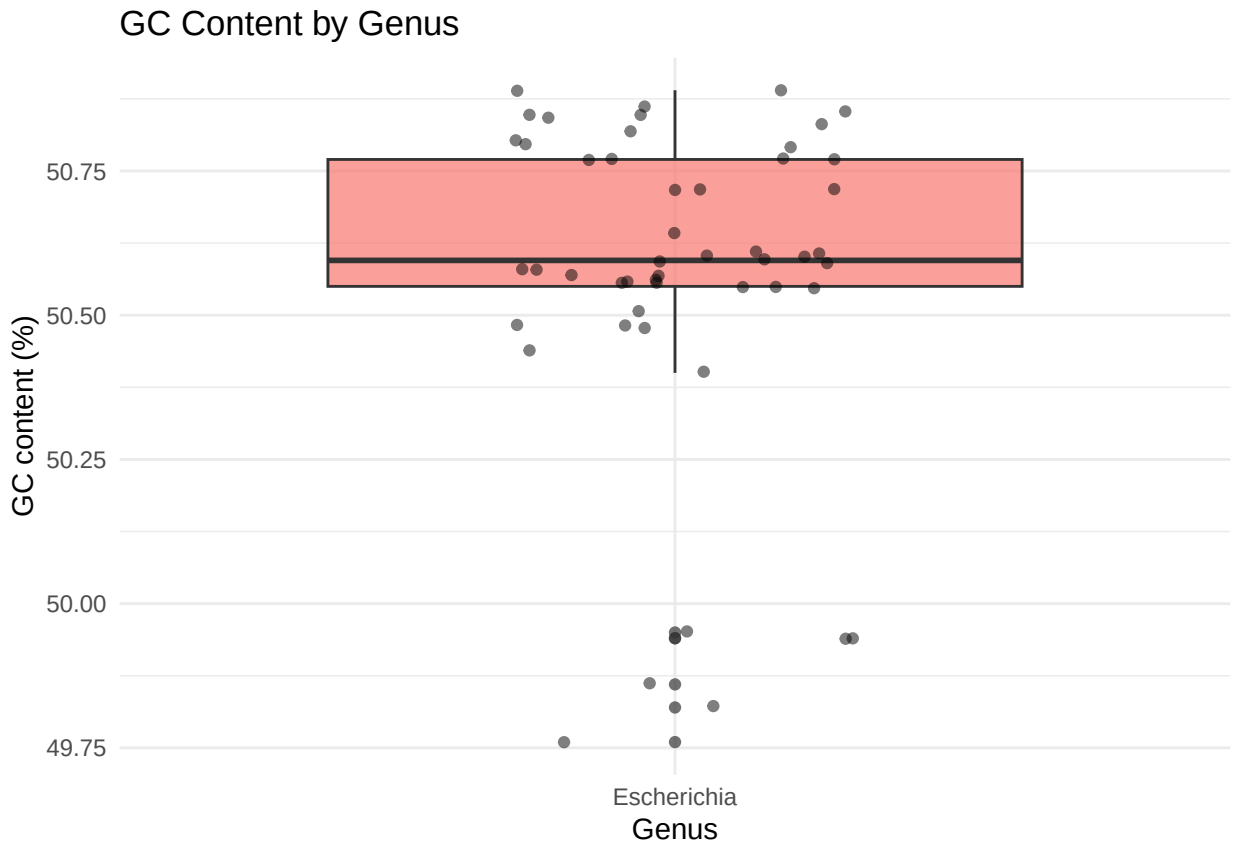

```
cor_test <- cor.test(
df$genome_size_bp,
df$gc_content_percent,
method = "spearman"
)

cor_test
```

```
##
##  Spearman's rank correlation rho
##
## data:  df$genome_size_bp and df$gc_content_percent
## S = 21137, p-value = 0.9176
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##         rho
## -0.01500289
```

7. GC Content by Genus

```
ggplot(df, aes(genus, gc_content_percent, fill = genus)) +
geom_boxplot(alpha = 0.7) +
geom_jitter(width = 0.2, alpha = 0.5) +
theme_minimal() +
labs(
title = "GC Content by Genus",
x = "Genus",
y = "GC content (%)"
) +
theme(legend.position = "none")
```



8. Hierarchical Clustering (Exploratory Phylogeny) 8.1 Feature Scaling and Distance Matrix

```
row.names(df) <- df$genome_id

features <- df %>%
select(gc_content_percent, genome_size_bp, n_contigs) %>%
scale()

dist_matrix <- dist(features, method = "euclidean")
```

8.2 Hierarchical Clustering

```
hc <- hclust(dist_matrix, method = "ward.D2")
```

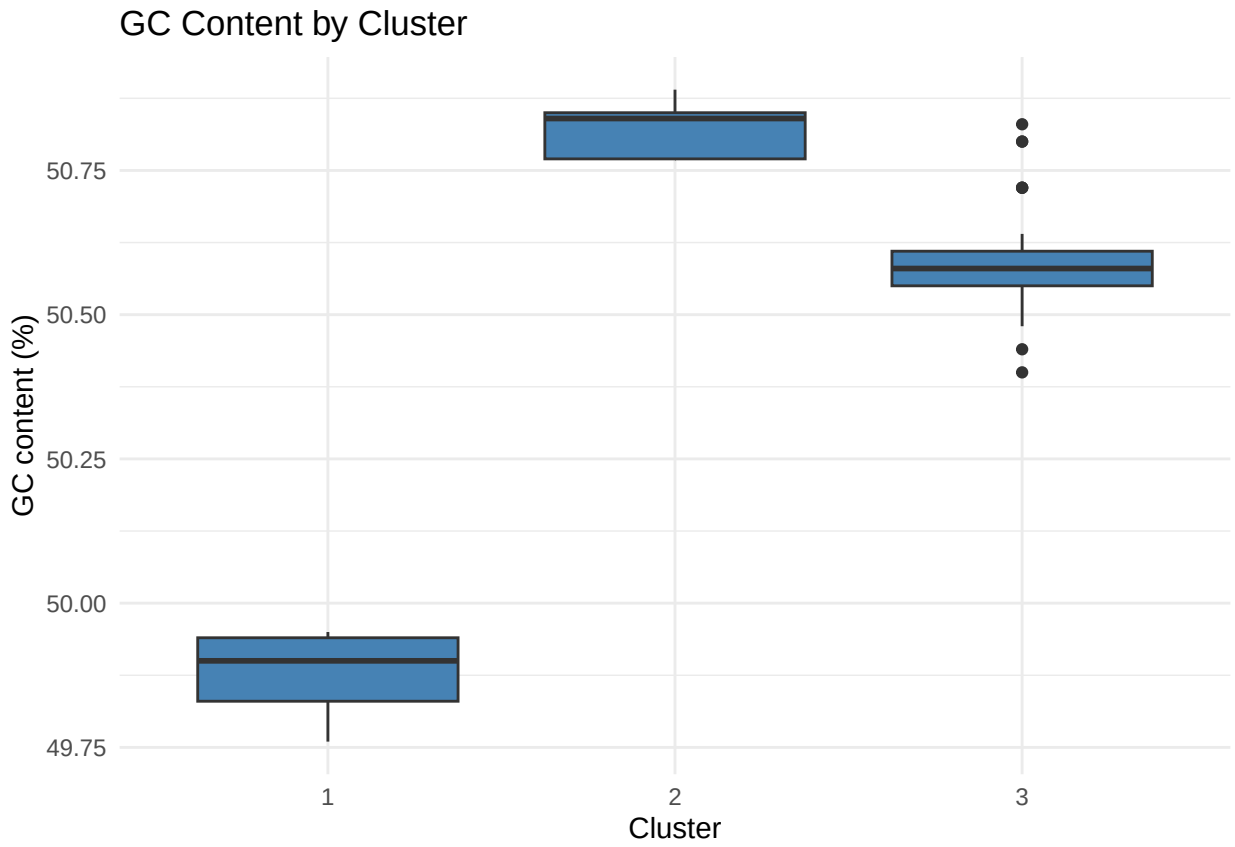
8.3 Dendrogram Visualization

```
fviz_dend(
hc,
k = 3,
rect = TRUE,
cex = 0.6,
main = "Genome Similarity Clustering"
)
```

8.4 Cluster Interpretation

```r
df$cluster <- factor(cutree(hc, k = 3))  # <- important : factor

ggplot(df, aes(x = cluster, y = gc_content_percent)) +
  geom_boxplot(fill = "steelblue") +
  theme_minimal() +
  labs(
    title = "GC Content by Cluster",
    x = "Cluster",
    y = "GC content (%)"
  )
```



9. Conclusion

This exploratory comparative genomics analysis highlights variability in GC content and genome size within the dataset. Hierarchical clustering reveals genome groups based on global genomic features. Such approaches provide a fast and informative way to explore genomic similarity without requiring heavy phylogenetic pipelines.