Untitled

Ann

2025-09-18

```r
library("R.utils")
```

```
## Loading required package: R.oo

## Loading required package: R.methodsS3

## R.methodsS3 v1.8.2 (2022-06-13 22:00:14 UTC) successfully loaded. See ?R.methodsS3 for help.

## R.oo v1.27.1 (2025-05-02 21:00:05 UTC) successfully loaded. See ?R.oo for help.

##
## Attaching package: 'R.oo'

## The following object is masked from 'package:R.methodsS3':
##
##     throw

## The following objects are masked from 'package:methods':
##
##     getClasses, getMethods

## The following objects are masked from 'package:base':
##
##     attach, detach, load, save

## R.utils v2.13.0 (2025-02-24 21:20:02 UTC) successfully loaded. See ?R.utils for help.

##
## Attaching package: 'R.utils'

## The following object is masked from 'package:utils':
##
##     timestamp

## The following objects are masked from 'package:base':
##
##     cat, commandArgs, getOption, isOpen, nullfile, parse, warnings
```

```r
URL="https://ftp.ensemblgenomes.ebi.ac.uk/pub/bacteria/release-62/fasta/bacteria_40_collection/mesomycop
#download.file(URL,destfile="Mesomycoplasma_cds.fa.gz")
#gunzip("Mesomycoplasma_cds.fa.gz")
list.files()
```

```
##  [1] "bioassignment"                  "condition_treated_results.csv"
##  [3] "ecoli_cds.fa"                   "gene_expression.tsv"
##  [5] "growth_data.csv"                "Mesomycoplasma_cds.fa"
##  [7] "Mesomycoplasma_cds.fa.gz"       "part2.pdf"
##  [9] "part2.Rmd"                      "partt1.pdf"
## [11] "partt1.Rmd"                     "week10.Rmd"
```

```
# -------------------------------
# R Markdown code to check file location
# -------------------------------

# 1. Check R's current working directory
getwd()  # This shows the folder R is currently using
```

```
## [1] "/home/s225654083/myrepo/bioassignment"
```

```
# 2. List all files in the working directory
list.files()  # Shows files in the current folder
```

```
##  [1] "bioassignment"                  "condition_treated_results.csv"
##  [3] "ecoli_cds.fa"                   "gene_expression.tsv"
##  [5] "growth_data.csv"                "Mesomycoplasma_cds.fa"
##  [7] "Mesomycoplasma_cds.fa.gz"       "part2.pdf"
##  [9] "part2.Rmd"                      "partt1.pdf"
## [11] "partt1.Rmd"                     "week10.Rmd"
```

```
# 3. If your files are in a different folder, specify the path
folder_path <- "C:/Users/Ann/Documents/myrepo/bioassignment"  # <-- replace with your actual path

# 4. List all files in that folder
list.files(folder_path)
```

```
## character(0)
```

```
# 5. Check if your specific FASTA files exist
file.exists(paste0(folder_path, "/e_coli_cds.fa"))
```

```
## [1] FALSE
```

```
file.exists(paste0(folder_path, "/mesomycoplasma_cds.fa"))
```

```
## [1] FALSE
```

```
library(seqinr)
```

```
##
## Attaching package: 'seqinr'
```

```
## The following object is masked from 'package:R.oo':
##
##     getName
```

```
# Read the files with exact names
ecoli_fasta <- read.fasta("ecoli_cds.fa")
Mesomycoplasma_fasta <- read.fasta("Mesomycoplasma_cds.fa")

# Calculate total CDS lengths
total_ecoli_length <- sum(sapply(ecoli_fasta, length))
total_Mesomycoplasma_length <- sum(sapply(Mesomycoplasma_fasta, length))

# Print results
cat("Total E.coli CDS length:", total_ecoli_length, "\n")
```

```
## Total E.coli CDS length: 3978528
```

```
cat("Total Mesomycoplasma CDS length:", total_Mesomycoplasma_length, "\n")
```

```
## Total Mesomycoplasma CDS length: 859086
```

```
library(seqinr)
# Count the sequences
num_ecoli <- length(ecoli_fasta)
num_myco <- length(Mesomycoplasma_fasta)


# Calculate total lengths
total_ecoli_length <- sum(sapply(ecoli_fasta, length))
total_Mesomycoplasma_length <- sum(sapply(Mesomycoplasma_fasta, length))

# Calculate number of genes
num_ecoli_genes <- length(ecoli_fasta)
num_Mesomycoplasma_genes <- length(Mesomycoplasma_fasta)

# Calculate average gene lengths
avg_ecoli_length <- total_ecoli_length / num_ecoli_genes
avg_Mesomycoplasma_length <- total_Mesomycoplasma_length / num_Mesomycoplasma_genes

# Create a data frame
gene_lengths_df <- data.frame(
  Organism = c("E.coli", "Mesomycoplasma"),
  Total_CDS_Length = c(total_ecoli_length, total_Mesomycoplasma_length),
  Number_of_Genes = c(num_ecoli_genes, num_Mesomycoplasma_genes),
  Average_Gene_Length = c(avg_ecoli_length, avg_Mesomycoplasma_length)
)

# Print the table
gene_lengths_df
```

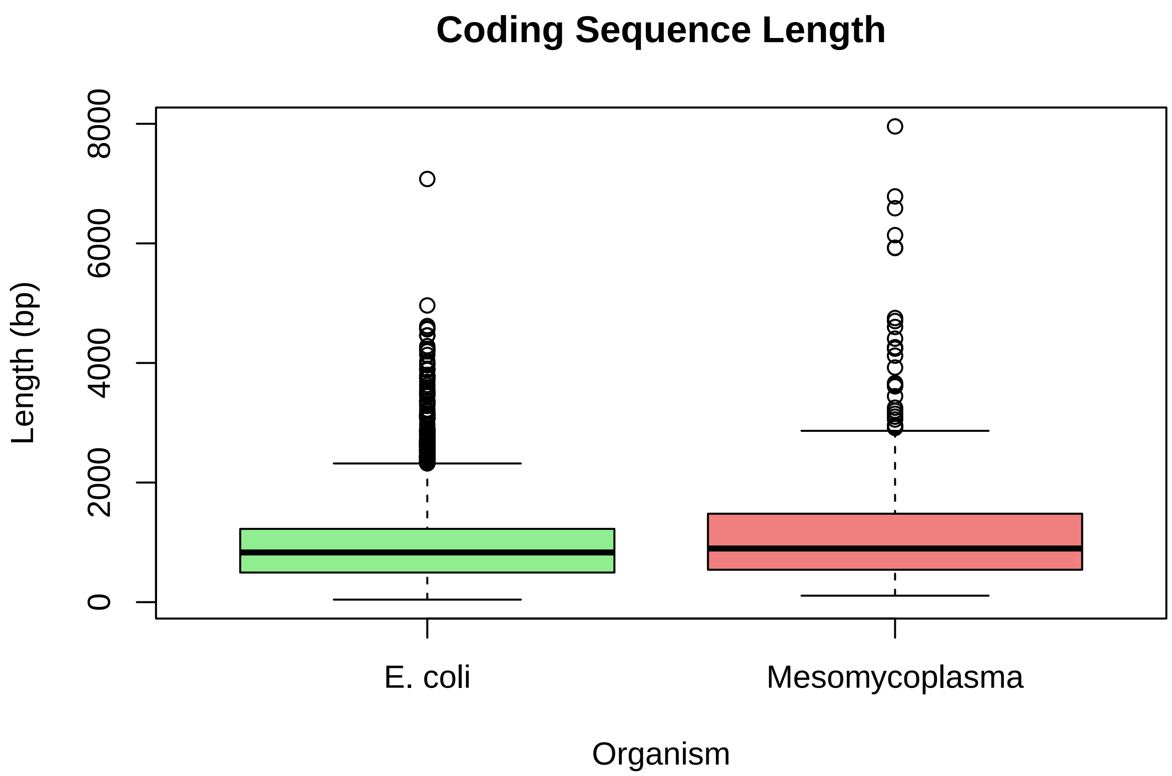
```
##         Organism Total_CDS_Length Number_of_Genes Average_Gene_Length
## 1         E.coli          3978528            4239            938.5534
## 2 Mesomycoplasma           859086             748           1148.5107
```

```
# Count the sequences
num_ecoli <- length(ecoli_fasta)
num_myco <- length(Mesomycoplasma_fasta)

# Get the total length of every sequence
total_length_ecoli <- sum(sapply(ecoli_fasta, length))
total_length_myco <- sum(sapply(Mesomycoplasma_fasta, length))

# Calculate average CDS length
avg_length_ecoli <- round(total_length_ecoli / num_ecoli, 1)
avg_length_myco <- round(total_length_myco / num_myco, 1)

# Make a table (showing total length in bp and kbp, plus average CDS length)
length_table <- data.frame(
  Organism = c("E. coli", "Mesomycoplasma"),
  Number_of_CDS = c(num_ecoli, num_myco),
  Total_Length_bp = c(total_length_ecoli, total_length_myco),
  Total_Length_kbp = round(c(total_length_ecoli/1000, total_length_myco/1000), 1),
```

```r
  Average_CDS_Length = c(avg_length_ecoli, avg_length_myco)
)

# Print the table
print(length_table)
```

```
##         Organism Number_of_CDS Total_Length_bp Total_Length_kbp
## 1        E. coli          4239         3978528           3978.5
## 2 Mesomycoplasma           748          859086            859.1
##   Average_CDS_Length
## 1              938.6
## 2             1148.5
```

```r
# Get the length of each individual sequence
ecoli_lengths <- sapply(ecoli_fasta, length)
myco_lengths <- sapply(Mesomycoplasma_fasta, length)

# Combine for plotting
all_lengths <- c(ecoli_lengths, myco_lengths)
all_organisms <- c(rep("E. coli", length(ecoli_lengths)), rep("Mesomycoplasma", length(myco_lengths)))
plot_data <- data.frame(Length = all_lengths, Organism = all_organisms)

# Make the boxplot
boxplot(Length ~ Organism, data = plot_data,
        main = "Coding Sequence Length",
        ylab = "Length (bp)",
        xlab = "Organism",
        col = c("lightgreen", "lightcoral"))
```



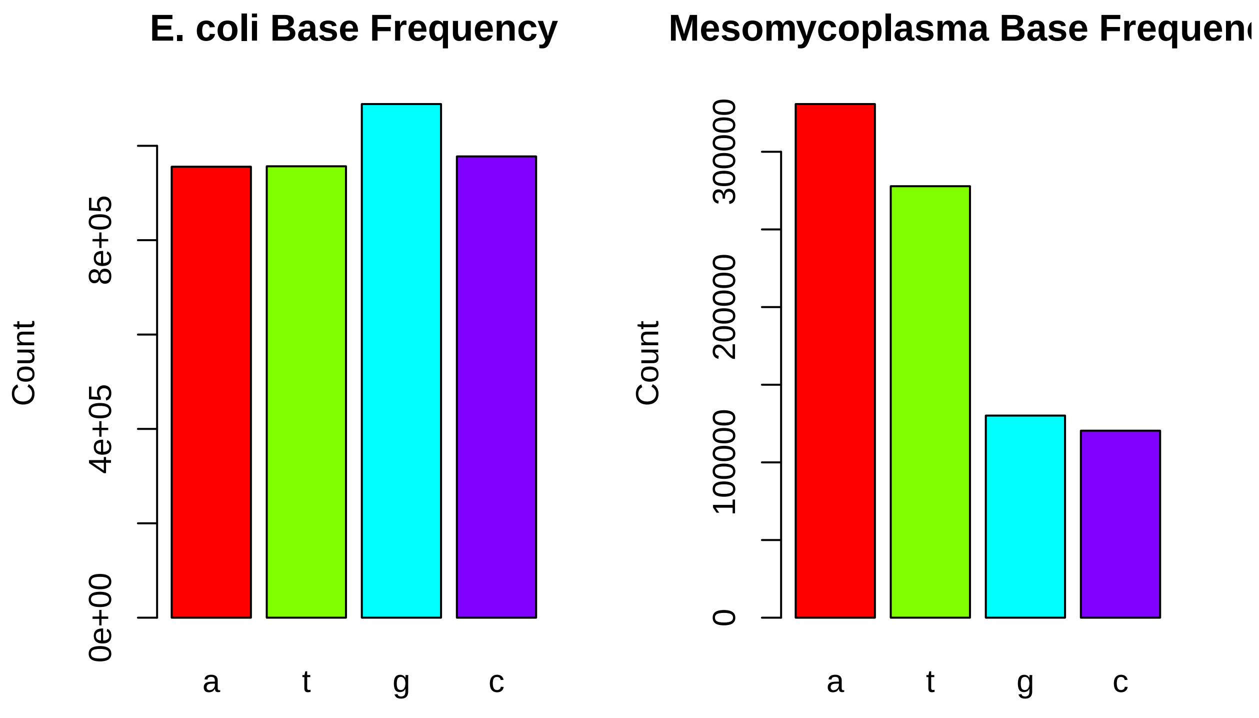
```r
# Calculate summary statistics (Mean and Median)
summary_table <- data.frame(
  Organism = c("E. coli", "Mesomycoplasma"),
  Mean_Length = round(c(mean(ecoli_lengths), mean(myco_lengths)), 1),
  Median_Length = c(median(ecoli_lengths), median(myco_lengths))
)

# Print the summary table
print(summary_table)
```

```
##         Organism Mean_Length Median_Length
## 1        E. coli       938.6           831
## 2 Mesomycoplasma      1148.5           897
```

```r
# Count all bases in all sequences for each organism
all_ecoli_bases <- unlist(ecoli_fasta)
all_myco_bases <- unlist(Mesomycoplasma_fasta)

# Count frequency of A, T, G, C
ecoli_base_freq <- table(all_ecoli_bases)
myco_base_freq <- table(all_myco_bases)

# Combine frequencies into a single data frame for easier comparison
base_freq_table <- data.frame(
  Base = c("A", "T", "G", "C"),
  E_coli = as.integer(ecoli_base_freq[c("a","t","g","c")]),
  Mesomycoplasma = as.integer(myco_base_freq[c("a","t","g","c")])
)

# Print the table
print(base_freq_table)
```

```
##   Base  E_coli Mesomycoplasma
## 1    A  955768         330716
## 2    T  956665         277818
## 3    G 1088501         130141
## 4    C  977594         120411
```

```r
# Make side-by-side barplots
par(mfrow = c(1, 2), mar=c(5,4,4,2)) # Two plots side-by-side, adjust margins

barplot(ecoli_base_freq[c("a","t","g","c")],
        main="E. coli Base Frequency",
        col=rainbow(4),
        ylab="Count")

barplot(myco_base_freq[c("a","t","g","c")],
        main="Mesomycoplasma Base Frequency",
        col=rainbow(4),
        ylab="Count")
```



```r
library(seqinr)
# Put your FASTA file in the working directory or provide the full path
ecoli_cds <- read.fasta(file = "ecoli_cds.fa", seqtype = "DNA")
```

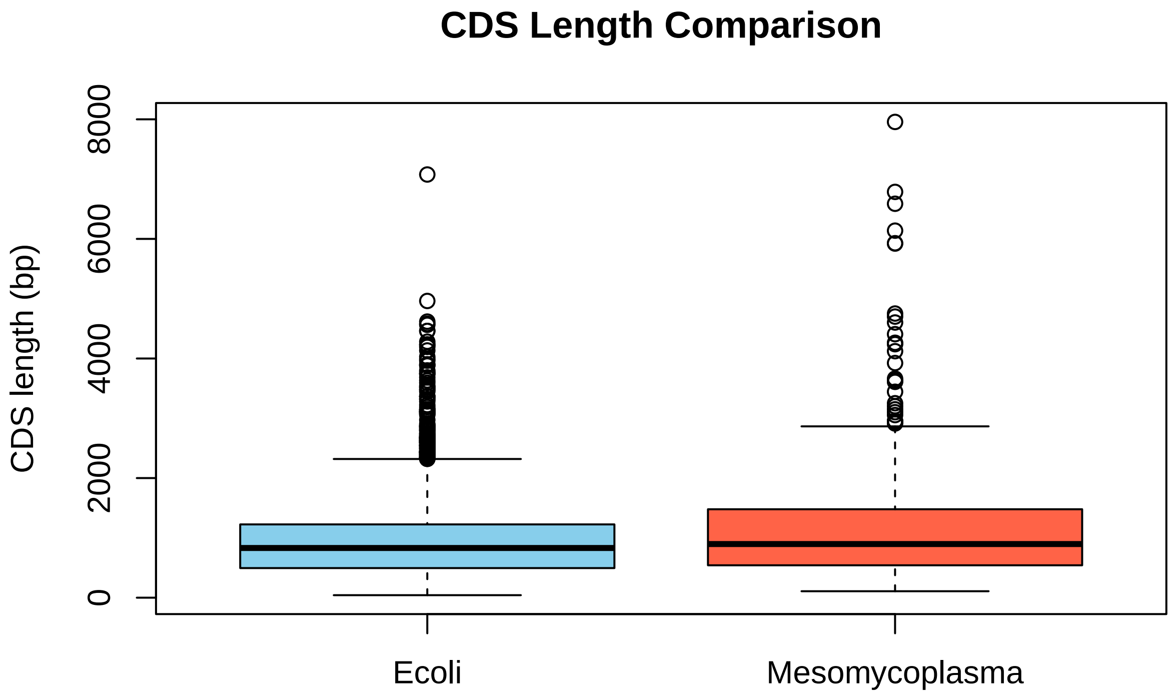
```
meso_cds <- read.fasta(file = "Mesomycoplasma_cds.fa", seqtype = "DNA")
str(head(ecoli_cds))    # Should show a list of sequences
```

```
## List of 6
##  $ AAC73112: 'SeqFastadna' chr [1:66] "a" "t" "g" "a" ...
##   ..- attr(*, "name")= chr "AAC73112"
##   ..- attr(*, "Annot")= chr ">AAC73112 cds chromosome:ASM584v2:Chromosome:190:255:1 gene:b0001 gene_
##  $ AAC73113: 'SeqFastadna' chr [1:2463] "a" "t" "g" "c" ...
##   ..- attr(*, "name")= chr "AAC73113"
##   ..- attr(*, "Annot")= chr ">AAC73113 cds chromosome:ASM584v2:Chromosome:337:2799:1 gene:b0002 gene
##  $ AAC73114: 'SeqFastadna' chr [1:933] "a" "t" "g" "g" ...
##   ..- attr(*, "name")= chr "AAC73114"
##   ..- attr(*, "Annot")= chr ">AAC73114 cds chromosome:ASM584v2:Chromosome:2801:3733:1 gene:b0003 gen
##  $ AAC73115: 'SeqFastadna' chr [1:1287] "a" "t" "g" "a" ...
##   ..- attr(*, "name")= chr "AAC73115"
##   ..- attr(*, "Annot")= chr ">AAC73115 cds chromosome:ASM584v2:Chromosome:3734:5020:1 gene:b0004 gen
##  $ AAC73116: 'SeqFastadna' chr [1:297] "g" "t" "g" "a" ...
##   ..- attr(*, "name")= chr "AAC73116"
##   ..- attr(*, "Annot")= chr ">AAC73116 cds chromosome:ASM584v2:Chromosome:5234:5530:1 gene:b0005 gen
##  $ AAC73117: 'SeqFastadna' chr [1:777] "a" "t" "g" "c" ...
##   ..- attr(*, "name")= chr "AAC73117"
##   ..- attr(*, "Annot")= chr ">AAC73117 cds chromosome:ASM584v2:Chromosome:5683:6459:-1 gene:b0006 ge
```

```
str(head(meso_cds))
```

```
## List of 6
##  $ ENSB:3oeifSflike0qpP: 'SeqFastadna' chr [1:804] "a" "t" "g" "c" ...
##   ..- attr(*, "name")= chr "ENSB:3oeifSflike0qpP"
##   ..- attr(*, "Annot")= chr ">ENSB:3oeifSflike0qpP cds primary_assembly:ASM476872v1:Chromosome:67445
##  $ ENSB:pNdelnjLN8iljUA: 'SeqFastadna' chr [1:1080] "a" "t" "g" "a" ...
##   ..- attr(*, "name")= chr "ENSB:pNdelnjLN8iljUA"
##   ..- attr(*, "Annot")= chr ">ENSB:pNdelnjLN8iljUA cds primary_assembly:ASM476872v1:Chromosome:95031
##  $ ENSB:JYa8GznKAoDE8Ks: 'SeqFastadna' chr [1:1788] "a" "t" "g" "g" ...
##   ..- attr(*, "name")= chr "ENSB:JYa8GznKAoDE8Ks"
##   ..- attr(*, "Annot")= chr ">ENSB:JYa8GznKAoDE8Ks cds primary_assembly:ASM476872v1:Chromosome:83845
##  $ ENSB:2w6OT_Sw05a4mWt: 'SeqFastadna' chr [1:195] "a" "t" "g" "g" ...
##   ..- attr(*, "name")= chr "ENSB:2w6OT_Sw05a4mWt"
##   ..- attr(*, "Annot")= chr ">ENSB:2w6OT_Sw05a4mWt cds primary_assembly:ASM476872v1:Chromosome:16366
##  $ ENSB:8mtr_HgKvIkH10u: 'SeqFastadna' chr [1:243] "a" "t" "g" "a" ...
##   ..- attr(*, "name")= chr "ENSB:8mtr_HgKvIkH10u"
##   ..- attr(*, "Annot")= chr ">ENSB:8mtr_HgKvIkH10u cds primary_assembly:ASM476872v1:Chromosome:42341
##  $ ENSB:h0wPVC7VAh6SW0a: 'SeqFastadna' chr [1:1518] "g" "t" "g" "a" ...
##   ..- attr(*, "name")= chr "ENSB:h0wPVC7VAh6SW0a"
##   ..- attr(*, "Annot")= chr ">ENSB:h0wPVC7VAh6SW0a cds primary_assembly:ASM476872v1:Chromosome:82661
```

```
ecoli_cds_lengths <- sapply(ecoli_cds, length)
meso_cds_lengths <- sapply(meso_cds, length)

boxplot(list(Ecoli = ecoli_cds_lengths, Mesomycoplasma = meso_cds_lengths),
        main = "CDS Length Comparison",
        ylab = "CDS length (bp)",
        col = c("skyblue","tomato"))
```



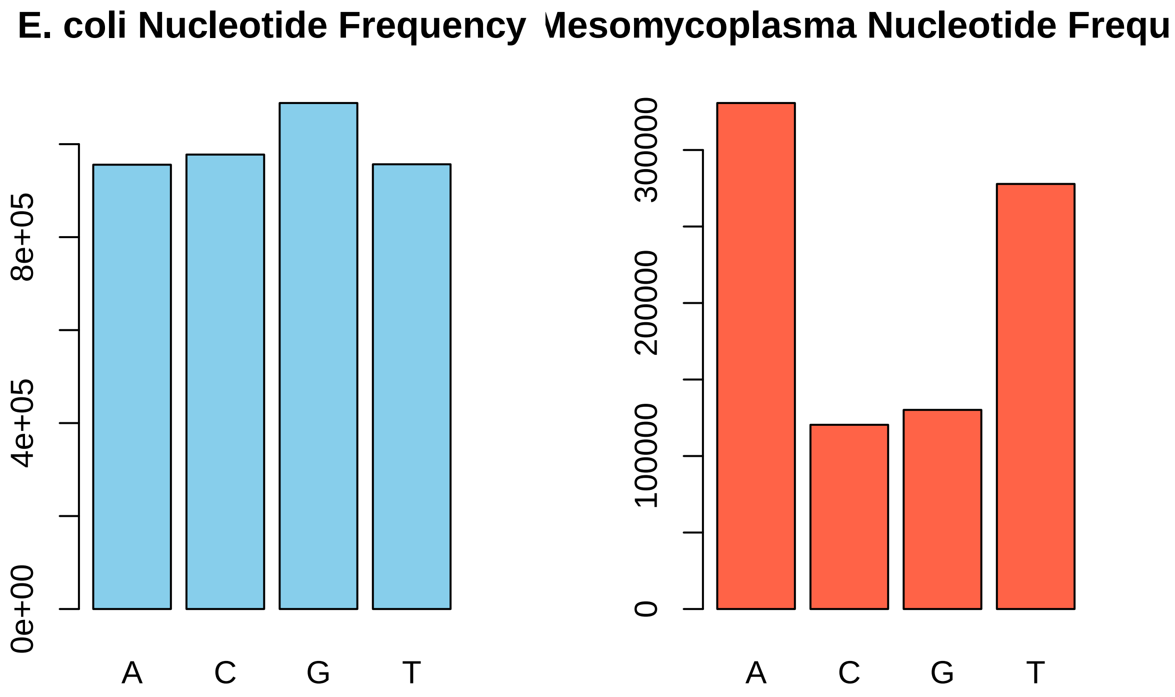
```r
# Mean and Median
mean_ecoli <- mean(ecoli_cds_lengths)
median_ecoli <- median(ecoli_cds_lengths)
mean_meso <- mean(meso_cds_lengths)
median_meso <- median(meso_cds_lengths)

cat("E. coli: mean =", mean_ecoli, ", median =", median_ecoli, "\n")
```

```
## E. coli: mean = 938.5534 , median = 831
```

```r
cat("Mesomycoplasma: mean =", mean_meso, ", median =", median_meso, "\n")
```

```
## Mesomycoplasma: mean = 1148.511 , median = 897
```

```r
# Nucleotide frequency:
ecoli_concat <- toupper(paste(sapply(ecoli_cds, function(x) paste(getSequence(x), collapse = "")), colla
meso_concat <- toupper(paste(sapply(meso_cds, function(x) paste(getSequence(x), collapse = "")), collaps

ecoli_nt_freq <- table(strsplit(ecoli_concat, "")[[1]])
meso_nt_freq <- table(strsplit(meso_concat, "")[[1]])

par(mfrow=c(1,2))
barplot(ecoli_nt_freq, main="E. coli Nucleotide Frequency", col="skyblue")
barplot(meso_nt_freq, main="Mesomycoplasma Nucleotide Frequency", col="tomato")
```

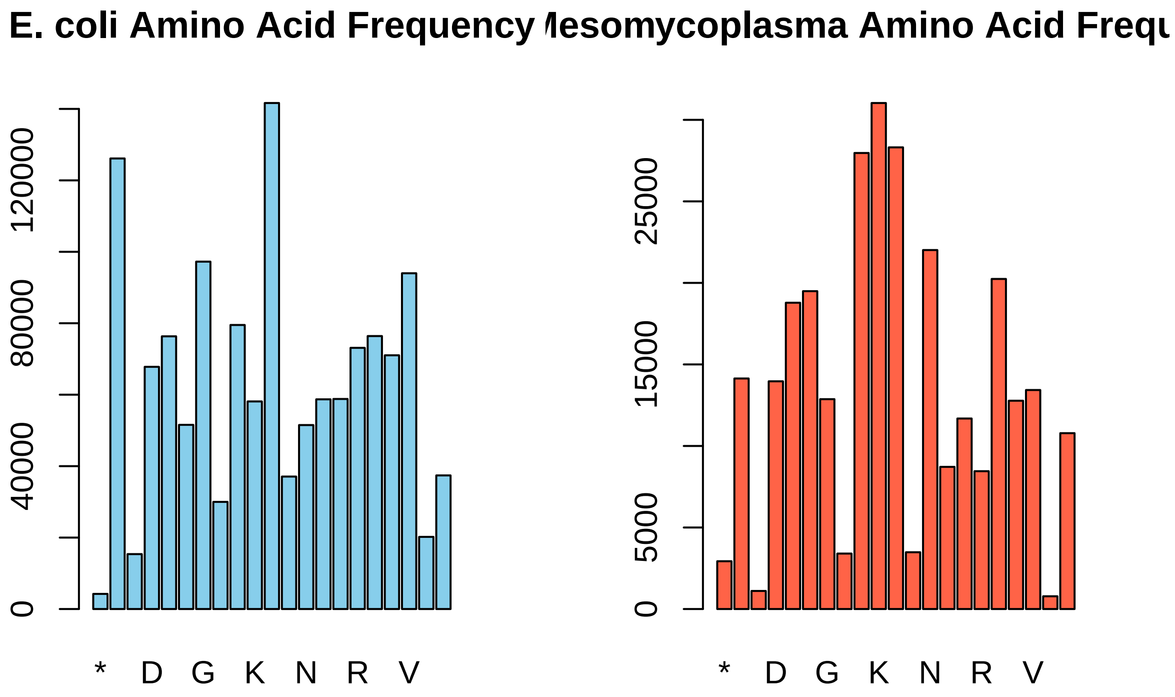


```
# Amino acid frequency (protein translation)
library(Biostrings)
```

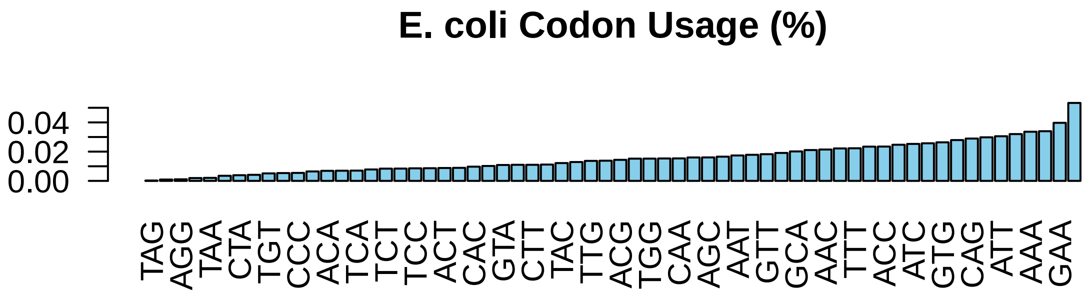
```
## Loading required package: BiocGenerics

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, basename, cbind, colnames,
##     dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,
##     union, unique, unsplit, which.max, which.min

## Loading required package: S4Vectors

## Loading required package: stats4

##
## Attaching package: 'S4Vectors'

## The following objects are masked from 'package:base':
##
```

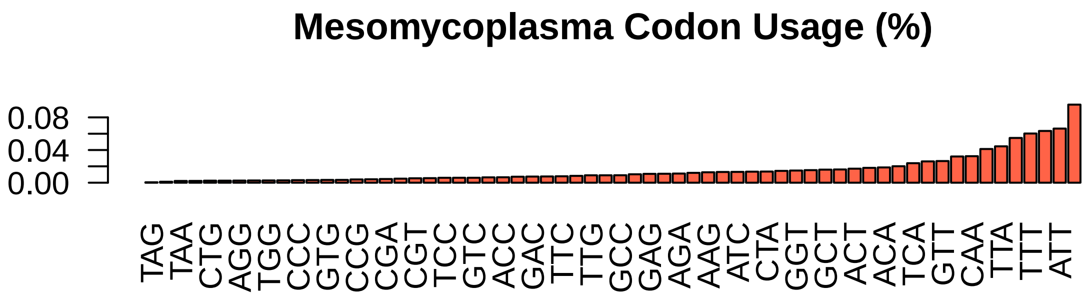
```
##     expand.grid, I, unname

## Loading required package: IRanges

##
## Attaching package: 'IRanges'

## The following object is masked from 'package:R.oo':
##
##     trim

## Loading required package: XVector

## Loading required package: GenomeInfoDb

##
## Attaching package: 'Biostrings'

## The following object is masked from 'package:seqinr':
##
##     translate

## The following object is masked from 'package:base':
##
##     strsplit
```

```
ecoli_AA <- lapply(ecoli_cds, function(x) as.character(translate(DNAString(paste(getSequence(x), collap
ecoli_AA_concat <- paste(unlist(ecoli_AA), collapse = "")
ecoli_aa_freq <- table(strsplit(ecoli_AA_concat, "")[[1]])

Mesomycoplasma_AA <- lapply(meso_cds, function(x) as.character(translate(DNAString(paste(getSequence(x)
Mesomycoplasma_AA_concat <- paste(unlist(Mesomycoplasma_AA), collapse = "")
Mesomycoplasma_aa_freq <- table(strsplit(Mesomycoplasma_AA_concat, "")[[1]])

par(mfrow=c(1,2))
barplot(ecoli_aa_freq, main="E. coli Amino Acid Frequency", col="skyblue")
barplot(Mesomycoplasma_aa_freq, main="Mesomycoplasma Amino Acid Frequency", col="tomato")
```



```r
# Codon frequency function
codon_table <- function(seq_list) {
  codons <- unlist(lapply(seq_list, function(x) {
    s <- toupper(paste(getSequence(x), collapse = ""))
    groups <- substring(s, seq(1,nchar(s)-2,3), seq(3, nchar(s), 3))
    groups
  }))
  table(codons)
}

ecoli_codon_freq <- codon_table(ecoli_cds)
meso_cds <- read.fasta(file = "Mesomycoplasma_cds.fa", seqtype = "DNA")
meso_cds_lengths <- sapply(meso_cds, length)
meso_codon_freq <- codon_table(meso_cds)

# Convert to proportion:
ecoli_codon_prop <- ecoli_codon_freq / sum(ecoli_codon_freq)
meso_codon_prop <- meso_codon_freq / sum(meso_codon_freq)

# Compare using barplots
par(mfrow=c(2,1))
barplot(sort(ecoli_codon_prop), las=2, main="E. coli Codon Usage (%)", col="skyblue")
barplot(sort(meso_codon_prop), las=2, main="Mesomycoplasma Codon Usage (%)", col="tomato")
```





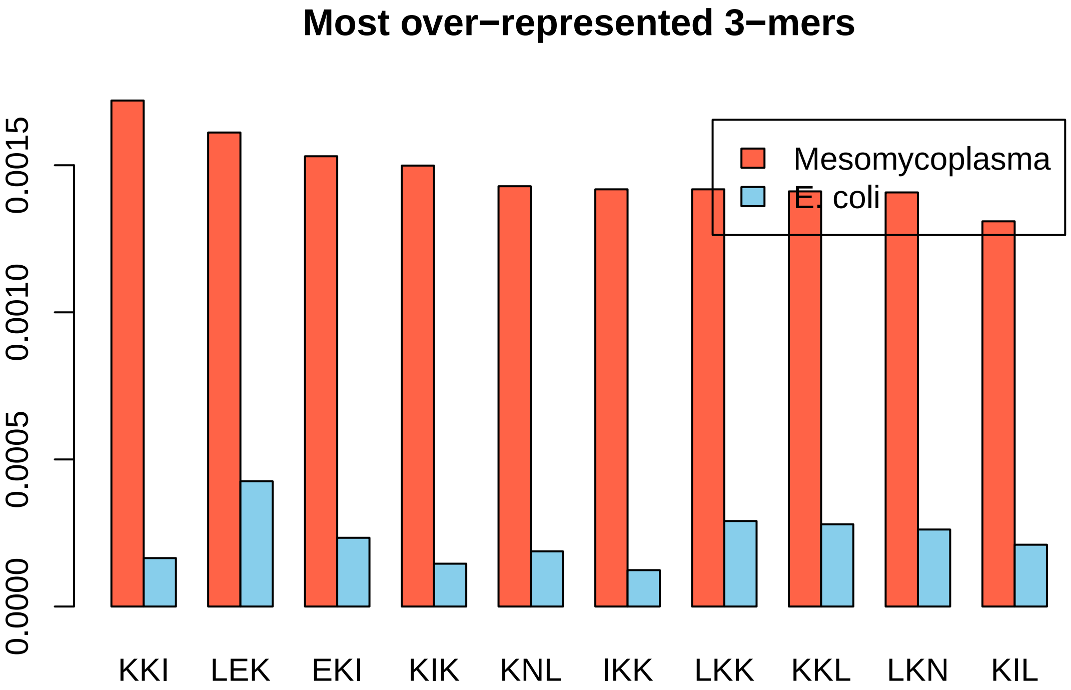
```r
# K-mer counting function (for example, k = 3)
kmer_counter <- function(seq_vec, k=3) {
  kmers <- unlist(lapply(seq_vec, function(s) {
    n <- nchar(s)
    if (n < k) return(character(0))
    sapply(1:(n-k+1), function(i) substr(s, i, i+k-1))
  }))
  table(kmers)
}
# For E. coli
ecoli_aa_strings <- sapply(ecoli_AA, paste, collapse = "")
k=3
ecoli_kmer <- kmer_counter(ecoli_aa_strings, k)
Mesomycoplasma_aa_strings <- sapply(Mesomycoplasma_AA, paste, collapse = "")
Mesomycoplasma_kmer <- kmer_counter(Mesomycoplasma_aa_strings, k)

ecoli_kmer_prop <- ecoli_kmer / sum(ecoli_kmer)
Mesomycoplasma_kmer_prop <- Mesomycoplasma_kmer / sum(Mesomycoplasma_kmer)

# Most over- and under-represented in Mesomycoplasma:
Mesomycoplasma_kmer_sorted <- sort(Mesomycoplasma_kmer_prop, decreasing=TRUE)
over_Mesomycoplasma <- head(Mesomycoplasma_kmer_sorted, 10)
under_Mesomycoplasma <- tail(Mesomycoplasma_kmer_sorted, 10)

print(over_Mesomycoplasma)
```

```
## kmers
##         KKI         LEK         EKI         KIK         KNL         IKK
## 0.001720107 0.001611284 0.001530544 0.001498950 0.001428742 0.001418211
##         LKK         KKL         LKN         KIL
## 0.001418211 0.001411190 0.001407679 0.001309388
```

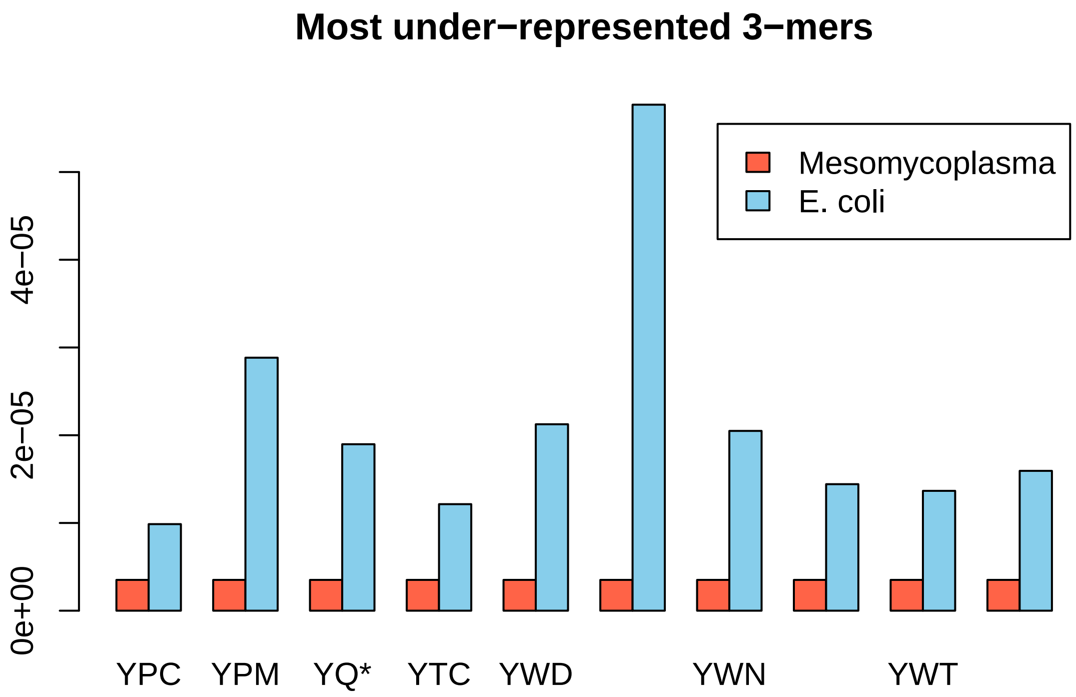
```
print(under_Mesomycoplasma)
```

```
## kmers
##          YPC          YPM          YQ*          YTC          YWD          YWL
## 3.510422e-06 3.510422e-06 3.510422e-06 3.510422e-06 3.510422e-06 3.510422e-06
##          YWN          YWP          YWT          YWY
## 3.510422e-06 3.510422e-06 3.510422e-06 3.510422e-06
```

```
# Compare to E. coli for the same k-mers:
ecoli_over_match <- ecoli_kmer_prop[names(over_Mesomycoplasma)]
ecoli_under_match <- ecoli_kmer_prop[names(under_Mesomycoplasma)]

# Plot comparisons:
barplot(rbind(as.numeric(over_Mesomycoplasma), as.numeric(ecoli_over_match)),
        beside=TRUE, names.arg=names(over_Mesomycoplasma),
        legend.text=c("Mesomycoplasma", "E. coli"),
        main="Most over-represented 3-mers", col=c("tomato","skyblue"))
```



```
barplot(rbind(as.numeric(under_Mesomycoplasma), as.numeric(ecoli_under_match)),
        beside=TRUE, names.arg=names(under_Mesomycoplasma),
        legend.text=c("Mesomycoplasma", "E. coli"),
        main="Most under-represented 3-mers", col=c("tomato","skyblue"))
```


Most under−represented 3−mers
Mesomycoplasma
E. coli
0e+00
2e−05
4e−05
YPC
YPM
YQ*
YTC
YWD
YWN
YWT