# Assessment 4

Nicole Hinojosa, Saqeena Thapa, Vanshika

2024-10-03

## Part 1: Importing files, data wrangling, mathematical operations, plots and saving code on GitHub.

### Introduction

This report analyzes RNA-seq count data for gene expression and tree circumference measurements at two different sites over a 20-year period.

### Task 1: RNA-seq count data for gene expression, high and low expression of 3 genes.

**1.1 Read in the file "gene_expression.tsv", making the gene identifiers the row names. Show a table of values for the first six genes.**

1) Load libraries

```
#RNA-seq Count Data Analysis
#Load necessary libraries
library(R.utils)
#use BiocManager::install("Biostrings") if it is not already installed in your Rstudio
library(Biostrings)
library(seqinr)
library(dplyr)
library(ggplot2)
library(readr)
library(tidyr)
```

2) Read in the gene expression data

```
#Download the data from the github link provided
URL = "https://raw.githubusercontent.com/ghazkha/Assessment4/refs/heads/main/gene_expression.tsv"
download.file(URL, destfile = "gene_expression.tsv")

# Read the downloaded TSV file into R
gene_expression <- read.table("gene_expression.tsv", header = TRUE, sep = "\t", row.names = 1)
```

3) 1st First 6 rows of the gene_expression data

```
##                                GTEX.1117F.0226.SM.5GZZ7 GTEX.1117F.0426.SM.5EGHI
## ENSG00000223972.5_DDX11L1                             0                        0
## ENSG00000227232.5_WASH7P                            187                      109
## ENSG00000278267.1_MIR6859-1                           0                        0
## ENSG00000243485.5_MIR1302-2HG                         1                        0
## ENSG00000237613.2_FAM138A                             0                        0
## ENSG00000268020.3_OR4G4P                             0                        1
##                               GTEX.1117F.0526.SM.5EGHJ
## ENSG00000223972.5_DDX11L1                            0
## ENSG00000227232.5_WASH7P                           143
## ENSG00000278267.1_MIR6859-1                          1
## ENSG00000243485.5_MIR1302-2HG                        0
## ENSG00000237613.2_FAM138A                            0
## ENSG00000268020.3_OR4G4P                             0
```

### 1.2 Make a new column which is the mean of the other columns. Show a table of values for the first six genes.

Calculate Mean Expression

```
##                               GTEX.1117F.0226.SM.5GZZ7 GTEX.1117F.0426.SM.5EGHI
## ENSG00000223972.5_DDX11L1                            0                        0
## ENSG00000227232.5_WASH7P                           187                      109
## ENSG00000278267.1_MIR6859-1                          0                        0
## ENSG00000243485.5_MIR1302-2HG                        1                        0
## ENSG00000237613.2_FAM138A                            0                        0
## ENSG00000268020.3_OR4G4P                             0                        1
##                               GTEX.1117F.0526.SM.5EGHJ mean_expression
## ENSG00000223972.5_DDX11L1                            0       0.0000000
## ENSG00000227232.5_WASH7P                           143     146.3333333
## ENSG00000278267.1_MIR6859-1                          1       0.3333333
## ENSG00000243485.5_MIR1302-2HG                        0       0.3333333
## ENSG00000237613.2_FAM138A                            0       0.0000000
## ENSG00000268020.3_OR4G4P                             0       0.3333333
```

### 1.3 List the 10 genes with the highest mean expression.

Identify Top 10 Genes

```
##                           GTEX.1117F.0226.SM.5GZZ7 GTEX.1117F.0426.SM.5EGHI
## ENSG00000198804.2_MT-CO1                    267250                  1101779
## ENSG00000198886.2_MT-ND4                    273188                   991891
## ENSG00000198938.2_MT-CO3                    250277                  1041376
## ENSG00000198888.2_MT-ND1                    243853                   772966
## ENSG00000198899.2_MT-ATP6                   141374                   696715
## ENSG00000198727.2_MT-CYB                    127194                   638209
## ENSG00000198763.3_MT-ND2                    159303                   543786
## ENSG00000211445.11_GPX3                     464959                    39396
## ENSG00000198712.1_MT-CO2                    128858                   545360
## ENSG00000156508.17_EEF1A1                   317642                    39573
##                           GTEX.1117F.0526.SM.5EGHJ mean_expression
## ENSG00000198804.2_MT-CO1                    218923        529317.3
## ENSG00000198886.2_MT-ND4                    277628        514235.7
## ENSG00000198938.2_MT-CO3                    223178        504943.7
## ENSG00000198888.2_MT-ND1                    194032        403617.0
## ENSG00000198899.2_MT-ATP6                   151166        329751.7
## ENSG00000198727.2_MT-CYB                    141359        302254.0
## ENSG00000198763.3_MT-ND2                    149564        284217.7
## ENSG00000211445.11_GPX3                     306070        270141.7
## ENSG00000198712.1_MT-CO2                    122816        265678.0
## ENSG00000156508.17_EEF1A1                   339347        232187.3
```

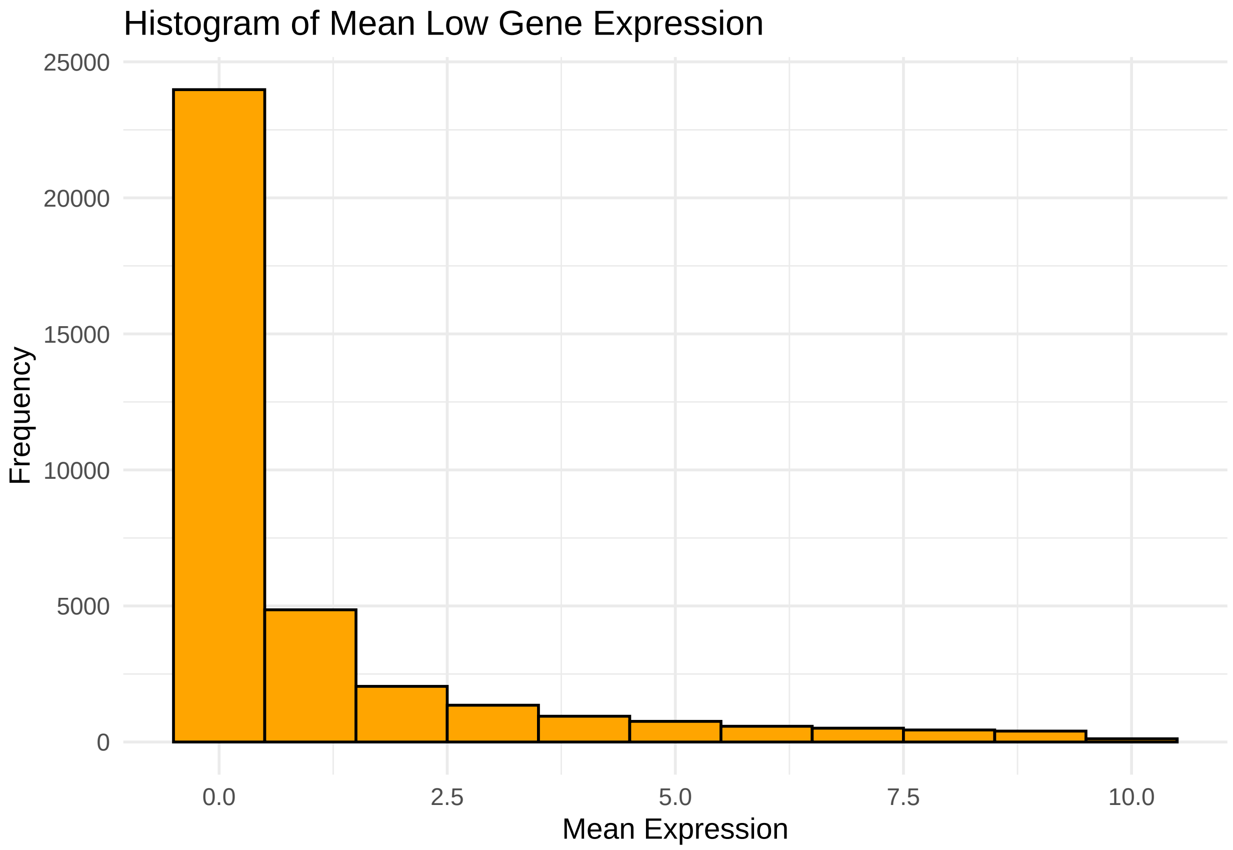
### 1.4 Determine the number of genes with a mean <10.

Count Genes with Low Expression (Mean < 10)

```
## [1] 35988
```

### 1.5 Make a histogram plot of the mean values and include it into your report.

Histogram of <10, Mean Values



```
## Saving 6.5 x 4.5 in image
```

## Task 2: Tree circumference measurements over 20 years.

### 2.1 Import "growth_data.csv" file into an R object. What are the column names?

Read Data to perform a Tree Circumference Data Analysis

```r
# Read in the growth data
#Download the data from the github link provided
URL = "https://raw.githubusercontent.com/ghazkha/Assessment4/refs/heads/main/growth_data.csv"
download.file(URL, destfile = "growth_data.csv")

# Read the downloaded TSV file into R
growth_data <- read.csv("growth_data.csv")

head(growth_data)
```

```
##        Site TreeID Circumf_2005_cm Circumf_2010_cm Circumf_2015_cm
## 1 northeast   A012            5.2           10.1           19.9
## 2 southwest   A039            4.9            9.6           18.9
## 3 southwest   A010            3.7            7.3           14.3
## 4 northeast   A087            3.8            6.5           10.9
## 5 southwest   A074            3.8            6.4           10.9
## 6 northeast   A008            5.9           10.0           16.8
##   Circumf_2020_cm
## 1            38.9
## 2            37.0
## 3            28.1
## 4            18.5
## 5            18.4
## 6            28.4
```

```
# Show column names
cat("The column names are:", colnames(growth_data))
```

```
## The column names are: Site TreeID Circumf_2005_cm Circumf_2010_cm Circumf_2015_cm Circumf_2020_cm
```

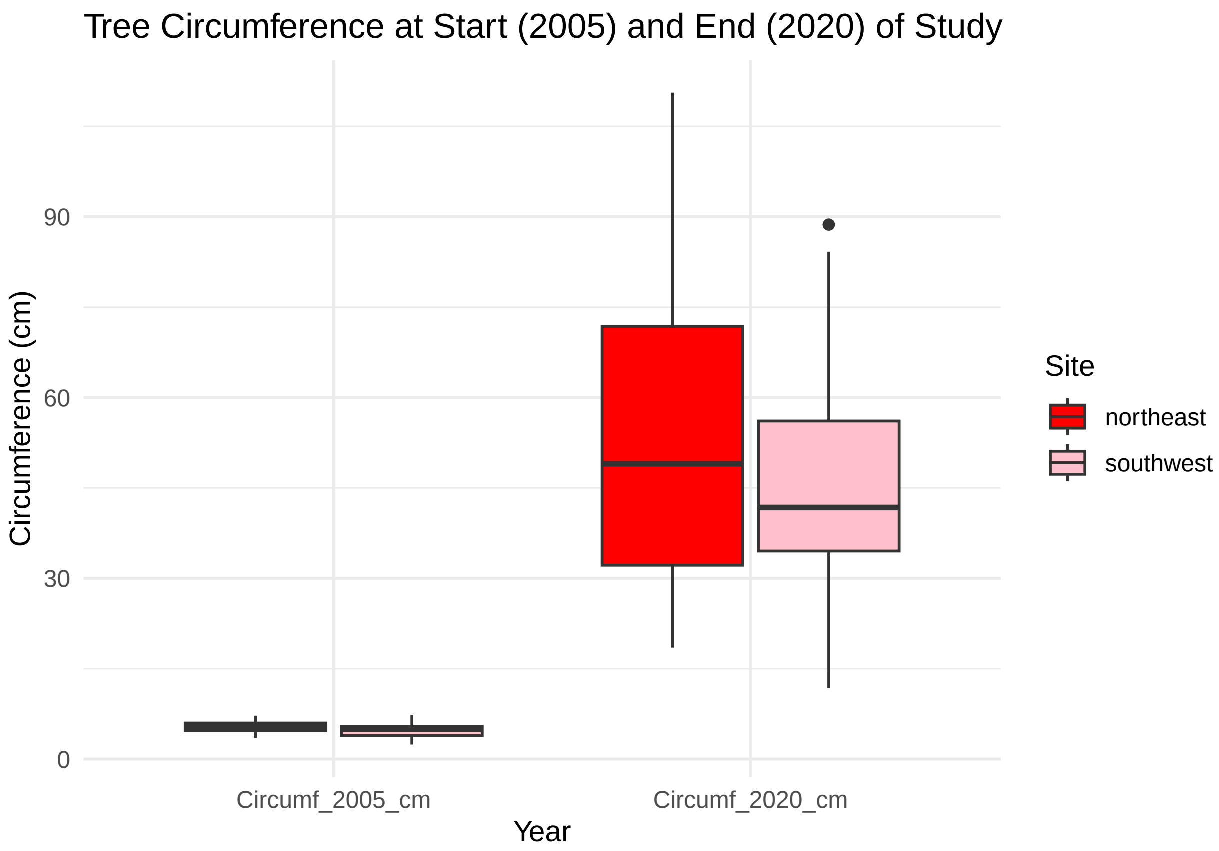
### 2.2 Calculate the mean and standard deviation of tree circumference at the start and end of the study at both sites.

Statistics

```
##   mean_start_2005_southwest sd_start_2005_southwest mean_start_2005_northeast
## 1                     4.862                1.147471                     5.292
##   sd_start_2005_northeast mean_end_2020_southwest sd_end_2020_southwest
## 1               0.9140267                  45.596              17.87345
##   mean_end_2020_northeast sd_end_2020_northeast
## 1                  54.228              25.22795
```

### 2.3 Make a box plot of tree circumference at the start and end of the study at both sites.

Boxplot of circumferences



```
## Saving 6.5 x 4.5 in image
```

**2.4 Calculate the mean growth over the last 10 years at each site.**

Mean Growth Calculation

```
## # A tibble: 2 x 2
##   Site      mean_growth
##   <chr>           <dbl>
## 1 northeast        42.9
## 2 southwest        35.5
```

**2.5 Use the t.test to estimate the p-value that the 10 year growth is different at the two sites.**

T-Test for Growth Difference

```
## 
##  Welch Two Sample t-test
## 
## data:  Growth_10_years by Site
## t = 1.8882, df = 87.978, p-value = 0.06229
## alternative hypothesis: true difference in means between group northeast and group southwest is not e
## 95 percent confidence interval:
##  -0.3909251 15.2909251
## sample estimates:
## mean in group northeast mean in group southwest 
##                   42.94                   35.49
```

**Interpretation:** p-value: The p-value of 0.06229 suggests that the difference in mean growth between the two sites is not statistically significant at the conventional alpha level of 0.05. However, it is close to this threshold, indicating a potential trend toward significance.

Mean Comparison: The mean growth in the northeast (42.94 cm) is higher than that in the southwest (35.49 cm). This suggests that trees in the northeast experienced greater growth compared to those in the southwest over the last 10 years.

Confidence Interval: The confidence interval includes zero, which means we cannot conclusively say that there is a true difference in growth between the two sites. The upper limit (15.29 cm) indicates that, while the northeast shows higher growth, it is possible that the actual difference might be minimal or even negative.

# Part 2: Examining biological sequence diversity

## Introduction

This report compares the sequence features of *Streptacidiphilus jiangxiensis* (GCA_900109465) with *Escherichia coli*. *Escherichia coli* is a Gram-negative, rod-shaped bacterium commonly found in the intestines of warm-blooded organisms, playing a vital role in gut health. While some strains of *Escherichia coli* can cause illnesses, the bacterium is extensively used as a model organism in molecular biology due to its relatively simple genome and well-studied genetics. In contrast, *Streptacidiphilus jiangxiensis* is a Gram-positive bacterium isolated from acidic environments (Huang et al., 2004). This species is characterized by its unique metabolic pathways and adaptations to specific ecological niches, which may hold potential for applications in bioremediation or antibiotic development.

## Questions:

**1) Download the whole set of coding DNA sequences for E. coli and your organism of interest. How many coding sequences are present in these organisms? Present this in the form of a table. Describe any differences between the two organisms.**

Sequences

```r
# URLs for the coding DNA sequences
URL_Ecoli <- "https://ftp.ensemblgenomes.ebi.ac.uk/pub/bacteria/release-59/fasta/bacteria_117_collectio
URL_Streptacidiphilus <- "https://ftp.ensemblgenomes.ebi.ac.uk/pub/bacteria/release-59/fasta/bacteria_5

# Downloading the sequences
download.file(URL_Ecoli, destfile = "e_coli_cds.fa.gz")
download.file(URL_Streptacidiphilus, destfile = "streptacidiphilus_cds.fa.gz")

#Decompress the files

gunzip("e_coli_cds.fa.gz")
gunzip("streptacidiphilus_cds.fa.gz")

# Reading the sequences
ecoli_seqs <- seqinr::read.fasta ("e_coli_cds.fa")
streptacidiphilus_seqs <- seqinr::read.fasta ("streptacidiphilus_cds.fa")
```

CDS count

```
##                         Organism Coding_Sequences
## 1               Escherichia coli             4931
## 2 Streptacidiphilus jiangxiensis             8650
```

**Answer:** *Escherichia coli* contains 4,931 coding sequences while *Streptacidiphilus jiangxiensis* has a substantially higher count (8,650) of coding sequences, points to a significant disparity in genetic diversity between the two bacterial species. This greater number of coding sequences in *Streptacidiphilus jiangxiensis* is indicative of a more complex genetic framework, which may translate into enhanced functional capabilities and a broader range of physiological adaptations (Wright, 1990, Malik et al., 2020).

The expanded repertoire of genes in *Streptacidiphilus jiangxiensis* likely contributes to its metabolic versatility, enabling it to thrive in diverse environments. This versatility allows *Streptacidiphilus jiangxiensis* to exploit various substrates, potentially including organic compounds found in soil or plant matter, that *Escherichia coli* might not utilize as efficiently. For example, *Streptacidiphilus jiangxiensis* may possess unique enzymes or metabolic pathways that allow it to break down complex carbohydrates, synthesize essential nutrients, or produce secondary metabolites, such as antimicrobial compounds, which can offer competitive advantages in its ecological niche (Wright, 1990).

Moreover, the increased number of coding sequences may also reflect evolutionary adaptations to environmental pressures. In soil ecosystems, where nutrient availability can fluctuate, the ability to produce a wide range of enzymes and metabolites could enhance survival and reproduction (Harman and Uphoff, 2019). *Streptacidiphilus jiangxiensis* may engage in symbiotic relationships with plants or other soil microorganisms, leveraging its genetic diversity to facilitate nutrient exchange or improve soil health (Cong et al., 2021). In contrast, *Escherichia coli*, while highly adaptable and successful in its own right, is often more specialized for life in nutrient-rich environments, such as the gastrointestinal tract of mammals. Thus, the greater coding sequence count in *Streptacidiphilus jiangxiensis* underscores its potential for metabolic innovation and ecological resilience, reflecting a sophisticated evolutionary response to its environment (Wright, 1990; Malik et al., 2020).

**2) How much coding DNA is there in total for these two organisms? Present this in the form of a table. Describe any differences between the two organisms.**

Total Coding DNA Length

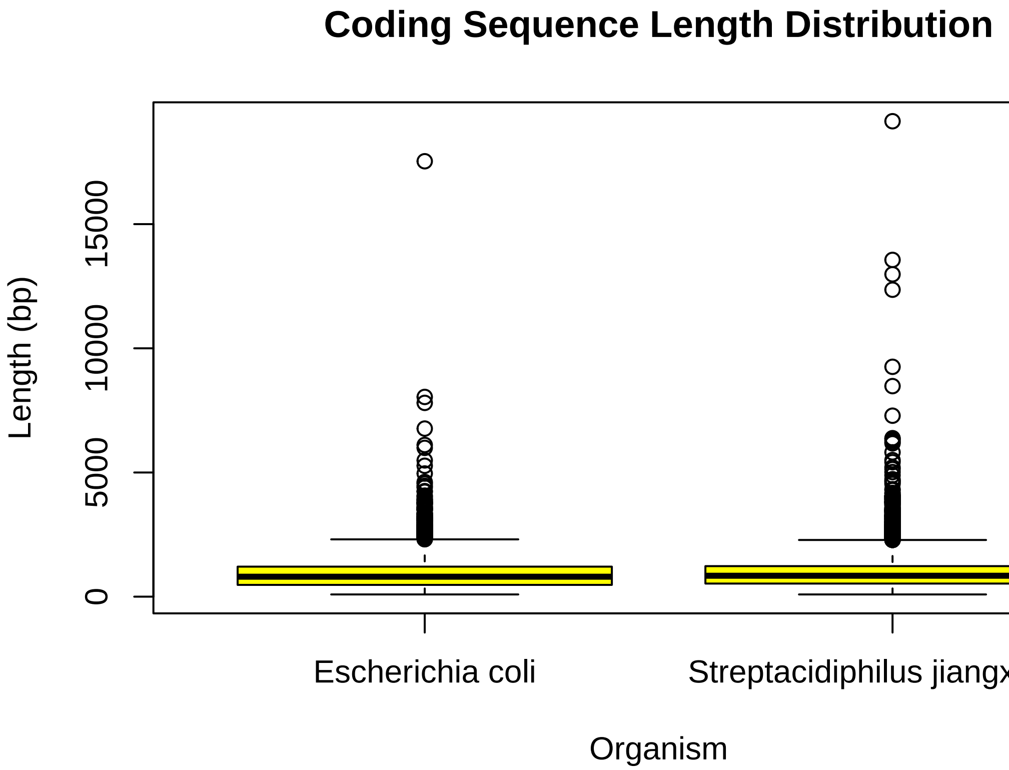
```
##                         Organism Total_Length
## 1               Escherichia coli      4593474
## 2 Streptacidiphilus jiangxiensis      8422779
```

**Answer:** The genomic analysis reveals that *Escherichia coli* comprises approximately 4,593,474 base pairs of coding DNA, whereas *Streptacidiphilus jiangxiensis* possesses a significantly larger genomic footprint of about 8,422,779 base pairs. This substantial difference in coding DNA indicates a more complex genome in *Streptacidiphilus*, which is often associated with greater metabolic diversity. A larger genomic content can provide a broader array of genes, facilitating the synthesis of diverse proteins that are crucial for various metabolic pathways and biochemical processes (Bentley, 2009).

The expanded coding capacity of *Streptacidiphilus* likely equips it with the ability to thrive in complex and potentially harsh environmental conditions. For instance, the organism might possess genes that allow it to metabolize a wider range of substrates, adapt to changes in nutrient availability, or withstand environmental stresses, such as acidity or high salinity (Rasko et al., 2008). Such metabolic versatility could enable *Streptacidiphilus* to exploit ecological niches that are less accessible to simpler organisms like *E. coli*, which tends to thrive in nutrient-rich environments typical of the mammalian gut (Lozica et al., 2022).

Furthermore, the increased genetic content in *Streptacidiphilus* may include genes responsible for specialized functions, such as biosynthesis of secondary metabolites, antibiotic resistance, or symbiotic interactions with other organisms (Jarocki et al., 2019). These attributes are particularly valuable for survival in competitive ecosystems where resource availability can be unpredictable. In essence, the larger coding DNA in *Streptacidiphilus jiangxiensis* reflects an evolutionary strategy that enhances its adaptability and functional repertoire, illustrating how genomic complexity can influence ecological success and resilience (Huang et al., 2004).

**3) Calculate the length of all coding sequences in these two organisms. Make a boxplot of coding sequence length in these organisms. What is the mean and median coding sequence length of these two organisms? Describe any differences between the two organisms.**



Coding Sequence Length Distribution

Mean and Median Coding Sequence Length

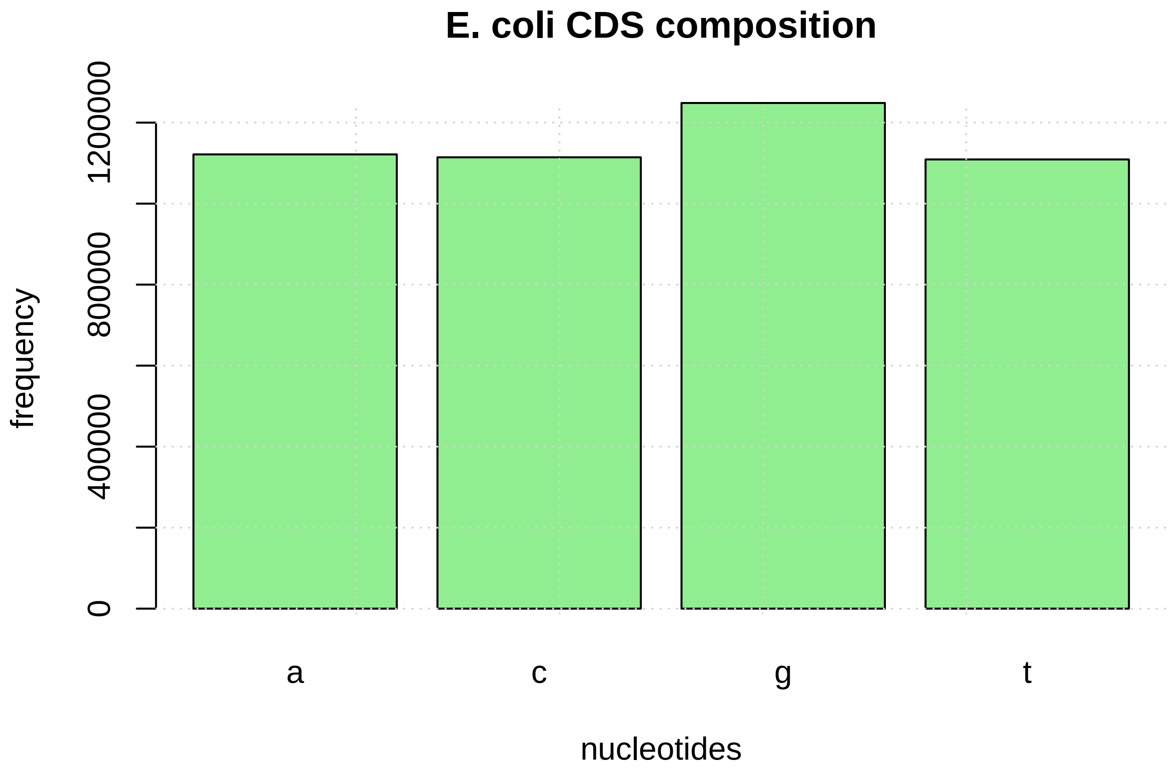
```
##                         Organism Mean_Length Median_Length
## 1               Escherichia coli    931.5502           804
## 2 Streptacidiphilus jiangxiensis    973.7317           843
```

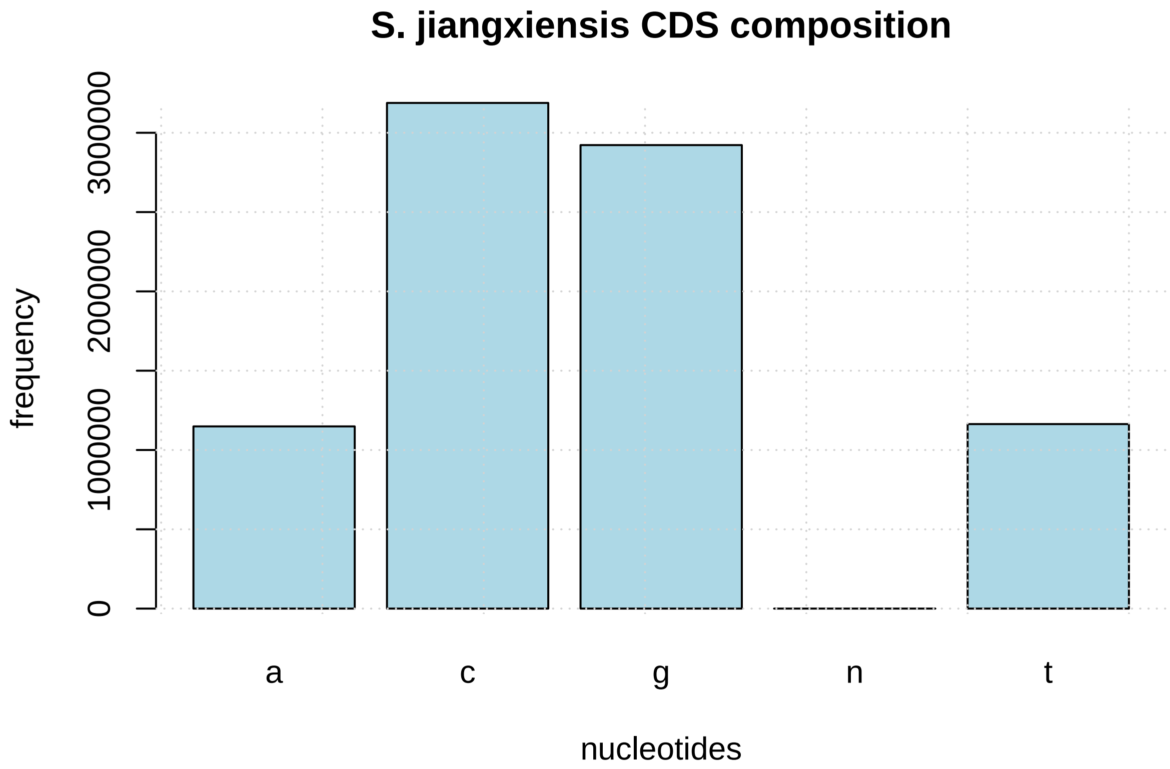
**Answer:** The comparison of coding sequence lengths between *Escherichia coli* and *Streptacidiphilus jiangxiensis* reveals significant insights into their genomic architectures and evolutionary adaptations. *Escherichia coli*, with a mean coding sequence length of 931.55 bp and a median of 804 bp, is known for its streamlined genome, which is optimized for rapid growth and efficient protein production in the nutrient-rich environments of mammalian intestines (Koonin, 2009). This bacterium's genetic simplicity has made it a model organism for studying fundamental biological processes, allowing researchers to explore genetic functions and interactions in a relatively uncomplicated context (Lizana and Schwartz, 2024).

In contrast, *Streptacidiphilus jiangxiensis* presents a longer mean coding sequence of 973.73 bp and a median of 843 bp, suggesting a more intricate gene structure. This complexity may be indicative of specialized adaptations to its unique ecological niche, which includes survival in acidic environments (Malik et al., 2020). Longer coding sequences can allow for the encoding of more extensive and functionally diverse proteins, potentially enabling the bacterium to exploit a broader range of substrates or cope with environmental stresses (Maharjan and Ferenci, 2014). The differences in coding sequence length may reflect distinct evolutionary pressures faced by these organisms, with *Escherichia coli* evolving for rapid proliferation and *Streptacidiphilus jiangxiensis* potentially developing complex metabolic pathways or interactions to thrive in less hospitable conditions (Wan et al., 2022). This variation underscores the diverse functional requirements and ecological strategies that shape the genomes of bacteria, highlighting how environmental factors influence genetic structure and complexity (Chan et al., 2018).

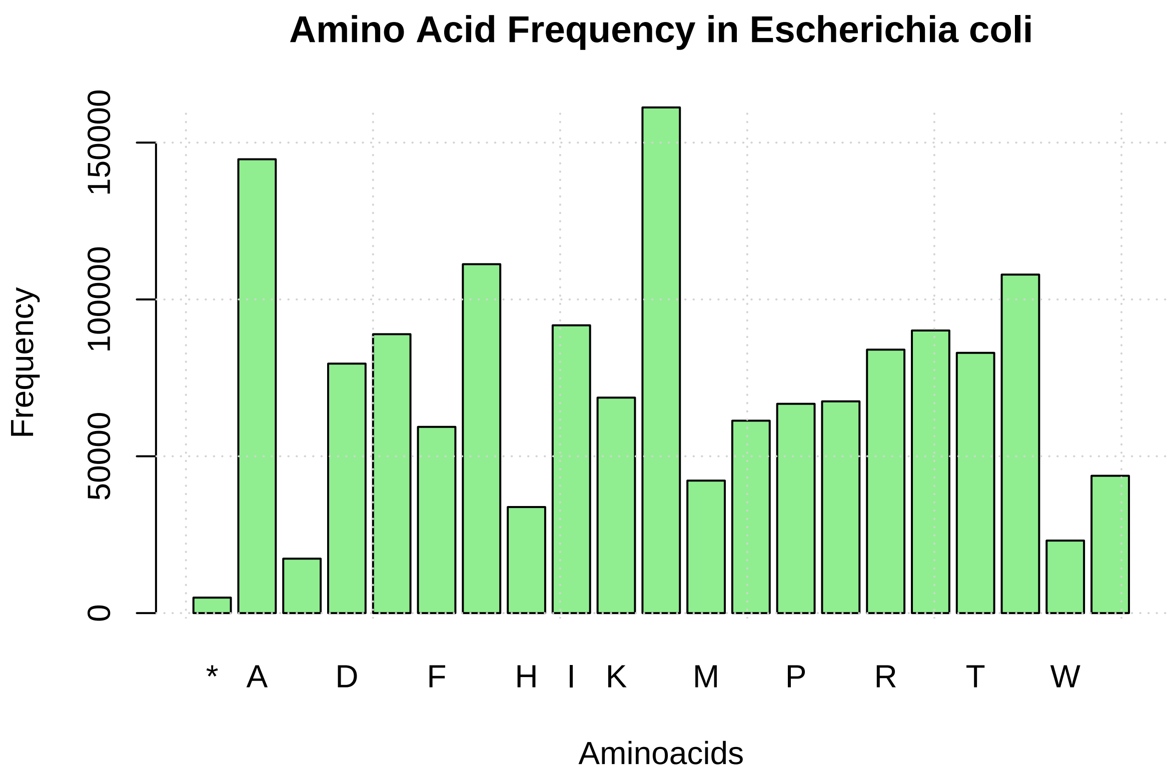
**4) Calculate the frequency of DNA bases in the total coding sequences for both organisms. Perform the same calculation for the total protein sequence. Create bar plots for nucleotide and amino acid frequency. Describe any differences between the two organisms.**

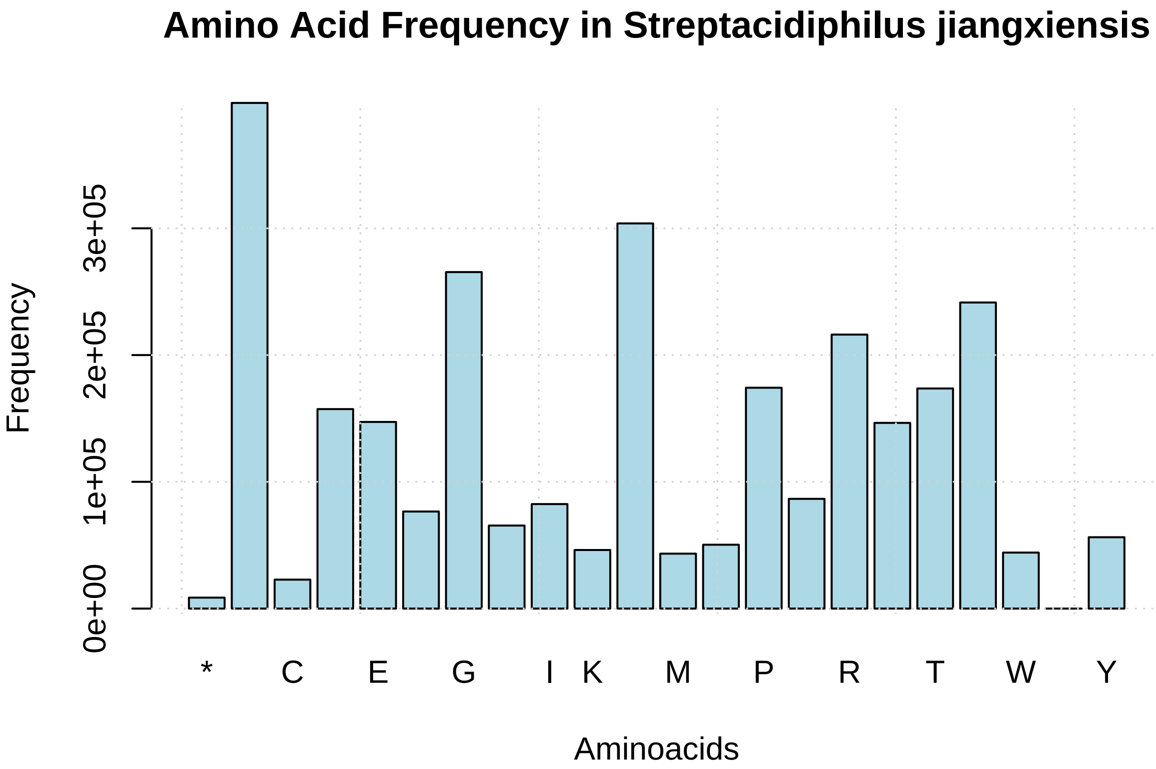
Frequency of DNA Bases





Amino Acid Frequency


Amino Acid Frequency in Escherichia coli
Frequency
0
50000
100000
150000
*
A
D
F
H
I
K
M
P
R
T
W
Aminoacids



**Answer:** The observation that *Escherichia coli* exhibits a relatively even distribution of nucleotides—where adenine (A), thymine (T), cytosine (C), and guanine (G) are present in comparable proportions—contrasts sharply with the significantly higher CG content observed in *Streptacidiphilus jiangxiensis*. This difference in nucleotide composition can have profound implications for the evolutionary strategies and ecological niches occupied by these two bacteria (Malik et al., 2020). The balanced nucleotide composition in *E. coli* reflects its adaptation to a diverse range of environments, particularly in the nutrient-rich gastrointestinal tracts of mammals, where rapid replication and metabolic efficiency are advantageous (Alteri and Mobley, 2012). In contrast, the elevated CG content in *Streptacidiphilus* may indicate specialized adaptations to more extreme or variable environmental conditions, such as those found in soil ecosystems. Higher GC content can contribute to increased stability of the DNA helix and may be linked to resistance against environmental stresses, such as desiccation or temperature fluctuations (Šmarda et al., 2014). Moreover, elevated CG content can affect gene expression and regulation, allowing *Streptacidiphilus* to potentially encode proteins that facilitate survival in nutrient-poor or competitive environments. Thus, the differences in nucleotide composition not only reflect the evolutionary histories of these bacteria but also underscore their distinct ecological strategies and adaptations to their respective habitats (Hu et al., 2022).

The differences in amino acid frequency between *Escherichia coli* and *Streptacidiphilus jiangxiensis* not only reveal evolutionary adaptations but also highlight functional specializations that may be crucial for each organism's survival in their respective environments. In *E. coli*, the five most frequent amino acids—leucine (L), alanine (A), glycine (G), valine (V), and isoleucine (I)—suggest a protein composition that supports its versatile metabolic functions and adaptability. Leucine, being the most abundant, is known for its role in protein synthesis and cellular signaling, indicating a potential emphasis on growth and reproduction (Maser et al., 2020).

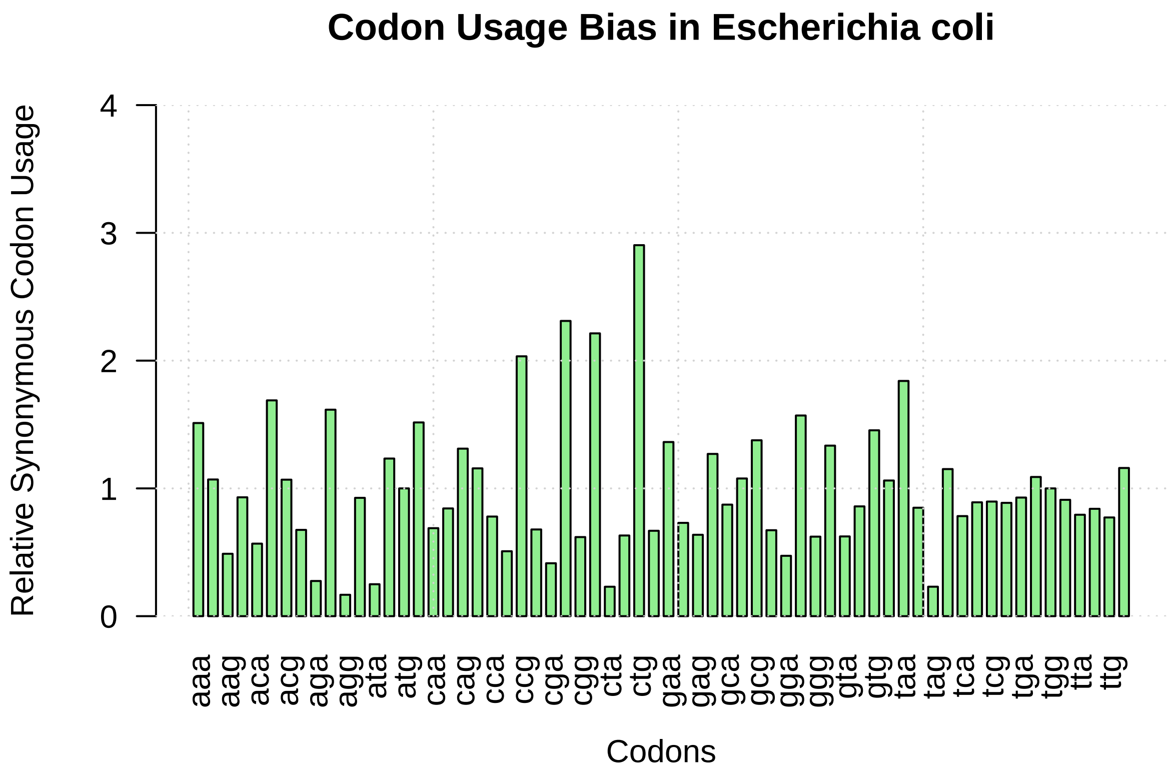
Conversely, *Streptacidiphilus jiangxiensis* shows a slightly different pattern, with alanine (A), leucine (L), glycine (G), valine (V), and arginine (R) as the five most frequent amino acids. The presence of arginine in this top five may indicate specific adaptations related to stress responses or metabolic processes unique to its ecological niche. The dominance of alanine, along with the other branched-chain amino acids, suggests that *Streptacidiphilus* might be optimized for particular biochemical pathways that enable it to thrive in potentially nutrient-poor or competitive environments (Koonin, 2009).

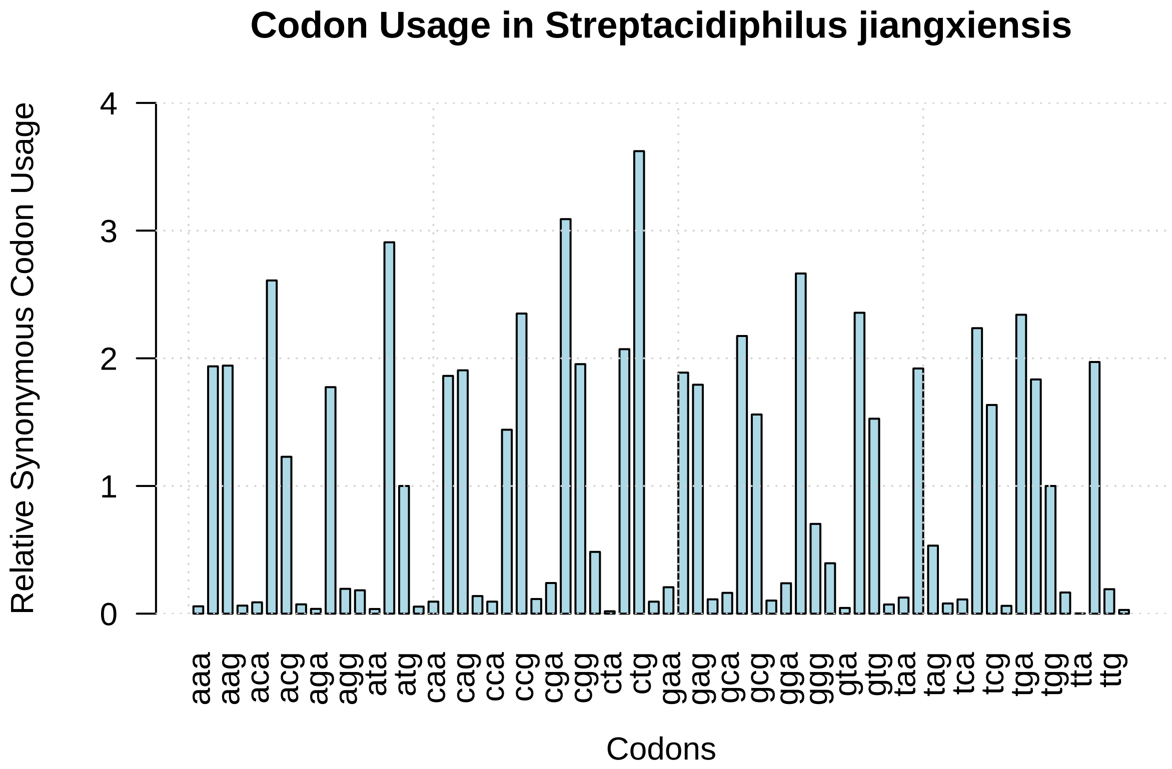
These variations in amino acid frequencies not only reflect the distinct metabolic capabilities of each organism but also offer insights into how they might respond to environmental challenges. For instance, the higher frequency of arginine in *Streptacidiphilus* may confer advantages in environments requiring enhanced nitrogen metabolism or stress resilience (Chen et al., 2021). Thus, examining the amino acid composition provides a deeper understanding of the functional diversity and ecological adaptations of these two bacteria, shedding light on their evolutionary trajectories in different ecological contexts (Monteiro et al., 2023).

**5) Create a codon usage table and quantify the codon usage bias among all coding sequences. Describe any differences between the two organisms with respect to their codon usage bias. Provide charts to support your observations.**

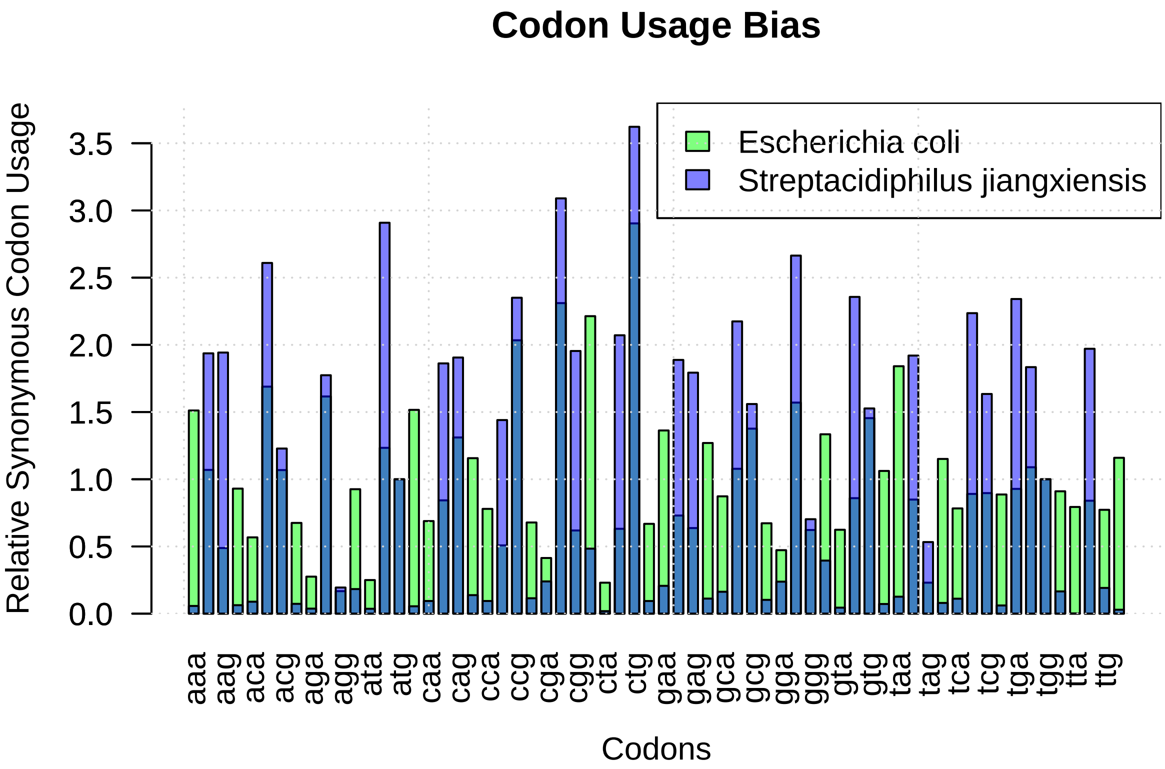
Codon Usage Bias

```
##      AA codon   eff        freq E. coli RSCU S. jiangxiensis RSCU
## aaa Lys   aaa 51923 0.033910935    1.5116526           0.05739639
## aac Asn   aac 32806 0.021425614    1.0696620           1.93692442
## aag Lys   aag 16774 0.010955107    0.4883474           1.94260361
## aat Asn   aat 28533 0.018634916    0.9303380           0.06307558
## aca Thr   aca 11773 0.007688952    0.5675786           0.08879213
## acc Thr   acc 35034 0.022880722    1.6889960           2.60962222
```





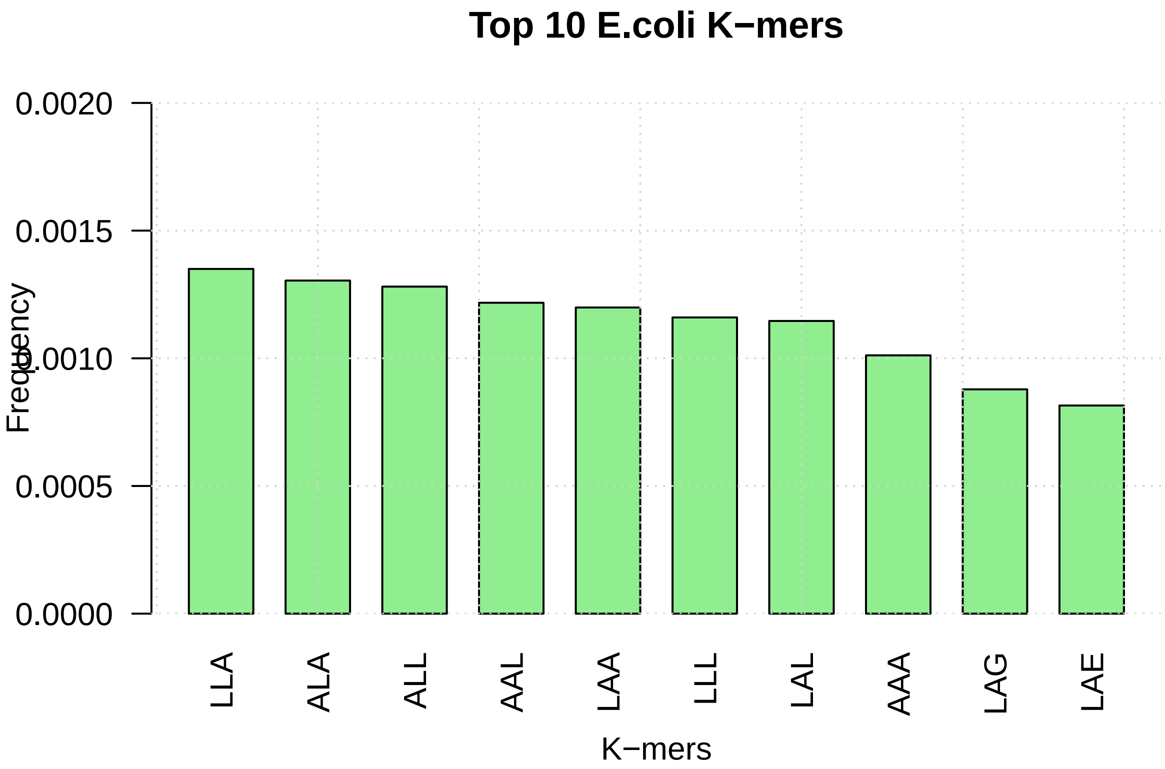
Overlay the two barplots to compare the RSCU between the two organisms.

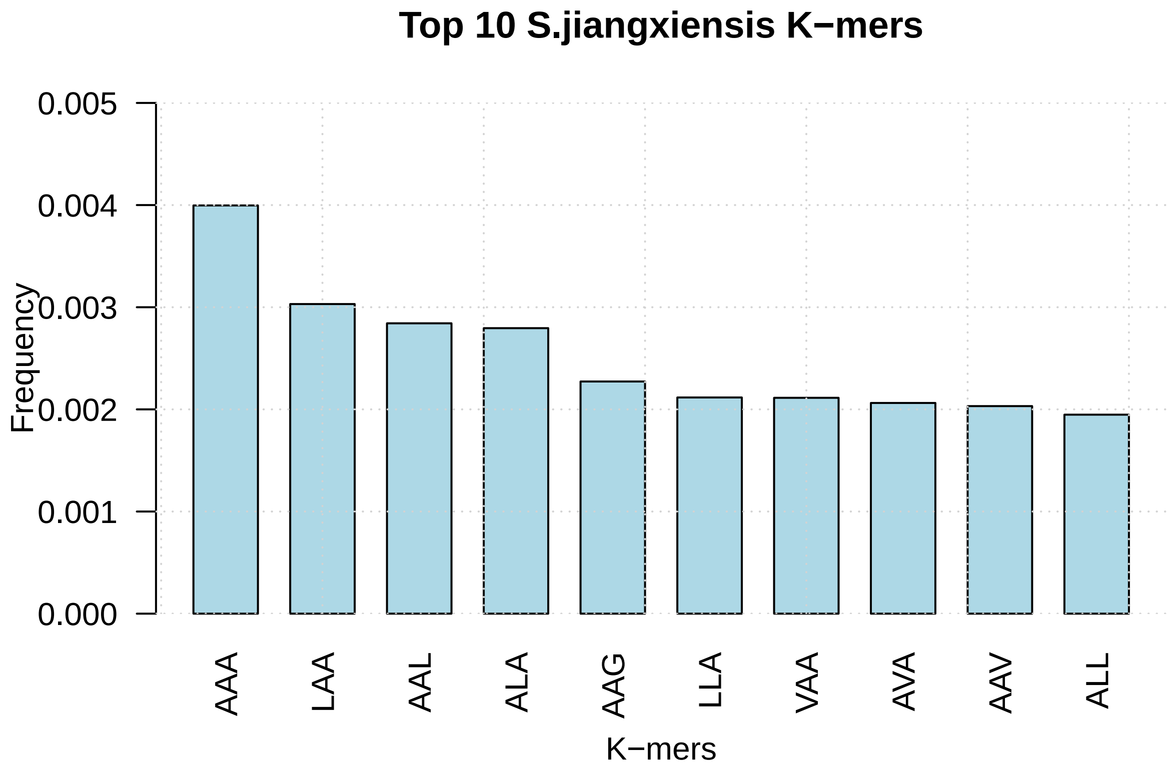


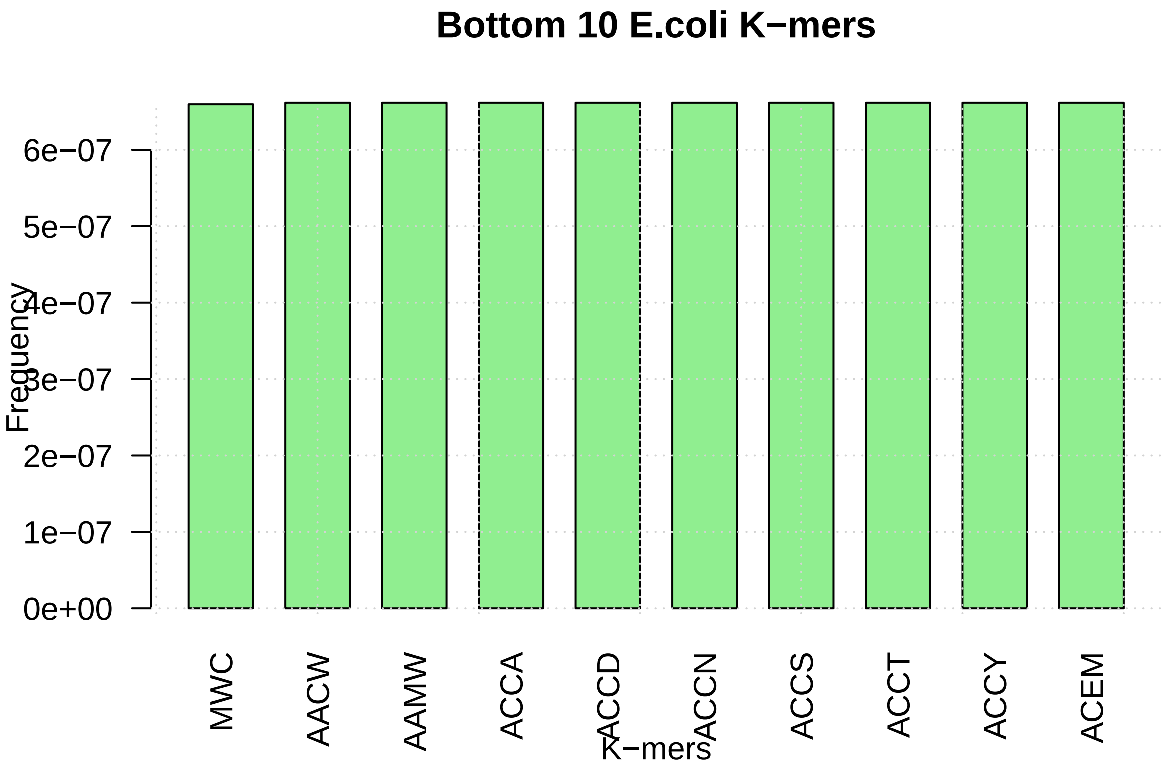
**Answer:** The comparison of codon usage bias between *Escherichia coli* and *Streptacidiphilus jiangxiensis* reveals intriguing insights into their evolutionary adaptations and functional capabilities. In *E. coli*, the Relative Synonymous Codon Usage (RSCU) is notably high for codons such as CTG, CGC, CGT, CCG, and TAA, indicating a preference for these codons in highly expressed genes (Stoletzki and Eyre-Walker, 2006). This pattern suggests that *E. coli* has evolved to optimize its translational efficiency and speed in its typical environments, which often include nutrient-rich conditions that favor rapid growth and division (Lipinszki et al., 2018). Conversely, *Streptacidiphilus jiangxiensis* exhibits a distinct codon usage bias with elevated RSCU values for CTG, CGC, ATC, ACC, and GGC. The presence of ATC and ACC among the preferred codons indicates a potential adaptation to a different ecological niche, possibly involving specialized metabolic functions or responses to varying environmental pressures. These differences in codon preference may reflect the organisms' evolutionary histories, metabolic versatility, and ecological roles (Tyagi et al., 2023), with *Streptacidiphilus* possibly relying on a broader array of substrates or exhibiting different growth strategies compared to the more straightforward, rapid proliferation of *E. coli.* Understanding these patterns can provide deeper insights into how these bacteria interact with their environments and adapt to the challenges they face (Chan et al., 2018).

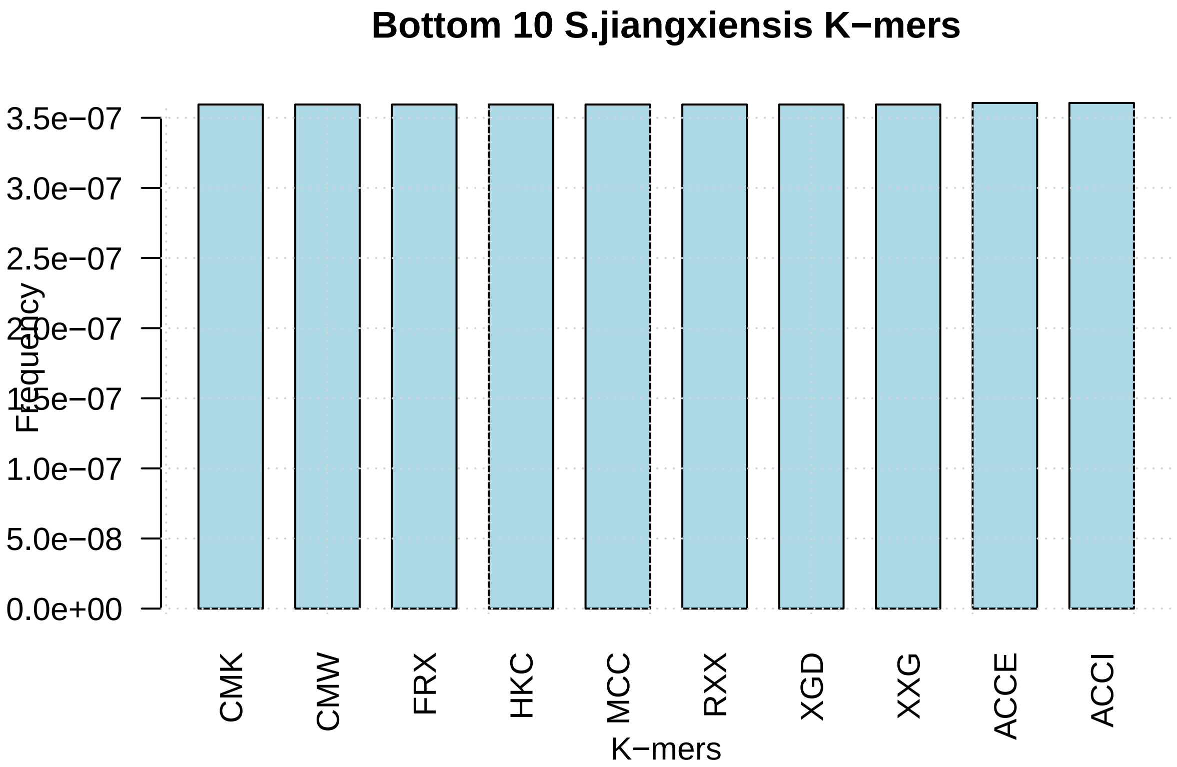
**6) In the organism of interest, identify 10 protein sequence k-mers of length 3-5 which are the most over- and under-represented k-mers in your organism of interest. Are these k-mers also over- and under-represented in E. coli to a similar extent? Provide plots to support your observations. Why do you think these sequences are present at different levels in the genomes of these organisms?** K-mer Analysis

```
##      Kmer   Frequency
## 3781  LLA 0.001349939
## 181   ALA 0.001304435
## 190   ALL 0.001280694
## 10    AAL 0.001217385
## 3601  LAA 0.001198920
## 3790  LLL 0.001160011

##      Kmer    Frequency
## 4362  MWC 6.594718e-07
## 8039 AACW 6.616233e-07
## 8219 AAMW 6.616233e-07
## 8421 ACCA 6.616233e-07
## 8423 ACCD 6.616233e-07
## 8432 ACCN 6.616233e-07

## Number of E. coli K-mers with 0 frequency: 2297663

##      Kmer   Frequency
## 1     AAA 0.003996559
## 3970  LAA 0.003030943
## 10    AAL 0.002842565
## 190   ALA 0.002795111
## 6     AAG 0.002272758
## 4159  LLA 0.002116375

##      Kmer    Frequency
## 660   CMK 3.594998e-07
## 670   CMW 3.594998e-07
## 2078  FRX 3.594998e-07
## 2816  HKC 3.594998e-07
## 4433  MCC 3.594998e-07
## 6593  RXX 3.594998e-07

## Number of S. jiangxiensis K-mers with 0 frequency: 3149491
```


Top 10 E.coli K−mers
Frequency
0.0020
0.0015
0.0010
0.0005
0.0000
LLA
ALA
ALL
AAL
LAA
LLL
LAL
AAA
LAG
LAE
K−mers


Top 10 S.jiangxiensis K−mers
Frequency
0.005
0.004
0.003
0.002
0.001
0.000
AAA
LAA
AAL
ALA
AAG
LLA
VAA
AVA
AAV
ALL
K−mers


Bottom 10 E.coli K−mers
Frequency
6e−07
5e−07
4e−07
3e−07
2e−07
1e−07
0e+00
MWC
AACW
AAMW
ACCA
ACCD
ACCN
ACCS
ACCT
ACCY
ACEM
K−mers



**Answer:** The analysis of k-mer frequency in proteins offers valuable insights into the evolutionary strategies and functional adaptations of *Escherichia coli* and *Streptacidiphilus jiangxiensis*. The top 10 k-mers in *E. coli*, which include sequences like LLA, ALA, and ALL, predominantly feature leucine (L) and alanine (A). This amino acid composition suggests a potential focus on proteins that require flexibility and structural stability, characteristics that are essential for E. coli's rapid growth and metabolic efficiency in nutrient-rich environments, such as the intestines of warm-blooded animals (Yang et al., 2020). The abundance of leucine, known for its role in protein synthesis and cellular metabolism, may enhance the organism's ability to adapt quickly to varying nutrient availability, providing a competitive advantage in diverse ecological niches (Tian et al., 2017).

In contrast, *Streptacidiphilus jiangxiensis* presents a different k-mer profile, with its top sequences such as AAA and AAG indicating a higher representation of alanine and a notable presence of arginine (R). The inclusion of arginine in its protein repertoire may enable *Streptacidiphilus* to engage in a wider range of metabolic pathways, including nitrogen metabolism and stress responses, which are critical for survival in the variable soil environments it inhabits. This greater diversity in k-mers suggests a more complex protein structure that may facilitate specialized functions, such as the synthesis of secondary metabolites that help in nutrient acquisition and competition with other soil microorganisms (Yang et al., 2020).

The contrasting least frequent k-mers further illuminate these differences. For example, *E. coli* contains k-mers like MWC and AACW, which may correspond to rare or less-utilized protein motifs, possibly limiting its capacity to produce certain proteins under specific environmental conditions. Meanwhile, *Streptacidiphilus* features k-mers such as CMK and FRX, which may represent unique adaptations or niche-specific proteins that enhance its survival in less competitive environments. The presence of such rare sequences could confer the ability to utilize specific substrates or produce bioactive compounds that aid in survival and competitive advantage in its soil habitat (Bussi et al., 2021).

Overall, the differences in k-mer frequency between these two organisms reveal not only their distinct evolutionary paths but also their strategies for thriving in their respective environments. E. coli's k-mer composition supports rapid growth and adaptability in rich environments, while Streptacidiphilus's diverse k-mer patterns likely enhance its metabolic versatility and ecological resilience in variable soil ecosystems.

## References

ALTERI, C. J. & MOBLEY, H. L. 2012. Escherichia coli physiology and metabolism dictates adaptation to diverse host microenvironments. Curr Opin Microbiol, 15, 3-9.

BENTLEY, S. 2009. Sequencing the species pan-genome. Nat Rev Microbiol, 7, 258-9.

BUSSI, Y., KAPON, R. & REICH, Z. 2021. Large-scale k-mer-based analysis of the informational properties of genomes, comparative genomics and taxonomy. PLoS One, 16, e0258693.

CHAN, C., PHAM, P., DEDON, P. C. & BEGLEY, T. J. 2018. Lifestyle modifications: coordinating the tRNA epitranscriptome with codon bias to adapt translation during stress responses. Genome Biology, 19, 228.

CHEN, Q., WANG, Y., ZHANG, Z., LIU, X., LI, C. & MA, F. 2021. Arginine Increases Tolerance to Nitrogen Deficiency in Malus hupehensis via Alterations in Photosynthetic Capacity and Amino Acids Metabolism. Front Plant Sci, 12, 772086.

CONG, W., YU, J., FENG, K., DENG, Y. & ZHANG, Y. 2021. The Coexistence Relationship Between Plants and Soil Bacteria Based on Interdomain Ecological Network Analysis. Front Microbiol, 12, 745582.

HARMAN, G. E. & UPHOFF, N. 2019. Symbiotic Root-Endophytic Soil Microbes Improve Crop Productivity and Provide Environmental Benefits. Scientifica (Cairo), 2019, 9106395.

HU, E. Z., LAN, X. R., LIU, Z. L., GAO, J. & NIU, D. K. 2022. A positive correlation between GC content and growth temperature in prokaryotes. BMC Genomics, 23, 110.

HUANG, Y., CUI, Q., WANG, L., RODRIGUEZ, C., QUINTANA, E., GOODFELLOW, M. & LIU, Z. 2004. Streptacidiphilus jiangxiensis sp. nov., a novel actinomycete isolated from acidic rhizosphere soil in China. Antonie van Leeuwenhoek, 86, 159-165.

JAROCKI, V. M., REID, C. J., CHAPMAN, T. A. & DJORDJEVIC, S. P. 2019. Escherichia coli ST302: Genomic Analysis of Virulence Potential and Antimicrobial Resistance Mediated by Mobile Genetic Elements. Front Microbiol, 10, 3098.

KOONIN, E. V. 2009. Evolution of genome architecture. Int J Biochem Cell Biol, 41, 298-306.

LIPINSZKI, Z., VERNYIK, V., FARAGO, N., SARI, T., PUSKAS, L. G., BLATTNER, F. R., POSFAI, G. & GYORFY, Z. 2018. Enhancing the Translational Capacity of E. coli by Resolving the Codon Bias. ACS Synthetic Biology, 7, 2656-2664.

LIZANA, L. & SCHWARTZ, Y. B. 2024. The scales, mechanisms, and dynamics of the genome architecture. Sci Adv, 10, eadm8167.

LOZICA, L., VILLUMSEN, K. R., LI, G., HU, X., MALJKOVIć, M. M. & GOTTSTEIN, Ž. 2022. Genomic Analysis of Escherichia coli Longitudinally Isolated from Broiler Breeder Flocks after the Application of an Autogenous Vaccine. Microorganisms, 10.

MAHARJAN, R. & FERENCI, T. 2014. Mutational Signatures Indicative of Environmental Stress in Bacteria. Molecular Biology and Evolution, 32, 380-391.

MALIK, A., KIM, Y. R. & KIM, S. B. 2020. Genome Mining of the Genus Streptacidiphilus for Biosynthetic and Biodegradation Potential. Genes (Basel), 11.

MASER, A., PEEBO, K., VILU, R. & NAHKU, R. 2020. Amino acids are key substrates to Escherichia coli BW25113 for achieving high specific growth rate. Res Microbiol, 171, 185-193.

MONTEIRO, L. D. F. R., GIRALDI, L. A. & WINCK, F. V. 2023. From Feasting to Fasting: The Arginine Pathway as a Metabolic Switch in Nitrogen-Deprived Chlamydomonas reinhardtii. Cells, 12, 1379.

RASKO, D. A., ROSOVITZ, M. J., MYERS, G. S. A., MONGODIN, E. F., FRICKE, W. F., GAJER, P., CRABTREE, J., SEBAIHIA, M., THOMSON, N. R., CHAUDHURI, R., HENDERSON, I. R., SPERANDIO, V. & RAVEL, J. 2008. The Pangenome Structure of Escherichia coli: Comparative Genomic Analysis of E. coli Commensal and Pathogenic Isolates. Journal of Bacteriology, 190, 6881-6893.

ŠMARDA, P., BUREš, P., HOROVá, L., LEITCH, I. J., MUCINA, L., PACINI, E., TICHý, L., GRULICH, V. & ROTREKLOVá, O. 2014. Ecological and evolutionary significance of genomic GC content diversity in monocots. Proc Natl Acad Sci U S A, 111, E4096-102.

STOLETZKI, N. & EYRE-WALKER, A. 2006. Synonymous Codon Usage in Escherichia coli: Selection for Translational Accuracy. Molecular Biology and Evolution, 24, 374-381.

TIAN, J., YAN, Y., YUE, Q., LIU, X., CHU, X., WU, N. & FAN, Y. 2017. Predicting synonymous codon usage and optimizing the heterologous gene for expression in E. coli. Scientific Reports, 7, 9926.

TYAGI, S., KABADE, P. G., GNANAPRAGASAM, N., SINGH, U. M., GURJAR, A. K. S., RAI, A., SINHA, P., KUMAR, A. & SINGH, V. K. 2023. Codon Usage Provide Insights into the Adaptation of Rice Genes under Stress Condition. Int J Mol Sci, 24.

WAN, T., GONG, Y., LIU, Z., ZHOU, Y., DAI, C. & WANG, Q. 2022. Evolution of complex genome architecture in gymnosperms. GigaScience, 11.

WRIGHT, F. 1990. The ‘effective number of codons’ used in a gene. Gene, 87, 23-9. YANG, Z., LI, H., JIA, Y., ZHENG, Y., MENG, H., BAO, T., LI, X. & LUO, L. 2020. Intrinsic laws of k-mer spectra of genome sequences and evolution mechanism of genomes. BMC Evol Biol, 20, 157.

```
## R version 4.1.2 (2021-11-01)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 22.04.4 LTS
## 
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.10.0
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.10.0
## 
## locale:
##  [1] LC_CTYPE=C.UTF-8       LC_NUMERIC=C           LC_TIME=C.UTF-8       
##  [4] LC_COLLATE=C.UTF-8     LC_MONETARY=C.UTF-8    LC_MESSAGES=C.UTF-8   
##  [7] LC_PAPER=C.UTF-8       LC_NAME=C              LC_ADDRESS=C          
## [10] LC_TELEPHONE=C         LC_MEASUREMENT=C.UTF-8 LC_IDENTIFICATION=C   
## 
## attached base packages:
## [1] stats4    stats     graphics  grDevices utils     datasets  methods  
## [8] base     
## 
## other attached packages:
##  [1] tidyr_1.3.1         readr_2.1.5         ggplot2_3.5.1      
##  [4] dplyr_1.1.4         seqinr_4.2-36       Biostrings_2.62.0  
##  [7] GenomeInfoDb_1.30.1 XVector_0.34.0      IRanges_2.28.0     
## [10] S4Vectors_0.32.4    BiocGenerics_0.40.0 R.utils_2.12.3     
## [13] R.oo_1.26.0         R.methodsS3_1.8.2  
## 
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.13            highr_0.11             compiler_4.1.2        
##  [4] pillar_1.9.0           bitops_1.0-8           tools_4.1.2           
##  [7] zlibbioc_1.40.0        digest_0.6.37          gtable_0.3.5          
## [10] evaluate_1.0.0         lifecycle_1.0.4        tibble_3.2.1          
## [13] pkgconfig_2.0.3        rlang_1.1.4            cli_3.6.3             
## [16] yaml_2.3.10            xfun_0.47              fastmap_1.2.0
## [19] GenomeInfoDbData_1.2.7 withr_3.0.1            knitr_1.48
## [22] hms_1.1.3              generics_0.1.3         vctrs_0.6.5
## [25] grid_4.1.2             tidyselect_1.2.1       ade4_1.7-22
## [28] glue_1.7.0             R6_2.5.1               fansi_1.0.6
## [31] rmarkdown_2.28         farver_2.1.2           purrr_1.0.2
## [34] tzdb_0.4.0             magrittr_2.0.3         scales_1.3.0
## [37] htmltools_0.5.8.1      MASS_7.3-55            colorspace_2.1-1
## [40] labeling_0.4.3         utf8_1.2.4             munsell_0.5.1
## [43] RCurl_1.98-1.16        crayon_1.5.3
```