FInal Report

2024-10-13

SLE777_Group11_Assessment4

PART 1

The file "gene_expression.tsv" contains RNA-seq count data for three samples of interest.

1. Read in the file, making the gene identifiers the row names. Show a table of values for the first six genes.

Downloaded the file through the provided URL to the directory and then read the dataframe using "read.table" command. Read the data table using "Head" command.

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

Added a new column using "gene_data$Mean_Expression <- rowMeans(gene_data)" and showed the first six genes using the head command.

3. List the 10 genes with the highest mean expression

Here listing the 10 genes with the highest mean expression means identifying the top 10 genes (typically represented as rows in a data frame) based on their average expression values across multiple samples or conditions (typically represented as columns). This is done by calculating the means, sorting the means and selecting. The results were, ENSG00000198804.2_MT-CO1 ENSG00000198886.2_MT-ND4 ENSG00000198938.2_MT-CO3 ENSG00000198888.2_MT-ND1 ENSG00000198899.2_MT-ATP6 ENSG00000198727.2_MT-CYB ENSG00000198763.3_MT-ND2 ENSG00000211445.11_GPX3 ENSG00000198712.1_MT-CO2 ENSG00000156508.17_EEF1A1

4. Determine the number of genes with a mean <10

Using the sorted data, here the genes with means which are less than 10 selected.

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

in this step, first the png package was installed and then using that the png image of histogram was inserted.

```r
install.packages("png")
```

```
## Installing package into '/home/s224168163/R/x86_64-pc-linux-gnu-library/4.1'
## (as 'lib' is unspecified)
```

```r
library(png)
```

#get the location of the image

```r
getwd()
```

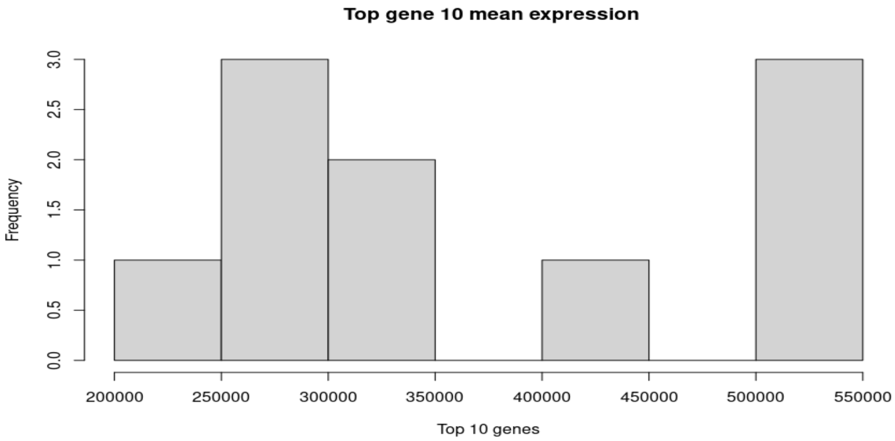
```
## [1] "/home/s224168163/assessment 4/test/SLE777_Group11_Assessment4"
```

#Read the image file for the histogram

```r
img <- readPNG("/home/s224168163/assessment 4/test/SLE777_Group11_Assessment4/Rplot.png")
```

#Plot the image

```r
plot(1:2, type="n", xlab="", ylab="", axes=FALSE)  # Create an empty plot
rasterImage(img, 1, 1, 2, 2)  # Display the image within the plot
```



6. Import this csv file into an R object. What are the column names?

The column names were, Site, TreeID, Circumf_2005_cm, Circumf_2010_cm, Circumf_2015_cm, Circumf_2020_cm

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

For this initially the data for start 2005 and and end 2020 of each site needed to be sub setted. Using the colmns for 2005 and 2020 in each site mean of the column was derived and from that mean the standard deviation was derived. According to that the results were, For southwest, means and standard deviation for 2005 is 4.862 cm, 1.147471 respectively and for 2020, 45.596 cm, 17.87345 respectively. For northeast means and standard deviation for 2005 is 5.292 cm, 0.9140267 respectively and for 2020, 54.228 cm, 25.22795 respectively.

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

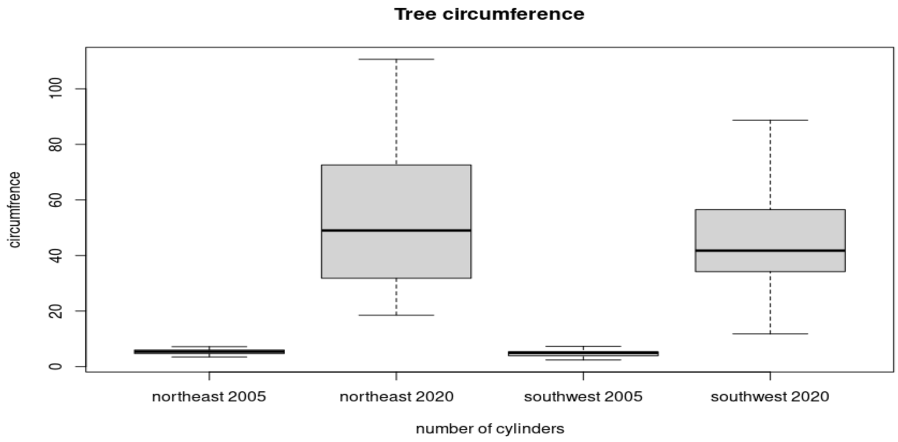
Using the command "BoxPlot" the plot was created and saved as a png image.

Read the image file for the histogram

```r
img <- readPNG("/home/s224168163/assessment 4/test/SLE777_Group11_Assessment4/Rplot01.png")
```

Plot the image

```
plot(1:2, type="n", xlab="", ylab="", axes=FALSE)  # Create an empty plot
rasterImage(img, 1, 1, 2, 2)  # Display the image within the plot
```



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

For this initially the data for last 10 years, starting from 2010 and ending in 2020, of each site needed to be subsetted. Another column was added to the end as "growth difference" which is the difference between the data of 2020 and 2010. Then the column mean was calculated and from the mean of growth difference mean growth over the last 10 years at each site was calculated. The results were, for northeast the mean was 42.940 cm and for southwest the mean was 35.490 cm.

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

Here "t.test" fuction was used and from that the p value was obtained as 0.06229 which is greater than 0.05 showing that the data are significantly different from each other and from that it can conclude that the there is a significant growth difference in last ten years in two sites.

PART 2

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.

E. coli has 4,239 coding sequences, while Campylobacter coli has 1,976. This shows that E. coli has over double the coding sequences, suggesting a more complex genome with potentially more diverse functions.

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.

E. coli has 3,978,528 base pairs of coding DNA, more than double that of Campylobacter coli at 1,726,818. It also surpasses Campylobacter coli in all nitrogen bases: adenine (955,768 vs. 623,683), cytosine (977,594 vs. 229,717), guanine (1,088,501 vs. 318,984), and thymine (956,665 vs. 554,434). This highlights E. coli's greater genetic complexity.

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.

E. coli has a mean length of 938.55, while Campylobacter coli has a mean length of 873.90, indicating that E. coli has a higher mean length. The median length for E. coli is 831, compared to 750 for Campylobacter coli, further demonstrating that E. coli generally has longer sequences than Campylobacter coli in both mean and median lengths.

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.

In Campylobacter, adenine (A) is the most abundant nucleotide in coding sequences (CDS), significantly surpassing cytosine (C), guanine (G), and thymine (T). In contrast, E. coli has higher frequencies of all amino acids, especially glycine (G), with the others nearly equal. This indicates a robust amino acid profile in E. coli, suggesting a metabolic advantage. Overall, E. coli benefits from its diverse amino acid availability, while Campylobacter relies heavily on adenine in its genetic structure. Campylobacter exhibits higher frequencies of amino acids compared to E. coli, particularly for lysine (K) and aspartic acid (D), indicating a distinct amino acid profile. While both have similar levels of alanine (A), E. coli shows lower overall counts. Additionally, less common amino acids like tryptophan (W) and tyrosine (Y) are slightly more abundant in Campylobacter. These differences suggest a more diverse amino acid composition in Campylobacter, potentially affecting its metabolic pathways and functions.

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.

Campylobacter shows lower RSCU values, with many codons used less frequently than expected. In contrast, E. coli has a broader spread of RSCU values, with codons often near 1.0, indicating expected usage. This 1highlights a strong codon bias in Campylobacter, which prefers only a few codons, while E. coli uses codons more evenly, with many falling in the 1.0 to 2.0 range. Codons with RSCU values above 1 are used more often in E. coli than in Campylobacter.

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?

Campylobacter has lower frequency limits, with some codons showing values as low as 0.00000 for three amino acids compared to E. coli. In contrast, E. coli does not exhibit a minimum codon frequency of 1.000. This suggests that the under-representation of certain codons in E. coli differs significantly from that in Campylobacter. In E. coli, A and L are predominantly used, while C and W are used less frequently. Meanwhile, in Campylobacter, G is used more often, along with A, L, and K. Based on the plot generated, the overall amino acid usage is not significantly different between the two species. Despite both species showing relatively higher usage of A and L compared to other amino acids, the differences in codon usage frequency and amino acid usage patterns likely explain why these sequences are present at different levels in the genomes