KI-RIKEN course
Medical Genomics
Autumn 2024

Assignment 1

Task 1

1. Read the research article of the hands-on working group you are assigned (see file "Student Groups.pdf" in the shared folder).

2. Answer the following questions
    a. What is the medically relevant insight from the article?

In this article, the population differences in the context of the mobile genetic elements were described. By comparing databases of different regions (Japan and East Asia, South Asia, Africa, America, Europe), they have discovered that the Alu elements in the genome are differentially distributed. Japanese population shows a much lower proportion of AluY compared to other ethnicities (Fig 1g); furthermore, the American and European populations have consistently lower counts of LINE-1 and SVA. The differences in the mobile elements in the genomes on the population scale suggest that they are not essential to the functioning of cell but can be used to track evolutionary changes.

The second biological insight of the paper comes from merging the information about the mobile element variants (MEV) and the expression quantitative trait loci (eQTL, representing the mRNA expression). The authors found that MEVs can influence gene regulation

    b. Which genomics technology/ technologies were used?

This paper showcased a newly developed tool, MEGAnE. By finding within the genome non-unique 32-mers, it has improved the recognition of the insertion and absence of the mobile elements in highly repetitive or otherwise complex genomic regions. It reduces the computational load of the searches by removing multi-mapping reads. Consequently, the genomes can be aptly searched for the presence/absence and location of the mobile elements which is a great improvement in the field.

Additionally, the authors correlate the mobile elements variants with the gene expression (expression quantitative trait loci) to link the genome composition with the gene functionality.

3. Further related research questions
    a. List and explain at least three questions/ hypotheses you can think of that extend the analysis presented in the paper.

1) Genome organisation context for mobile element changes: Where in the nucleus are those alterations more prevalent and most different between populations? By integrating with existing data on the human nucleus (e.g., Hi-C), it could be resolved whether those events happen in euchromatin or across the nucleus. (Expansion of the Fig 2A).
2) Functional context for the integration of the MEVs and eQTLs: Which genes are the most affected between populations? The paper briefly mentions various diseases associated with the Japanese population, but can it be done more systematically?

3) Cell cycle context: In Japanese population, the Alu insertions are correlated with early replication. Merging with 1) and cell-biology based studies (CRISPR with insertions, S-phase disturbance), one could investigate whether the mobile elements are making the nucleus more or less resistant to replication stress.

b. [Optional] Devise a computational analysis strategy for (some of) the listed questions under 3a.

For 2), a database of diseases affecting particular populations should be curated. The following correlations between the local regions and MEs should be comparably straightforward.

For 3), the correlation between the RepliSeq data and MEVs presence between populations could be performed to target the most sensitive and obvious sites. By genome modifications (either expanding such regions to ridiculous sizes or deleting them), one could see the effect of further genome changes on the replication timings.

Task 4

Using R example datasets

1. Use the R internal CO2 dataset ("data(CO2)").
2. Describe briefly the content of the CO2 dataset using the help function.

CO2 dataset is a list of 84 measurements from an experiment on cold tolerance of plant and its effect on the CO2 metabolism. It has information on the plant, origin of the plant, treatment (chilled vs non-chilled), concentration of carbon dioxide measured, and the uptake rate of CO2. The data comes from a paper: Potvin, C., Lechowicz, M. J. and Tardif, S. (1990) "The statistical analysis of ecophysiological response curves obtained from experiments involving repeated measures", *Ecology*, 71, 1389–1400.

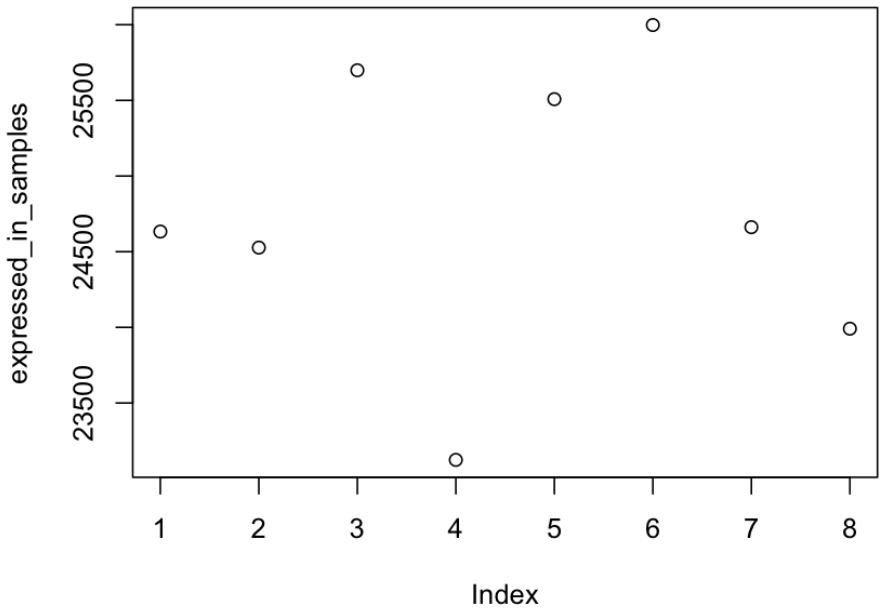
3. What is the average and median CO2 uptake of the plants from Quebec and Mississippi?

Mean = 27.2131
Median = 28.3

4. [Optional] In the "airway" example data from Bioconductor, how many genes are expressed in each sample? How many genes are not expressed in any sample?

Genes not expressed at all: 30208 (count the 0 value rows)

How many genes are expressed in each sample:



Task 5

R Functions

1. Write a function that calculates the ratio of the mean and the median of a given vector. This is a helpful measure to detect data with outlying values. Note: See Reference for R language

mean(df)/median(df)

where df is a vector

2. Write a function that ignores the lowest and the highest value from a given vector and calculate the mean.

```
df_nominmax <- df[df != max(df) & df != min(df)]
mean(df_nominmax)
```

3. Read about piping from here:https://r4ds.had.co.nz/pipes.html#pipes (you don't have to learn everything, a basic understanding of the usage is enough). Write a short (max. 300 characters, no spaces) explanation of why, how, and when not to use pipes.

Pipes are a streamlining method to manage multiple operations. In a pipe, LHS argument of an operation is fed into the RHS operation without the need to name each sub-argument separately, keeping the code clean and easy to read. Pipes are unsuitable when interim steps need to be inspected, or when multiple steps or arguments are needed.

4. Familiarize yourself with the apply-family of functions (apply, lapply, sapply etc.) http://uc-r.github.io/apply_family Write a short explanation (max. 300 characters, no spaces) of why they could be useful in your work.

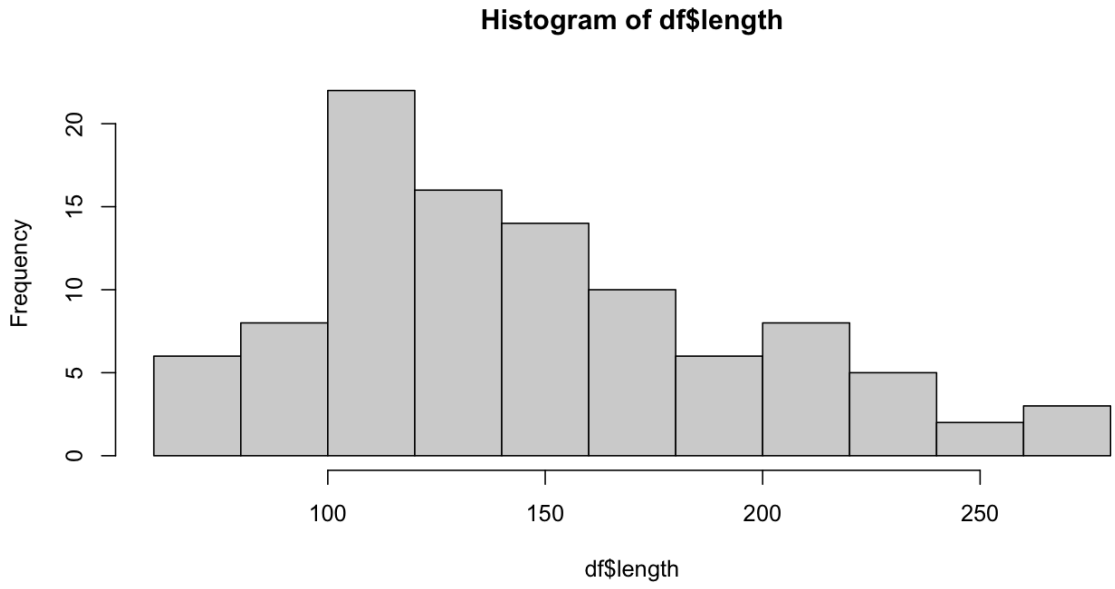
Apply family of functions applies a function over a specified argument (data frame, vector, list) without a need for a loop. It is more concise and easy to write with each building block defined separately.

Task 6

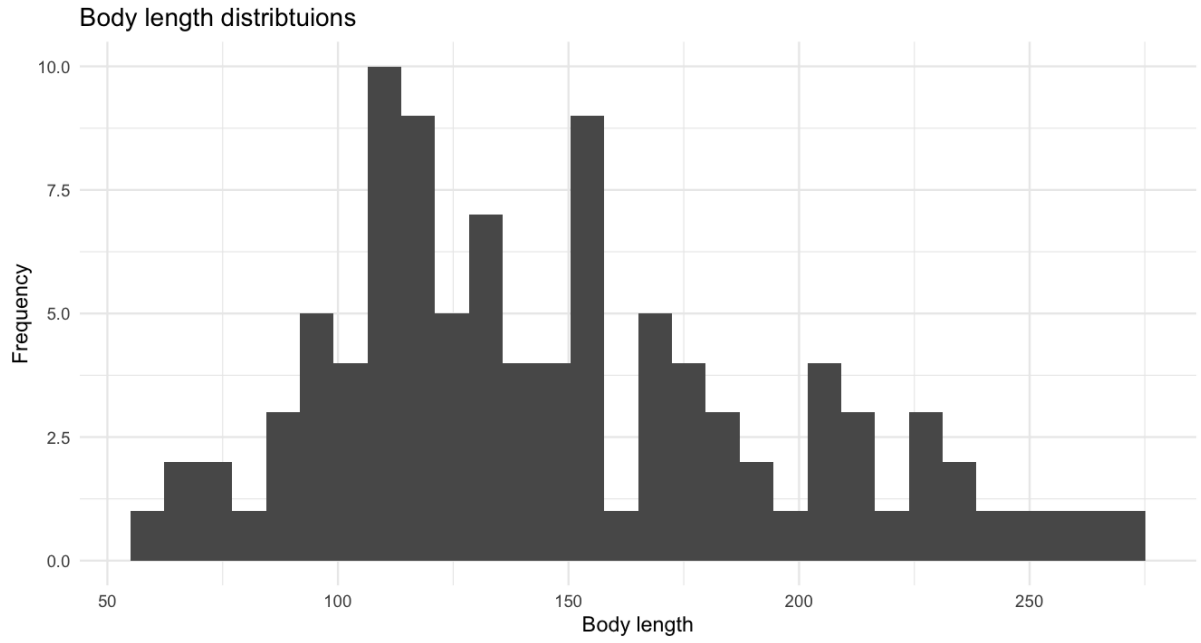
Basic visualization with R

1. Compare the distributions of the body heights of the two species from the 'magic_guys.csv' dataset graphically
    a. using the basic 'hist' function as well as ‘ggplot’ and ‘geom_histogram’ functions from the ggplot2 package. Optimize the plots for example by trying several different 'breaks'. Note that ggplot2-based functions give you many more options for changing the visualization parameters, try some of them.

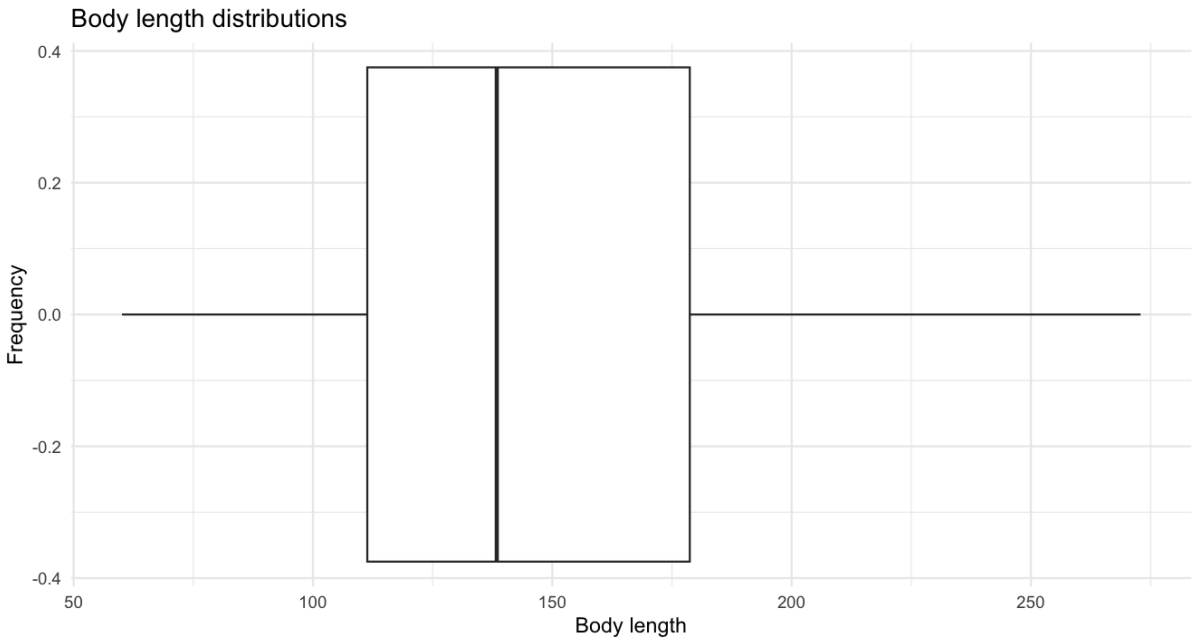
Standard ‘hist’ histogram



ggplot2



b. Do the same comparison as in a. but with boxplots. If you want to use the ggplot2-package, use the functions ‘ggplot’ and ‘geom_boxplot’.



c. Save the plots with the ‘png’, ‘pdf’, and ‘svg’ formats. In which situation would you use which file format?

Png
- No loss of quality (raster/pixel-based)
- Transparency
- Fixed resolution (it is not possible to scale easily)
- Large size

Svg
- Scalable vector graphics – unlike png, it is possible to scale high-resolution images
- Can have a large size for complex plots

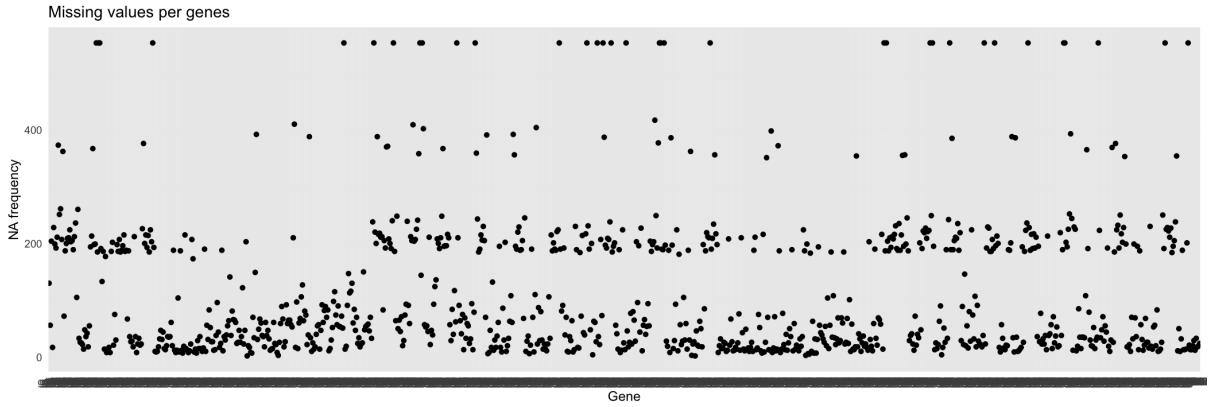
Pdf
- Similarly to svg, vector based: high quality and detail, scalable
- Good for printing (vector based images don’t pixelate) and multi-page works
- Universally supported format

2. Load the gene expression data matrix from the ‘microarray_data.tab’ dataset provided in the shared folder, it is a big tabular separated matrix.
   a. How big is the matrix in terms of rows and columns?

533 rows and 1000 columns

b. Count the missing values per gene and visualize this result.

Assuming columns are genes (g in the name) and rows are observations, we need to count NA in each column and plot it as a bar chart or a scatter plot or similar:



c. Find the genes for which there are more than X% (X=10%, 20%, 50%) missing values.

Note: I do not understand why some values are negative

Method

- Find the percentage of NA values for each column, store it as a vector
- Compare the value to a threshold X%, store larger values as a new vector
- Print out new vectors

10%

```
 [1] "g1"   "g2"   "g3"   "g4"   "g5"   "g10"  "g11"  "g12"  "g14"  "g15"  "g16"  "g18"
[13] "g21"  "g22"  "g23"  "g24"  "g25"  "g26"  "g28"  "g29"  "g35"  "g36"  "g37"  "g38"
[25] "g39"  "g40"  "g41"  "g42"  "g44"  "g45"  "g46"  "g47"  "g48"  "g49"  "g50"  "g51"
[37] "g52"  "g53"  "g54"  "g55"  "g56"  "g57"  "g58"  "g59"  "g60"  "g61"  "g62"  "g63"
[49] "g64"  "g65"  "g66"  "g67"  "g68"  "g69"  "g70"  "g71"  "g72"  "g73"  "g74"  "g75"
[61] "g76"  "g77"  "g78"  "g79"  "g80"  "g81"  "g82"  "g83"  "g84"  "g85"  "g86"  "g87"
[73] "g88"  "g89"  "g90"  "g91"  "g92"  "g93"  "g94"  "g95"  "g96"  "g97"  "g98"  "g99"
[85] "g100" "g101" "g102" "g103" "g104" "g105" "g106" "g107" "g108" "g109" "g110" "g111"
[97] "g112" "g113" "g114" "g115" "g116" "g117" "g118" "g119" "g120" "g130" "g131" "g132"
[109] "g133" "g134" "g135" "g136" "g137" "g138" "g139" "g140" "g142" "g147" "g148" "g151"
[121] "g152" "g153" "g154" "g155" "g156" "g157" "g158" "g159" "g160" "g165" "g171" "g172"
[133] "g173" "g174" "g175" "g176" "g177" "g178" "g179" "g180" "g194" "g196" "g200" "g204"
[145] "g210" "g221" "g223" "g233" "g239" "g241" "g242" "g243" "g244" "g252" "g258" "g260"
[157] "g263" "g264" "g268" "g274" "g280" "g281" "g284" "g285" "g286" "g287" "g288" "g290"
[169] "g292" "g294" "g295" "g296" "g297" "g298" "g299" "g301" "g309" "g311" "g312" "g313"
[181] "g314" "g315" "g316" "g320" "g321" "g322" "g323" "g324" "g327" "g329" "g331" "g332"
[193] "g333" "g334" "g335" "g336" "g337" "g339" "g344" "g347" "g348" "g351" "g352" "g353"
[205] "g354" "g355" "g356" "g357" "g358" "g359" "g360" "g361" "g362" "g363" "g364" "g365"
[217] "g366" "g367" "g368" "g369" "g370" "g372" "g374" "g377" "g378" "g379" "g380" "g381"
[229] "g382" "g383" "g384" "g385" "g386" "g387" "g388" "g389" "g390" "g391" "g392" "g396"
[241] "g400" "g401" "g402" "g403" "g404" "g405" "g406" "g407" "g408" "g409" "g410" "g411"
[253] "g413" "g415" "g416" "g417" "g418" "g419" "g421" "g423" "g425" "g429" "g430" "g431"
[265] "g432" "g433" "g434" "g435" "g436" "g437" "g438" "g439" "g440" "g445" "g450" "g453"
[277] "g455" "g459" "g460" "g461" "g462" "g463" "g464" "g465" "g466" "g467" "g468" "g469"
[289] "g470" "g476" "g477" "g479" "g481" "g483" "g491" "g492" "g493" "g494" "g495" "g496"
[301] "g497" "g498" "g499" "g500" "g502" "g507" "g510" "g513" "g514" "g515" "g518" "g519"
[313] "g520" "g522" "g524" "g525" "g527" "g530" "g531" "g532" "g533" "g534" "g535" "g536"
[325] "g537" "g538" "g539" "g540" "g541" "g544" "g547" "g553" "g555" "g556" "g558" "g559"
```

[337] "g561" "g563" "g564" "g567" "g569" "g570" "g571" "g572" "g573" "g574" "g575" "g576"
[349] "g577" "g578" "g579" "g580" "g583" "g585" "g586" "g589" "g591" "g592" "g595" "g597"
[361] "g599" "g607" "g610" "g611" "g612" "g613" "g614" "g615" "g616" "g617" "g618" "g619"
[373] "g620" "g631" "g638" "g650" "g653" "g657" "g660" "g663" "g666" "g669" "g672" "g681"
[385] "g689" "g691" "g694" "g696" "g700" "g707" "g709" "g711" "g715" "g718" "g719" "g722"
[397] "g724" "g726" "g743" "g744" "g747" "g748" "g749" "g751" "g752" "g753" "g754" "g755"
[409] "g756" "g757" "g758" "g759" "g760" "g761" "g762" "g763" "g764" "g765" "g766" "g767"
[421] "g768" "g769" "g770" "g776" "g781" "g782" "g783" "g784" "g785" "g786" "g787" "g788"
[433] "g789" "g790" "g795" "g796" "g801" "g802" "g803" "g804" "g805" "g806" "g807" "g808"
[445] "g809" "g810" "g812" "g814" "g818" "g820" "g822" "g824" "g831" "g832" "g833" "g834"
[457] "g835" "g836" "g837" "g838" "g839" "g840" "g843" "g849" "g850" "g851" "g854" "g861"
[469] "g862" "g863" "g864" "g865" "g866" "g867" "g868" "g869" "g870" "g872" "g874" "g882"
[481] "g884" "g891" "g892" "g893" "g894" "g895" "g896" "g897" "g898" "g899" "g900" "g903"
[493] "g909" "g910" "g911" "g919" "g922" "g926" "g931" "g932" "g933" "g934" "g935" "g936"
[505] "g937" "g938" "g939" "g940" "g945" "g947" "g948" "g951" "g952" "g956" "g963" "g965"
[517] "g970" "g971" "g972" "g973" "g974" "g975" "g976" "g977" "g978" "g979" "g980" "g985"
[529] "g989"

20%
 [1] "g1"   "g14" "g18" "g21" "g22" "g24" "g25" "g26" "g29" "g39" "g40" "g41"
[13] "g48" "g51" "g52" "g53" "g54" "g55" "g56" "g57" "g58" "g59" "g60" "g61"
[25] "g62" "g63" "g64" "g65" "g66" "g67" "g68" "g69" "g70" "g71" "g72" "g73"
[37] "g74" "g75" "g76" "g77" "g78" "g79" "g80" "g81" "g82" "g83" "g84" "g85"
[49] "g86" "g87" "g88" "g89" "g90" "g91" "g92" "g93" "g94" "g95" "g96" "g97"
[61] "g98" "g99" "g100" "g101" "g102" "g103" "g104" "g105" "g106" "g107" "g108" "g109"
[73] "g110" "g111" "g112" "g113" "g114" "g115" "g116" "g117" "g118" "g119" "g120" "g130"
[85] "g131" "g132" "g133" "g134" "g135" "g136" "g137" "g138" "g139" "g140" "g142" "g147"
[97] "g148" "g151" "g152" "g153" "g154" "g155" "g156" "g157" "g158" "g159" "g160" "g165"
[109] "g171" "g172" "g173" "g174" "g175" "g176" "g177" "g178" "g179" "g180" "g196" "g200"
[121] "g204" "g210" "g233" "g252" "g260" "g290" "g297" "g301" "g321" "g329" "g332" "g333"
[133] "g334" "g335" "g344" "g351" "g352" "g353" "g354" "g355" "g356" "g357" "g358" "g359"
[145] "g360" "g361" "g362" "g363" "g364" "g365" "g366" "g367" "g368" "g369" "g370" "g379"
[157] "g381" "g382" "g383" "g384" "g385" "g386" "g387" "g388" "g389" "g390" "g391" "g396"
[169] "g400" "g401" "g402" "g403" "g404" "g405" "g406" "g407" "g408" "g409" "g410" "g415"
[181] "g417" "g418" "g431" "g432" "g433" "g434" "g435" "g436" "g437" "g438" "g439" "g440"
[193] "g445" "g450" "g455" "g461" "g462" "g463" "g464" "g465" "g466" "g467" "g468" "g469"
[205] "g470" "g476" "g491" "g492" "g493" "g494" "g495" "g496" "g497" "g498" "g499" "g500"
[217] "g510" "g513" "g515" "g518" "g519" "g520" "g522" "g527" "g531" "g532" "g533" "g534"
[229] "g535" "g536" "g537" "g538" "g539" "g540" "g544" "g547" "g558" "g561" "g569" "g570"
[241] "g571" "g572" "g573" "g574" "g575" "g576" "g577" "g578" "g579" "g580" "g583" "g585"
[253] "g586" "g591" "g599" "g611" "g612" "g613" "g614" "g615" "g616" "g617" "g618" "g619"
[265] "g620" "g650" "g657" "g663" "g669" "g689" "g691" "g694" "g744" "g751" "g752" "g753"
[277] "g754" "g755" "g756" "g757" "g758" "g759" "g760" "g761" "g762" "g763" "g764" "g765"
[289] "g766" "g767" "g768" "g769" "g770" "g781" "g782" "g783" "g784" "g785" "g786" "g787"
[301] "g788" "g789" "g790" "g801" "g802" "g803" "g804" "g805" "g806" "g807" "g808" "g809"
[313] "g810" "g814" "g831" "g832" "g833" "g834" "g835" "g836" "g837" "g838" "g839" "g840"
[325] "g849" "g850" "g851" "g854" "g861" "g862" "g863" "g864" "g865" "g866" "g867" "g868"
[337] "g869" "g870" "g872" "g891" "g892" "g893" "g894" "g895" "g896" "g897" "g898" "g899"
[349] "g900" "g910" "g919" "g926" "g931" "g932" "g933" "g934" "g935" "g936" "g937" "g938"
[361] "g939" "g940" "g947" "g951" "g970" "g971" "g972" "g973" "g974" "g975" "g976" "g977"
[373] "g978" "g979" "g980" "g985" "g989"

```
50%
[1] "g18"  "g48"  "g55"  "g58"  "g60"  "g66"  "g73"  "g79"  "g83"  "g91"  "g93"  "g94"  "g99"
[14] "g105" "g109" "g132" "g135" "g137" "g138" "g172" "g260" "g290" "g301" "g329" "g352"
"g355"
[27] "g362" "g363" "g368" "g383" "g388" "g389" "g390" "g391" "g406" "g417" "g431" "g432"
"g440"
[40] "g461" "g462" "g498" "g519" "g527" "g531" "g532" "g538" "g572" "g575" "g576" "g577"
"g585"
[53] "g615" "g619" "g663" "g669" "g751" "g753" "g766" "g768" "g788" "g802" "g804" "g838"
"g851"
[66] "g854" "g864" "g892" "g893" "g898" "g919" "g932" "g971" "g980"
```

d. Replace the missing values by the average expression value for the particular gene. (Note: Imputing data has to be used with caution!)

Note: I do not understand why some values are negative.

```
# Replace NA with the mean value for each gene
gene_means <- sapply(df2, function(x) mean(x, na.rm = TRUE))
df2[]<- lapply(seq_along(df2), function(i) {
  x <- df2[[i]]
  x[is.na(x)] <- gene_means[i]
  return(x)
})
```

3. Visualize the data in the CO2 dataset in a way that gives you a *deeper understanding* of the data. What do you see?

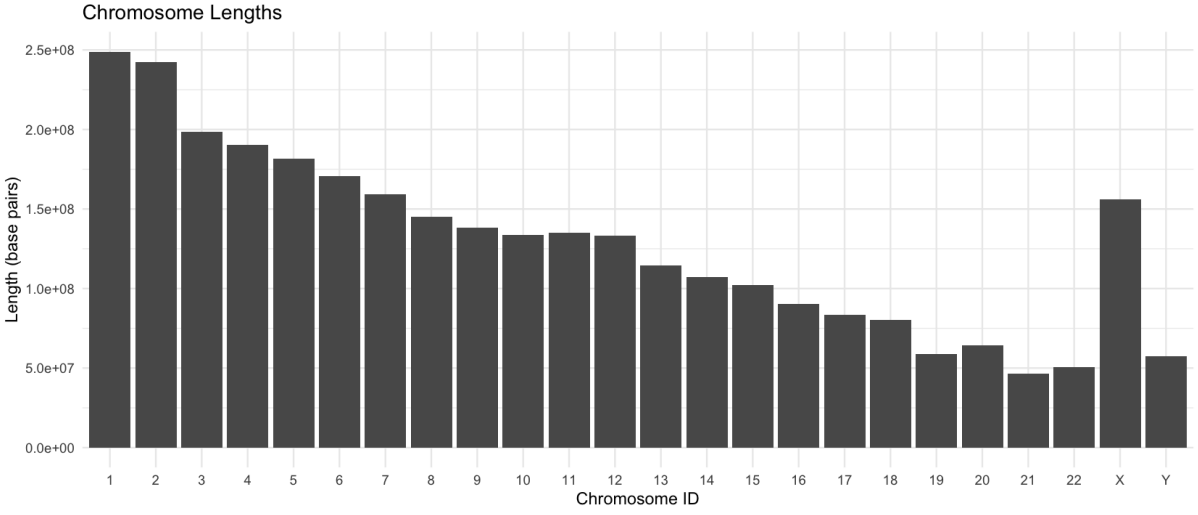
I do not understand what does deeper understanding mean here and why is it a CO2 dataset.

Task 7

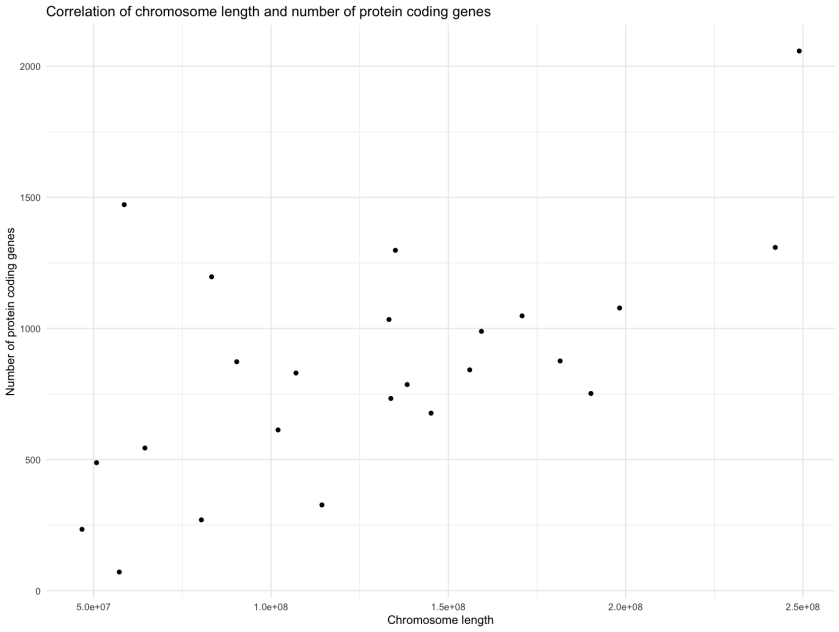
1. Install the Tidybiology package, which includes the data ‘chromosome’ and ‘proteins’
   devtools::install_github("hirscheylab/tidybiology")
   a. Extract summary statistics (mean, median and maximum) for the following variables from the ‘chromosome’ data: variations, protein coding genes, and miRNAs. Utilize the tidyverse functions to make this as simply as possible.

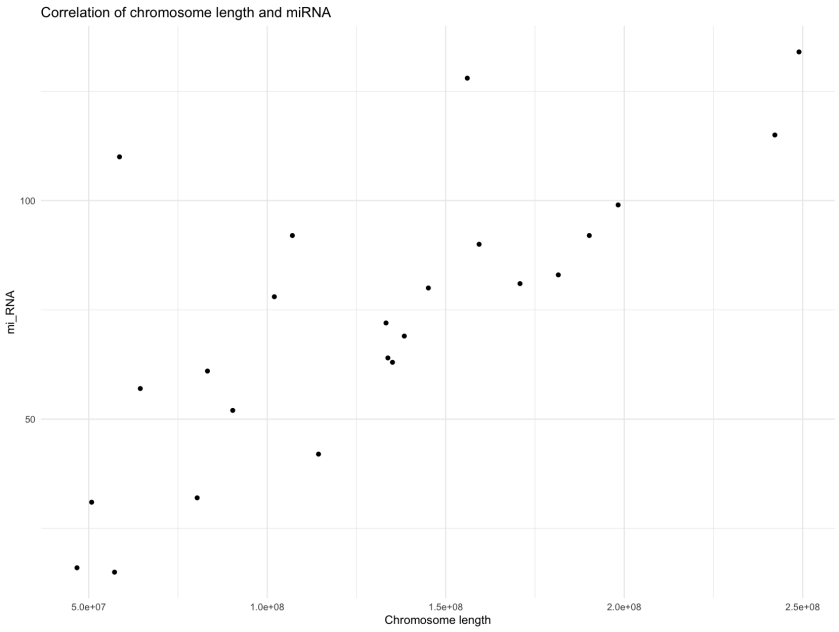
```
>summary(chromosome)
      id       length_mm         basepairs          variations       protein_codinggenes
 1     : 1   Min.   :16.00   Min.   : 46709983   Min.   :  211643   Min.   :  71.0
 2     : 1   1st Qu.:27.75   1st Qu.: 82536402   1st Qu.: 4395298   1st Qu.: 595.8
 3     : 1   Median :45.50   Median :133536366   Median : 6172346   Median : 836.0
 4     : 1   Mean   :43.83   Mean   :128677910   Mean   : 6484572   Mean   : 850.0
 5     : 1   3rd Qu.:55.00   3rd Qu.:162210974   3rd Qu.: 8742592   3rd Qu.:1055.5
 6     : 1   Max.   :85.00   Max.   :248956422   Max.   :12945965   Max.   :2058.0
 (Other):18
  pseudo_genes    totallongnc_rna  totalsmallnc_rna     mi_rna          r_rna
 Min.   : 185.0   Min.   :  71.0   Min.   : 30.0    Min.   : 15.00   Min.   : 5.00
 1st Qu.: 445.8   1st Qu.: 439.0   1st Qu.:167.0    1st Qu.: 55.75   1st Qu.:13.00
 Median : 590.5   Median : 633.5   Median :220.5    Median : 75.00   Median :23.00
 Mean   : 608.3   Mean   : 613.6   Mean   :208.9    Mean   : 73.17   Mean   :22.08
 3rd Qu.: 772.5   3rd Qu.: 751.0   3rd Qu.:236.0    3rd Qu.: 92.00   3rd Qu.:27.25
 Max.   :1220.0   Max.   :1200.0   Max.   :496.0    Max.   :134.00   Max.   :66.00

    sn_rna          sno_rna         miscnc_rna     centromereposition_mbp
 Min.   : 17.00   Min.   :  3.00   Min.   :  8.00   Min.   : 12.50
 1st Qu.: 49.75   1st Qu.: 36.75   1st Qu.: 66.50   1st Qu.: 18.73
 Median : 77.00   Median : 59.50   Median : 94.50   Median : 38.40
 Mean   : 81.00   Mean   : 63.38   Mean   : 92.21   Mean   : 43.36
 3rd Qu.:106.00   3rd Qu.: 76.00   3rd Qu.:107.50   3rd Qu.: 55.25
 Max.   :221.00   Max.   :145.00   Max.   :192.00   Max.   :125.00
```

   b. How does the chromosome size distribute? Plot a graph that helps to visualize this by using ggplot2 package functions.



c. Does the number of protein coding genes or miRNAs correlate with the length of the chromosome? Make two separate plots to visualize these relationships.





d. Calculate the same summary statistics for the 'proteins' data variables length and mass. Create a meaningful visualization of the relationship between these two variables by utilizing the ggplot2 package functions. Play with the colors, theme- and other visualization parameters to create a plot that pleases you.

```
> summary(proteins$length)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    2.0   251.0   414.0   557.2   669.0 34350.0
> summary(proteins$mass)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
260	27940	46140	62061	74755	3816030