Task 7

2023-10-06

#Tidybiology

Extract summary statistics (mean, median, maximum) for the following variable from the chromosom data: variations, protein, coding genes, and mRNA

```
#load the package needed
library(tidyverse)
library(gridExtra)
library(tidybiology)
data("chromosome")
```

```
#Extract the summary statistics of chromosome data for selected columns (variations, protein_codinggene
variable<-chromosome%>%
select(variations, protein_codinggenes, mi_rna)
summary(variable)
```

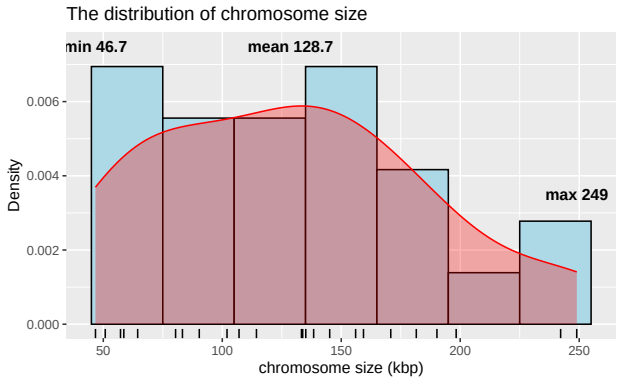
```
##    variations       protein_codinggenes     mi_rna
##  Min.   :  211643   Min.   :  71.0      Min.   : 15.00
##  1st Qu.: 4395298   1st Qu.: 595.8      1st Qu.: 55.75
##  Median : 6172346   Median : 836.0      Median : 75.00
##  Mean   : 6484572   Mean   : 850.0      Mean   : 73.17
##  3rd Qu.: 8742592   3rd Qu.:1055.5      3rd Qu.: 92.00
##  Max.   :12945965   Max.   :2058.0      Max.   :134.00
```

How does the chromosome size distribute?

```
#Add kilobasepair column
new_chromosome <-chromosome%>%
mutate(kbp=basepairs/1000000)

#Annotation to ggplot
annotation <-data.frame(x=c(round(min(new_chromosome$kbp),1), round(max(new_chromosome$kbp),1), round(m
y=c(0.0075, 0.0035, 0.0075),
label=c("min", "max", "mean"))

#generate the distribution graph
ggplot(new_chromosome, aes(kbp))+
geom_histogram(aes(y=..density..), color="black",fill="lightblue", binwidth =30)+
geom_rug()+
geom_density(color="red", fill="red", alpha=0.3)+
geom_text(data=annotation, aes(x=x, y=y, label=paste(label,x)), fontface="bold")+
labs(title = "The distribution of chromosome size", x="chromosome size (kbp)", y="Density")
```



According to this graph, the distribution of chromosome size is right-skewed, where most of the chromosome has a size around 46-200 kbp.

Does the number of protein coding genes or miRNA correlate with the length of the chromosome?

```
#Calculate the correlation
corPL <- cor(chromosome$protein_codinggenes, chromosome$length_mm)
corPL
```

```
## [1] 0.6060185
```

```
corML <- cor(chromosome$mi_rna, chromosome$length_mm)
corML
```

```
## [1] 0.7366973
```

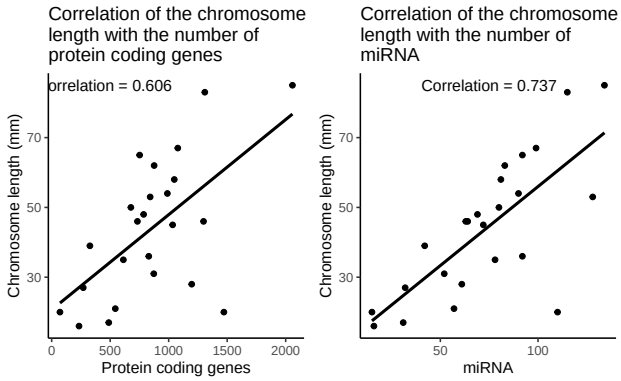
```
#define the title
titlePL <-str_wrap("Correlation of the chromosome length with the number of protein coding genes", widt
titleML <-str_wrap("Correlation of the chromosome length with the number of miRNA", width=30)

#generate the graph
plotProtein <- ggplot(chromosome, aes(x = protein_codinggenes, y = length_mm)) +
geom_point() +
geom_smooth(method = lm, color="black", se=FALSE)+
labs(title = titlePL, x="Protein coding genes", y="Chromosome length (mm)")+
theme_classic()

plotMiRNA <- ggplot(chromosome, aes(x = mi_rna, y = length_mm)) +
geom_point() +
geom_smooth(method = lm, color="black", se=FALSE)+
labs(title = titleML, x="miRNA", y="Chromosome length (mm)")+
theme_classic()

#Give the correlation information on the graph
plotPL <- plotProtein + annotate("text", x=450, y=max(chromosome$length_mm), label="Correlation = 0.606
plotML <-plotMiRNA + annotate("text", x=75, y=max(chromosome$length_mm), label="Correlation = 0.737")

grid.arrange(plotPL, plotML,ncol=2)
```



According to this graph, length of the chromosome shows moderately positive correlation (correlation coefficient > 0.5) with the number of protein coding genes and miRNA.

Calculate the summary statistics for the protein data variables length and mass. And create the correlation plot between two parameters

```
data("proteins")
variable<-proteins%>%
select(length,mass)
summary(variable)
```

```
##      length             mass
##  Min.   :    2.0   Min.   :    260
##  1st Qu.:  251.0   1st Qu.:  27940
##  Median :  414.0   Median :  46140
##  Mean   :  557.2   Mean   :  62061
##  3rd Qu.:  669.0   3rd Qu.:  74755
##  Max.   :34350.0   Max.   :3816030
```

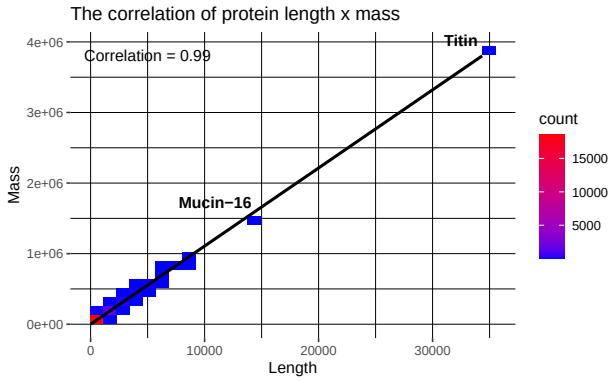
Plot the correlation between proteinlength and protein mass

```
#calculate the correlation
cor(proteins$length, proteins$mass)
```

```
## [1] 0.9991674
```

```
#generate the plot
proteinplot<- ggplot(proteins, aes(x = length, y = mass)) +
stat_bin2d()+
scale_fill_gradient(low="blue", high="red")+
geom_smooth(method = lm, color="black")+
labs(title = "The correlation of protein length x mass", x="Length", y="Mass")+
theme(
panel.grid.major=element_line(color="black", linewidth = 0.3),
panel.grid.minor=element_line(color="black", linewidth = 0.1),
panel.background = element_rect(fill="white"))

#label the selected protein
selected_points <-proteins[proteins$protein_name%in% c("Mucin-16 ","Titin "), ]
proteinplot+geom_text(data=selected_points, aes(label=protein_name), vjust=-0.8, hjust=1, size=4, fontfa
  annotate("text", x=5000, y=max(proteins$mass), label="Correlation = 0.99")
```



According to this graph, the correlation between protein length and mass is strongly positive (0.99). Also, we could see that two proteins, Mucin-16 and Titin, larger in size compared with the rest. We also could log10 all the variables to make each point not overlap. But I personally want to show the 2 large proteins here.