QBS 103 Final Project

Chichi Illoh

2023-07-24

Variables used to generate plots

Gene:LCN2
Continuous variable: Age
Categorical Variables: Sex and Disease Status

The LCN2 gene encodes the protein Lipocalin 2 which plays a role in innate immunity by limiting bacterial growth through the sequestering of iron-containing siderophores. This protein is thought to be be involved in multiple cellular processes: the maintenance of skin homeostasis and suppression of invasiveness and metastasis.

In relation to COVID-19, researchers found that LCN2 is overexpressed in COVID-19 patients. They also found that this gene is linked to neutrophil and virus response activities,so higher expression of this gene leads to inflammatory responses and cilium movement.

Sources:

https://www.genecards.org/cgi-bin/carddisp.pl?gene=LCN2

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7714049/

```
setwd("C:/Users/Student/Desktop/QBS103/")# set working directory
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
meta_data = read.csv("QBS103_finalProject_metadata.csv")
#head(meta_data)
gene_data = read.csv("QBS103_finalProject_geneExpression.csv")
#head(gene_data)

# convert into vertical data and extract specific gene: LCN2
gene_expression = matrix(gene_data[5,2:127])
head(gene_expression)
```

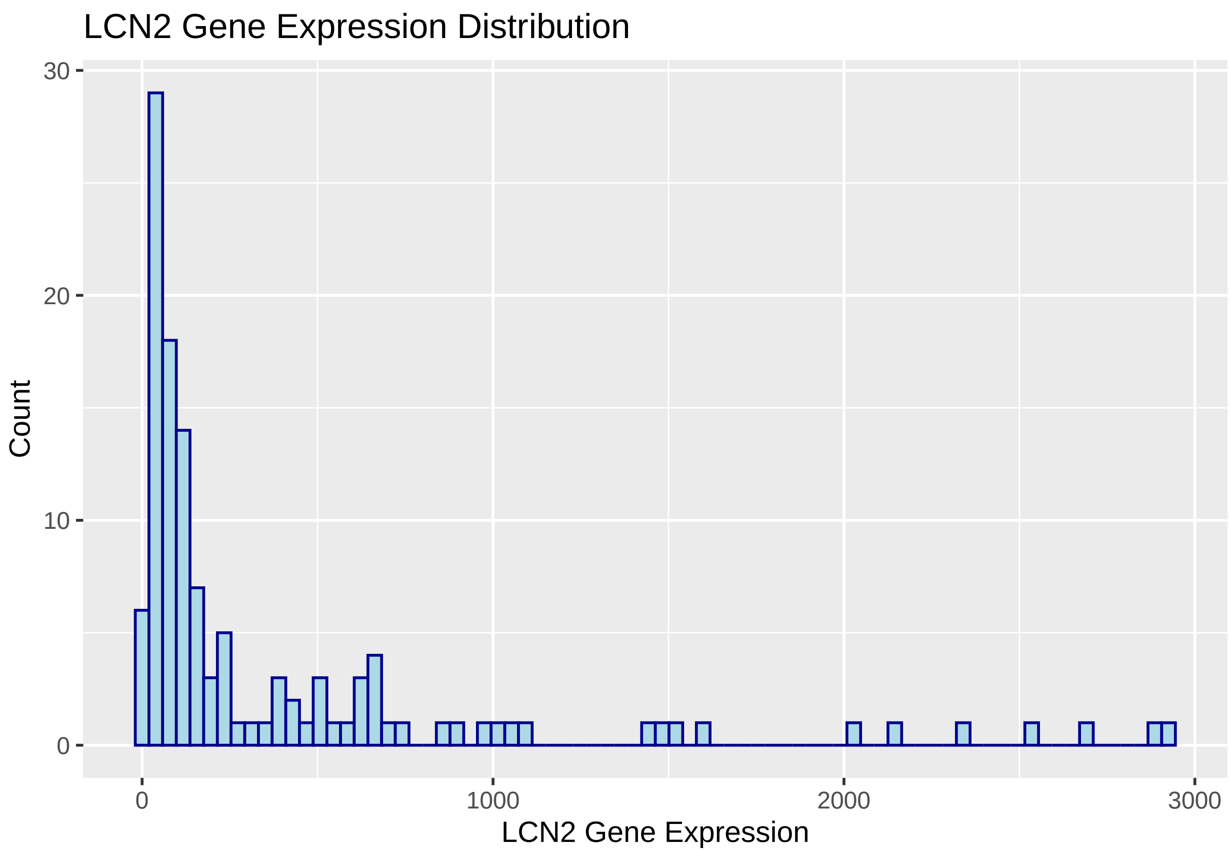
```
##      [,1]
## [1,] 87.71
## [2,] 662.59
## [3,] 121.86
## [4,] 31.73
## [5,] 16.42
## [6,] 447.98
```

```
# new data frame created to link variables from both data sets into one
myData <- data.frame('Subject ID' = seq(1:126), "Gene.Expression" = gene_expression,'Age' = meta_data$ag
head(myData)
```

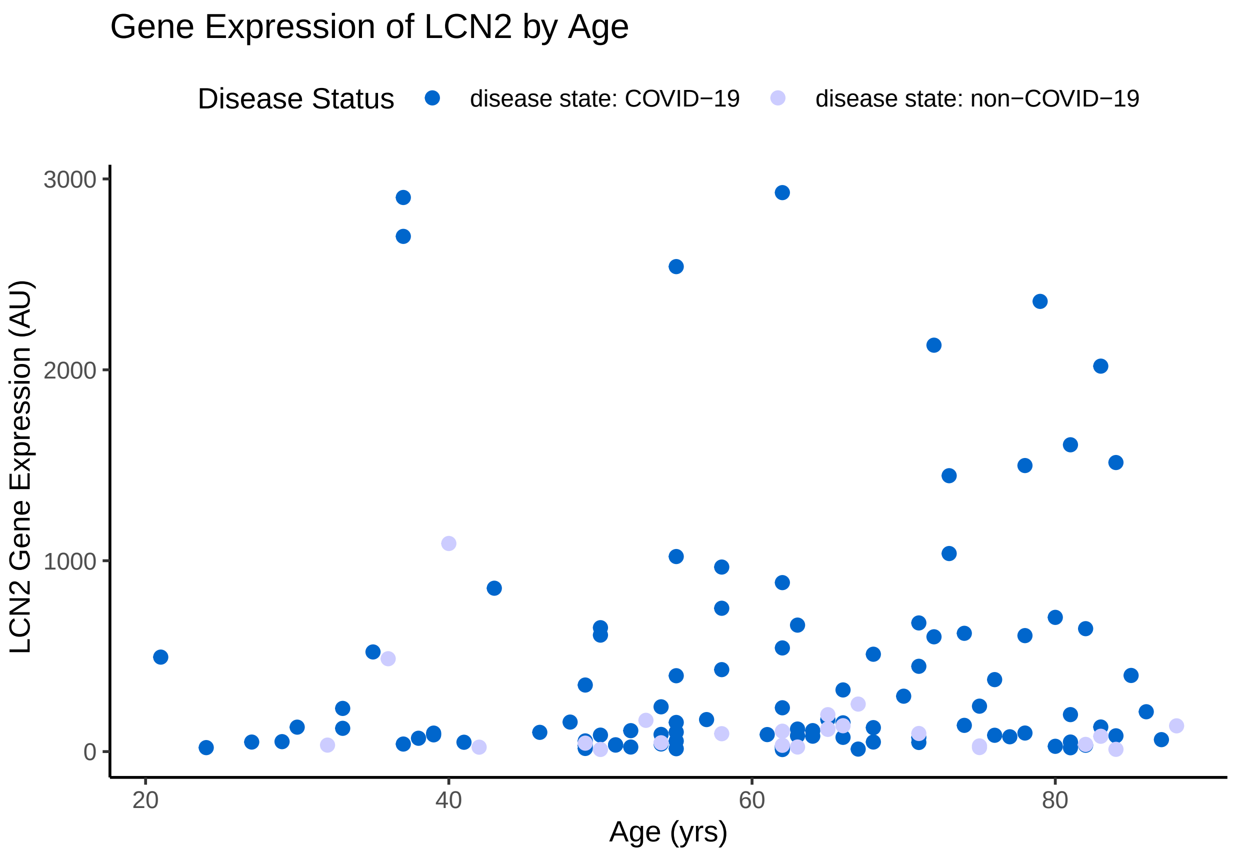
```
##   Subject.ID Gene.Expression Age   Sex          Disease.Status ICU.Status
## 1          1           87.71  39  male disease state: COVID-19         no
## 2          2          662.59  63  male disease state: COVID-19         no
## 3          3          121.86  33  male disease state: COVID-19         no
## 4          4           31.73  49  male disease state: COVID-19         no
## 5          5           16.42  49  male disease state: COVID-19         no
## 6          6          447.98   :  male disease state: COVID-19         no
```

```
myData <- myData[!(row.names(myData) %in% c("6","86","104","115")),] # filter out data that do not have
# Source: https://sparkbyexamples.com/r-programming/drop-dataframe-rows-in-r/?expand_article=1


myData$Age<- as.numeric(myData$Age)# Convert age variable to numeric
myData$Gene.Expression<- as.numeric(myData$Gene.Expression) #convert gene expression variable to numeri
```

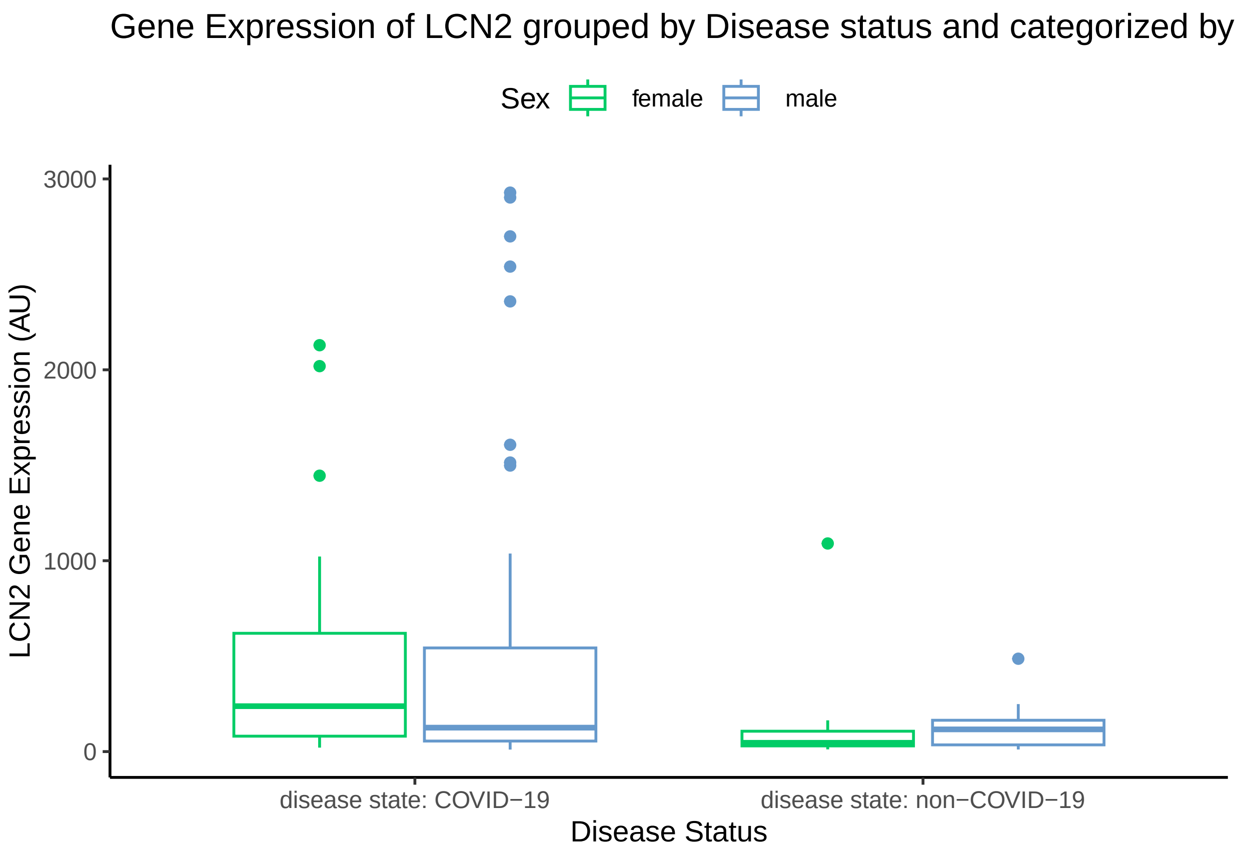
```
# Histogram for gene expression
ggplot(myData, aes(x=Gene.Expression)) +

  geom_histogram(binwidth = 39,color = "darkblue", fill = "lightblue")+

  labs(title = "LCN2 Gene Expression Distribution", x = "LCN2 Gene Expression", y = "Count")
```



```r
BlankTheme <- theme(# Remove borders and grid lines
        panel.border = element_blank(), panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        # Define my axis
        axis.line = element_line(colour = "black", linewidth = rel(1)),
        # Set plot background to white
        plot.background = element_rect(fill = "white"),
        panel.background = element_blank(),
        legend.key = element_rect(fill = 'white'),
        # Move legend to the top
        legend.position = 'top')
# Scatterplot for gene expression and continuous covariate
ggplot(myData, aes(x = Age, y = Gene.Expression, color = Disease.Status))+

  geom_point(size = 2)+

  labs(title = "Gene Expression of LCN2 by Age", x = "Age (yrs)", y = " LCN2 Gene Expression (AU)", col

  scale_color_manual(values = c("#0066CC", "#CCCCFF")) +

  BlankTheme
```



```
BlankTheme <- theme(# Remove borders and grid lines
        panel.border = element_blank(), panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        # Define my axis
        axis.line = element_line(colour = "black", linewidth = rel(1)),
        # Set plot background to white
        plot.background = element_rect(fill = "white"),
        panel.background = element_blank(),
        legend.key = element_rect(fill = 'white'),
        # Move legend to the top
        legend.position = 'top')
# Box plot of gene expression separated by sex and disease status

ggplot(myData, aes(x = Disease.Status, y = Gene.Expression, color = Sex))+

  geom_boxplot()+

  scale_color_manual(values = c('#00CC66','#6699CC')) +

  labs(title = "Gene Expression of LCN2 grouped by Disease status and categorized by Sex", x = "Disease

  BlankTheme
```


Gene Expression of LCN2 grouped by Disease status and categorized by
Sex
female
male
3000
2000
1000
0
LCN2 Gene Expression (AU)
disease state: COVID−19
disease state: non−COVID−19
Disease Status