R_final

Tanya Budhiraja

2025-07-10

import data

```
genes <- read.csv("~/Desktop/QBS103_GSE157103_genes.csv", row.names = 1)
# i am choosing gene: ABHD5 (row 97)
#continuous covariate: Age
#categorical covariates: sex & ICU status
metadata <- read.csv("~/Desktop/QBS103_GSE157103_series_matrix-1.csv")
#head(genes)
#head(metadata)
```

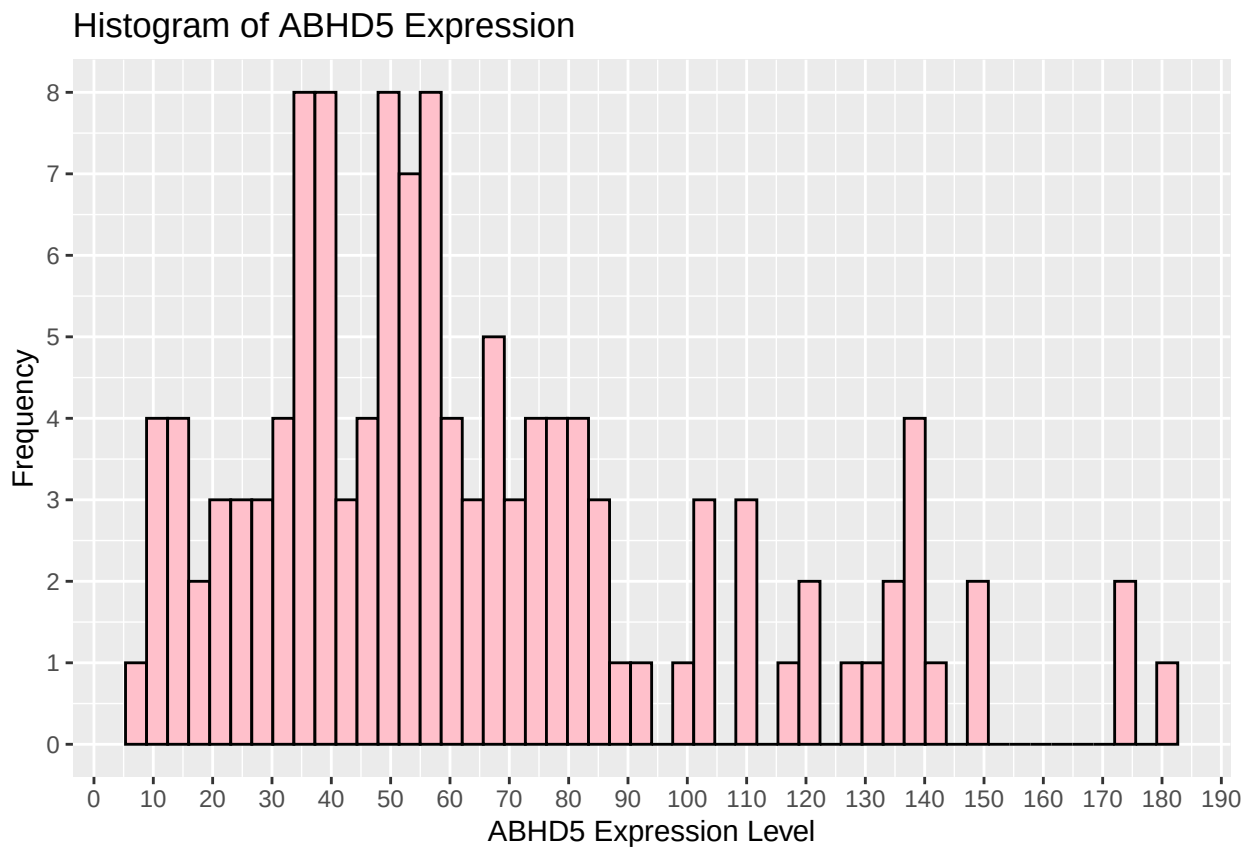
extract my gene's row from the gene df and add to metadata

```
metadata$ABHD5 <- as.numeric(genes["ABHD5", ])
#head(metadata)
```

Histogram for gene expression

```
library(ggplot2) #import lib

ggplot(metadata, aes(x = ABHD5)) +
  geom_histogram(bins = 50, fill = "pink", color = "black") + #initialize
  scale_x_continuous( #set x range
    breaks = seq(0, 190, by = 10),
  )+
  scale_y_continuous( #set y range
    breaks = seq(0, 10, by = 1),
  )+ #labels
  labs(title = "Histogram of ABHD5 Expression", x = "ABHD5 Expression Level",
    y = "Frequency"
  )
```



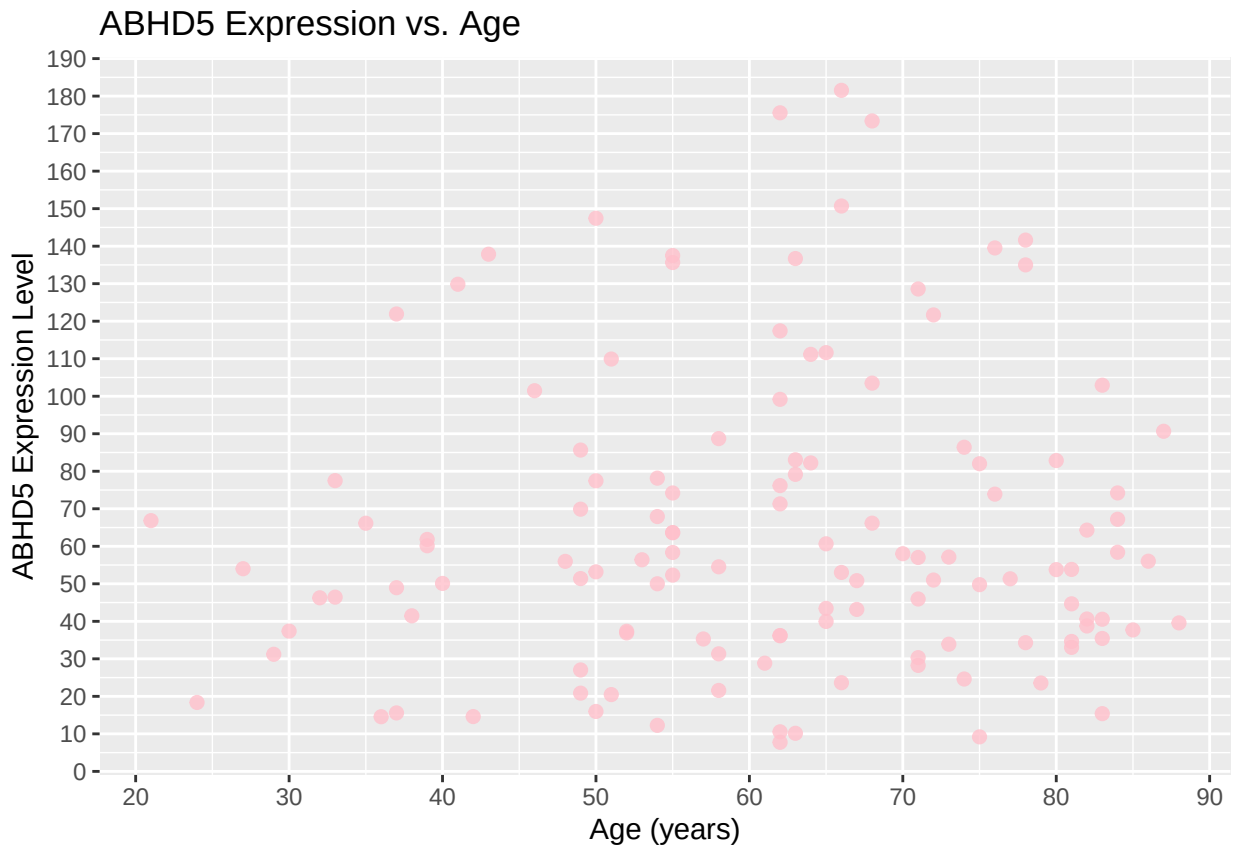
Scatterplot for gene expression and continuous covariate

```
#make age numeric
metadata$age <- as.numeric(metadata$age)
```

```
## Warning: NAs introduced by coercion
```

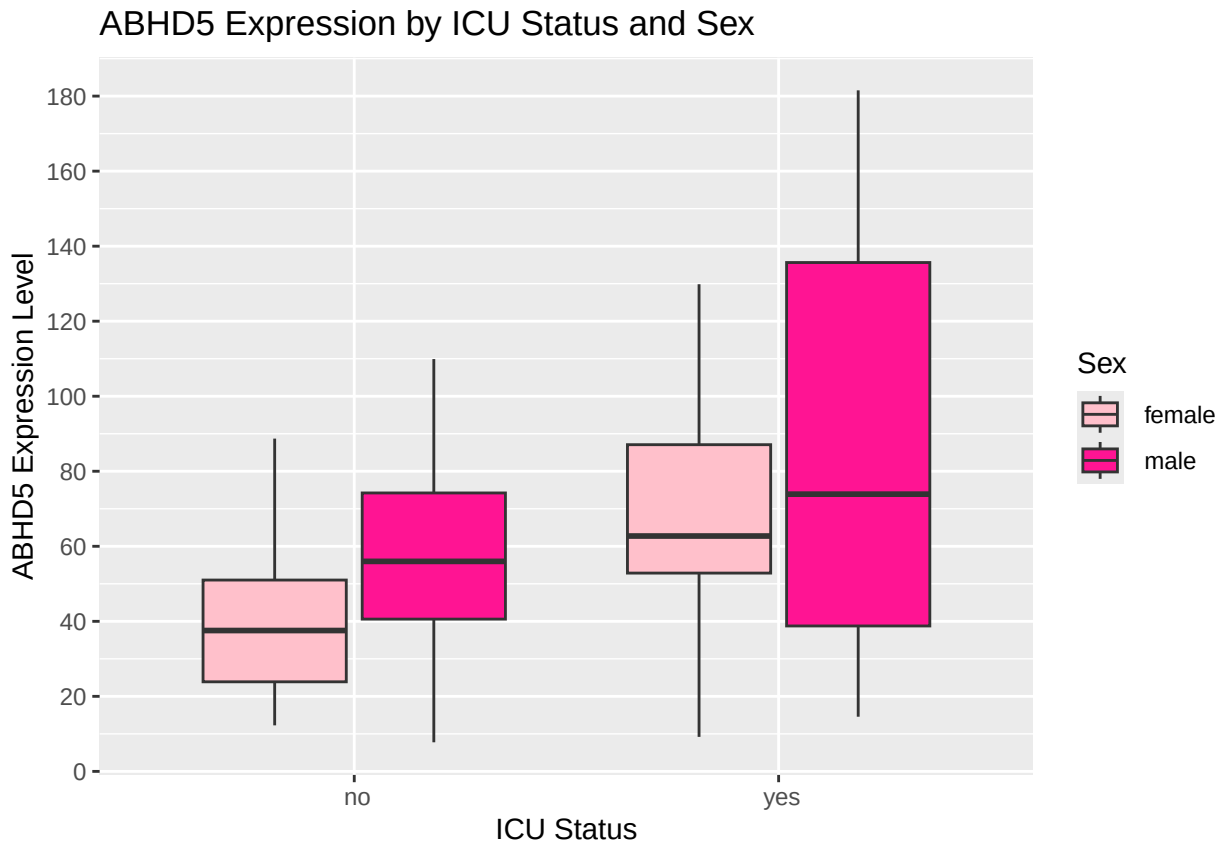
```
#clean data- remove na's
metadata_clean <- metadata[ !(is.na(metadata$age) | is.na(metadata$ABHD5)), ]

ggplot(metadata_clean, aes(x = age, y = ABHD5)) +
  scale_x_continuous( # set x range
    breaks = seq(10, 100, by = 10),
  ) +
  scale_y_continuous( # set y range
    breaks = seq(0, 200, by = 10),   # use full ABHD5 range
  ) +
  geom_point(size = 2, colour = "pink", alpha = 0.8) +     # data points
  labs( # axis labels & title
    title = "ABHD5 Expression vs. Age",
    x = "Age (years)",
    y = "ABHD5 Expression Level"
  )
```



Boxplot of gene expression separated by both categorical covariates

```
#clean data
#remove where sex is unknown
#find categories for sex
#unique(metadata$sex)
metadata_clean <- metadata_clean[metadata_clean$sex != " unknown", ]
#plot
ggplot(metadata_clean, aes(x = icu_status, y = ABHD5, fill = sex)) +
  geom_boxplot() +
  scale_fill_manual(values = c(" female" = "pink", " male" = "deeppink")) +
  labs(
    title = "ABHD5 Expression by ICU Status and Sex",
    x = "ICU Status",
    y = "ABHD5 Expression Level",
    fill = "Sex"
  ) +
  scale_y_continuous(
    breaks = seq(0, 200, by = 20),
  )
```



analysis:

histogram- shows the distribution ABHD5's expression across all the samples. The distribution is skewed towards the right with majority of values falling between 30 and 80 but a few samples showed higher expression (150–180). From this we can draw that while ABHD5 is expressed in most of the sampled individuals at a similar level, some individuals can have significantly elevated levels.

scatterplot- shows the relationship between patient age and ABHD5 expression. There doesn't appear to be a linear correlation between expression and age. The expression values are widely dispersed across all the measured ages indicating that expression does not vary consistently with age. From the graph we can also draw that at all ages there are extreme expression values. This may suggest that there are other factors other than age influence ABHD5 expression.

boxplot- compares ABHD5 expression across ICU status and seperates by sex.Non ICU patients tended to show lower expression with less variability for both sexes. Among ICU patients expression appears higher with more variability (especially for males). This may show a potential correlation with elevated ABHD5 expression and ICU status -> possibly meaning higher gene expression is correlated with more severe COVID-19 cases. The difference based on sex may show natural biological variation in expression.