Submission1_CN

CN

2024-07-29

R Markdown

```
#import counts and series matrix as GSE157103_counts and GSE157103_series
GSE157103_counts <- read.csv("~/Downloads/QBS103_GSE157103_genes.csv")
GSE157103_series <- read.csv("~/Downloads/QBS103_GSE157103_series_matrix.csv")

#make the imported files dataframes
GSE157103_counts <- as.data.frame(GSE157103_counts)
GSE157103_series <- as.data.frame((GSE157103_series))


#set the gene names as rownames of GSE157103_counts
rownames(GSE157103_counts) <- GSE157103_counts$X
GSE157103_counts <- GSE157103_counts[,-1]
```

```
#check dimensions of data
dim(GSE157103_counts)
```

```
## [1] 100 126
```

```
dim(GSE157103_series)
```

```
## [1] 126  25
```

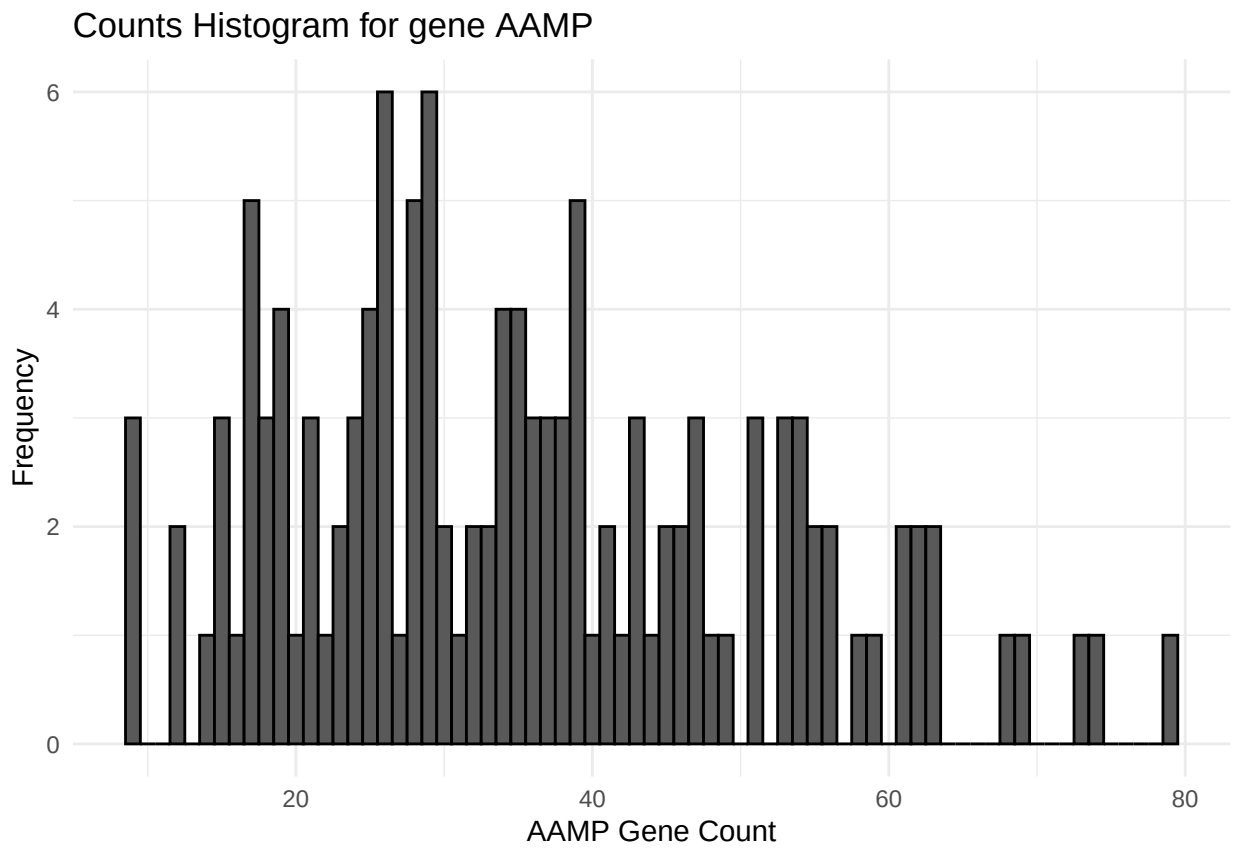
```
#subset AAMP as the gene of interest
#AAMP is tumor-progression associated
GSE157103_counts_AAMP <- as.data.frame(t(GSE157103_counts[rownames(GSE157103_counts)== "AAMP",]))
GSE157103_counts_AAMP$counts <- GSE157103_counts_AAMP$AAMP
GSE157103_counts_AAMP$ID <- colnames(GSE157103_counts)

#assign IDs to each count for the AAMP gene
rownames(GSE157103_counts_AAMP) <- 1:126

#generate histigram for "AAMP"
#load ggplot2
library(ggplot2)


ggplot(GSE157103_counts_AAMP, aes(x = counts)) +
```

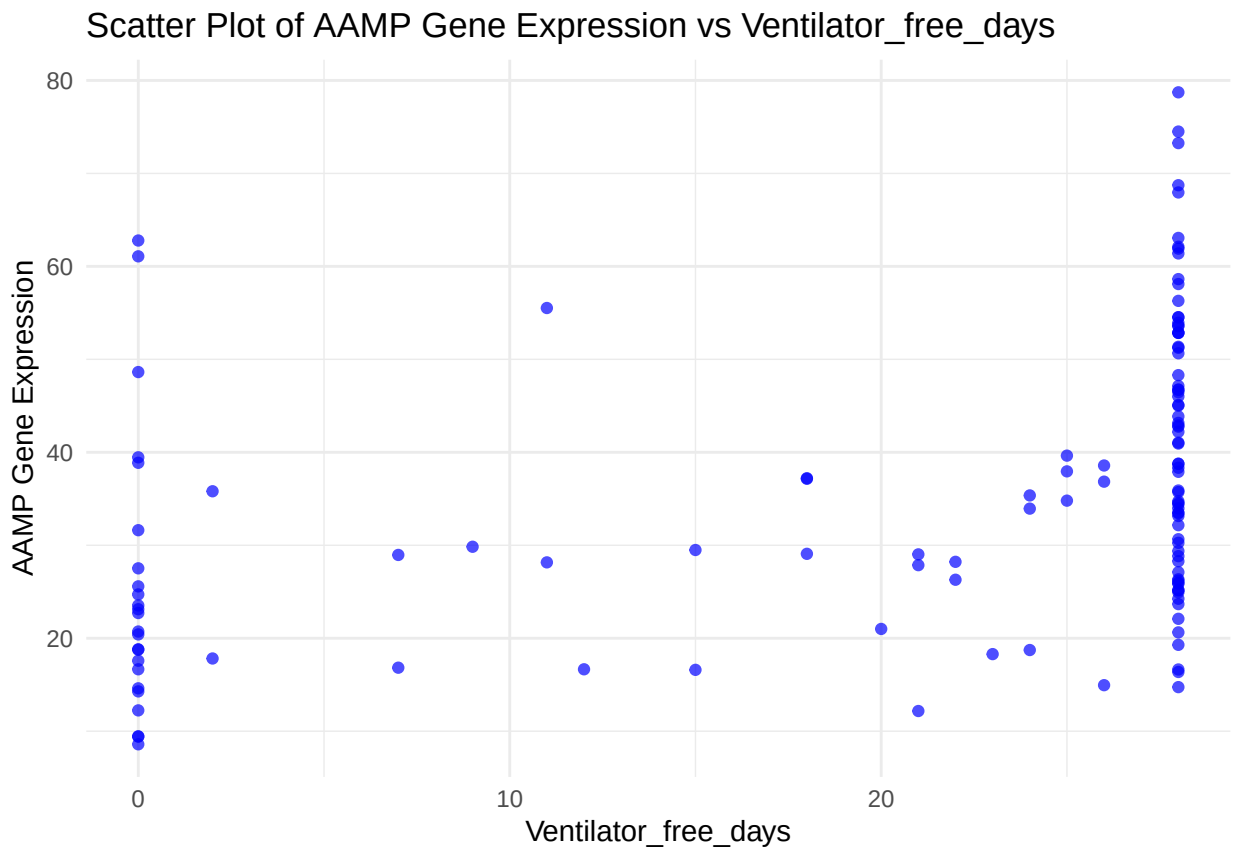
```
  geom_histogram(binwidth = 1, color = "black") +
  labs(title = "Counts Histogram for gene AAMP",
       x = "AAMP Gene Count",
       y = "Frequency") +
  theme_minimal()
```



#plot gene expression and ventilator-free days

```
#make a df that includes gene expression counts and ventilator-free datys data
GSE157103_counts_AAMP$ventilator.free_days <- GSE157103_series$ventilator.free_days

ggplot(GSE157103_counts_AAMP, aes(x = ventilator.free_days, y = counts)) +
  geom_point(alpha = 0.7, color = "blue") +
  labs(title = "Scatter Plot of AAMP Gene Expression vs Ventilator_free_days",
       x = "Ventilator_free_days",
       y = "AAMP Gene Expression") +
  theme_minimal()
```




```
#Boxplot of gene expression separated by both categorical covariates (5 pts)
#Gene expression, Disease State Sex
GSE157103_counts_AAMP$disease_status <- GSE157103_series$disease_status
GSE157103_counts_AAMP$sex <- GSE157103_series$sex



newBlankTheme <- theme(# Remove all the extra 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
plot.background = element_rect(fill = "white"),
panel.background = element_blank(),
legend.key = element_rect(fill = 'white'),
# Move legend
legend.position = 'top')

#exclude the datapoint whose sex is unknown
ggplot(GSE157103_counts_AAMP[-115,],aes(x = sex,y = counts, color = disease_status)) +
# Add box plot
geom_boxplot() +
# Define colors
scale_color_manual(values = c('darkgreen','grey50')) +
```

```
# Change labels
labs(x = 'Sex',y = 'Gene Expression Counts',color = 'Disease Status') +
# Set theme
newBlankTheme
```