R Submission 2

Ruiqi Li

2025-07-22

Overview

In this submission, I: 1. Write a function (`create_plot`) that generates the three figures from Presentation 1, taking as input

(a) a data frame,
(b) a list of >=1 gene names,
(c) one continuous covariate, and
(d) two categorical covariates.

2. Loop over 3 selected genes to produce all plots.
3. Prepare one boxplot to discuss in class.

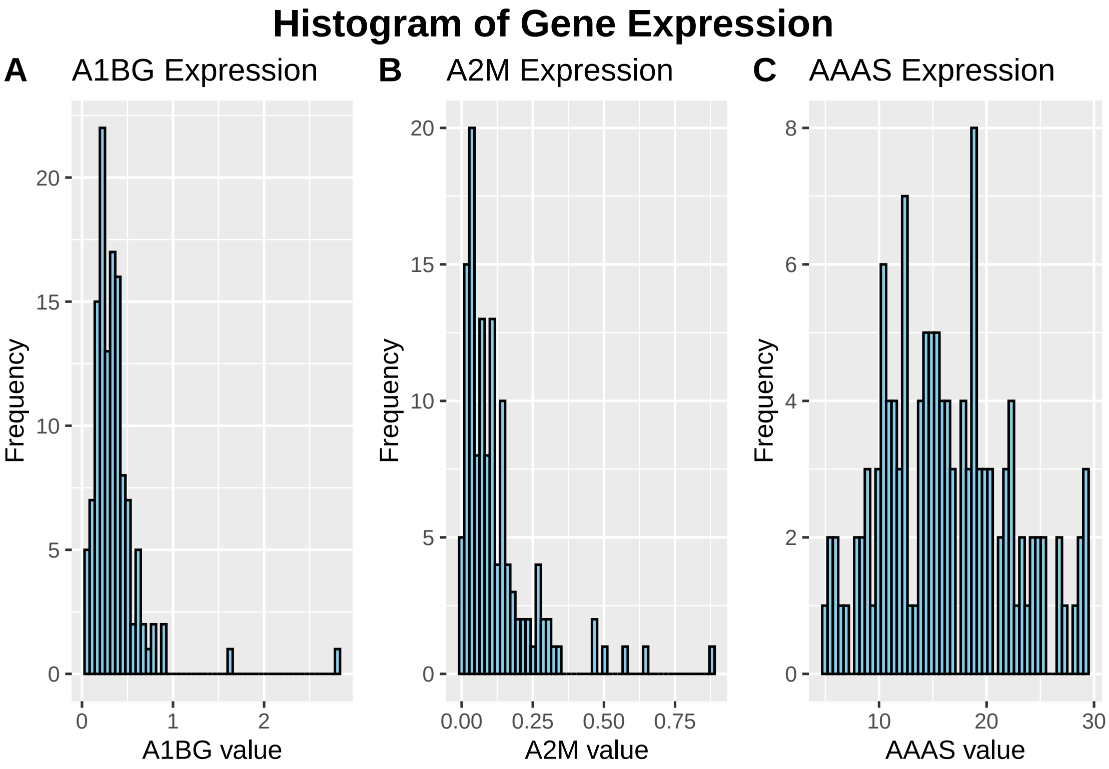
Data used: `QBS103_GSE157103_genes.csv` (gene expression) and `QBS103_GSE157103_series_matrix-1.csv` (metadata).

```
#submission 2
#Load gene expression data
gene_data <- read.csv(file = "QBS103_GSE157103_genes.csv",row.names=1)

#Load metadata for participants
series <- read.csv(file = "QBS103_GSE157103_series_matrix-1.csv")

# Build plots for a list of genes
create_plot<-function(gene_data, metadata,gene_list,
                      continue_v,categorical_v1,categorical_v2){
  histog_list=list()
  scatter_list=list()
  boxplot_list=list()
  for(i in gene_list){
     # 1. Pull one gene row (expression across all samples)
    new_gene <- gene_data[i, ]

    # 2. Histogram for that gene
    histog_list[[i]]<-ggplot(data.frame(value = as.numeric(new_gene)),
                             aes(x = value)) +
    geom_histogram(bins = 50, color = "black", fill = "skyblue") +
    labs(title = paste(i, "Expression"),
         x     = paste(i, "value"),
         y     = "Frequency")
```

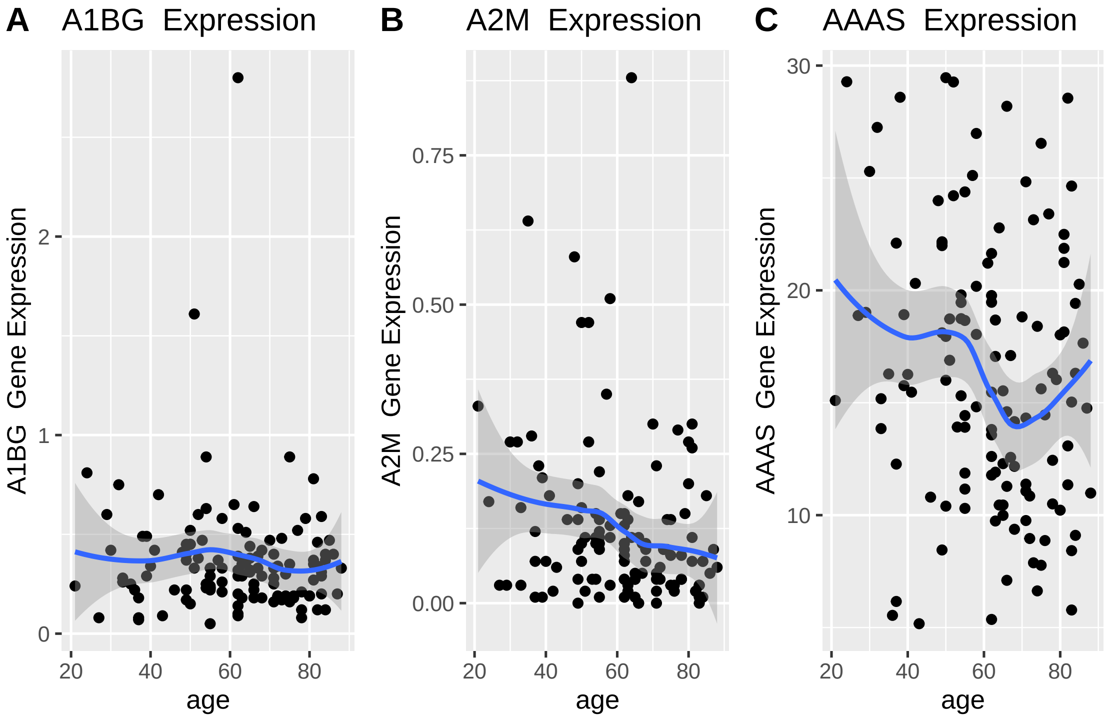
```
   # 3. Long format + merge with metadata
   gene_line1 <- new_gene %>%
    pivot_longer(cols = everything(),
                 names_to = "participant_id",
                 values_to = "expr")

   new_df<-merge(metadata,gene_line1,by="participant_id")
   new_df$expr <- as.numeric(new_df$expr)
   new_df$cont_num <- suppressWarnings(as.numeric(new_df[[continue_v]]))
   new_df <- filter(new_df, !is.na(cont_num), !is.na(expr))

   # 4. Scatterplot: expression vs continuous covariate
   scatter_list[[i]]<-ggplot(new_df, aes(x = as.numeric(cont_num),
                                         y = expr)) +
     geom_point() +
     geom_smooth(method = "loess", se = TRUE, formula = y ~ x) +
     labs(title = paste(i," Expression"),
          x = continue_v,
          y = paste(i," Gene Expression"))

   # 5. Boxplot: expression by two categorical covariates
   boxplot_list[[i]]<-ggplot(new_df,aes(x=.data[[categorical_v1]],
                                        y=expr,
                                        color=.data[[categorical_v2]]))+
     geom_boxplot()+
     labs(
       title = paste(i," Expression"),
       x = categorical_v1,
       y =paste( i," Gene Expression"),
       color = categorical_v2)

  }

    arrange_plot(histog_list,"Histogram of Gene Expression")
    arrange_plot(scatter_list,paste("Scatterplot of Gene Expression versus ",
                                    continue_v))
    arrange_plot(boxplot_list,paste("Gene Expression by ",
                                    categorical_v1,
                                    " and ",
                                    categorical_v2))
}
```

```
# Arrange a list of ggplots and add a common title
arrange_plot<-function(plot_list,title_text){
  n <- length(plot_list)
  final_plot <- do.call(
  ggarrange,
  c(
    plot_list,
    list(
      labels = LETTERS[1:n],
      ncol = n,
      nrow = 1,
      common.legend = TRUE,
      legend = "right"
      )
    )
  )
  final_plot_with_title <- annotate_figure(
    final_plot,
    top = text_grob(title_text,
                  face = "bold", size = 16)
    )
  print(final_plot_with_title)
}
```

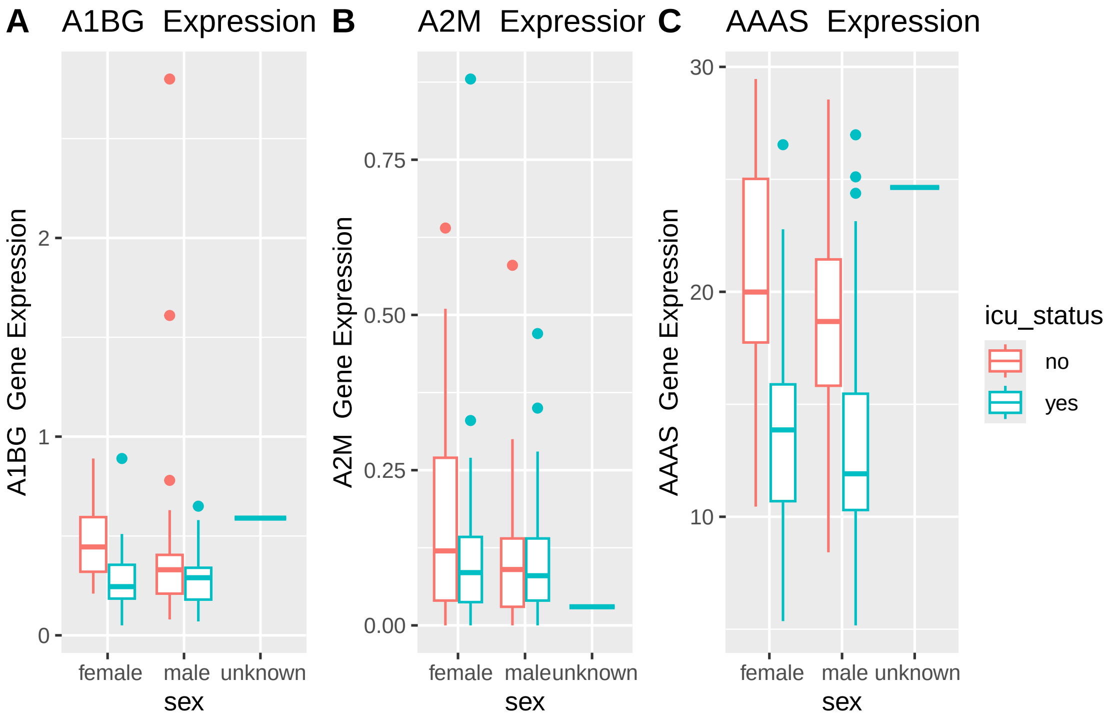
```
# Pick three genes by row index (1, 3, 8)
gene_list=list(rownames(gene_data)[1],
               rownames(gene_data)[3],
               rownames(gene_data)[8])

# Create all required plots: df, gene list, 1 continuous ("age"),
# 2 categorical ("sex","icu_status")
create_plot(gene_data, series,gene_list,"age","sex","icu_status")
```



Scatterplot of Gene Expression versus age



Gene Expression by sex and icu_status


A
A1BG Expression
B
A2M Expressior
C
AAAS Expression
A1BG Gene Expression
A2M Gene Expression
AAAS Gene Expression
female
male
unknown
sex
icu_status
no
yes