RNA Clustering

Michael Callahan

2024-04-03

Load data

```
# library(here)
# library(dplyr)
#
# current_project_path <- here()
#
# file_path <- file.path(current_project_path, "Data", "CMP_RNA_1.tsv")
# CMP_RNA_1 <- read.delim(file_path)
#
# file_path <- file.path(current_project_path, "Data", "CMP_RNA_2.tsv")
# CMP_RNA_2 <- read.delim(file_path)
#
# file_path <- file.path(current_project_path, "Data", "CFU-E_RNA_1.tsv")
# CFUE_RNA_1 <- read.delim(file_path)
#
# file_path <- file.path(current_project_path, "Data", "CFU-E_RNA_2.tsv")
# CFUE_RNA_2 <- read.delim(file_path)
#
# file_path <- file.path(current_project_path, "Data", "Erythroblast_RNA_1.tsv")
# EBLST_RNA_1 <- read.delim(file_path)
#
# file_path <- file.path(current_project_path, "Data", "Erythroblast_RNA_2.tsv")
# EBLST_RNA_2 <- read.delim(file_path)
#
# file_path <- file.path(current_project_path, "Data", "HSC_RNA_1.tsv")
# HSC_RNA_1 <- read.delim(file_path)
#
# file_path <- file.path(current_project_path, "Data", "HSC_RNA_2.tsv")
# HSC_RNA_2 <- read.delim(file_path)
# ```
#
# ## Data preprocessing
#
# ```{r preprocess}
# #FYI - gene_ids are unique
#
# #CREATE A TPM Dataframe
# # Add prefixes to the tpm columns in each dataframe
# CFUE_RNA_1_tpm <- CFUE_RNA_1 %>% select(gene_id, CFUE_RNA_1_tpm = TPM)
# CFUE_RNA_2_tpm <- CFUE_RNA_2 %>% select(gene_id, CFUE_RNA_2_tpm = TPM)
```

```r
# CMP_RNA_1_tpm <- CMP_RNA_1 %>% select(gene_id, CMP_RNA_1_tpm = TPM)
# CMP_RNA_2_tpm <- CMP_RNA_2 %>% select(gene_id, CMP_RNA_2_tpm = TPM)
# EBLST_RNA_1_tpm <- EBLST_RNA_1 %>% select(gene_id, EBLST_RNA_1 = TPM)
# EBLST_RNA_2_tpm <- EBLST_RNA_2 %>% select(gene_id, EBLST_RNA_2 = TPM)
# HSC_RNA_1_tpm <- HSC_RNA_1 %>% select(gene_id, HSC_RNA_1 = TPM)
# HSC_RNA_2_tpm <- HSC_RNA_2 %>% select(gene_id, HSC_RNA_2 = TPM)
#
# # Join the dataframes together on gene_id
# tpm <- inner_join(CFUE_RNA_1_tpm, CFUE_RNA_2_tpm, by = "gene_id") %>%
#        inner_join(CMP_RNA_1_tpm, by = "gene_id") %>%
#        inner_join(CMP_RNA_2_tpm, by = "gene_id") %>%
#        inner_join(EBLST_RNA_1_tpm, by = "gene_id") %>%
#        inner_join(EBLST_RNA_2_tpm, by = "gene_id") %>%
#        inner_join(HSC_RNA_1_tpm, by = "gene_id") %>%
#        inner_join(HSC_RNA_2_tpm, by = "gene_id")
#
# rm(
#   CFUE_RNA_1_tpm, CFUE_RNA_2_tpm,
#   CMP_RNA_1_tpm, CMP_RNA_2_tpm,
#   HSC_RNA_1_tpm, HSC_RNA_2_tpm,
#   EBLST_RNA_1_tpm, EBLST_RNA_2_tpm,
#   CMP_RNA_1, CMP_RNA_2,
#   CFUE_RNA_1, CFUE_RNA_2,
#   HSC_RNA_1, HSC_RNA_2,
#   EBLST_RNA_1, EBLST_RNA_2
#   )
#
# tpm <- tpm %>%
#   rename_at(vars(2:5), ~ sub("_tpm$", "", .))
#
#
# # Set 'column_name' as row names
# rownames(tpm) <- tpm$gene_id
#
# # Remove 'column_name' from dataframe (optional)
# tpm <- tpm[, -which(names(tpm) == 'gene_id')]
#
# #Drop all rows with only 0s
# tpm <- tpm[rowSums(tpm != 0) > 0, ]
#
# save(tpm, file = 'all_tpm.Rdata')
```

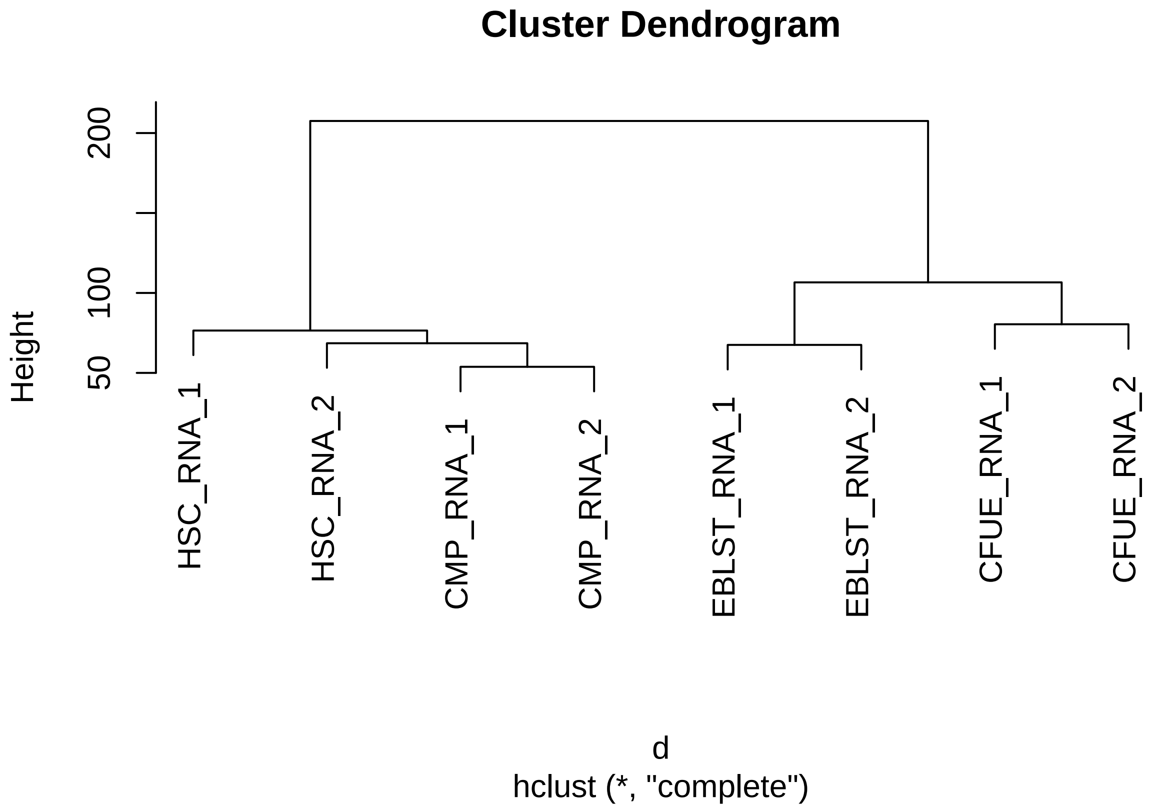
Clustering

```r
library(here)
```

```
## Warning: package 'here' was built under R version 4.3.3
```

```
## here() starts at C:/Users/mgcal/OneDrive/Documents/School/Courses/Stat 555/Project/Stat555Project
```

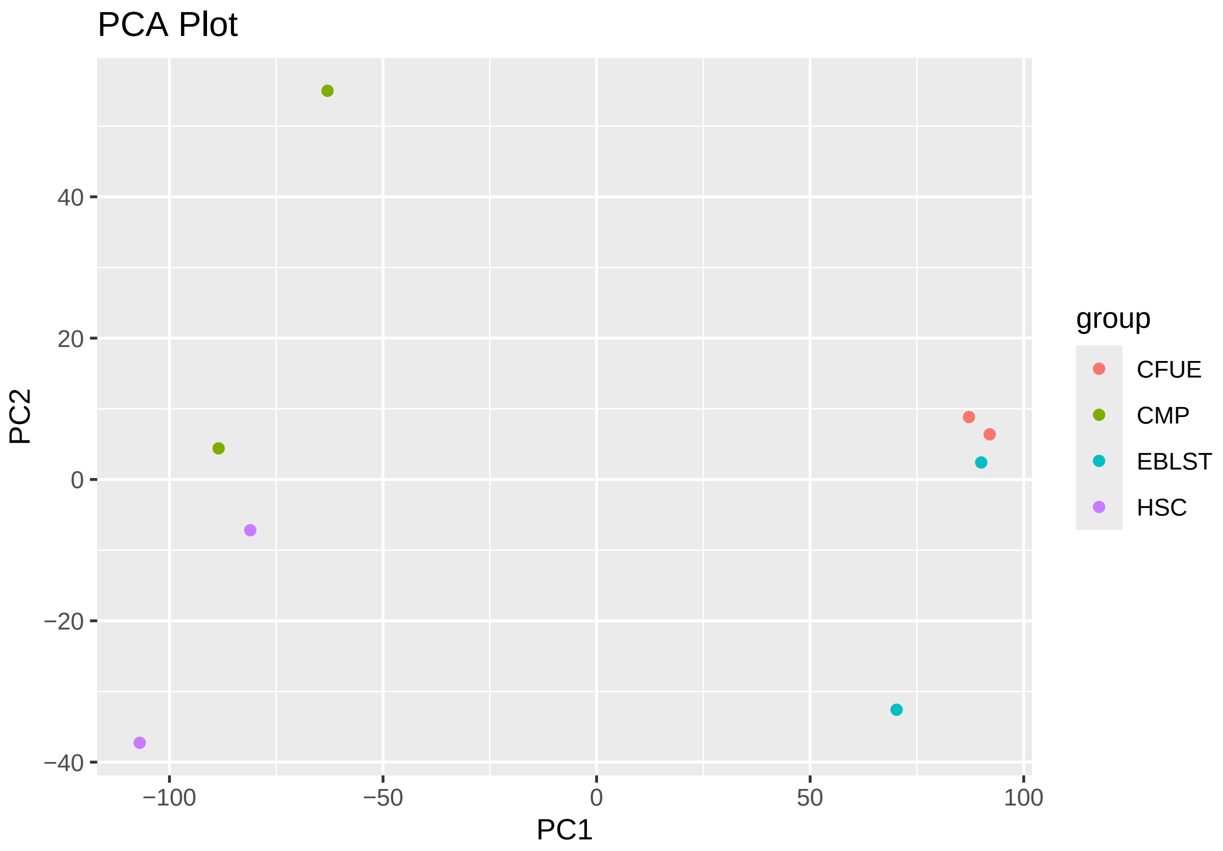
```r
library(stats)

current_project_path <- here()

file_path <- file.path(current_project_path, "Data", "all_tpm.Rdata")
load(file_path)

row_variances <- apply(tpm, 1, var)
tpm <- tpm[order(row_variances, decreasing = TRUE)[1:5000], ]

tpm <- log2(tpm + 1)

d=dist(t(tpm))
hc=hclust(d,method="complete")
plot(hc)
```



```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```r
pcaResults <- prcomp(t(tpm))
pc <- data.frame(PC1 = pcaResults$x[,1],
                 PC2 = pcaResults$x[,2],
                 PC3 = pcaResults$x[,3],
                 PC4 = pcaResults$x[,4],
                 PC5 = pcaResults$x[,5]
                 )

#Create a colData matrix
colData = data.frame(
  source_name = c('CMP_RNA_1','CMP_RNA_2',
                  'CFUE_RNA_1','CFUE_RNA_2',
                  'HSC_RNA_1','HSC_RNA_2',
                  'EBLST_RNA_1','EBLST_RNA_2'
                  ),
  group = c('CMP', 'CMP', 'CFUE', 'CFUE', 'HSC', 'HSC', 'EBLST', 'EBLST')
  )

# Add sample metadata from colData
pc <- cbind(pc, colData)

# Plot PCA results using ggplot2
ggplot(pc, aes(x = PC1, y = PC2, color = group)) +
  geom_point() +
  labs(title = "PCA Plot", x = "PC1", y = "PC2")
```



```r
summary(pcaResults)
```

```
## Importance of components:
```

```
##                            PC1      PC2      PC3      PC4     PC5      PC6
## Standard deviation     91.7525 28.46336 23.46976 19.62037 15.1122 13.73010
## Proportion of Variance  0.7853  0.07557  0.05138  0.03591  0.0213  0.01759
## Cumulative Proportion   0.7853  0.86089  0.91227  0.94818  0.9695  0.98707
##                             PC7       PC8
## Standard deviation     11.77356 8.776e-14
## Proportion of Variance  0.01293 0.000e+00
## Cumulative Proportion   1.00000 1.000e+00
```