Alzheimer EDA

Paul K. Yu

2024-11-21

Exploratory data analysis

Do this once (set eval to TRUE only once)

```
# Install required libraries if not already installed
if (!requireNamespace("pheatmap", quietly = TRUE)) install.packages("pheatmap")
if (!requireNamespace("ggplot2", quietly = TRUE)) install.packages("ggplot2")
```

Get data

```
library(data.table)

# Get RNA data
epxression <- as.data.frame(fread("alzheimer/GSE125583_raw_counts_GRCh38.p13_NCBI.tsv", header = TRUE, s

# Remove NA
epxression <- na.omit(epxression)

# Remove duplicate genes
epxression <- epxression[!duplicated(epxression$GeneID), ]

# Make gene ID as row name
rownames(epxression) <- epxression[[1]]  # Set the first column as row names

# Remove gene ID column
epxression <- epxression[, -c(1)]

n <- ncol(epxression) ## number of subjects
g <- nrow(epxression) ## number of gene features

# Get outcome data
# Manually select only ID_REF and stage
outcome <- as.data.frame(t(fread("alzheimer/GSE125583_series_matrix.txt", header = FALSE, sep = "\t")))
outcome <- setNames(outcome, c("ID_REF", "stage"))
outcome <- outcome[-c(1), ]
outcome$stage <- as.factor(outcome$stage)
levels(outcome$stage)
```

```
## [1] "braak.score: I"   "braak.score: II"  "braak.score: III" "braak.score: IV"
## [5] "braak.score: NA"  "braak.score: V"   "braak.score: VI"
```

```r
levels(outcome$stage) <- c(1,2,3,4,0,5,6)

# Make ID_REF as row name
rownames(outcome) <- outcome[[1]]  # Set the first column as row names

# Remove NA
outcome <- na.omit(outcome)
matches <- colnames(epxression) %in% rownames(outcome)
epxression <- epxression[, matches]
matches <- rownames(outcome) %in% colnames(epxression)
outcome <- outcome[matches,]
```

Summary of the dataset

Get summary of outcome data

```r
summary(outcome)
```

```
##     ID_REF          stage
##  Length:289         1: 4
##  Class :character   2:15
##  Mode  :character   3:44
##                     4:99
##                     0: 2
##                     5:65
##                     6:60
```

Check for missing values

```r
sum(is.na(epxression))
```

```
## [1] 0
```

```r
sum(is.na(outcome))
```

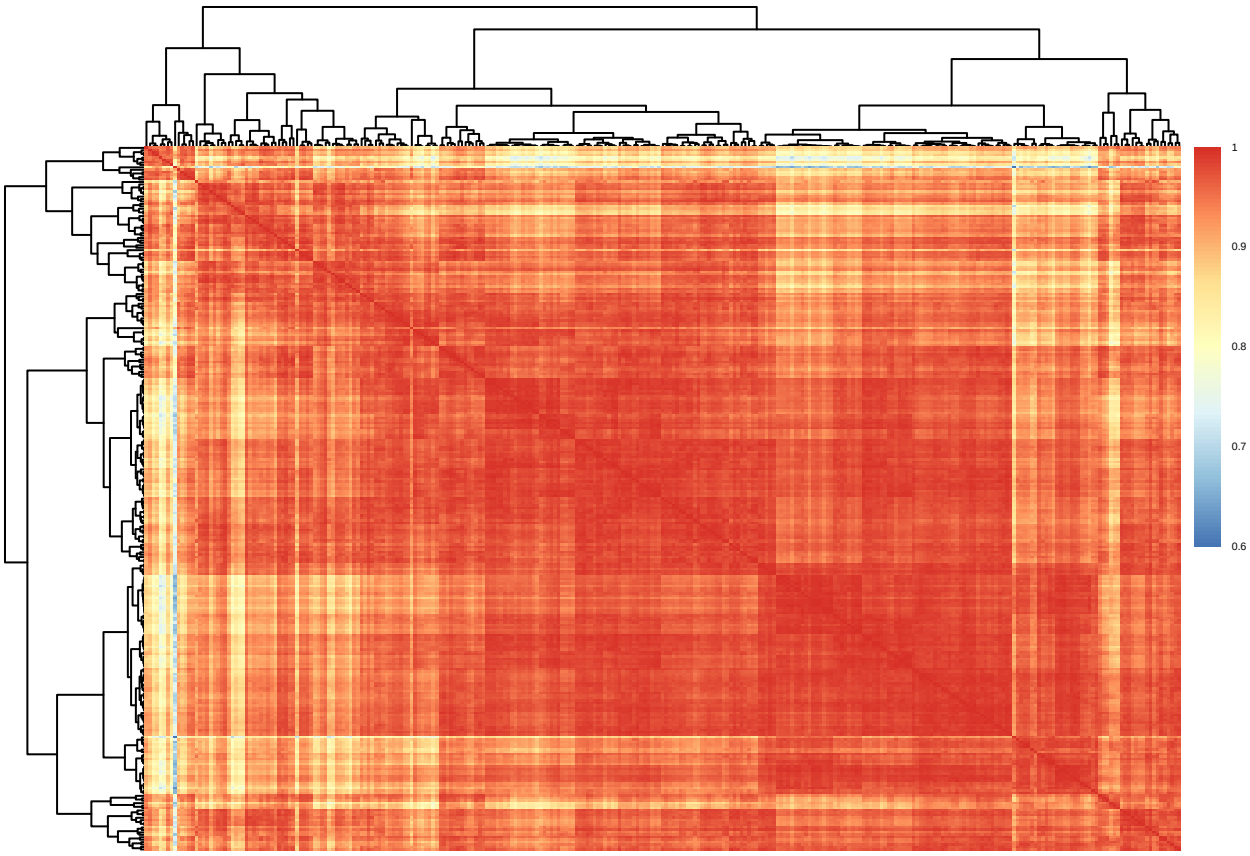
```
## [1] 0
```

Correlation analysis

```r
library(pheatmap)

# Compute correlations
cor_matrix <- cor(epxression, use = "pairwise.complete.obs")

# Visualize correlations as a heatmap
pheatmap(cor_matrix,
         show_rownames = FALSE,
         show_colnames = FALSE,
         clustering_distance_rows = "correlation",
         clustering_distance_cols = "correlation",
         fontsize = 4)
```


1
0.9
0.8
0.7
0.6