

Metabric EDA

Paul K. Yu

2024-11-19

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("metabric/data_mrna_illumina_microarray.txt", header = TRUE, sep = "\t"))

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

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

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

# Create a table of gene name and ID
gene_names <- epxression[, c(1, 2)]

# Remove gene name and ID columns
epxression <- epxression[, -c(1, 2)]

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

# Get outcome data
outcome <- fread("metabric/brca_metabric_clinical_data.tsv", header = TRUE, sep = "\t")
outcome <- outcome[, c("Patient ID", "Overall Survival Status")]
outcome$`Overall Survival Status` <- as.factor(outcome$`Overall Survival Status`)
levels(outcome$`Overall Survival Status`) <- c(0,1)

# Remove NA
outcome <- na.omit(outcome)
```

```
matches <- colnames(epxression) %in% outcome$`Patient ID`  
epxression <- epxression[, matches]  
matches <- outcome$`Patient ID` %in% colnames(epxression)  
outcome <- outcome[matches, ]
```

Summary of the dataset

Get summary of outcome data

```
summary(outcome)
```

```
##   Patient ID      Overall Survival Status  
##   Length:1980      0: 837  
##   Class :character 1:1143  
##   Mode  :character
```

Check for missing values

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

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

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

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

Correlation analysis

```
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,  
         main = "Correlation Heatmap")
```

