# Data Analytics II/III: In Class Solutions

QBS 103: Foundations of Data Science

August 6, 2024

**In Class Acitivity**

1. Write a function to calculate the relative abundance of each miRNA in each sample. Verify that each sample has a total relative abundance of 1.

```
# Load data (available on canvas)
load('EVmiRNA.RData')

# Build function
calculateRelAbun <- function(x) { # x reflects a df with samples in columns and miRNA in rows
  # Calculate total miRNA for each sample
  total.miRNA <- apply(x,MARGIN = 2,FUN = sum)
  # Generate empty table for relative abundance
  relAbun <- as.data.frame(matrix(ncol = ncol(x),nrow = nrow(x)),
                           row.names = row.names(x))
  colnames(relAbun) <- colnames(x)
  # Loop through samples
  for (sample in colnames(x)) {
    relAbun[,sample] <- x[,sample]/total.miRNA[sample]
  }
  relAbun
}

# Calculate for our data frame
relAbun <- calculateRelAbun(miRNA)

# Check dimensions (should still be 798 x 25)
dim(relAbun)
```

```
## [1] 798  25
```

```
# Check all columns sum to 1
apply(relAbun,MARGIN = 2,FUN = sum)
```

```
##  Subject1  Subject2  Subject3  Subject4  Subject5  Subject6  Subject7  Subject8
##         1         1         1         1         1         1         1         1
##  Subject9 Subject10 Subject11 Subject12 Subject13 Subject14 Subject15 Subject16
##         1         1         1         1         1         1         1         1
## Subject17 Subject18 Subject19 Subject20 Subject21 Subject22 Subject23 Subject24
##         1         1         1         1         1         1         1         1
```

```
## Subject25
##         1
```

2. Using the apply function, identify the highest relative abundance each miRNA has in a single sample.

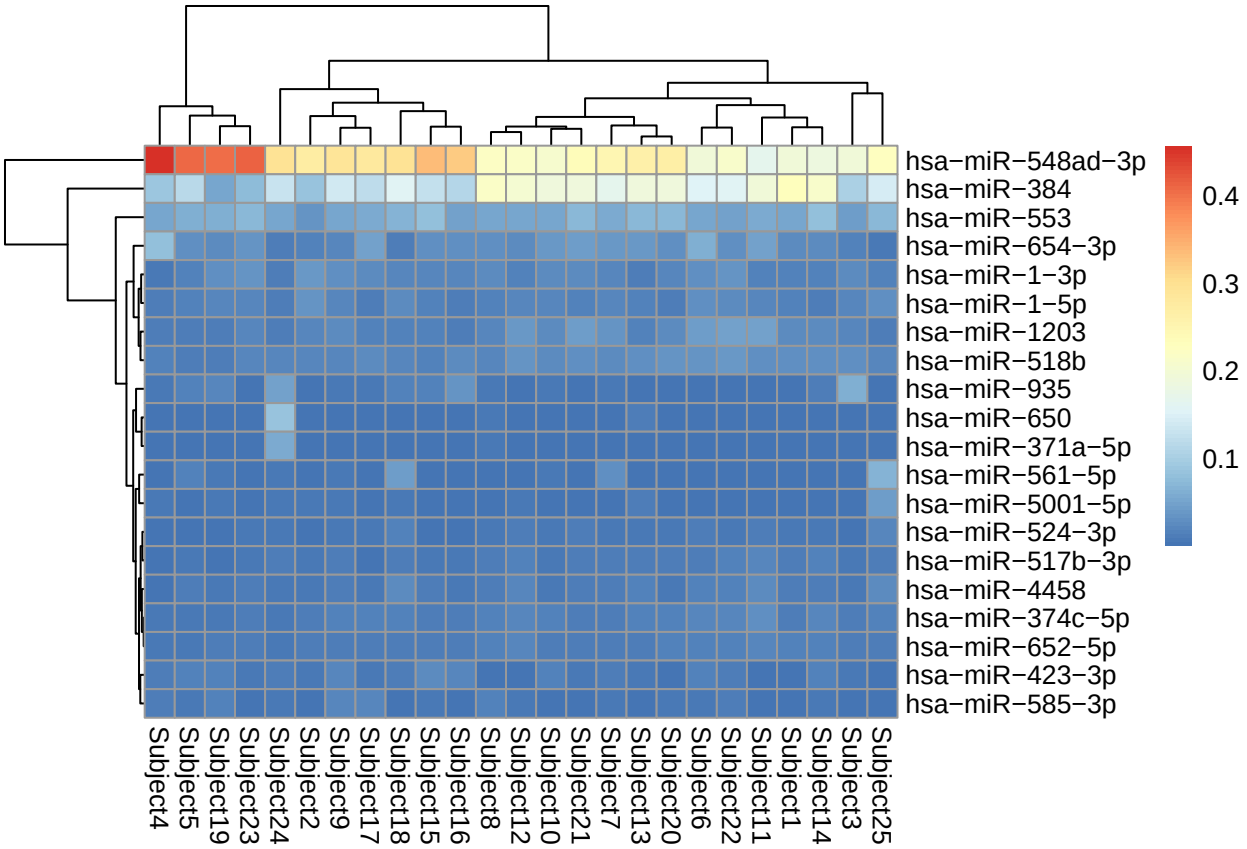
```
# Apply max() function on all rows to return max value for each miRNA
topRelAbun <- apply(relAbun,MARGIN = 1,FUN = max)
```

3. Sort the dataset by miRNA with the highest relative abundance and generate a heatmap of the relative abundance (not the absolute counts) of each miRNA, including the top 20 miRNA by single-sample relative abundance.

```
# Reorder relative abundance data set by "topRelAbun" with highest values on top
relAbun <- relAbun[order(topRelAbun,decreasing = T),]

# Load package
library(pheatmap)

# Generate heatmap
pheatmap(relAbun[1:20,],
         clustering_distance_cols = 'euclidean',
         clustering_distance_rows = 'euclidean')
```




4. Generate a random binary variable for sex and a categorical variable for age group using distributions and age cutoffs (hint: use the *cut()* function) of your chosing. Add tracking bars to your plot for your generated variables.

```r
# Define covariate for tracking bar
set.seed(9876)
annotationData <- data.frame(row.names = colnames(miRNA),
                             'Status' = c(factor(rbinom(n = 25,size = 1,prob = 0.6),labels = c('Disease
                             'Sex' = c(factor(rbinom(n = 25,size = 1,prob = 0.5),labels = c('Male','Fema
                             'Age' = c(cut(runif(n = 25,min = 30,max = 60),breaks = c(30,40,50,60),
                                           labels = c('30-40','40-50','50-60'))))
# Define color palette
annotationColors <- list(Status = c('Disease X' = 'aquamarine4',
                                    'Healthy' = 'deepskyblue4'),
                         Sex = c('Male' = 'royalblue4',
                                 'Female' = 'magenta4'),
                         Age = c('30-40' = 'gold',
                                 '40-50' = 'orange2',
                                 '50-60' = 'firebrick'))

# Generate heatmap
pheatmap(relAbun[1:20,],
         clustering_distance_cols = 'euclidean',
         clustering_distance_rows = 'euclidean',
         annotation_col = annotationData,
         annotation_colors = annotationColors)
```