ICA Report

By using R programming, the data analysis on genes given was performed. This report will provide an overview of the code, the purpose of which was to produce two heatmaps from given data. Those two heatmaps will give an overview of gene expression profiles, as well as information regarding treatment.

Getting Started

In the beginning setwd command is used to get to the folder with all the documents needed. Several libraries are then downloaded to be used. Pheatmap to create heatmaps and Viridis for better data visualization.

```
setwd("~/ICA")
library(readr)
library(dplyr)
library(pheatmap)
library(viridis)
```

Functions:

Also, three functions are written to make sure:

1. The files are in the directory, in case if any files are missing a message will be shown: "FILES MISSING'.
2. Checking whether gene_annotation has the column named 'LongName' as this column contains actual names of the genes.
3. Checking through the data_all document ensuring all samples are present in the data.

Function 1

Checking all the files are oon the directory:

```
file_check <- function(file_paths) {
  if (!all(sapply(file_paths, file.exists))) {
    stop("FILES MISSING")
  }
}
```

Function 2

Checking needed information is present in the gene_annotation:

```
check_long_name_column <- function(gene_annotation) {
  stopifnot(
    is.data.frame(gene_annotation),
    "LongName" %in% colnames(gene_annotation)
  )
}
```

Function 3

Checking needed information is present in the data_all:

```
check_samples_columns <- function(data_all) {
  stopifnot(
    is.data.frame(data_all),
    "A" %in% colnames(data_all),
    "B" %in% colnames(data_all),
    "C" %in% colnames(data_all),
    "D" %in% colnames(data_all),
    "E" %in% colnames(data_all),
    "F" %in% colnames(data_all),
    "G" %in% colnames(data_all),
    "H" %in% colnames(data_all),
    "I" %in% colnames(data_all),
    "J" %in% colnames(data_all),
    "K" %in% colnames(data_all),
    "L" %in% colnames(data_all)
  )
}
```

Getting the tables

There are 3 documents available in the folder. Here converting those documents into the tables:

```
data_all <- read_csv("data_all.csv")
gene_annotation <- read_csv("gene_annotation.csv")
sample_annotation <- read_csv("sample_annotation.csv")
files <- c("data_all.csv", "gene_annotation.csv", "sample_annotation.csv", "genelist_76.txt")
file_check(files)
check_long_name_column(gene_annotation)
check_samples_columns(data_all)
```

Converting genelist.txt into csv, so can have another table:

```
#get genelist.txt convert it to the csv file
genes <- read.table("genelist_76.txt", header=TRUE)
head(genes)
```

```
##     x
## 1  43
## 2  86
## 3  63
## 4  59
## 5  34
## 6 109
```

DATA FRAME

Obtaining results from data_all:

In order to create two heatmaps need to get the information needed. Firstly, rename one of the columns in the data_all to the gene_name. Data_all contains all of the genes, however for analysis only need those in the genelist provided. Hence, making the x column in genes into a vector as this column has numbers of genes

that have to be analyzed. Creating a new data frame with genes taken from data_all. Creating another data frame, where information from data_all into a log (numbers). Merging both tables, to get one data frame.

```
#rename the column in the data_all, so it's a gene number
colnames(data_all)[1] <- "gene_no"
#as gene_no corresponds to x in genes, make genes x vector
genes_new <- genes$x
#take the data from data_all that corresponds to the genelist given by using %in%
df_new <- data_all[data_all$gene_no %in% genes_new, ]
#convert numbers to log
df_new_log <- log2(df_new[2:13]+1)
#merge tables
df_final <- cbind(df_new[,1, drop=FALSE], df_new_log)
#merging according two data frames as correspond to each other
gene_ann <- merge(df_final, gene_annotation, by.x = "gene_no", by.y = "Gene")
#getting rid of column that contains same information
df_ann <- gene_ann[, -14]
```

Getting new data frame:

After having all the results from data_all, now getting information from gene_annotation. Creating a new data frame where merge according to the columns: gene_no and Gene as correspond to each other. Then adjust the new data frame, having the information of Type and gene names together, making them row names, for heatmaps. Finally, create a separate data frame for treatment, to make this annotation in heatmaps.

```
#moving type and longname in the beginning of the table
df_ann_new <- df_ann %>% relocate("Type", .after = "gene_no")
df_ann_final <- df_ann_new %>% relocate("LongName", .after = "gene_no")
#removing unwanted column
df_ann_final_1 <- df_ann_final[, -1]
#renaming column
colnames(df_ann_final_1)[1] <- "Gene_name"
#getting new column with merging names
df_ann_final_1$d <- paste(df_ann_final_1$Gene_name, df_ann_final_1$Type, sep="_")
#removing those columns (as have one that contains information)
df_ann_final_1 <- df_ann_final_1[, -c(1,2)]
#making this column a rownames for the  heatmaps
row.names(df_ann_final_1) <- df_ann_final_1$d
#removing unwanted column
df_ann_final_1 <- df_ann_final_1[,-13]

#creating new data frame for annotations using sample_annotation column
sample_annotation <- sample_annotation[, -1]
data_annotation <- data.frame(row.names = sample_annotation$SampleName, sample_annotation$TreatmentGrou
colnames(data_annotation)[1] <- "Treatment"
```

Getting annotations colours: Just before creating two heatmaps, now need to adjust colours for annotation so it's different from the heatmap itself. Here the unique function was used to get unique values from data_annotation. Those values are then calculated and the colour is generated based on a number of treatments (which is 4). The annotation list is created so that each treatment (1-4) corresponds to a certain number.

```
#getting unique values
treatments <- unique(data_annotation$Treatment)
#calculating lenght
num_treatments <- length(treatments)
```

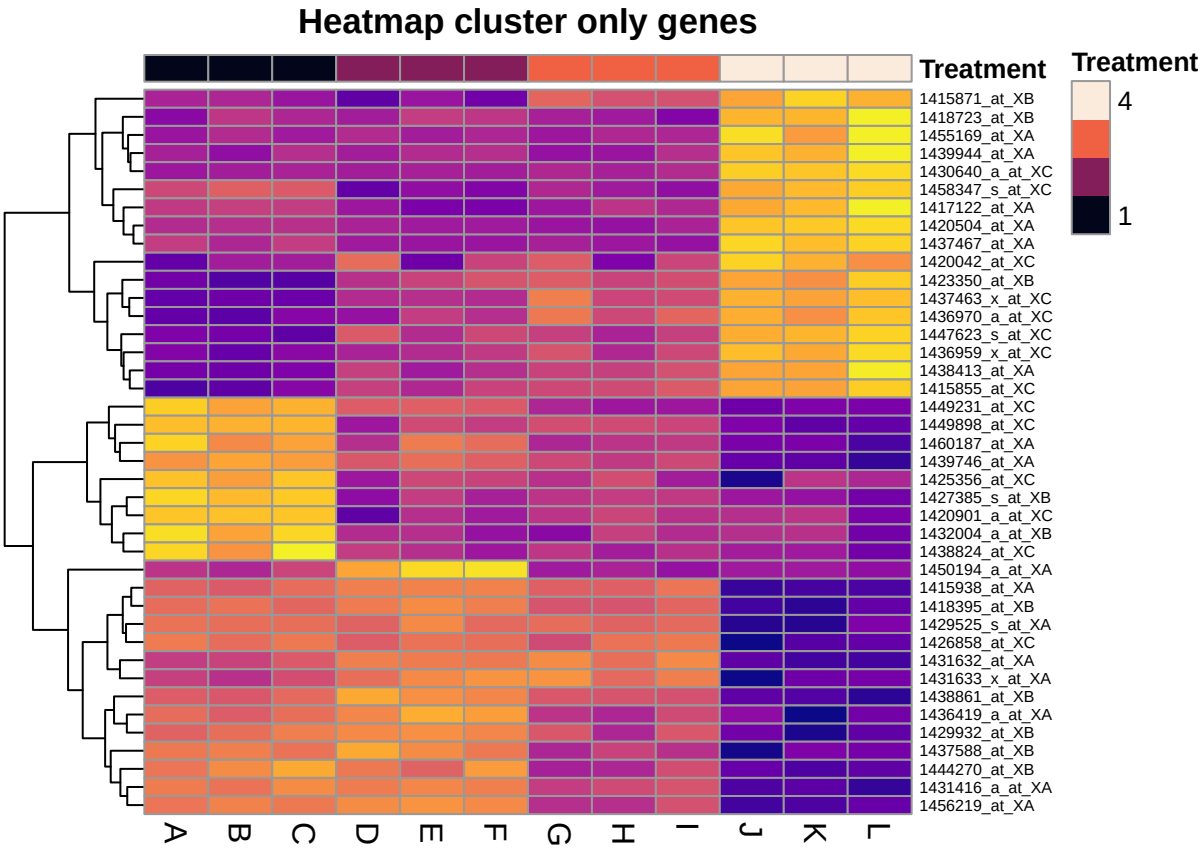
```
#using viridis to assign specific colours
treatment_colors <- viridis(num_treatments, option = "rocket")
#making a list
annotation_colors <- list(Treatment = treatment_colors[data_annotation$Treatment])
```

HEATMAPS:

Heatmaps are a great tool that is usually used to visualize a complex dataset. It allows us to visualize hierarchical clustering. So, the numerical data from the data frame will be transported to the colour data instead, where darker colours will represent high values and lighter colours will represent lower values. Both columns and rows of the data matrix will be arranged according to the results of hierarchical clustering. The data matrix will be displayed as a colour scheme, which will make it easier to find similarities between different elements in the provided dataset.

First heatmap:

```
pheatmap(
  df_ann_final_1,
  scale='row',
  annotation_col= data_annotation,
  annotation_colors = annotation_colors,
  cluster_col=FALSE,
  main = 'Heatmap cluster only genes',
  legend = FALSE,
  fontsize_row = 6,
  fontsize_col = 12,
  color = viridis(100, option = "plasma"),
)
```



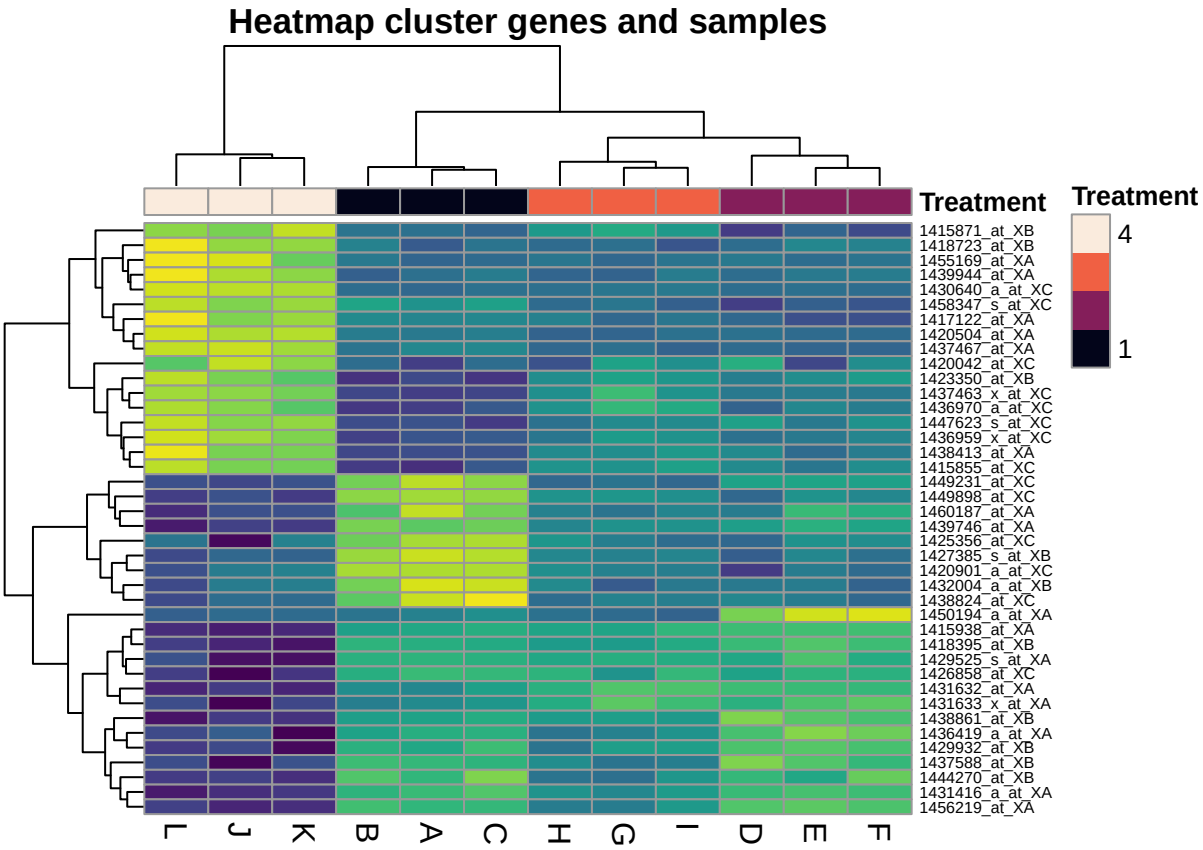
The first heatmap represents only genes being clustered. The genes are now grouped according to their similarity. The colour scale in the heatmap gives an indication of relative expression levels. The darker colour will suggest higher values (higher expression levels). There is also treatment annotation present. There were four different treatments for 12 samples.

It can be seen that for treatment four in samples J, K and L there are very high and low values across genes. For the other treatments and samples, the colour scheme seems to be more mixed, suggesting there are a lot of average values. For treatment one and samples A, B and C it can be seen that there are also a few genes with high or low expression levels, but there are also lots of average values for other genes.

As the rows of the heatmap are grouped based on their similarity, it can be seen in the dendrogram.

Second heatmap:

```
pheatmap(
  df_ann_final_1,
  scale='row',
  annotation_col= data_annotation,
  annotation_colors = annotation_colors,
  cluster_col=TRUE,
  main = 'Heatmap cluster genes and samples',
  legend = FALSE,
  fontsize_row = 6,
  fontsize_col = 12,
  color = viridis(100, option= "viridis")
)
```



In the second heatmap, both genes and samples are clustered. There are now two dendrograms. Still, it can be seen that samples K, J and L in treatment four, have very high or low values, while other samples have a more mixed average range of values. The same can be said regarding treatment 1 and samples A, B and C. As in the first heatmap, there is a group of genes with low expression levels, however, other genes show average results.