ICA Data Processing Report - B236494

Description of the processing pipeline with outputs

Storing Data that is to be processed into Variables

We need to store the files we have been given into data frames so they can be easily manipulated in R.

We're probably more interested in doublings of expression than more subtle changes, so we will flip our values to the log2 scale. We'll need an offset to prevent nasty NA errors in R when log2 is applied to 0 values.

```
# Read the data_all.csv file into a variable
all_data <- read.csv("data_all.csv")

# Remove the first column
all_data <- all_data[, -1]

# Convert the data to log scaled data with an offset of 1
all_data_log2_1 <- log2(all_data + 1)
```

Creating the genelist variable

This chunk of code will be used to define the rows of interest (i.e. the genes) to be plotted onto the heatmap ensuring duplicates are removed.

```
# Read the genelist_68.txt file to be plotted to a variable,
# Remove the 'x' row,
# Rename column to "Gene"
genelist_68 <- read.table("genelist_68.txt")
genelist_68 <- data.frame(genelist_68[-c(1), ])
colnames(genelist_68) <- paste0("Gene")


# Remove the duplicate genes to obtain final list of genes
genelist_final <- unique(genelist_68)
```

Creating the gene annotation variable

This variable will be the 'annotation_row' parameter of the heatmaps we create

```
# Read the gene annotation file into a variable
gene_annotation <- read.csv("gene_annotation.csv")

# Remove the first column
gene_annotation <- gene_annotation[, -1]

# Select the rows we need to plot from the genelist_68.txt file
# Create a vector of the rows we need to plot in heatmap
rows_to_plot <- as.numeric(genelist_final$Gene)
```

```r
# Select the rows needed for gene annotation,
# Rename the rownames to use in pheatmap
# Rename the column name for easy understanding
gene_annotation_to_plot <- data.frame(gene_annotation[rows_to_plot, 2])
rownames(gene_annotation_to_plot) <- gene_annotation[rows_to_plot, 3]
colnames(gene_annotation_to_plot) <- paste0("GeneType")
```

Creating the sample annotation variable

This variable will be the 'annotation_col' parameter of the heatmaps we create

```r
# Sample Annotation csv file to a variable
sample_annot <- read.csv("sample_annotation.csv")

# Remove the extra column,
# Rename rowname to sample names,
# Then remove sample names row
sample_annot <- data.frame(sample_annot[, -1])
rownames(sample_annot) <- sample_annot[, 1]
final_sample_annot <- subset(sample_annot, select = -c(SampleName))
```

Creating the variable to use in pheatmap

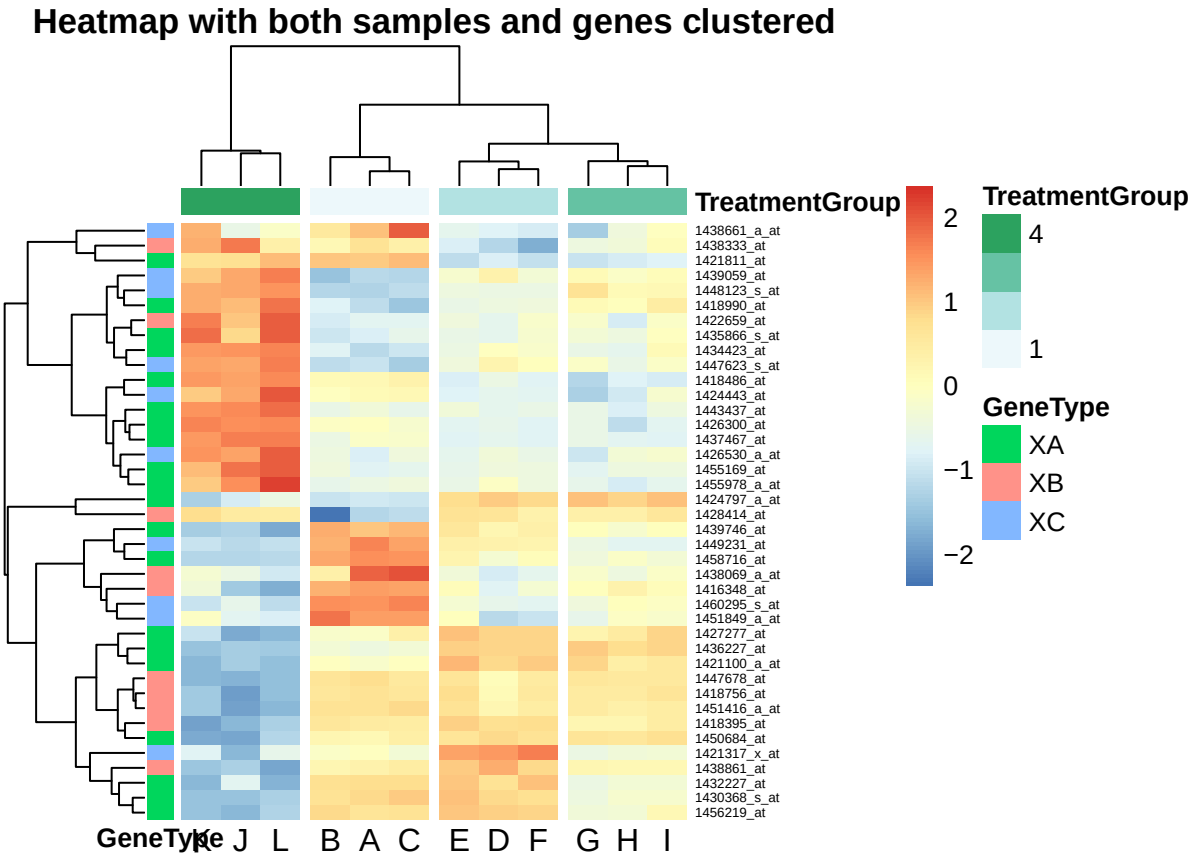
This is the variable which will be the matrix of the log scaled data from earlier for which we will be plotting the heatmaps of.

```r
# Change rownames to the LongName + Type
rownames(all_data_log2_1) <- paste(gene_annotation$LongName, sep = "_")

# Select the rows we need to plot from the genelist_68.txt file
# Create a vector of the rows we need to plot in heatmap
rows_to_plot <- as.numeric(genelist_final$Gene)

# Store the selected rows with the log scaled data
genes_to_plot <- all_data_log2_1[rows_to_plot, ]

# Convert to dataframe to use in pheatmaps
genes_to_plot <- data.frame(genes_to_plot)
```

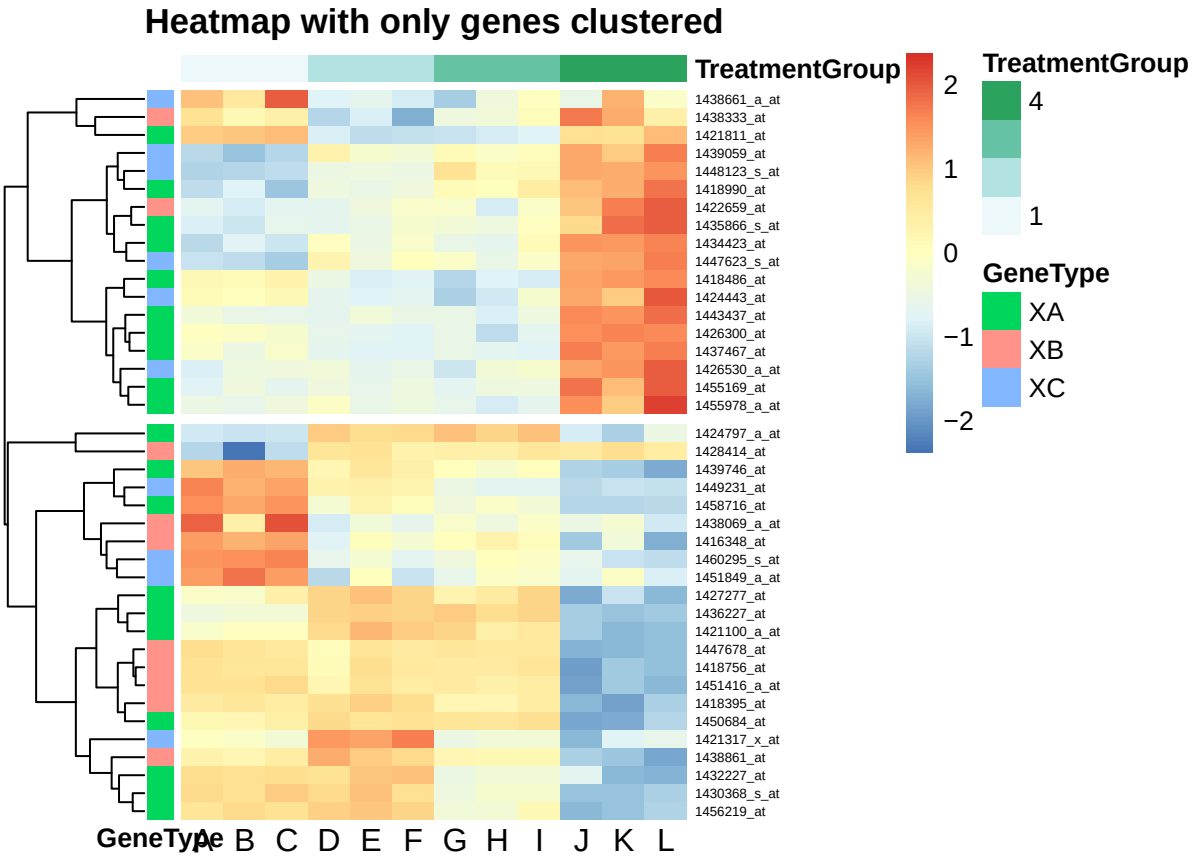
Plotting the Heatmaps

This chunk of code is where it all comes together to give us the pretty heatmaps.

```
library(pheatmap)
# This will give us the heatmap where both samples and genes are clustered
heatmap1 <- pheatmap(
  genes_to_plot,
  main = "Heatmap with both samples and genes clustered",
  scale = "row",
  cluster_col = TRUE,
  fontsize_number = 4,
  number_color = "black",
  display_numbers = FALSE,
  legend_breaks = c(-2, -1, 0, 1, 2),
  legend = TRUE,
  annotation_row = gene_annotation_to_plot,
  annotation_col = final_sample_annot,
  angle_col = 0,
  cutree_cols = 4,
  cutree_rows = 1,
  fontsize_row = 5,
  fontsize_col = 12,
  border_color = FALSE,
  na_col = "black"
)
```



```
library(pheatmap)
# This will give us the heatmap where only genes are clustered.
heatmap2 <- pheatmap(
  genes_to_plot,
  main = "Heatmap with only genes clustered",
  scale = "row",
  cluster_col = FALSE,
  fontsize_number = 4,
  number_color = "black",
  display_numbers = FALSE,
  legend_breaks = c(-2, -1, 0, 1, 2),
  legend = TRUE,
  annotation_row = gene_annotation_to_plot,
  annotation_col = final_sample_annot,
  angle_col = 0,
  cutree_cols = 1,
  cutree_rows = 2,
  fontsize_row = 5,
  fontsize_col = 12,
  border_color = FALSE,
  na_col = "black"
)
```



Interpretation of the heatmap plots

Main Inferences:

1. Heatmap with both samples and genes clustered:

- The clustering of samples here is such that, for each 'TreatmentGroup' (TG), the samples have a strong relation with the samples that come from the same TG.
- There is very high positive trend for expression values for certain genes in TG 4 and also the highest negative trend for the TG's other samples.

2. Heatmap with only genes clustered:

- By clustering with only genes, in the upper block, there is a clear higher expression for some genes but only in TG 4 and a comparatively lesser expression for the same genes in TG 1, TG 2 and TG 3.
- For the genes in the lower block, there is a comparatively higher expression value in TG 1, TG 2 and TG 3 and a clearly lower expression value trend in TG 4.