PRACTICAL BIOINFORMATICS WINTER 2023

ASSIGNMENT 01

BIO221

/ Submitted by: Ankit Kumar
/ 2021015
/ April 03, 2023

1. Dataset for this assignment

Series **GSE67255**	
Title	Effects of Systemically Administered Hydrocortisone on the Human Immunome
Organism	Homo sapiens
Classes	1. dose..250.mg.hydrocortisone 2. dose..50.mg.hydrocortisone
Sample Count	108
Link	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi
GEO2R	https://www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE67255

1. Download any microarray data of interest from GEO with at least 100 samples and two classes.

Loading GEOquery to directly download the archive of the dataset named "**GSE67255_series_matrix.txt.gz**"
"and then extracting I'm RStudio using getGEO() function.

```
library(GEOquery)
gse <- getGEO("GSE67255")
gse <- gse[[1]]
exprs  <- exprs(gse)
pData <- pData(gse)
fData <- fData(gse)
```

2. After performing EDA and preprocessing, list the data attributes, including pdata and fdata.

pData will look like following:

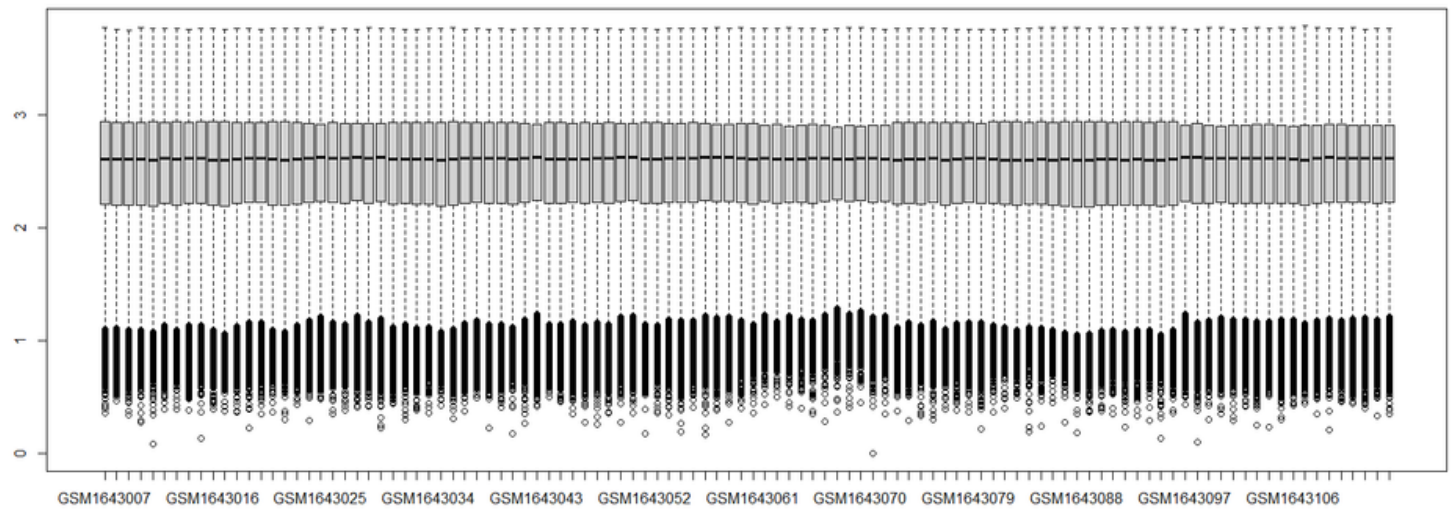
	title	geo_accession	status	submission_date	last_update_date	type	channel_count	source_name_ch1	org
GSM1643007	PBMC_s001_0hr	GSM1643007	Public on Mar 14 2016	Mar 25 2015	Mar 14 2016	RNA	1	Individual s001 at baseline	
GSM1643008	PBMC_s001_1hr	GSM1643008	Public on Mar 14 2016	Mar 25 2015	Mar 14 2016	RNA	1	Individual s001 at 1 hour	
GSM1643009	PBMC_s001_4hr	GSM1643009	Public on Mar 14 2016	Mar 25 2015	Mar 14 2016	RNA	1	Individual s001 at 4 hours	
GSM1643010	PBMC_s001_8hr	GSM1643010	Public on Mar 14 2016	Mar 25 2015	Mar 14 2016	RNA	1	Individual s001 at 8 hours	
GSM1643011	PBMC_s001_12hr	GSM1643011	Public on Mar 14 2016	Mar 25 2015	Mar 14 2016	RNA	1	Individual s001 at 12 hours	
GSM1643012	PBMC_s001_24hr	GSM1643012	Public on Mar 14 2016	Mar 25 2015	Mar 14 2016	RNA	1	Individual s001 at 24 hours	
GSM1643013	PBMC_s002_0hr	GSM1643013	Public on Mar 14 2016	Mar 25 2015	Mar 14 2016	RNA	1	Individual s002 at baseline	
GSM1643014	PBMC_s002_1hr	GSM1643014	Public on Mar 14 2016	Mar 25 2015	Mar 14 2016	RNA	1	Individual s002 at 1 hour	
GSM1643015	PBMC_s002_4hr	GSM1643015	Public on Mar 14 2016	Mar 25 2015	Mar 14 2016	RNA	1	Individual s002 at 4 hours	
GSM1643016	PBMC_s002_8hr	GSM1643016	Public on Mar 14 2016	Mar 25 2015	Mar 14 2016	RNA	1	Individual s002 at 8 hours	
GSM1643017	PBMC_s002_12hr	GSM1643017	Public on Mar 14 2016	Mar 25 2015	Mar 14 2016	RNA	1	Individual s002 at 12 hours	
GSM1643018	PBMC_s002_24hr	GSM1643018	Public on Mar 14 2016	Mar 25 2015	Mar 14 2016	RNA	1	Individual s002 at 24 hours	
GSM1643019	PBMC_s003_0hr	GSM1643019	Public on Mar 14 2016	Mar 25 2015	Mar 14 2016	RNA	1	Individual s003 at baseline	

Showing 1 to 13 of 108 entries, 46 total columns

Accessing basic details of dataset and performing EDA

```r
# the dimensions of the dataset
dim(data)

# the structure of the dataset
str(data)

# summary statistics of the dataset
summary(data)

# missing values in the dataset
sum(is.na(data))

# if negative values then dealing with them
min_value <- min(data, na.rm = TRUE)
data_shifted <- data - min_value + 1

# Step 5: Preprocessing - Log transformation
data_log <- log2(data_shifted)
boxplot(data_log)
```

Boxplot of Data_log:



dimension of data: 33297 x 108

3. State the effects after completing the log transformation of microarray data.

Log transformation converts the raw data values into logarithmic scale, which compresses the dynamic range of the data and brings the values closer together. This is particularly useful for data with high variability, such as gene expression data.

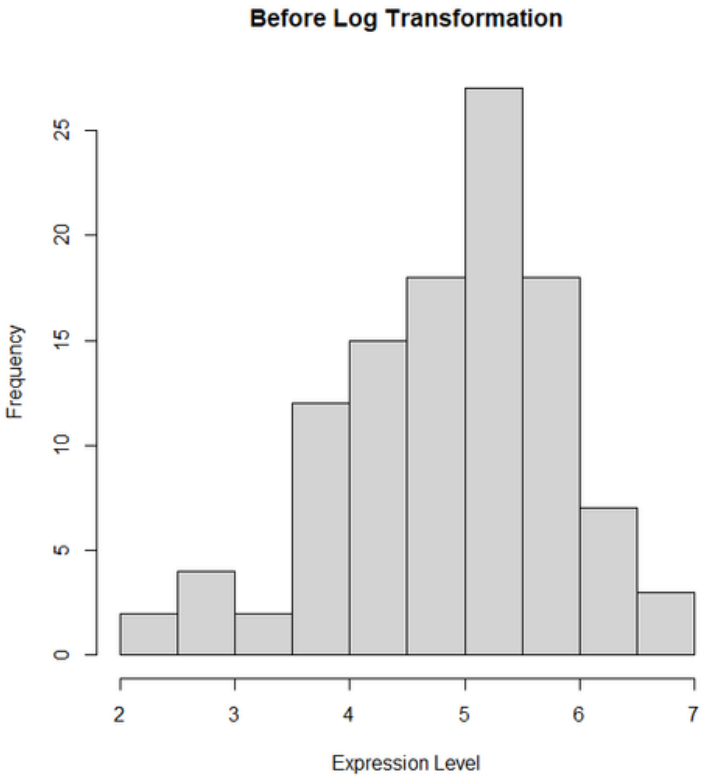
After log transformation, the data distribution becomes more symmetric and closer to a normal distribution. As a result, log-transformed data is more amenable to statistical analyses such as hypothesis testing and clustering, which assume normality and homogeneity of variance.

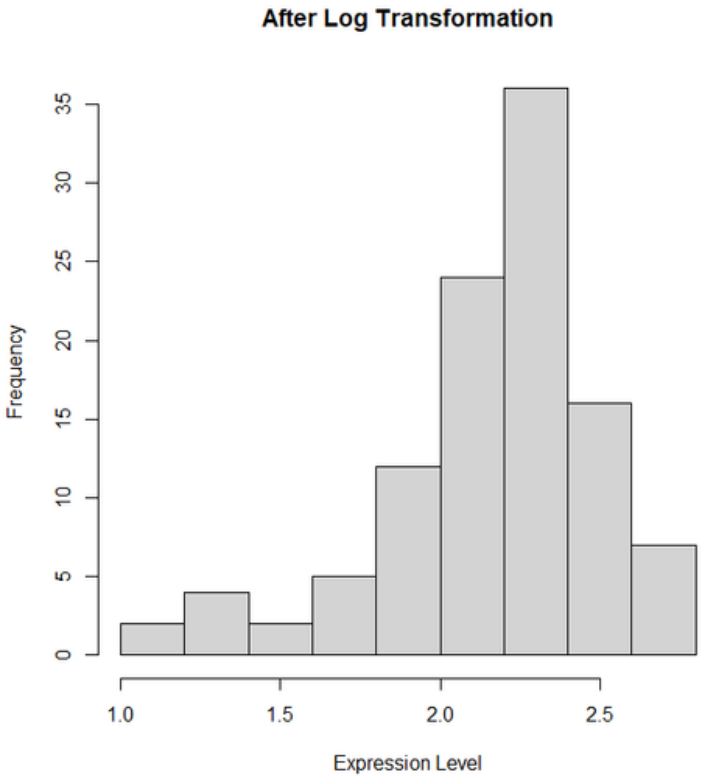
Overall, the log transformation of microarray data helps to reduce noise, improve accuracy, and increase the sensitivity of downstream analyses.

This can be seen in the following histograms which represent variable and distribution of data before and after Log Transform

```
##### 3. State the effects after completing the log transformation of microarray data.
####
par(mfrow = c(1, 2))
hist(data[1, ], main = "Before Log Transformation", xlab = "Expression Level")
hist(data_log[1, ], main = "After Log Transformation", xlab = "Expression Level")
```

Tough there is not much visible difference but more of the redundant values and negative/NaN values are removed to make data more consistent

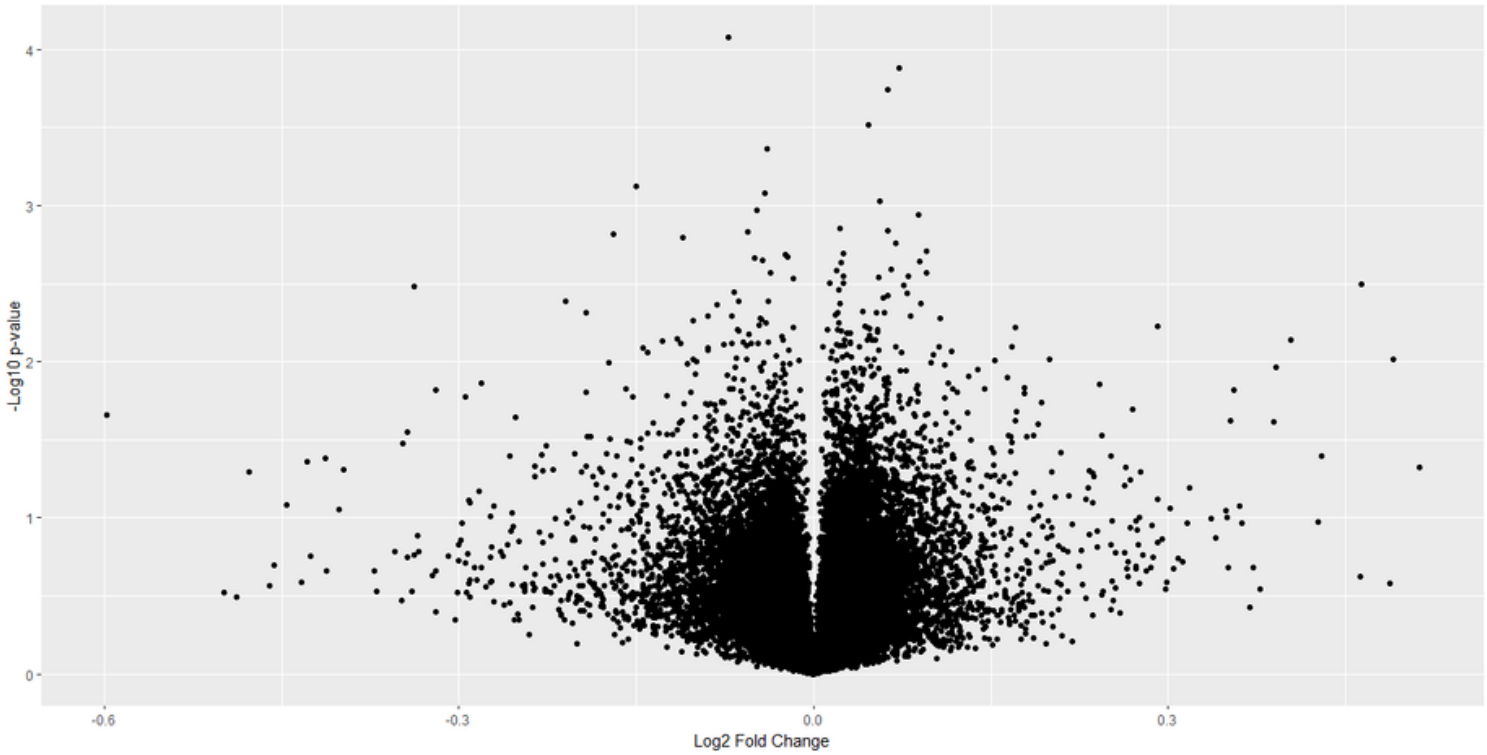




4. Perform differential expression analysis using simple t-test, log fold change, and correct p values using Holm correction. Draw a volcano plot.

```r
#4
# Define classes
class1 <- data[,1:3]
class2 <- data[,4:6]
print(class1)
print(class2)

# Perform t-test
ttest <- apply(data, 1, function(x) t.test(x[1:3], x[4:6])$p.value)

# Calculate log fold change and -log10 p-value
logFC <- apply(data, 1, function(x) log2(mean(x[1:3]) / mean(x[4:6])))
logP <- -log10(ttest)

# Create data frame for plotting
plot_data <- data.frame(logFC = logFC, logP = logP)

# Create volcano plot
ggplot(plot_data, aes(x = logFC, y = logP)) +
  geom_point() +
  xlab("Log2 Fold Change") +
  ylab("-Log10 p-value")
```

performing the analysis for first and last 3 elements

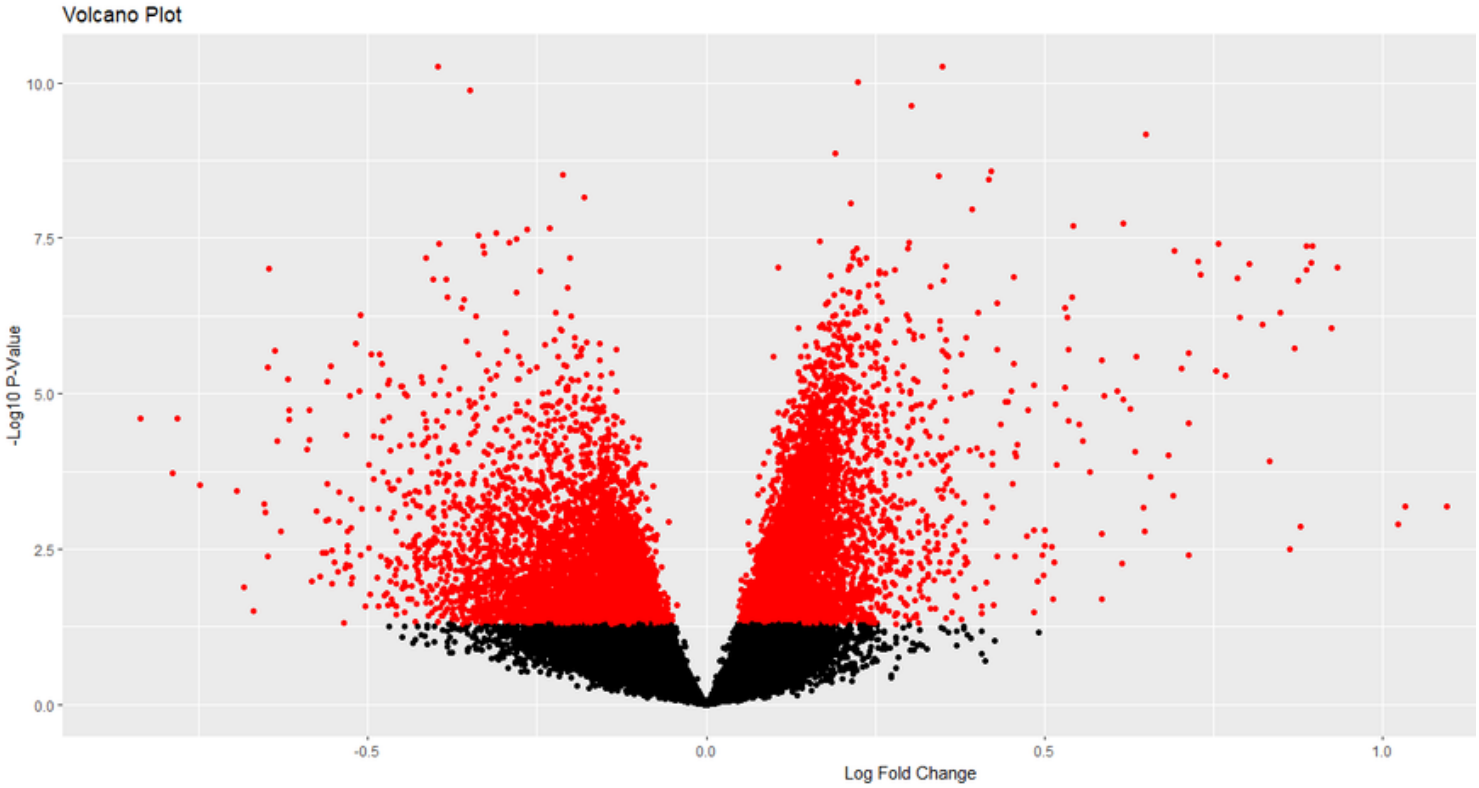


I performed a t-test on gene expression data for two defined classes, calculated the log fold change and -log10 p-value for each gene, and created a volcano plot to visualize the results. The volcano plot shows the log fold change on the x-axis and the negative logarithm of the p-value on the y-axis. The plot_data data frame is created with the log fold change and -log10 p-value for each gene,

5. Perform differential expression analysis using the limma package. Draw a volcano plot.

```r
##### 5 #####
library(limma)
colnames(pData(gse))
#number of classes in ch1.1

pData(gse)$characteristics_ch1.1 <- factor(pData(gse)$characteristics_ch1.1)
Group1 <- levels(pData(gse)$characteristics_ch1.1)[1]
Group2 <- levels(pData(gse)$characteristics_ch1.1)[2]
new_levels <- make.names(levels(pData(gse)$characteristics_ch1.1))

design <- model.matrix(~ 0 + pData(gse)$characteristics_ch1.1)
colnames(design) <- new_levels
cont.matrix <- makeContrasts(dose..250.mg.hydrocortisone - dose..50.mg.hydrocortisone,
levels = design)
gg <- exprs(gse)
fit <- lmFit(gg, design)

#  Calculate the moderated t-statistics:
fit2 <- contrasts.fit(fit, cont.matrix)
fit2 <- eBayes(fit2, trend = TRUE)



#########################
fit <- lmFit(data, design)
fit <- contrasts.fit(fit, cont.matrix)
fit <- eBayes(fit)

# Extract the log-fold change and p-value
results <- topTable(fit, adjust.method = "holm", sort.by = "none", number = Inf)
logFC <- results$logFC
PValue <- results$P.Value

## Create a volcano plot
plot_data <- data.frame(logFC, -log10(PValue))
ggplot(plot_data, aes(x = logFC, y = -log10(PValue))) +
  geom_point(aes(color = ifelse(PValue < 0.05, "red", "black"))) +
  scale_color_identity() +
  ggtitle("Volcano Plot") +
  xlab("Log Fold Change") +
 ylab("-Log10 P-Value")
```

Results after differential expression analysis using limma



- I used the limma package to perform differential expression analysis. First, I defined the experimental design using model.matrix(), create the contrast matrix using makeContrasts(), fit a linear model using lmFit(), and performed empirical Bayes moderation using eBayes().
- I created a volcano plot using ggplot() and geom_point(). The x-axis represents the log2 fold change, while the y-axis represents the negative log10-transformed p-value. The red dots represent the genes that are significantly differentially expressed between the two groups with a corrected p-value less than 0.05. These genes are considered statistically significant and are often of greater interest in downstream analyses. The black dots, on the other hand, represent genes that are not significantly differentially expressed. The position of a dot in the plot can indicate the magnitude of the gene expression difference and the statistical significance of the difference. Typically, the farther a dot is from the origin, the greater the magnitude of the difference in expression levels, and the higher it is on the y-axis, the more significant the difference.

6. Choose a significant cutoff based on log(FC) and p-values and justify why you chose those values as the cutoff.

The cutoff for logFC is set to 1, which corresponds to a two-fold change in expression between the two classes. For p-value, I used cutoff as 0.05, which corresponds to a false discovery rate of 5%.

Why?
This cutoff is biologically relevant because it ensures that only genes with significant changes in expression are considered. A fold change of 2 is also a commonly used threshold in the our PB books.

7) Perform Enrichment analysis using the set of genes that you have obtained using the Gene set enrichment analysis method.

```r
library(dplyr)
library(clusterProfiler)
library(org.Hs.eg.db)
# 7

results = topTable(fit, adjust.method = "holm", sort.by = "none", number = Inf)

results = rownames_to_column(results, "Gene_ID")

de_genes = results %>%
  filter(abs(logFC) >= 1, PValue <= 0.05) %>%
  dplyr::select(Gene_ID)
de_genes_entrez = unique(results$Gene_ID)

de_genes_entrez = as.list(factor(de_genes_entrez))
class(de_genes_entrez)

de_genes_entrez = unlist(de_genes_entrez)
de_genes_entrez = sort(de_genes_entrez)
de_genes_entrez = as.list(factor(de_genes_entrez))
go_enrichment = enrichGO(de_genes_entrez, OrgDb="org.Hs.eg.db",
keyType="ENTREZID", ont="BP", pvalueCutoff=0.05, qvalueCutoff=0.1)
```

I performed gene ontology (GO) enrichment analysis using the enrichGO function from the **clusterProfiler** package.

I extracted the expression data from the GEO dataset and stored it in the 'data' variable. Then divided the data into two groups, 'class1' and 'class2'and performed a t-test on each gene to compare the two groups and calculates the log fold change and -log10 p-value for each gene.

after performing steps like in step 5 and 6, I Converted the gene symbols to Entrez IDs and performed Gene Ontology (GO) enrichment analysis for biological processes using the enrichGO function.

Results:

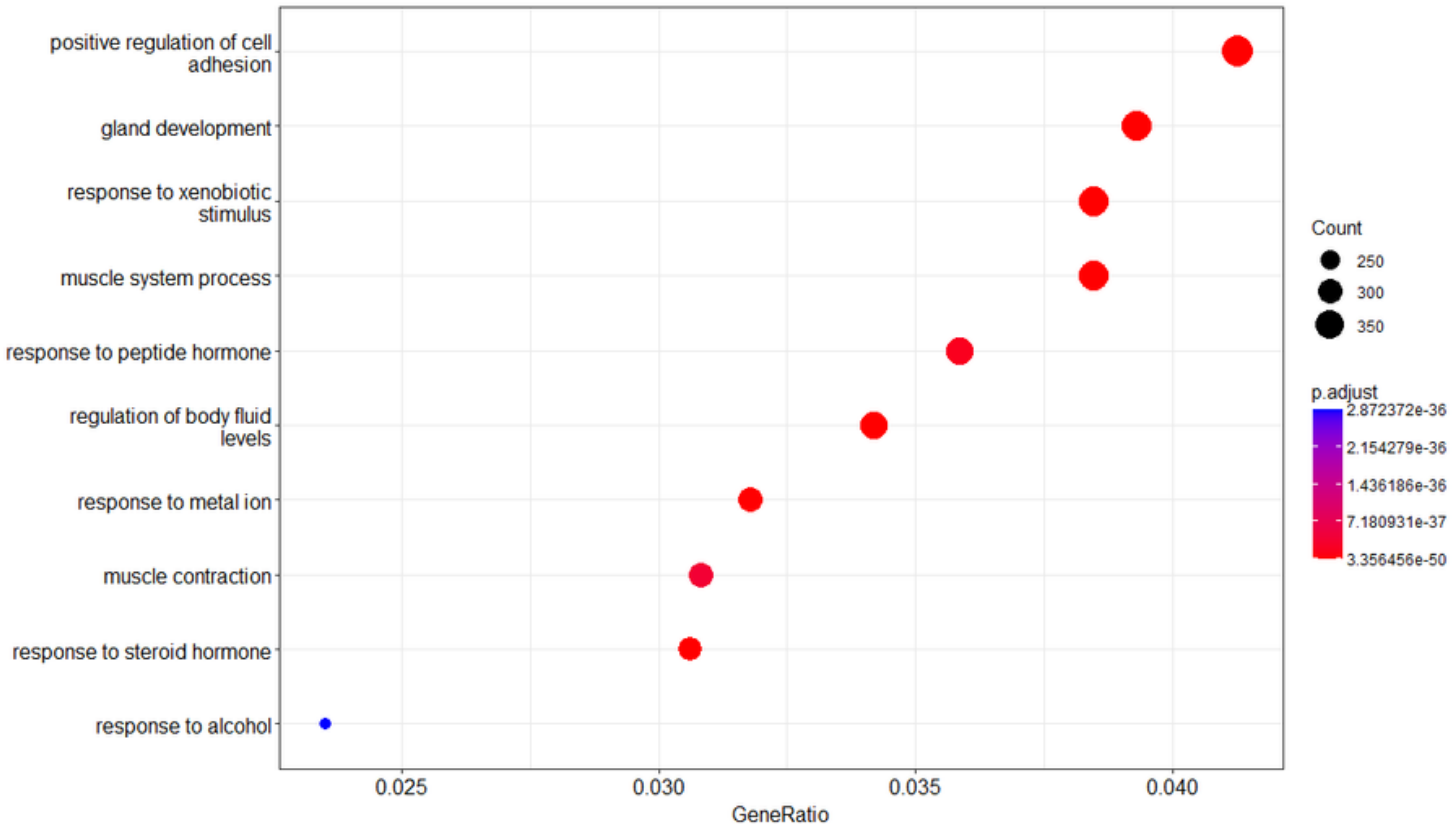


fig : dotplot() of the variable go_enrichment with top 10 pathways

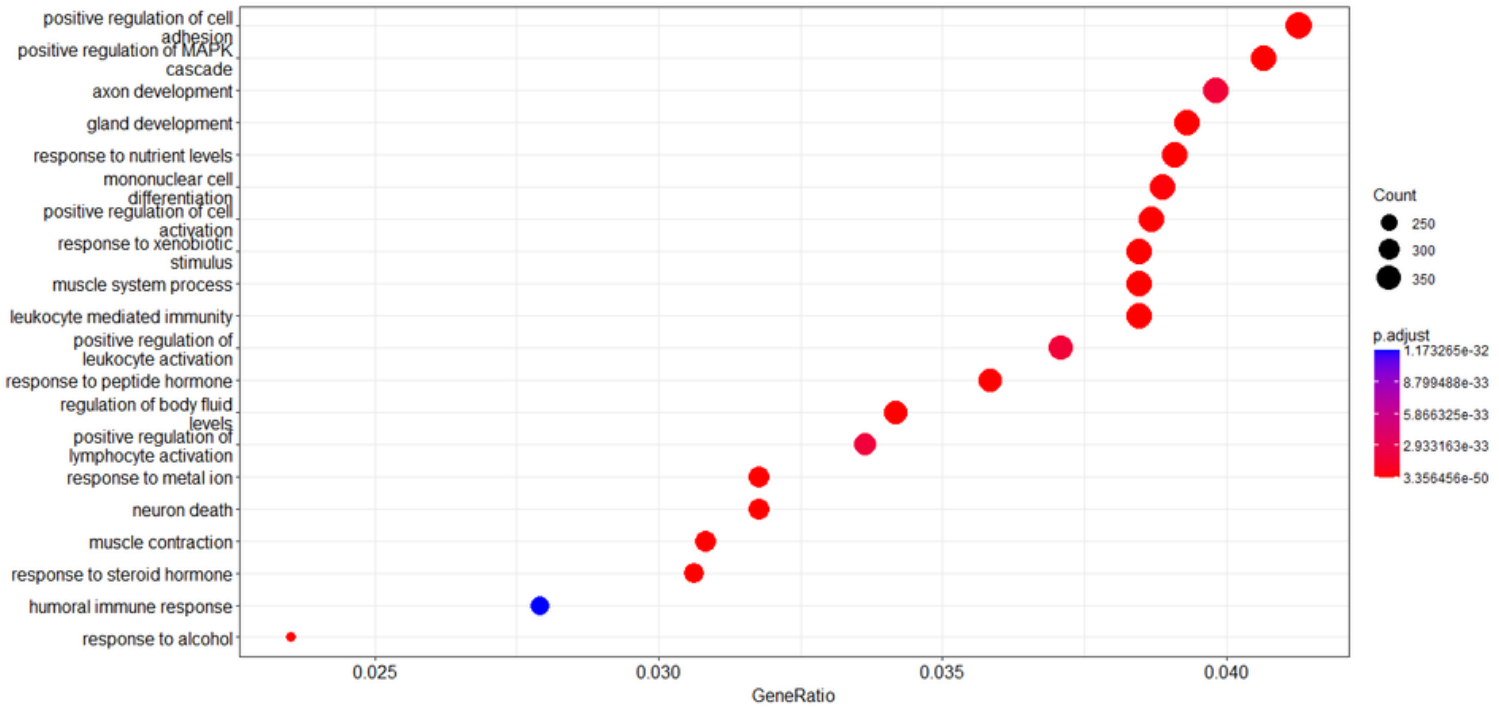


fig : dotplot() of the variable go_enrichment with top 20 pathways

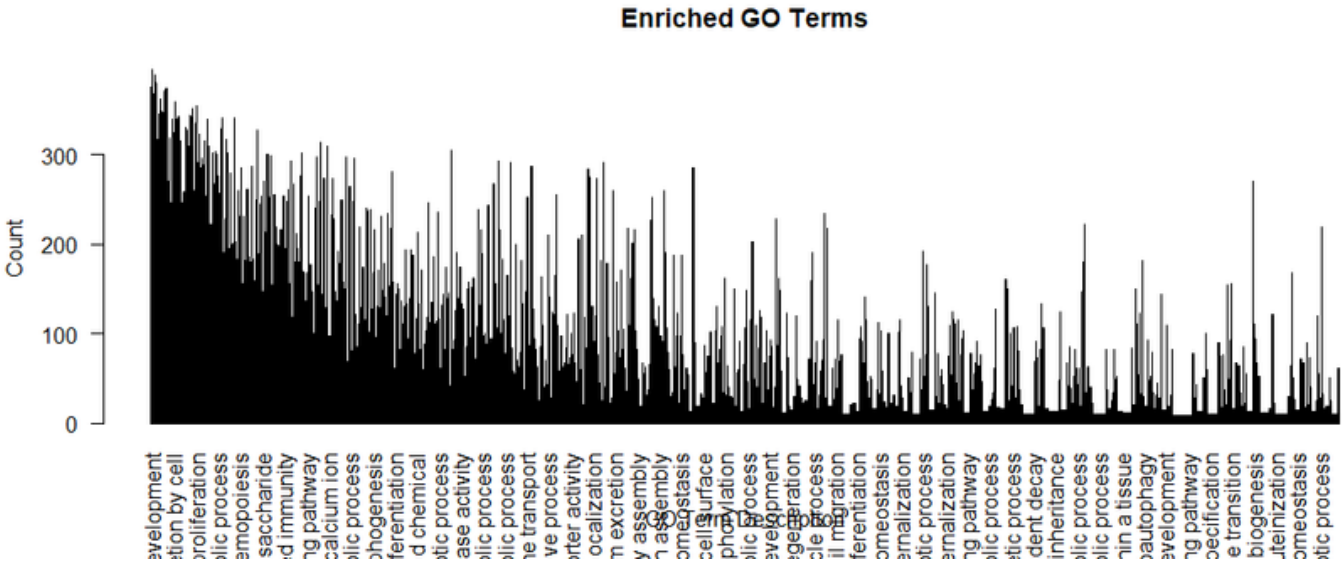


fig : barplot() of the variable go_enrichment with all pathways

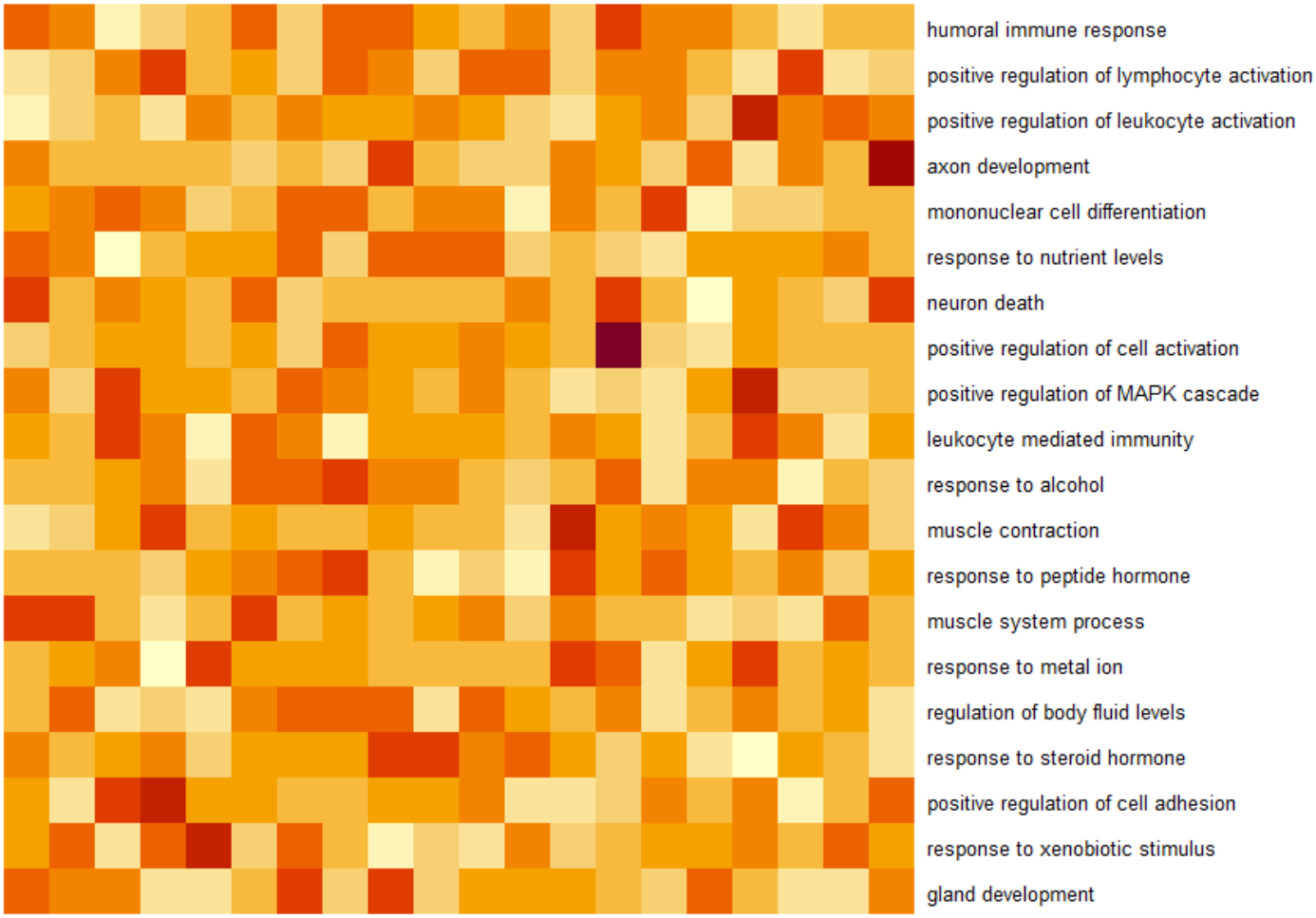


fig : heatmap() of the variable go_enrichment with 20 pathways

9) Observe and analyze the pathways which you obtained.

```
# View top enriched terms
head(go_enrichment)
pathways = head(go_enrichment@result, n=20)
pathways$Description
```

- To analyze the pathways obtained from Gene Ontology (GO) enrichment analysis, we first need to understand what these pathways represent. GO is a widely used ontology for describing the biological processes, molecular functions, and cellular components of genes and gene products.

- After performing gene ontology enrichment analysis, we obtained the top 20 enriched pathways. The pathways are ranked based on the p-value obtained from the statistical analysis. The pathways and their descriptions are as follows:

1. Gland development - refers to the process by which glands, specialized structures that produce and secrete substances, are formed and mature.
2. Response to xenobiotic stimulus - refers to the organism's response to exposure to a foreign substance, such as a drug, pollutant, or toxin.
3. Positive regulation of cell adhesion - refers to the processes by which cells are encouraged to stick together or to other surfaces, and is important for many cellular processes, including tissue formation and wound healing.
4. Response to steroid hormone - refers to the cellular response to steroid hormones, which play a role in a variety of physiological processes such as development, reproduction, and metabolism.
5. Regulation of body fluid levels - refers to the maintenance of appropriate levels of fluids in the body, which is important for maintaining homeostasis.
6. Response to metal ion - refers to the cellular response to metal ions, which can have toxic effects on the body at high concentrations.
7. Muscle system process - refers to the processes involved in the development, function, and regulation of muscle tissue, including muscle contraction and relaxation.
8. Response to peptide hormone - refers to the cellular response to peptide hormones, which play a role in a variety of physiological processes such as growth and development, metabolism, and stress response.

and so on...