Stage 1 Task

```
library(gplots)
```

```
## Warning: package 'gplots' was built under R version 4.3.3
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(RColorBrewer)
```

```
# load dataset
url <- "https://raw.githubusercontent.com/HackBio-Internship/public_datasets/main/Cancer2024/glioblastom
gene_data <- read_csv(url)
```

```
## New names:
## Rows: 582 Columns: 11
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (1): ...1 dbl (10): TCGA-19-4065-02A-11R-2005-01, TCGA-19-0957-02A-11R-2005-01,
## TCGA-0...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```r
head(gene_data)
```

```
## # A tibble: 6 x 11
##   ...1      TCGA-19-4065-02A-11R~1 TCGA-19-0957-02A-11R~2 TCGA-06-0152-02A-01R~3
##   <chr>                      <dbl>                  <dbl>                  <dbl>
## 1 ENSG0000~                    763                   4526                    683
## 2 ENSG0000~                   2759                   8384                   2763
## 3 ENSG0000~                    939                    850                   1250
## 4 ENSG0000~                    231                   1266                    817
## 5 ENSG0000~                    540                    512                    655
## 6 ENSG0000~                   1282                    720                   1694
## # i abbreviated names: 1: 'TCGA-19-4065-02A-11R-2005-01',
## #   2: 'TCGA-19-0957-02A-11R-2005-01', 3: 'TCGA-06-0152-02A-01R-2005-01'
## # i 7 more variables: 'TCGA-14-1402-02A-01R-2005-01' <dbl>,
## #   'TCGA-14-0736-02A-01R-2005-01' <dbl>, 'TCGA-06-5410-01A-01R-1849-01' <dbl>,
## #   'TCGA-19-5960-01A-11R-1850-01' <dbl>, 'TCGA-14-0781-01B-01R-1849-01' <dbl>,
## #   'TCGA-02-2483-01A-01R-1849-01' <dbl>, 'TCGA-06-2570-01A-01R-1849-01' <dbl>
```

```r
# set genes as rownames
gene_data <- gene_data %>% column_to_rownames(var = '...1')
```

```r
names(gene_data)
```

```
##  [1] "TCGA-19-4065-02A-11R-2005-01" "TCGA-19-0957-02A-11R-2005-01"
##  [3] "TCGA-06-0152-02A-01R-2005-01" "TCGA-14-1402-02A-01R-2005-01"
##  [5] "TCGA-14-0736-02A-01R-2005-01" "TCGA-06-5410-01A-01R-1849-01"
##  [7] "TCGA-19-5960-01A-11R-1850-01" "TCGA-14-0781-01B-01R-1849-01"
##  [9] "TCGA-02-2483-01A-01R-1849-01" "TCGA-06-2570-01A-01R-1849-01"
```

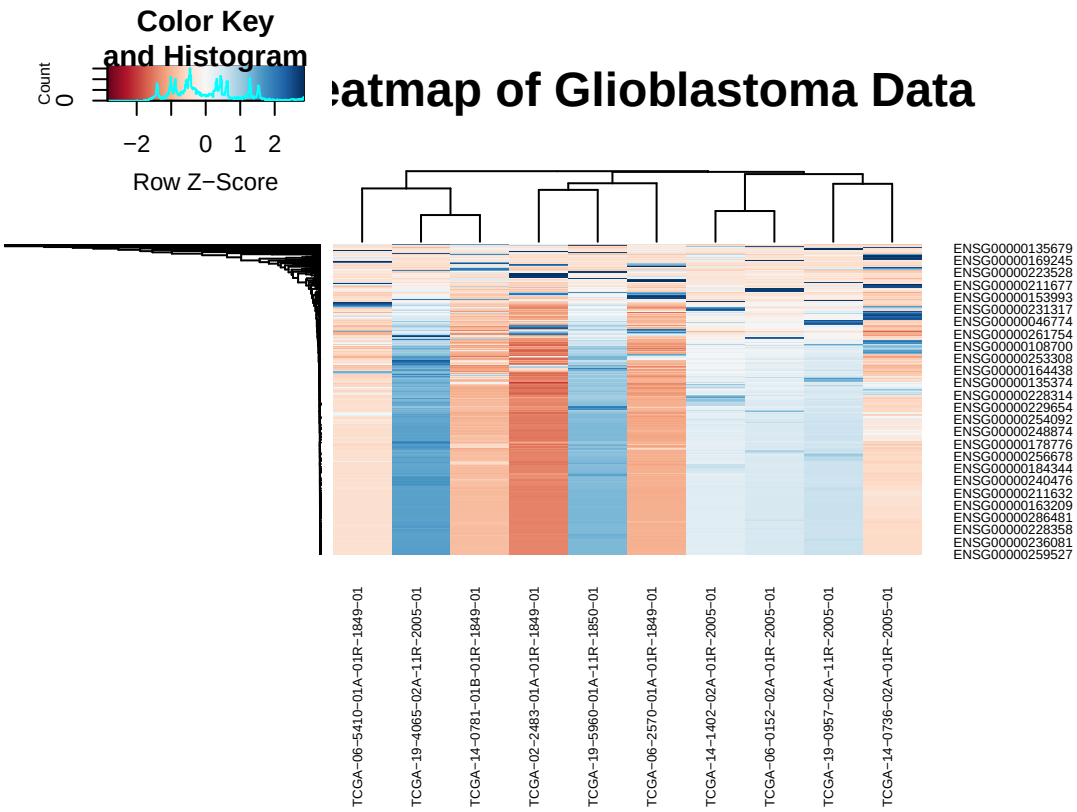
Visualise expression levels using heatmap and showing clusters

Create heatmap

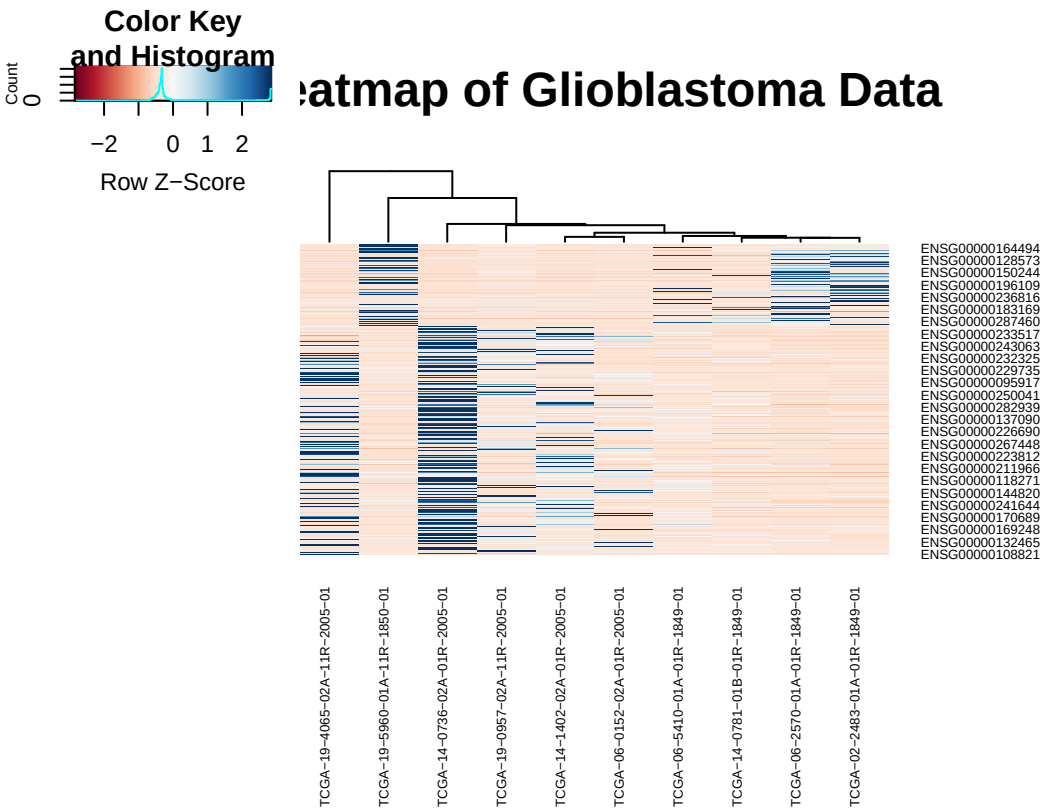
```r
# Scale the data
scaled_data <- scale(gene_data)

# Define color palette for heatmap
diverging_palette <- colorRampPalette(brewer.pal(11, "RdBu"))(256)
```

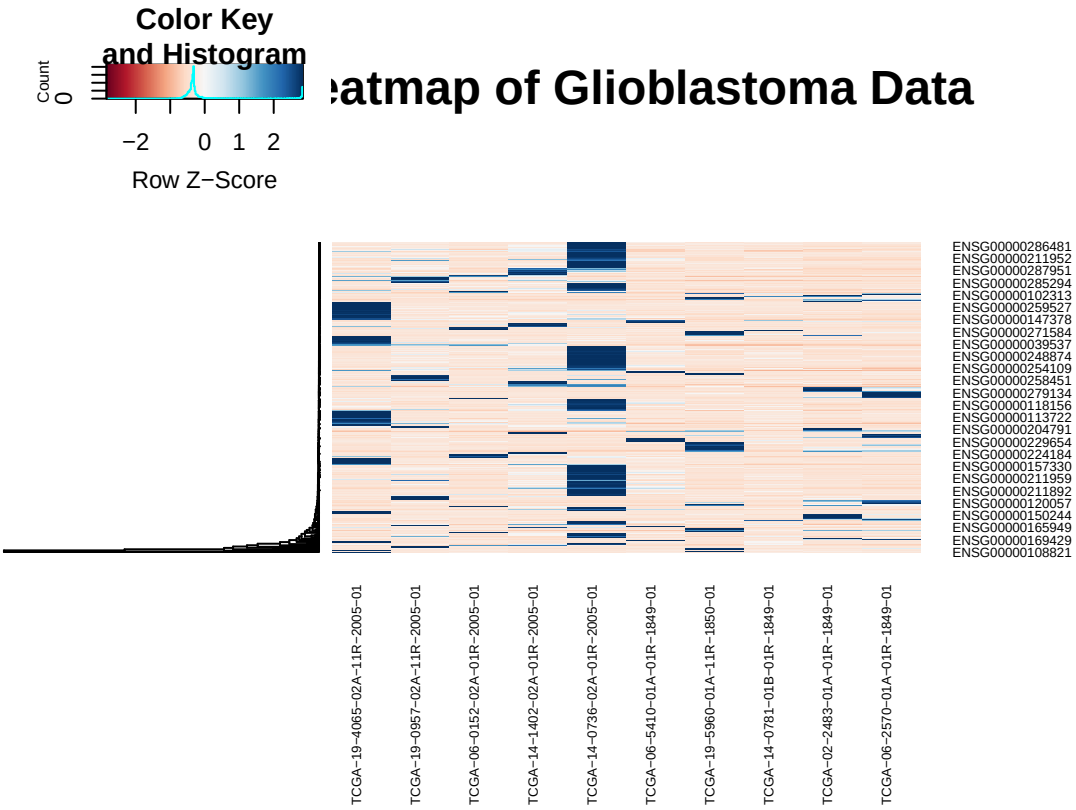
```r
# Create heatmap both samples and genes
heatmap.2(as.matrix(scaled_data),
          col = diverging_palette,
          trace = "none",
          dendrogram = "both",  # Cluster both rows and columns
          scale = "row",        # Scale data by row
          margins = c(10, 10),  # Margins around the plot
          cexRow = 0.5,         # Size of row labels
          cexCol = 0.5,         # Size of column labels
          main = "Heatmap of Glioblastoma Data")
```



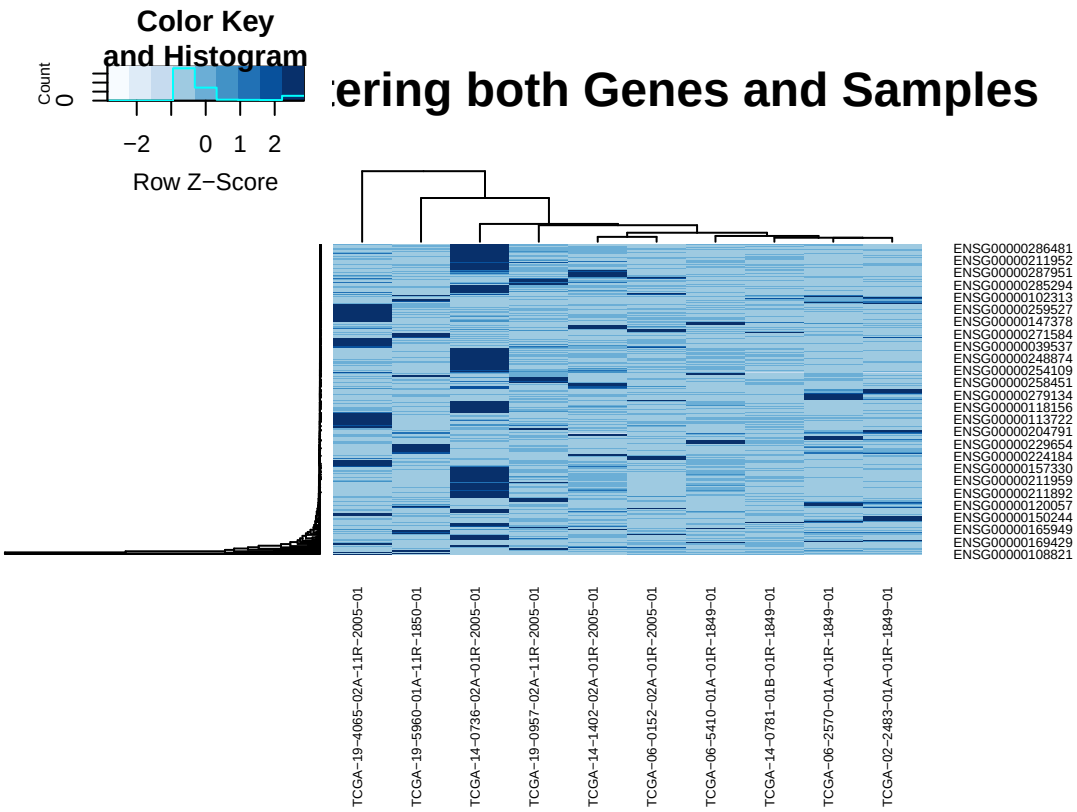

```r
# Create heatmap across samples only
heatmap.2(as.matrix(gene_data),
          col = diverging_palette,
          Colv=TRUE,
          Rowv=FALSE,
          trace = "none",
          dendrogram = "col",  # Cluster columns
          scale = "row",         # Scale data by row
          margins = c(10, 10),  # Margins around the plot
          cexRow = 0.5,          # Size of row labels
          cexCol = 0.5,          # Size of column labels
          main = "Heatmap of Glioblastoma Data")
```




```
# Create heatmap across genes only
heatmap.2(as.matrix(gene_data),
          col = diverging_palette,
          Colv=FALSE,
          Rowv=TRUE,
          trace = "none",
          dendrogram = "row",  # Cluster rows
          scale = "row",         # Scale data by row
          margins = c(10, 10),  # Margins around the plot
          cexRow = 0.5,          # Size of row labels
          cexCol = 0.5,          # Size of column labels
          main = "Heatmap of Glioblastoma Data")
```




```
# cluster columns and rows (samples and genes) - sequential
heatmap.2(as.matrix(gene_data),
          Colv = T, Rowv = T,
          dendrogram = 'both',
          col = RColorBrewer::brewer.pal(9, 'Blues'),
          trace='none', scale='row',
          margins = c(10, 10),  # Margins around the plot
          cexRow = 0.5,         # Size of row labels
          cexCol = 0.5,         # Size of column labels
          main="Clustering both Genes and Samples")
```




```
# creating metadata

# all samples with 2005 grouped in same group
metadata <- data.frame(
  row.names = colnames(gene_data),
  groups=ifelse(grepl('2005', names(gene_data)), 'group1', 'group2' )
  )

metadata$groups <- relevel(factor(metadata$groups), ref='group2')

metadata
```

```
##                                groups
## TCGA-19-4065-02A-11R-2005-01 group1
## TCGA-19-0957-02A-11R-2005-01 group1
## TCGA-06-0152-02A-01R-2005-01 group1
## TCGA-14-1402-02A-01R-2005-01 group1
## TCGA-14-0736-02A-01R-2005-01 group1
## TCGA-06-5410-01A-01R-1849-01 group2
## TCGA-19-5960-01A-11R-1850-01 group2
## TCGA-14-0781-01B-01R-1849-01 group2
## TCGA-02-2483-01A-01R-1849-01 group2
## TCGA-06-2570-01A-01R-1849-01 group2
```

```
group1 <- gene_data[, which(metadata$groups == 'group1')]
group2 <- gene_data[, which(metadata$groups == 'group2')]
```

```r
# function that run a t-test
run_differential_genes <- function(row){
  t.test(row[names(group1)], row[names(group2)])$p.value
}
```

```r
# calculate pvalues, log fold change and mean difference
pvalues <- apply(cbind(group1, group2), 1, run_differential_genes)
logfoldchange <- log2(rowMeans(group1)) - log2(rowMeans(group2))
mean_diff <- rowMeans(group1) - rowMeans(group2)
```

```r
results <- as.data.frame(cbind(mean_diff, logfoldchange, pvalues))
```

```r
# selecting upregulated and downregulated genes
upregulated <- results %>%
  filter(pvalues < 0.05, logfoldchange >= 1.5)

downregulated <- results %>%
  filter(pvalues < 0.05, logfoldchange <= 1.5)
```

```r
upregulated
```

```
##                 mean_diff logfoldchange    pvalues
## ENSG00000243955       9.0      3.614710 0.02982752
## ENSG00000095917      14.0      3.459432 0.03200868
## ENSG00000231107      17.4      2.544321 0.03591137
## ENSG00000254092      38.6      2.703607 0.01661109
## ENSG00000172236      49.8      3.095157 0.02709551
## ENSG00000197253      55.0      2.946229 0.02529988
## ENSG00000172116      68.2      2.803735 0.04879771
## ENSG00000162598      98.8      2.753644 0.01702327
## ENSG00000256193      94.6      2.743902 0.04132131
## ENSG00000160183     123.8      2.602592 0.04541860
```

```r
downregulated
```

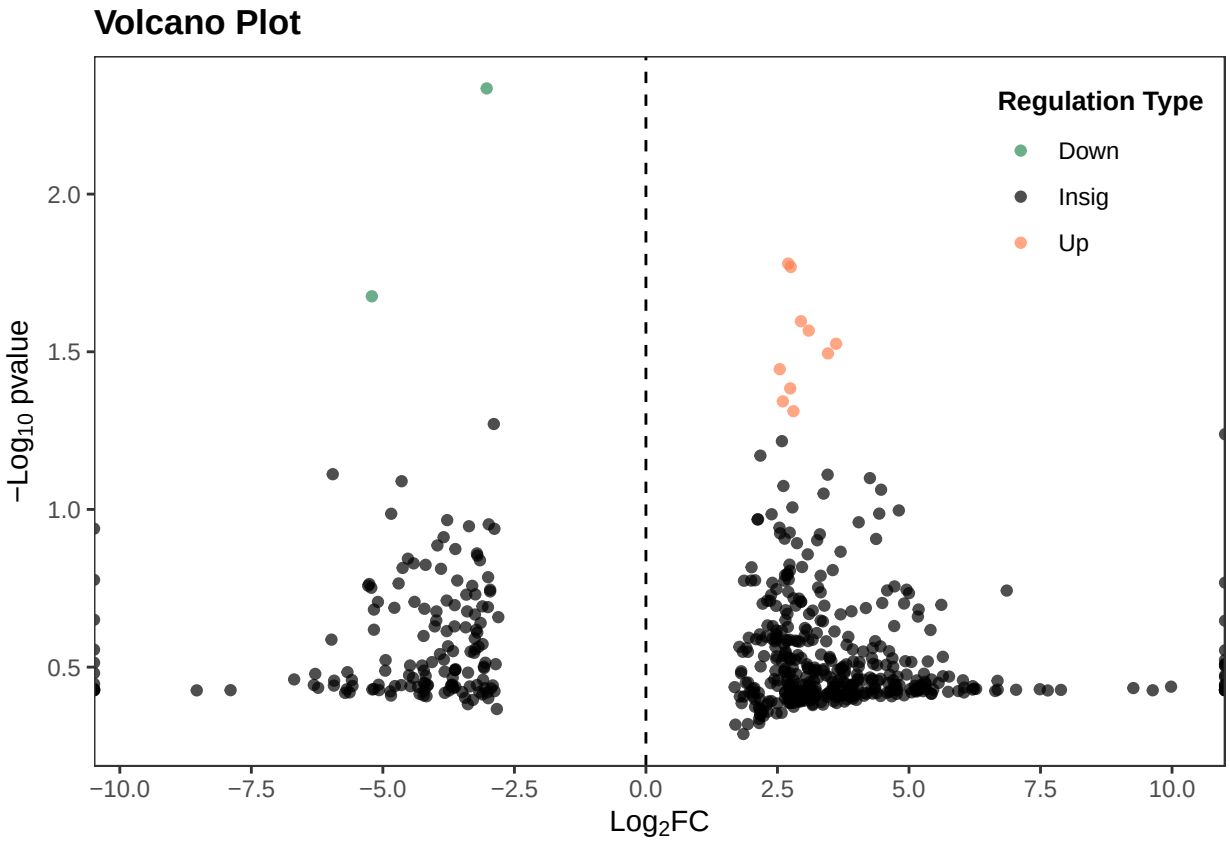
```
##                 mean_diff logfoldchange     pvalues
## ENSG00000241945    -227.4     -3.026967 0.004622371
## ENSG00000279104      -7.2     -5.209453 0.021089610
```

```r
downregulated %>% write.csv('../Data/downregulated_genes.csv', row.names = T)
upregulated %>% write.csv('../Data/upregulated_genes.csv', row.names = T)
```

Visualisation

```r
# volcano plots
results %>%
  mutate(neg_log_pval = -log10(pvalues)) %>%
  mutate(group = case_when(
    rownames(gene_data) %in% rownames(upregulated) ~ 'up',
```

```
  rownames(gene_data) %in% rownames(downregulated) ~ 'down',
  .default = 'insig'
)) %>%
ggplot(aes(logfoldchange, neg_log_pval, color=group)) +
geom_point(alpha=0.7) +
theme_bw() +
theme(legend.key = element_blank(),
      legend.position = 'inside',
      legend.title = element_text(face='bold', size=10),
      plot.title = element_text(face='bold'),
      panel.grid = element_blank(),
      legend.position.inside = c(0.89, 0.83)) +
geom_vline(xintercept = 0, linetype='dashed') +
scale_color_manual(values=c('seagreen', 'black', 'coral'),
                   labels=c('Down', 'Insig', 'Up')) +
labs(title='Volcano Plot', x=expression('Log'[2]*'FC'),
     y=expression('-Log'[10]*' pvalue'), color='Regulation Type') +
scale_x_continuous(breaks=seq(-12.5,12.5,2.5), expand = c(0.1,0.1,0.05,0.1))
```



Visualisation of functional analysis

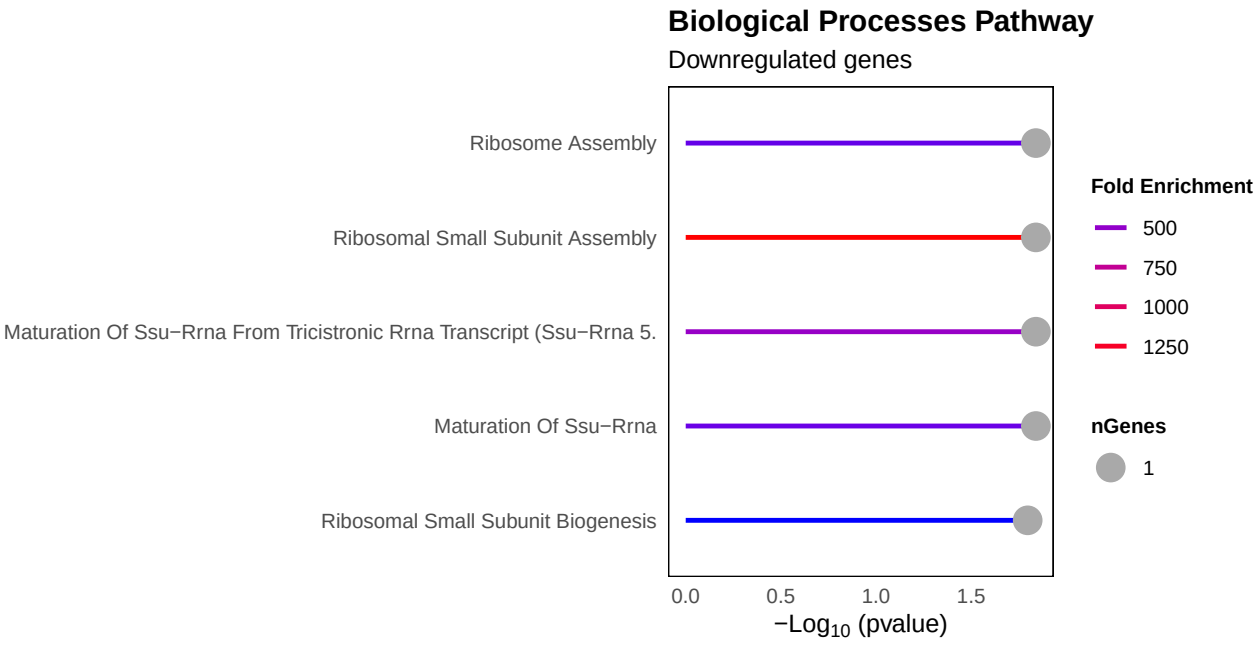
Biological processes

Here, we will visualise the biological process result obtained from functional annotation

```
BP_upregulated <- read_csv(
  '../Data/Upregulated enrichment biological processes.csv',
  show_col_types = FALSE)

BP_downregulated <- read_csv(
  '../Data/Downregulated enrichment biological processes.csv',
  show_col_types = FALSE)
```

```
BP_downregulated %>%
  # filter by statistically significant result
  filter(`Enrichment FDR` < 0.05) %>%
  slice_min(`Enrichment FDR`, n=5) %>%
  mutate(`Enrichment FDR` = -log10(`Enrichment FDR`)) %>%
  mutate(Pathway = str_to_title(str_trim(str_remove(Pathway, 'GO:\\d+\\s')))) %>%
  ggplot(aes(x=`Enrichment FDR`, y=fct_reorder(Pathway, `Enrichment FDR`))) +
    geom_segment(aes(x=0, xend=`Enrichment FDR`, y=Pathway, yend=Pathway,
                     color=`Fold Enrichment`), linewidth=1) +
    geom_point(aes(size=nGenes), color='darkgray') +
  theme_minimal() +
  theme(panel.grid=element_blank(),
        plot.title=element_text(face='bold'),
        axis.text.y = element_text(size=9),
        legend.title = element_text(size=9, face='bold'),
        panel.background = element_rect(fill='white')) +
  scale_color_gradient(low='blue', high='red') +
  scale_size(range = c(4,6)) +
  guides(color=guide_legend(title='Fold Enrichment'),scale='none') +
  labs(title='Biological Processes Pathway',
       subtitle = 'Downregulated genes', y='', x=expression("-Log"[10]*" (pvalue)"))
```



```r
BP_upregulated %>%
  # filter by statistically significant result
  filter(`Enrichment FDR` < 0.05) %>%
  slice_min(`Enrichment FDR`, n=5) %>%
  mutate(`Enrichment FDR` = -log10(`Enrichment FDR`)) %>%
  mutate(Pathway = str_to_title(str_trim(str_remove(Pathway, 'GO:\\d+\\s')))) %>%
  ggplot(aes(x=`Enrichment FDR`, y=fct_reorder(Pathway, `Enrichment FDR`))) +
    geom_segment(aes(x=0, xend=`Enrichment FDR`, y=Pathway, yend=Pathway,
                     color=`Fold Enrichment`), linewidth =1) +
    geom_point(aes(size=nGenes), color='darkgray') +
  theme_minimal() +
  theme(panel.grid=element_blank(),
        plot.title=element_text(face='bold'),
        axis.text.y = element_text(size=9),
        legend.title = element_text(size=9, face='bold'),
        panel.background = element_rect(fill='white')) +
  scale_color_gradient(low='blue', high='red') +
  scale_size(range = c(4,5)) +
  guides(color=guide_legend(title='Fold Enrichment'),scale='none') +
  labs(title='Biological Processes Pathway',
       subtitle = 'Upregulated genes', y='', x=expression("-Log"[10]*" (pvalue)"))
```