Task 3: RNA-seq Differential Gene Expression & RNA-seq Functional Profiling

Simon Safron [Mn: 11923407], Alexander Veith [Mn: 12122739], Miriam Überbacher [Mn:01627576]

2025-12-11

Contents

1 Introduction **2**

2 Raw Reads and Mapping QC **2**

2.1 FastQC: 2

2.2 HTSeq-count: 3

3 Dataset **4**

3.1 Raw read counts 4

3.2 Sample information 4

4 Preprocessing of the data **5**

4.1 Filtering of the data 5

5 DGE Analysis **5**

5.1 Differential Expression Analysis 5

5.2 Extracting results 5

6 Vizualising data **8**

6.1 Experimental QC - Clustering of samples (PCA) 8

6.2 Viewing counts for a single geneID 9

7 Part 2: Functional Analysis and Vizualisation **11**

7.1 Setup 11

8 Gene annotation (databases) **11**

9 Running functional analysis **13**

9.1 Overrepresentation analysis 13

10 Vizualization **16**

10.1 Vulcano Plot 16

10.2 Dotplot 18

10.3 Clustering by GO relations 22

References **25**

1 Introduction

2 Raw Reads and Mapping QC

2.1 FastQC:

a) According to FastQC: What was the minimum and the maximum number of read pairs sequenced per sample?

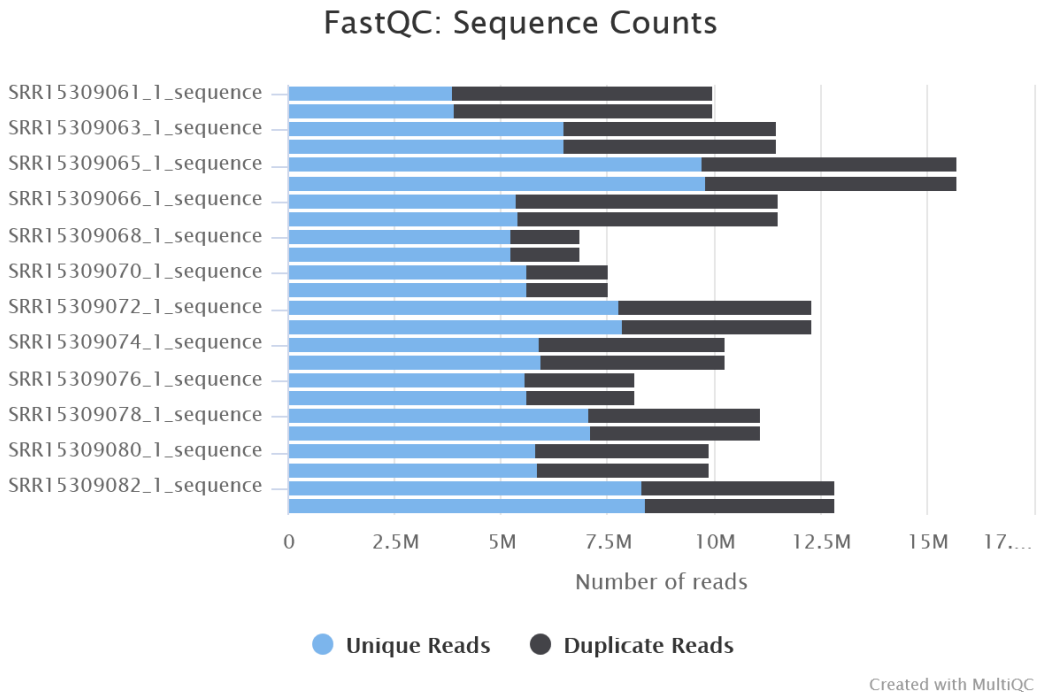

Figure 1: Barplot of the number of read pairs per sample.

Answer: As can be seen in Figure 1 the minimum number of reads per sample is around 6.8 million and the maximum number of reads per sample is around 15.7 million.

b) What is the most overrepresented sequence (string of nucleotides) that was found by FastQC?

Answer: According to the MultiQC report the most overrepresented sequence was:

```
"AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA"
```

c) What might be the reason there is so much of this specific sequence/homopolymer ?

Answer: This could be from A tailed adapter dimers and PCR slippage products that outcompete genuine RNA fragments in the library.

2.2 HTSeq-count:

d) According to HTSeq Count: What was the minimum and the maximum number of read pairs reported per sample?

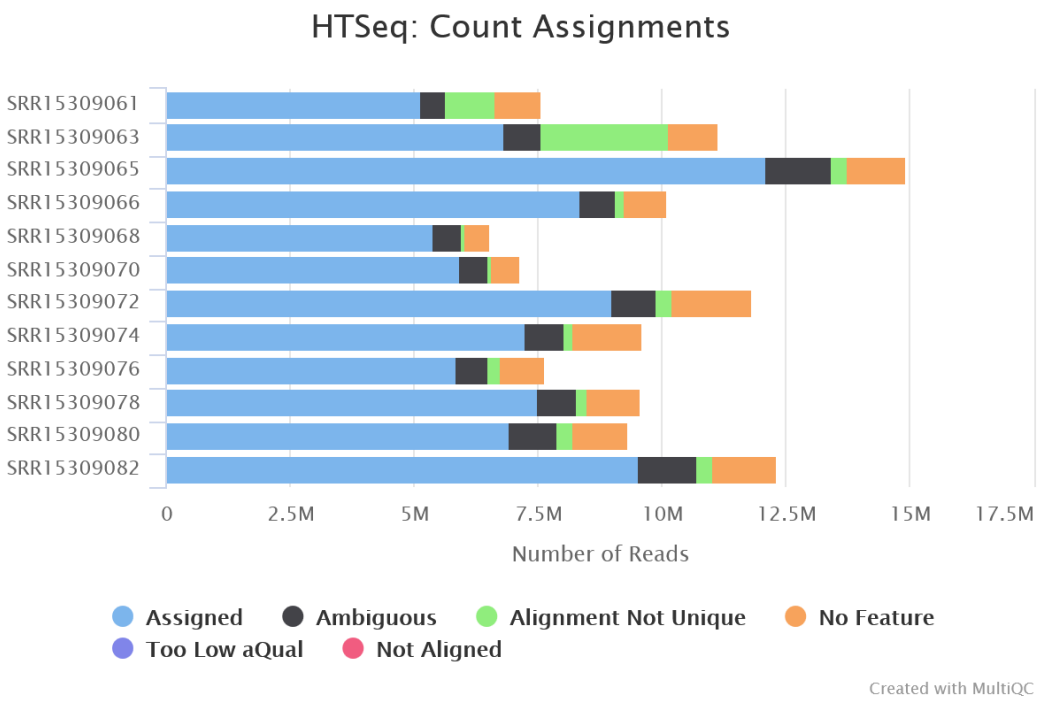

Figure 2: Barplot of the number of read pairs, per sample HTSeq.

Answer: As can be seen in Figure 2 the minimum number of reads per sample is around 5.9 million and the maximum number of reads per sample is around 14.8 million.

f) Which was the minimum and maximum percentage of reads uniquely assigned to a gene, as reported by HTSeq-count?

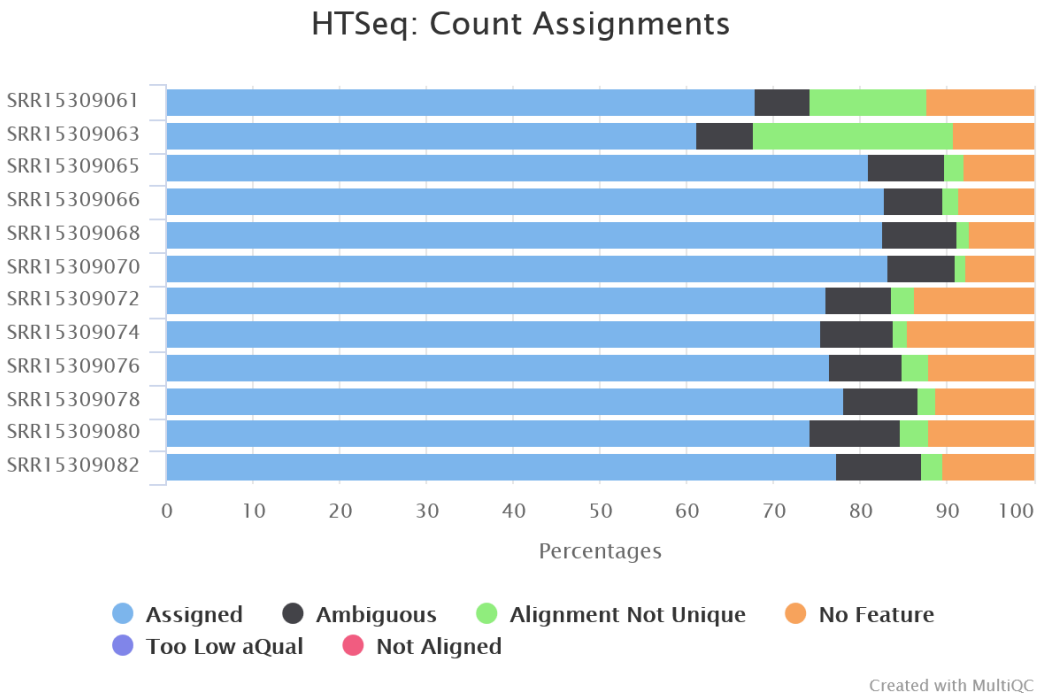

Figure 3: HTSeq assignment plot in percentages.

Answer: As can be seen in Figure 3 the minimum percentage of reads uniquely assigned to a gene is 61.3% and the maximum percentage of reads uniquely assigned to a gene is around 83.3%.

3 Dataset

3.1 Raw read counts

3.2 Sample information

a) Which columns define the used base strain and substrain (WT or TGT mutant), respectively? Can you spot an error in one of those columns?

Answer: The genotype column shows if the wild type or tgt mutant was used. The strain column shows wich *E.coli* strain was used. The error is in the strain column suggesting that for every experiment a different strain was used. But comparing this with the strains and method supsection of the methods section of the paper shows that only "Escherichia coli K-12 MG1655 was used as the WT strain.".[1]

b) Which column defines if nickel was added to the media?

Answer: The treatment column defines if nickel was added to the media.

c) Find the following information on the SRA Study page:

Which type of Illumina machine was used for sequencing?

Answer: Illumina HiSeq 2500

What was the library layout?

Answer: PAIRED

When was the data released?

Answer: 2022-05-16

4 Preprocessing of the data

4.1 Filtering of the data

a) For how many genes did we originally retrieve count data?

```
dim(rawCounts)
#[1] 4295   12
```

Answer: Originally count data were retrieved for 4295 genes

b) How many will be left after applying the filter?

```
dim(rawCounts[rowSums(rawCounts) > 10, ])
#[2] 4221   12
```

Answer: After applying the filter 4221 genes will be left.

5 DGE Analysis

5.1 Differential Expression Analysis

5.2 Extracting results

Interpreting the summary:

a) How many genes are significantly up-regulated and how many are significantly down-regulated in the nickel treated WT strain as compared to the untreated WT strain, using the default cutoff for the adjusted p-value?

```
# Extract results with default alpha (0.1)
DESeq2Results_WT_nickel <- results(DESeq2Data,
                                   contrast = c("group","WT.Nickel","WT.none"))

# View summary
summary(DESeq2Results_WT_nickel)


# out of 4221 with nonzero total read count
# adjusted p-value < 0.1
# LFC > 0 (up)       : 1069, 25%
# LFC < 0 (down)     : 1063, 25%
# outliers [1]       : 6, 0.14%
# low counts [2]     : 0, 0%
# (mean count < 1)
# [1] see 'cooksCutoff' argument of ?results
# [2] see 'independentFiltering' argument of ?results
```

Answer: 1069 genes are significantly up-regulated and 1063 genes are significatnly down-regulated in the nickel treated WT strain as compared to the untreated WT strain, using the default cutoff for the adjusted p-value of 10%.

b) What is the standard cutoff used for the significance level (adjusted p-value), if we don't change it?

Answer: The standard cutoff used for the significance level (adjusted p-value), if we don't change it is 10% (p-value < 0.1).

c) How many significantly differentially expressed genes does that make in total?

```
# Extract TGT.Nickel vs TGT.none with alpha = 0.1
DESeq2Results_TGT_nickel <- results(DESeq2Data,
                                    contrast = c("group","TGT.Nickel","TGT.none"),
                                    alpha = 0.1)

# View summary
summary(DESeq2Results_TGT_nickel)


# out of 4221 with nonzero total read count
# adjusted p-value < 0.1
# LFC > 0 (up)       : 986, 23%
# LFC < 0 (down)     : 877, 21%
# outliers [1]       : 6, 0.14%
# low counts [2]     : 164, 3.9%
# (mean count < 5)
```

```
# [1] see 'cooksCutoff' argument of ?results
# [2] see 'independentFiltering' argument of ?results
```

Answer: If we add up the up und down regulated genes we get the total amount of significantly differentially expressed genes. Looking at the summary this would be 1863 genes.

Changing the alpha factor:

d) For the comparison of the genotypes under standard contitions. How many significantly differentially expressed genes in total are reported for a significance level of 0.05? (Go to the RStudio Help and search for "results" function, to identify the attribute you have to change.)

```
# Extract genotype comparison (TGT vs WT, no nickel) with alpha = 0.05
DESeq2Results_genotype <- results(DESeq2Data,
                                  contrast = c("group","TGT.none","WT.none"),
                                  alpha = 0.05)

# View summary (this shows the numbers you need)
summary(DESeq2Results_genotype)


# out of 4221 with nonzero total read count
# adjusted p-value < 0.05
# LFC > 0 (up)       : 187, 4.4%
# LFC < 0 (down)     : 275, 6.5%
# outliers [1]       : 6, 0.14%
# low counts [2]     : 0, 0%
# (mean count < 1)
# [1] see 'cooksCutoff' argument of ?results
# [2] see 'independentFiltering' argument of ?results
```

Answer: If we add up the up und down regulated genes we get the total amount of significantly differentially expressed genes. Looking at the summary this would be 462 genes at a p-value $<$ 0.05.

Comparing the nickel treatment to no treatment in the TGT-mutant:

e) Repeat the steps above for the comparison of the TGT-mutant strain treated with nickel to the TGT-mutant strain not treated with nickel. How many significantly differentially expressed genes in total are reported for a significance level of 0.05?

```
# Extract TGT mutant nickel effect with alpha = 0.05
DESeq2Results_TGT_nickel <- results(DESeq2Data,
                                    contrast = c("group","TGT.Nickel","TGT.none"),
                                    alpha = 0.05)
```

```
# View summary (this shows the numbers you need)
summary(DESeq2Results_TGT_nickel)

# out of 4221 with nonzero total read count
# adjusted p-value < 0.05
# LFC > 0 (up)       : 826, 20%
# LFC < 0 (down)     : 752, 18%
# outliers [1]       : 6, 0.14%
# low counts [2]     : 82, 1.9%
# (mean count < 3)
# [1] see 'cooksCutoff' argument of ?results
# [2] see 'independentFiltering' argument of ?results
```

Answer: If we add up the up und down regulated genes we get the total amount of significantly differentially expressed genes. Looking at the summary this would be 1578 genes at a p-value $<$ 0.05.

6 Vizualising data

6.1 Experimental QC - Clustering of samples (PCA)

Figure 4: Clustering of samples PCA dot plot.

a) Do the groups of replicates behave as expected?

Answer: Looking at the plot in figure 4 we can observe that the individual groups are clearly separated from each other. It also makes sense that there is a big distance between untreated and Nickel treated strains (WT as well as TGT strains).

b) Which sample would you identify as an outlier?

Answer: Looking at figure 4 we would identify the WT strain untreated as an outlier as it does not completely cluster with its biological replicates but one point also clusters with the TGT untreated cluster.

6.2 Viewing counts for a single geneID

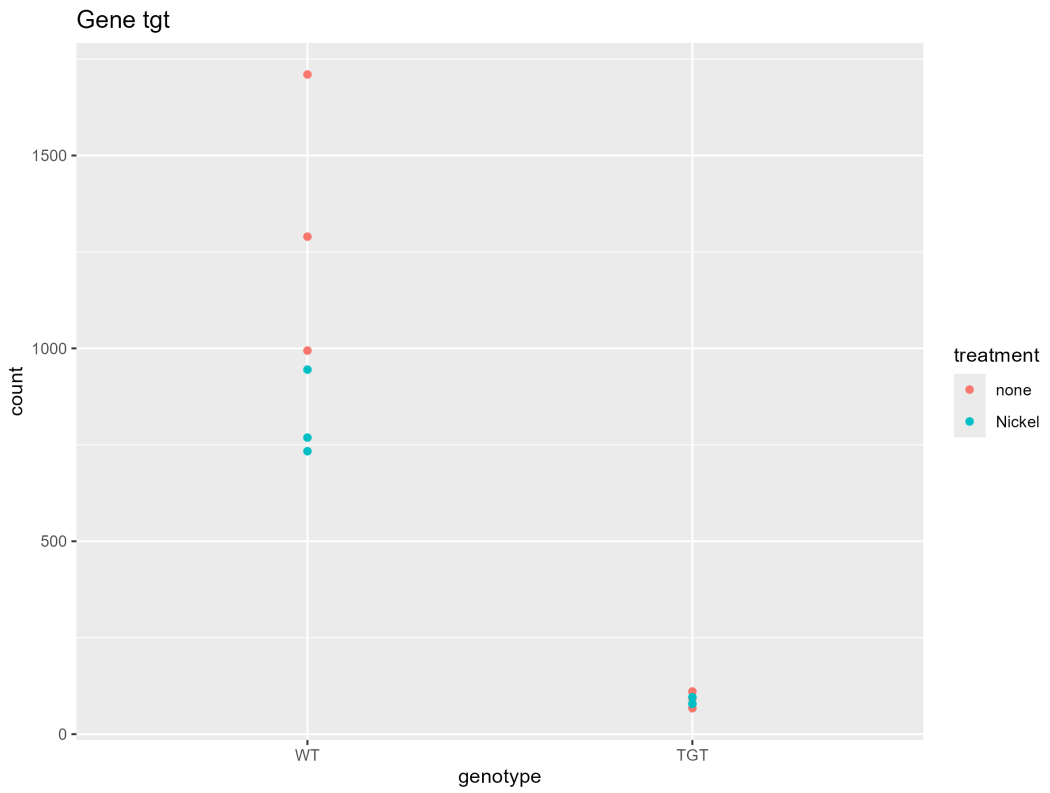

Figure 5: Viewing counts for a single geneID in a dot plot.

c) Think about the mutations in the E. coli strains and how that influences the transcripts of a gene. Are the read counts for the tgt gene in the wild type and the knockout strain what you expected? Explain why.

Answer:

Wild type + none: tgt encodes tRNA guanine transglycosylase, an essential enzyme for queuosine synthesis. It is expressed under normal growth conditions to modify tRNAs. Expected Counts are going to be high.

Wild type + Nickel: tgt is transcriptionally repressed by nickel stress. Expected Counts are going to be lower than that of Wt + none but still higher than the expected counts of the knock out mutants.

tgt + none: The tgt gene is deleted so No functional tgt should be transcribed. Expected counts are going to be very low to near zero.

tgt + Nickel: The tgt knock out stais the same but now Nickel is added. This should not have a significant affect on the already very low to near zero counts.

7 Part 2: Functional Analysis and Vizualisation

7.1 Setup

```
#BiocManager::install(c("clusterProfiler"))
#install.packages("tidyverse")

#install.packages("devtools")
#devtools::install_github('kevinblighe/EnhancedVolcano')
```

```
library(EnhancedVolcano)
library(clusterProfiler)
library(tidyverse)
library(ggplot2)
library(dplyr)
library(DESeq2)
```

```
### load the data with read delim, because it is tab seperated
annotatedRawCounts <- read_delim("Counts_raw.tsv")

#head(annotatedRawCounts)

annotatedRawCounts <- annotatedRawCounts  %>%
  column_to_rownames(var = "ID")

### split data into rawCounts and Annotations
rawCounts <- annotatedRawCounts[,3:14]
annotations <- annotatedRawCounts[,1:2]

DESeq2ResultsDF <- read_delim("DESeq2Result_treatment.tsv")
```

8 Gene annotation (databases)

a) If you look closely at the 21 OrgDB packes, there are 2 different packages for E. coli. Which one should we use?

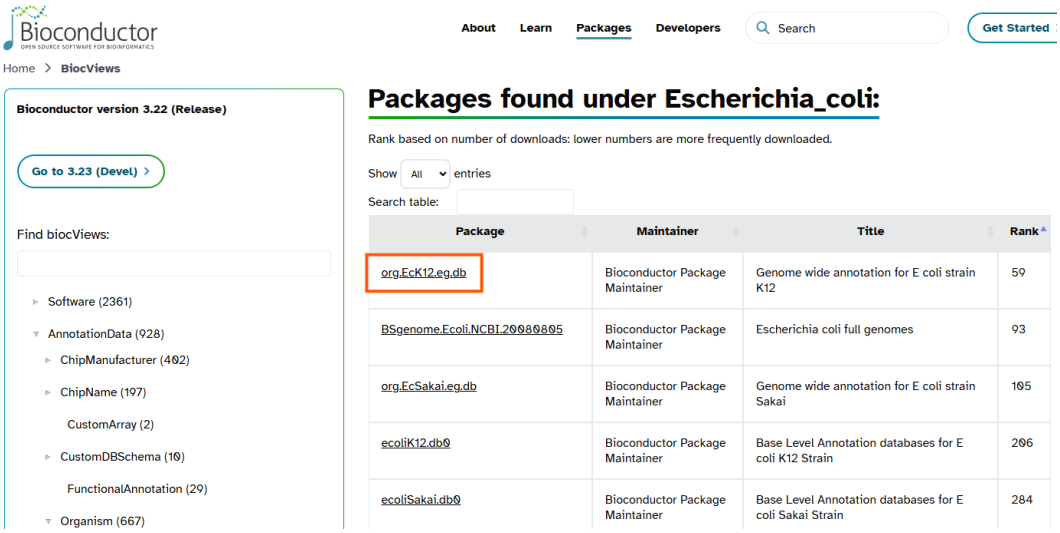

Figure 6: Website of the used database for E.coli K12.

```
BiocManager::install("org.EcK12.eg.db")

library(org.EcK12.eg.db)
```

Answer: Since we are working with E. coli K12 MG1655 it would make sense to use the genome wide annotation package for E.coli K12 (org.EcK12.eg.db)

c) List the different available keytypes/identifiers.

```
genes <- annotations$gene
head(genes,10)
```

```
##  [1] "thrL" "thrA" "thrB" "thrC" "yaaX" "yaaA" "yaaJ" "talB" "mog"  "satP"
```

```
keytypes(org.EcK12.eg.db)
```

```
##  [1] "ACCNUM"      "ALIAS"       "ENTREZID"    "ENZYME"      "EVIDENCE"
##  [6] "EVIDENCEALL" "GENENAME"    "GO"          "GOALL"       "ONTOLOGY"
## [11] "ONTOLOGYALL" "PATH"        "PMID"        "REFSEQ"      "SYMBOL"
```

Answer: The different keytypes and identifiers are shown above.

d) Which of the keytypes/identifiers include the gene names given in the annotation?

```
keys <- keys(org.EcK12.eg.db, keytype = "SYMBOL")
head(keys)
```

```
## [1] "yjhR"  "nfrA"  "thrL"  "insB1" "sspA"  "yaaJ"
```

```
sum(genes %in% keys)
```

```
## [1] 4257
```

Answer: After iterating through all different keytypes, SYMBOL was found to contain the genes.

9 Running functional analysis

9.1 Overrepresentation analysis

a) How many significantly up- and downregulated genes are left, after applying the LFC cutoff?

```
res_up <- dplyr::filter(DESeq2ResultsDF, padj < 0.05 & log2FoldChange > 1)
res_down <- dplyr::filter(DESeq2ResultsDF, padj < 0.05 & log2FoldChange < -1)

genes_up_id <- res_up$ID
genes_up <- annotations[genes_up_id,"gene"]

genes_down_id <- res_down$ID
genes_down <- annotations[genes_down_id,"gene"]

genes_de <- c(genes_up,genes_down)

length(genes_up)
```

```
## [1] 660
```

```
length(genes_down)
```

```
## [1] 582
```

```
length(genes_de)
```

```
## [1] 1242
```

Answer: With the *length* command we can show the number of genes up- or downregulated. In total, 660 genes were up- and 582 genes were downregulated. In sum, 1242 genes are deregulated.

b) How many significantly over-represented biological processes (GO terms) are there, per subset of genes (all differentially expressed genes, up-regulated genes only, down-regulated genes only.)

```
EC <- "org.EcK12.eg.db"
EC_KEY <- "SYMBOL"

orBP <- enrichGO(genes_de,
                 EC,
                 ont="BP",
                 keyType = EC_KEY,
                 pvalueCutoff=0.05)


orUpBP <- enrichGO(genes_up,
                 EC,
                 ont="BP",
                 keyType = EC_KEY,
                 pvalueCutoff=0.05)


orDownBP <- enrichGO(genes_down,
                 EC,
                 ont="BP",
                 keyType = EC_KEY,
                 pvalueCutoff=0.05)

#To show the number of statistically significant GO terms
n_BP_all <- nrow(orBP)
n_BP_up <- nrow(orUpBP)
n_BP_down <- nrow(orDownBP)

print(n_BP_all)
```

```
## [1] 98
```

```
print(n_BP_up)
```

```
## [1] 82
```

```
print(n_BP_down)
```

```
## [1] 100
```

Answer: There are in summary 100 GO terms deregulated, 59 up and 93 down. For a quick look into the first few results, take a look at the tables below.

Table 1: Overrepresented BP terms - All DE genes

	Description	p.adjust	Count
GO:0006935	chemotaxis	0e+00	34
GO:0042330	taxis	0e+00	34
GO:0040011	locomotion	0e+00	42
GO:0001539	cilium or flagellum-dependent cell motility	2e-07	49
GO:0071973	bacterial-type flagellum-dependent cell motility	2e-07	49

Table 2: Overrepresented BP terms - Upregulated genes

	Description	p.adjust	Count
GO:0046377	colanic acid metabolic process	0.00e+00	19
GO:0009242	colanic acid biosynthetic process	0.00e+00	16
GO:0072348	sulfur compound transport	7.00e-07	23
GO:0006857	oligopeptide transport	2.30e-06	21
GO:0042938	dipeptide transport	3.58e-05	15

Table 3: Overrepresented BP terms - Upregulated genes

	Description	p.adjust	Count
GO:0046377	colanic acid metabolic process	0.00e+00	19
GO:0009242	colanic acid biosynthetic process	0.00e+00	16
GO:0072348	sulfur compound transport	7.00e-07	23
GO:0006857	oligopeptide transport	2.30e-06	21
GO:0042938	dipeptide transport	3.58e-05	15

10 Vizualization

10.1 Vulcano Plot

a) Which is the most down-regulated gene and which is the most up-regulated gene, in terms of LFC?

```
# Set a boolean column for significance
DESeq2ResultsDF$significant <- ifelse(!is.na(DESeq2ResultsDF$padj) & DESeq2ResultsDF$padj < .05

## Add gene annotations
DESeq2ResultsDF$gene <- annotations[DESeq2ResultsDF$ID,"gene"]

EnhancedVolcano(DESeq2ResultsDF,
                lab =  DESeq2ResultsDF$gene,
                x = 'log2FoldChange',
                y = 'padj',
                pCutoff = 0.05)
```

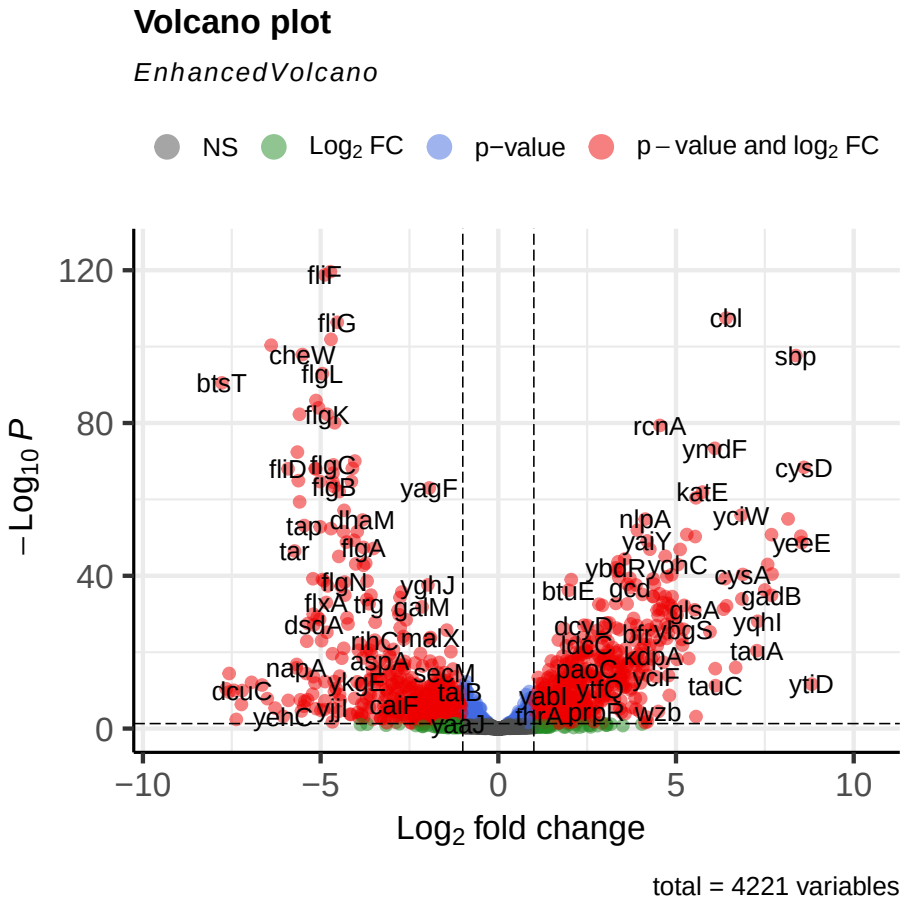

Figure 7: Vulcano plot

```r
#show the entry with the highest log2FoldChange
DESeq2ResultsDF[DESeq2ResultsDF$log2FoldChange == max(DESeq2ResultsDF$log2FoldChange),]
```

```
## # A tibble: 1 x 9
##   baseMean log2FoldChange lfcSE  stat   pvalue     padj ID    significant gene
##      <dbl>          <dbl> <dbl> <dbl>    <dbl>    <dbl> <chr> <lgl>       <chr>
## 1     41.6           8.82  1.20  7.33 2.25e-13 2.33e-12 b4721 TRUE        ytiD
```

```r
#show the entry with the lowest log2FoldChange
DESeq2ResultsDF[DESeq2ResultsDF$log2FoldChange == min(DESeq2ResultsDF$log2FoldChange),]
```

```
## # A tibble: 1 x 9
##   baseMean log2FoldChange lfcSE  stat   pvalue     padj ID    significant gene
##      <dbl>          <dbl> <dbl> <dbl>    <dbl>    <dbl> <chr> <lgl>       <chr>
## 1   38216.          -7.78 0.378 -20.6 7.41e-94 3.13e-91 b4354 TRUE        btsT
```

```r
ggsave('figs/volcano_plot.png', width = 12, height = 8)
```

Answer: The most upregulated gene is *ytiD* with a L2FC of 8.82. The most downregulated gene is *btsT* with a L2FC of -7.78.

b) How many times are those genes higher or lower expressed in treated compared to untreated WT strain? (Calculate the fold change from the LFC)

```r
DESeq2ResultsDF <- read_delim("DESeq2Result_treatment.tsv")
res_df <- as.data.frame(DESeq2ResultsDF)

most_up_idx <- which.max(res_df$log2FoldChange)
most_up <- res_df[most_up_idx, ]

most_down_idx <- which.min(res_df$log2FoldChange)
most_down <- res_df[most_down_idx, ]

gene_name_up <- annotations[most_up$ID, "gene"]
gene_name_down <- annotations[most_down$ID, "gene"]


most_up$foldChange <- 2^most_up$log2FoldChange
most_down$foldChange <-  2^most_down$log2FoldChange

most_up$gene <- gene_name_up
most_down$gene <- gene_name_down

print(c(most_up$log2foldChange, most_up$foldChange ,most_up$gene ))
```

```
## [1] "451.110505802606" "ytiD"
```

```
print(c(most_down$log2foldChange, most_down$foldChange, most_down$gene))
```

```
## [1] "0.00455492673103271" "btsT"
```

Answer: For a comprehensive comparison, we can do this manually by thinking of how the LFC is defined. Since it's a logarithmic function with base 2, we can calculate the FoldChange: FC = 2^LFC and compare the results manually. For faster and automated calculation, one may use the code above, which takes the data from the DESeq dataset for treated and untreated WT and calculates the FC. Gene *ytiD* is 451.1 times higher transcribed than in untreated WT. Gene *btsT* is 217.4 times lower transcribed than in untreated WT.

c) Which of the genes shows the most significant deregulation (the lowest p-value)?

```
min_pvalue_idx <- which.min(res_df$pvalue)
gene_min_pvalue <- res_df[min_pvalue_idx, ]

gene_id_sign <- gene_min_pvalue$ID
gene_name_sign <- annotations[gene_id_sign, "gene"]

most_sign <- gene_min_pvalue
most_sign$gene <- gene_name_sign

print(c(gene_min_pvalue$pvalue, most_sign$gene))
```

```
## [1] "5.66136528139081e-124" "fliM"
```

Answer: The gene with the lowest p-value (the most significant deregulation): *fliM* with a **p-value = 5.66e-124**.

10.2 Dotplot

a) Which is the most significantly enriched GO term ?

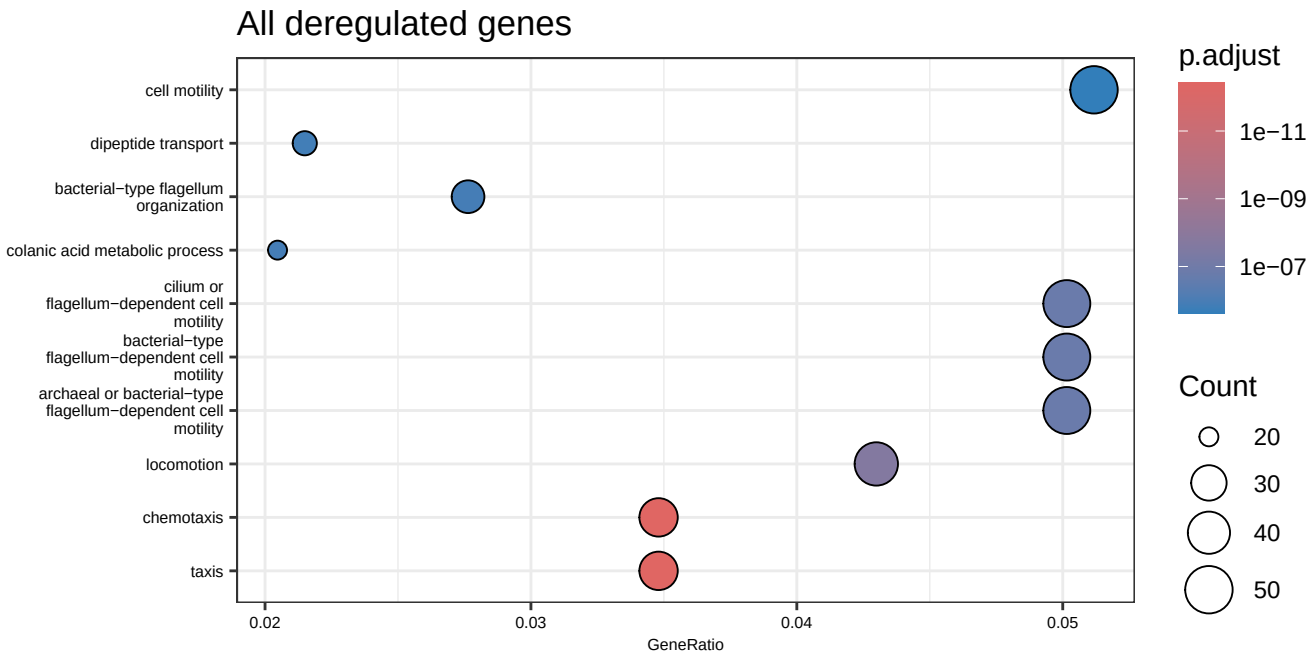

Figure 8: All deregulated genes

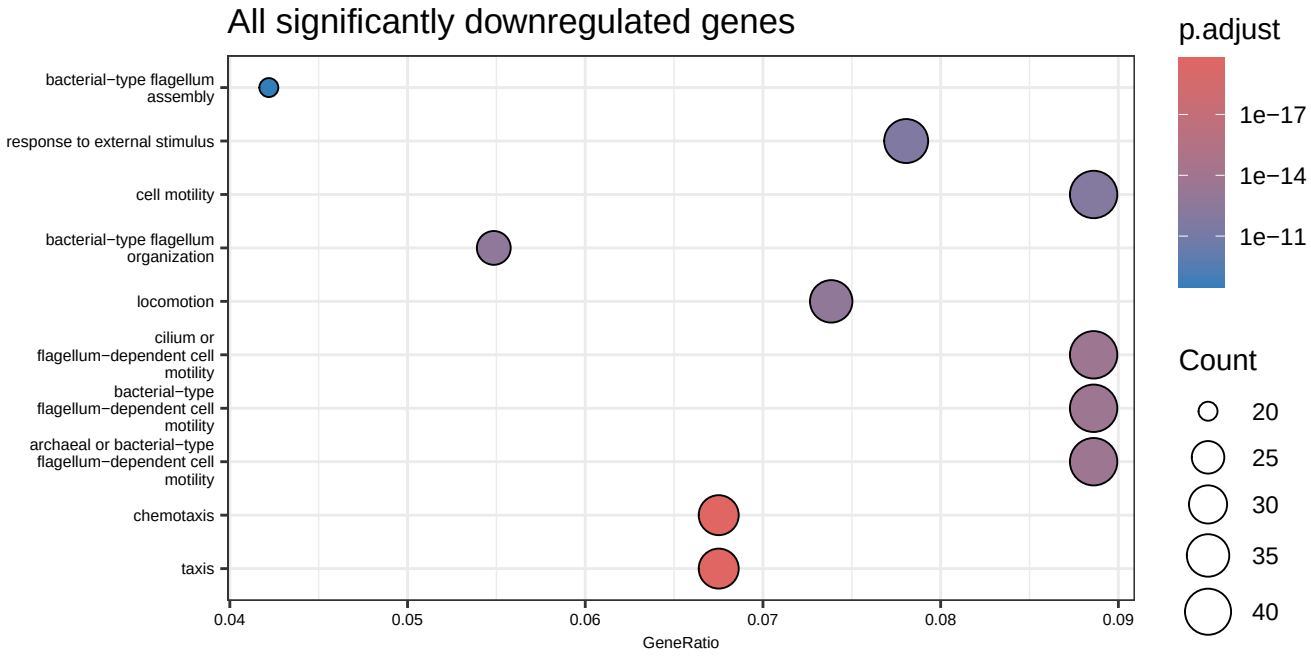

Figure 9: All significantly downregulated genes

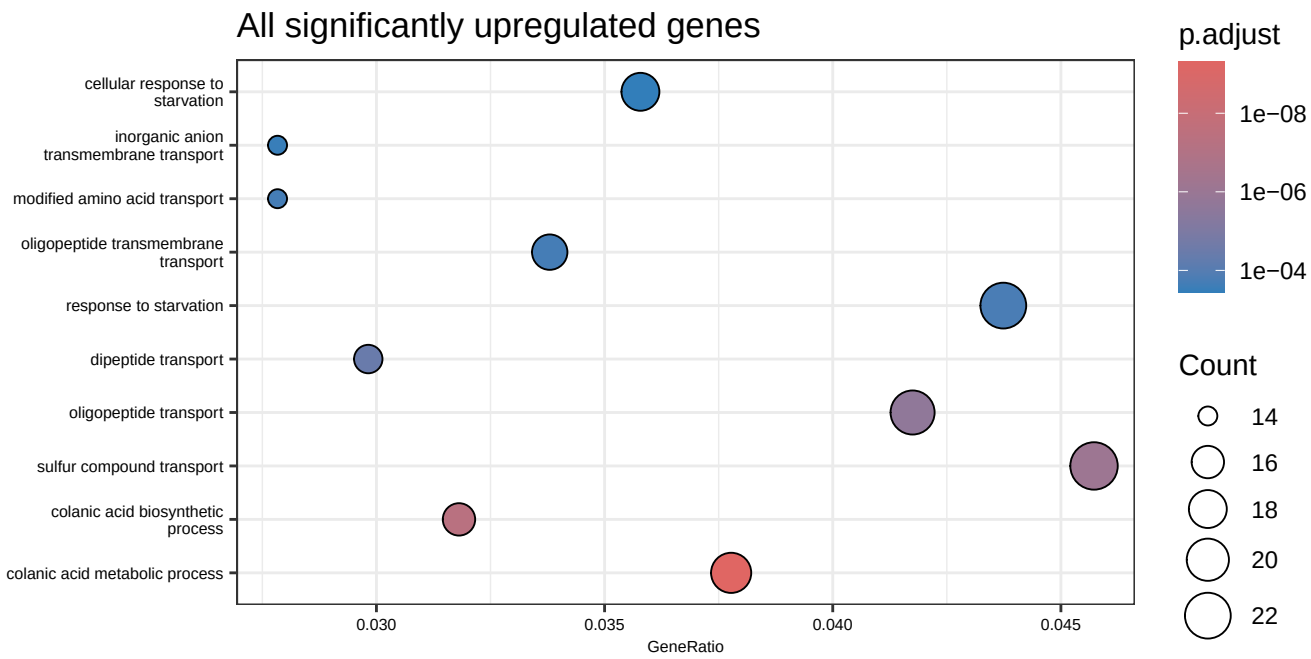

Figure 10: All significantly upregulated genes

a) Which is the most significantly enriched GO term in terms of upregulated genes?

Answer: By looking at the graphs,which are sorted by adjusted p-values, we can see that the **genes for response to starvation** have the lowest p-value, therefore are the most significant upregulated GO term.

b) Which is the most significantly enriched GO term in terms of downregulated genes?

Answer: By looking at the graphs,which are sorted by adjusted p-values, we can see that the **genes for iron import into the cell** have the lowest p-value, therefore are the most significant downregulated GO term.

c) Which of the shown GO terms has the highest GeneRatio ?

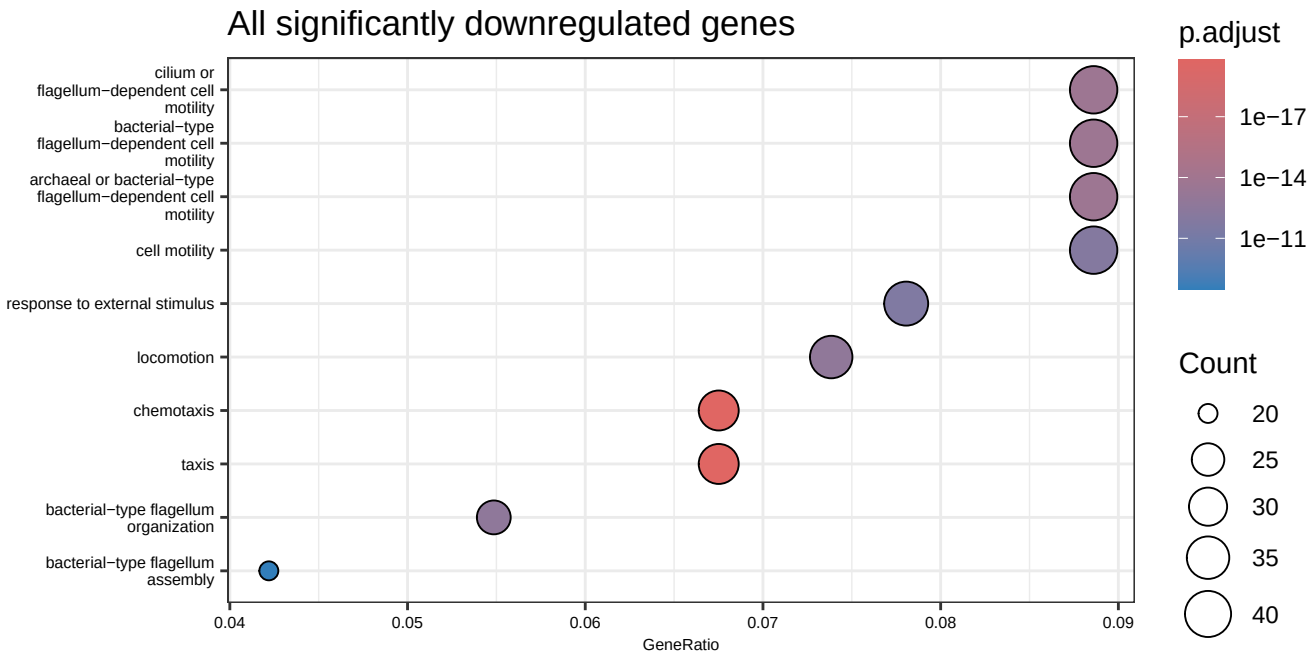

Figure 11: All significantly downregulated genes

Answer: By looking at the graphs,which are sorted by GeneRatio, we can see that **gene terms with mobility functions like cell motility or flagellum-dependent cell motility** rank highest in terms of GeneRatio.

d) Which of the shown GO terms has the highest Count ?

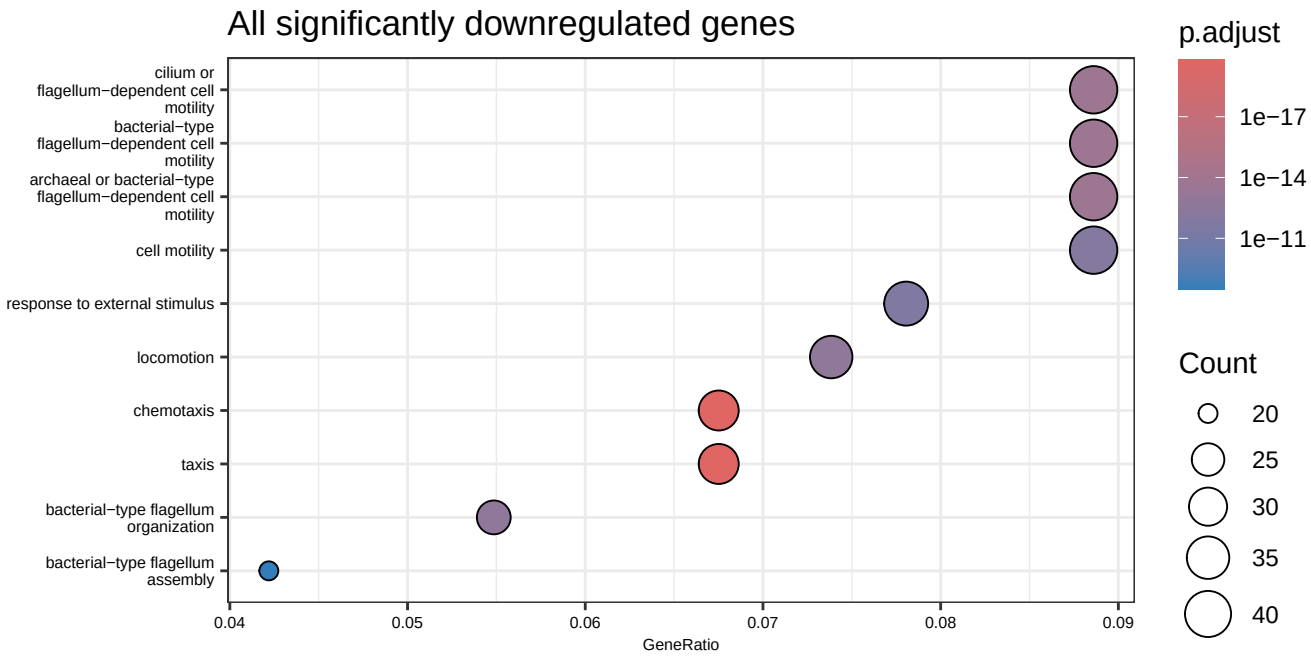

Figure 12: All significantly downregulated genes

Answer: By looking at the graphs,which are sorted by Counts, we can see that **the same gene terms with mobility functions like cell motility or flagellum-dependent cell motility** rank also highest in terms of Counts.

10.3 Clustering by GO relations

a) How many clusters can you identify when showing 10, 20, and 30 categories?

```
#Setting group=TRUE also does a visual clustering with labels
enrichedBP <- enrichplot::pairwise_termsim(orDownBP)
emapplot(enrichedBP, showCategory = 30, node_label="group", layout=igraph::layout_with_kk)
```

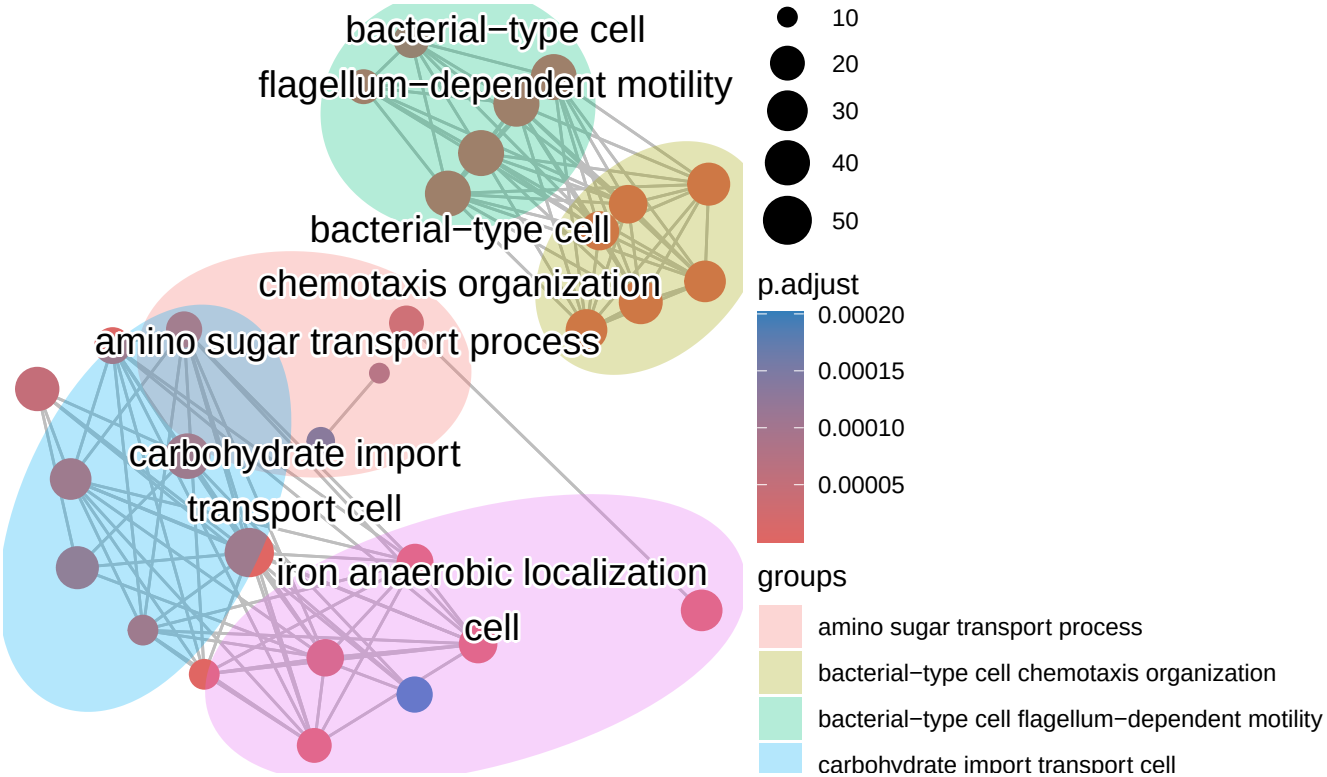

Figure 13: Enrichment plot with visual clustering for easier analyis. Used categories: 30

Answer: For easier counting, the node labels were grouped using nodel_label="group".

1) For 10 Categories, we count 3 clusters.
2) For 20 Categories, we count 4 clusters.
3) For 30 Categories, we count 5 clusters.

b) What are in your opinion the main clusters and which processes are they related to?

Answer: Enabling nodel_label="group" vizualises the different clusters. This clusters and a short biological explanation is represented below:

1) Iron anaerobic localization: Under nickel exposure, iron metabolism is broadly perturbed, and several studies show transcriptional repression of iron uptake and related functions as part of a protective response to prevent further metal overload and mismetallation of iron enzymes. [[2]][3] In this setting, this is represented by the shift away from pathways that include iron transport.

2) carbohydrate import transport: Carbohydrate transport systems are major entry points for carbon and are under strong catabolite repression and stress regulation. During metal stress, cells often reduce uptake and catabolism of some carbon sources, lowering metabolic flux and thereby reducing production of reactive by-products. Down-regulation of carbohydrate import fits a program, in which only the most efficient or stress-compatible substrates are prioritized while others are transcriptionally repressed. [4]

3) amino sugar transport process: Amino sugar uptake and utilization (e.g. N-acetylglucosamine) for cell wall biosynthesis is a process that is energy-intensive and tightly adjusted to growth rate. Under metal stress and reduced growth, expression of transporters for alternative nutrients, including amino sugars, is commonly repressed as part of a general reduction in anabolic activity. Thus, the observed down-regulation of amino sugar transport genes likely reflects both lower growth rate and a reprioritization of nutrient use under nickel-induced stress.[4] [5]

4) bacterial-type cell flagellum-dependent motility: Flagella and motility are among the most expensive structures for bacteria to build and run and many stresses (metals, antibiotics, nutrient limitation) trigger coordinated repression of flagellar gene expression. In addition, motile planktonic growth is often disadvantageous under stress, while bbiofilm-like states provide protection. Nickel, for example, can promote biofilm formation in E. coli where non-motile cells have a fitness advantage. Down-regulation of flagellar genes is therefore consistent with an adaptive shift from an energy-intensive motile lifestyle towards a more stationary state.[6] [7] [8] [9] [5]

5) bacterial-type cell chemotaxis organization: Chemotaxis systems coordinate gradient sensing and flagellar rotation; they are functionally coupled to motility and are typically co-regulated with flagellar genes. When flagellar assembly and rotation are repressed under metal or nutrient stress, chemotaxis signal transduction becomes less useful and is often down-regulated in parallel. The reduced expression of chemotaxis and chemotaxis-organization genes thus likely reflects this coordinated shut-down of the motility module in favor of conserving ATP and redirecting resources to metal detoxification and stress-response pathways.[8]

References

[1] L. Pollo-Oliveira *et al.*, "The absence of the queuosine tRNA modification leads to pleiotropic phenotypes revealing perturbations of metal and oxidative stress homeostasis in escherichia coli K12," *Metallomics*, vol. 14, no. 9, p. mfac065, 2022.

[2] C. L. Washington–Hughes, G. T. Ford, A. D. Jones, K. McRae, and F. W. Outten, "Nickel exposure reduces enterobactin production in escherichia coli," *MicrobiologyOpen*, vol. 8, no. 4, p. e00691, 2019.

[3] S. Semsey, A. M. Andersson, S. Krishna, M. H. Jensen, E. Massé, and K. Sneppen, "Genetic regulation of fluxes: Iron homeostasis of escherichia coli," *Nucleic acids research*, vol. 34, no. 17, pp. 4960–4967, 2006.

[4] E. Sonnleitner, "A comparative analysis: Molecular mechanisms of carbon catabolite repression in bacteria," *Annual Review of Microbiology*, vol. 79, no. 1, pp. 241–262, 2025.

[5] Z. Khan, A. Rehman, M. A. Nisar, S. Zafar, and I. Zerr, "Biosorption behavior and proteomic analysis of escherichia coli P4 under cadmium stress," *Chemosphere*, vol. 174, pp. 136–147, 2017.

[6] C. Perrin *et al.*, "Nickel promotes biofilm formation by escherichia coli k-12 strains that produce curli," *Applied and environmental microbiology*, vol. 75, no. 6, pp. 1723–1733, 2009.

[7] R. Marques Mendonca, T. Fulton, C. Blackwood, and D. Costello, "Sublethal nickel toxicity shuts off manganese oxidation and pellicle biofilm formation in pseudomonas putida GB-1," *Environmental Microbiology*, vol. 25, no. 12, pp. 3639–3654, 2023.

[8] S. Lin, R. Liang, T. Zhang, Y. Yuan, S. Shen, and H. Ye, "Microarray analysis of the transcriptome of the escherichia coli (e. Coli) regulated by cinnamaldehyde (CMA)," *Food and Agricultural Immunology*, vol. 28, no. 3, pp. 500–515, 2017.

[9] W. Benyoussef, M. Deforet, A. Monmeyran, and N. Henry, "Flagellar motility during e. Coli biofilm formation provides a competitive disadvantage which recedes in the presence of co-colonizers," *Frontiers in Cellular and Infection Microbiology*, vol. 12, p. 896898, 2022.

[10] M. Love, S. Anders, and W. Huber, *DESeq2: Differential gene expression analysis based on the negative binomial distribution*. 2025. Available: https://github.com/thelovelab/DESeq2

[11] H. Wickham *et al.*, *ggplot2: Create elegant data visualisations using the grammar of graphics*. 2025. Available: https://ggplot2.tidyverse.org

[12] Y. Xie, *Knitr: A general-purpose package for dynamic report generation in r*. 2025. Available: https://yihui.org/knitr/

[13] J. Allaire *et al.*, *Rmarkdown: Dynamic documents for r*. 2025. Available: https://github.com/rstudio/rmarkdown

[14] H. Wickham, *Tidyverse: Easily install and load the tidyverse*. 2023. Available: https://tidyverse.tidyverse.org

[15] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biology*, vol. 15, p. 550, 2014, doi: 10.1186/s13059-014-0550-8.

[16] H. Wickham, *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York, 2016. Available: https://ggplot2.tidyverse.org

[17] Y. Xie, *Dynamic documents with R and knitr*, 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC, 2015. Available: https://yihui.org/knitr/

[18] Y. Xie, "Knitr: A comprehensive tool for reproducible research in R," in *Implementing reproducible computational research*, V. Stodden, F. Leisch, and R. D. Peng, Eds., Chapman; Hall/CRC, 2014.

[19] Y. Xie, J. J. Allaire, and G. Grolemund, *R markdown: The definitive guide*. Boca Raton, Florida: Chapman; Hall/CRC, 2018. Available: https://bookdown.org/yihui/rmarkdown

[20] Y. Xie, C. Dervieux, and E. Riederer, *R markdown cookbook*. Boca Raton, Florida: Chapman; Hall/CRC, 2020. Available: https://bookdown.org/yihui/rmarkdown-cookbook

[21] H. Wickham *et al.*, "Welcome to the tidyverse," *Journal of Open Source Software*, vol. 4, no. 43, p. 1686, 2019, doi: 10.21105/joss.01686.