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 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?

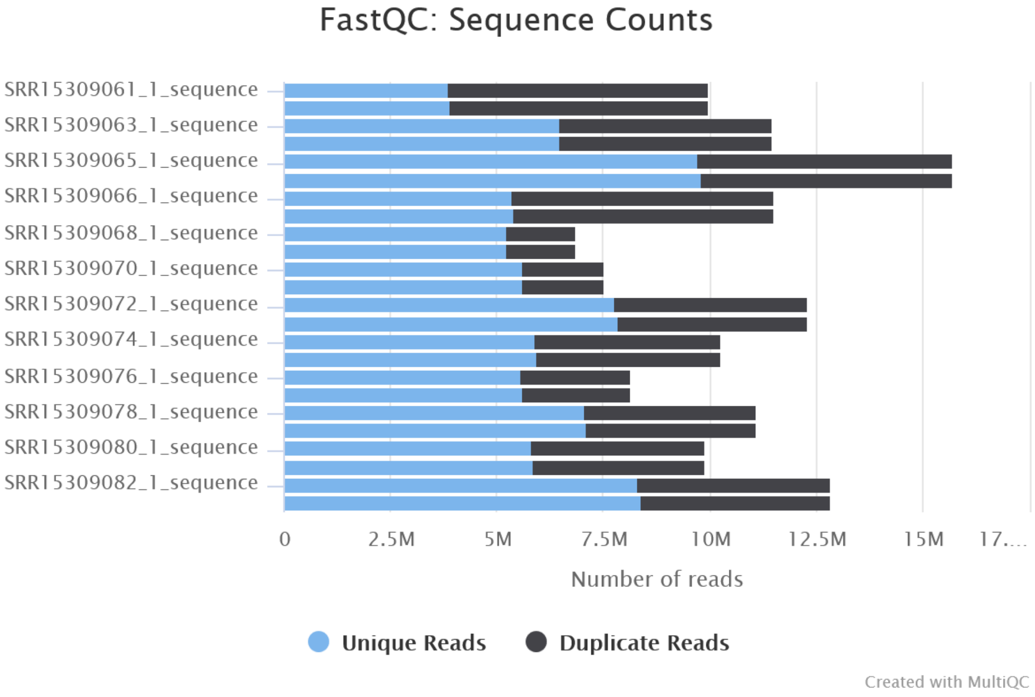


Figure 1: Barplot of the number of read pairs per sample.

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?

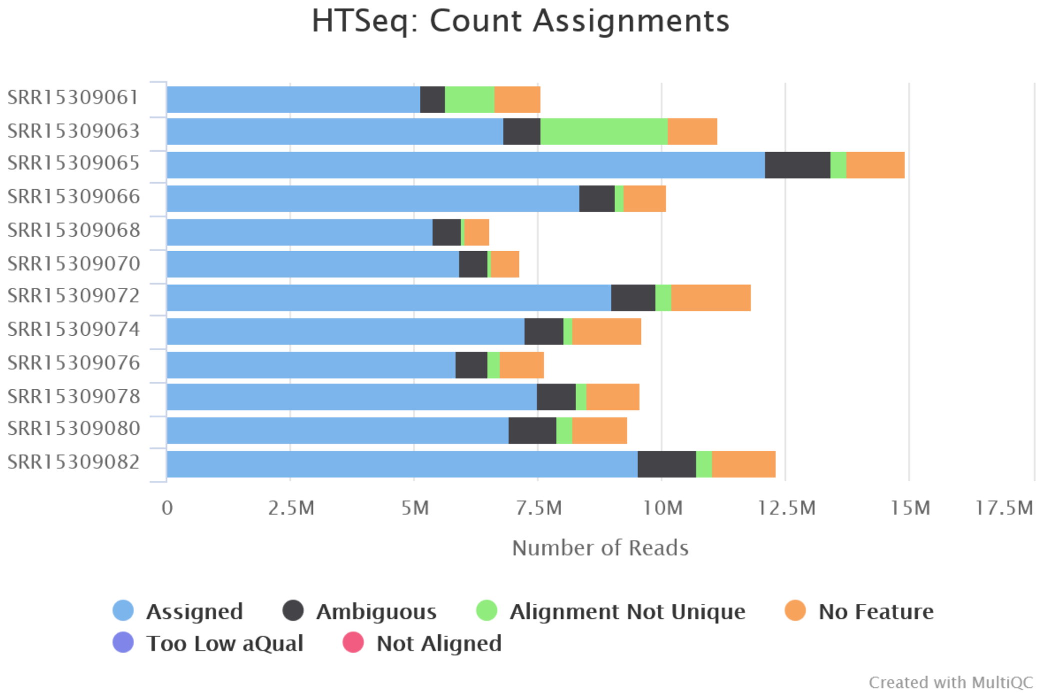


Figure 2: Barplot of the number of read pairs, per sample HTSeq.

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?

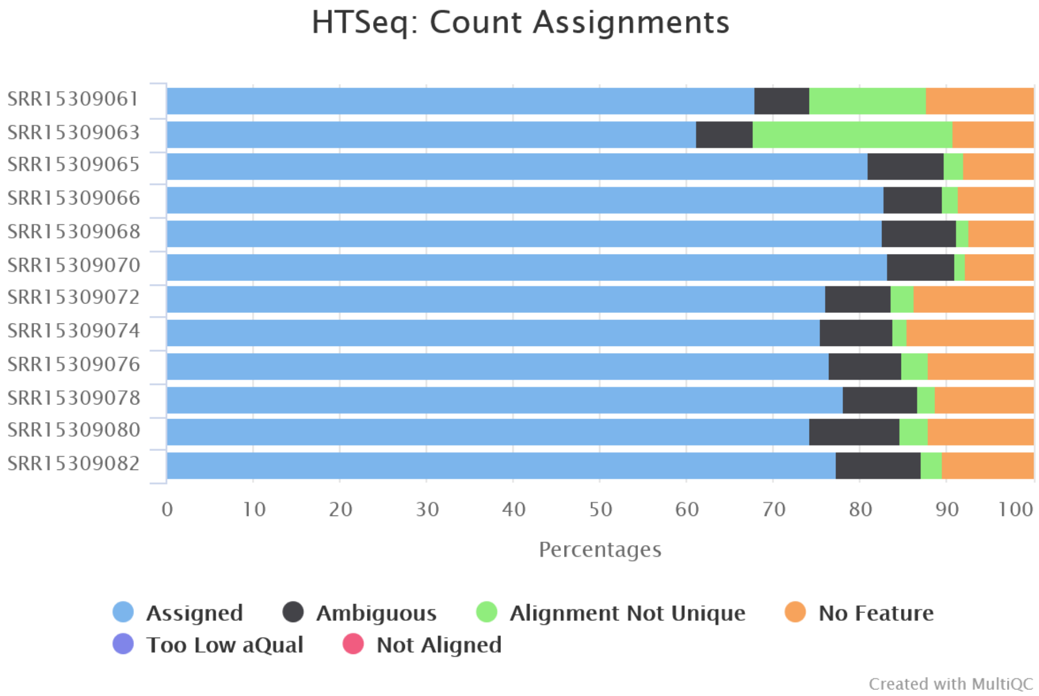


Figure 3: HTSeq assignment plot in percentages.

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)

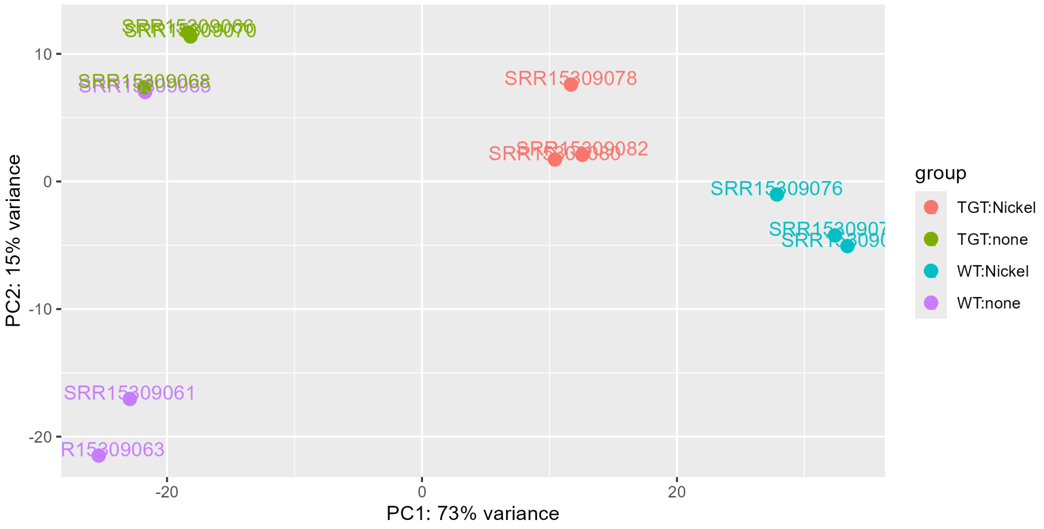


Figure 4: Clustering of samples PCA dot plot.

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

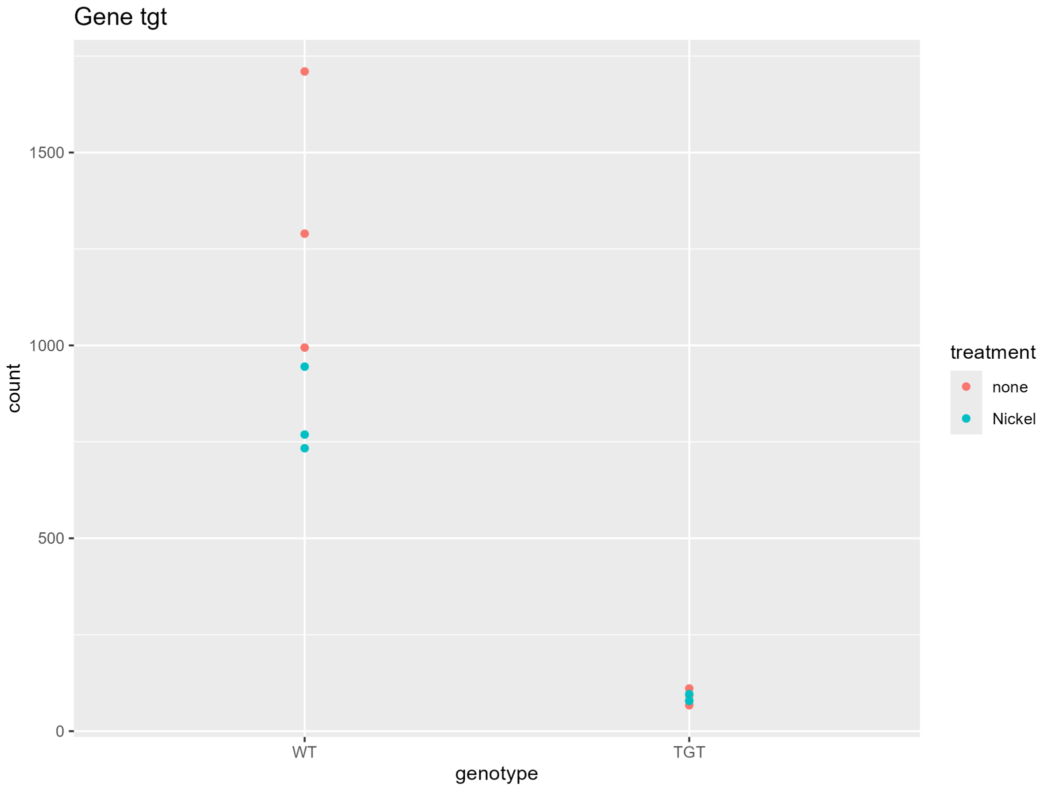


Figure 5: Viewing counts for a single geneID in a dot plot.

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)
```

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

head(annotatedRawCounts)
```

```
## # A tibble: 6 x 15
##   ID    product              gene  SRR15309076 SRR15309078 SRR15309074 SRR15309066
##   <chr> <chr>                <chr>       <dbl>       <dbl>       <dbl>       <dbl>
## 1 b0001 thr operon leader~ thrL          113         432         119         107
## 2 b0002 fused aspartate k~ thrA         2910       27851        3646        2378
## 3 b0003 homoserine kinase  thrB         1274        6956        1497        1476
## 4 b0004 threonine synthase thrC         1814       10638        2256        2462
## 5 b0005 DUF2502 domain-co~ yaaX          124         216         171         124
## 6 b0006 peroxide stress r~ yaaA          405         443         530         741
## # i 8 more variables: SRR15309082 <dbl>, SRR15309080 <dbl>, SRR15309070 <dbl>,
## #   SRR15309065 <dbl>, SRR15309072 <dbl>, SRR15309068 <dbl>, SRR15309063 <dbl>,
## #   SRR15309061 <dbl>
```

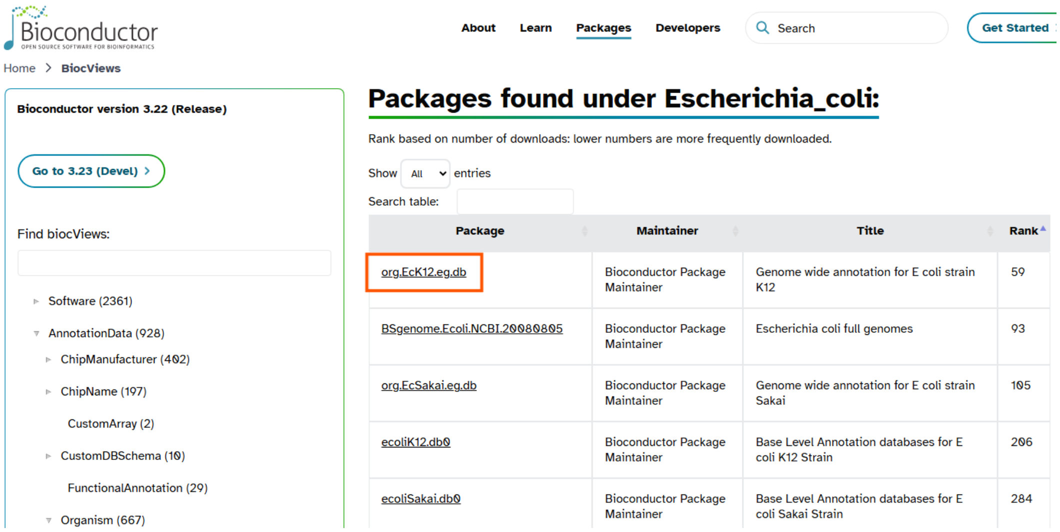
```
annotatedRawCounts <- annotatedRawCounts  %>%
  column_to_rownames(var = "ID")

### 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?



Bioconductor version 3.22 (Release)

Go to 3.23 (Devel) >

Find biocViews:

- Software (2361)
- AnnotationData (928)
  - ChipManufacturer (402)
  - ChipName (197)
  - CustomArray (2)
  - CustomDBSchema (10)
  - FunctionalAnnotation (29)
  - Organism (667)

Packages found under Escherichia_coli:

Rank based on number of downloads: lower numbers are more frequently downloaded.

Show All entries

Search table:

Package	Maintainer	Title	Rank
org.EcK12.eg.db	Bioconductor Package Maintainer	Genome wide annotation for E coli strain K12	59
BSgenome.Ecoli.NCBI.20080805	Bioconductor Package Maintainer	Escherichia coli full genomes	93
org.EcSakai.eg.db	Bioconductor Package Maintainer	Genome wide annotation for E coli strain Sakai	105
ecoliK12.db0	Bioconductor Package Maintainer	Base Level Annotation databases for E coli K12 Strain	206
ecoliSakai.db0	Bioconductor Package Maintainer	Base Level Annotation databases for E coli Sakai Strain	284

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.

```
#gene names/identifiers
genes <- annotations$gene
head(genes,10)
```

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

```
#View all available keytypes
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"
```

```
#make sure our IDs are contained in the keys
#prints the number of genes that are found within the keys
sum(genes %in% keys)
```

```
## [1] 4274
```

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?

```
# filter our results by padj and log2FoldChange
res_up <- dplyr::filter(DESeq2ResultsDF, padj < 0.05 & log2FoldChange > 1)
res_down <- dplyr::filter(DESeq2ResultsDF, padj < 0.05 & log2FoldChange < -1)

# extract gene IDs for upregulated genes
genes_up_id <- res_up$ID
```

```
genes_up <- annotations[genes_up_id,"gene"]

# extract gene names for downregulated genes
genes_down_id <- res_down$ID
genes_down <- annotations[genes_down_id,"gene"]

# append both lists to get all deregulated genes
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)
```

Table 1: Overrepresented BP terms - All DE genes

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

```
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] 100
```

```
print(n_BP_up)
```

```
## [1] 59
```

```
print(n_BP_down)
```

```
## [1] 93
```

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.

#Vizualization

Table 2: Overrepresented BP terms - Upregulated genes

	Description	p.adjust	Count
GO:0046377	colanic acid metabolic process	0.00e+00	18
GO:0009242	colanic acid biosynthetic process	2.00e-07	15
GO:0072348	sulfur compound transport	4.00e-07	23
GO:0006857	oligopeptide transport	2.20e-06	21
GO:0042938	dipeptide transport	3.42e-05	15

Table 3: Overrepresented BP terms - Downregulated genes

	Description	p.adjust	Count
GO:0006935	chemotaxis	0	32
GO:0042330	taxis	0	32
GO:0001539	cilium or flagellum-dependent cell motility	0	42
GO:0071973	bacterial-type flagellum-dependent cell motility	0	42
GO:0097588	archaeal or bacterial-type flagellum-dependent cell motility	0	42

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] M. Love, S. Anders, and W. Huber, *DESeq2: Differential gene expression analysis based on the negative binomial distribution.* 2025. doi: 10.18129/B9.bioc.DESeq2.

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

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

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

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

[7] 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.

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

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

[10] 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.

[11] 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

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

[13] 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.