
TH Bingen Technical University of Applied Sciences
Department 2 - Technology, Informatics and Economics
Applied Bioinformatics (B.Sc.)

# Gene Expression Analyses on yeast heat shock experiments

Prüfungsleistung DATR Mining with R
Abgegeben am: ???
Name: Robin Ender Matrikelnummer: 2184737

# List of Figures

# List of Tables

# Contents

# 1 Introduction

Heat stress poses a significant challenge to organisms by damaging cellular structures and functions. In response, cells activate the heat shock response, an ancient and conserved mechanism involving the production of heat shock proteins. These molecular chaperones prevent protein aggregation, assist in protein refolding, and restore cellular homeostasis. Triggered by protein unfolding rather than direct temperature sensing, this response mitigates widespread cellular disruptions, including cytoskeletal damage, organelle disorganization, and ATP depletion. [1]
This work focuses on the analyses of gene expression data provided by a yeast heat shock experiment that concentrated on the binding protein Mip6. The researchers that conducted this experiment tested the heat shock response on a single culture of *Saccharomyces cerevisiaes* that was split into three groups. The control group, that was maintained at 30°C and two groups that were incubated at 39°C for 20 (Condition 1) and 120 (Condition 2) minutes. The RNA-seq data gathered in this experiment will be analysed here. [2]
In our analyses we will cover the genetic response of yeast to heat shock by determining which classes of genes were up- and down-regulated. We will also invest if there are clusters of genes that show a similar response and examine their general function using the R package WGCNA. We will also compare the results of the tools Kallisto and Bowtie2 used to create the alignments as well as the results for differential gene expression created by EdgeR and Limma.

# 2 Material

As described in the introduction, we used existing data from a heat shock experiment. The origin of this data as well as of the reference data is shown in table 1.

Table 1: Data used in this work.

| Description | Database | Identifier | Version or Date created |
|---|---|---|---|
| RNA-seq data | NCBI GEO | GSE135568 | Mar 09, 2020 |
| Reference genome | SGD | S288C | version=R64-4-1 2024-05-29 |
| Gene Ontology | UniProt | GO | 2024-12-15 |

# 3 Methods

The analysis was conducted on the BioServer of the TH Bingen running Ubuntu 22.04.2 (x86) as well as on a private laptop running MacOS 15.1.1 (arm).

## 3.1 Programs

The programs and versions listet in table 2 were used for the analysis.

Table 2: Programs used for the analysis

| Program | Version | Operating System |
|---|---|---|
| bash | 5.1.16 | Ubuntu |
| bowtie2 | 2.4.4 | Ubuntu |
| fastqc | 0.11.9 | Ubuntu |
| Stringtie | 2.2.1 | Ubuntu |
| gffread | 0.12.8 | Ubuntu |
| kallisto | 0.46.2 | Ubuntu |
| R | 4.4.0 | Ubuntu |
| Samtools | 1.13 | Ubuntu |
| slurm-wlm | 21.08.5 | Ubuntu |
| Trimmomatic | 0.39 | Ubuntu |
| bash | 5.2.37 | MacOS |
| ggVennDiagram | 1.5.2 | MacOS |
| ggplot2 | 3.5.1 | MacOS |
| wordcloud | 2.6 | MacOS |
| WGCNA | 1.73 | MacOS |
| tximport | 1.34.0 | MacOS |
| stringr | 1.5.1 | MacOS |
| edgeR | 4.4.1 | MacOS |
| RColorBrewer | 1.1-3 | MacOS |
| limma | 3.62.2 | MacOS |
| tidyverse | 2.0.0 | MacOS |
| corrplot | 0.95 | MacOS |
| rtracklayer | 1.66.0 | MacOS |
| tidyverse | 2.0.0 | MacOS |

## 3.2 Reference Transcriptome

To create a reference transcriptome we used gffread on our reference genome. We calculated the number of generated reference transcriptomes, checked if all transcripts in the gff file have been translated to reference transcripts in the output file and analysed the number of splice variants for each gene.

## 3.3 Quality Filtering

To ensure that only reads with a sufficient quality were used for our analysis we used Trimmomatic to remove adapter sequences, low-quality bases, and other contaminants. We calculated the number of genes filtered out and compared the quality of the reads before and after this step with FastQC.

## 3.4 Gene Expression Quantification

We quantified the gene expression using Kallisto and Bowtie2. Samtools was used to create a binary file from the result file of bowtie2. Stringtie was used to generate the gene counts for our bowtie2 results. For Kallisto we calculated the total number of expressed genes as well as the percentage of expressed genes relative to the total in the transcriptome. Furthermore, we also analysed some basic statistics and calculated the Correlation between the samples. For bowtie2 we also calculated the percentage of aligned reads as well as analysed the alignment quality statistics. For both methods we extracted the TPM values on a gene basis and compared them. We then calculated the log fold change for the TPM values and also compared them between the methods.

## 3.5 Weighted Gene Co-expression Network Analysis

Using the results of the Kallisto gene expression quantification we performed a Weighted Gene Co-expression Network Analysis to construct a gene network and identify modules of high correlation. We also identified gene modules showing significant correlations with heat-shock conditions and created Wordclouds using the Gene Ontology Term Annotations.

## 3.6 Differential Gene Expression

For the differential gene expression analysis we also used the results of Kallisto. We conducted this analysis with edgeR as well as Limma with the voom normalisation. We identified increased and decreased genes in these differentially expressed gene and again created word clouds. We higlighted the three most extremely changed genes in bar charts.

# 4 Results

## 4.1 Reference Transcriptome

From the 6585 genes of *Saccharomyces cerevisiae*, 11599 transcripts have been generated. Of these 6585 genes, 2575 had one splice variant and 4512 had two.

## 4.2 Quality Filtering

Only a small number of reads were filtered out (up to $\sim$ 1 percent). The quality improvement was not visible in the data provided by FastQC.

## 4.3 Gene Expression Quantification

### 4.3.1 Kallisto

For each replicate around ten thousand (mean of 10622) of the 11599 trancripts have been found. This corresponds to over 90 percent. The TPM values between the samples were comparable, with. The control condition had slightly lower median values (between 9 and 12) compared two condition 1 (between 15 and 16) and condition 2 (between 14 and 18). They ranged between 0 and 86654.40. To visualize them the logarithm of the TPM value plus one was used, see figure 1.

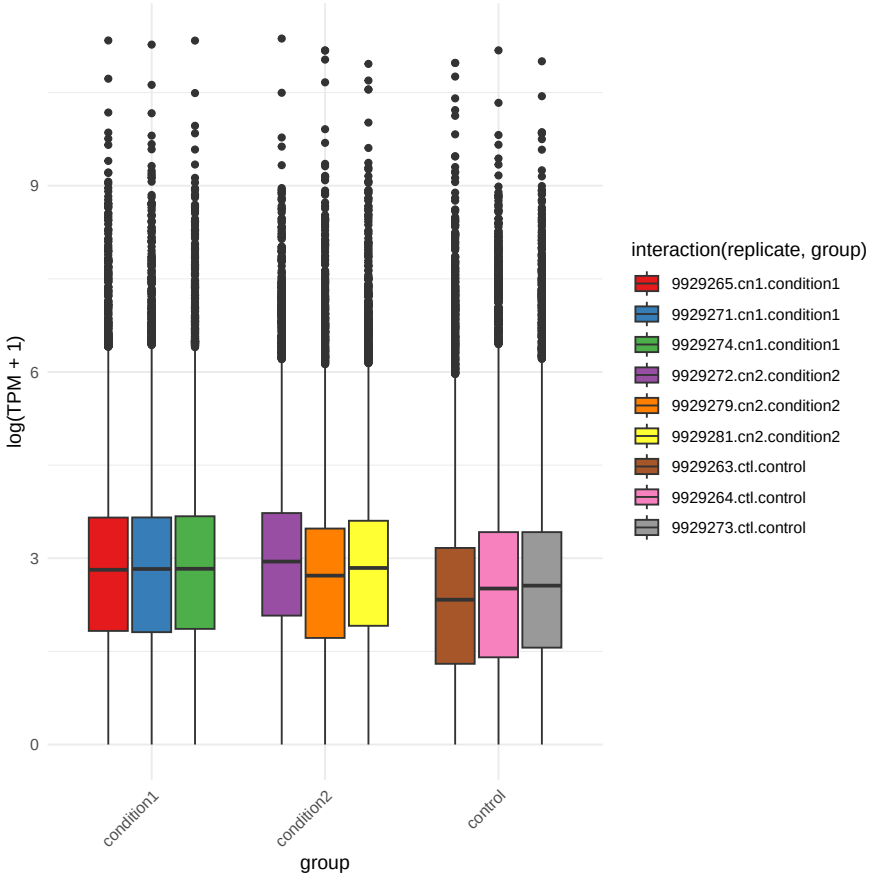



Figure 1: Box plots of the TPM values for the Kallisto analysis

The correlation between the technical and biological replicates of the Kallisto alignment was calculated and visualized in figure 2. The samples of condition 1 showed a high correlation between each other. The replicates for the control condition as well as for condition 2 differed more in comparison.

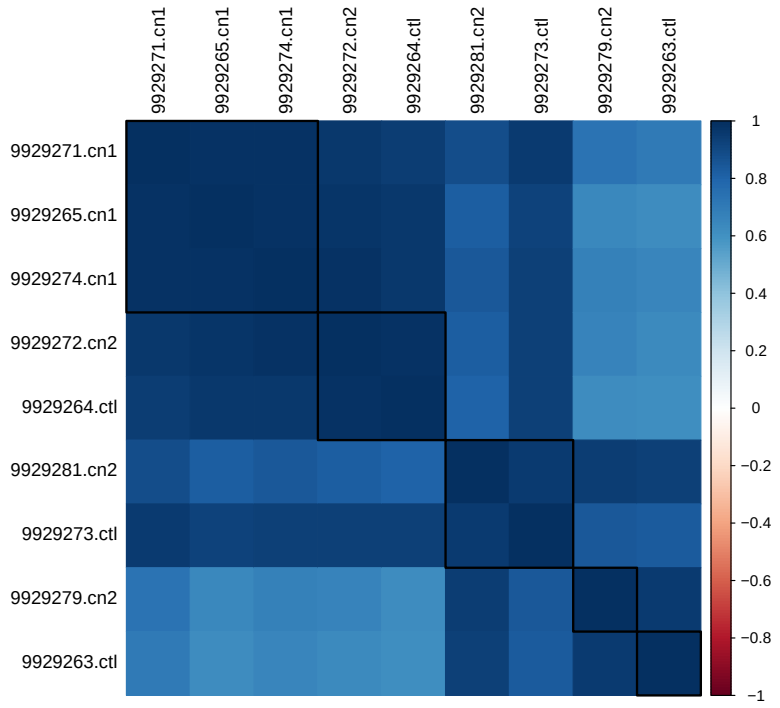



Figure 2: Correlation plot for the technical and biological replicates of the Kallisto analysis

### 4.3.2 Bowtie2

For the Bowtie2 alignment we analysed how many reads were mapped once, multiple times or were not aligned at all. All replicates showed similar ratios. The mapping of the reads of Bowtie2 is shown in figure 3

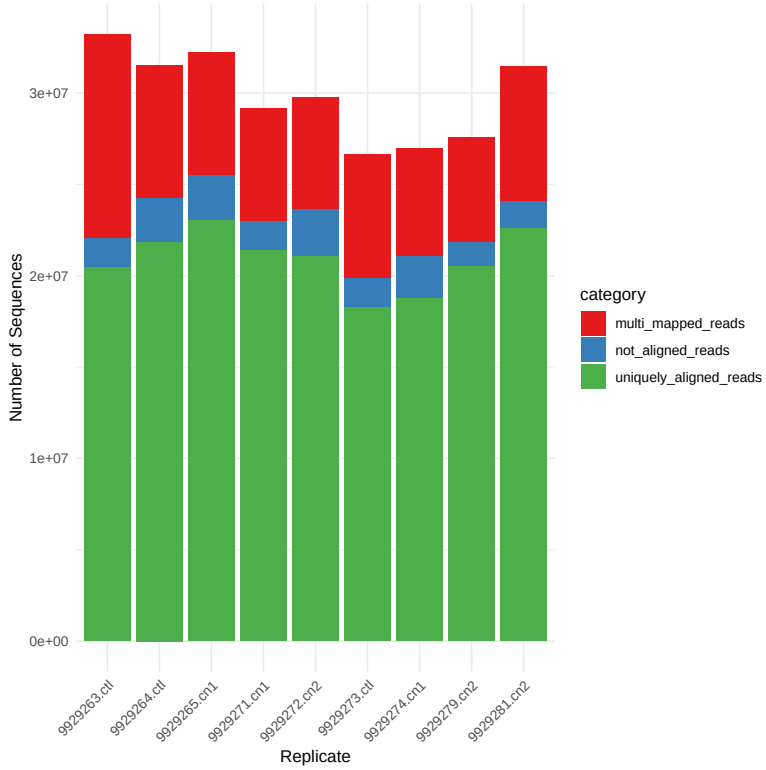



Figure 3: Mapping of the reads of the Bowtie2 alignment

### 4.3.3 Comparison Kallisto and Bowtie2

The distribution of the TPM values obtained by Kallisto and Bowtie2 were similar. The Bowtie2 results also showed slightly lower median for the control group. For both programs and for both conditions the log fold changes have been calculated and the distribution was plotted in figure 4. The distribution of TPM values looks comparable between the programs. S shift to the positive log fold changes can be seen.

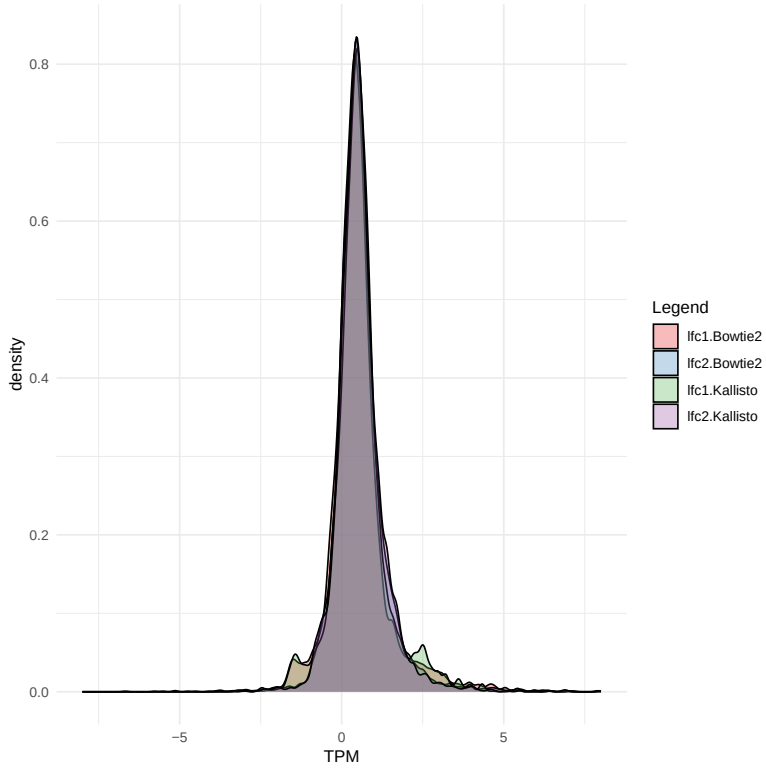



Figure 4: Grouped box plots of the three biological replicates per group

## 4.4 Weighted Gene Co-expression Network Analysis

The word cloud generated for the genes in the significant module is shown in figure 5

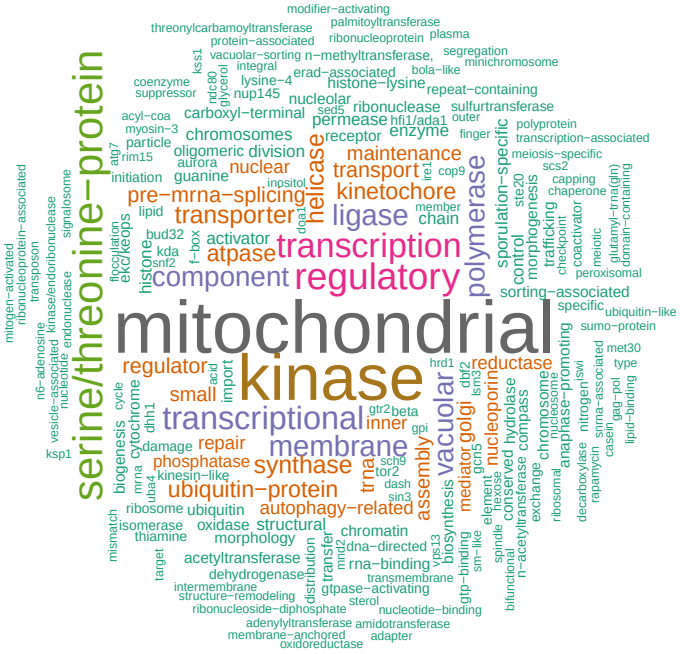
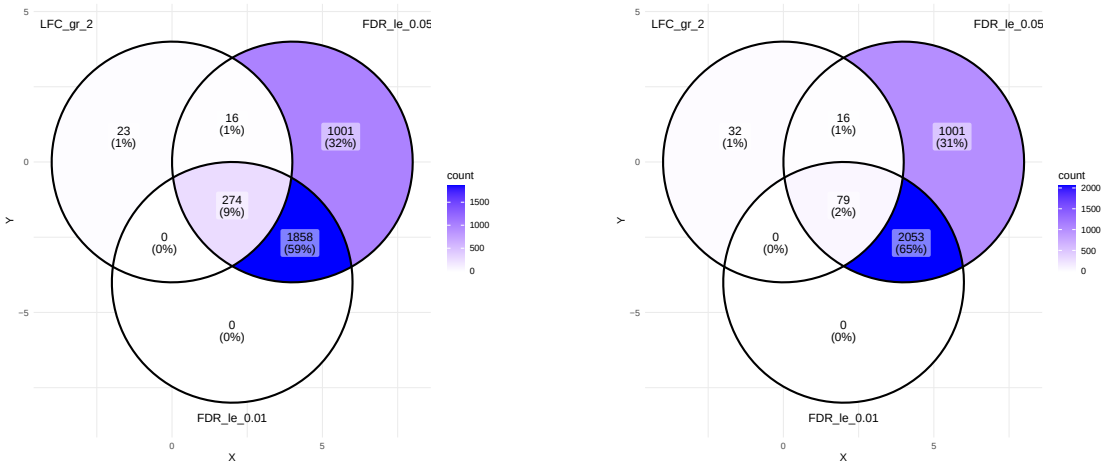


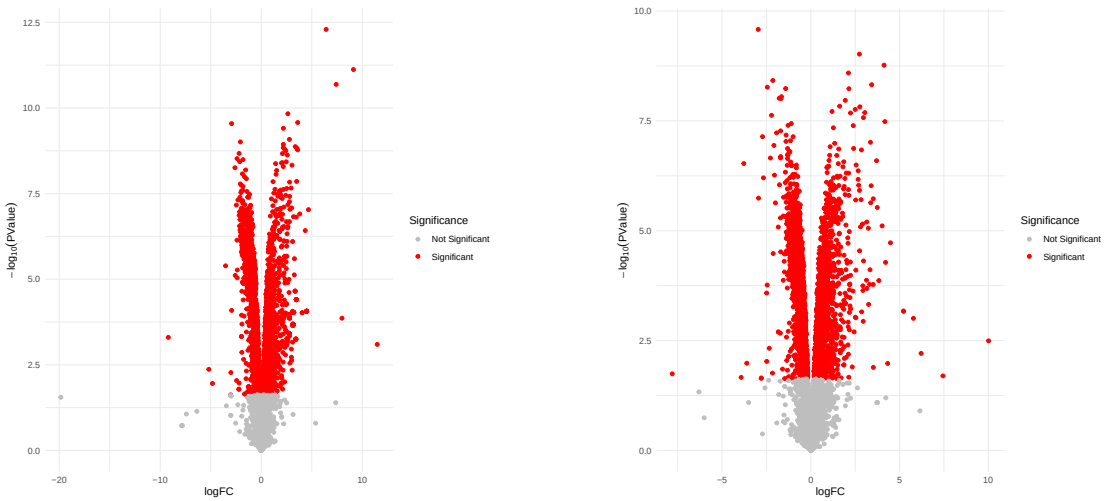

Figure 5: Word Cloud for the Module Eigengen 2

## 4.5 Differential Gene Expression




(a) (b)

Figure 6: Caption for the entire figure.




(a) (b)

Figure 7: Caption for the entire figure.

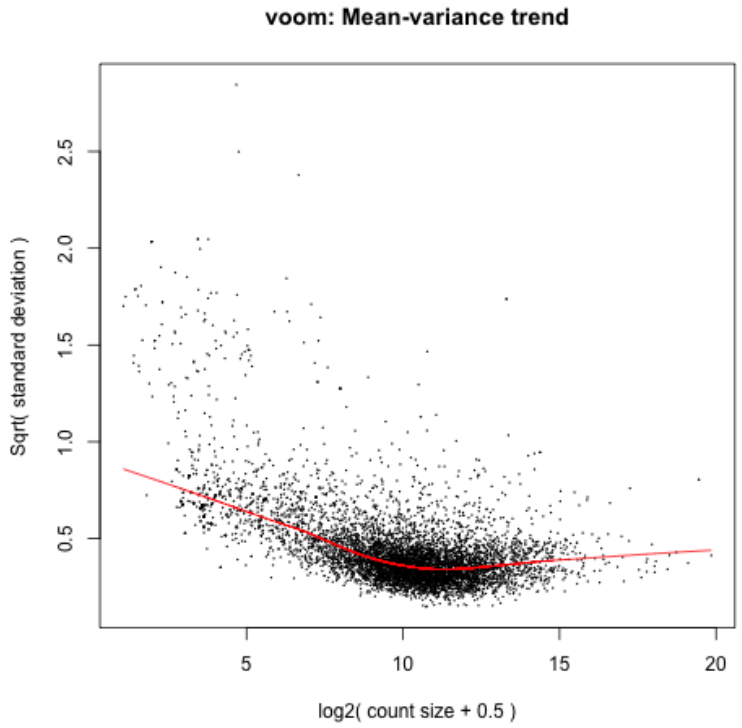
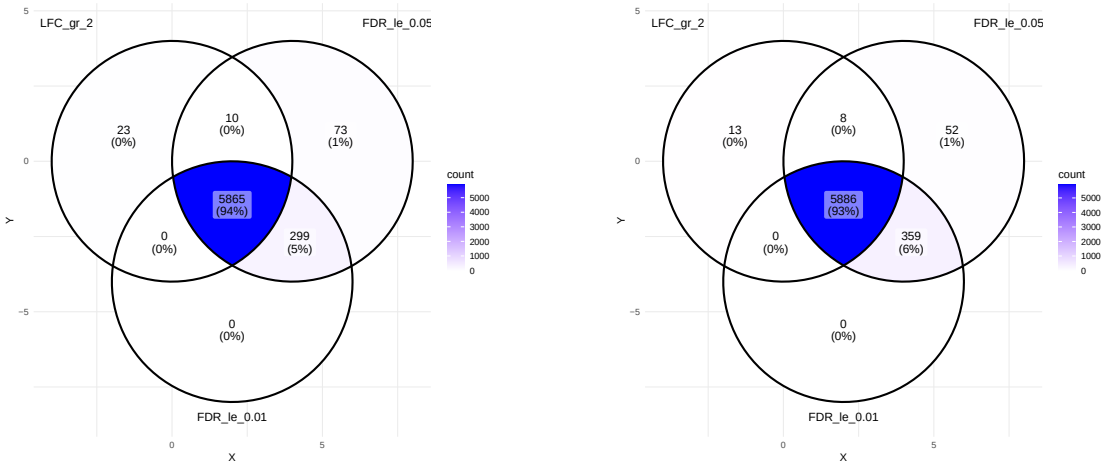
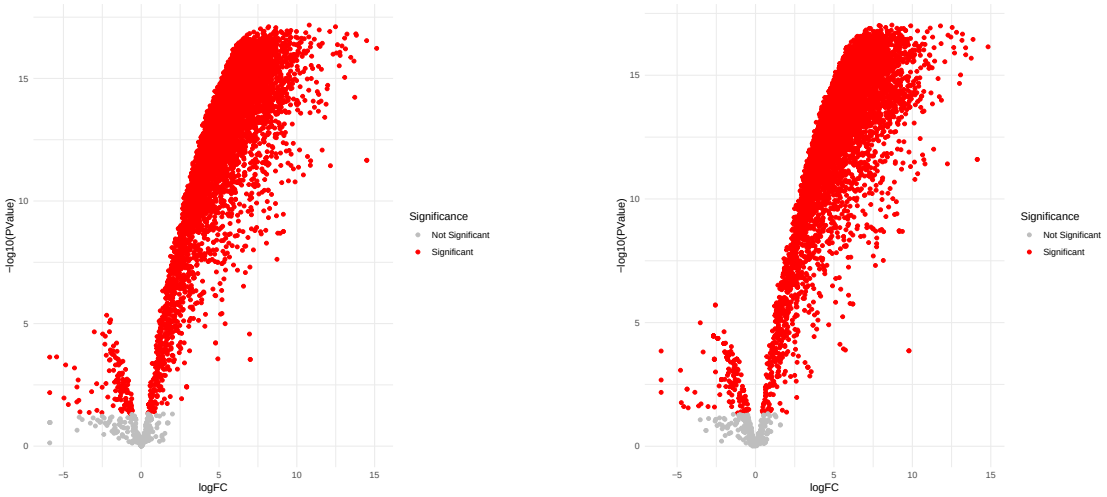



Figure 8: Voom plot



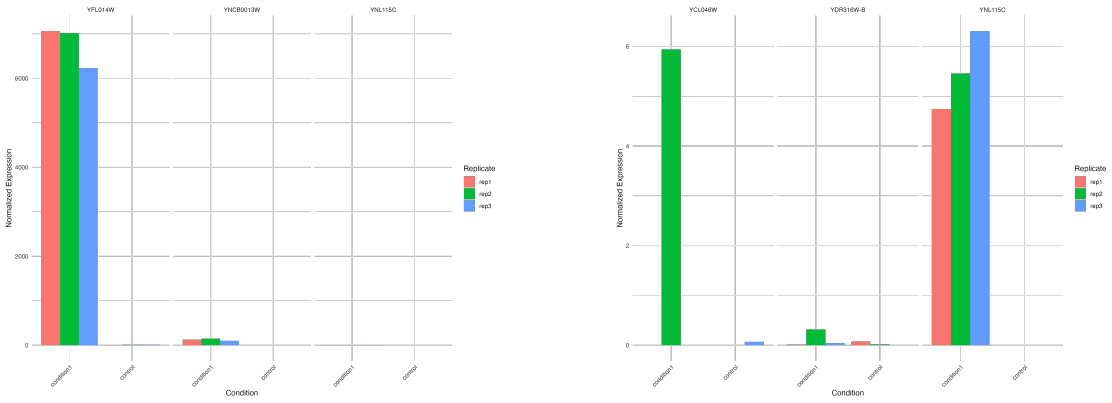

(a) (b)

Figure 9: Caption for the entire figure.




(a) (b)

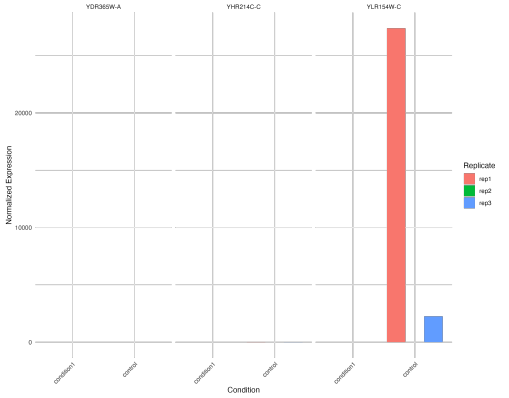
Figure 10: Caption for the entire figure.



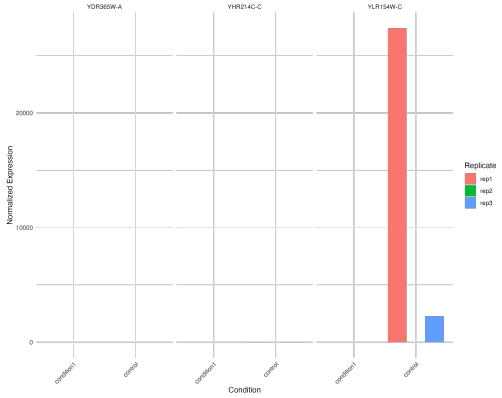

(a)




(b)




(c)

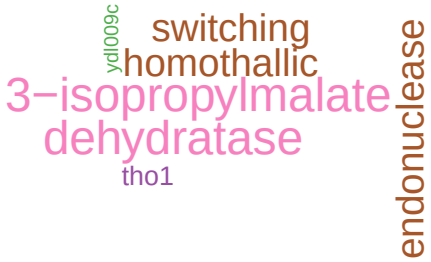



(d)

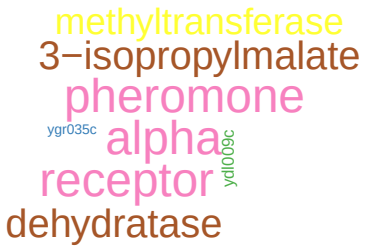
Figure 11: Results from the four experiments: (a) First experiment, (b) Second experiment, (c) Third experiment, (d) Fourth experiment.


(a)


(b)

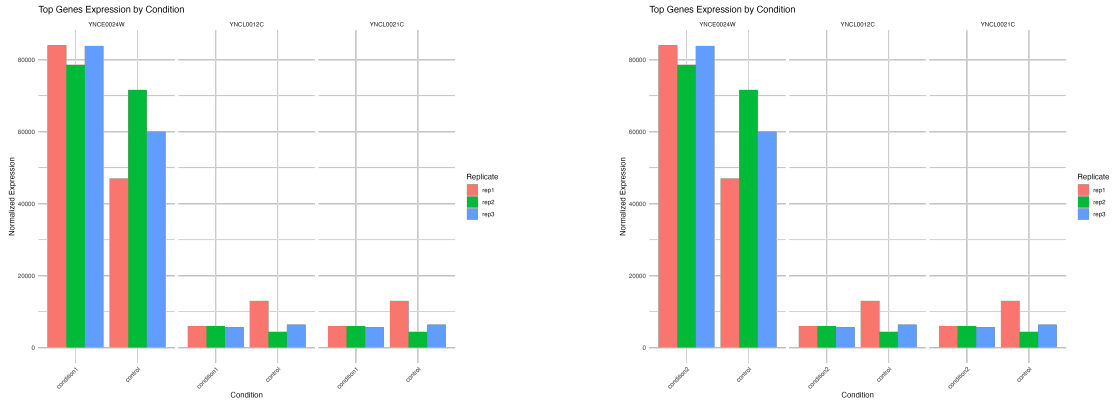



(c)




(d)

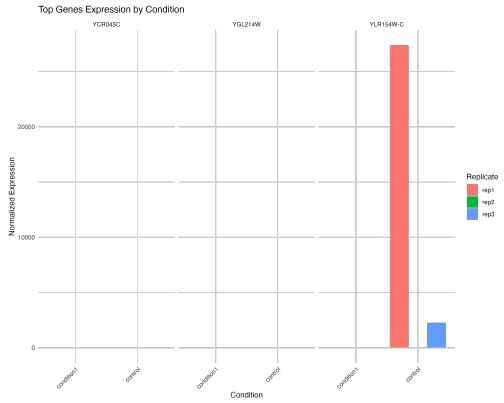
Figure 12: Results from the four experiments: (a) First experiment, (b) Second experiment, (c) Third experiment, (d) Fourth experiment.



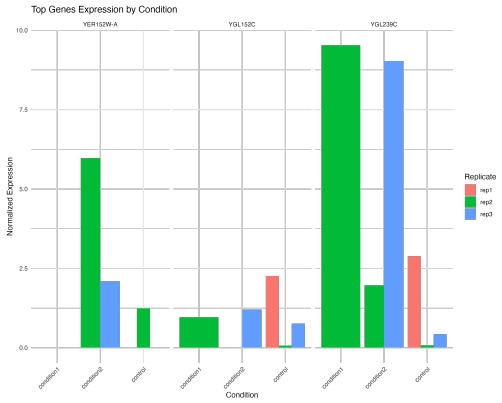

(a)




(b)




(c)

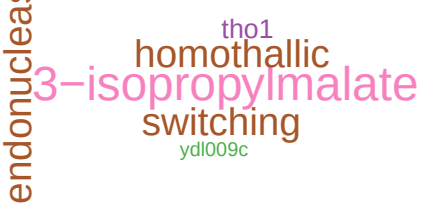



(d)

Figure 13: Results from the four experiments: (a) First experiment, (b) Second experiment, (c) Third experiment, (d) Fourth experiment.


(a)


(b)




(c)


(d)

Figure 14: Results from the four experiments: (a) First experiment, (b) Second experiment, (c) Third experiment, (d) Fourth experiment.

# 5 Discussion

# Bibliography

[1] Klaus Richter, Martin Haslbeck, and Johannes Buchner. "The Heat Shock Response: Life on the Verge of Death". In: *Molecular Cell* 40.2 (Oct. 22, 2010). Publisher: Elsevier, pp. 253–266. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2010.10.006. URL: https://www.cell.com/molecular-cell/abstract/S1097-2765(10)00782-3 (visited on 01/19/2025).

[2] Carme Nuño-Cabanes et al. "A multi-omics dataset of heat-shock response in the yeast RNA binding protein Mip6". In: *Scientific Data* 7.1 (Feb. 27, 2020), p. 69. ISSN: 2052-4463. DOI: 10.1038/s41597-020-0412-z. URL: https://www.nature.com/articles/s41597-020-0412-z (visited on 01/19/2025).

**Erklärung zur Originalität der Arbeit**

Hiermit bestätige ich, dass die abgegebene Arbeit das Original ist und von mir ohne weitere Hilfe geschrieben wurde. Wenn Arbeit anderer referenziert oder genutzt wurde, wurde dies angemessen kenntlich gemacht. Meine Arbeit wurde noch nicht bewertet oder veröffentlicht. Die elektronisch abgegebene Version stimmt mit der elektronischen überein.

| | |
|---|---|
| Unterschrift | Ort und Datum |

**Erklärung zum Eigentum und Urheberrecht**

Hiermit erkläre ich meine Zustimmung, dass die Technische Hochschule Bingen diese Arbeit anderen Studierenden und interessierten Dritten zur Verfügung stellen und in meinem Namen (Robin Ender) veröffentlichen darf.

| | |
|---|---|
| Unterschrift | Ort und Datum |