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: Data Mining with R
Abgegeben am: 11.02.2025
Name: Robin Ender
Matrikelnummer: 2184737

List of Figures

List of Tables

Contents

Acronyms

TPM Transcripts Per Million. 4, 5, 8

WGCNA Weighted Gene Co-expression Network Analysis. 4, 10, 13

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 investigate, 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
R	4.4.2	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

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 Transcripts Per Million (TPM) values on a per-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 WGCNA

Using the results of the Kallisto gene expression quantification we performed a Weighted Gene Co-expression Network Analysis (WGCNA) 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 a Wordcloud 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.

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. The control condition had slightly lower median values (between 9 and 12) compared to condition 1 (between 15 and 16) and condition 2 (between 14 and 18). They ranged between 0 and 86654.40. To visualize them the $log(TPM + 1)$ was used, see Figure 1.

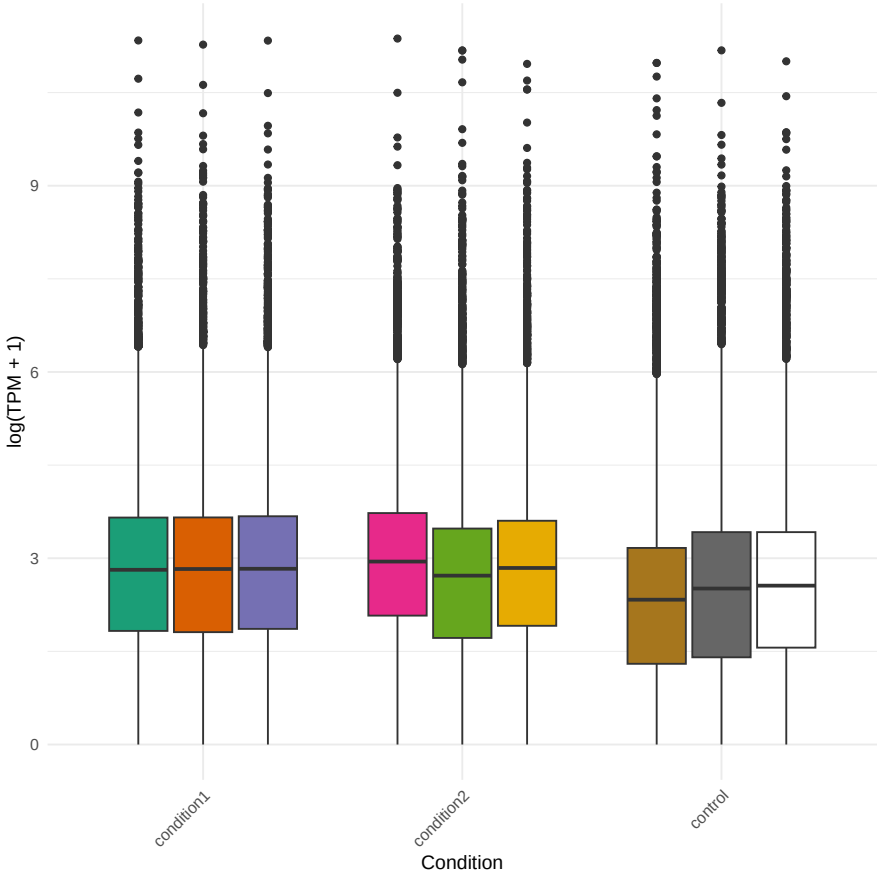


Figure 1: Box plots of the TPM values for the three replicates ot the Kallisto analysis.

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 and did not form clusters.

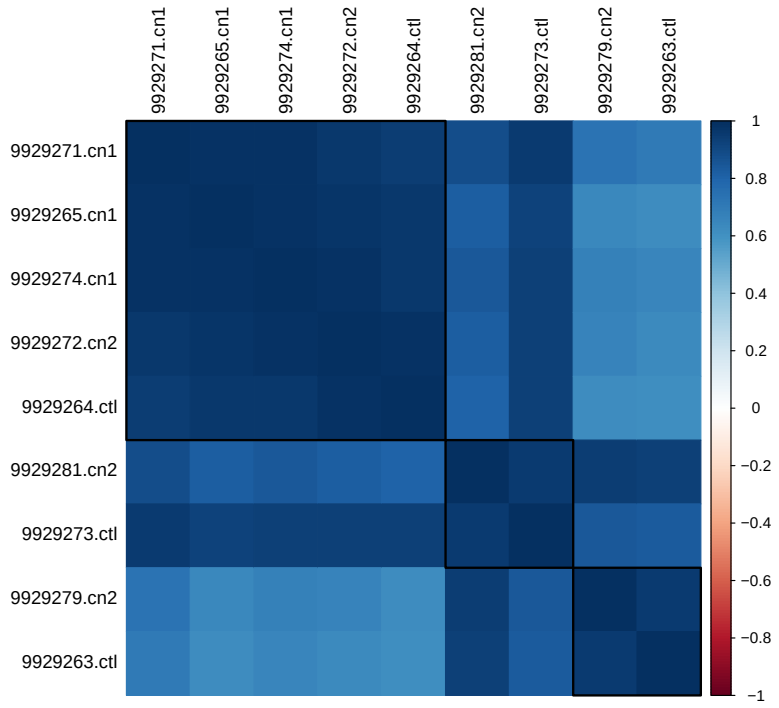


Figure 2: Correlation plot for the technical and biological replicates of the Kallisto analysis. Clusters are shown by black borders.

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

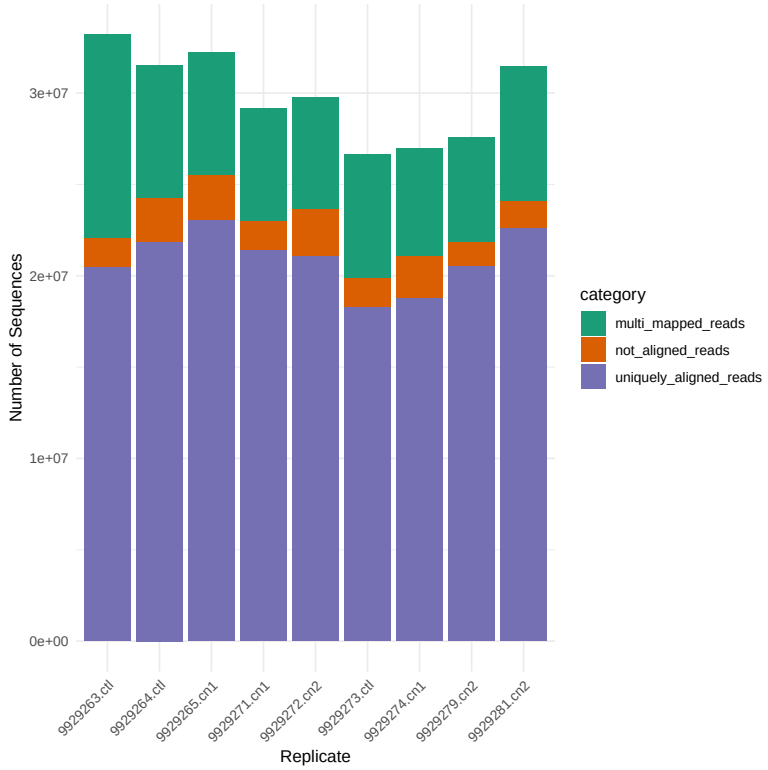


Figure 3: Mapping of the reads of the Bowtie2 alignment.

4.3.3 Comparison Kallisto and Bowtie2

The distributions of the TPM values obtained by Kallisto and Bowtie2 were similar. The Bowtie2 results also showed a 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. A shift towards positive log-fold changes can be seen.

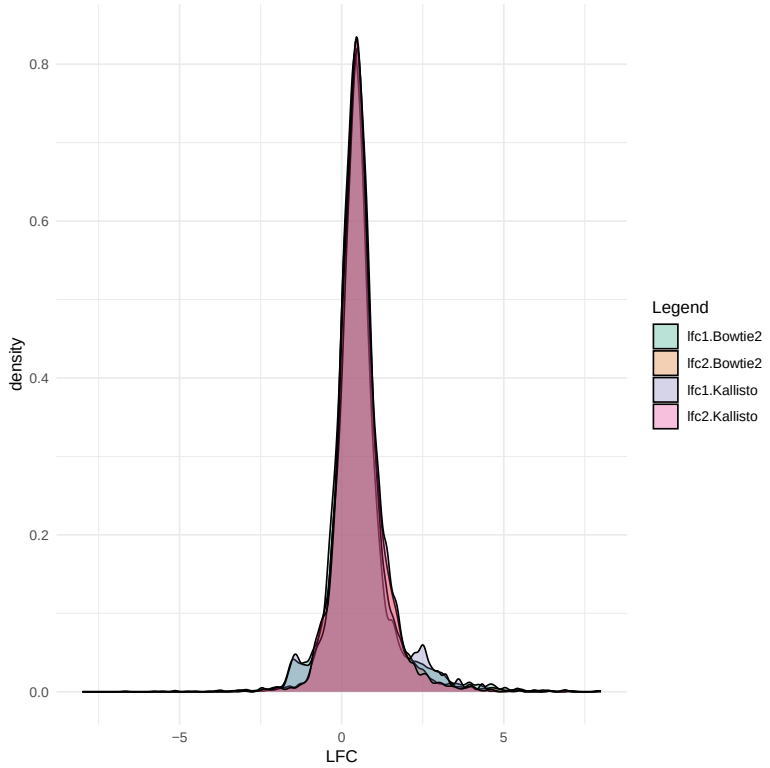


Figure 4: Grouped box plots of the three biological replicates per condition and per program.

To see if Kallisto and Bowtie2 identified the same genes as significantly changed, we calculated the Jaccard similarity coefficient for genes with an absolute log-fold change > 2 for condition 1 (0.60) and for condition 2 (0.40).

4.4 WGCNA

During the WGCNA one significant module eigenegene was identified. To see which genes are found in this module a word cloud using GO annotations was created, see Figure 5.

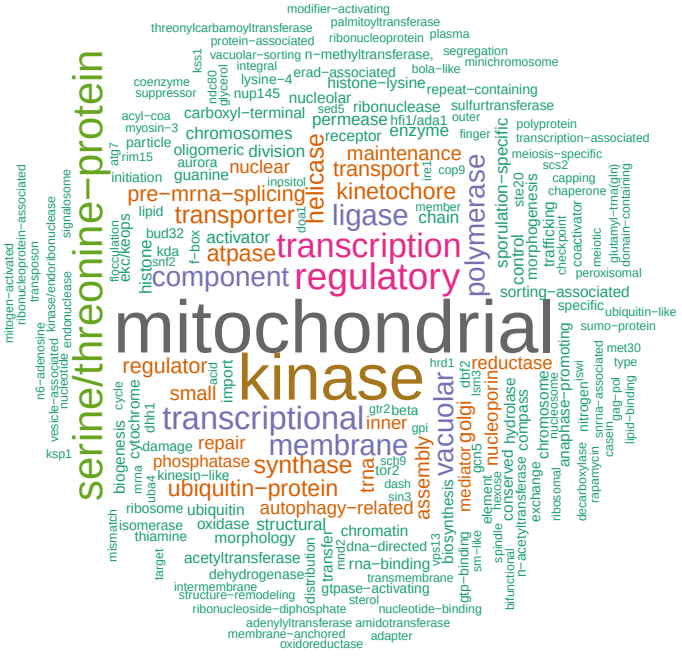
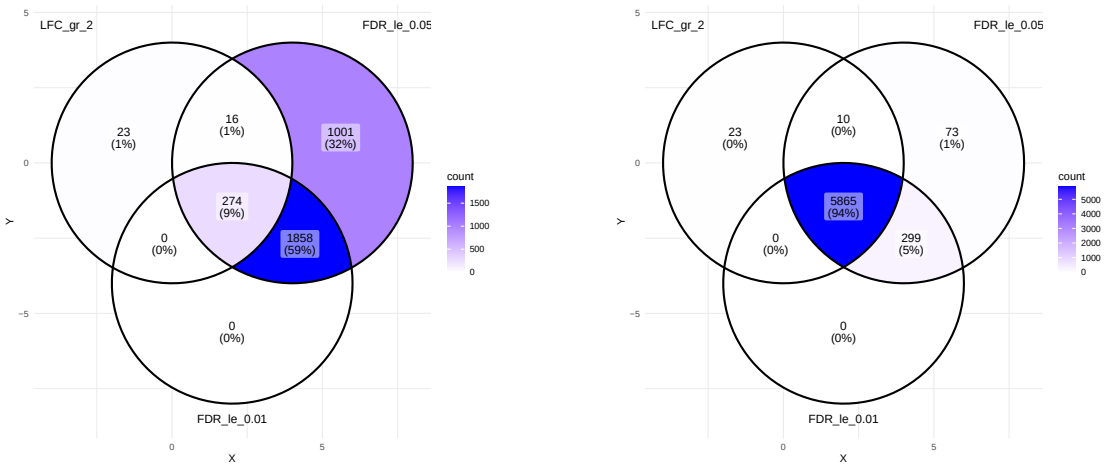


Figure 5: Word Cloud for the Module Eigengen 2.

4.5 Differential Gene Expression

4.5.1 Comparing edgeR and Limma

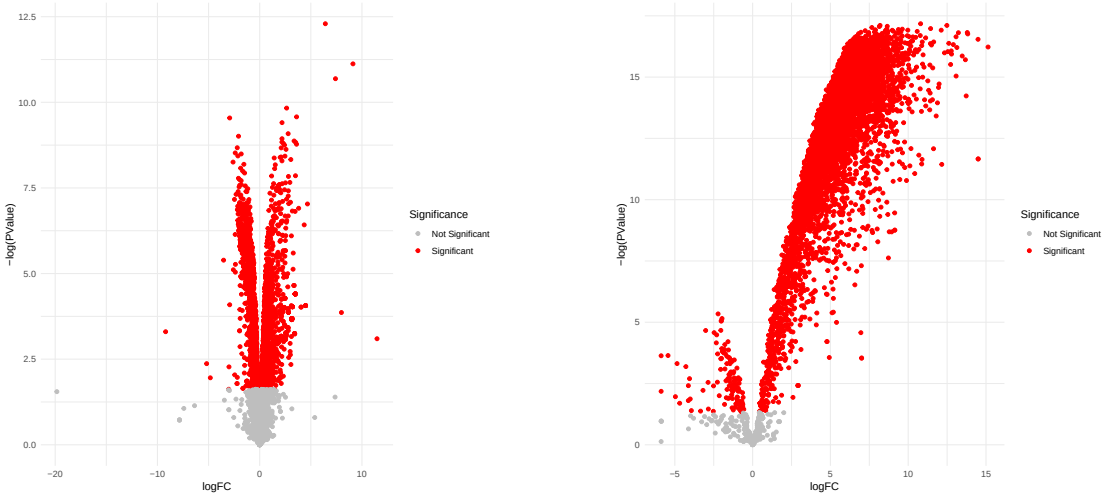
Venn diagrams were created showing the relation of the log-fold change and the adjusted p-values. As seen in Figure 6, edgeR and Limma behaved differently. While in edgeRs results only a few genes show a log-fold change greater than 2, almost all genes in Limmas results do. The Limma genes also seem to have lower adjusted p-values.



(a) (b)

Figure 6: Venn diagrams displaying relations between the adjusted p-value (FDR) and the log-fold change (LFC) for the edgeR results (a) and Limma results (b) for condition 1.

Viewing the volcano plots edgeR shows a relatively symmetric log-fold change, with mostly low values. The volcano plot for the Limma results on the other hand shows a clear tendency towards positive and higher log-fold changes.



(a) (b)

Figure 7: Volcano plots visualizing the log-fold change and significance for each gene for edgeR (a) and Limma (b) for condition 1.

Jaccard similarity between edgeR and limma was calculated for the top twenty up and

down regulated genes. No overlap was found for the up regulated genes. However, this was not the case for the down regulated genes. For condition 1, a Jaccard similarity coefficient of 0.48 was calculated, for condition 2, a Jaccard similarity coefficient of 0.18 was calculated.

4.5.2 Differentially expressed genes

For the top twenty up and down regulated genes word clouds were generated. Both edgeR and limma show an up regulation of proteins associated with heat shock, limma also shows an up regulation of chaperone associated proteins. For the down regulated genes the words 3-isopropylmalate and dehydratase were often found.
Using the descriptions provided by the *Saccharomyces* Genome Database [3] we looked at the most up and down regulated genes. We found up regulation for genes related to plasma membrane organization during stress conditions (YFL014W), as well as genes that take part in the production of proteins like tRNAs (YNCB0013W) or rRNA(YNCL0012C, YNCL0021C). For down regulated genes we found genes associated with retrotransposons (YHR214C-C, YDR365W-A) and several dubious open reading frames (YER152W-A, YGL152C, YGL239C).

5 Discussion

5.1 Genetic Response to Heat Shock in Yeast

The yeast *Saccharomyces cerevisiae* responds to heat shock through significant changes in gene expression. Heat shock predominantly upregulated genes involved in protein folding and stress response, such as chaperones or components crucial for ribosomal activity, while downregulated genes were often associated with retrotransposons and certain dubious open reading frames. Up regulation generally occured more often and stronger.

5.2 Clusters of Genes and Their Functions

Weighted Gene WGCNA identified one significant gene module correlating with heat shock conditions. Using a word cloud it was possible to get a overview of functions of the genes within this cluster. Important keywords such as transcription, synthase or polymerase hint at a strong relation of this cluster with the creation of new proteins.

5.3 Comparison of Tools and Agreement of Results

Gene expression quantification using Kallisto and Bowtie2 yielded comparable TPM distributions, with both methods indicating higher expression levels under heat shock conditions. However, discrepancies arose in identifying significantly altered genes, as evidenced by Jaccard similarity coefficients (0.60 and 0.40 for conditions 1 and 2, respectively). Differential expression analysis also highlighted differences: Limma (with voom normalization) exhibited a broader detection of high log-fold changes compared to EdgeR, which identified fewer but more symmetric log-fold changes. The absent overlap in identified upregulated genes but moderate agreement for downregulated genes (Jaccard coefficients of 0.48 and 0.18 for conditions 1 and 2) underscores the variability introduced by different analytical pipelines.

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).

[3] *Saccharomyces Genome Database — SGD*. URL: https://www.yeastgenome.org/ (visited on 01/27/2025).

Supplements

The data analysis was performed before some changes to the exercise were made. To have comparable results, the result data was taken from another student, where the data analysis was done after all changes were made to the exercises:
/media/BioNAS/ag_hallab/DATR/Beck/

The original results for my analysis can be found here:
/media/BioNAS/ag_hallab/DATR/ender/.

The analysis of the data pipeline was conducted on the personal computer. To make the analysis available, they (scripts and results) are copied to the supplements folder of the zip file. They are also available on Github https://github.com/derRiesenOtter/datr_pl

The README.md file can also be found inside the supplements folder, or on Github.

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