Analysis of differential gene expression in wild and cultivated rice under drought stress

Ingo Giebel[1]

[1]Math.-Nat. Fakultät, Heinrich-Heine-Universität Düsseldorf
[1]QBio304: Applied Bioinformatics
[2]RG Network Analysis and Modelling, Leibniz Institute of Plant Genetics and Crop Plant Research
[2]IBG-4: Bioinformatik, Forschungszentrum Jülich GmbH
[1]Prof. Dr. Björn Usadel
[2]Dr. Jedrzej Jakub Szymanski

May 1, 2023

Contents

1 Abstract **3**

2 Introduction **3**

3 Materials and methods **3**
- 3.1 Selection of the RNA-seq data . . . 3
- 3.2 Quality evaluation . . . 4
- 3.3 RNA-seq preprocessing . . . 5
- 3.4 Transcripts abundances quantification . . . 5
- 3.5 Statistical evaluation and differential expression analysis . . . 6
  - 3.5.1 Import of the kallisto transcript-level estimates . . . 6
  - 3.5.2 Filtering and normalization . . . 7
  - 3.5.3 Hierarchical cluster analysis . . . 8
  - 3.5.4 Principal component analysis . . . 8
  - 3.5.5 Identification of differentially expressed genes . . . 10
- 3.6 Functional enrichment analysis . . . 10

4 Results **10**
- 4.1 Quality evaluation . . . 10
- 4.2 Reads mapped to the reference transcriptome . . . 11
- 4.3 Hierarchical cluster analysis . . . 11
- 4.4 Principal component analysis . . . 14
- 4.5 Differentially expressed genes . . . 14
- 4.6 Functional enrichment analysis . . . 18

5 Discussion **20**

6 Online content and license notice **21**

References **22**

Index **24**

List of Figures

1 Exemplary FastQC quality assessment of the per base sequence content - raw vs trimmed FASTQ file . . . 5
2 Exemplary FastQC quality assessment of the per base sequence quality - raw vs trimmed FASTQ file . . . 5
3 TPM statistics about the imported kallisto data . . . 7
4 Log2(CPM) distribution of the unfiltered, non-normalized data . . . 7
5 Log2(CPM) distribution of the filtered, non-normalized data . . . 8
6 Log2(CPM) distribution of the filtered, normalized data . . . 8
7 Log2(CPM) distribution of the filtered, normalized data in comparison with the non-normalized and the unfiltered data . . . 9
8 Trimmomatic preprocessing . . . 11
9 FastQC quality assessment of the preprocessed FASTQ files . . . 12
10 kallisto pseudoalignments . . . 12
11 Hierarchical cluster analysis of the *O. nivara* RNA-seq data . . . 13
12 Hierarchical cluster analysis of the *O. sativa* RNA-seq data . . . 13
13 PCA of the log2(CPM) data - *O. nivara* . . . 14
14 PCA of the log2(CPM) data - *O. sativa* . . . 15
15 Volcano plot of the DEGs - *O. nivara* . . . 15
16 Volcano plot of the DEGs - *O. sativa* . . . 16
17 Venn diagrams of the DEGs . . . 16
18 Heatmap of the DEGs - *O. nivara* . . . 17
19 Heatmap of the DEGs - *O. sativa* . . . 17
20 Manhattan plot with the first 10 top-ranked GO terms highlighted - *O. nivara* 19
21 Manhattan plot with the first 10 top-ranked GO terms highlighted - *O. sativa* 19

List of Tables

1 Number of genes with no reads, and number of genes with CPMs ≥ 1 . . . 7

1 Abstract

Rice is the most widely consumed staple food for a large part of the world's human population, and drought is the most imperative and major limitation for rice production in rainfed ecosystems. Therefore, there is a great requirement of rice varieties with drought tolerance, and much research is done to improve the drought tolerance of rice. The understanding of drought stress responses on the gene level may substantially contribute to this goal.

Here, the impact of drought stress on the gene expression of wild and cultivated rice is examined with the following statistical methods: hierarchical clustering analysis (HCA), principal component analysis (PCA), functional enrichment analysis. The analyzed RNA-seq data was publicly submitted by the Institute of Botany, Chinese Academy of Sciences, in January 2021.

This analysis reveals that both the wild and the cultivated rice species react strongly to drought stress by down- and up-regulating their gene activity. The hierarchical cluster analysis of the two different wild rice cultivars further uncovers that the examined cultivar is an important confounding factor. When considering the ontology of the down- and up-regulated genes, both rice species largely coincide with respect to the significantly enriched GO terms. Another result is that on the gene level the impact of drought stress may be best described and categorized in terms of biological processes (BP) as opposed to the molecular functions (MF) of the genes and as opposed to the cellular components (CC) in which the gene products are physically located.

2 Introduction

RNA-sequencing is used to analyze the transcriptome, indicating which of the genes encoded in the DNA are turned on or off and to what extent. This study examines the impact of drought stress on the gene expression of wild and and cultivated rice, using statistical methods. To this end, publicly available RNA-seq data from the Institute of Botany, Chinese Academy of Sciences, submitted on January 1st, 2021, is analyzed with respect to differential gene expression. This data allows for a direct comparison of normal and drought stress conditions for the two closely related species of wild and cultivated rice.

3 Materials and methods

3.1 Selection of the RNA-seq data

This study uses publicly available paired-end RNA-seq data of wild and cultivated rice, submitted in January 1, 2021 by the Institute of Botany, Chinese Academy of Sciences. This data allows to compare rice grown under normal conditions with rice grown under

drought stress conditions. Furthermore, the data allows for an interspecies comparison of wild rice (*Oryza nivara*, cultivars BJ278 and BJ89) with cultivated rice (*Oryza sativa*, cultivar Nipponbare).

All samples were uniformly taken from seedlings (leaf tissue) at the age of twelve days. Used sequencing platform: Illumina HiSeq 2000.

All this makes the data well suited for a targeted analysis of drought stress responses.

3.2 Quality evaluation

The quality of the raw and trimmed RNA-seq data was assessed using FastQC ("Babraham Bioinformatics," 2023). FastQC is a quality control analysis tool for high throughput sequencing data. It provides information about

- *basic statistics*: some simple composition statistics for the FASTQ file analyzed
- *per base sequence quality*: an overview of the range of quality values across all bases at each position in the FASTQ file
- *per tile sequence quality*: an overview of the per tile sequence quality in case an Illumina library was used
- *per sequence quality scores*: an overview of how the overall quality scores of the sequences are distributed
- *per base sequence content*: an overview of the proportion of each base position in a FASTQ file for which each of the four normal DNA bases has been called
- *per sequence GC content*: the GC content across the whole length of each sequence in a file compared with a normal distributed GC content
- *per base N content*: an overview of the N content at each position across all bases
- *sequence length distribution*: an overview of how the sequence lengths are distributed
- *sequence duplication levels*: an overview of the degree of duplication for every sequence in a library
- *over-represented sequences*: a list of over-represented sequences matched against common contaminants
- *adapter content*: a check for significant amounts of adapter sequences the FASTQ file

The results of the separate FastQC analyses (of the raw and trimmed FASTQ files), the results of the Trimmomatic trimming and the information about the kallisto pseudoalignments were summarized in an interactive MultiQC HTML-report. See (Ewels et al., 2016).

3.3 RNA-seq preprocessing

The initial quality assessment of the raw FASTQ files revealed that about roughly the first 12 base pairs of the reads were of low quality. Therefore, the data was preprocessed/trimmed using Trimmomatic, a ”flexible and efficient preprocessing tool, which could correctly handle paired-end data” (Ewels et al., 2016).

That trimming substantially improved the per base sequence content and the per base sequence quality. Figures 1 and 2 show exemplary a comparison of the FastQC assessments of a raw FASTQ file vs the corresponding trimmed FASTQ file.

Figure 1: FastQC quality assessment of the per base sequence content of the raw FASTQ file CRR240976_f1.fastq.gz vs the trimmed file CRR240976_f1.trim.p.fastq.gz

i) Raw data **ii)** Trimmed data

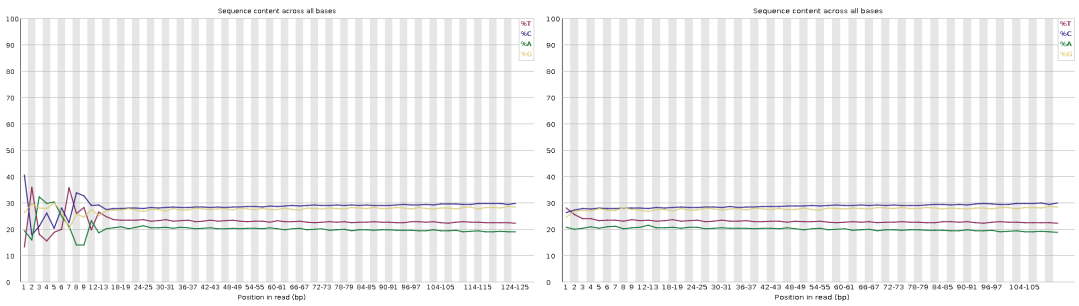
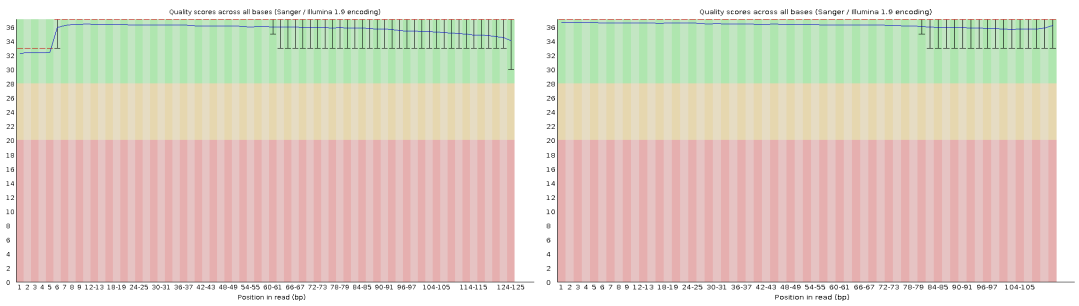


Figure 2: FastQC quality assessment of the per base sequence quality of the raw FASTQ file CRR240976_f1.fastq.gz vs the trimmed file CRR240976_f1.trim.p.fastq.gz

i) Raw data **ii)** Trimmed data



3.4 Transcripts abundances quantification

The mapping/pseudoalignment of the RNA-seq reads and the abundances quantification of the transcripts was done using kallisto. According to (Bray et al., 2016), kallisto offers the following advantages over other alignment and quantification software:

> kallisto is a program for quantifying abundances of transcripts from RNA-Seq data, or more generally of target sequences using high-throughput sequencing reads. It is based on the novel idea of pseudoalignment for rapidly determining the compatibility of reads with targets, without the need for alignment. On benchmarks with standard RNA-Seq data, kallisto can quantify 30 million human bulk RNA-seq reads in less than 3 minutes on a Mac desktop computer using only the read sequences and a transcriptome index that itself takes than 10 minutes to build. Pseudoalignment of reads preserves the key information needed for quantification, and kallisto is therefore not only fast, but also comparably accurate to other existing quantification tools. In fact, because the pseudoalignment procedure is robust to errors in the reads, in many benchmarks kallisto significantly outperforms existing tools.

kallisto requires a reference genome/transcriptome for aligning the RNA-seq data. To this end, reference FASTA cDNA dumps of *O. nivara*, cultivar BJ278 and *O. sativa*, cultivar Nipponbare were downloaded from Ensembl. The Ensembl project delivers reference data for genome interpretation for any species: genome assemblies from public archive are annotated with genes, regulatory regions, variants and comparative data to provide a foundation for scientific research and genome interpretation (Cunningham et al., 2021).

3.5 Statistical evaluation and differential expression analysis

The preprocessed and aligned data was further evaluated and analyzed using an R-script (R Core Team, 2023) executed within RStudio (Posit team, 2023). The following sections describe the analysis steps in detail.

3.5.1 Import of the kallisto transcript-level estimates

The kallisto transcript-level estimates were imported using the R-package `tximport` (Love et al., 2022; Soneson et al., 2015). Thereby, the abundances, counts, and transcript lengths were summarized to the gene level.

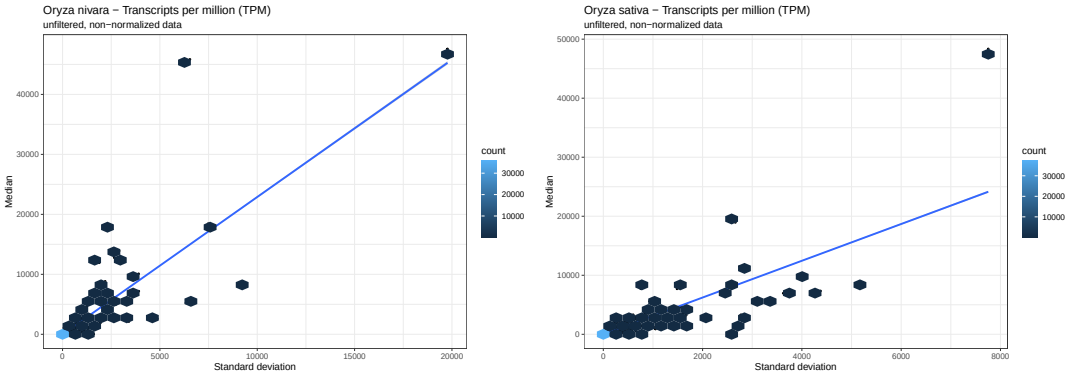
Scaling method: average transcript length over samples and then the library size (parameter `lengthScaledTPM`).

The mapping of the transcript IDs (used within the kallisto `abundance.tsv` files) to the corresponding gene IDs was done using the BioMart database `plants_mart` hosted at https://plants.ensembl.org. Datasets: `nivara_eg_gene` and `osativa_eg_gene` for *O. nivara* and *O. sativa*, respectively. See (Durinck & Huber, 2023; Durinck et al., 2009).

Figure 3 shows some basic transcripts per million (TPM) statistics about the imported kallisto files.

Figure 3: TPM statistics about the imported kallisto data

i) *O. nivara* **ii)** *O. sativa*

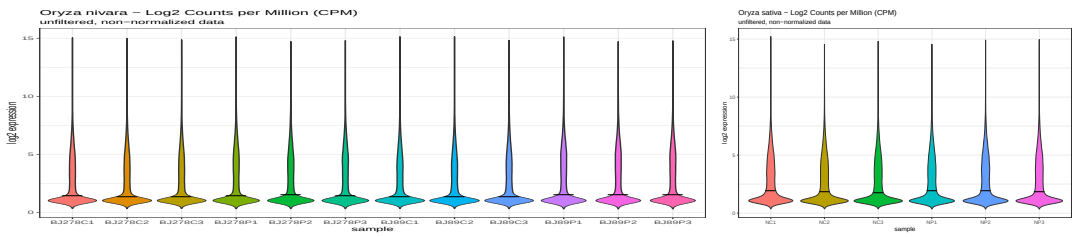


3.5.2 Filtering and normalization

For further analysis, `DGEList`-objects with counts per million (CPM) and log2(CPM) values were created using the R-package `edgeR` (Chen et al., 2023; Robinson et al., 2010). Figure 4 shows the distribution of the log2(CPM) values.

Figure 4: Log2(CPM) distribution of the unfiltered, non-normalized data

i) *O. nivara* **ii)** *O. sativa*



In order to assess reasonable values for filtering the data, the number of genes with no reads at all (in none of the samples) and conversely the number of genes with CPMs ≥ 1 in at least 1, 2, 3, … of the samples were computed. Table 1 summarizes the results.

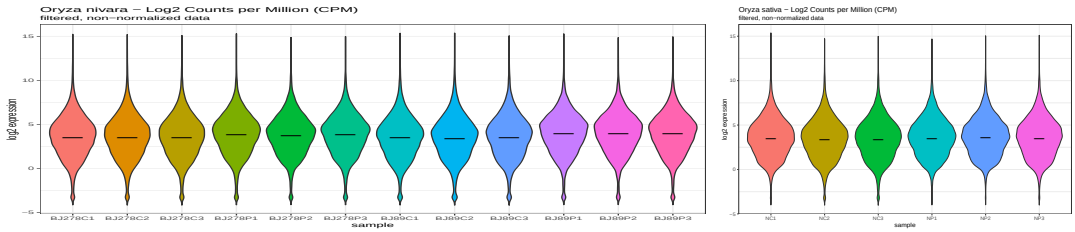
Table 1: Number of genes with no reads at all, and conversely number of genes with CPMs ≥ 1 in at least n = 1, 2, 3, … of the samples

			Genes with CPMs ≥ 1 in at least n samples					
Species	Σ genes	Σ no reads	1	2	3	4	5	6
O. nivara	36313	7115	21001	20355	19891	19302	18895	18481
O. sativa	37967	4699	23934	22510	21609	20592	19715	18656

Filtering out genes with low reads (< 1 CPM in at least half of the samples) resulted in the distribution of the log2(CPM) values shown in figure 5.

Figure 5: Log2(CPM) distribution of the filtered (< 1 CPM in at least half of the samples), non-normalized data

i) *O. nivara* **ii)** *O. sativa*



Finally, the filtered data was normalized using the `edgeR` function `calcNormFactors` which calculates scaling factors to convert raw library sizes into effective library sizes. Used normalization method: `TMM`. The results of the normalization are shown in figure 6.

Figure 6: Log2(CPM) distribution of the filtered, normalized data

i) *O. nivara* **ii)** *O. sativa*

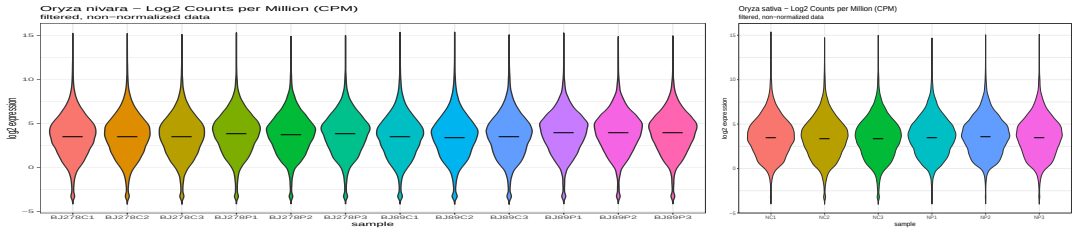


Figure 7 provides an overview of the filtering and normalization results.

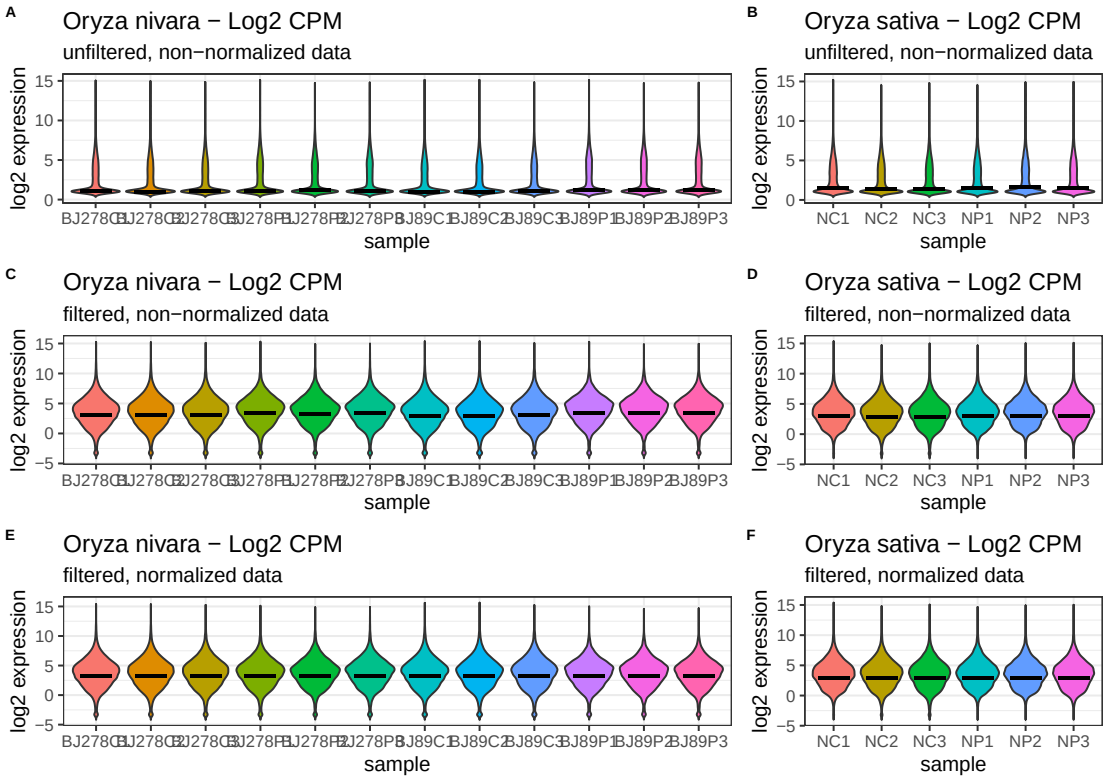
3.5.3 Hierarchical cluster analysis

A hierarchical cluster analysis (HCA) was performed via the `stats` function `hclust` on the Euclidean distance matrix of the log2(CPM) values. Used agglomeration method: `method = complete`.

3.5.4 Principal component analysis

A principal component analysis (PCA) of the filtered and normalized log2(CPM) values was performed via the `stats` function `prcomp`.

Figure 7: Log2(CPM) distribution of the filtered, normalized data in comparison with the non-normalized and the unfiltered data



3.5.5 Identification of differentially expressed genes

In order to identify differentially expressed genes (DEGs), design and contrast matrices were created using the `stats` function `model.matrix` and the `limma` function `makeContrasts`.

The design matrices were used to create linear model fits for each gene via the `limma` functions `voom` and `lmFit`.

The linear model fits and the contrast matrices were used to calculate estimated coefficients and standard errors (contrasts) via the `limma` function `contrasts.fit`.

The contrasts were used to "calculate moderated t-statistics, moderated F-statistic, and log-odds of differential expression by empirical Bayes moderation of the standard errors towards a global value" (empirical Bayes statistics) via the `limma` function `eBayes` (Smyth et al., 2023).

Finally, the empirical Bayes statistics were used to extract a table of the top-ranked genes via the `limma` function `topTable` (sorted by the `LogFC`-values).

Venn diagrams of the DEGs (according to the calculated empirical Bayes statistics) were created via the `gprofiler2` function `decideTests` (with parameters `method = "global"`, `adjust.method = "BH"`, `p.value = 0.01`, `lfc = 7`) and the `limma` function `vennDiagram`.

See (Ritchie et al., 2015) for details on the statistical foundations implemented by `limma`.

3.6 Functional enrichment analysis

A functional enrichment analysis of the 100 "top-ranked" genes was performed via the `gprofiler2` function `gost` (with `correction_method = "fdr"` (Kolberg & Raudvere, 2021).

4 Results

4.1 Quality evaluation

The RNA-seq data was initially assessed with FastQC, and according to this assessment the data was preprocessed/trimmed with Trimmomatic and afterwards assessed again with FastQC. Finally, an overall report on the data preprocessing, the quality assessments and the kallisto pseudoalignments was created with MultiQC.

Figure 8 shows the MultiQC report on the Trimmomatic preprocessing (the surviving reads).

Figure 8: Trimmomatic preprocessing

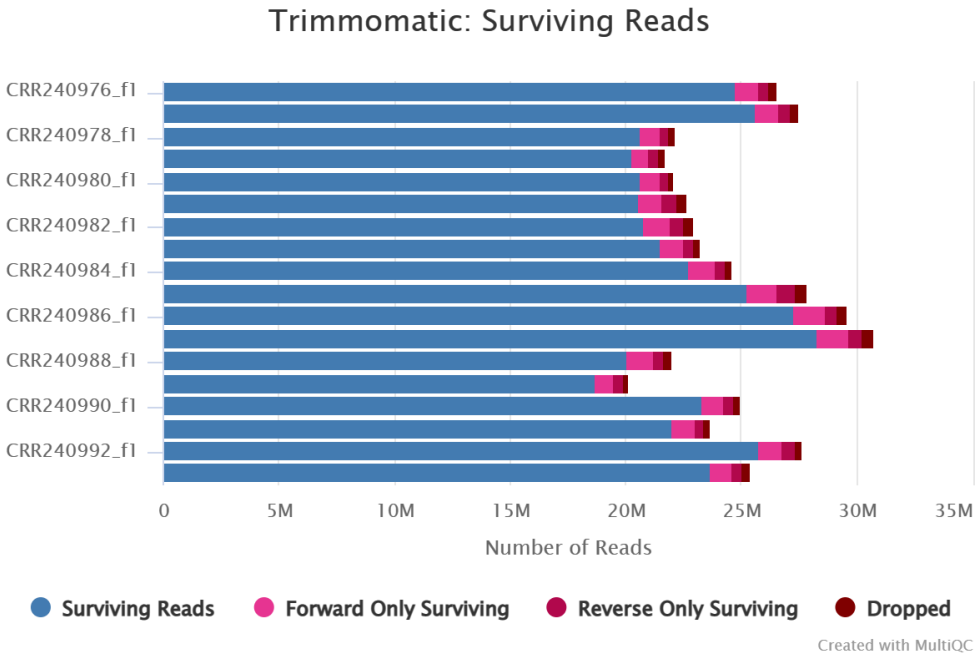


Figure 9 shows the MultiQC overview of the FastQC quality assessments of the trimmed FASTQ files.

According to these quality assessments, the (trimmed) RNA-seq data may be regarded as good quality for the purpose of this research.

4.2 Reads mapped to the reference transcriptome

For most of the FASTQ files, kallisto pseudoaligned well above 80 % of the (preprocessed) RNA-seq reads. Figure 10 shows a MultiQC overview of the kallisto pseudoalignments.

4.3 Hierarchical cluster analysis

The hierarchical cluster analysis reveals that the normal condition groups and the drought stress condition groups are closely related (grouped together). But the data for *O. nivara* also shows that the cultivar has an even greater impact on the clustering then the drought stress condition (see figures 11 and 12).

Figure 9: FastQC quality assessment of the preprocessed FASTQ files

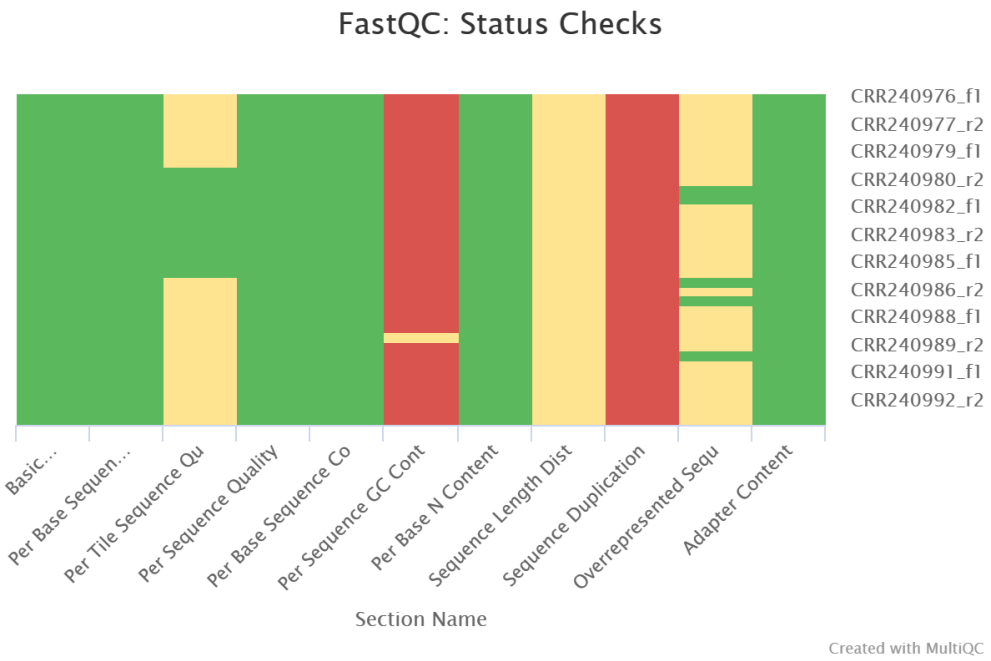


Figure 10: kallisto pseudoalignments

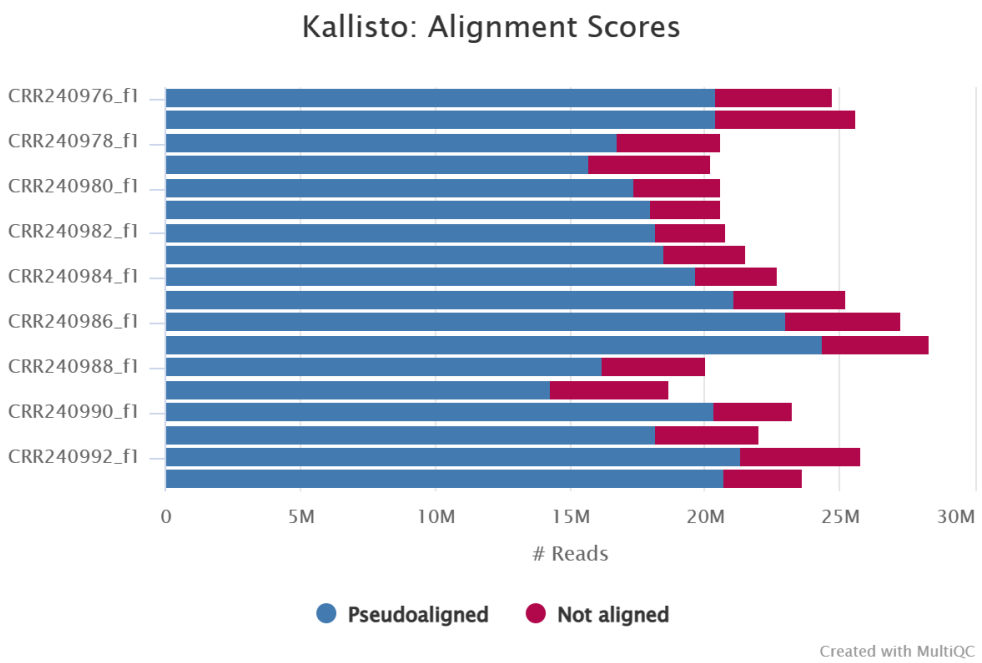


Figure 11: Hierarchical cluster analysis of the *O. nivara* RNA-seq data

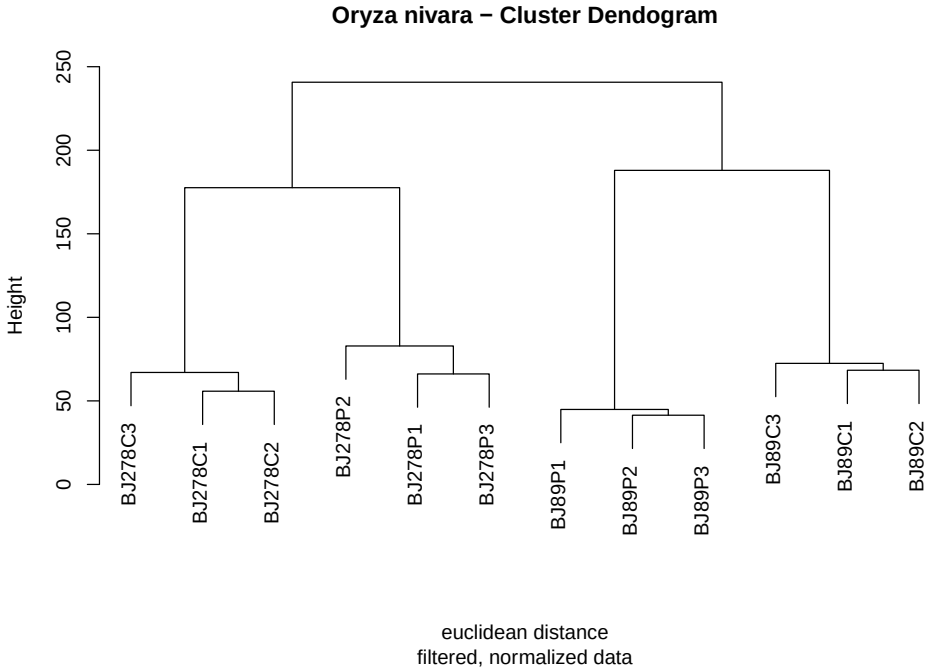
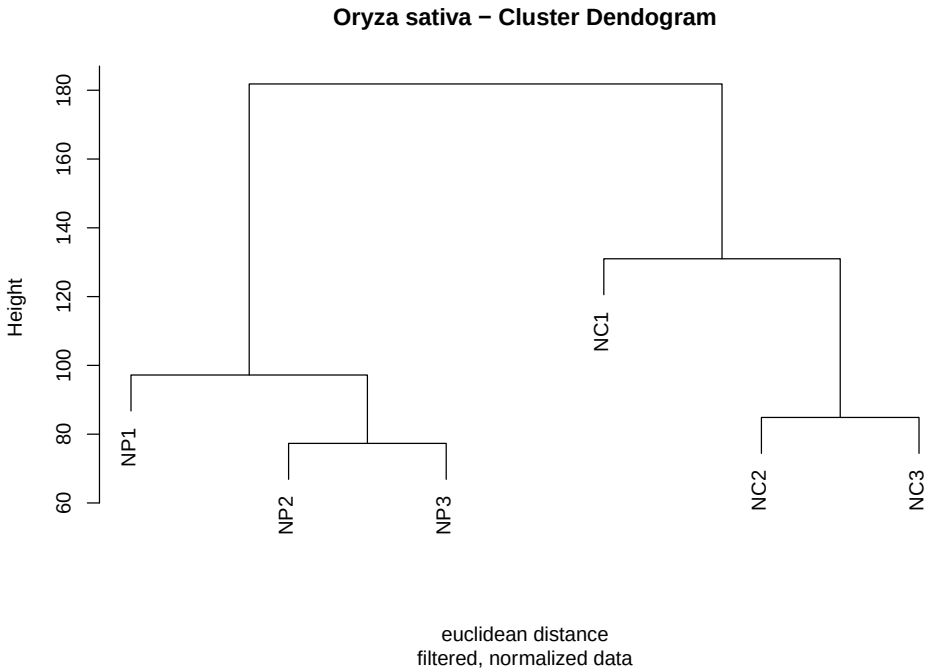


Figure 12: Hierarchical cluster analysis of the *O. sativa* RNA-seq data



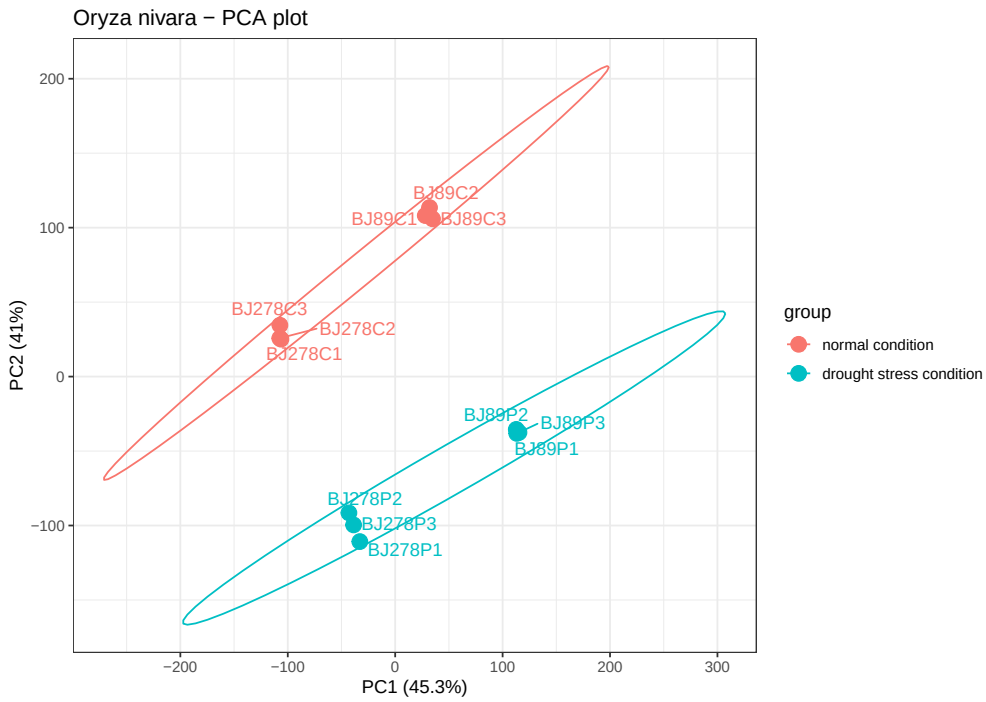
4.4 Principal component analysis

The principal component analysis (PCA) reveals that for both *O. nivara* and *O. sativa* the first two principal components account for more than 80 % of the variance in the gene expression. Figures 13 and 14 show the contribution percentage of the samples to the first two principal components (PCs) for the two species.

For *O. nivara*, the samples from the same condition (normal vs drought stress) cluster together, with a tight clustering of the two different cultivars. This means that the different conditions might well explain the differences in gene expression, with the cultivar being an important confounding factor.

For *O. sativa*, the samples from the same condition cluster together with the exception of sample "NC1". This might be due to a batch effect.

Figure 13: PCA of the log2(CPM) data - *O. nivara*



4.5 Differentially expressed genes

The volcano plots 15 and 16 provide a quick overview of the genes with large fold changes that are also statistically significant. These may be the biologically most significant genes.

According to the limma tests with a p-value of one percent, there are 15 down- vs 33 up-regulated genes for *O. nivara* under drought stress conditions, and 2 down- vs 16

Figure 14: PCA of the log2(CPM) data - *O. sativa*

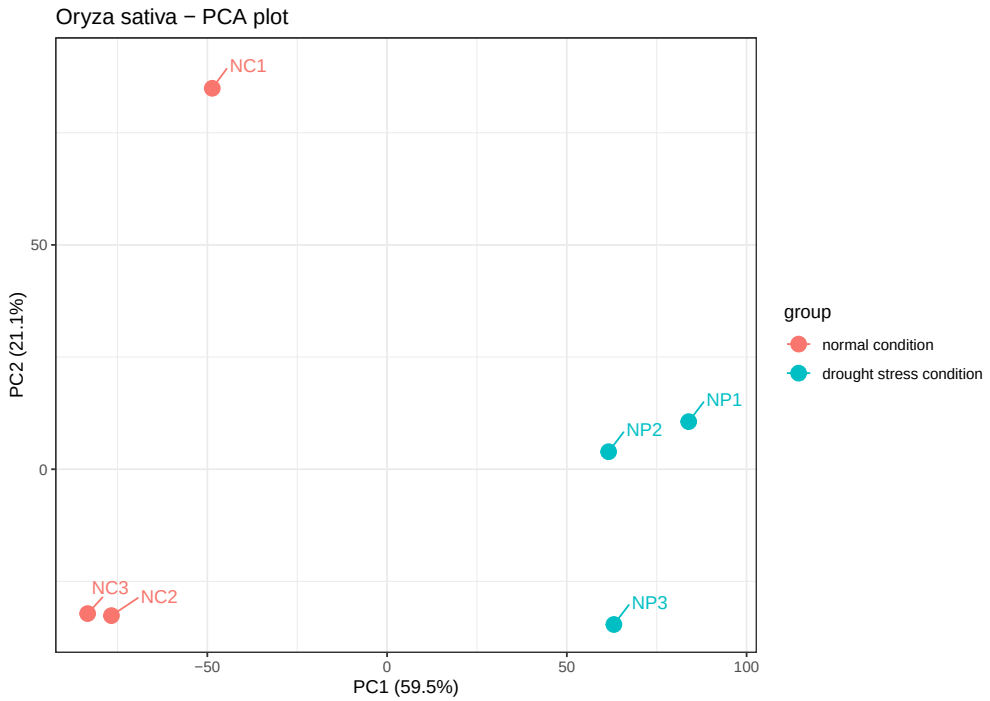


Figure 15: Volcano plot of the DEGs - *O. nivara*

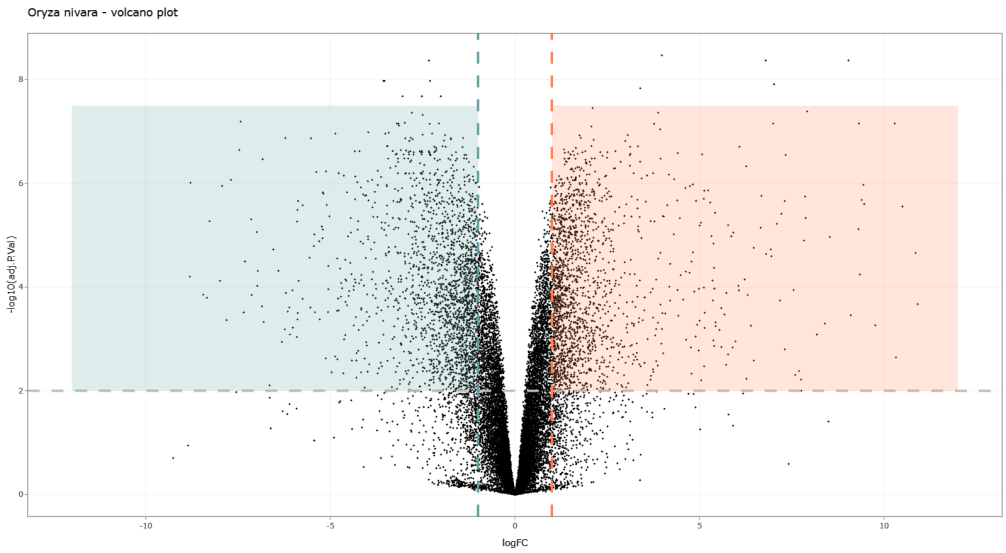
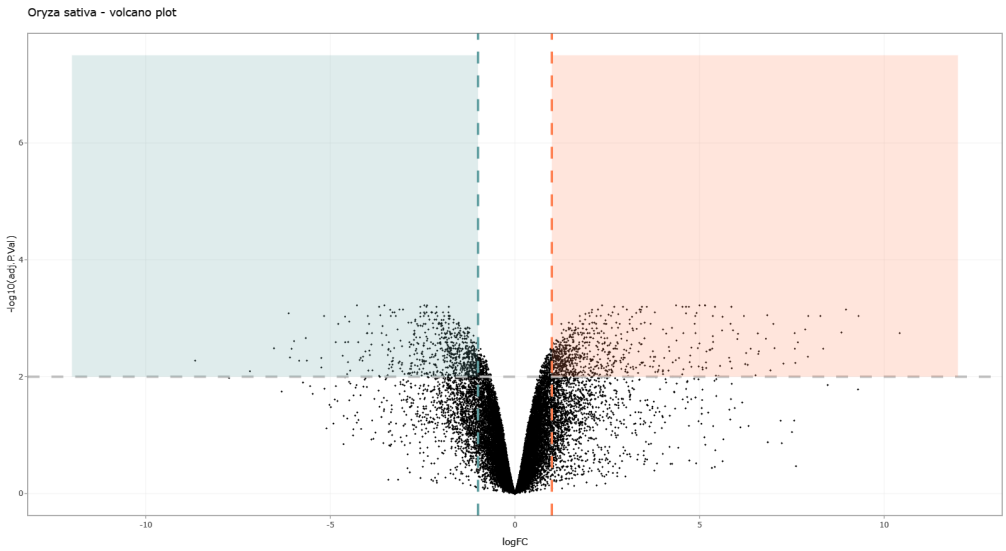


Figure 16: Volcano plot of the DEGs - *O. sativa*



up-regulated for *O. sativa*. Figure 17 shows the respective Venn diagrams. Figures 18 and 19 present heatmaps of these genes.

Figure 17: Venn diagrams of the DEGs

i) *O. nivara* **ii)** *O. sativa*

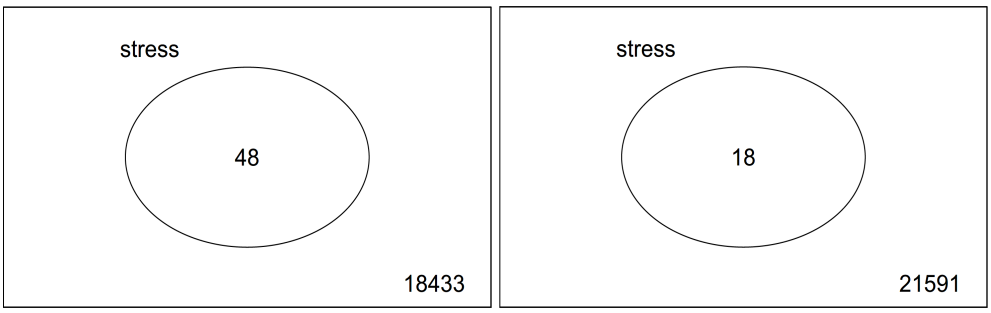


Figure 18: Heatmap of the DEGs - *O. nivara*

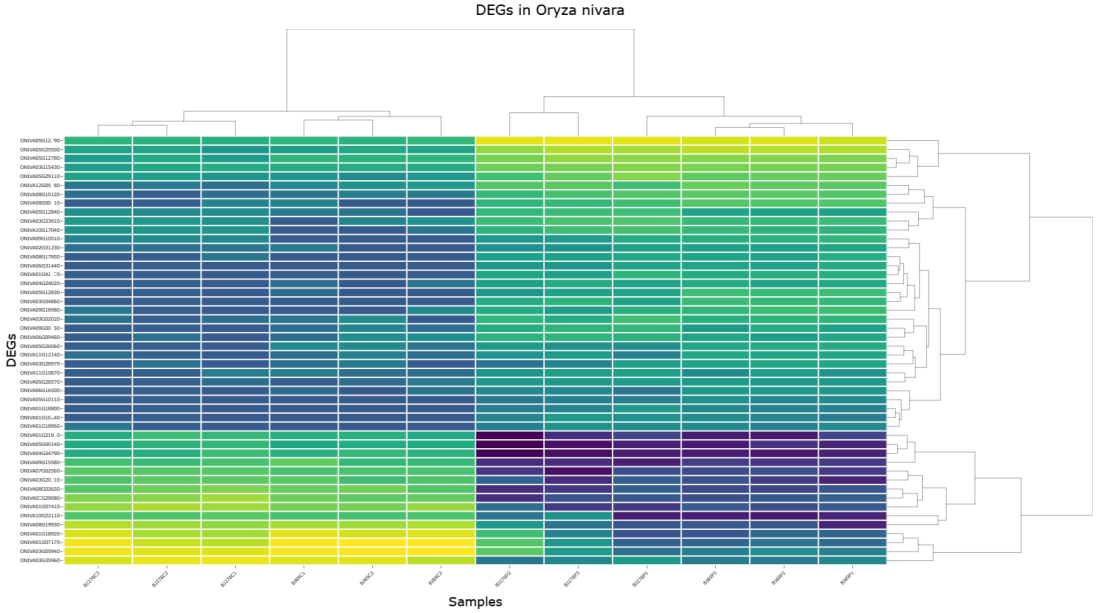
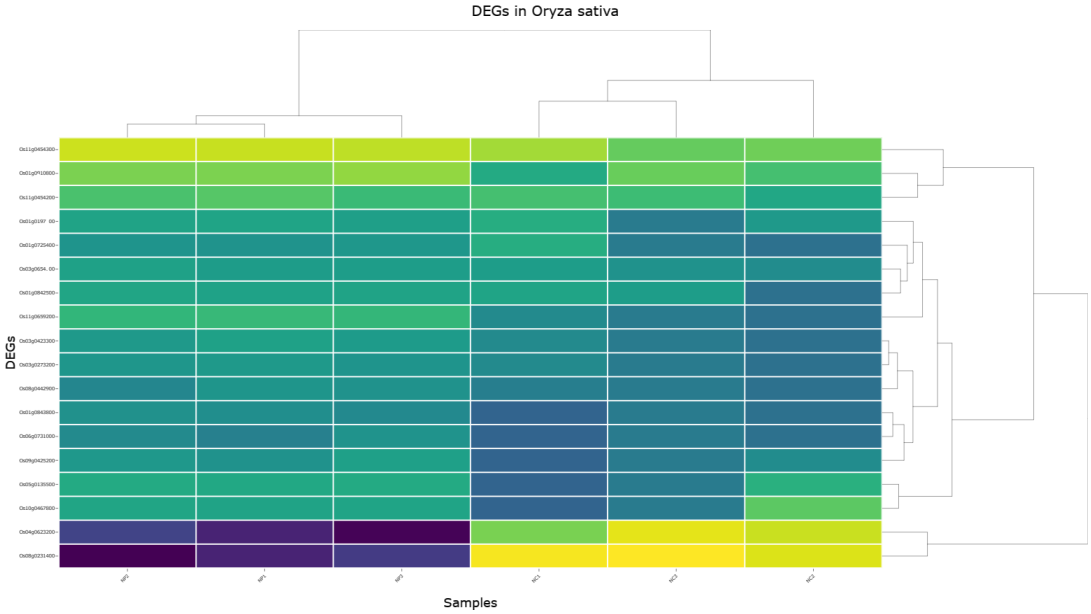


Figure 19: Heatmap of the DEGs - *O. sativa*



Down- and up-regulated genes are for *O. nivara*:

ONIVA01G07170, ONIVA01G07410, ONIVA01G16640, ONIVA01G18800, ONIVA01G18920, ONIVA01G18960, ONIVA01G21950, ONIVA01G41350, ONIVA02G31230, ONIVA03G02020, ONIVA03G04860, ONIVA03G05940, ONIVA03G15430, ONIVA03G20710, ONIVA03G28970, ONIVA03G33910, ONIVA03G35960, ONIVA04G24020, ONIVA04G24790, ONIVA05G00140, ONIVA05G10110, ONIVA05G12780, ONIVA05G12790, ONIVA05G12830, ONIVA05G12840, ONIVA05G25500, ONIVA05G26060, ONIVA05G28570, ONIVA05G29080, ONIVA05G29110, ONIVA06G09460, ONIVA06G16500, ONIVA06G31440, ONIVA07G02560, ONIVA08G00310, ONIVA08G02650, ONIVA08G10120, ONIVA08G17850, ONIVA08G19930, ONIVA09G00350, ONIVA09G10510, ONIVA09G15580, ONIVA09G18980, ONIVA10G17040, ONIVA10G22110, ONIVA11G10870, ONIVA11G12140, ONIVA12G05380

For *O. sativa* these genes are:

Os01g0197700, Os01g0725400, Os01g0842500, Os01g0843800, Os01g0910800, Os03g0273200, Os03g0423300, Os03g0654700, Os04g0623200, Os05g0135500, Os06g0731000, Os08g0231400, Os08g0442900, Os09g0425200, Os10g0467800, Os11g0454200, Os11g0454300, Os11g0659200

A description and the location of all these genes can be found at https://plants.ensembl.org. For a deeper analysis of these genes and their potential biological significance, the available information appears to be insufficient.

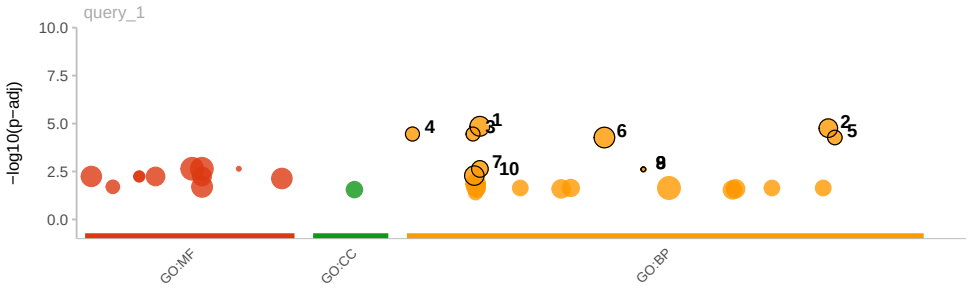
4.6 Functional enrichment analysis

Finally, for the one-hundred top-ranked genes a statistical functional enrichment analysis was performed to identify over- or under-represented information from gene ontology (GO) terms. Figures 20 and 21 present the results as Manhattan plots.

For both *O. nivara* and *O. sativa*, the top ten GO terms belong to the major subontology "biological process (BP)". It appears that on the gene level the impact of drought stress may be best described and categorized in terms of biological processes as opposed to the molecular functions (MF) of the genes and as opposed to the cellular components (CC) in which the gene products are physically located.

Furthermore, for both rice species the following GO terms are significantly enriched (when just comparing the top ten GO terms): response to organic substance, response to water, response to acid chemical, response to salt, response to chemical, lignin catabolic process, phenylpropanoid catabolic process. This confirms the close relationship of the two species, and it also confirms that the GO terms are suitable to describe genetic responses to environmental factors.

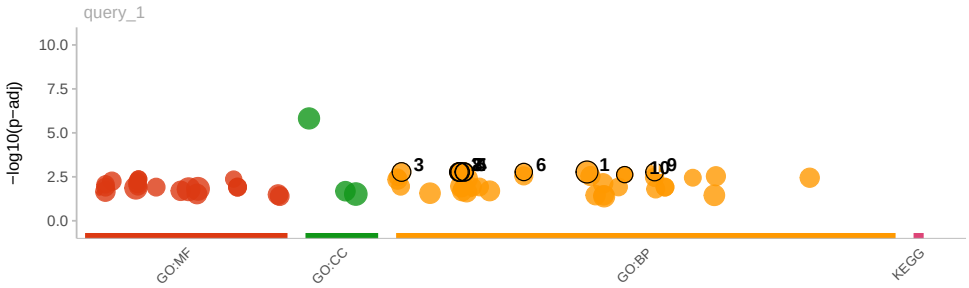
Figure 20: Manhattan plot with the first 10 top-ranked GO terms highlighted - *O. nivara*



id	source	term_id	term_name	term_size		p_value
1	GO:BP	GO:0010033	response to organic substance	699		1.4e−05
2	GO:BP	GO:1901700	response to oxygen−containing compound	419		1.7e−05
3	GO:BP	GO:0009415	response to water	84		3.5e−05
4	GO:BP	GO:0001101	response to acid chemical	87		3.5e−05
5	GO:BP	GO:1902074	response to salt	100		5.4e−05
6	GO:BP	GO:0042221	response to chemical	1038		5.4e−05
7	GO:BP	GO:0010035	response to inorganic substance	198		2.3e−03
8	GO:BP	GO:0046274	lignin catabolic process	24		2.4e−03
9	GO:BP	GO:0046271	phenylpropanoid catabolic process	24		2.4e−03
10	GO:BP	GO:0009628	response to abiotic stimulus	604		5.2e−03

g:Profiler (biit.cs.ut.ee/gprofiler)

Figure 21: Manhattan plot with the first 10 top-ranked GO terms highlighted - *O. sativa*



id	source	term_id	term_name	term_size		p_value
1	GO:BP	GO:0042221	response to chemical	1487		1.7e−03
2	GO:BP	GO:0009250	glucan biosynthetic process	121		1.7e−03
3	GO:BP	GO:0001101	response to acid chemical	123		1.7e−03
4	GO:BP	GO:0009832	plant−type cell wall biogenesis	109		1.7e−03
5	GO:BP	GO:0009834	plant−type secondary cell wall biogenesis	58		1.7e−03
6	GO:BP	GO:0030244	cellulose biosynthetic process	51		1.7e−03
7	GO:BP	GO:0009414	response to water deprivation	116		1.7e−03
8	GO:BP	GO:0009415	response to water	118		1.7e−03
9	GO:BP	GO:0051274	beta−glucan biosynthetic process	62		1.7e−03
10	GO:BP	GO:0046274	lignin catabolic process	26		2.4e−03

g:Profiler (biit.cs.ut.ee/gprofiler)

5 Discussion

This analysis reveals that both species *O. nivara* and *O. sativa* react strongly to drought stress by down- and up-regulating their gene activity. According to the limma tests (p-value = 1 %), the number of significantly down- and up-regulated genes is much higher for the wild rice than for the cultivated rice: 48 genes in the wild rice vs 18 genes in the cultivated rice. These results are consistent with current research results: Drought tolerance is regulated by several genes (Joshi et al., 2016). Furthermore, wild and cultivated species of rice have distinctive proteomic responses to drought stress (Hamzelou et al., 2020).

The hierarchical cluster analysis of the two *O. nivara* cultivars further uncovers that the examined cultivar might have an even greater impact on the gene expressions than drought stress conditions. However, when considering the ontology of the down- and up-regulated genes, both rice species largely coincide with respect to the significantly enriched GO terms. Another result is that on the gene level the impact of drought stress may be best described and categorized in terms of biological processes (BP) as opposed to the molecular functions (MF) of the genes and as opposed to the cellular components (CC) in which the gene products are physically located.

The initial quality assessment indicate a good quality of the examined RNA-seq data. However, the principal component analysis indicates that one of the three *O. sativa* samples might have a batch effect. Irrespective of this, the general results of this analysis should be well reproducible.

Rice is the most widely consumed staple food for a large part of the world's human population, and drought is the most imperative and major limitation for rice production in rainfed ecosystems. Therefore, there is a great requirement of rice varieties with drought tolerance, and much research is done to improve the drought tolerance of rice (Panda et al., 2021).

This study analyses leaf tissues from seedlings at the age of twelve days. For a better understanding of the adaptation mechanisms to drought stress conditions, it would be necessary to also examine and compare other rice species and cultivars, especially drought-tolerant versus drought-sensitive cultivars. Furthermore, different tissues and ages should be examined. The function of the differentially expressed genes need to be analyzed in detail.

The research of drought stress responses on the gene level may substantially contribute to the breeding for drought tolerant rice varieties.

6 Online content and license notice

This document, its LaTeX source files, all figures and tables and all supplemental R and shell scripts are available at the public GitHub repository https://github.com/IngoGiebel/qbio304-student-work. Furthermore, the data subfolder contains the following files:

- the `abundance.tsv` files created by kallisto
- the `*_fastqc.html` report files created by FastQC
- the log-files created by Trimmomatic and kallisto
- the study design file for the data: `studydesign-PRJCA004229.tsv`

The data folder does not contain: the raw and trimmed FASTQ files, the reference genome files for *O. nivara* and *O. sativa*, the index files for these reference genomes created by kallisto, the `*_fastqc.zip` files created by MultiQC. Information on how to get this data is provided in the `README.md` of the data subfolder.

The `README.md` of the scripts subfolder contains information about the external programs required to execute the shell scripts.

This work, its LaTeX source files, all figures and tables and all provided R and shell scripts are under the MIT License: https://github.com/IngoGiebel/qbio304-student-work/blob/main/LICENSE.md.

References

Babraham bioinformatics. (2023, April 30). https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527. https://doi.org/10.1038/nbt.3519

Chen, Y., Lun, A. T., McCarthy, D. J., Ritchie, M. E., Phipson, B., Hu, Y., Zhou, X., Robinson, M. D., & Smyth, G. K. (2023). *Edger: Empirical analysis of digital gene expression data in r* [http://bioinf.wehi.edu.au/edgeR].

Cunningham, F., Allen, J. E., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Austine-Orimoloye, O., Azov, A. G., Barnes, I., Bennett, R., Berry, A., Bhai, J., Bignell, A., Billis, K., Boddu, S., Brooks, L., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., … Flicek, P. (2021). Ensembl 2022. *Nucleic Acids Research*, *50*(D1), D988–D995. https://doi.org/10.1093/nar/gkab1049

Durinck, S., & Huber, W. (2023). *Biomart: Interface to biomart databases (i.e. ensembl)* [R package version 2.54.1].

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature Protocols*, *4*, 1184–1191.

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Hamzelou, S., Kamath, K. S., Masoomi-Aladizgeh, F., Johnsen, M. M., Atwell, B. J., & Haynes, P. A. (2020). Wild and cultivated species of rice have distinctive proteomic responses to drought. *International Journal of Molecular Sciences*, *21*(17). https://doi.org/10.3390/ijms21175980

Joshi, R., Wani, S. H., Singh, B., Bohra, A., Dar, Z. A., Lone, A. A., Pareek, A., & Singla-Pareek, S. L. (2016). Transcription factors and plants response to drought stress: Current understanding and future directions. *Frontiers in Plant Science*, *7*. https://doi.org/10.3389/fpls.2016.01029

Kolberg, L., & Raudvere, U. (2021). *Gprofiler2: Interface to the g:profiler toolset* [R package version 0.2.1]. https://CRAN.R-project.org/package=gprofiler2

Love, M., Soneson, C., Robinson, M., Patro, R., Morgan, A. P., Thompson, R. C., Shirley, M., & Srivastava, A. (2022). *Tximport: Import and summarize transcript-level estimates for transcript- and gene-level analysis* [R package version 1.26.1]. https://github.com/mikelove/tximport

Panda, D., Mishra, S. S., & Behera, P. K. (2021). Drought tolerance in rice: Focus on recent mechanisms and approaches. *Rice Science*, *28*(2), 119–132. https://doi.org/10.1016/j.rsci.2021.01.002

Posit team. (2023). *Rstudio: Integrated development environment for r*. Posit Software, PBC. Boston, MA. http://www.posit.co/

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). Edger: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Smyth, G., Hu, Y., Ritchie, M., Silver, J., Wettenhall, J., McCarthy, D., Wu, D., Shi, W., Phipson, B., Lun, A., Thorne, N., Oshlack, A., de Graaf, C., Chen, Y., Langaas, M., Ferkingstad, E., Davy, M., Pepin, F., & Choi, D. (2023). *Limma: Linear models for microarray data* [R package version 3.54.2]. http://bioinf.wehi.edu.au/limma

Soneson, C., Love, M. I., & Robinson, M. D. (2015). Differential analyses for rna-seq: Transcript-level estimates improve gene-level inferences. *F1000Research*, *4*. https://doi.org/10.12688/f1000research.7563.1

Index

B
batch effect, 14, 20

C
counts per million (CPM), 7
cultivar
- BJ278, 4
- BJ89, 4
- Nipponbare, 4

D
DNA
- cDNA, 6

differentially expressed genes (DEGs), 10

E
Ensembl, 6

F
FASTA, 6
FASTQ, 4
FastQC, 4, 10f
functional enrichment analysis, 3, 10, 18

G
gene
- ID, 6

gene ontology (GO), 18
- biological process (BP), 3, 18, 20
- cellular component (CC), 3, 18, 20
- molecular function (MF), 3, 18, 20

genome/transcriptome
- reference, 6

H
hierarchical cluster analysis (HCA), 3, 8, 11, 20
- Euclidean distance matrix, 8

I
Illumina
- HiSeq 2000, 4

K
kallisto, 4f, 10f

L
LogFC, 10

M
MultiQC, 4, 10

O
Oryza
- nivara, 4
- sativa, 4

P
principal component analysis (PCA), 3, 8, 14, 20
- principal component (PC), 14

R
R, 6
- RStudio, 6
- package
  - edgeR, 7
  - gprofiler2, 10
  - limma, 10
  - stats, 8, 10
  - tximport, 6

RNA-seq, 3
- adapter sequence, 4
- contaminant, 4
- paired-end, 3
- quality assessment, 4

T
transcript
- ID, 6
- abundances quantification, 5
- pseudoalignment, 4
- transcripts per million (TPM), 6

transcriptome, 3
Trimmomatic, 4f, 10