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

List of Figures

List of Tables

1 Abstract

Provide a brief summary of the purpose of the assignment, the methods used, the main findings, and the significance of the results. Limit the abstract to 200-250 words.

2 Introduction

2.1 Background

Introduce and explain the study and give a rationale for the RNA-seq analysis. Discuss why of RNA sequencing in understanding gene expression and regulation in plants and why it is used in this study.

2.2 Objectives

State the specific aims of the assignment, which include obtaining plant RNA-seq data, evaluating its quality, mapping to a respective genome, performing statistical evaluation, differential expression analysis, functional enrichment analysis, and critically evaluating and discussing the results.

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).

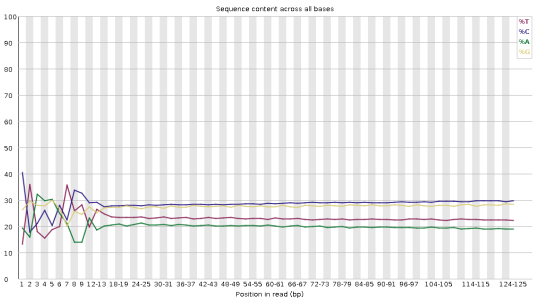
3.3 RNA-seq preprocessing

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

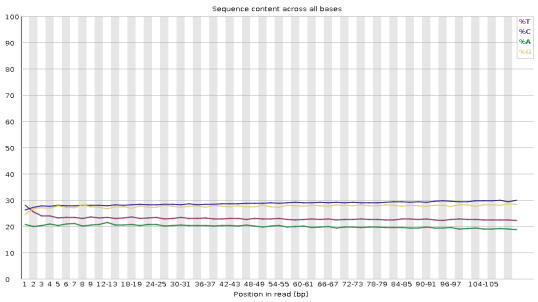
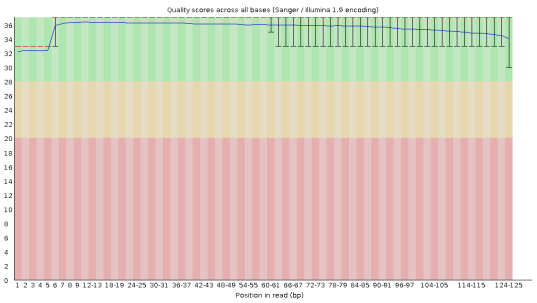
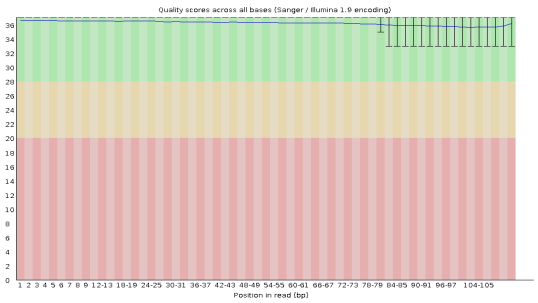


Figure 2: FastQC quality assessment of the per base sequence quality of the raw FASTQ file CRR240976_f1.fastq.gz vs the trimmed file CRR240976_f1.trim.p.fastq.gz

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 Oryza nivara, cultivar BJ278 and Oryza 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 (Bolger et al., 2014).

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 `txiimport` (Love et al., 2022; Soneson et al., 2015). Thereby, the abundances, counts, and transcript lengths were summarized to the gene level.

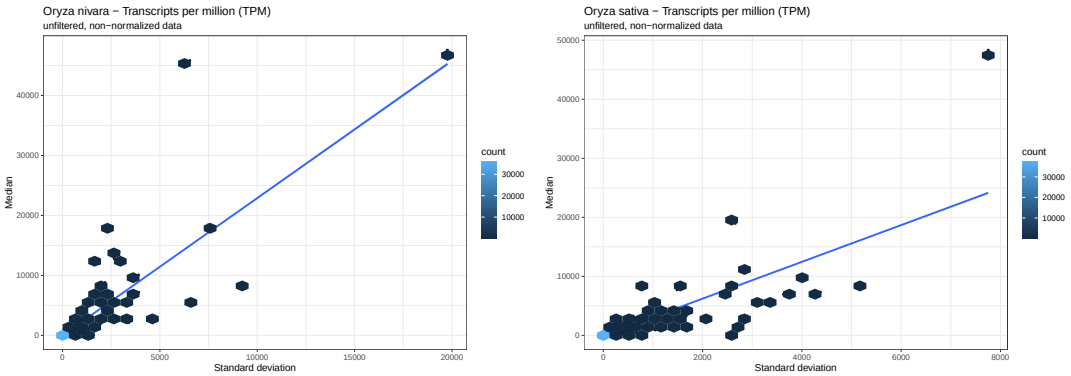
Scaling method: average transcript length over samples and then the library size (parameter `lengthScaledTPM`).

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.

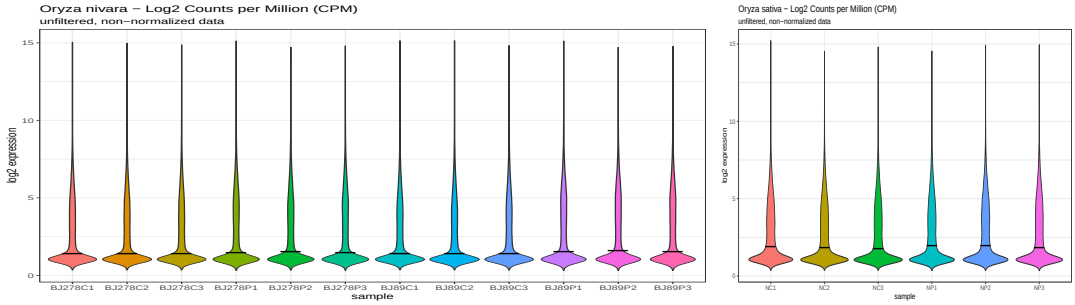
Figure 3: TPM statistics about the imported kallisto data



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



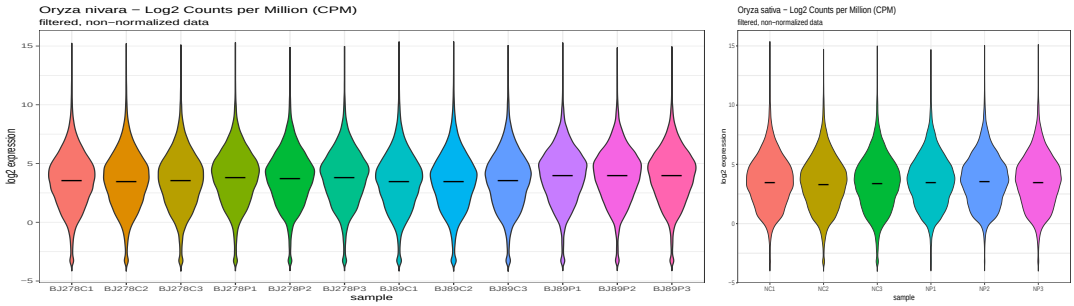
In order to assess reasonable values for filtering the data, the number of genes with no reads at all (in none of the samples) and conversely the number of genes with CPMs ≥ 1 in at least 1, 2, 3, … of the samples were computed. Table 1 summarizes the results.

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.

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

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



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

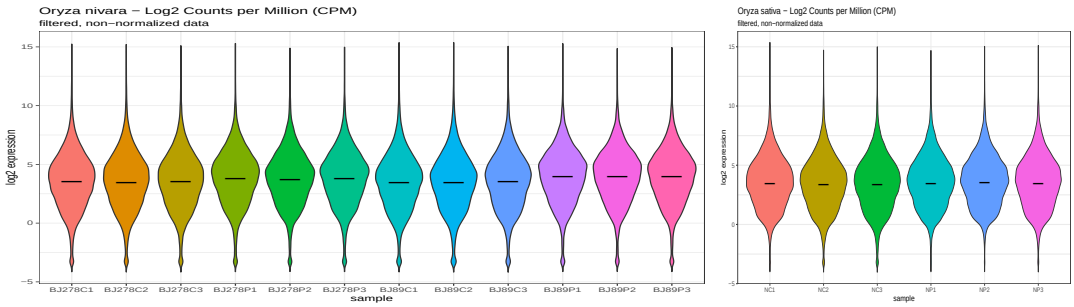
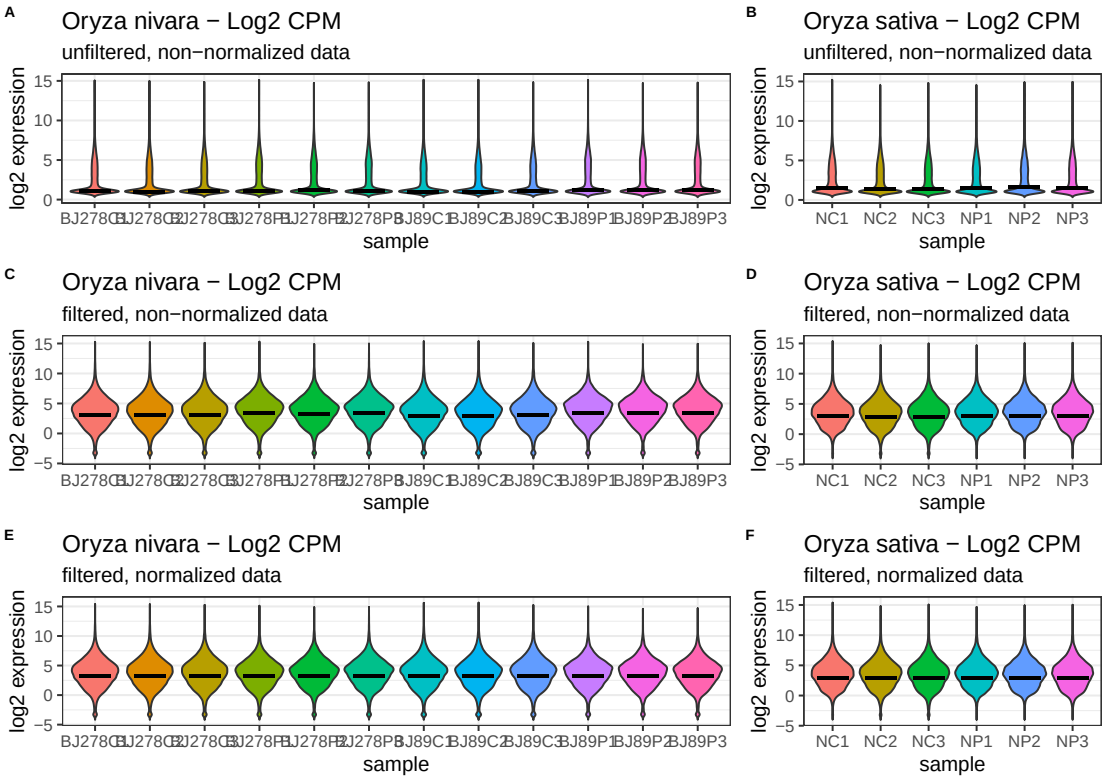


Figure 7 provides an overview of the filtering and normalization results.

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`.

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



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`.

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 (see section 3.3). 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

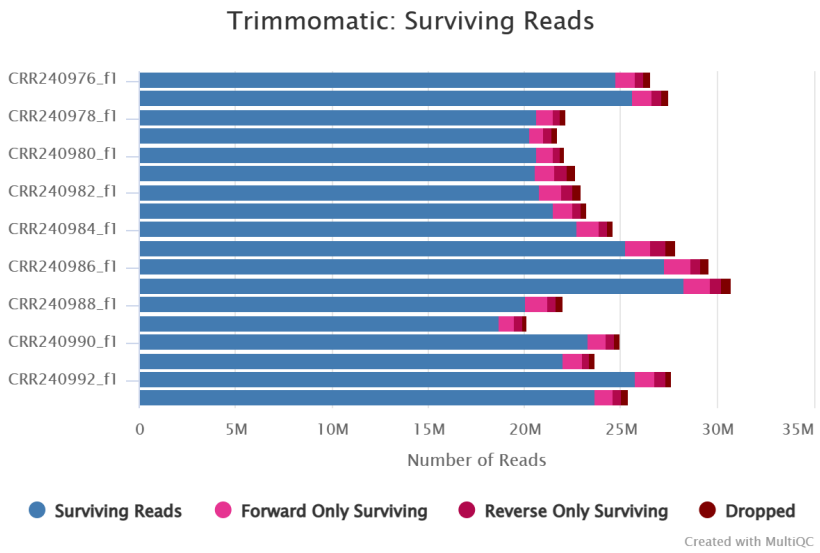


Figure 9 shows the MultiQC overview of the FastQC quality assessments of the trimmed FASTQ files.

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 (see section 3.5.3) 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

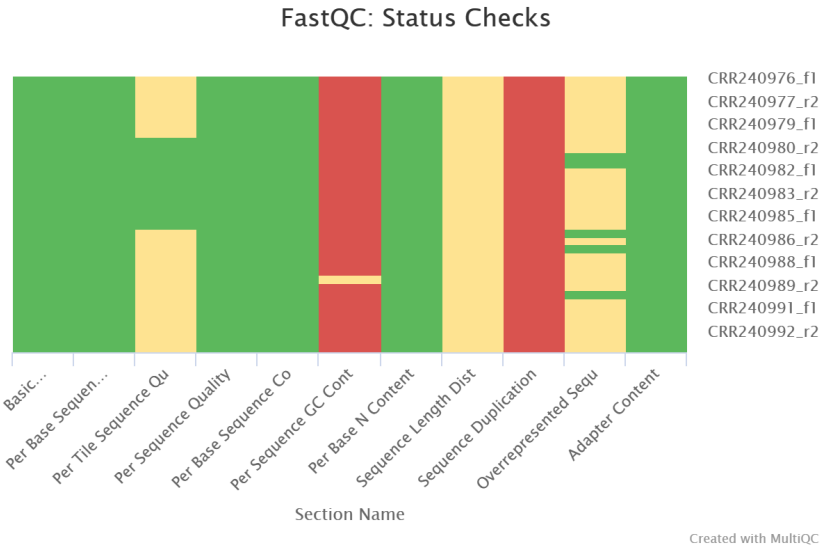


Figure 10: kallisto pseudoalignments

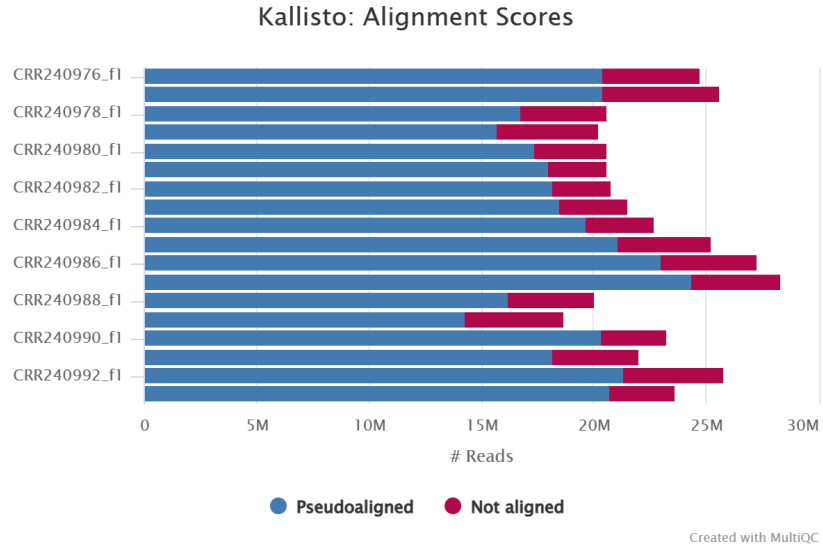


Figure 11: Hierarchical cluster analysis of the O. nivara RNA-seq data

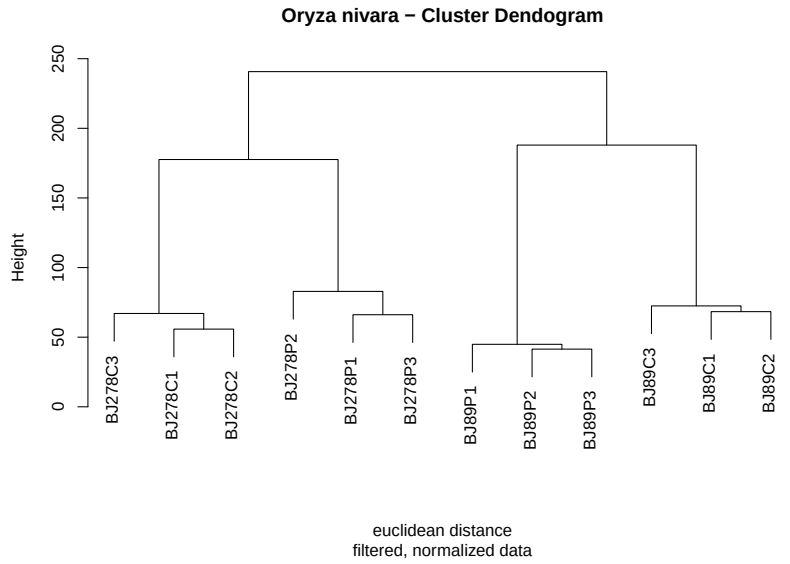
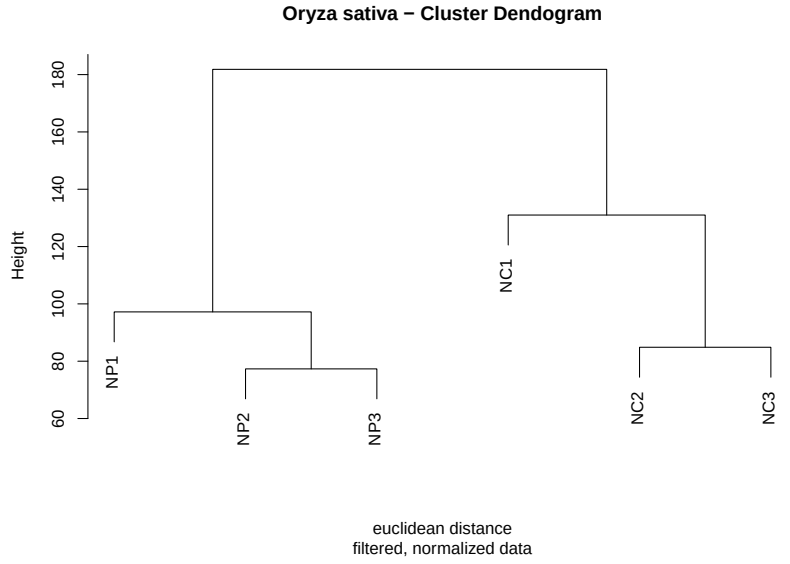


Figure 12: Hierarchical cluster analysis of the O. sativa RNA-seq data



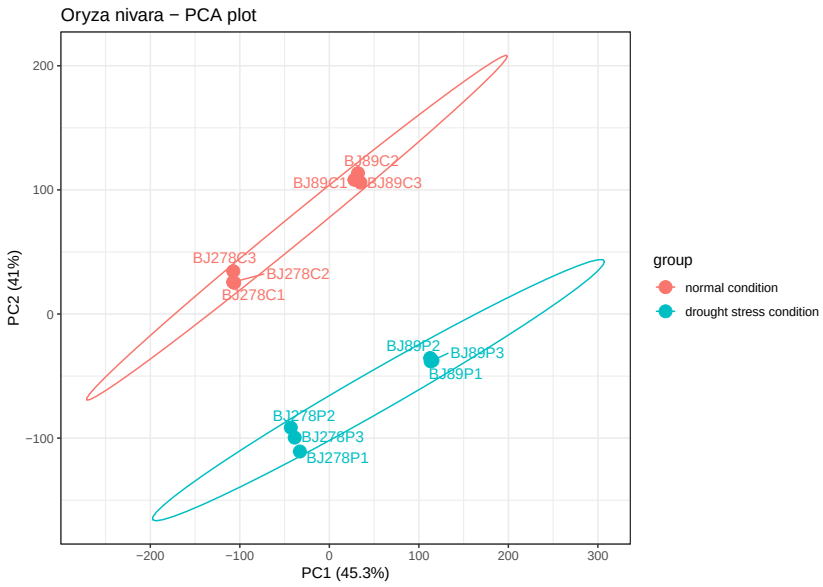
4.4 Principal component analysis

The principal component analysis (PCA) (see section 3.5.4) 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

TODO: Discuss the identified differentially expressed genes and their potential biological significance.

4.6 Functional enrichment analysis

TODO: Present the results of the functional enrichment analysis, highlighting the enriched functional categories.

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

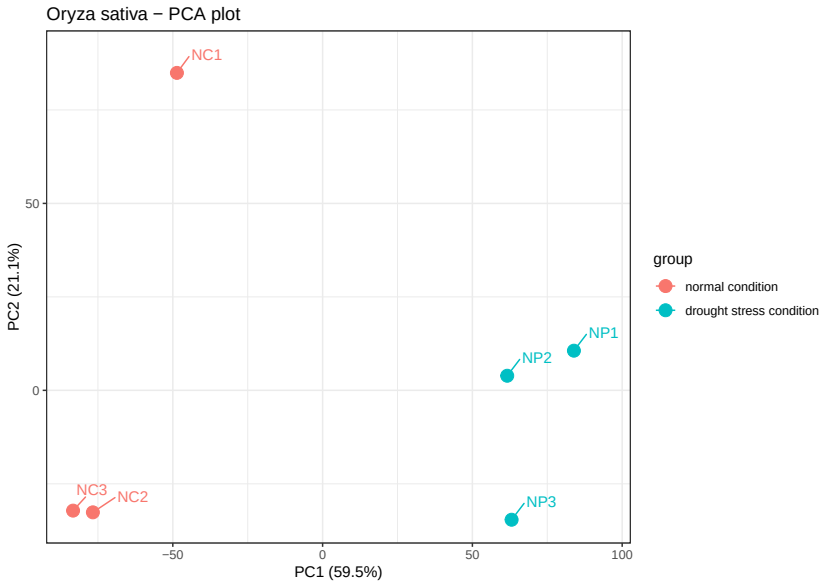


Figure 15: Volcano plot of the DEGs - O. nivara

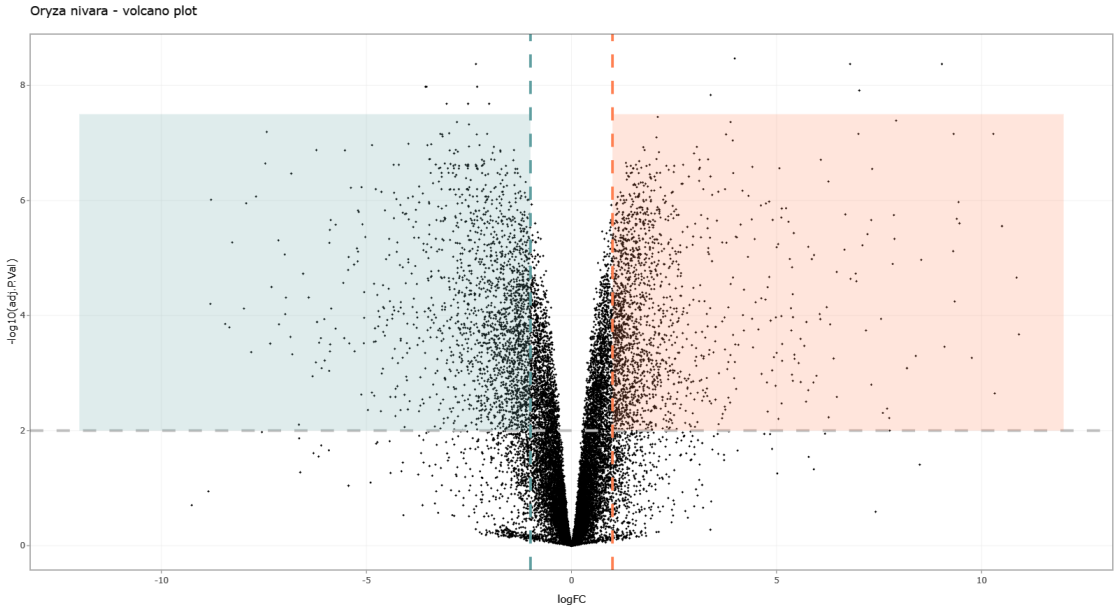


Figure 16: Volcano plot of the DEGs - O. sativa

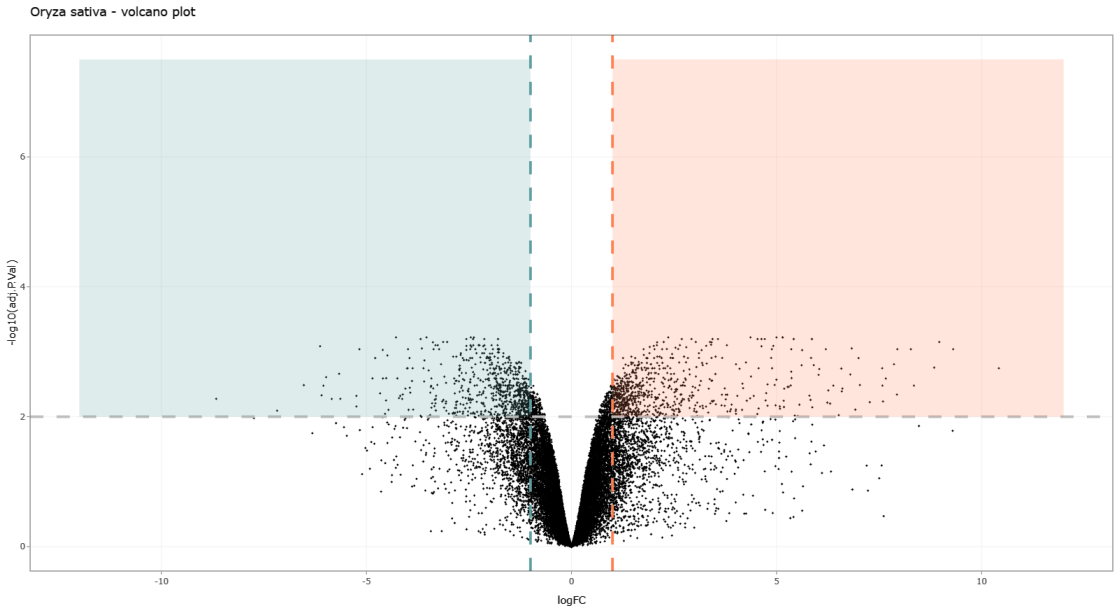
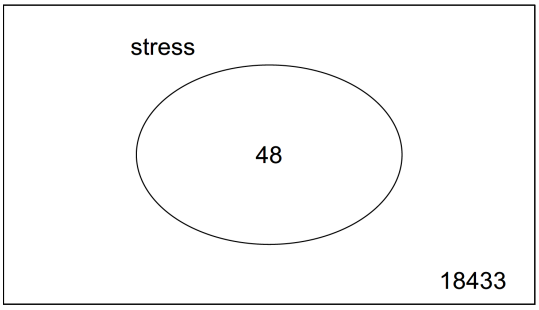


Figure 17: Venn diagrams of the DEGs

i) O. nivara



ii) O. sativa

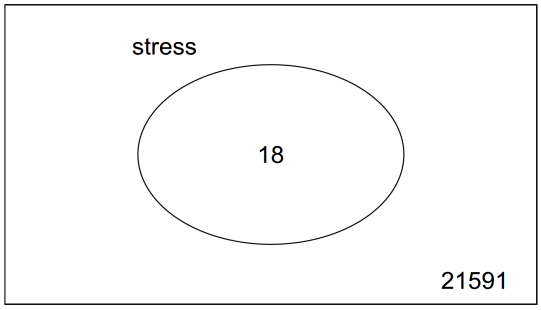


Figure 18: Heatmap of the DEGs - O. nivara

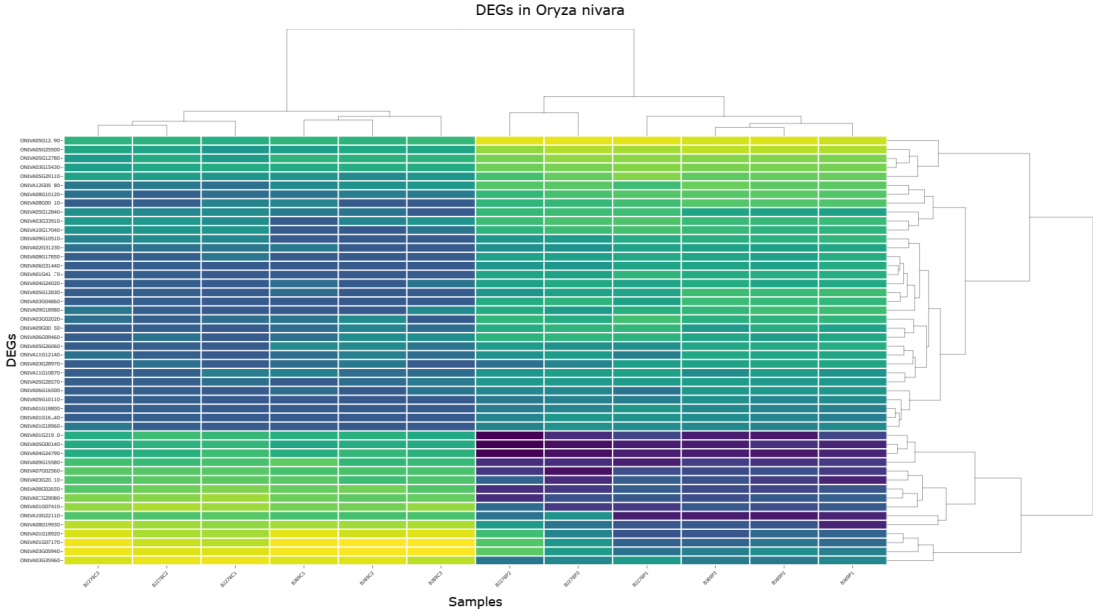


Figure 19: Heatmap of the DEGs - O. sativa

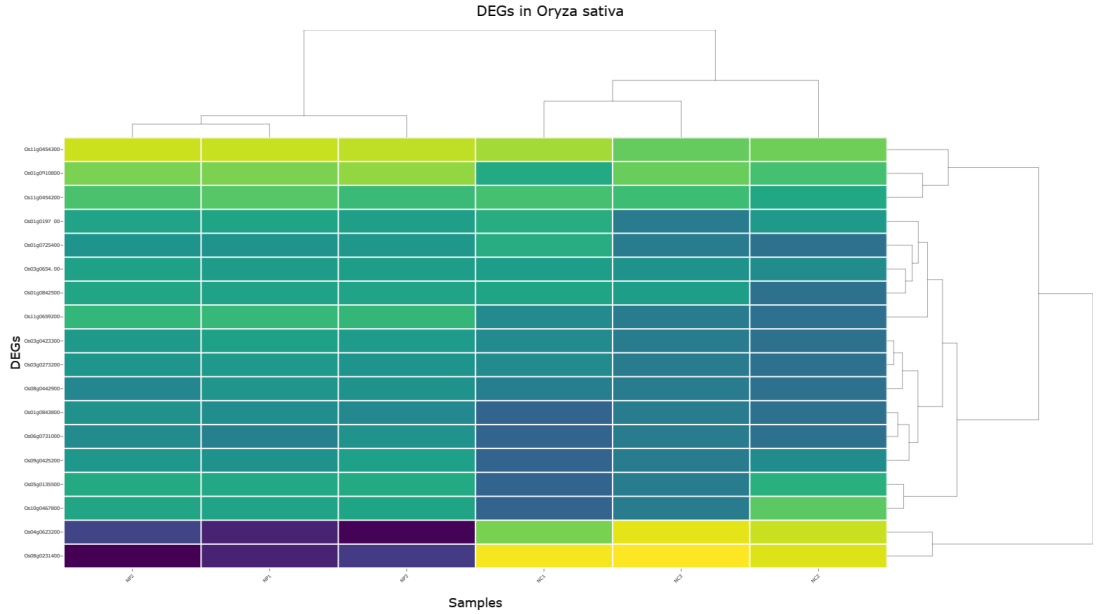
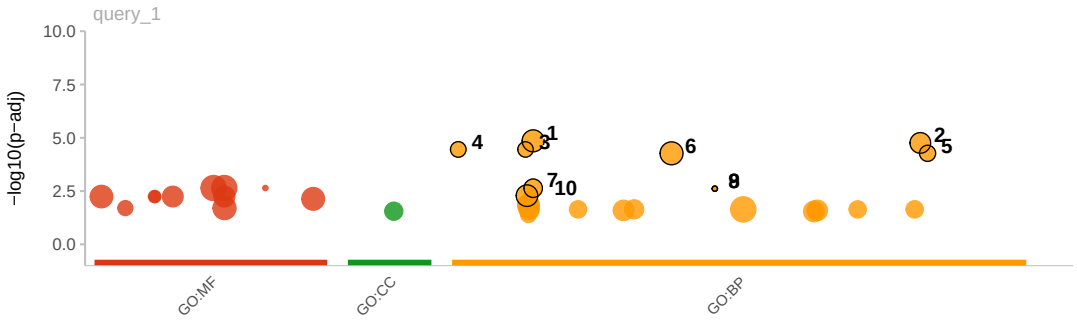


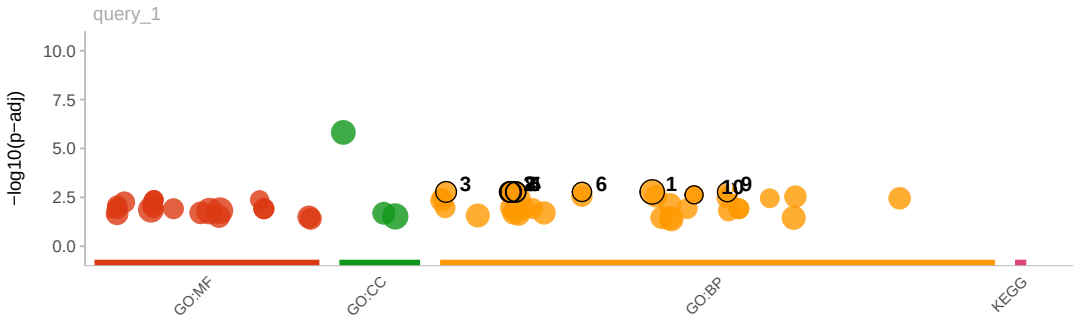
Figure 20: Manhattan plot with the first 10 top-ranked GO terms highlighted - O. nivara



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

5.1 Critical evaluation of the results

Discuss the quality and reliability of the RNA-seq data and the downstream analyses.

5.2 Biological implications

Discuss the potential implications of the findings for plant biology and the broader scientific community.

5.3 Limitations and future directions

Address the limitations of the current analysis and suggest possible future directions to expand on the findings.

6 Conclusion

Summarize the main findings of the assignment, reiterating the significance of the results, and provide a final statement on the overall outcome of the study.

References

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

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

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].

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

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

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