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

Ingo Giebel

Math.-Nat. Fakultät, Heinrich-Heine-Universität Düsseldorf

QBio304: Applied Bioinformatics

Prof. Dr. Björn Usadel

Dr. Jedrzej Jakub Szymanski

May 1, 2023

Contents

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.

Therefore, the data is 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 checks if the reads in the FastQ file contain a significant amount of adapter sequences

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

3.3 Mapping to the respective genome

Detail the reference genome used and the bioinformatics tools employed for the mapping process.

3.4 Statistical evaluation and differential expression analysis

Explain the statistical methods and software used for evaluating the data and identifying differentially expressed genes.

3.5 Functional enrichment analysis

Describe the tools and databases used to perform functional enrichment analysis to interpret the biological significance of the differentially expressed genes.

4 Results

4.1 Quality evaluation

Present the findings from the quality evaluation of the selected RNA-seq data.

4.2 Mapping efficiency and coverage

Report the results of the mapping process, including the mapping efficiency and coverage.

4.3 Exploratory data analysis

Discuss the dominating variance components, reproducibility, possible batch effects and confounding variables.

4.4 Differentially expressed genes

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

4.5 Functional enrichment analysis

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

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/

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

Index