Transcriptomic Analysis of Salt Stress Responses Across Time in Capsicum annuum

George David Apostolidis[1*]

[1]Wageningen University & Research, The Netherlands .

Corresponding author(s). E-mail(s): george.apostolidis@wur.nl;

Abstract

Salinity stress is a significant abiotic factor affecting agricultural output globally. That's why a better knowledge of the molecular mechanisms behind plant stress responses is necessary. In the following experiment, transcriptional responses to Capsicum annuum were examined using an automated RNA-seq analysis pipeline. The pipeline combines quality control, read alignment, gene-level quantification, differential expression analysis and functional interpretation using Gene Ontology enrichment. Clear separation between salt-stressed and mock-treated samples were revealed through the application of the pipeline to a publicly available dataset. The results indicate strong stress-induced transcriptional reprogramming. Differential expression analysis identified numerous genes with significant expression changes, while enrichment analysis highlighted biological processes related to oxidative stress, osmotic regulation, water deprivation, and chloroplast function. Statistical analyses showed that the main biological conclusions were stable across different aligners and analysis settings. In conclusion, the developed pipeline can capture biologically meaningful stress responses and provide a framework for transcriptomic analysis that can be extended to other stress conditions of plant species.

Keywords: RNA-seq, salt stress, Capsicum annuum

1 General Introduction

Plants are constantly exposed to a variety of environmental pressures that have a significant impact on their development, growth, and productivity. Stresses like these can be divided into two categories: abiotic (such as salt, drought, and high temperatures) and biotic (such as pathogens and herbivores). Stress exposure causes extensive

molecular changes in transcript processing, gene expression, and regulatory pathways, allowing plants to adjust to unfavorable circumstances.

Salinity stress is one of the most difficult abiotic challenges affecting global agriculture. Salinization of soil reduces water uptake, disrupts ion homeostasis, and interferes with vital metabolic functions, which eventually result in lower agricultural yields. Salinity stress is predicted to worsen in many agricultural areas due to increased irrigation, climate change, and soil degradation. Therefore, it is important to identify the genes and biological processes that contribute to salt tolerance.

Capsicum annuum (pepper) is a well-researched crop species known for its nutritional and economic value. It has also become a valuable model for studying plant stress responses because it consists of high-quality reference genomes, extensive transcriptomic datasets, and well-curated functional annotation resources. Large-scale RNA sequencing experiments have been produced for pepper under several stress conditions, providing a robust foundation for systematic analysis of stress-induced transcriptional changes. It is commonly known that RNA sequencing has emerged as a powerful technology for measuring genome-wide gene expression with high sensitivity and resolution, enhancing the identification of differentially expressed genes. Carefully designed computational pipelines are needed to extract biologically meaningful insights from RNA sequence results. These pipelines can consist of quality control, alignment, statistical modeling, and functional interpretation in a reproducible and scalable manner.

The primary focus of this experiment is to develop an automated RNA sequence analysis pipeline. By applying this tool we are trying to capture patterns in transcriptional responses to salt stress in Capsicum annuum. In particular, the project tries to identify genes that are differentially expressed under salt stress conditions. It also determines which biological processes and pathways are significantly affected by these expression changes. We want to try to move beyond lists of genes and toward a system-level understanding of how salt stress reshapes cellular functions by combining differential expression analysis with gene ontology enrichment.

The central research questions guiding this work are

- Which genes are differentially expressed in Capsicum annuum under salt stress conditions?
- Which Gene Ontology (GO) categories and biological pathways are significantly enriched as a result of these transcriptional changes?

In order to answer the above questions, we implemented a reproducible computational workflow, capable of processing raw RNA-seq data through to biologically interpretable results. The workflow includes handling of sequencing quality, read alignment to the reference genome, accurate quantification of gene expression, and statistical identification of differential expression. Lastly, functional enrichment analysis is used to place expression changes in context within known biological processes relevant to stress adaptation.

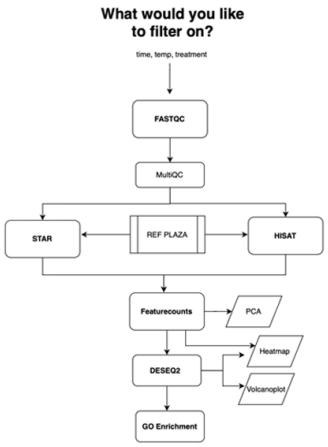

Fig. 1 The RNA sequence pipeline

2 Design & Implementation

The design and the implementation of the pipeline emphasize modularity and reproducibility. Specifically, we allow each analytical step to be independently validated while maintaining a coherent end-to-end process. All the stages of the pipeline were developed independently as discrete components with clearly defined inputs and outputs. This ensures that the results can be inspected easily, alternative datasets can be adapted, or settings can be changed. A structured design, like the above, ensures that the workflow remains strong to changes in experimental design. Examples can be the inclusion of additional samples or time points, while preserving consistency across all analyses.

We were given access to a high-performance computing system, so the pipeline was designed to run analysis on this system because of the size of the RNA sequence databases. It was implemented in Python with explicit checks, logging, and directory management, ensuring solid execution across samples and conditions. Version control using GitLab was integrated into the workflow from the beginning, with a structured branching strategy (main, dev, and individual contributor branches) that allowed parallel development, iterative testing, and controlled integration of new features.

The pipeline architecture, as shown in Figure 1, follows the standard RNA-seq analysis workflow, consisting of quality control, read mapping, read counting, normalization, differential expression analysis, and functional interpretation. Each step was implemented as a distinct module, allowing individual components to be tested, replaced, or optimized independently without disrupting the overall workflow.

The first step to assess the integrity of the raw FASTQ files was implemented as the quality control. This step makes sure that the analysis was not confounded by poor sequencing quality. In general, read quality remained consistently high across positions. In such a case, applying trimming would shorten reads and reduce alignment accuracy. So, in order to preserve maximum sequence information and avoid introducing bias,

raw reads were used directly for downstream analyses, in line with recommended RNA sequence best practices.

For our next step, read mapping, we gave the option to the user to choose between two aligners, HISAT2 and STAR. HISAT2 was selected as the primary aligner for our analysis due to its high accuracy, efficient memory usage, and resilience when aligning RNA sequence reads to large genomes, and it's a perfect suit for execution on the Linux Server. STAR is an alternative option for mapping for comparison and validation purposes. The HISAT2 implementation included explicit index generation based on the Capsicum annuum reference genome, with appropriate handling of splice junctions and transcript annotations to ensure accurate alignment. We also added absolute path resolution and binary availability checks to ensure that everything runs smoothly on the system.

Read quantification was performed using featureCounts, a tool known for its efficiency and compatibility with downstream differential expression tools. In order to generate consistent gene identifiers and sample ordering, the pipeline creates automated count matrices across all samples and experimental conditions. It was very important to correctly extract and store experimental metadata, like time points and conditions, because this data has a direct impact on the design matrix that was used for differential expression analysis.

PyDESeq2, a Python-based reimplementation of the DESeq2 statistical framework, was used to perform differential expression analysis. Maintaining a fully Python-based workflow while utilizing a proven statistical model for RNA-seq count data was the driving force behind this decision. In order to guarantee that experimental contrasts are established uniformly across runs, the pipeline automatically creates the necessary count matrices and metadata tables.

To reduce the false discovery rate, significant differentially expressed genes were chosen using modified p-values. Flexible downstream interpretation and visualization are made possible by the pipeline's delivery of both filtered gene lists and complete result tables. Functional enrichment analysis was added as a downstream module to convert differential expression results into biological insight. Using Capsicum annuum-specific GO annotations from the Pepper GO and PLAZA resources, Gene Ontology enrichment was carried out.

To enable consistent integration of expression results with functional databases, additional scripts were developed to map gene identifiers between various annotation systems. In order to support the project's main objective of finding stress-responsive genes and pathways important for crop improvement, enrichment analyses concentrated on identifying biological processes and pathways linked to salt stress responses.

The pipeline's initial implementation concentrated on achieving end-to-end functionality, which allowed for the processing of raw sequencing data all through to differential expression findings. Refactoring and optimization took a lot of work after a functional version was created. This involved reorganizing the codebase for maintainability, clarity, and modularity; enhancing file existence checks and directory handling; and making sure that intermediate steps could be omitted if already finished, minimizing needless recomputation. Significant improvements were made to STAR mapping

scripts, including better error handling, index generation optimization, and parameter tuning. In the same way, to address edge cases and guarantee compatibility with PyDESeq2, the logic for count matrix generation and metadata handling was iteratively improved. Multiple levels of pipeline validation were carried out. While complete pipeline runs were utilized to confirm end-to-end consistency, individual modules were tested separately to ensure proper input–output behavior. The reliability was evaluated by comparing various mapping strategies (STAR vs. HISAT2) and repeating the process on the same datasets. Results could be linked to particular pipeline runs and parameter settings thanks to logging and directory structure conventions.

3 Applications & Results

In a publicly available Capsicum annuum dataset, salt-stressed samples were compared to mock-treated controls at different time intervals (0, 3, 6, 12, 24, and 72 hours) using the finished RNA-seq processing methodology. The objectives were to identify transcriptional changes caused by salt stress and determine which biological processes are affected over time. All analyses were performed utilizing the automated approach described in the Methods section to guarantee uniform processing across samples and settings.

The initial FastQC quality examination of raw sequencing data revealed that all samples had generally high sequencing quality. Per-base quality scores were consistently high, and there was no sign of significant adapter contamination or systematic quality degradation. According to the MultiQC report, read quality and sequencing depth were comparable for salt-stressed and mock-treated samples, indicating that technical artifacts are unlikely to be the cause of downstream expression differences.

The readings were aligned to the Capsicum annuum reference genome (Pepper Genome v1.6) using HISAT2. The reference genome's indexes were pre-generated on the server, and all samples were mapped using the default alignment parameters. Salt stress did not cause systemic biases that affected read mappability, as evidenced by the consistently high mapping efficiency across all samples and the lack of significant differences between treatments or time periods.

Aligned readings were processed using StringTie in order to create reference-guided transcripts. Transcript assembly was performed for each sample with the discovery of novel transcripts disabled to guarantee that only annotated genes were quantified. The merging function in StringTie was then used to create a single transcriptome for every sample. The merged GTF contained StringTie-specific transcript identifiers (MSTRG IDs), each of which was linked to a corresponding reference gene. These identifiers were substituted with reference gene IDs using a custom Python script to guarantee consistent gene-level interpretation. The gene-level read counts that were obtained from the combined annotation using featureCounts were used in all subsequent studies.

Principal component analysis was performed on log2-transformed gene counts to assess global expression patterns and sample clustering. The first two primary components accounted for 22.8% and 18.3% of the overall variation, respectively. PCA revealed a clear difference between salt-stressed and mock-treated samples along the first main component, indicating that salt stress is a significant factor impacting

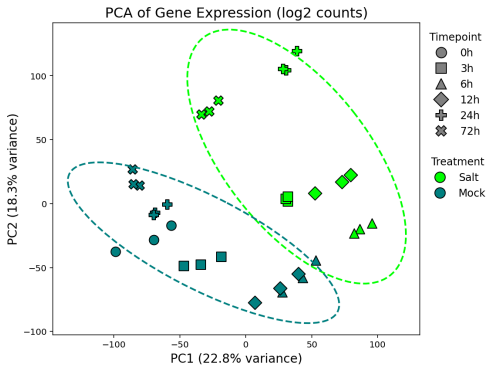

Fig. 2 PCA

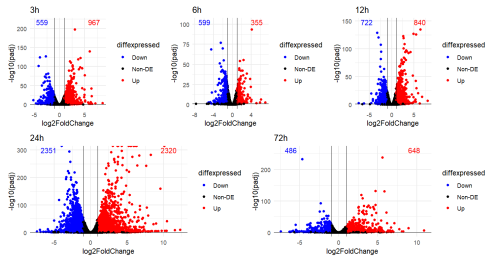

Fig. 3 Differential gene expression analysis

transcriptional variation 2. Temporal structure was also observed along the second main component. In mock-treated samples, time points at 0, 24, and 72 hours were closely grouped, which is consistent with circadian modulation of gene expression. Nevertheless, samples treated with salt at 24 and 72 hours did not cluster together, suggesting that long-term salt stress overrides circadian effects and results in distinct transcriptional states across time.

Differential expression analysis was carried out using DESeq2, comparing salt-stressed samples to sham controls at each time point independently. If a gene's corrected p-value (FDR) was less than 0.05, it was considered differentially expressed.

The distribution of log2 fold changes and statistical significance at each time point are shown by volcano plots 3. Numerous genes were both up-regulated and down-regulated in response to salt stress, with the number and amount of differentially expressed genes changing over time. At 24 hours, the greatest transcriptional response was seen, with the greatest number of genes exhibiting notable changes in expression. Fewer genes showed altered expression at earlier time points (3–12 hours), suggesting a gradual activation of stress-responsive pathways. The number of genes with differential expression dropped after 72 hours, indicating a partial adaptation to extended stress.

To better explore expression dynamics, a heatmap was created using the top 30 most significant differentially expressed genes across all time points (absolute log2 fold change ¿ 1 and adjusted p-value ¡ 0.05). Gene expression levels were log2-transformed and standardized using z-scores to facilitate cross-sample comparisons.

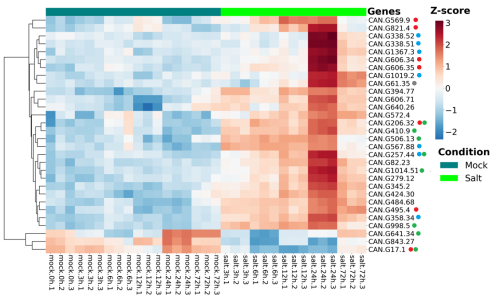

Fig. 4 Expression patterns of highly responsive genes (heatmap)

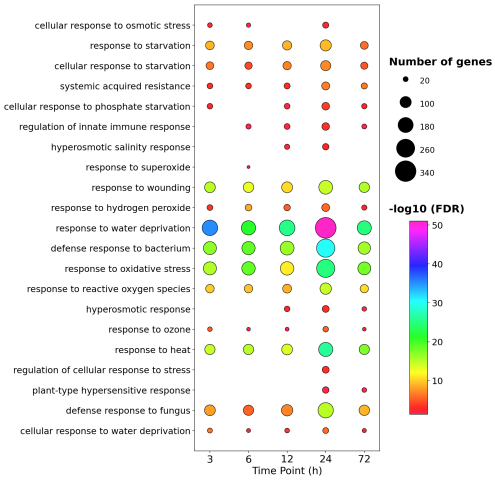

Fig. 5 Gene Ontology enrichment analysis

Hierarchical clustering demonstrated a clear separation between the salt-stressed and mock-treated samples 4. Most of the selected genes had low or baseline expression in mock samples, which was significantly increased after salt stress, particularly at 24 and 72 hours. Clusters of genes with similar expression profiles indicate coordinated regulation of stress-response pathways rather than distinct gene-specific effects.

To explain the results of differential expression, Gene Ontology (GO) enrichment analysis was performed using PLAZA annotations specific to capsicum annuum. Without using an additional log2 fold-change criterion, all genes with an adjusted p-value less than 0.05 were categorized as differentially expressed in order to maximize sensitivity.

GO enrichment revealed a notable overrepresentation of biological processes associated with oxidative stress, dehydration, bacterial defense, and response to reactive oxygen species 5. These methods are consistent with the known physiological effects of salt stress, which include osmotic imbalance and ionic stress, which increase the production of reactive oxygen species (ROS) and activate defense mechanisms.

Several GO keywords, including defense reaction to fungus and response to heat, were most common during the 24-hour period. These responses most likely reflect a general stress-response mechanism rather than stress-specific signaling, as genes often have many specified roles. The biggest enrichment signals and the largest number of associated genes that were consistently observed at 24 hours may indicate the peak of transcriptional reprogramming following salt exposure.

Overall, the enriched GO keywords that closely resemble those identified in the reference study include defense-related activities and oxidative stress. The greater number of genes associated with these terms in this study is likely due to the use of GO annotations exclusive to Capsicum annuum rather than cross-species annotation.

These results collectively demonstrate that Capsicum annuum undergoes substantial, time-dependent transcriptional changes as a result of salt stress. Global expression patterns significantly differentiate between salt-stressed and control samples, with the most pronounced effects occurring at 24 hours. Differential expression and enrichment analyses often reveal osmotic stress, oxidative stress, and defense mechanisms as important components of the salt-stress response. These findings confirm that the pipeline reliably detects physiologically important stress responses and provides an integrated platform for transcriptome study.

4 Discussion & Conclusions

The primary objective of this experiment was to reproduce and extend published findings on salt stress-induced transcriptional responses in Capsicum annuum using an independently developed RNA-seq analytic process. The results of the study are broadly consistent with previous findings in plant stress biology, particularly the stimulation of genes associated with oxidative stress, osmotic regulation, and water restriction. The enrichment of Gene Ontology categories linked to these processes indicates a considerable degree of reproducibility at the level of biological interpretation, suggesting that the pipeline successfully captures canonical stress-response pathways.

Even though the project did not try to replicate a specific published paper in a detailed, gene-by-gene manner, the overlap in functional categories and global expression trends supports the conclusion that the fundamental biological conclusions are repeatable. Slight differences in individual gene rankings or effect sizes, which are expected given variances in alignment techniques, annotation versions, and statistical implementations, do not affect the overall consistency of the results.

In addition to identifying differentially expressed genes, the pipeline enabled additional exploratory research that supported biological interpretation. Principal component analysis revealed a clear difference between salt-stressed and control samples, indicating that stress condition and time are the primary drivers of transcriptional variance. This work supports the hypothesis that salt stress results in coordinated, system-wide changes rather than distinct gene-level effects.

Visualization tools such as heatmaps and volcano plots further supported this notion by highlighting ordered expression patterns across functionally related genes. These methods provided complementary perspectives on the data, allowing effect magnitude and statistical significance to be taken into account at the same time.

Importantly, the consistency of the data from multiple analytical viewpoints increases confidence in the biological importance of the results.

An essential part of this study was evaluating the robustness of the results with respect to methodological choices. The pipeline supported many aligners, and a comparison between STAR and HISAT2 revealed that the mapping tool choice had no discernible impact on the primary findings. Despite slight variations in absolute counts and individual gene rankings, global expression patterns, sample clustering, and enriched biological processes were consistent between aligners.

Similarly, using PyDESeq2 for differential expression analysis did not significantly deviate from the expected DESeq2-based behavior. Although minor differences in implementation details may affect dispersion estimates or p-values for borderline genes, the overall set of substantially enriched pathways remained intact.

Parameter settings, particularly those related to alignment and information encoding, were found to influence technical aspects of the analysis when properly adjusted, but they had minimal impact on high-level biological interpretation. These results demonstrate the pipeline's resilience to minor methodological variation and emphasize the importance of careful parameter selection and validation.

The pipeline that was developed has several benefits. Its modular design allows for transparency, reproducibility, and ease of expansion; its interaction with version control allows for coordinated collaboration and iterative improvement. Automated inspections, logging, and directory management improved usability and reduced the likelihood of silent failures on an HPC system.

However, the pipeline is not without its constraints. Although it allows for numerous aligners and downstream analysis, time constraints prohibited thorough benchmarking across tools and parameter spaces. Additionally, the pipeline focuses mostly on gene-level expression rather than fully exploring alternative splicing or isoform-level regulation, which may be critical in stress responses. Functional interpretation was restricted to Gene Ontology enrichment, but it might be expanded to include pathway-level modeling or network-based analysis.

Additionally, even if the pipeline is repeatable in its current form, containerization or workflow management tools like Snakemake or Nextflow would enhance reproducibility across different computing systems. These improvements would further increase portability and scalability for future study.

5 Author Contributions

The RNA-seq analysis pipeline was developed through a collaborative effort involving George, Demian, Laurens, Ilse, and Elif, with clearly defined responsibilities across different stages of the workflow.

Demian was in charge of implementing the read-mapping workflow using HISAT2 and converting StringTie and HISAT2 execution into Python-based scripts and Snakemake-compatible components. This work generated BAM and GTF files for all salt and mock samples and formed the basis for the final mapping strategy. Demian also helped combine StringTie outputs for further analysis and created the initial version of the GO enrichment analysis script.

Ilse was primarily responsible for testing and benchmarking STAR-based alignment, which included command optimization, index construction, and multi-sample test runs. In order to give time-point comparisons and project-specific data, Ilse also expanded and modified an existing R methodology to develop the volcano plot display for differential expression results.

Elif designed and tested workflows for quality control and filtering, including FastQC processing, and wrote scripts for read-level quality evaluation. These components ensured data integrity prior to mapping and further analyses.

Laurens focused on downstream and functional analysis, including writing scripts for KEGG gene-pathway retrieval, assessing pathway enrichment, and determining whether gene identifiers in PLAZA, NCBI, and KEGG databases are compatible. Laurens also performed BLAST-based mapping of PLAZA genes to the Capsicum annuum NCBI genome and KEGG pathway enrichment on the resultant differentially expressed gene sets.

George was responsible for coordinating the execution of backend analyses, creating read counts using featureCounts, and building and integrating the PyDESeq2 differential expression analysis pipeline. In addition to managing pipeline integration and merging, George resolved inter-script dependencies, debugged and stabilized the end-to-end workflow across development branches, and contributed to documentation and presentation materials.

Each author took part in testing, debugging, debating methodological choices, and assessing the final pipeline and results. All authors approved the final analysis and workflow.

When difficulties developed during the development of this work, ChatGPT and Gemini, two generative AI tools, were only used for background information, conceptual explanation of biological processes, and general coding guidance. No text, code, or analysis generated by AI was specifically included in the final pipeline or report. All of the coding, analysis, and written content used in this study were generated, reviewed, and validated by the authors.

The usage of generative AI was limited, transparent, and documented through individual logbooks and group discussions in accordance with the group's AI use agreement. AI techniques were not used for reference generation or citation; all scientific references were selected and independently verified by the authors. The authors are solely responsible for the study's final content, methodology, and findings.