Identification of Stress-responsive Genes in Differential Co-expression Networks

Application in Rice and Sugarcane

Camila Riccio-Rengifo

Advisors:

Camilo Rocha

Jorge Finke

Thesis presented as a partial requirement to opt for the title of Doctor in Engineering and Applied Sciences

Facultad de Ingeniería y Ciencias

Pontificia Universidad Javeriana

Cali, Colombia

November 1, 2023

Identification of Stress-responsive Genes in Differential Co-expression Networks

Abstract

Understanding how organisms respond to environmental stress is crucial in the field of biology. Stress-responsive genes play a pivotal role in this adaptation and their identification is a significant challenge. Traditional approaches to find these genes has limitations in capturing the complex genetic interactions governing stress responses. This thesis introduces the Control-Stress Data Integration with Overlapping Clustering (CSI-OC) Workflow. The proposed approach combines statistical and network-based methods to identify stress-responsive genes. It uses gene expression change metrics and constructs differential co-expression networks to analyze gene interactions. Its most distinctive feature is the ability to detect overlapping gene modules and select those related to phenotypic traits. The workflow is applied in rice and sugarcane, two important crops in agriculture, unveiling key genes related to stress response. These overlapping modules prove vital in understanding co-expression networks and stress response. Experimentation with synthetic data validates the workflow's reliability. Ultimately, this disertation enriches the field of biology by providing a robust analytical tool for identifying stress-responsive genes. The detection of overlapping modules makes a notable progression, mirroring the intricate dynamics of gene interactions. Moreover, the potential applications of this workflow extend beyond biology, encompassing areas like economy and social sciences, where understanding interactions is key to comprehending systemic phenomena.

Identificación de Genes que Responden al Estrés en Redes de Co-expresión Diferencial

Resumen

Comprender cómo los organismos responden al estrés ambiental es crucial en el campo de la biología. Los genes que responden al estrés desempeñan un papel fundamental en esta adaptación y su identificación es un desafío significativo. Los enfoques tradicionales para encontrar estos genes tienen limitaciones para capturar las complejas interacciones genéticas que gobiernan las respuestas al estrés. Esta tesis introduce el flujo de trabajo de Integración de Datos de Control-Estrés con Agrupación Superpuesta (CSI-OC por sus siglas en inglés). La aproximación propuesta combina métodos estadísticos y basados en redes para identificar genes que responden al estrés. Utiliza métricas de cambio en la expresión génica y construye redes de co-expresión diferencial para analizar las interacciones génicas. Su característica más distintiva es la capacidad de detectar módulos de genes superpuestos y seleccionar aquellos relacionados con rasgos fenotípicos. El flujo de trabajo se aplica en arroz y caña de azúcar, dos cultivos importantes en la agricultura, revelando genes clave relacionados con la respuesta a estrés. Estos módulos superpuestos resultan vitales para comprender las redes de co-expresión y la respuesta al estrés. La experimentación con datos sintéticos valida la fiabilidad del flujo de trabajo. En última instancia, esta disertación enriquece el campo de la biología al proporcionar una herramienta analítica sólida para identificar genes que responden al estrés. La detección de módulos superpuestos representa un progreso notable, reflejando la intrincada dinámica de las interacciones génicas. Además, las aplicaciones potenciales de este flujo de trabajo se extienden más allá de la biología y abarcan áreas como la economía y las ciencias sociales, donde entender las interacciones es clave para comprender los fenómenos sistémicos.

Acknowledgments

I want to express my gratitude to my advisors, Jorge Finke and Camilo Rocha, for their guidance, support, and mentorship throughout the research and writing process. In particular, I would like to express my admiration for Professor Rocha, who exemplifies a remarkable sense of responsibility, methodical approach, and efficiency.

I also want to express my profound appreciation for my defense committee, who generously provided invaluable insights and constructive feedback on this project. Thanks also to the Pontificia Universidad Javeriana Cali and all the professors who contributed to my academic training. Aditionally, this endeavor would not have been possible without the support from the OMICAS program, who financed my research.

My gratitude extends to everyone with whom I had the opportunity to collaborate on research projects. A special thanks to Dr. Mauricio Ramirez, who, in his postdoctoral role within the OMICAS program, provided invaluable support and research skills that were fundamental in interpreting the biological aspects of my work's results.

I am also grateful to my lab mates, colleagues, and research team for the fun times we had working and socializing together. Special thanks to my friends for their unwavering support, encouragement, and for making the journey more enjoyable. Above all, my heartfelt gratitude goes to those who, even from a distance, provided emotional support during my darkest times.

Lastly, I owe an immeasurable debt of gratitude to my family: my parents, Marisol Rengifo and Giancarlo Riccio, and my sister, Giulliana Riccio. They have always been my primary motivation, my most honest critics, and an endless source of unconditional love and support.

Contents

1 Introduction **10**

1.1 Motivation and Problem 10

1.2 Approach 13

1.3 Summary of Chapters and Contributions 14

2 Preliminaries **16**

2.1 Mathematical Foundations of the Workflow 16

2.1.1 Log Fold Change (LFC) 16

2.1.2 Differential co-expression network (DCN) 17

2.1.3 Hierarchical Link Clustering (HLC) 17

2.1.4 Least Absolute Shrinkage Selector Operator (Lasso) 18

2.2 Workflow Evaluation Approaches 19

2.2.1 Computational validation 19

2.2.2 Knowledge-based evaluation 20

2.3 Synthetic data generation 21

2.3.1 GAN 21

2.3.2 cGAN 22

3 The CSI-OC Workflow **23**

3.1 Data Pre-processing 24

3.2 DCN Construction 25

3.2.1 Pearson-based Co-expression Networks 26

3.2.2 Lasso-based Co-expression Networks 27

3.3 Overlapping Module Detection 28

3.4 Relevant Module Selection 29

3.5 Discussion 30

4 Identifying Salt-Responsive Genes in Rice 32

4.1 Origin of Data 33
4.2 Summary of Selected Genes 34
4.3 Overlapping Genes Characterization 35
4.4 Computational Validation 36
4.5 Knowledge-based Evaluation 37
4.6 Discussion 39

5 Identifying Drought-Responsive Genes in Sugarcane 42

5.1 Origin of Data 44
5.2 Pearson vs Lasso Networks 45
5.3 Summary of Selected Genes 46
5.4 Overlapping Genes 47
5.5 Computational Validation 48
5.6 Knowledge-based Evaluation 49
5.7 Discussion 53

6 CSI-OC Workflow Performance and Validation with Synthetic Data 58

6.1 CoSynthEx 60
6.2 Real input data Statistics 61
6.3 CSI-OC with CoSynthEx 63
6.4 Results 65
6.5 Discussion 68

7 CSI-OC Validation in a Non-biological Context 70

7.1 Origin of the Data 72
7.2 Preliminary Statistical Analysis 73
7.3 Summary of Selected Products 74
7.4 Overlapping Products Characterization 74
7.5 Computational Validation 75
7.6 Knowledge-based Evaluation 76
7.7 Discussion 78

8 Conclusion and Future Work **80**

8.1 Conclusion 80

8.2 Future Work 81

Nomenclature **83**

Bibliography **101**

List of Figures

3.1 CSI-OC workflow 23
3.2 CSI-OC data pre-processing 25
3.3 Pearson-based DCN construction 26
3.4 Lasso-based DCN construction 27
3.5 Relevant modules selection 29

4.1 Salt-responsive rice genes selection process 34
4.2 Distribution of rice genes across overlapping modules. 35
4.3 Computational validation of salt-responsive rice genes 36

5.1 Drought-responsive sugarcane genes in leaf and root across stress levels 47
5.2 Classifier validation of drought-responsive sugarcane genes 49
5.3 Regressor validation of drought-responsive sugarcane genes 50
5.4 DEGs vs CSI-OC selected genes in sugarcane 51
5.5 Enriched BPs from CSI-OC selected sugarcane genes in CML 52
5.6 Enriched BPs from CSI-OC selected sugarcane genes in root 53

6.1 CoSynThEx cGAN architecture 60
6.2 Violin plots of rice expression data 61
6.3 Violin plots of rice phenotypic traits 62
6.4 Framework for CSI-OC validation with synthetic data 64
6.5 Loss Curves of CoSynthEx cGAN Training 66
6.6 Performance heatmap of CSI-OC workflow across synthetic scenarios 67

7.1 Distribution of transaction data across stores 73
7.2 Enrichment of overlapping products with product families. 75

7.3 Computational validation of weekend-responsive products 76
7.4 Selected products enrichment and upsetplot 77
7.5 Enrichment of CSI-OC selected products with product families. 77

1. Introduction

In the intricate realm of plant biology, deciphering the molecular mechanisms that underlie responses to environmental stressors stands as a formidable challenge. Stress-responsive genes play a pivotal role in orchestrating a plant's ability to adapt and thrive under adverse conditions, making them subjects of profound scientific interest. Stress-responsive genes are the molecular architects of a plant's resilience. They encode proteins that facilitate cellular, physiological, and metabolic adjustments, ultimately determining a plant's survival and productivity. Harnessing the understanding of these genes offer a path toward crop improvement and agricultural sustainability.

1.1 Motivation and Problem

In this pursuit of stress tolerance, rice (*Oryza sativa*) and sugarcane (*Saccharum* spp. hybrid) take the center stage. These two crops have emerged as primary subjects of interest due to their relevance in addressing agricultural challenges. As the global population burgeons, and climate change exacerbates, the cultivation of stress-resilient crops is no longer a mere aspiration, but an imperative. Rice, serving as a dietary staple for billions, faces the onslaught of factors such as salinity, which jeopardize its yield and, consequently, global food security (Zeng and Shannon, 2000; Chang et al., 2019; Taratima et al., 2022). Sugarcane, a source of renewable energy and vital in the production of sugar and biofuels, confronts its own set of environmental challenges, with drought as a potent adversary (Contiliani et al., 2022, 2023; Kaura et al., 2022; Li et al., 2023; Shrestha et al., 2023; Verma et al., 2022b; Zahoor and Babar, 2023). The focus in rice and sugarcane crops, particularly in Colombia, aligns with the objectives of the OMICAS program (Jaramillo-Botero et al., 2022) that encompasses this research.

Building on this foundation, this research addresses the central challenge of comprehensively and systematically identifying stress-responsive genes using an in-silico approach, with a particular application focus on rice and sugarcane. Despite significant advancements in genomics

and molecular biology, the problem of pinpointing these crucial genes persists. The conventional method for identifying stress-responsive genes primarily relies on statistics, involving the identification of differentially expressed genes (DEGs). This approach hinges on detecting significant changes in mean gene expression between control and stress conditions (Pan, 2002; Reiner et al., 2003). However, this statistical approach has inherent limitations, as it may not fully capture the intricate and many interactions between genes responsible for stress responses.

To address this limitation, network-based approaches have emerged as valuable tools for capturing the complex relationships between genes within co-expression networks. These co-expression networks consist of nodes representing genes and edges representing relationships between gene expression patterns (Rao and Dixon, 2019). Gene relationships are often quantified using correlation metrics, with the Pearson correlation coefficient being a widely used method (Pearson, 1920). Nevertheless, researchers continue to explore alternative metrics and approaches for connecting nodes to enhance the inference of network topology (Reshef et al., 2011; Song et al., 2012; Mateos et al., 2019).

In light of this, numerous algorithms have been developed for analyzing co-expression networks, with one of the most widely used being Weighted Gene Co-expression Network Analysis (WGCNA) (Langfelder and Horvath, 2008). WGCNA has gained popularity in bioinformatics and biomedical research, allowing for the construction of gene co-expression networks and the identification of co-expressed genes. Its applications extend to disease biomarker discovery, gene function prediction, protein-protein interaction inference, and genetic variant detection in cancer research (Sundarrajan and Arumugam, 2016; Yuan et al., 2018; Li et al., 2017; Kanonidis et al., 2016; Hou et al., 2021). When comparing control and stress conditions, including in WGCNA and similar network-based approaches, two common methods are employed. One stacks the control and stress data to create a unified dataset, preserving the unique characteristics of each condition, but potentially increasing computational demands, particularly with substantial datasets. Alternatively, researchers can conduct the analysis separately for control and stress data, followed by the combination of control and stress networks through a statistical method to create a differential co-expression network (Choi et al., 2005; Fukushima, 2013; Yang et al., 2013; Liu, 2018). This second method offers a more detailed understanding of the differences in gene behavior between the two conditions. However, it requires running the co-expression network analysis twice, once for each condition.

Furthermore, a crucial aspect of co-expression network analysis is the identification of co-

expressed genes organized into groups, known as modules. Genes within the same module often share the same transcriptional regulatory pathways, are functionally related, or are members of the same pathway or metabolic complex (Fionda, 2019). While various algorithms for detecting modules in co-expression networks have been developed and assessed (Saelens et al., 2018), many of these methods tend to focus on identifying disjoint (non-overlapping) modules. Nevertheless, addressing the issue of overlapping modules is essential due to the increasing evidence supporting the highly combinatorial nature of gene regulation, where gene products participate in multiple pathways, exhibit multiple biological functions, and belong to several protein complexes simultaneously (Oeckinghaus et al., 2011; van de Peppel and Holstege, 2005; Gavin et al., 2002). The flexibility to accommodate these overlapping modules is critical for a more accurate representation of network structures and functions.

Another primary objective in co-expression analyses is the identification of biologically or clinically relevant modules (Langfelder and Horvath, 2008). This process involves condensing the node profiles within a specific module into a representative entity known as the eigengene. The eigengene, typically corresponding to the hub gene or the first principal component of the gene profiles, is crucial for assessing the relevance of its module. Module relevance is determined by its impact on sample traits, such as phenotypic traits in the context of plants. Simple linear regression models (James et al., 2023) are the traditional choice for measuring the relationship between eigengenes and traits. They determine the module's impact on the trait. However, there is an opportunity for further enhancement using advanced regression techniques (James et al., 2021) to improve the precision and robustness of the intricate relationships between modules and traits.

In current approaches, non-overlapping module detection often requires large module sizes to prevent an overwhelming number of modules from being detected. Consequently, when selecting the relevant modules, the total number of genes associated with these modules remains significantly high. To streamline the focus onto pertinent genes, specific genes exhibiting exceptional topological characteristics within each module, such as hub nodes (Vandereyken et al., 2018), are typically chosen. However, it's worth noting that a prior study has demonstrated that co-expression network analyses featuring a greater number of modules and smaller module sizes tend to exhibit higher biological coherence (Abbassi-Daloii et al., 2020). This observation hints at an opportunity for refining current methodologies and potentially optimizing the detection of relevant genes.

Lastly, it is important to highlight that differential co-expression networks are not limited to biological contexts and can be applied to various non-biological domains where coordinated interac-

tions between entities under specific conditions are of interest. Furthermore, modules that overlap can provide a more realistic representation of the substructures not only in biological networks, but also in many other networks. Forcing a node into one community will fail to accommodate multiple relationships and functions that a node may have, resulting in an erroneous representation of the network structure Jin et al. (2015). Likewise, selecting relevant modules concerning external sample traits in non-biological domains aids in interpreting the functional relevance of those modules and their contribution to the system's behavior.

1.2 Approach

To address the knowledge gaps and harness opportunities for refinement, this research's central objective is the creation of an in-silico workflow for identifying stress-responsive genes. This workflow, named Control-Stress Data Integration with Overlapping Clustering (CSI-OC), seeks to more accurately capture the system's true characteristics.

The CSI-OC workflow takes a comprehensive approach, blending statistical and network-based methods to pinpoint stress-responsive genes. It utilizes the Log Fold Change (LFC) metric for merging control and stress expression data through one-to-one comparisons. This process explicitly captures information related to differential expression before constructing networks, resulting in a more focused dataset, eliminating the need for a separate network analysis. Furthermore, it explores an alternative to the traditional Pearson-based network construction method, namely the Lasso-based network construction. This approach is especially valuable when handling datasets with numerous genes and relatively few samples, as it yields network structures that accentuate the most crucial gene connections, rendering gene interactions more transparent and interpretable.

Additionally, CSI-OC places substantial emphasis on comprehending gene interactions within overlapping modules and their impact on phenotypic traits. Leveraging Lasso regression (Tibshirani, 1996) as a selection method, the workflow aims to identify modules that exert significant influence on phenotypic traits. Since these modules can overlap, the genes they encompass offer noteworthy insights into the complex interplay among genes in stress responses. Consequently, the need for a predetermined minimum module size can be discarded, regardless of the number of modules detected. Thus, the modules identified as relevant through Lasso are likely to be smaller in size and can be considered in their entirety, removing the necessity to restrict the selection of

relevant genes solely to hub nodes. This approach ensures that relevant genes are chosen based on their collective contributions to the dynamics of stress response, focusing on the genes whose interactions intricately shape the response to stress.

To ensure the systematic and comprehensive design and validation of the proposed workflow, a set of four specific objectives has been established. While these objectives may differ in wording and order from those originally presented in the candidacy proposal, they effectively encompass the original goals. The new objectives not only maintain the essence of the original ones, but also provide greater specificity and refinement.

i. To introduce and elucidate the CSI-OC workflow as a tool for the identification of stress-responsive genes.

ii. To apply the CSI-OC workflow to rice and sugarcane under specific stress conditions, namely salinity and drought.

iii. To assess the workflow's performance through the use of synthetic data, shedding light on its limitations and strengths.

iv. To explore the adaptability of the CSI-OC workflow beyond biological contexts through its application in a non-biological domain.

By achieving these objectives, this research aims to enhance the understanding of stress-responsive genes in plants, particularly in rice and sugarcane, refine the workflow's capabilities, and demonstrate its versatility.

1.3 Summary of Chapters and Contributions

This dissertation contributes to several ongoing research efforts in multiple areas, including genomics, plant biology, computational biology, network analysis, and data integration.

Chapter 3: In this chapter, the CSI-OC workflow is introduced and described in detail. The workflow integrates diverse data sources, captures differential expression information and enables the identification of key gene modules. This workflow serves as the centerpiece of the dissertation and offers insights into its development and functioning. It is implemented in Python, has been registered as software with the Dirección Nacional de Derechos de Autor (DNDA) (DNDA, 2023), and is available at (Riccio-Rengifo et al., 2021a).

Chapter 4: This chapter showcases a practical application of the CSI-OC workflow, utilizing the Pearson-based network construction method. This application is executed on publicly accessible rice data under salt stress. The findings from this case study underscore the workflow's effectiveness in identifying stress-responsive genes in a specific biological context. Furthermore, this main findings has been published in (Riccio-Rengifo et al., 2021b).

Chapter 5: This section delves into the application of the CSI-OC workflow using the Lasso-based network construction method. It conducts a comprehensive analysis on sugarcane data provided by Cenicaña, systematically assessing gene responses across diverse plant tissues and increasing levels of drought stress. The case study provides additional evidence of the workflow's utility in a different biological context. The findings are presented in (Riccio-Rengifo et al., 2023b).

Chapter 6: This chapter presents a comprehensive framework for testing and validating the CSI-OC workflow using synthetic data to shed light on its capabilities and limitations under controlled conditions. It highlights the the pivotal role of phenotypic data characteristics in influencing the workflow's performance. Moreover, the chapter introduces the method named CoSynthEx, designed to generate synthetic gene expression data that incorporates information about phenotypic traits and sample conditions. CoSynthEx has been registered as software with the DNDA and is available at (Riccio-Rengifo et al., 2023a).

Chapter 7: This chapter showcases the adaptability of the CSI-OC workflow by extending its application into a non-biological context, specifically within the setting of a supermarket chain. This explorative application illustrates how the workflow can be adapted beyond traditional biological applications, highlighting its broad range of potential uses.

Chapter 8: This final chapter summarizes the key findings, contributions, and implications of the research. It provides a comprehensive overview of the work and suggests directions for future research.

2. Preliminaries

This chapter establishes the fundamental background on which the following sections of this document are built. In Section 2.1, it starts with a comprehensive exploration of mathematical concepts, structures and algorithmic techniques, which collectively will form the core of a workflow designed for the identification of stress-responsive genes in plants. Section 2.2 follows, providing a description of the methodologies that will be employed to assess the significance and meaningfulness of the genes selected by the workflow. Finally, in Section 2.3, we delve into the theoretical foundations of the conditional Generative Adversarial Network (cGAN) model, which would serve as basis for the generation of synthetic data, allowing to systematically evaluate the workflow to explore its limitations.

2.1 Mathematical Foundations of the Workflow

2.1.1 Log Fold Change (LFC)

The Log Fold Change (LFC) is a measure commonly used in statistics and bioinformatics to quantify the relative difference in values between two conditions or groups. It is typically applied to gene expression data or any other dataset where you want to compare the magnitude of change between two sets of measurements.

The Log Fold Change for a particular value or observation is calculated as the logarithm (usually base 2) of the ratio between the values in the two conditions. The formula is as follows:

$$\text{LFC} = \log_2(\text{Condition}\mathrm{A}/\text{Condition}\mathrm{B}) \tag{2.1}$$

where Condition A (Condition B) is the measurement or value in the sample under condition A (condition B).

The Log Fold Change provides a concise representation of how much a particular measure-

ment has changed between two conditions. A positive LFC indicates an increase in the measurement in Condition A compared to Condition B, while a negative LFC indicates a decrease. A LFC of 0 means there is no difference between the two conditions. The logarithm to base 2 is most commonly used (Oeckinghaus et al., 2011) as it is easy to interpret, e.g., a doubling in the Condition A scaling is equal to a LFC of 1, a quadrupling is equal to a LFC of 2 and so on.

Researchers often use Log Fold Change in differential expression analysis to identify genes, proteins, or other biomolecules that show significant changes in expression or abundance between experimental conditions. It simplifies interpretation by expressing changes on a logarithmic scale, which is particularly useful when dealing with data that spans a wide range of values.

2.1.2 Differential co-expression network (DCN)

A differential co-expression network (DCN) is an undirected graph $G = (V, E)$, where the set of N nodes $V = \{v_1, v_2, \ldots, v_N\}$ represents genes and a link $(v_i, v_j) \in E$ indicates a common alteration in the expression pattern of genes v_i and v_j when changing between two particular conditions (e.g., control and stress).

Differential co-expression networks represent a specialized aspect of co-expression analysis that focuses on changes in gene relationships across different conditions or states. While traditional co-expression networks reveal consistent associations between genes, differential co-expression networks detect alterations in these associations when transitioning between experimental conditions, tissues, or disease states. This approach allows researchers to identify condition-specific or context-specific gene interactions, providing insights into regulatory mechanisms that are active only under specific circumstances. Differential co-expression networks are instrumental in understanding how gene relationships adapt to different biological contexts, shedding light on the dynamic nature of gene regulation and its relevance to various biological phenomena.

2.1.3 Hierarchical Link Clustering (HLC)

The Hierarchical Link Clustering (HLC) algorithm partitions groups of links (rather than nodes), where each node inherits all memberships of its links and can belong to multiple, overlapping modules Ahn et al. (2010).

More specifically, HLC evaluates the similarity between links if they share a particular node. Consider a pair of incident links e_{ik} and e_{jk} to node k. The similarity between e_{ik} and e_{jk} is defined

by the Jaccard index as

$$S(e_{ik}, e_{jk}) = \frac{| \eta(i) \cap \eta(j) |}{| \eta(i) \cup \eta(j) |}, \tag{2.2}$$

where $\eta(v)$ denotes the set containing the node v and its neighbors, for any $v \in V$. The algorithm uses single-linkage hierarchical clustering to build a dendrogram where each leaf is a link from the network and branches represent linked communities.

The threshold to cut the dendrogram is defined based on the average density of links inside communities (i.e., partition density). For $G = (V, E)$ and a partition of the links into c subsets, the partition density is computed as

$$D = \frac{2}{|E|} \sum_{c} |E_c| \frac{|E_c| - |V_c| + 1}{(|V_c| - 1)(|V_c| - 2)}. \tag{2.3}$$

Note that, in most cases, the partition density D has a single global maximum along the dendrogram. If the dendrogram is cut at the top, then D represents the average link density of a single giant module. If the dendrogram is cut at the bottom, then most modules consist of a single link. By computing D at each level of the dendrogram, the level that maximizes partition density can be found (nonetheless, meaningful structure could exist above or below the threshold).

2.1.4 Least Absolute Shrinkage Selector Operator (Lasso)

Lasso Tibshirani (1996) is a regularized linear regression technique. By combining aregression model with a procedure of contraction of some parameters towards 0, Lasso imposes a restriction (or a penalty) on regression coefficients. In other words, Lasso solves the least squares problem with restriction on the ℓ_1-norm of the coefficient vector. In particular, the approach is especially useful in scenarios where the number of variables c is much greater than the number of samples m (i.e., $c \gg m$).

Consider a dataset of m samples, consisting each of c covariates and a single outcome. Let y_i be the outcome and $x_i := (x_{i1}, ..., x_{ic})$ be the covariate vector for the i-th sample. The objective of Lasso is to solve

$$\min \left\{ \sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{c} \alpha_j x_{ij} \right)^2 \right\} \quad , \quad \text{subject to} \quad \sum_{j=1}^{c} |\alpha_j| \leq s, \tag{2.4}$$

where s is the regularization penalty. Equivalently, in the Lagrangian form, Lasso minimizes

$$\sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{c} \alpha_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{c} |\alpha_j| \,, \tag{2.5}$$

where $\lambda \geq 0$ is the corresponding Lagrange multiplier. Since the value of the regularization parameter λ determines the degree of penalty and the accuracy of the model, cross-validation is used to select the regularization parameter that minimizes the mean-squared error. Lasso is preferred in the proposed workflow because it tends to outperform other methods such as ordinary least squares regression and Ridge Muthukrishnan and Rohini (2016).

2.2 Workflow Evaluation Approaches

2.2.1 Computational validation

Computational validation of a group of elements selected by a computational workflow is imperative, particularly when accessing additional supporting information is challenging. In the absence of a gold standard for comparison, randomly selected elements serve as a baseline.

Specifically, our focus lies in evaluating the significance of a group of genes selected for their relevance to stress responses based on their interactions and influence on plant phenotypic traits. To account for the genes' varying behavior between control and stress conditions, we subjected them to a classification task. The expression profiles of these selected genes serve as inputs for a Random Forest classification model (Biau and Scornet, 2016), tasked with classifying samples as either control or stress based solely on the gene expression profiles. This classification is performed a total of one hundred times, and each time, the performance of the selected genes in terms of accuracy is computed. The same classification procedure is repeated using a random set of network genes, having the same number of genes as in the selected set. This approach allows us to assess the selected genes' ability to discriminate between control and stress samples.

Furthermore, considering that the selected genes are expected to influence plant phenotypic traits, we evaluate them through a regression task. The process mirrors the classification evaluation, with the expression profiles of the selected genes utilized as input features for a Random Forest Regressor (Biau and Scornet, 2016). In this case, the response variable is a phenotypic trait. The model's performance is assessed in terms of the mean squared error (MSE) and compared against the MSE of a model created with a random set of genes. This approach enables us to assess the selected genes' capacity to predict phenotypic traits.

In both scenarios, we employ a Wilcoxon signed-rank test (Woolson, 2007) to determine whether the performance of the selected genes surpasses that of randomly chosen genes. The presence of a significant difference between the medians of the selected and random genes serves

as an indicator of the superior performance of the selected genes in comparison to the random selection. For the classification task, this significant difference should be attributed to a higher median accuracy among the selected genes, as higher accuracy corresponds to an enhanced performance in classification models. In the regression task, the significant difference should be attributed to a lower median MSE among the selected genes, given that a lower MSE aligns with improved performance in regression models. Observing these patterns within the selected gene set signifies the meaningfulness of the chosen genes and reaffirms the satisfactory functioning of the gene selection tool.

2.2.2 Knowledge-based evaluation

Functional enrichment is a critical step in understanding the underlying processes contributing to phenotype or stress responses. The evaluation of a specific gene set can be achieved through Gene Ontology (GO) enrichment analysis (Ashburner et al., 2000; Consortium, 2021). In this context, a special emphasis is placed on the GO category of biological processes (BP), as these annotations are particularly straightforward to interpret within the scope of this study. For instance, terms related to stress responses or defense mechanisms fall within this category.

Enrichment analysis leverages the power of statistical methodologies, primarily Fisher's Exact Test (Sprent, 2011), in combination with a robust False Discovery Rate (FDR) correction for multiple testing (Korthauer et al., 2019; Benjamini and Hochberg, 1995) and the methodology is implemented using custom Python scripts. This approach effectively identifies GO terms that are either over-represented or under-represented in a given gene set compared to the background set's annotations.

Furthermore, the evaluation of gene sets extends beyond GO enrichment to encompass the significant representation of genes with specific characteristics, such as differential expression or those acting as transcription factors. The presence of these gene types within the set of genes under investigation is determined using a Fisher test. For instance, transcription factor genes can be extracted from specialized databases like PlantTFDB (Jin et al., 2016). Differentially Expressed Genes (DEGs) are identified based on DESeq2 results Love et al. (2014), using a predefined threshold, often set at 1 or higher for a more stringent analysis. Within the DESeq2 results, DEGs are further categorized into up-regulated genes (LFC $> t$, FDR < 0.05) and down-regulated genes (LFC $< -t$, FDR < 0.05).

This comprehensive knowledge-based evaluation approach allows for a deeper understanding

of the functional and regulatory aspects of different gene sets.

2.3 Mathematical Foundations of the Synthetic Data Generation

2.3.1 Generative Adversarial Network (GAN)

Conditional Generative Adversarial Networks (cGANs) are a class of deep generative models that expand upon the conventional Generative Adversarial Network (GAN) framework (Creswell et al., 2018; Aggarwal et al., 2021; Gui et al., 2021). The GAN consists of two neural networks: a generator and a discriminator. Broadly speaking, the generator is responsible for generating new data, while the discriminator is responsible for distinguishing between real and generated data. The two networks are trained together in an adversarial manner. The generator is trying to fool the discriminator into thinking that the generated data is real, while the discriminator is trying to distinguish between real and generated data.

Formally, a GAN employs differentiable functions $\mathcal{D}$ and $\mathcal{G}$ to represent the discriminator and generator, utilizing real data x and random variables z as their respective inputs. The generated sample, denoted as $\mathcal{G}(z)$, adheres to the distribution p_{data} of real data. When the discriminator $\mathcal{D}$ receives input from real data x, its objective is to correctly classify it as true and assign a label of 1. Conversely, if the input originates from $\mathcal{G}(z)$, $\mathcal{D}$ should classify it as false and label it as 0. The role of $\mathcal{D}$ is to accurately determine the data source, while the primary goal of $\mathcal{G}$ is to ensure that the performance of generated data $\mathcal{G}(z)$ when assessed by $\mathcal{D}$ (i.e., $\mathcal{D}(\mathcal{G}(z))$) aligns with the performance of real data x when evaluated by $\mathcal{D}$ (i.e., $\mathcal{D}(x)$). Through adversarial optimization, both $\mathcal{D}$ and $\mathcal{G}$ gradually enhance their performance. Ultimately, when $\mathcal{D}$ achieves a high level of discriminatory ability but can no longer distinguish the data source accurately, it signifies that the generator $\mathcal{G}$ has successfully captured the distribution of real data.

The training cost of a GAN is assessed using a cross-entropy function that relies on both the generator and the discriminator. The optimization of GAN can be formulated as a minimax problem:

$$\min_{\mathcal{G}} \max_{\mathcal{D}} \mathbb{E}_{p_{\text{data}}(x)}[\log \mathcal{D}(x)] + \mathbb{E}_{p_z(z)}[\log (1 - \mathcal{D}(\mathcal{G}(z)))] \tag{2.6}$$

where $\mathbb{E}$ represents the expectation, p_z is the prior distribution of the input noise variable z, and p_{data} is the distribution of the real data x.

2.3.2 Conditional Generative Adversarial Network (cGAN)

GANs can be extended to a conditional model when both the discriminator and generator are conditioned on additional information denoted as y. This extension is called conditional GAN (cGAN) and its corresponding objective function is as follows:

$$\min_{\mathcal{G}} \max_{\mathcal{D}} \mathbb{E}_{p_{\text{data}}(x)}[\log \mathcal{D}(x|y)] + \mathbb{E}_{p_z(z)}[\log (1 - \mathcal{D}(\mathcal{G}(z|y)))] \tag{2.7}$$

The cGAN framework allows for the generation of data samples that are not only indistinguishable from real data but also coherent with the provided conditions. The cGAN training process involves a game between the generator, striving to generate data indistinguishable from real data under the specified conditions, and the discriminator, which aims to accurately classify real and synthetic samples while considering the conditional information.

In the context of gene expression data analysis, cGANs can be used to generate synthetic gene expression profiles that are consistent with certain conditions or biological contexts. The generator network in the cGAN is designed to take both random noise and the conditional information as input. By training the cGAN on real expression data, it learns to generate synthetic expression profiles that match the specified conditions. This allows researchers to control and manipulate gene expression patterns based on the provided conditions.

3. The Control-Stress Data Integration with Overlapping Clustering (CSI-OC) Workflow

The workflow for identifying candidate stress-responsive genes, using gene expression and phenotypic traits measured under control and stress conditions, is described here. It is called Control-Stress data Integration with Overlapping Clustering (CSI-OC). The approach relies on identifying overlapping gene modules, within a differential co-expression network, that are highly related to phenotypic traits. The general structure of the CSI-OC workflow is schematically illustrated in Fig 3.1.

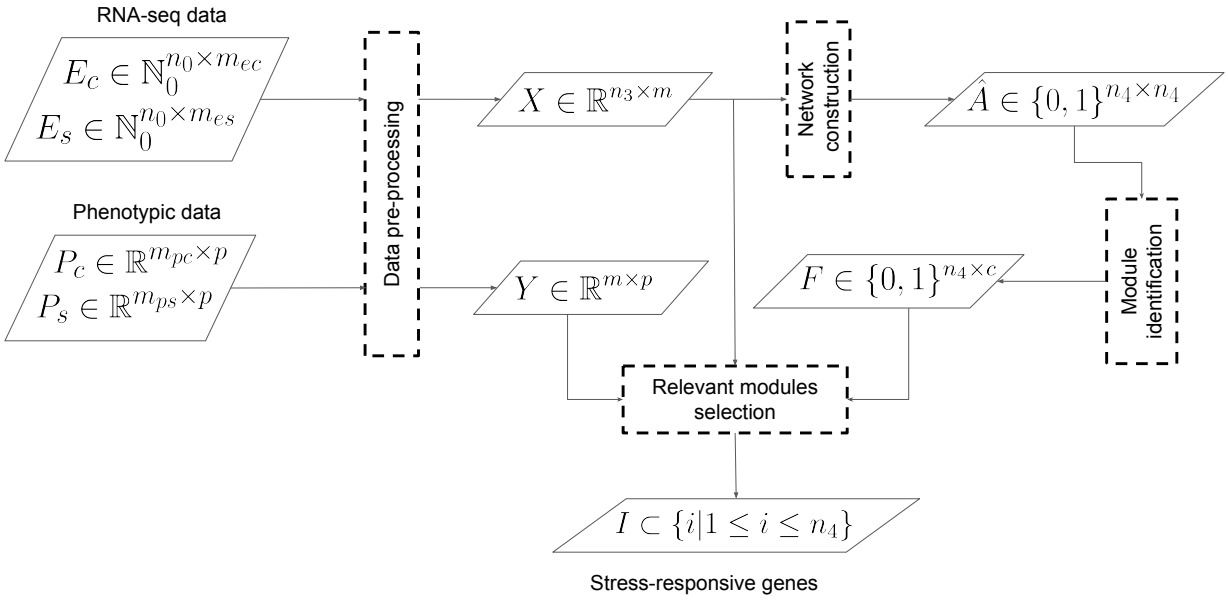

Figure 3.1: CSI-OC workflow. It takes as input expression and phenotypic data, and outputs a set of stress-responsive genes.

The workflow input correspond to gene expression (denoted as E_c for control and E_s for stress)

and phenotypic traits (denoted as P_c for control and P_s for stress) measured under control and stress conditions. It can be broken down into four steps. The first step computes the changes, from control to stress condition, with the Log Fold Change (LFC) metric for the expression data (denoted as X) and phenotypic traits (denoted as Y). The second step, corresponds to the differential co-expression network (DCN) construction (denoted as A) from the LFC data. The third step identifies the overlapping gene modules (denoted as F) within the DCN. The fourth and final step pins down the modules most highly related to the LFC in phenotypic traits using Lasso regression. The union of the genes that belong to the selected modules (denoted as I) are those that are called candidate stress-responsive genes. These four steps will be explained in detail in this chapter.

3.1 Data Pre-processing

The goal of the data pre-processing step is to build matrices X and Y representing, respectively, the changes in expression levels and phenotypic values between control and stress condition. Fig 3.2 shows the flow of data through the pre-processing pipeline.

A normalization process is applied to interpret RNA-seq data and handle possible biases affecting the quantification of results. Here, DESeq2 (Love et al., 2014) is used to correct the library size and RNA composition bias. To do so, the control $E_c \in \mathbb{N}_0^{n_0 \times m_{ec}}$ and stress $E_s \in \mathbb{N}_0^{n_0 \times m_{es}}$ data are joined in the matrix $D_0 \in \mathbb{N}_0^{n_0 \times (m_{ec}+m_{es})}$ before normalization. The normalized data is represented as a matrix $D_1 \in \mathbb{R}^{n_0 \times (m_{ec}+m_{es})}$. In each expression matrix (E_c and E_s) the rows represent the genes and the columns the samples. Note that the number of control samples m_{ec} may be different from the number of stress samples m_{es}, due to the replicates that the m genotypes may have. The biological replicates of each genotype and the gene replicates are averaged and represented as a matrix $D_2 \in \mathbb{R}^{n_1 \times 2m}$. Note that the previous step keeps the same number of unique genes, but reduces gene dimensionality to $n_0 \leq n_1$ by removing duplicates. The genes exhibiting low variance or low expression are removed from D_2. Consequently, this stage reduces the set of genes from a pool of size n_1 to a restricted pool of size $n_2 \leq n_1$. The control and stress data are again separated into the matrices $E_c \in \mathbb{R}^{n_2 \times m}$ and $E_s \in \mathbb{R}^{n_2 \times m}$, respectively. The matrix entries $E_c(i, j)$ and $E_s(i, j)$ represent the normalized expression level of gene i in accession j under control and stress condition, respectively. The initial phenotypic data, under control ($P_c \in \mathbb{R}^{m_{pc} \times p}$) and stress ($P_s \in \mathbb{R}^{m_{ps} \times p}$), also go through the process of averaging replicates, obtaining the P_c and P_s matrices of dimensions $m \times p$.

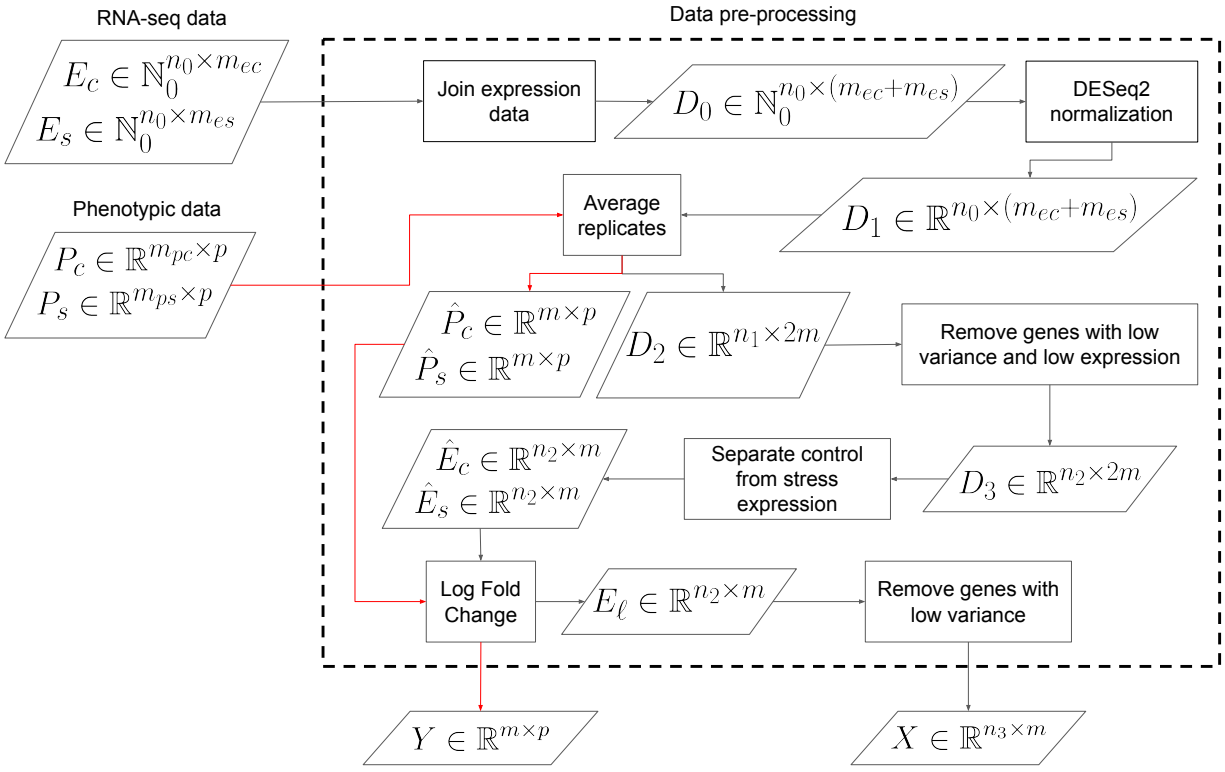

Figure 3.2: CSI-OC data pre-processing. The red lines indicate the path of the phenotypic data and the black lines indicate the path of the expression data.

In the above configuration, the changes in expression levels and phenotypic values between control and stress conditions are measured in terms of logarithmic ratios with the LFC metric. In the case of expression levels, the log ratios are represented in the Log Fold Change matrix $E_\ell \in \mathbb{R}^{n_2 \times m}$, where $E_\ell(i,j) = \log_2(E_s(i,j)/E_c(i,j))$. Similarly, the log ratios of the phenotypic data are computed and represented in the $Y \in \mathbb{R}^{m \times p}$ matrix.

The final stage of pre-processing is to filter E_ℓ by removing rows (e.g., genes) with low variance in the differential expression patterns, thus obtaining a new matrix X of dimensions $n_3 \times m$, with $n_3 \leq n_2$.

3.2 Differential Co-expression Network (DCN) Construction

Differential co-expression networks are of biological interest because adjacent nodes in the network represent genes that respond similarly to a so-called stress condition. A differential co-expression network (DCN) is an undirected graph $G = (V, E)$, where the set of N nodes $V =$

$\{v_1, v_2, \ldots, v_N\}$ represents genes and a link $(v_i, v_j) \in E$ indicates a common alteration in the expression pattern of genes v_i and v_j when changing between two particular conditions (e.g., control and stress).

Two approaches to construct differential co-expression networks are described here. The first approach is based on Pearson correlation values, while the second approach is based on LASSO regression coefficients. The choice of which approach to use depends on the specific dataset and the goals of the study. The network construction process described below applies both to differential co-expression networks and to co-expression networks, where the only difference is the input data. Differential co-expression networks are built using differential expression profiles (e.g., LFC) instead of the expression profiles themselves as in co-expression networks.

3.2.1 Pearson-based Co-expression Networks

The Log Fold Change matrix X is used as input to build a Pearson-based DCN $\hat{A}$ as shown in Figure 3.3.

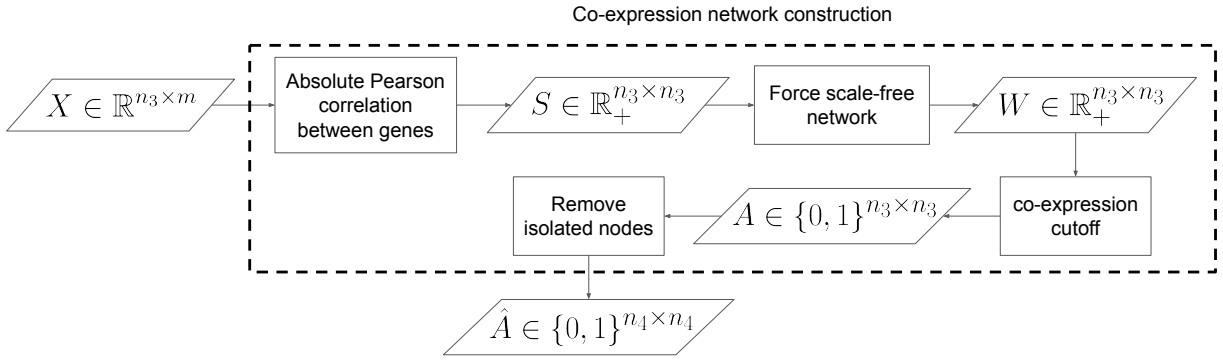

Figure 3.3: Pearson-based differential co-expression network construction.

First, the level of concordance between gene differential expression profiles across samples is measured. To this end, the absolute value of the Pearson Correlation Coefficient (PCC) is used as the similarity measure between genes, meaning that pairs of nodes with strong negative correlation are considered connected with the same strength as nodes with strong positive correlation (Song et al., 2012). The resulting values are stored in the similarity matrix $S \in \mathbb{R}_+^{n_3 \times n_3}$.

Second, the matrix S is transformed into an adjacency matrix $W \in \mathbb{R}_+^{n_3 \times n_3}$ where each entry $W(i,j) = S(i,j)^\beta$ encodes the connection strength between each pair of genes. In other words, the elements of the adjacency matrix are the similarity values up to the power $\beta > 1$ so that the

degree distribution will fit a scale-free network. In a strict scale-free network, the logarithm of $P(k)$ (i.e., the probability of a node having degree k) is approximately inversely proportional to the logarithm of k (i.e., the degree of a node). The parameter β is chosen to be the smallest value for which the R^2 of the linear regression between $log_{10}(p(k))$ and $log_{10}(k)$ is closest to 1 (e.g., $R^2 > 0.8$).

Third, the adjacency matrix W is transformed into an unweighted network $A \in \{0,1\}^{n_3} \times n_3$. To this end, the PCC cutoff is determined using the approach described in (Aoki et al., 2007). The number of nodes, edges, and the network density is determined for different PCC cutoffs. In a neighborhood of the optimal PCC cutoff, the number of nodes presents a linear decrease and the density of the network reaches its minimum, while below this value the number of edges rapidly increases. Following this observation, a cutoff is selected such that gene pairs having a correlation score higher than the threshold are considered to have a significant level of co-expression. The entries of W become 1 above the cutoff and 0 otherwise.

Finally, isolated nodes are removed from the network A. These nodes do not have any edges to other nodes, and they can introduce noise into the network. Therefore, the resulting matrix $\hat{A}$ is of dimensions $n_4 \times n_4$ where $n_4 \leq n_3$.

3.2.2 Lasso-based Co-expression Networks

This subsection explains how building differential co-expression networks using Lasso regression coefficients. Figure 3.4 shows a step-by-step diagram of this approach.

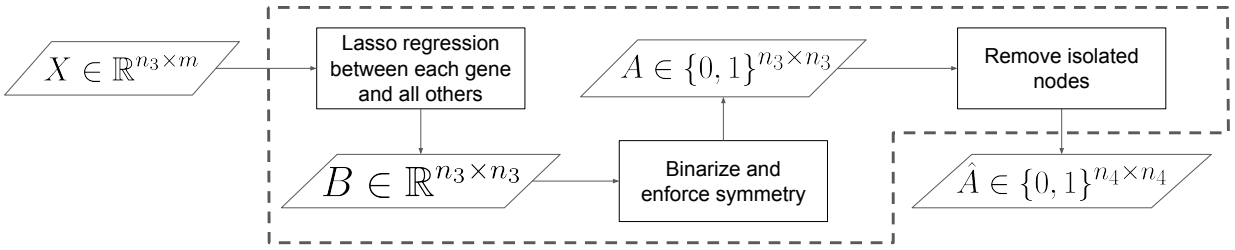

Figure 3.4: Lasso-based differential co-expression network construction.

Note that building a network $G = (V, E)$, that is, a representation of pairwise relationships over a set of vertices, is equivalent to inferring a neighborhood for each vertex (i.e., the set of vertices to which it is connected). Given a LFC matrix $X \in \mathbb{R}^{n_3 \times m}$, the set of neighbors of vertex $v_i \in V$, denoted $V(v_i) := \{v_j : (v_i, v_j) \in E\}$, is inferred by regressing x_i against all other variables $x_{\setminus i} := [x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{n_3}]^T \in \mathbb{R}^{n_3 - 1}$. The result is a matrix $B \in \mathbb{R}^{n_3 \times n_3}$ whose diagonal is

zero and the remaining $n_3 - 1$ entries of a row i correspond to the coefficients of the regression of x_i against $x_{\setminus i}$. Each entry $B(i,j)$ represents the strength of the relationship between vertices v_i and v_j, where zero strength indicates no connection. For each variable x_i the regression problem has the form:

$$\underset{\beta_i}{\text{minimize}} \left\| X_i - X_{\setminus i}\beta_i \right\|_2^2 + \lambda \left\| \beta_i \right\|_1 , \tag{3.1}$$

where X_i and $X_{\setminus i}$ represent the observations on x_i (i.e., the transpose of the i-th row of X) and the rest of the variables, respectively. The vector $\beta_i \in \mathbb{R}^{n_3-1}$ is a vector of coefficients for x_i. In Eq. 3.1, the first term can be interpreted as a local log-likelihood of β_i and the second term corresponding to a ℓ_1 penalty (absolute value of the magnitude of coefficients) is added to enforce sparsity. The regularization parameter λ balancing the two terms. Lasso is repeated for all the variables leading to a set of $n_3 \times n_3$ coefficients that are computed from $\beta_1, \ldots, \beta_{n_3}$. Note that there is no guarantee that $B(i,j) \neq 0$ implies $B(j,i) \neq 0$. Therefore, the information in $V(v_i)$ and $V(v_j)$ is combined to enforce symmetry: an edge (v_i, v_j) is meaningful, if $B(i,j)$ and $B(j,i)$ are both non-zero. In this way, the symmetric matrix $A \in \{0,1\}^{n_3 \times n_3}$ is created. Finally, as in the Pearson-based approach, the number of nodes is reduced from n_3 to n_4 by removing the isolated nodes from A if there is any, resulting in the matrix $\hat{A}$.

Note also that including the ℓ_1 penalty allows Lasso to identify the variables that are strongly associated with the response variable (i.e., variable selection). Since the value of the regularization parameter λ determines the degree of penalty and the accuracy of the model, cross-validation is used to select a regularization parameter that minimizes the mean-squared error. If the degree of penalty λ is equal to zero, the solution is the same as least-squares (LS) linear regression (Björck, 1990). For larger values of λ, larger coefficients are shrunk towards zero. Lasso's advantage over LS is that the later does not yield non-zero estimates, which would result in a fully connected network and the problem of setting a threshold for edge significance. Lasso avoids this additional step as it simultaneously performs parameter estimation and variable selection by forcing the least significant coefficients to zero through the ℓ_1 penalty.

3.3 Overlapping Module Detection

The next step in the workflow is to identify overlapping modules from the differential co-expression network represented by $\hat{A}$. The idea is to cluster genes with similar patterns of differential expres-

sion change. Membership in these modules may overlap in biological contexts, because modules may be related to specific molecular, cellular, or tissue functions, and the biological components (i.e., genes) may be involved in multiple functions. This motivates the use of the Hierarchical Link Clustering (HLC) algorithm (Ahn et al., 2010) within the CSI-OC workflow to detect overlapping rather than disjoint gene communities. Broadly speaking, the HLC algorithm partitions groups of links (rather than nodes) within the co-expression network, allowing each node (gene) to inherit memberships from its multiple links, consequently belonging to multiple, overlapping modules.

Inside the CSI-OC workflow, the HLC algorithm organizes the n_4 genes of matrix $\hat{A}$ into c modules, where each gene can belong to zero or multiple modules. This information is represented as an affiliation matrix, $F \in \{0,1\}^{n_4 \times c}$, where $F(i,u) = 1$ if node i is a member of module u and $F(i,u) = 0$ otherwise). This matrix serves as a fundamental resource for downstream analyses, offering valuable insights into the intricate relationships between genes and their multifaceted functional roles within the biological context under study.

3.4 Relevant Module Selection with Respect to Phenotypic Traits

To determine which sets of genes most influence the phenotypic response, a selection process is carried out using Lasso regression. Fig. 3.5 summarizes the data processing at this stage of the CSI-OC workflow.

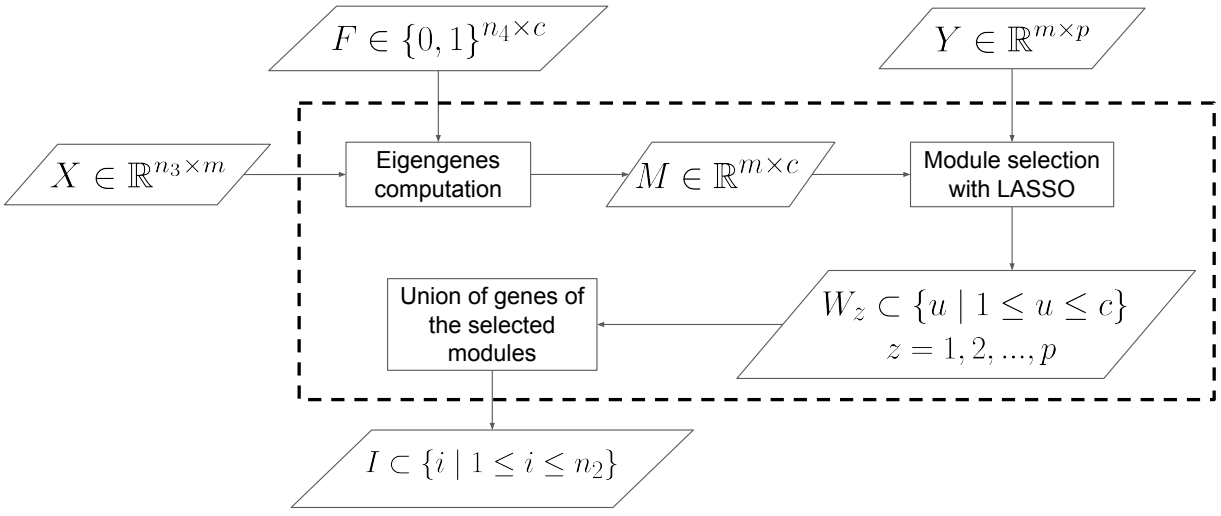

Figure 3.5: Relevant modules selection within the CSI-OC workflow.

Each network module is represented by an eigengene, which is defined as the first principal

component of the belonging genes. An eigengene is an average differential expression profile for each module. It is computed from the Log Fold Change Matrix X and the affiliation matrix F. Given a module u, the affiliation matrix F is used to identify the genes belonging to u. The corresponding rows of the matrix X are then selected to compute the first principal component of u. Each principal component becomes a column of the matrix $M \in \mathbb{R}^{m \times c}$. These profiles are associated with each phenotypic trait using LASSO as a feature selection mechanism (Fonti and Belitser, 2017). To identify the most relevant modules associated with the phenotypic response to the specific stress, the eigengenes (i.e., the columns of M) act as regressor variables and each phenotypic trait (i.e., each column of Y) is used as an outcome variable. LASSO is applied p times, once for each phenotypic trait. Therefore, for each trait Y_z, with $z \in 1, 2, ..., p$, the regression has the form:

$$\underset{\alpha_z}{\text{minimize}} \left\| Y_z - M\alpha_z \right\|_2^2 + \lambda \left\| \alpha_z \right\|_1, \tag{3.2}$$

where the weight $\alpha_z(i)$ represents the importance of the i-th module in the phenotypic response of the z-th trait. The regularization parameter λ, tuned with cross-validation, determines the number of modules to be selected. The weights α evolve with each LASSO iteration, by trying to optimize Equation 3.2, until the designed number of modules with non-zero weight is found. Intuitively, the use of Lasso in the workflow achieves the goal of neglecting (i.e., reducing to zero) the weights associated to modules with non-essential effects in the phenotypic response and, at the same time, enhancing the weights associated to modules with significant effects.

The output after the application of LASSO is a set W_z of modules for each phenotypic trait z, where $W_z \subseteq \{u | 1 \leq u \leq c\}$ for $z = 1, 2, .., p$. A candidate stress-responsive gene in I for downstream analysis is any gene belonging to a selected module; that is, $I = \bigcup_{z=1}^{p} W_z$, where $I \subseteq \{i | 1 \leq i \leq n_2\}$.

3.5 Discussion

In this chapter, a comprehensive workflow has been elucidated for the identification of candidate stress-responsive genes. The workflow involves four key steps: data pre-processing, construction of differential co-expression networks, detection of overlapping gene modules, and the selection of relevant modules with respect to phenotypic traits. Each step plays a critical role in uncovering the genes that are most closely associated with specific stress conditions. The utilization of tools such

as Log Fold Change matrices, Pearson-based and Lasso-based network construction, alongside the application of the Hierarchical Link Clustering (HLC) algorithm, collectively constitute a robust analytical framework for the elucidation of stress-responsive genes within biological systems.

Regarding the Pearson-based network construction method, using a linear correlation coefficient (PCC) to express relationships between genes across samples is a common and reasonable practice in many biological analyses, especially in gene expression studies. However, its appropriateness depends on several factors:

- Linearity Assumption: PCC assumes a linear relationship between variables. If the relationship between genes is non-linear, PCC might not accurately capture the association.
- Outlier Sensitivity: PCC can be sensitive to outliers, potentially skewing the correlation value and affecting the interpretation of gene relationships.
- Data Distribution: PCC assumes a bivariate normal distribution. If the data distribution significantly deviates from normality, it might affect the reliability of the correlation results.
- Biological Interpretation: While PCC quantifies the strength and direction of the linear relationship between genes, it might not capture more complex biological interactions or regulatory mechanisms.

Thus, while PCC is a widely used and valuable measure to explore relationships between genes, it's important to consider its limitations.

The results presented in this chapter will serve as the foundation for the subsequent chapters of this research. The workflow developed will be applied to two distinct biological case studies: one involving rice under salt stress and the other focusing on sugarcane under drought stress.Through the application of this workflow, the aim is to identify and validate stress-responsive genes specific to each case. Additionally, experiments with synthetic data will be conducted to explore the boundaries and limitations of the workflow, facilitating its refinement for application in different contexts. Beyond its biological applications, the workflow's principles and methodologies will be generalized for use in other domains. Specifically, this analytical framework will be applied in a marketing case study, demonstrating its adaptability in revealing patterns and correlations in diverse datasets. Through adaptation and expansion of the workflow to diverse contexts, the objective is to demonstrate its broader applicability in addressing complex research inquiries and real-world challenges across various fields of study.

4. Identifying Salt-Responsive Genes in Rice (*Oryza sativa*)

Stresses are key factors that influence plant development, often associated to extensive losses in agricultural production (Mesterházy et al., 2020; Shrivastava and Kumar, 2015). Soil salinity is one of the most devastating abiotic stresses. According to Shrivastava and Kumar (2015), soil salinity contributes to a significant reduction in areas of cultivable land and crop quality. The study estimates that 20% of the total cultivated land worldwide and 33% of the total irrigated agricultural land is affected by high salinity. By the end of $2\,050$, areas of high salinity are expected to reach 50% of the cultivated land (Shrivastava and Kumar, 2015). In dry and semiarid regions, soil salinity emerges as the primary environmental challenge that significantly constrains plant productivity (Hussain et al., 2009). In sensitive plant species, salt stress inhibits growth and development by reducing leaf area, photosynthesis, respiration rate, protein synthesis, nitrogen fixation, yield, and biomass (Ghoname et al., 2023; Nguyen et al., 2023; Önay and Demirbas, 2023).

Rice (*Oryza sativa*), is a staple food for over half of the global population. Especially in Asia, it is considered a glycophytic plant and is among the most vulnerable cereal crops to salinity stress (Zeng and Shannon, 2000; Chang et al., 2019; Taratima et al., 2022). In the context of Colombian agriculture, rice holds a prominent position as the third-largest crop in terms of national production value, following coffee and sugarcane. It engages approximately two million people throughout the value chain and plays a crucial role in the country's food security and rural consumption, with a total production of 2.5 million tonnes a year (Lacambra et al., 2020).

Salinity tolerance and susceptibility in rice are the result of elaborated interactions between morphological, physiological, and biochemical processes, which are regulated by multiple genes in various parts of the plant genome (Reddy et al., 2017). One of the primary responses to salt stress involves limiting the uptake of toxic ions, such as sodium (Na+), into the plant. To counteract the osmotic stress caused by high salt levels in the soil, rice accumulates compatible solutes in its

cells to maintain cellular turgor pressure and prevent water loss, enabling the plant to withstand water scarcity associated with salinity. Rice also adapts its root architecture in response to salt stress, developing longer roots or increasing the density of lateral roots to explore a larger soil volume in search of less saline areas (Acosta-Motos et al., 2017; Zhao et al., 2020; Hannan et al., 2020; Vázquez-Glaría et al., 2021). Understanding these multifaceted mechanisms of salt stress response in rice and identifying the genes involved, is crucial for developing crop improvement strategies that enhance its resilience to salinity, thereby ensuring stable rice production in regions affected by high soil salinity levels.

In this chapter,we delve into the intricate realm of salt stress responses in rice (*Oryza sativa*) by leveraging the CSI-OC workflow. To ensure a coherent and systematic exploration, the chapter is organized into distinct sections. Section 4.1 elucidates the data's origin, detailing its sources and the variables employed in this study. Section 4.2 explores the dynamic evolution of gene numbers throughout the comprehensive selection process within the CSI-OC workflow. Section 4.3 delves into the characteristics of the overlapping modules identified. Section 4.4 presents the results of computational validation, while Section 4.5 showcases the knowledge-based evaluation. Finally, Section 4.6 engages in an insightful discussion, synthesizing the implications of the findings and their potential impact on rice crop improvement. This structured approach aims to provide a comprehensive and in-depth exploration of salt stress responses in rice, bolstered by the innovative CSI-OC workflow.

4.1 Origin of Data

The RNA-seq data was accessed from the GEO database (Clough and Barrett, 2016), accession number GSE98455. It corresponds to $n_0 = 57\,845$ gene expression profiles of shoot tissues measured for control and salt conditions in $m = 92$ accessions of the Rice Diversity Panel 1 (Eizenga et al., 2014), with $r = 2$ biological replicates. A total of $p = 3$ phenotypic traits were used: shoot potassium content (K_shoot), shoot biomass (BM_shoot), and root biomass (BM_root). These traits were measured for the same 92 genotypes, under control and stress conditions, and can be found in the supplementary information for Campbell et al. (2017).

4.2 Summary of Selected Genes

For this case study in rice, the CSI-OC workflow with the Pearson-based network construction method was applied. As result, the workflow identified 19 rice genes that seem relevant in the response to salt stress. They are distributed across six modules: three modules, each grouping together three genes, are associated to shoot K content; two modules of three genes are associated to shoot biomass; and one module of four genes is associated to root biomass. These genes represent target genes for the improvement of salinity tolerance in rice. Figure 4.1 depicts in a Venn diagram how the number of genes selected at different stages evolve. It summarizes how, from the initial $n_0 = 57\,845$ genes (with no gene replicates), the workflow of chapter 3 identified a reduced set of 19 genes. First, $48\,431$ genes are discarded after filtering the normalized expression data D_2, then 486 additional genes are discarded when filtering the Log Fold Change matrix X, next $3\,118$ genes identified as isolated nodes in A are also discarded, to finally arrive at 19 selected genes, of which 16 are differentially expressed ($|\text{LFC}| > 2$).

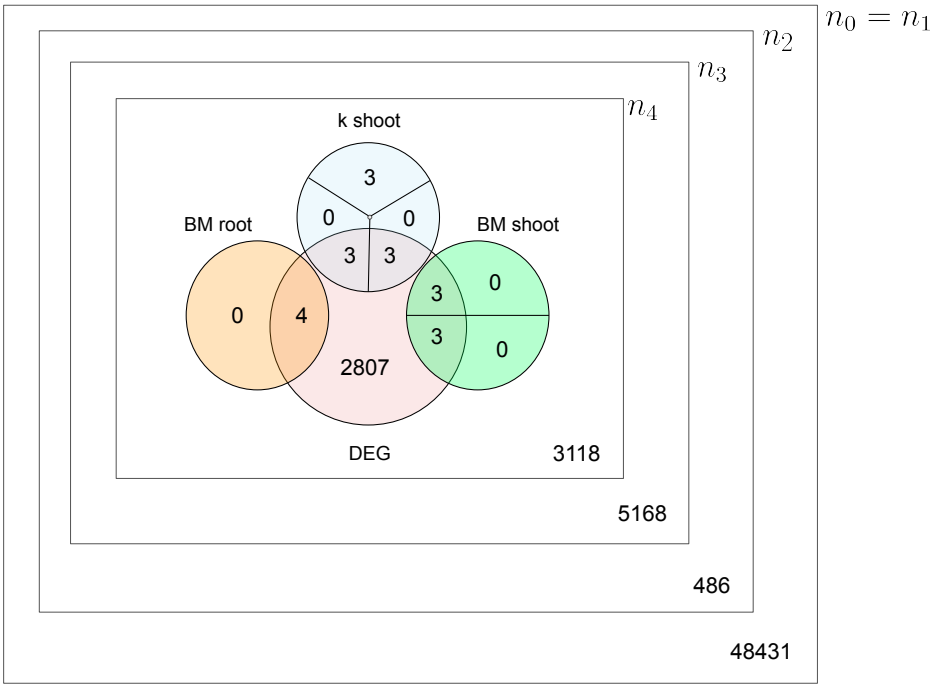

Figure 4.1: Venn diagram illustrating the progressive gene selection process within the CSI-OC workflow for the rice case study, from an initial set of $57\,845$ genes to the final selection of 19 salt-responsive genes.

4.3 Overlapping Genes Characterization

The resulting adjacency matrix $\hat{A}$ had $5\,810$ connected genes and accounted for $614\,501$ edges. After applying the HLC algorithm, a total of $4\,131$ genes were distributed in $c = 5\,143$ overlapping modules of at least 3 genes. Figure 4.2 presents a histogram of the overlapping percentage of these genes, measured as the proportion of modules to which each gene belongs. The first bar of the histogram represents the genes with zero overlap, corresponding to 28% of the total genes; the remaining 72% represents the genes belonging to more than one module.

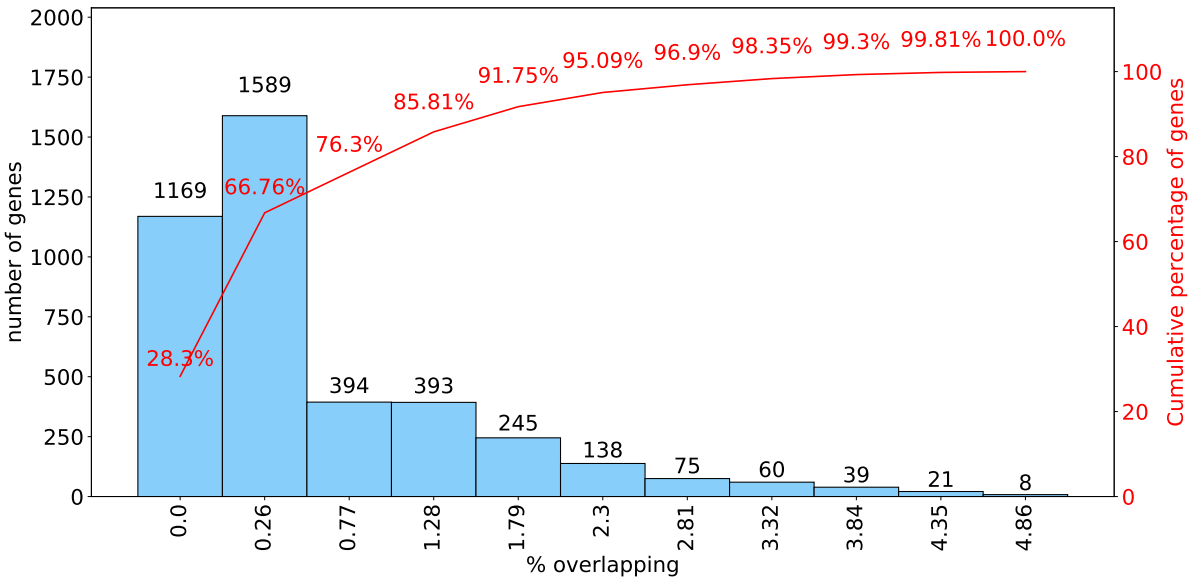

Figure 4.2: Distribution of rice genes across overlapping modules.

Out of the $4\,131$ genes that were categorized into at least one module, $1\,169$ genes were uniquely associated with a single module, while the remaining $2\,962$ genes exhibited overlaps across multiple modules. Among these overlapping genes, a notable subset comprising 168 genes stood out as statistically significant (validated by a Fisher exact test with a p-value < 0.05). Intriguingly, these genes are recognized as Transcription Factors (TFs) that have been drawn from the PlantTFDB database (Jin et al., 2016). This finding lends strong support to the biological relevance of the overlapping modules.

Transcription Factors (TFs) are pivotal in regulating gene expression, impacting various pathways with diverse functions (Renkawitz, 2006). Since TFs wield control over different biological functions, it is expected that they would be found within overlapping modules. This observation underscores the importance of these modules in representing interconnected regulatory networks

that encompass a range of biological processes, all under the orchestration of these transcriptional regulators.

4.4 Computational Validation

The ability of the selected genes to discriminate between control and stress samples was evaluated using a random forest classifier, which was executed 100 times for each set of genes (CSI-OC selected and randomly selected). The difference in the median accuracy values between the two sets was assessed through a Wilcoxon test. While the test revealed significant differences in median accuracy (p-value < 0.05) between the selected and random genes, it is important to note that both medians hovered around 0.5. This level of accuracy is generally considered suboptimal for a robust classification model.

However, the situation improved considerably when evaluating the selected genes' capacity to predict phenotypic traits. As depicted in Figure 4.3, the expression profiles of the selected genes exhibited better predictive performance (lower MSE values) for all phenotypic traits compared to random genes. These results were further reinforced by a Wilcoxon signed-rank test based on the MSE values, revealing statistically significant population differences (p-value < 0.05).

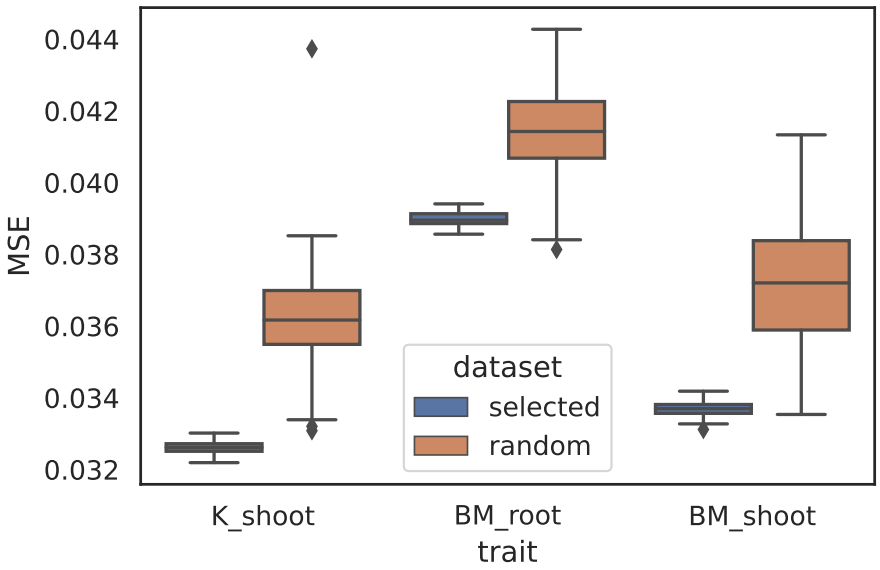

Figure 4.3: Boxplots of the MSE for the phenotypic trait predictions, contrasting the performance of CSI-OC selected genes (in blue) with random genes (in orange).

In summary, the results highlight the effectiveness of the CSI-OC workflow in identifying genes

whose expression profiles demonstrated a stronger association with phenotypic traits compared to random gene sets. While distinguishing between control and stress samples presented challenges, the ability to predict phenotypic outcomes underscores the potential of this approach for gaining valuable insights from gene expression data.

4.5 Knowledge-based Evaluation

From the 19 genes selected by LASSO, 16 genes (84%) are also identified as diferentially expressed ($|\text{LFC}| \geq 2$) for at least one of the 92 accessions.

In general, there are $3\,741$ unselected diferentially expressed genes and $5\,168$ unselected non-diferentially expressed ones, for a total of $8\,909$ genes. Therefore, differentially expressed genes are significantly more likely to be selected by the workflow, as checked by a Fisher exact test with p-value less than 10^{-3}.

The 19 selected genes were also enriched by contrasting them with findings reported in other works, namely, Du et al. (2019); Lahiri et al. (2021); McGowan et al. (2021); Yu et al. (2021), which applied different approaches to study the RNA-seq dataset GSE98455 (i.e., the one used in the case study). In McGowan et al. (2021), 11 of the 19 selected genes are reported to have conserved heritability for both control and salt stress conditions. The identifiers for the 19 genes are listed in Table 4.1. Note that some genes are marked with (*): in the DEG column are those diferentially expressed, and in the H column are those having heritable expression under control and salt stress (as reported in the literature).

Salinity tolerance comes from genes that limit the rate of salt uptake from the soil and the transport of salt throughout the plant, adjust the ionic and osmotic balance of cells in roots and shoots, and regulate leaf development and the onset of senescence Munns (2005). GO terms related to these characteristics, and therefore relevant to salt stress, are found in this case study to be associated with some selected genes. For example, gene LOC_Os12g37260 is annotated with response to abiotic stimulus and response to stress, and gene LOC_Os12g10280 is annotated with response to extracellular stimulus, channel activity, and transmembrane transport. Genes LOC_Os04g12499, LOC_Os04g12530, and LOC_Os12g10280 are annotated with transporter activity, while gene LOC_Os04g35010 is annotated with multicellular organism development.

In-vivo experiments, reported by independent authors, provide evidence on the relationship with salt stress of five genes among the ones selected in the case study (26%). Gene LOC_Os04g12530

Phenotypic trait	Module	TU ID	LOC_Os ID	DEG	H
K_shoot	1	13101.t01457	LOC_Os01g16124	*	*
		13101.t01458	LOC_Os01g16130	*	*
		13104.t01366	LOC_Os04g16230	*	
	2	13104.t01068	LOC_Os04g12520	*	*
		13104.t01069	LOC_Os04g12530	*	*
		13104.t01066	LOC_Os04g12499	*	*
	3	13101.t00913	LOC_Os01g10400		
		13102.t03795	LOC_Os02g41820		*
		13103.t00468	LOC_Os03g05870		*
BM_shoot	4	13101.t02836	LOC_Os01g33450	*	*
		13102.t01261	LOC_Os02g14520	*	
		13107.t03589	LOC_Os07g39390	*	
		13112.t00905	LOC_Os12g10280	*	
BM_root	5	13101.t05133	LOC_Os01g58100	*	
		13112.t02444	LOC_Os12g27254	*	
		13112.t03421	LOC_Os12g37260	*	*
	6	13104.t03155	LOC_Os04g35010	*	*
		13108.t03971	LOC_Os08g42310	*	*
		13109.t01501	LOC_Os09g17049	*	

Table 4.1: Salt-responsive rice genes.

is reported as an up-regulated gene in rice plants tolerant to salt stress (Razzaque et al., 2019). Gene LOC_Os12g10280 encodes an aquaporin nodulin 26-like intrinsic membrane (NIP3;5) protein (Hsieh et al., 2018); it has been shown that NIPs play an important role in salt stress responses and in maintaining plant water balance (Kapilan et al., 2018). Gene LOC_Os04g35010 encondes a protein from the bHLH domain, which has been shown to be part of multiple cellular processes, including salt stress signaling pathways (Qian et al., 2021). Gene LOC_Os12g27254 encodes spermidine hydroxycinnamoyltransferase 2 (SHT2) protein. This protein contributes to the natural variation of spermidine-based phenolamides in rice cultivars, whose activity promotes tolerance to saline stress (Bassard et al., 2010; Roychoudhury et al., 2011; Gupta et al., 2013; Peng et al., 2019). Gene LOC_Os12g37260 encodes Lipoxygenase protein, which is considered to correlate directly with salt tolerance in rice (Mittova et al., 2002; Mostofa et al., 2015; Hou et al., 2015). Note that the STRING database reports a protein-protein interaction of the last two mentioned proteins, namely SHT2 and Lipoxygenase, supporting their membership within the same module, as seen in Table 4.1. In relation to the five genes mentioned, there are 387 other genes known to be involved in salt stress (Razzaque et al., 2019; Liu et al., 2019; Chen et al., 2020). Therefore, it can be said that the number of genes selected by the workflow that are related to salt stress is significant, as checked by a Fisher exact test with p-value less than 10^{-2}.

4.6 Discussion

The results obtained through the CSI-OC workflow showcase its potential in identifying stress-responsive genes in the context of salinity stress in rice. This workflow, employing a Pearson-based network construction method, stands out for its ability to pinpoint 19 rice genes that exhibit relevance in the response to salt stress. These genes are distributed across various modules, shedding light on their association with different aspects of salinity response.

A notable feature of the proposed CSI-OC workflow lies in its module detection technique. Traditionally, the co-expression network analyses have employed methods like Weighted Gene Co-expression Network Analysis (WGCNA) to identify distinct gene modules. However, the CSI-OC workflow implement the Hierarchical Link Clustering (HLC) algorithm, an approach capable of detecting overlapping communities within the network. Many existing module detection techniques primarily focus on non-overlapping communities, and those that deal with overlaps tend to utilize methods like bi-clustering and decomposition (Saelens et al., 2018). Yet, real-world networks, in-

cluding biological ones, often exhibit overlapping community structures (Palla et al., 2005). The CSI-OC workflow, therefore, presents a significant advancement by offering a generalization of previous approaches, like WGCNA, with the potential to capture genes associated with multiple biological processes simultaneously. This adaptability is especially valuable when studying complex biological systems where genes often participate in multiple pathways and functions.

The presence of transcription factors (TFs) among the overlapping genes identified in this study underscores their pivotal role in orchestrating various biological processes, including the response to salt stress. Abiotic stress signals trigger intricate cascades that enlist diverse TFs to modulate the expression of stress-responsive genes, thus enhancing stress tolerance in plants (Khan et al., 2018). These TFs, a subset of the overlapping genes, emerged as statistically significant, further substantiating the biological relevance of the overlapping modules. TFs are well-known for their capacity to regulate the expression of multiple genes, influencing a wide array of pathways with diverse functions (Renkawitz, 2006). Given their versatility, it is entirely expected that TFs would be found within overlapping modules, emphasizing their role as central regulators that govern interconnected biological processes. This observation strengthens the significance of the identified overlapping modules as integral components of the regulatory networks governing salt stress responses in rice.

The computational validation of the CSI-OC workflow's performance provides valuable insights into its capabilities. While the ability of the selected genes to discriminate between control and stress samples exhibited room for improvement, the workflow excelled when assessed based on its capacity to predict phenotypic traits. The expression profiles of the selected genes demonstrated superior predictive performance compared to random gene sets, as evident from the lower Mean Squared Error (MSE) values. This validation highlights the utility of the CSI-OC approach in identifying genes whose expression profiles are closely associated with phenotypic traits, a crucial aspect when studying complex biological responses.

The knowledge-based evaluation of the 19 selected genes provides further validation of the CSI-OC workflow's effectiveness. A significant proportion (84%) of these genes was also identified as differentially expressed in at least one of the 92 accessions studied. This alignment with existing knowledge underscores the workflow's ability to identify genes relevant to salt stress responses. Moreover, contrasting the selected genes with findings from other studies revealed conserved heritability for a subset of these genes under both control and salt stress conditions. This knowledge-based validation strengthens the reliability of the workflow's results and reinforces

the notion that it can extract biologically significant genes associated with stress responses.

In conclusion, the findings presented in this chapter offer compelling evidence for the effectiveness of the CSI-OC workflow in identifying stress-responsive genes, particularly in the context of salinity stress in rice. The ability to detect overlapping communities within co-expression networks represents a significant advancement, as it accommodates genes associated with multiple biological processes. While many of the identified genes align with existing knowledge of salt stress responses, further investigations are needed to elucidate the precise biological functions of the remaining genes, which have not yet been reported in the literature as related to salt stress response. These uncharted genes hold the potential to provide critical insights into the intricate mechanisms governing plant responses to salt conditions in rice and, by extension, other stressors in diverse plant systems.

5. Identifying Drought-Responsive Genes in Sugarcane (*Saccharum spp.* Hybrid)

Agricultural sustainability is under threat due to the changing climate. In light of changing weather patterns, various strategies need to be identified to help plants cope with different levels of stress, such as drought or dehydration. Drought is the most significant abiotic stress (Riyazuddin et al., 2022; Upadhyay, 2019), accounting for a global yield loss of $\$30$ billion USD (Gupta et al., 2020). Crop production is projected to face serious challenges by 2050, according to predictions (Kour et al., 2022). Sugarcane (*Saccharum spp.*) is one of the primary crops worldwide, used to make a variety of products including sugar, bioethanol, and renewable bioenergy (Verma et al., 2022a). In Colombia, sugarcane represents 1.3% of total production, placing it in 11th position worldwide, with a production of 2.1 million Tons of sugar, 347 million liters of bioethanol, and $1,745$ Gwh of electric energy between 2022 and 2023. In total, sugarcane contributes 0.6% of the Gross Domestic Product (GDP) (Asocaña, 2023). Its productivity is negatively impacted by abiotic stress, as it is considered a high water-exhausting crop (Lakshmanan and Robinson, 2013). To this end, most sugarcane plantations are located in southwest Colombia, specifically along the Cauca river, covering $238\,134$ hectares and are classified according to the moisture: semi-dry, humid and foothill (Trujillo-Montenegro et al., 2021).

Responses to drought stress in plants are the result of complex interactions between morphological, physiological, and biochemical processes. They are regulated by multiple genes distributed across various regions of the plant genome (Riccio-Rengifo et al., 2021b). In general, its consequences at the molecular or cellular levels, manifests in phenomic characters such as carbon assimilation rates, turgor pressure, leaf gas exchange, photosynthesis, respiration, carbohydrate metabolism and nutrient uptake (Hussain et al., 2018). For example, one of the most distinctive effects of drought stress is the overproduction of reactive oxygen species (ROS), which leads to alterations in the plant's cellular membranes, ionic imbalance, and oxidation of the biomolecules

(Hussain et al., 2016, 2018). The response to drought stress varies based on the photosynthetic pathways of the plants, with marked differences in C3 and C4 plants (Karami et al., 2023; Lopes et al., 2011; Tahmasebi and Niazi, 2021), such as sugarcane. For example, Tahmasebi and Niazi (2021) suggested several specific mechanisms to respond to drought stress in C4 plants (maize), when compared to C3 plants (rice). Their findings revealed that, while expression related with response to stress, metabolic pathways and photosynthesis are common across both types, other expressions distinctly vary.

Sugarcane is one of the most extensively studied crops for drought stress, underscoring the complexity of responses at all levels (Contiliani et al., 2022, 2023; Kaura et al., 2022; Li et al., 2023; Shrestha et al., 2023; Verma et al., 2022b; Zahoor and Babar, 2023). This chapter aims to apply the CSI-OC workflow to identify relevant genes associated with drought stress in Colombian sugarcane cultivars. To achieve this, we analyzed the gene expression profiles of different cultivars and compared the leaves and roots of three cultivars under three conditions: normal irrigation (control), drought with up to 25% moisture, and drought until the plant's wilting point.

In summary, this chapter is structured to provide a comprehensive exploration of the CSI-OC workflow's application to identify key genes associated with drought stress in Colombian sugarcane cultivars. Section 5.1 delves into the origin of the data, highlighting the RNA-seq and phenotypic data sources, as well as the experimental details that form the foundation of the analysis. In Section 5.3, the outputs from the CSI-OC workflow are presented, covering the number of selected genes for each treatment, their intersections, and differences between leaves and roots. Section 5.4 discusses the significance of overlapping genes within modules and their biological implications, emphasizing genes crucial for plant adaptation under drought stress. Section 5.5 focuses on computational validation, confirming the relevance of CSI-OC selected genes in distinguishing control from stress samples and predicting phenotypic traits. Section 5.6 provides a knowledge-based evaluation, highlighting enriched gene ontologies of selected genes and their importance in stress response. Finally, Section 5.7 offers a comprehensive discussion of these results, elucidating the intricate responses of sugarcane to drought stress at both the molecular and phenotypic levels, enhancing our understanding of plant adaptation to adverse environmental conditions.

5.1 Origin of Data

RNA-seq read counts were provided by Cenicaña (Centro de Investigación de la Caña de Azúcar de Colombia). The estimate of the RNA-seq read counts was handled by Trujillo-Montenegro et al. (2021), where the raw RNA-seq reads were processed. The study focused on three genotypes associated with drought stress: S67, S137, S165, (Saavedra-Díaz et al., 2023). The genotypes were analyzed under three conditions: normal irrigation, medium stress (i.e., dehydration with up to 25% of moisture), and severe stress (i.e., dehydration until reaching plant wilting point). For the transcriptome analysis, samples comprised 5g of leaves and roots sourced from five-month plants under each treatment, with biological duplicates ($r_1 = 6$). The RNA extraction and Illumina paired-end RNA-seq (2x100bp) processing were previously conducted by Cenicaña. Pre-processing and read counts were estimated using FeatureCounts (Liao et al., 2014) by Trujillo-Montenegro et al. (2021) and made available for this study. The reference genome was the cultivar CC-01940 (Trujillo-Montenegro et al., 2021). In terms of phenotypic data, metrics such as the leaf temperature, stomatal conductance, fluorescence of photosystem II and chlorophyll content were recorded, with nine biological and three technical replicates ($r_2 = 27$ by treatment). The instruments used were the High Temperature IR Thermometer (Extech Instruments, USA), porometer SC-1 (Decagon Devices, USA), FluorPenFP 100 model Z990 (Qubit Systems, Canada) and chlorophyllometer SPAD 502 Konica Minolta (Sensing Americas, USA), respectively. This dataset was previously documented by Cenicaña (Riascos-Arcos et al., 2015). DEG analysis had been discussed by Riascos-Arcos et al. (2015) and Trujillo-Montenegro et al. (2021). The phenotypic data supported certain hypotheses in Riascos-Arcos et al. (2015), but were not integrated into their analysis.

For the enrichment analysis, a database comprising gene-GO annotations was essential. Annotations for the CC-01940 genome were provided by Cenicaña. Further annotations were sourced from INTERPROscan (Blum et al., 2021), Monocot plaza (Van Bel et al., 2022), and OMA (Altenhoff et al., 2018), leveranging the CDS of the corresponding genome. We also looked into annotations of genes homologous to Saccharum spontaneum via g:profiler (Raudvere et al., 2019). Ultimately, we amassed $12\,688$ annotations spanning $10\,259$ genes and 519 GO terms, providing a comprehensive backdrop for our enrichment analysis.

5.2 Comparison of Pearson-based and Lasso-based Networks

To establish a comprehensive understanding of gene interactions in response to varying stress levels and tissue types, multiple experimental scenarios were devised using the available dataset, which encompassed different stress levels denoted as C (control), M (medium stress), and S (severe stress), along with distinct tissue types labeled as L (leaf) and R (root). Employing the pairwise comparison approach facilitated by the CSI-OC workflow, focus was placed on six specific cases: CML, CMR, CSL, CSR, MSL, and MSR. For example, the designation CML represents the comparison between control and medium stress conditions specifically in leaf tissue.

Initially, an attempt was made to construct networks using the Pearson-based method. However, the resulting networks proved to be impractical for subsequent workflow steps due to computational constraints. Consequently, the Lasso-based method was opted for.

A comparison between co-expression networks constructed using the Pearson-based and Lasso-based methods revealed significant differences in network topology across all cases (CML, CMR, CSL, CSR, MSL, and MSR) for sugarcane, as presented in Table 5.1. Lasso-based networks consistently demonstrated fewer nodes and edges, indicating a more sparsely connected structure compared to their Pearson-based counterparts. This characteristic aligns with Lasso's capability to prioritize essential gene connections while minimizing noise and spurious correlations.

The average degree, representing the average number of connections per node in the network, offers insights into overall connectivity and complexity. Pearson-based networks exhibit higher average degree values, ranging from $2\,300$ to $9\,332$, indicating denser connectivity with nodes having more connections. In contrast, Lasso-based networks show lower average degree values, ranging from 900 to $1\,809$, reflecting sparser connectivity.

Density, a measure of network connectedness, further underscores these differences. Pearson-based networks consistently display higher densities (ranging from 0.015 to 0.866) compared to Lasso-based networks (ranging from 0.00039 to 0.00103). This pattern suggests that Lasso-based networks are consistently sparser, with fewer connections relative to the total possible connections, than Pearson-based networks across all cases.

The choice of the Lasso-based network construction method for the sugarcane case study is further justified considering the computational demands of subsequent steps in the workflow, particularly the overlapping module identification. The HLC algorithm utilized for this task exhibits

Table 5.1: Comparison of topological properties between Pearson-based and Lasso-based sugarcane networks across all cases (CML, CMR, CSL, CSR, MSL, and MSR)

Case	Network	Number of Nodes	Number of Edges	Average Degree	Density
CML	Pearson	11,227	5,964,920	3,338	0.094
CMR	Pearson	14,096	9,370,059	4,188	0.094
CSL	Pearson	10,871	5,715,989	3,243	0.097
CSR	Pearson	14,204	10,220,720	4,270	0.101
MSL	Pearson	10,921	5,412,059	3,225	0.091
MSR	Pearson	13,652	10,324,297	4,168	0.111
CML	Lasso	4,643	9,327	1,162	0.00086
CMR	Lasso	4,976	8,622	1,245	0.00069
CSL	Lasso	3,597	6,692	900	0.00103
CSR	Lasso	7,234	10,370	1,809	0.00039
MSL	Lasso	3,751	6,163	939	0.00087
MSR	Lasso	6,171	11,633	1,544	0.00061

computational complexity that scales at least quadratically with the number of edges in the network. In scenarios where networks possess dense connectivity, as observed in Pearson-based networks, the computational burden imposed by the HLC algorithm can become prohibitively intensive. By leveraging the sparsity-enhancing properties of the Lasso-based approach, the resultant networks exhibit reduced density and fewer edges. Consequently, the computational requirements for overlapping module identification are significantly mitigated, enabling more efficient processing of large-scale networks, such as those encountered in the study of sugarcane gene expression data. Thus, the adoption of the Lasso-based network construction method aligns with the overarching goal of optimizing computational resources while maintaining the analytical rigor necessary for comprehensive network analysis within the CSI-OC workflow.

5.3 Summary of Selected Genes

In this specific case study focused on sugarcane, the CSI-OC workflow was employed using the Lasso-based network construction method. The workflow was systematically applied to evaluate gene responses to drought stress under various conditions, with a focus on both leaf and root

tissues. The analyses encompassed comparisons of control versus medium stress (CM), control versus severe stress (CS), and medium stress versus severe stress (MS). A comprehensive visualization of the findings can be seen in Figure 5.1, which depicts the cardinalities of selected genes for each experiment and their respective intersections.

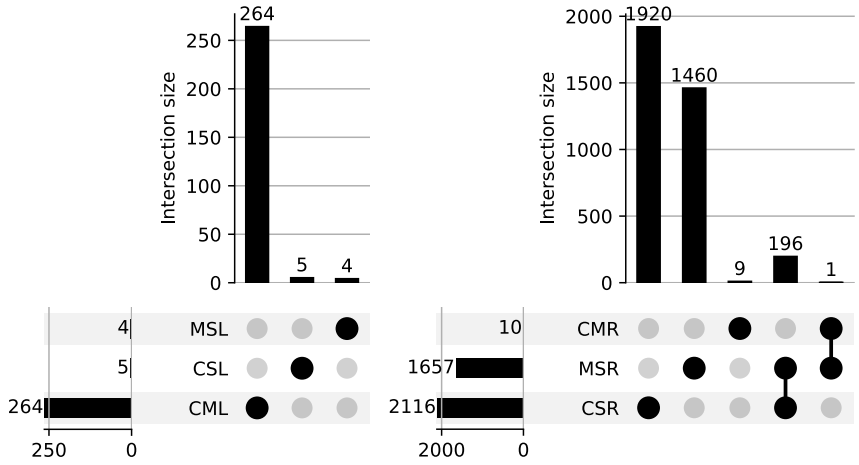

Figure 5.1: Upsetplot illustrating the selected sugarcane genes in leaf tissues (A) and root tissues (B) across all experiments, contrasting Control condition (C), Medium stress (M), and Severe stress (S).

When the moisture level decreased to 25% in leaves (referred to as CML), a high number of genes were selected as stress-responsive (264). However, the selected genes were only five when the stress was maximum (when the plant reached the wilting point, termed CSL). Interestingly, none of the selected genes were shared between treatments in leaves, indicating distinct responses across these stress levels. On the other hand, for roots (denoted as CMR), only ten genes were selected when the moisture level decreased to 25%, while a staggering 2116 genes were highlighted when the plants reached the wilting point (CSR). In terms of this treatment, a comparison between these two stress levels (referred to as MSR) reveled a set of 196 genes that were common to both.

5.4 Overlapping Genes Characterization

In the CML, MSL, and CMR experiments, all selected genes showed an overlap. In contrast, in the remaining experiments, a mere 20% of the selected genes had overlap. The presence of overlapping genes among the selected as stress-responsive suggests a high degree of co-

regulation among these genes, potentially indicating their pivotal roles in the shared response to stress.

When centering our analysis on genes that overlap, before pinpointing those responsive to stress, some interesting findings surfaced. Beyond the anticipated annotations tied to stress responses, there was a marked and statistically significant (p-value < 0.05) representation of the term “regulation of DNA-templated transcription”. This term is intrinsically linked to transcription factors, pivotal in guiding gene expression. It stands to reason that such transcription factors would emerged in multiple modules concurrently, acting as primary regulators in a myriad of biological pathways. Additional noteworthy annotations within the overlapping genes are tied to transport mechanisms: protein transport, proton transmembrane transport, lipid transport, and photosynthetic electron transport in photosystem II. Genes with these transport-related annotations gain prominence when seen in several modules at once, as they are presumed to occupy central roles in integrating, orchestrating, and fine-tuning diverse biological functions necessary for a plant’s resilience to stress. Their ubiquity across modules underscores their importance in bolstering plant survival and adaptability under adverse environmental settings. These findings indicate accounting for overlapping communities in the CSI-OC workflow provides a more comprehensive perspective on intricate gene interplays.

5.5 Computational Validation

The efficacy of the selected genes in distinguishing between control and stress samples was assessed. The experiments comparing control vs medium stress, both in leaf and root, outperformed sets of random genes, as depicted in Figure 5.2. This was substantiated by a Wilcoxon signed-rank test based on the accuracy values, which revealed median significant differences (p-value < 0.05) for those experiments. These results indicate that the CSI-OC workflow allows discriminating genes that classify well for the two control and medium stress conditions.

Furthermore, the efficiency of selected genes in predicting phenotypic traits was evaluated, as shown in Figure 5.3. The expression profiles of the selected genes proved more adept at predicting stomatal conductance (indicated by lower Mean Square Error, MSE, values) than their random counterparts in all experiments. In predicting the SPAD trait, the selected genes outperformed in all experiments except CML. For chlorophyll fluorescence, the he selected genes showed superior prediction capability in the CML and CSR experiments. And for the leaf temperature, the

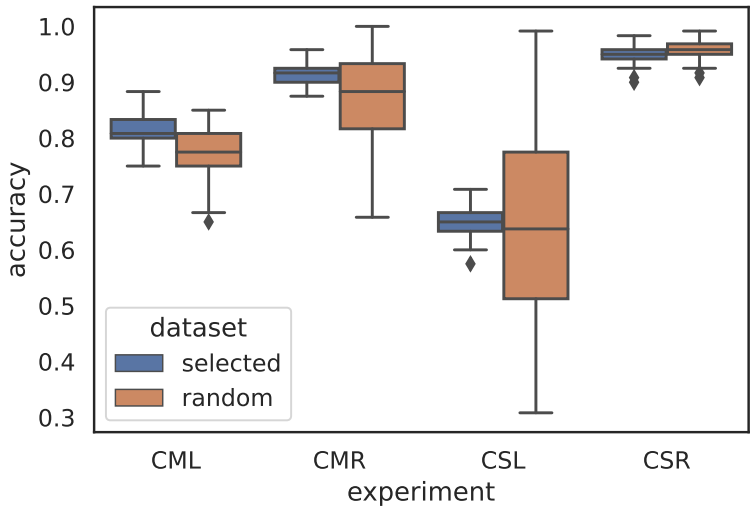

Figure 5.2: Boxplots illustrating the accuracy of the control/stress classification, contrasting CSI-OC selected genes (in blue) with random genes (in orange).

selected genes were more predictive in the CML experiment. These findings were corroborated by a Wilcoxon signed-rank test focusing on the MSE values, which showed significant population differences (p-value < 0.05) for the previously mentioned experiments and traits. This underscores the capability of the CSI-OC workflow in pinpointing genes whose expression patterns more accurately mirror phenotypic traits, as opposed to random gene groups.

5.6 Knowledge-based Evaluation

Figure 5.4 provides a comparison between genes selected using CSI-OC and those up- or down-regulated as identified by DESeq2. This is for leaves and roots subjected to drought stress, either up to 25% of moisture (as depicted in Figure 5.4-A), or up to the wilting point (as shown in Figure 5.4-B). When examining the control vs medium stress in leaves (CML), we identified 29 that were common between CSI-OC and up-regulated ($LFC > 1$) genes for CML. Out of these, seven have unknown functions. The rest are a mix of transcription factors (like ERF09 and ERF053-like, MYB-like, bZIP43-like), negative regulators (such as NGR-like, FOR1-like, LEA17-like), membrane transporters (including BOR4, GATL9), proteins linked to photosynthesis (COX3, psbC, OEP162) and tyrosine kinase receptors (like SIT2, At2g1913). There are also genes associated with photosynthesis, cell wall biosynthesis, and fatty acids production. Specifically, two genes are tied to stress response. based on our results, CSI-OC picks about 10% of the up-regulated genes (264

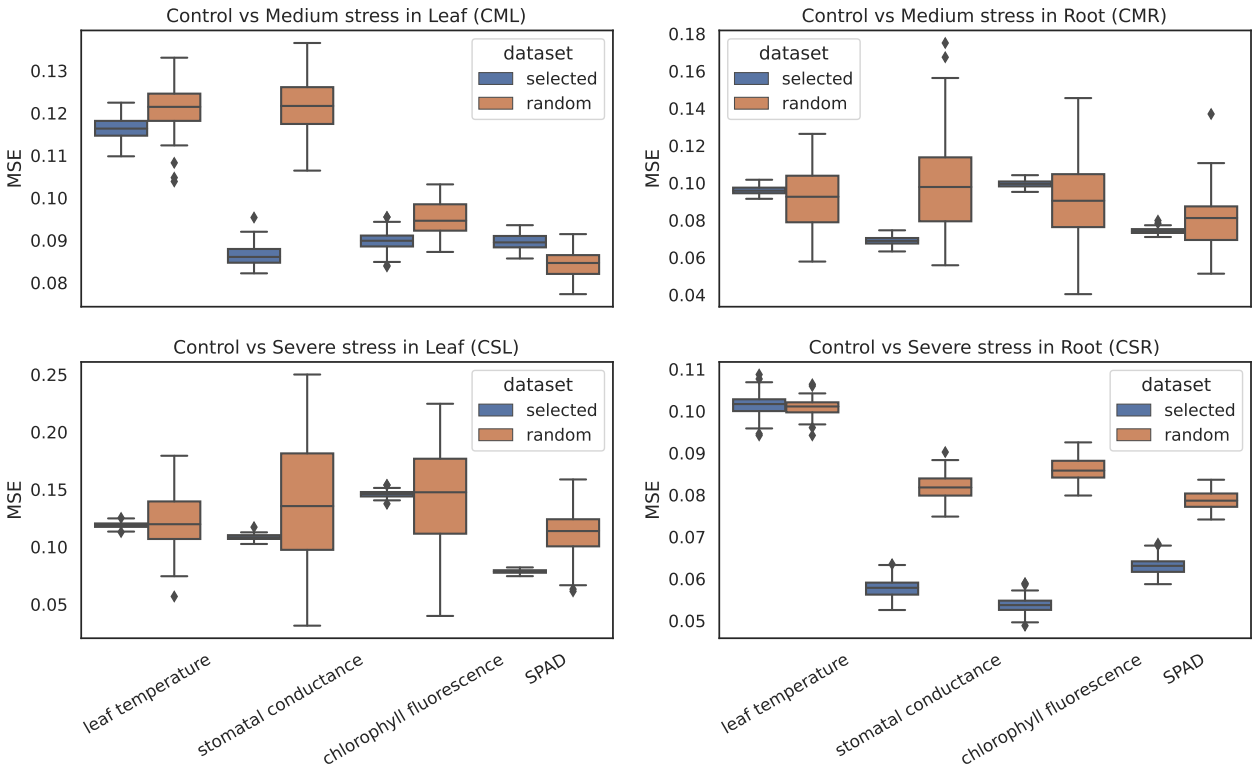

Figure 5.3: Boxplots illustrating the Mean Squared Error (MSE) for the phenotypic trait predictions, contrasting CSI-OC selected genes (in blue) with random genes (in orange) across experiments: CML, CSL, CMR, and CSR.

CSI-OC vs $2\,296$), with a 10% overlap, implying that CSI-OC identifies pertinent genes that might not be strictly correlated with DEG, but still react to the given phenotype. Interestingly, n genes chosen for CML overlapped with down-regulated genes ($LFC < -1$).

With respect to roots, only three genes overlapped with the down-regulated genes identified by DESeq2. These genes encode for an intramembrane protease (IMP, Type II CAAX prenyl endopeptidase Rce1-like, Sobic.002G200200.3), a transcriptional regulator called PHR (Phosphate starvation Response, cc_00024952) and a MLO-like protein (cc_00027902, GO:0006952, GO:0016021). Furthermore, CSI-OC did not reveal any genes common between leaf and root. This indicates that the correlation between gene expression and phenomic data results in distinct modules based on the tissue.

For severe stress (when the stress reaches the wilting point), CSI-OC identified only five genes for leaves. Among them, a leucine zipper transcription factor (cc_00031585) overlapped with down-regulated genes from DeSeq2 results. IN contrast, the root presented 247 CSI-OC coincid-

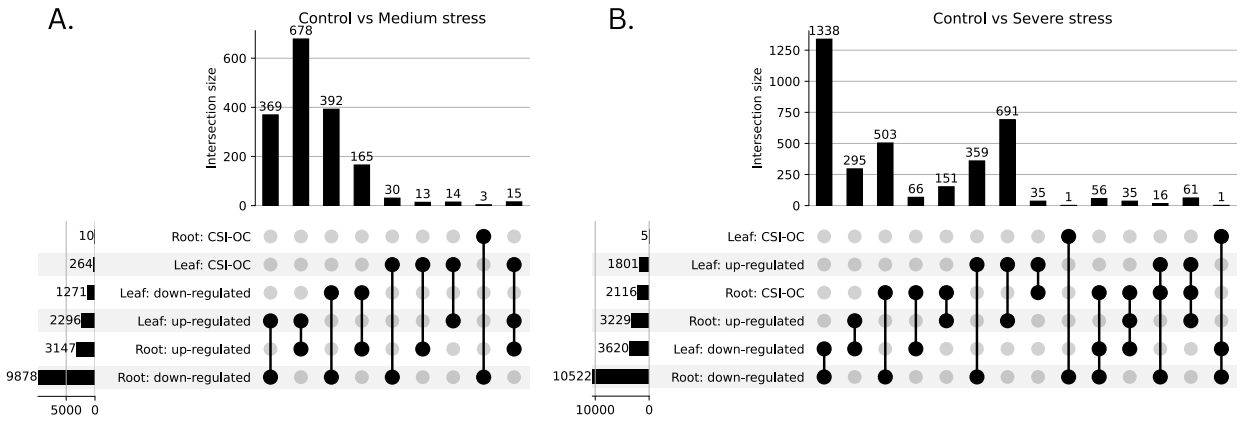

Figure 5.4: Upsetplots comparing up-regulated, down-regulated and CSI-OC selected genes, in leaves and roots under two stress levels: (A.) Control vs Medium stress, (B.) Control vs Severe stress.

ing with DeSeq's up-regulated genes, and another 448 with its down-regulated genes. In total, XSI-OC pinpointed $1\,605$ distinct genes for the CSR experiment.

In terms of enriched gene ontologies (biological processes) from differentially expressed genes selected by CSI-OC, Figure 5.5 reveals nine GOs common between CSI-OC and up-regulated genes for CML. Conversely, three GOs coincided with down-regulated genes, and two GOs were common for both with up- and down-regulated genes. Among these, the most predominant associated biological process was the regulation of transcription, involving 15 CSI-OC genes. This was followed by proteolysis (with 8 genes) and protein ubiquitination (with 6 genes). Interestingly, the most significant GOs enrichment p-values corresponded to proton transmembrane transport, phosphorylation, light reaction in photosynthesis and protein ubiquitination. Aditionally, other GOs have a direct association to drought stress and the evaluated phenotypes, including defense response, photosynthetic electron transport chain, signal transduction, and DNA repair.

For CSL, out of five genes pinpointed by CSI-OC (Figure 5.1), only two gene ontologies emerged as significant (p-value < 0.05): polyamine transmembrane transport and RNA-mediated gene silencing. Upon comparing both stress experiments, a mere four genes were selected by CSI-OC, with two BPs being significant: tRNA aminoacylation for protein translation and proteolysis.

Figure 5.6 showcases the GOs for CMR (Figure 5.6-A) and CSR (Figure 5.6-B). Here, CSI-OC unveiled 40 ontologies with minimal overlap with up or down-regulated genes. Only two over-

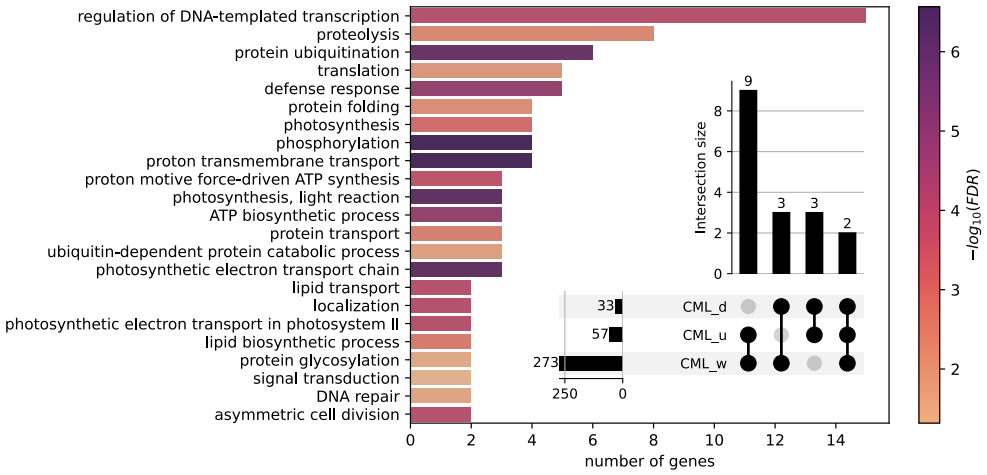

Figure 5.5: Barplot of enriched Biological Processes (BPs) from CSI-OC selected sugarcane genes in Leaves under Control vs Medium stress (CML). Upsetplot comparing up-regulated (u), down-regulated (d), and CSI-OC selected (w) genes.

lapped with up-regulated genes, one coincided with down-regulated ontologies, and two were in common between the ontologies for up and down-regulated genes. The biological process presenting the lowest FDR was response to stimulus. Other noteworthy ontologies included response to stress, defense response, protein processing (e.g., maturation and excision of three aminoacids at carboxy terminal position).

For control vs severe stress in roots (CSR), out of 78 biological processes enriched for genes pinpointed by CSI-OC, four ontologies overlapped with up-regulated genes, five with down-regulated and 12 coincided with both categories. Several shared ontologies such as protein phosphorylation (> 80 genes), oxidation-reduction processes (80 genes), transmembrane transport (40 genes) were highlighted. Notably, the hydrogen peroxide catabolic process registered the lowest FDR.

Finally, in contrasting both stress experiments for root, though $1\,657$ genes were highlighted by CSI-OC, only 80 biological processes met the significance threshold (p-value < 0.05). Among these, seven BPs were in tandem with down-regulated genes, and two aligned with both up and down-regulated genes. None of the ontologies exclusively linked to up-regulated genes. Nevertheless, distinct ontologies between treatments were observed, with the lipid biosynthetic process displaying lowest FDR, followed by protein deubiquitination and glucan metabolic process. The predominantly observed BP was DNA-templated transcription, encompassing nine genes.

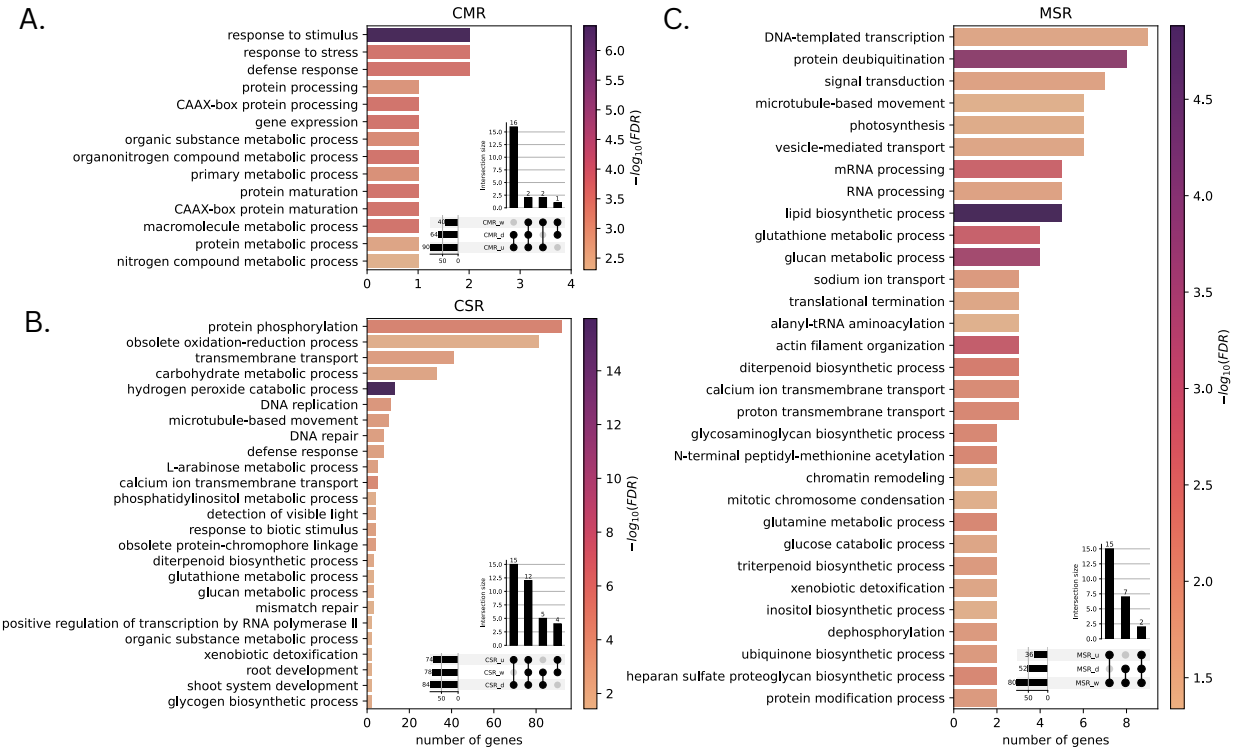

Figure 5.6: Barplot of enriched Biological Processes (BPs) in root tissue from CSI-OC selected sugarcane genes under different stress levels:(A) Control vs Medium stress (CMR), (B) Control vs Severe stress (CSR), and (C) Medium stress vs Severe stress (MSR). The upsetplot compares up-regulated (u), down-regulated (d), and CSI-OC selected (w) genes.

5.7 Discussion

CSI-OC offers a comprehensive workflow for investigating the relationship between specific genes and their role in response to different treatments, while also considering the interconnected genes that have the closest associations with phenotypic traits. This holistic approach allows for a more in-depth analysis of gene interactions and their contribution to stress response mechanisms. The application of the workflow to hybrid sugarcane cultivars posed a unique challenge. The reference genome for the sugarcane cultivars used in this study, CC 01-1940, is notably larger and more complex, containing a total of $63\,724$ predicted genes (Trujillo-Montenegro et al., 2021). This represents three times more genes than the rice nipponbare genome. As a result, when constructing co-expression networks for sugarcane based on gene counts, the traditional Pearson-based method proved to be computationally intensive and challenging primarily due to the requirement of

a threshold to determine the most robust connections for network edges. In fact, Pearson-based networks for sugarcane were found to be unfeasible to process. Two approaches were tried to define the Pearson threshold (Aoki et al., 2007), but both resulted in particularly high threshold values. These high thresholds led to the formation of very dense networks, which, in turn, significantly increased the computational complexity of the subsequent workflow step: overlapping clustering using HLC (Ahn et al., 2010).

This limitation was overcome with the Lasso-based approach to infer sparse co-expression networks within the CSI-OC workflow. This approach not only results in sparse networks, but also provides accurate parameter estimates even when dealing with limited sample sizes (Holmes Finch and Hernandez Finch, 2019). The latter is also a significant limitation of the Pearson correlation coefficient, since it demands large sample sizes for statistically reliable results (Bujang and Baharum, 2016; Liesecke et al., 2019; Ovens et al., 2020), which are often prohibitive due to time and cost constraints in extracting the experimental data. Therefore, the Lasso-based approach for network construction within the CSI-OC workflow is particularly advantageous and reliable in cases with large numbers of genes and few samples, as in the case with sugarcane.

Other sugarcane cultivars have been evaluated with a similar network-based approach. For example, WGCNA, has been used to assess sugarcane tiller seedlings under drought stress (Tang et al., 2023). In this study, they used WGCNA to identify disjoint gene modules, mainly focusing on the hub nodes of the modules significantly associated with physiological or biological traits. In contrast, the CSI-OC goes beyond identifying hub nodes (Vandereyken et al., 2018). It emphasizes overlapping modules, ensuring that multiple genes with interconnected roles are not overlooked. The CSI-OC workflow looks at those genes that, even if they are not central, significantly influence the phenotypic response due to their connections with multiple modules. In terms of computational cost, WGCNA is primarily designed to analyze a single dataset at a time. However, when comparing control and stress conditions, WGCNA employs a technique of stacking the control and stress datasets to create an unified dataset (Langfelder and Horvath, 2008). While this method retains the unique characteristics of each condition, it may result in a larger dataset, potentially increasing computational requirements, especially when dealing with substantial datasets. In contrast, CSI-OC takes a faster approach by calculating the Log2 Fold Change (LFC) for each gene within each genotype, effectively quantifying the gene expression change from control to stress conditions. This method explicitly captures the information related to differential expression. Utilizing a LFC matrix often yields a more concise and focused dataset, emphasizing genes with the

most significant expression alterations (Love et al., 2014). If the objective is to comprehend gene co-expression networks in response to stress conditions, utilizing LFC data can provide a more informative and efficient solution in this sugarcane case study. Finally, CSI-OC avoids the definition of an arbitrary threshold value of Pearson's coefficient value. In contrast, as a result of module selection using Lasso, the output only includes coefficient values different to zero, i.e. relevant genes.

It's noteworthy that the number of CSI-OC-selected genes in leaves was less than in roots. Such a trend implies that CSI-OC might exhibit increased stringency when phenotype data closely aligns with the expression data, as observed when both datasets originate from the same tissue (leaf). This heightened stringency may arise from the analysis becoming more tailored to a specific tissue which in turn, accentuates the expressions patterns of stress-responsive genes within that designated area. Furthermore, this observed behavior can be attributed to the fact that roots are often the primary tissue impacted by most abiotic stresses in grasses (Kang et al., 2022; Kul et al., 2021). This observation corroborates the effectiveness of the proposed workflow. As a result, genes that are distinctly associated with stress response in roots become more prominent. This leads to a refined selection of genes that are pivotal in facilitating stress adaptation within that tissue.

The phenotypic traits used with CSI-OC workflow encompass a range of parameters, including leaf temperature, stomatal conductance, fluorescence of photosystem II, and chlorophyll content. However, a wide range of additional possibilities could be considered, including physiological parameters (e.g., transpiration rate), physical measurements (e.g., internode or root height, stem or leaf diameter), and crop performance indicators (e.g., dry biomass weight, extraction yield, productivity). Moreover, phenotypic data can also encompass metabolite quantification, such as carbohydrates, nitrogen levels, or specific secondary metabolites. By leveraging the existing high-throughput technologies in both molecular and phenomic fields, the workflow integrates these diverse data types to enhance the understanding of plant responses under several environmental stresses.

The computational validation demonstrates the meaningfulness of genes identified by CSI-OC in each treatment. These genes outperformed randomly selected gene sets in both classification and regression tasks. In classification, CSI-OC-selected genes consistently excelled in distinguishing control from stress samples. In regression, they demonstrated superior predictive capabilities for phenotypic traits. Wilcoxon signed-rank tests (Woolson, 2007) confirmed significant

differences between the selected genes and random gene groups, affirming the effectiveness of the CSI-OC workflow in identifying stress-related genes with close ties to phenotypic traits. Functional enrichment analysis, discussed below, further supports the involvement of CSI-OC-selected genes in vital biological processes and the regulation of stress-related phenotypic traits.

The results displayed a contrasting pattern in the number of CSI-OC selected genes between leaves and roots. In the leaf tissue, a decrease in the count of responsive genes with escalating stress level was found. Conversely, root tissue demonstrated an opposite behavior, marking an uptick in gene counts at the wilting point. This dichotomy points towards a sequential flow of gene responses, transitioning from leaves to roots, as the plant encounters increased stress levels. The data hints that the timing of the stress response within a particular tissue might have implications on the overall reaction, suggesting that the plant coordinates responses between leaves and roots to ensure a swift, cohesive adaptation to drought conditions (Hsiao and Xu, 2000; Kang et al., 2022; Kul et al., 2021; Min et al., 2020; Vives-Peris et al., 2020). This alignment is evident in the synchronized gene expression patterns across tissues, which mirrors analogous biological processes noted in prior research on carbohydrate metabolism in plants such as (Du et al., 2020), Vicia sativa (Min et al., 2020), and Rose (Li et al., 2021).

Regarding the biological functions of selected genes by CSI-OC, some of these genes were found to be shared with up-, down- or both-regulated genes. However, unique expressed genes were also found to be not significantly different from the control experiment but having relevant gene expression network interactions, including the SAP (Stress Associated Protein) genes (Muthuramalingam et al., 2021). The importance of these types of genes is highlighted due to their relevance for drought stress response and the risk of overlooking them if solely relying on DESeq2 results for analysis. In leaves, for example, CSI-OC selected cc_00006300, a Knottin motif, whose function is associated with enhanced immunity in plants, defense against biotic stress, with antifungal or insecticide activities (Tavormina et al., 2015; Zhang et al., 2023). The gene cc_00002124 has a Gcn5-related N-acetyltransferase (GNAT) domain profile that was upregulated and associated with drought tolerance in sorghum (Devnarain et al., 2019). However, GNAT was negatively affected by a miRNA in Populus trichocharpa under temperature stress (Das et al., 2021), suggesting that this mechanism of regulation could lead to its exclusion in DESeq2 results. Nevertheless, CSI-OC methodology successfully identified the cc_00002124 gene. Similarly, cc_00008263, a gene that codes for an Argonaute protein (AGO1) and strongly associated with dehydration and RNA silencing (Cui et al., 2020), was surprisingly not detected by DESeq2.

Genes associated with BHLH transcription factor (Helix-loop-helix DNA-binding domain) were selected by CSI-OC, but not by DESeq2. Tang et al. (2023) and Contiliani et al. (2023) also reported this transcription factor by WGCNA and as an up-regulated gene in DESeq2, respectively, as a key regulator associated with drought stress in sugarcane, regulating vertical growth associated with ABA (abscisic acid), but also hypothetically in an independent manner (Tang et al., 2023). The ABA dependent response should explain our results, where the first response (determined by higher selected genes) occurred in the leaves than roots (comparing the absolute number of selected genes), and after, when the stress is at its maximum, the roots become responsible for almost all drought stress response (Surya Krishna et al., 2023). Contiliani et al. (2023) also highlighted the presence of up-regulated GSTT, a glutathione S-transferase, as a key gene for redox homeostasis in sugarcane.

In the roots, when looking at the genes exclusively selected by CSI-OC under medium stress conditions with moisture levels at 25%, gene cc_00001481 was the only one that exhibited significant and distinct biological processes related to drought stress, including the regulation of DNA and RNA, as well as responses to stimuli and stress. The most relevant GO was response to stimulus, suggesting that genes cc_00001481 and cc_00027902 are highly responsive in root. By contrast, when the stress reached the wilting point, BPs were related to response to oxidative stress (14 exclusive selected genes), defense response (15 genes), DNA reparation (13), protein degradation (14 genes) or xenobiotic transport or detoxification through the plasma membrane (12 genes). Finally, when assessing the comparison between stress-level treatments, it was observed that the lipid biosynthetic process (involving six genes) gained relevance during the transition from low moisture to extreme moisture scarcity, indicating a potential increase in the production of fatty acids within root cells in response to various drought-related stresses.

6. CSI-OC Workflow Performance and Validation with Synthetic Data

The generation of synthetic expression data provides innovative solutions to address critical challenges across various scientific disciplines, including issues related to data availability and experimentation (Savage, 2023). Synthetic data plays a pivotal role in enriching real-world datasets by augmenting them with additional data points and features. This augmentation significantly enhances the performance of various applications, including machine learning models, statistical analyses, data visualization, and hypothesis testing, especially in scenarios where real data is limited (Chen et al., 2021).

Cost-effectiveness is a compelling advantage of synthetic expression data generation. It becomes particularly evident, when compared to the comprehensive expenses associated with extracting real gene expression data, especially in plants. Extracting real gene expression data involves a multitude of costs, including those for sample collection, preparation, and preservation, as well as the selection of appropriate sequencing technologies and the sequencing process itself. Moreover, the need for specialized equipment and skilled personnel for RNA extraction, library preparation, and data analysis further escalates expenses (Conesa et al., 2016; Shendure et al., 2017). Additionally, gene expression experiments often require rigorous experimental design, increasing costs associated with multiple replicates, treatments, and time points. In contrast, synthetic expression data generation offers a cost-efficient alternative, circumventing many of these challenges. It enables researchers to simulate complex biological scenarios, reduce data collection and storage costs, and streamline experimental design while still providing valuable insights and advancing scientific knowledge in genomics (D'amico et al., 2023).

In addition to cost savings, synthetic expression data generation offers researchers a higher degree of control and flexibility. Researchers can tailor synthetic datasets to simulate specific scenarios, rare events, or experimental conditions that may be challenging or expensive to replicate

in real experiments. This level of control not only accelerates research, but also can allow for the exploration of a wide range of hypotheses without the constraints of real-experiment data. Synthetic data serves as a valuable tool for testing and validating novel methods in gene expression analysis, for it allows researchers to thoroughly assess the strengths and weaknesses of these methods before their application to real-world datasets, ensuring robust and reliable results.

As part of this research, a primary focus is the utilization of synthetic expression data for comprehensive testing and the identification of limitations within the CSI-OC workflow. As previously outlined in Chapter 3, this workflow relies on input data that encompasses gene expression and phenotypic traits measured under both control and stress conditions. The specific interest lies in evaluating the extent to which the CSI-OC workflow is capable of producing reliable results in relation to the degree of discrepancy between control and stress input data. To achieve this, it is imperative to have synthetic expression data that accurately mirrors various stress levels, as reflected in phenotypic trait values.

While numerous methods exist for generating synthetic expression data, previous research (Maier et al., 2013) has indicated that they often struggle to faithfully replicate the key properties of real gene expression data, such as, non-linearity, variability, and contextual information. In response to this challenge and to better align the model for synthetic expression data generation with the present research needs, we have developed our own model named CoSynthEx. This model focuses on the conditional generation of synthetic expression data, designed to create a more authentic simulation of gene expression data by integrating additional contextual information, including phenotypic traits and sample conditions (e.g., control or stress).

In summary, this chapter introduces CoSynthEx, a model designed to create synthetic expression data, and elucidates its role in assessing the CSI-OC workflow. To this end, Section 6.1 provides a comprehensive overview of the CoSynthEx model, emphasizing its reliance on real expression and phenotypic data. The real data used in this context corresponds to the rice dataset introduced in Chapter 4. Consequently, Section 6.2 offers a detailed statistical analysis of the rice data to enhance the understanding of the patterns distinguishing the control and stress data. Moving forward, Section 6.3 outlines the procedural framework for generating and testing synthetic expression data within the CSI-OC workflow. Section 6.4 presents the outcomes of the experiments and, finally, Section 6.5 provides the corresponding discussion of the main results.

6.1 CoSynthEx

CoSynthEx employs a conditional generative adversarial network (cGAN) as its foundational model (for additional background details, refer to Section 2.3). The utilization of a cGAN model enables the creation of synthetic expression data tailored to specific conditions and biological contexts. The architecture of the CoSynthEx cGAN is depicted in Figure 6.1. This cGAN operates by taking phenotypic trait data, denoted as P, and sample condition information (e.g., control and stress), denoted as Q, as conditional information. The generator network within the cGAN receives this conditional information along with random noise data, represented as R. In response, the generator produces synthetic expression data denoted as E_{synth}. The synthetic data is then passed on to the discriminator along with real expression data (denoted as E_{real}) and their corresponding conditional information (P and Q). The discriminator's role in this adversarial duo is to distinguish between real and generated data, while the generator aims to deceive the discriminator by creating synthetic data that closely resembles the real one. This dynamic process is essential for the successful operation of the cGAN and the generation of highly convincing synthetic expression data.

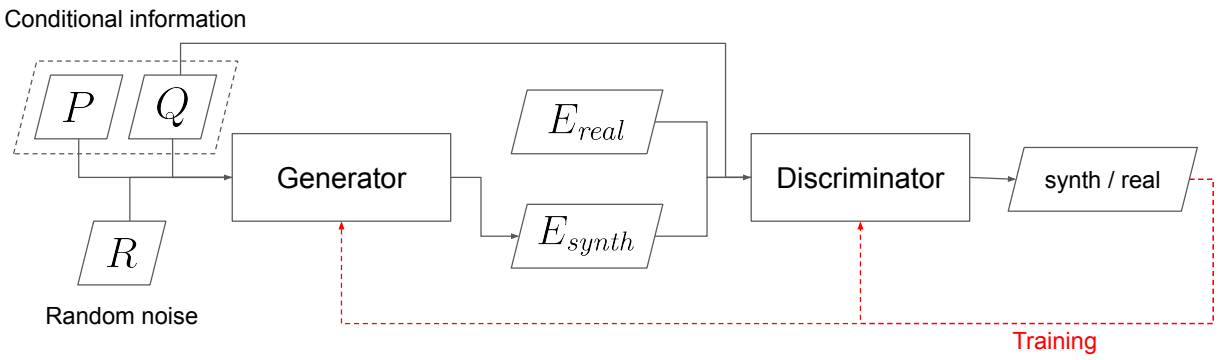

Figure 6.1: CoSynThEx cGAN architecture. The red line illustrates the feedback loop where the discriminator's outputs influence both its own learning and that of the generator throughout the training process.

The generator network consists of four layers: the input layer, two hidden layers, and the output layer. The activation function for the hidden layers is ReLU (Bai, 2022), while the output layer employs the sigmoid activation function (Rasamoelina et al., 2020; Kaloev and Krastev, 2021). Similarly, the discriminator follows a comparable architecture, featuring these same four layers and

activation functions. However, the discriminator introduces dropout in its hidden layers, to enhance robustness, generalization capabilities, and overall stability. This, in turn, leads to improved training dynamics and contributes to the generation of high-quality synthetic samples by the generator.

Through a process of rigorous training and parameter optimization, the model's loss curves should be finely tuned to exhibit ideal behavior. The generator's loss should begin at a relatively high point and progressively decrease during the training process, eventually converging to a low value. This trend indicates that the generator is effectively learning to generate data that closely resembles the distribution of real data. On the other hand, the discriminator's loss, at the outset, should start relatively high but gradually decreasing as well, ultimately stabilizing at a low value. This dynamic illustrates the success of the training process, where the discriminator faces challenges in distinguishing between real and synthetic data due to the generator's increasing ability to produce convincingly realistic data.

6.2 Statistical Analysis of Real Input Data

Recall from Section 4.1, rice expression data encompasses a total of $57,845$ gene expression profiles measured across 184 samples under both salt stress and control conditions. To initiate the analysis, we conducted a visual exploration employing violin plots (see Figure 6.2). These plots allowed us to examine various statistical aspects, including measures of central tendency (e.g., the median), data spread (as indicated by the interquartile range), and skewness, particularly focusing on the distinctions between control and stress conditions.

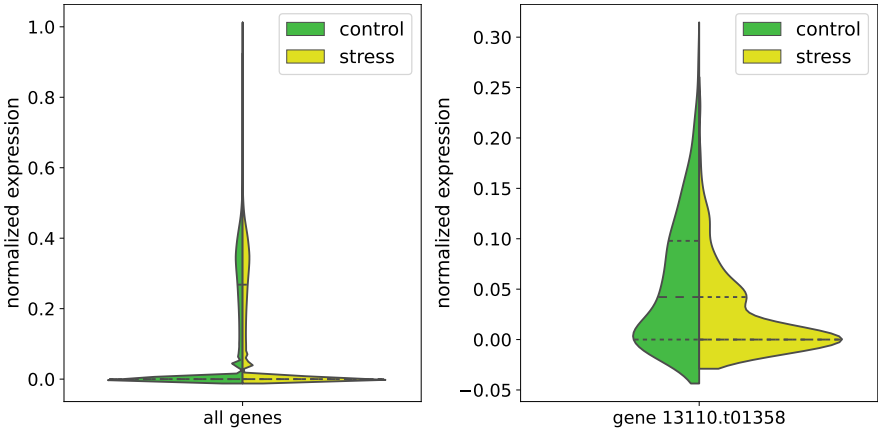

Figure 6.2: Violin plots of rice expression data under salt stress (yellow) and control conditions (green).

As illustrated in Figure 6.2, the expression data did not conform to a normal distribution. Furthermore, the left side of the figure demonstrated that discerning differences between the control and stress distributions in the overall dataset, considering all genes collectively, was challenging. However, a closer examination of individual genes, as depicted on the right hand side of Figure 6.2, revealed observable distinctions in the expression data. For example, the distribution of expression data for gene 13110.t01358 displayed noticeable variations between control and stress conditions. These distinctions were subsequently confirmed to be statistically significant through a Mann Whitney U test (p-value < 0.05). This analysis suggests that the differences in control and stress expression data are more pronounced at a specific gene level, rather than being general trends across the entire dataset.

Regarding the phenotypic traits, these include shoot potassium content (K_shoot), shoot biomass (BM_shoot), and root biomass (BM_root) measured for the same 184 samples as expression data. In this case, we also begin with a visual exploration of the data using violin plots to discern differences between the control and stress conditions for the trhee phenotypic traits. Figure 6.3 suggests that, in contrast to the expression data, significant differences are evident in the values of these phenotypic traits when comparing stress and control conditions.

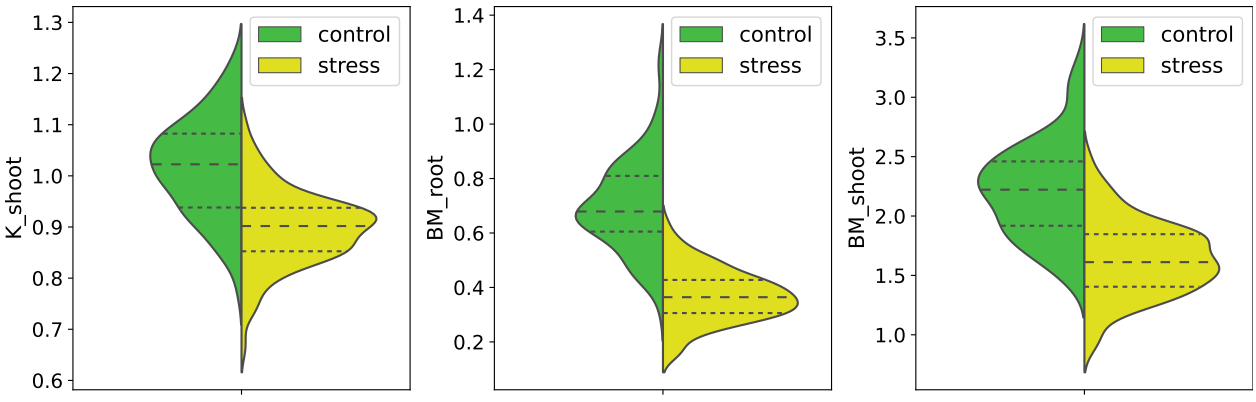

Figure 6.3: Violin plots of rice phenotypic traits (K_shoot, BM_root, and BM_shoot) under salt stress (yellow) and control conditions (green).

Note also from Figure 6.3 that the data distributions exhibit a shape reminiscent of a normal distribution. This observation is substantiated by a Shapiro-Wilk test, which confirms that the data distributions are not significantly different from a normal distribution in all cases (p-value > 0.001). The data's normality enables the application of an F-test to validate the significant difference in standard deviations between the control and stress data for all the phenotypic traits (p-value < 0.05).

With these assumptions met, a Welch t-test is applied to finally corroborate the initial observation. The results affirm that the difference between the means of the control and stress data is indeed significant (p-value < 0.05) for all phenotypic traits.

The statistical analyses conducted in this section underscore the intricate nature of the expression data. The pronounced disparities observed in the control vs. stress phenotypic data provide substantial evidence that these three traits are closely linked to the stress response in rice under salt stress conditions. Furthermore, these findings emphasize the significance of external data, such as phenotypic information, in effectively distinguishing between control and stress datasets.

6.3 Framework for CSI-OC Validation with CoSynthEx

The primary goal is to evaluate the extent to which the CSI-OC workflow yields meaningful results. This evaluation involves testing CSI-OC in various scenarios and employing computational validation to determine in which scenarios the selected genes hold meaningful significance.

To ensure a comprehensive evaluation, it is essential to consider multiple scenarios. Given the challenges associated with obtaining real data, resorting to the generation of synthetic data becomes necessary. Utilizing phenotypic traits as the foundation for scenario design is particularly advantageous, as they directly reflect differences between control and stress conditions. Real phenotypic data exhibiting normality further enables the simulation of various scenarios or stress levels by employing a normal distribution function with distinct parameters, such as mean and standard deviation. To complement the synthetic data, the CoSynthEx model facilitates the generation of expression values corresponding to the simulated phenotypic data for each scenario.

To generate expression data corresponding to each scenario, the CoSynthEx model should be previously trained with real data. After the training is completed, the discriminator is no longer required, and only the generator part of the network needs to be retained to produce new synthetic expression data for each scenario.

The step-by-step framework for testing the CSI-OC workflow with synthetic data for a specific scenario is outlined in Figure 6.4. From a global perspective, each scenario is characterized by mean and standard deviation values for the stress phenotypic data. The process unfolds in four steps, ultimately resulting in a p-value that allows to determine whether the CSI-OC workflow detects meaningful results in that particular scenario.

The first step involves taking the parameters of the respective scenario (μ and σ) as input and

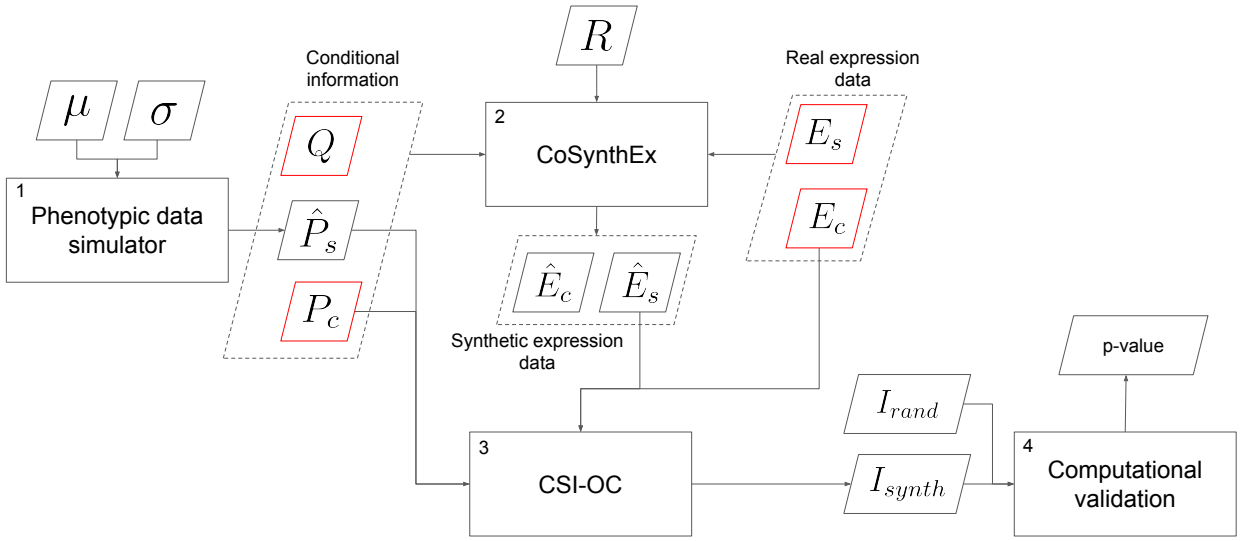

Figure 6.4: Framework for CSI-OC validation with synthetic data. The squares represent the four individual steps of the process. Trapezoids symbolize the inputs and outputs, with the red trapezoids denoting the real data that remains constant across all scenarios.

simulating phenotypic data for the stress condition ($\hat{P}_s$) using a normal distribution function. In the second step, synthetic expression data is generated using the CoSynthEx model. This step incorporates the simulated data from the previous step ($\hat{P}_s$), random noise (R), and the remaining inputs derived from real data. These inputs encompass the expression data (both under control and stress, denoted as E_c and E_s, respectively), the phenotypic data under control (P_c), and the sample condition vector (Q). It's worth noting that these inputs, sourced from real data, remain constant across all scenarios. The outcome of the second step includes both control and stress synthetic expression data ($\hat{E}_c$ and $\hat{E}_s$, respectively). Nevertheless, the objective is to maintain the control data unaltered, so we exclusively use the synthetic expression data under stress for the subsequent step.

The third step entails the application of the CSI-OC workflow, where the input involves real control data (both expression and phenotypic, represented as E_c and P_c) and synthetic stress data (both expression and phenotypic, denoted as $\hat{E}s$ and $\hat{P}s$, respectively). The output of this workflow consists of a set of selected genes, labeled as I_{synth}. To assess the significance of this gene selection, we compare it to a set of genes randomly selected (I_{rand}) with an equivalent cardinality. Subsequently, both sets undergo the fourth and final step, corresponding to the computational validation process. The result is a p-value; when this p-value falls below the significance level,

it indicates that the genes selected by the CSI-OC workflow outperform the randomly selected genes and the CSI-OC workflow yield meaninful results for the corresponding scenario.

6.4 Results

The framework was effectively applied to evaluate the CSI-OC workflow under a range of diverse scenarios using synthetic data. This testing process involved several key steps. Firstly, the CoSynthEx model was meticulously trained using authentic data. Subsequently, a comprehensive set of 25 distinct scenarios was systematically created. Finally, the CSI-OC workflow was tested and evaluated in these scenarios.

Figure 6.5 illustrates the loss curves associated with the CoSynthEx cGAN components during their training. The generator and discriminator were assessed using the Binary Cross Entropy (BCE) loss function, spanning a total of 200 epochs. The graphs reveal valuable insights into the training dynamics. It's noteworthy that during the first 20 epochs (or so), the discriminator's loss rapidly decreases and the generator's loss increases. This observation implies that the discriminator can easily differentiate between synthetic and real samples in the early stages, while the generator initially produces low-quality samples. However, as the training progresses, the generator's performance improves, making it increasingly challenging for the discriminator to distinguish between real and synthetic data. This adversarial interplay continues until the system reaches a state of equilibrium approximately after the initial 100 epochs. At this point, a balance is achieved, with the generator producing high-quality samples that can effectively deceive the discriminator, which still maintains its robust performance.

In the scenario design phase, a total of 25 unique scenarios were created. Each scenario was defined by the normal distribution parameters for stress phenotypic data, while keeping the control data unchanged. These scenarios were constructed by dividing the range between the mean of the real stress data (μ_a) and the mean of the real control data (μ_b) in four equal segments, resulting in five mean values for each phenotypic trait. For instance, considering a trait $z \in \{\text{K\_shoot}, \text{BM\_root}, \text{BM\_shoot}\}$, we derive $\mu_z = [\mu_{a,z} = \mu_{0,z}, \mu_{1,z}, \mu_{2,z}, \mu_{3,z}, \mu_{4,z} = \mu_{b,z}]$. The same process is applied to standard deviations. To facilitate subsequent handling, the means are rearranged, yielding, for example, $\mu_0 = [\mu_{0,\text{K\_shoot}}, \mu_{0,\text{BM\_root}}, \mu_{0,\text{BM\_shoot}}]$ (and a similar arrangement for standard deviations). With this organization, the different scenarios can be envisioned as a 5×5 mesh, denoted as Δ, where rows correspond to means and columns represent standard

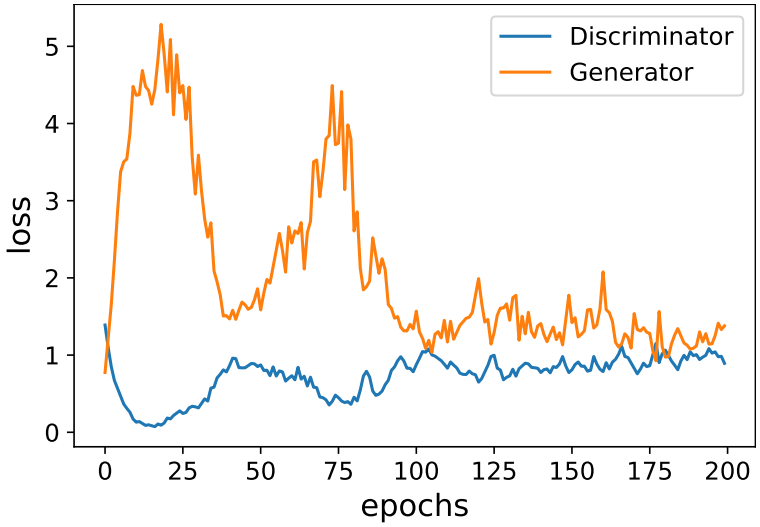

Figure 6.5: Loss Curves of CoSynthEx cGAN Training. The graph displays the loss curves of the CoSynthEx cGAN components during training. Binary Cross Entropy (BCE) was used as the loss function, and the training spanned 200 epochs.

deviations. In the scenario $\Delta(\mu_0, \sigma_0)$, the difference between control and stress parameters (both in terms of mean and standard deviation) is at its maximum, resembling the real phenotypic data. Conversely, the scenario $\Delta(\mu_4, \sigma_4)$ indicates no distinction between the parameters of phenotypic traits under control and stress conditions, representing the opposite extreme.

Figure 6.6 provides a comprehensive overview of the results obtained by running the workflow illustrated in Figure 6.4. The aggregated findings are represented in a heatmap that describes the scenario mesh. It consolidates 10 iterations within each of the 25 scenarios. Within this heatmap, each cell value represents the number of times, out of the 10 iterations, in which the resulting p-value was less than 0.05. It's key to recall that a p-value below this significance threshold indicates that the genes chosen by the CSI-OC workflow outperform randomly selected genes in a binary classification task that distinguishes between control and stress samples. Performance on this classification task determines the effectiveness or failure of the CSI-OC workflow.

Several significant patterns emerge from this heatmap analysis. The first notable pattern is the effect of proximity to real data. Scenarios located above the secondary diagonal, closer to scenario $\Delta(\mu_0, \sigma_0)$, representing the parameters of the real data, tend to yield values closer to 10. This pattern indicates that the workflow consistently produces meaningful results when the scenarios resemble the real data. In contrast, scenarios located below the secondary diagonal, closer to scenario $\Delta(\mu_4, \sigma_4)$, where there is no difference between the control and stress distributions of

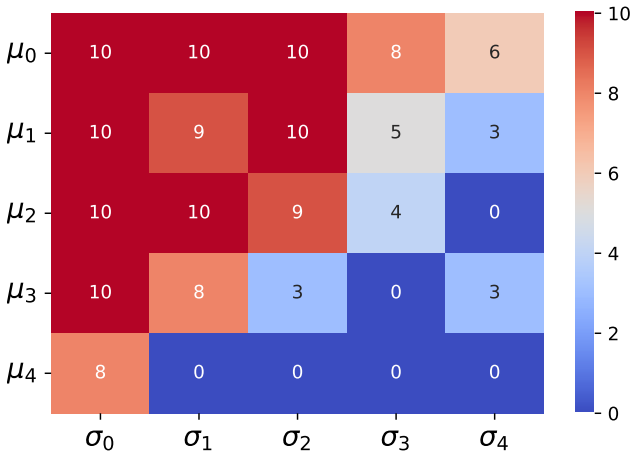

Figure 6.6: This heatmap summarizes the results of running the CSI-OC workflow 10 times in each of the 25 synthetic scenarios. Each cell value represents the number of times the p-value was less than 0.05, indicating the workflow's effectiveness in distinguishing between control and stress samples.

the phenotypic data, tend to result in values close to zero. This outcome suggests that the genes selected by the CSI-OC workflow behave similarly to random genes in these scenarios.

Another interesting pattern is the impact of equal means between control and stress phenotypic data. This condition, represented by scenarios associated with μ_4, is a critical factor influencing the CSI-OC workflow's performance. In most cases, when the phenotypic data's means under control and stress conditions are indistinguishable, the workflow fails to produce meaningful results. An exception occurs in the scenario $\Delta(\mu_4, \sigma_0)$, where the standard deviation difference between control and stress phenotypic data mirrors the difference in the real data. This scenario results in evidence that the CSI-OC workflow performs successfully.

Finally, another noteworthy pattern becomes evident in scenarios linked to σ_0, where the standard deviation difference between control and stress phenotypic data precisely matches the difference in the real data. In all scenarios associated with σ_0 robust evidence suggests that the workflow performs effectively under these conditions. These scenarios consistently yield high success rates, with nearly all values equal to 10. This pattern highlights the successful performance of the CSI-OC workflow when the spread of phenotypic data closely mirrors the characteristics of the real data.

6.5 Discussion

In this chapter, we have presented a comprehensive framework for testing and validating the CSI-OC workflow using synthetic data. This validation process is crucial for assessing the reliability and robustness of the CSI-OC workflow when applied to real data.

Using synthetic data, we had the advantage of controlling various parameters, allowing us to evaluate how well the workflow performs under different conditions. Phenotypic synthetic data was simulated using a normal probability distribution, a choice supported by the behavior of real data. However, the generation of synthetic expression data proved to be more intricate, leading to the development of the CoSynthEx model. CoSynthEx was designed to create a more authentic simulation of gene expression data by integrating information about phenotypic traits and sample conditions (e.g., control and stress). This model is founded on a conditional Generative Adversarial Network (cGAN) that trains a generator network to produce synthetic expression data while a discriminator network distinguishes between real and synthetic data.

The loss curves of the CoSynthEx cGAN components provided valuable insights into the training dynamics. During the initial phases of training, the discriminator had the upper hand as it could easily discern between real and synthetic samples. Meanwhile, the generator struggled to produce high-quality samples. However, as training progressed, the generator improved its performance, leading to an equilibrium in which it could generate convincing samples after approximately 100 epochs.

The chapter also focused on the design of scenarios. Each scenario was characterized by the normal distribution parameters of stress phenotypic data, while the control data remained constant. This systematic approach allowed us to assess the impact of different phenotypic data distributions on the workflow's performance.

Experimenting with just one scenario instead of multiple scenarios can limit the understanding of the workflow's performance under diverse conditions. Different scenarios represent diverse distributions of phenotypic data, offering a comprehensive view of how the workflow adapts to various conditions. Exploring only one scenario limits the understanding of the workflow's adaptability and performance range. It might not reflect the workflow's behavior in other scenarios, potentially overlooking limitations or strengths present in different settings. Focusing on a single scenario can still yield valuable insights and certain types of knowledge. For instance, assessing multiple runs or iterations within the same scenario can provide insights into the stability or variability of results

produced by the workflow under consistent conditions.

The scenarios differ based on the parameters used to generate synthetic data. These parameters include the mean (μ) and standard deviation (σ) of the phenotypic trait data under stress conditions. Within each scenario, the key difference between the 10 iterations lies in the random initialization and training of the CoSynthEx model, which generates synthetic expression data based on the specified parameters of the scenario. Despite using the same parameters for mean and standard deviation to define the scenario, each iteration involves a distinct random seed or initialization for the CoSynthEx model, leading to variations in the generated synthetic expression data. These variations can arise due to the stochastic nature of the training process and the inherent randomness involved in generating synthetic data using Generative Adversarial Networks (GANs). As a result, each iteration may produce slightly different synthetic expression data, even though they all correspond to the same scenario.

The heatmap representing the results of testing the workflow in various scenarios revealed distinct patterns. Scenarios closely resembling the parameters of real data consistently produced meaningful outcomes, with p-values less than 0.05 in nearly all 10 iterations. In contrast, scenarios where there was no difference between control and stress phenotypic data displayed performance similar to random gene selection. This analysis underscored the significance of phenotypic data characteristics in determining the CSI-OC workflow's performance. The similarity of scenario parameters to those of real data and the presence of distinct control-stress phenotypic data distributions emerged as critical factors for achieving meaningful results.

In summary, this chapter's validation framework provides essential insights into the behavior of the CSI-OC workflow under various conditions. The results derived from this chapter can guide the application of the CSI-OC workflow to real datasets, aiding researchers in understanding its limitations and strengths in different scenarios.

7. CSI-OC Validation in a Non-biological Context

The CSI-OC workflow, as described in Chapter 3, draws inspiration from a biological context. Its primary objective is to identify genes that play a pivotal role in a plant's response to stress, with the potential to become targets for enhancing plant tolerance to the corresponding stressors. To achieve this goal, the workflow analyzes gene expression data collected under both control and stress conditions, in a set of plant genotypes. In essence, the workflow implements network analysis to detect overlapping gene modules, providing insights into the complex regulatory networks involved in the stress response. Additionally, it integrates phenotypic data, which provides observable traits at a more macroscopic level in plants. These traits serve as a crucial resource for identifying genes that are likely to be integral in a plant's response to stress.

While initially designed for biological contexts, the CSI-OC workflow holds the potential for broader applications, transforming into a versatile analytical framework. In a more general context, the primary objective of the CSI-OC workflow is to identify pivotal components within a system's response to external stressors. Stressors, in this context, encompass any factors or stimuli that exert pressure or demands on a system, prompting an adaptive response. In a general application, the genotypes, for which gene expression is measured, correspond to diverse samples on which the system's components are evaluated. Furthermore, phenotypic data in a broader context can be viewed as observable traits or characteristics of these samples at a macro level, serving as a valuable resource for pinpointing central components in the system's adaptation to stressors.

To effectively implement the CSI-OC workflow in any context, certain input prerequisites must be met. These requirements encompass the following key elements: the system under consideration should consist of n components that can be systematically evaluated across a designated set of m samples. These m samples must be amenable to assessment under both a baseline, neutral condition referred to as "control" and a defined stress condition. Additionally, there should be a

set of p external traits that are measured across the same set of m samples, again under both control and stress conditions. By fulfilling these specific input criteria, the CSI-OC workflow can be applied across diverse contexts to gain insights into system responses under varying conditions.

In this chapter the CSI-OC workflow is applied to a scenario involving a supermarket chain. Here, the system encompasses a diverse range of products across numerous stores, with unit sales serving as the primary metric for evaluation. The stores themselves represent the samples and the workflow can be employed to assess the response of products to varying conditions. Weekdays, spanning from Monday to Thursday, can be considered the control condition, representing typical shopping days, while weekends, from Friday to Sunday, serve as the stress condition due to heightened customer traffic and potentially altered purchasing patterns. In this scenario, one of the external observable traits can be the number of transactions recorded by each store. The goal is to validate if the CSI-OC workflow can identify key products that demonstrate varying sales patterns under the control and stress conditions. This insight can be invaluable for inventory management, stock allocation, and marketing strategies, helping the supermarket chain optimize its operations to meet customer demands more effectively during different shopping periods.

The application of the CSI-OC workflow in non-biological domains, as exemplified in this chapter's exploration of a supermarket chain scenario, serves as a compelling motivation for its adoption. CSI-OC workflow offers a groundbreaking approach to understanding and managing complex systems' responses to external stressors. Traditional analyses in non-biological settings often fall short in providing a comprehensive view of how various components within a system coordinately react to changing conditions. The CSI-OC workflow fills this critical gap by combining gene expression data, or the equivalent in non-biological scenarios, with observable external traits to identify central elements driving a system's adaptation to stress. Additionally, this novel approach allows for the exploration of overlapping relationships, which is essential for capturing the multifaceted dynamics of real-world networks. The result is a wealth of insights that can guide decision-making, optimize processes, and uncover hidden patterns in fields as diverse as economics, logistics, and social sciences.

The chapter's structure has been designed to facilitate a comprehensive exploration of the CSI-OC workflow's adaptation and application, transitioning from its biological origins to a non-biological context. The chapter begins with an introduction to the data's source in Section 7.1, laying the groundwork for a case study centered on a supermarket chain. Section 7.2 focuses on

the preliminary statistical analysis, underscoring the significance of data validation for workflow suitability. Section 7.3 delves into a detailed summary of the selected products identified through the workflow, discussing their organization into modules, relevance to system responses, and substantial overlap within these modules. In Section 7.4, the focus shifts to the workflow's unique feature, identifying overlapping modules and emphasizing the adaptability of products. Section 7.5 highlights computational validation results, demonstrating the reliability of the workflow's outputs. The knowledge-based evaluation of selected products, their product families, and the implications for inventory management is covered in Section 7.6. The chapter concludes with Section 7.7, reflecting on the workflow's successful adaptation, utility in analyzing supermarket data, and its potential for diverse non-biological applications.

7.1 Origin of the Data

The data employed in this chapter is publicly available on the Kaggle platform and correspond to sales records from "Corporación Favorita" (Corporación Favorita, 2017). This Ecuador-based grocery retailer operates an extensive network of hundreds of supermarkets, offering a wide array of over $200\,000$ distinct products. In January 2018, "Corporación Favorita" challenged the Kaggle community, granting access to the dataset with the primary objective of enhancing product sales forecasting accuracy. In our context, we repurpose this data for an alternative purpose. To apply the CSI-OC workflow, we selectively extract relevant data to identify key products whose interactions are pertinent to the differential sales behavior during weekends. The dataset we utilize consists of unit sales records for $n_0 = 4\,036$ products, across $m = 54$ stores. Additionally, we incorporate to our analysis data related to the average daily transaction counts for these same 54 stores.

It is worth noting that the products are labeled with unique ID numbers, making it impossible to determine the specific product identity. However, the original dataset includes information on the product family, encompassing a total of 33 distinct categories, including beverages, cleaning products, and meats. These categories, to some extent analogous to the functional annotations of genes, will facilitate knowledge-based analysis of the sets of interest. It serves as a proxy for functional annotations, similar to how it works with genes in biological contexts.

7.2 Preliminary Statistical Analysis

It is crucial to conduct statistical tests on transaction data, which is analogous to phenotypic data in the biological context, before applying the workflow. Its importance is underscored by insights derived from the simulations with synthetic data in the Chapter 6. The simulations revealed that the performance of the CSI-OC workflow is significantly influenced by the characteristics of the external traits associated with the samples. Scenarios lacking a discernible difference between control and stress distributions in the external traits resulted in performance similar to random selection of the key components. Therefore, conducting statistical tests on transaction data serves as a critical preparatory step to assess the presence of distinct control-stress patterns in the data.

Figure 7.1 presents a violin plot illustrating the distribution of transaction data by store under both control and stress conditions. Visually, the data appear to display the desired differences in distribution. Nevertheless, this behavior should be confirmed through the relevant statistical tests.

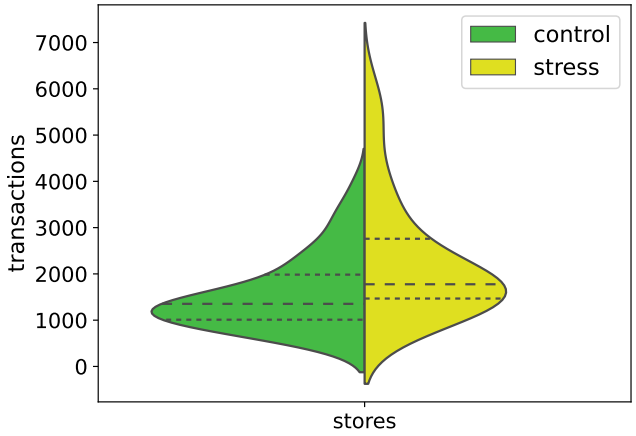

Figure 7.1: Violin plot illustrating the distribution of transaction data across stores under control and stress conditions

The difference in standard deviations was validated using Levene's test, resulting in a significant difference (p-value < 0.05) between the standard deviations of the control and stress conditions in the transaction data. Subsequently, a t-test, also confirmed a significant difference (p-value < 0.05) between the means of the control and stress transaction data. Ensuring that the transaction data exhibits these patterns enhances the likelihood of obtaining meaningful and actionable results when applying the CSI-OC workflow.

7.3 Summary of Selected Products

In this supermarket chain case study, the CSI-OC workflow was implemented using the Pearson-based network construction method. This choice was made prior to considering the Lasso-based method due to the ample sample size of 54 stores and a manageable number of products. As a result, the workflow successfully pinpointed 51 products that exhibit relevance in response to weekend days. These products are organized across 17 modules, each containing between 3 to 7 products. Remarkably, out of the selected genes, 42 (82%) demonstrate overlap within the larger set of 507 identified modules. This finding underscores the meaningful contribution of these products to the system's response to varying conditions and external factors.

7.4 Overlapping Products Characterization

The CSI-OC workflow offers a distinctive feature in the form of identifying overlapping modules. These modules are identified within the network constructed using the system's components. In the specific context of a supermarket chain, the network's nodes represent individual products and their connections are established based on correlations in their sales patterns. These connections enable the grouping of products into various modules. Each module can be thought of as a collection of complementary products (Avgeropoulos et al., 2015), meaning they are typically sold separately, but are frequently used together. In this context, products that overlap between multiple modules can be considered highly versatile. Their utility often depends on the presence of accompanying products, highlighting their adaptability and flexibility in meeting diverse customer needs.

After applying the CSI-OC workflow to the "Corporación Favorita" dataset, a total of 507 modules emerged from the 976 products remaining within the Pearson-based constructed network. Among all the products in the network, 612 were assigned to at least one module, and notably, 63% of them exhibited overlap. To visualize the enrichment of overlapping products and their associations with product families, Figure 7.2 illustrates a color bar graph. In this figure, product families with substantial representation within the overlapping gene set are depicted. The intensity of each bar reflects its significance as determined by an exact Fisher test, considering the products within the network as the background set. Additionally, the black and white plot within the figure provides an upsetplot showcasing the overlapping genes and the top 10 product families that demonstrate

significant representation within this set.

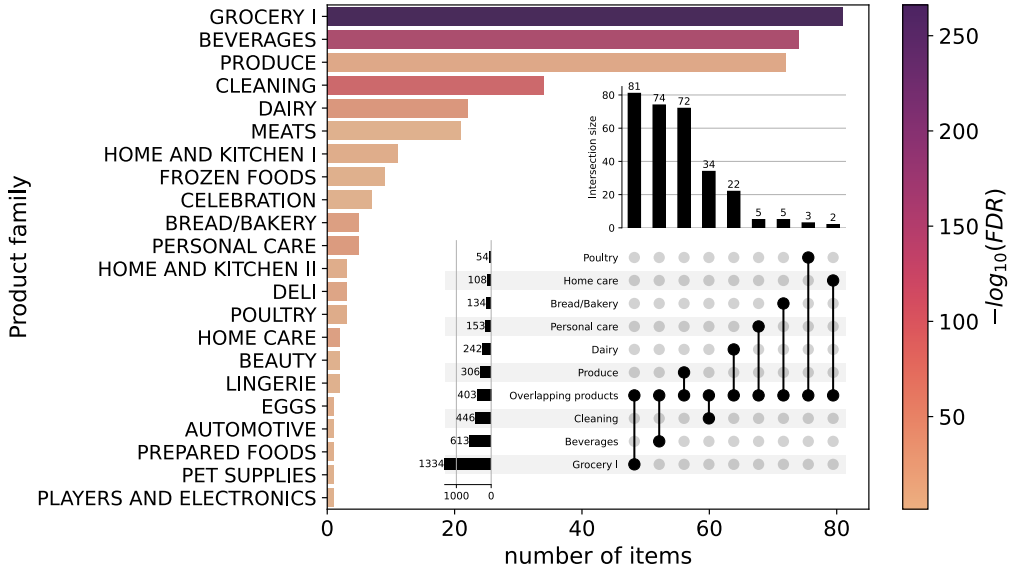

Figure 7.2: Enrichment of overlapping products with product families. The color bar plot illustrates the significance of product families within the overlapping products set. The upsetplot highlights the intersections between overlapping products and the top 10 product families with the most significant representation.

Understanding the significance of these product families can have several implications for inventory and stocking decisions. Highly versatile products within these families may experience fluctuating demand due to their associations with different modules. This knowledge can provide guidance for inventory management to avoid both overstocking and understocking. Additionally, the prevalence of these product families within the overlapping gene set supports market basket analysis. This type of analysis can uncover correlations between different product categories represented in the modules, which in turn assists in making informed decisions regarding product placement, pricing strategies, and promotional campaigns.

7.5 Computational Validation

The meaningfulness of the workflow results was assessed through a systematic comparison of products selected by the CSI-OC workflow and randomly selected products. This computational validation encompassed both classification and regression tasks. In the classification task, the products chosen by the CSI-OC workflow consistently outperformed randomly selected prod-

ucts in distinguishing between workdays (control) and weekends (stress). In the regression task, these selected products exhibited superior predictive capabilities for store transactions. Figure 7.3 presents the performance of both sets (selected and random products) over 100 trials for both tasks, as depicted in the boxplots. Notably, there is a substantial difference in the classification task, with the selected products displaying higher accuracy compared to the random ones. While the distinction in the regression task is not as pronounced, the selected products, on average, achieved a lower Mean Squared Error (MSE), indicating their superior performance over the random products.

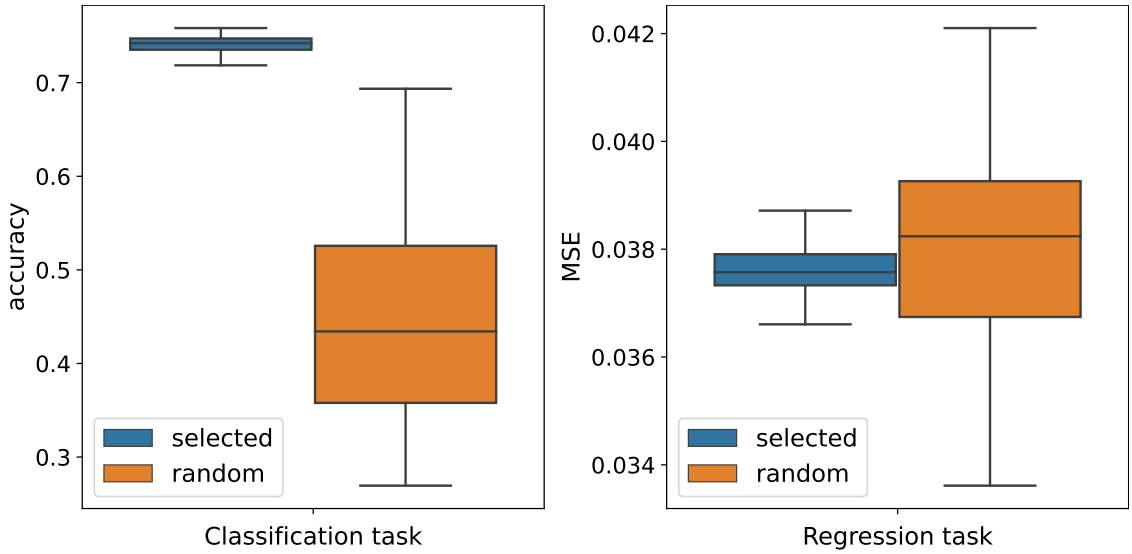

Figure 7.3: Computational validation of weekend-responsive products. (Left) Boxplots illustrating Accuracy for Classifying Workdays/Weekend Samples. (Right) Boxplots presenting the Mean Squared Error (MSE) for Predicting Store Transactions. Both panels show a comparison between CSI-OC selected products (in blue) and random products (in orange).

The Wilcoxon signed-rank test confirmed significant differences (p-value < 0.05) in the performance medians of the selected and random products for both tasks. These results strongly affirm the effectiveness of the CSI-OC workflow in identifying weekend-related products closely associated with store traits, such as store transactions.

7.6 Knowledge-based Evaluation

The 51 products identified by the CSI-OC workflow were analyzed in terms of their respective product families. Figure 7.5 visually represents all the product families associated with these selected

products using a color bar graph. The intensity of each bar corresponds to the significance of the product family, determined through an exact Fisher test. This analysis considers the products within the Pearson-based network as the background set. Additionally, within the same figure, a black and white plot is provided, showcasing the intersections between selected products and the products belonging to product families with statistically significant representation (FDR < 0.05). This graphical representation enhances the understanding of the relationships between selected products and their respective product families.

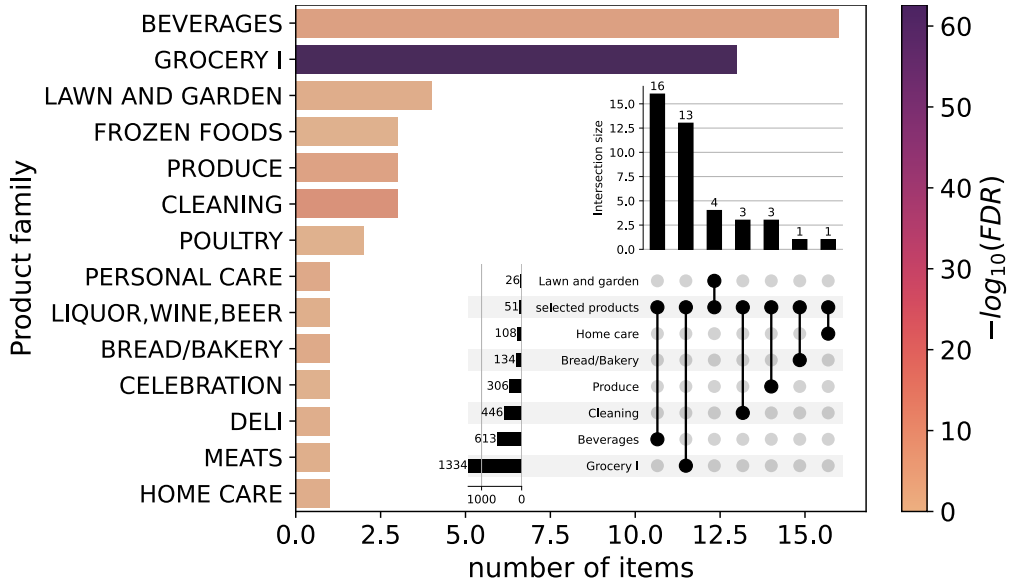

Figure 7.4: Selected products enrichment and upsetplot of the product families with significant representation.

Figure 7.5: Enrichment of CSI-OC selected products with product families. The color bar plot illustrates the significance of product families within the weekend-responsive products set selected by the CSI-OC workflow. The upsetplot displays intersections between selected products and the top 10 product families with the most significant representation.

The presence of “BEVERAGES” and “BREAD/BAKERY” suggests that customers tend to purchase these items during weekends, possibly for special occasions or leisure activities. Notably, the category of “BEVERAGES” may encompass alcoholic drinks, pointing to their association with weekend relaxation and social gatherings. Furthermore, the “GROCERY I” and “CLEANING” product families consist of essential household items, likely in demand during weekends, highlighting the link between weekend chores and the need for these products. Additionally, the presence of

"LAWN AND GARDEN" and "PRODUCE" suggests that customers are inclined toward outdoor improvements and engaging in DIY (Do It Yourself) activities during weekends. This could encompass tasks like tending to gardens, lawns, or carpentry projects. Meanwhile, the availability of "PERSONAL CARE" and "HOME CARE" products indicates that customers prioritize personal grooming and household maintenance during weekends.

These inferences highlight that the selected products belonging to these product families are likely to cater to the specific needs and preferences of customers during weekends, which can significantly impact store transactions. By understanding these relationships, supermarkets can optimize their product placement, promotions, and inventory management to better serve their customers during weekends and enhance their shopping experience.

7.7 Discussion

The adaptation of the CSI-OC workflow from its biological origins to a non-biological context has proven to be an analytical tool. Initially designed for identifying pivotal genes in a plant's response to stress, the workflow has found application in the analysis of data from a supermarket chain. Here, its primary aim was to identify pivotal products associated with differential sales patterns during weekends.

This successful application underscores the flexibility and broad utility of the CSI-OC workflow, which can be tailored to diverse non-biological contexts. By meeting specific input criteria, it offers valuable insights into how a system responds to external stressors. In this case, the supermarket chain scenario, where different products exhibited varying responses under weekend stress conditions, is just one of numerous potential applications across various domains.

The application of the CSI-OC workflow to the supermarket chain context has revealed interesting insights and patterns. The workflow's capacity to identify pivotal products, comprehend their relationships with various product families, and assess their performance under different conditions presents valuable opportunities for enhancing business operations. This understanding of how products are associated with customer behavior during weekends can enable supermarkets to optimize product placement, promotions, and inventory management, ultimately enriching the shopping experience for customers.

Additionally, the identification of products that overlap between multiple modules is another interesting finding. Such products showcase remarkable versatility, as their usefulness hinges

on the presence of complementary items. In a non-biological context, the ability to effectively identify versatile products holds the potential to guide inventory management, stock allocation, and marketing strategies. These insights enable the supermarket chain to meet customer demands more efficiently during various shopping periods.

Looking ahead, further enriching this analysis with information about the price and demand for each product could facilitate cross-price elasticity of demand analysis (Avgeropoulos et al., 2015; Capps Jr and Dharmasena, 2019). This analysis would help confirm whether products belonging to the same module truly exhibit complementary behavior. In economic terms, complementary products display a negative cross-elasticity of demand, meaning that as the price of one product rises, the demand for the complementary product falls.

In summary, the results underscore the adaptive potential of the CSI-OC workflow in non-biological contexts. Within the supermarket chain context, it presents a data-driven approach to optimizing operations, enhancing the customer experience, and improving decision-making efficiency. This chapter not only showcases the workflow's adaptability, but it also highlights the potential for its application across a diverse range of non-biological domains.

8. Conclusion and Future Work

8.1 Conclusion

This dissertation has primarily focused on the development of an in-silico workflow aimed at identifying stress-responsive genes while addressing critical challenges associated with capturing the real characteristics of the biological system. To address this objective, the Control-Stress Data Integration with Overlapping Clustering (CSI-OC) workflow was designed and presented in Chapter 3. The CSI-OC workflow serves as a comprehensive analytical framework that integrates both gene expression and phenotypic data from control and stress conditions. One of its primary functions is to capture differential gene expression information within a network, thereby facilitating a deeper understanding of the coordinated response to stress. This workflow encompasses a series of well-defined steps, including data pre-processing, the construction of differential co-expression networks, the detection of overlapping gene modules, and the selection of relevant modules based on their influence on specific phenotypic traits. Each step contributes to the overall goal of identifying stress-responsive genes and unraveling their intricate role within the biological system.

In the context of biology, CSI-OC has been employed to uncover stress-responsive genes in rice (Chapter 4) and sugarcane (Chapter 5), two essencial crop species known for their susceptibility to environmental stressors. Through the application of both Pearson-based and Lasso-based network construction methods, in conjunction with the Hierarchical Link Clustering (HLC) algorithm and Lasso selection method, the workflow unveiled genes intricately linked to specific stress conditions. Moreover, the identification of overlapping modules highlighted their significance as integral components within co-expression networks that govern stress responses, characterized by a substantial representation of Transcription Factors (TFs) or TF-related annotations. The computational validation of the workflow's performance underscored its effectiveness in identifying meaningful genes, outperforming randomly selected genes in tasks related to control/stress classification and phenotypic trait prediction. Knowledge-based evaluations validated the biological

relevance of the CSI-OC selected genes. The findings open up opportunities for the discovery of novel genes associated with stress responses, shedding light on the intricate mechanisms governing plant resilience to a variety of stressors. The understanding of this mechanisms can yield immediate benefits in the realm of agriculture, potentially leading to increased crop yields and sustainability. Furthermore, it may facilitate the development of stress-tolerant crop varieties, aligning with the global pursuit of food security and environmental resilience.

The validation of the CSI-OC workflow using synthetic data, presented in Chapter 6, served to reinforce its robustness and reliability. This synthetic data testing enabled controlled experiments, clearly demonstrating that meaningful results can be achieved, especially when scenarios exhibit significant differences between the normal parameters of control and stress phenotypic data. This comprehensive validation framework offered valuable insights into how the CSI-OC workflow behaves under varying conditions, providing guidance for its practical application to real-world datasets. Additionally, the development of the CoSynthEx model has provided an invaluable tool for generating synthetic gene expression data that is tailored to specific phenotypic traits and sample conditions.

Furthermore, the adaptability of the CSI-OC workflow extends beyond its original biological context. By applying the workflow to a non-biological domain, specifically to a supermarket chain setting, the workflow has demonstrated its versatility and broad applicability in Chapter 7. It identified pivotal products associated with differential sales patterns during weekends, which may facilitate data-driven decision-making in business operations. While the case study in the supermarket chain represents an initial exploration of the application in a non-biological context, further in-depth validation and research are needed to fully ascertain its potential and effectiveness in such diverse fields. Nevertheless, the adaptability of CSI-OC beyond the real of biology broadens its utility and opens doors to applications in various domains, having the potential to make a valuable contribution to the wider scientific community.

8.2 Future Work

One avenue for future work involves extending CSI-OC to incorporate a broader range of data types beyond gene expression and phenotypic data. While the current workflow effectively integrates these two crucial data sources, biological systems are complex, and integrating additional types of data can provide a more comprehensive understanding of the underlying mechanisms.

For instance, incorporating epigenetic data, such as DNA methylation and histone modification patterns, could shed light on how stress-responsive genes are regulated at the epigenetic level. By adding metabolomic data, it is possible to explore the metabolic pathways and small molecule interactions involved in stress responses. Furthermore, incorporating proteomic data could help identify key proteins that mediate stress responses. The integration of multi-omics data would enable a more holistic view of the molecular processes involved in stress adaptation. This future work would involve extending the CSI-OC workflow to handle diverse data types effectively, developing methodologies for data integration, and providing a more comprehensive analysis of the biological system.

Expanding the application of the CSI-OC workflow to organisms beyond plants and a broader spectrum of stress conditions is another promising avenue for future research. While this thesis focused on stress responses in plants, particularly rice and sugarcane, stress responses are pervasive in various organisms, including humans. Diseases such as cancer can be viewed as a form of stress on human cells, prompting changes in gene expression patterns and signaling pathways. Adapting the CSI-OC workflow to analyze stress-responsive genes in humans and other organisms could have profound implications for understanding the molecular basis of diseases and identifying potential therapeutic targets. Moreover, exploring different types of environmental stressors, such as heat stress or pathogen attacks, in various organisms would enhance the applicability and generalizability of the workflow. Future work would involve customizing the workflow to accommodate different organisms and types of stress, while also considering the unique characteristics of the data generated in each context.

While this research provided a preliminary exploration of the CSI-OC workflow's application in a non-biological context, there is ample room for more in-depth research in this area. Specifically, applying the workflow to non-biological contexts beyond the supermarket chain case study would expand its utility. Other non-biological domains, such as finance, social sciences, and engineering, can benefit from the analytical capabilities of the CSI-OC workflow. Future work involves conducting detailed case studies and extensive validations in these non-biological contexts to gauge the performance, robustness, and adaptability of the workflow more comprehensively. This entails developing domain-specific adaptations and understanding how the workflow can effectively address the unique challenges and questions posed by each field. Ultimately, exploring the CSI-OC workflow's potential in diverse non-biological domains would contribute to its broader impact on data-driven decision-making and knowledge discovery outside the realm of biology.

Nomenclature

Gene names

cc_00001481 Sugarcane gene. Gene related to transcription regulation and responses to stress and stimuli (Surya Krishna et al., 2023).

cc_00002124 Sugarcane gene. Encodes for a GNAT (N-acetyltransferase GNAT)-related protein, associated with drought tolerance in sorghum (Devnarain et al., 2019; Das et al., 2021).

cc_00006300 Sugarcane gene. Knottin motif, associated with enhanced immunity in plants and defense against biotic stress (Tavormina et al., 2015; Zhang et al., 2023).

cc_00008263 Sugarcane gene. Encodes for an Argonaute protein (AGO1), implicated in dehydration and RNA silencing (Cui et al., 2020; Tang et al., 2023).

cc_00027902 Sugarcane gene. MLO-like protein, potentially involved in stress response (Contiliani et al., 2023).

LOC_Os04g12530 Rice gene that reported as an up-regulated gene in rice plants tolerant to salt stress (Razzaque et al., 2019).

LOC_Os04g35010 Rice gene that encodes a protein from the bHLH domain, which has been shown to be part of multiple cellular processes, including salt stress signaling pathways (Qian et al., 2021).

LOC_Os12g10280 Rice gene that encodes an aquaporin nodulin 26-like intrinsic membrane (NIP3;5) protein (Hsieh et al., 2018); it has been shown that NIPs play an important role in salt stress responses and in maintaining plant water balance (Kapilan et al., 2018).

LOC_Os12g27254 Rice gene that encodes spermidine hydroxycinnamoyltransferase 2 (SHT2) protein. This protein contributes to the natural variation of spermidine-based phenolamides

in rice cultivars, whose activity promotes tolerance to saline stress (Bassard et al., 2010; Roychoudhury et al., 2011; Gupta et al., 2013; Peng et al., 2019).

LOC_Os12g37260 Rice gene that encodes Lipoxygenase protein, which is considered to correlate directly with salt tolerance in rice (Mittova et al., 2002; Mostofa et al., 2015; Hou et al., 2015).

Math variables

α_z Weights representing module importance in Lasso regression.

β Value to power the correlation of the genes to produce a higher similarity with a scale-free network.

β_i Weights representing gene importance in Lasso regression.

Δ Synthetic scenario parameters.

ℓ_1 L1 norm corresponding to the sum of the absolute vector values.

$\hat{A}$ Final Differential Co-expression Network (DCN) after removing isolated nodes.

$\hat{E}_c$ Synthetic gene expression data under control conditions.

$\hat{E}_s$ Synthetic gene expression data under stress conditions.

$\hat{P}_s$ Synthetic phenotypic data under stress conditions.

λ Regularization parameter in Lasso regression.

$\mathcal{D}$ Discriminator neural network in Generative Adversarial Networks (GANs).

$\mathcal{G}$ Generator neural network in Generative Adversarial Networks (GANs).

μ Mean of stress phenotypic data.

μ_z Mean values for each phenotypic trait.

σ Standard deviation of stress phenotypic data.

σ_z Standard deviation values for each phenotypic trait.

A Unweighted and symmetric adjacency matrix.

B Matrix of coefficients in Lasso-based Differential Co-expression Network (DCN) construction.

c Number of modules detected in the Differential Co-expression Network (DCN).

D Partition density of links in a hierarchical clustering dendrogram.

D_0 Matrix combining control and stress expression data before normalization.

D_1 Normalized expression data matrix.

D_2 Normalized expression matrix after averaging biological and gene replicates.

D_3 Normalized expression matrix after removing genes with low variance or expression.

E Set of unordered pair of vertices within a graph.

E_ℓ Log Fold Change (LFC) matrix for expression data.

E_c Gene expression data under control conditions.

E_s Gene expression data under stress conditions.

E_{real} Real gene expression data.

E_{synth} Synthetic gene expression data.

F Affiliation matrix indicating the membership of genes within modules.

I Set of the union of the genes that belong to the final selected modules, also called stress-responsive genes.

I_{rand} Set of randomly selected genes.

I_{synth} Set of selected genes by the CSI-OC workflow.

M Matrix of eigengenes representing modules.

m_{ec} Number of control samples.

m_{es} Number of stress samples.

N Number of nodes in the Differential Co-expression Network (DCN).

n_0 Original number of genes.

n_1 Reduced number of unique genes after averaging replicates.

n_2 Number of genes after removing low-variance or low-expression ones.

n_3 Final gene pool size after filtering low variance in the differential expression patterns.

n_4 Reduced gene pool size after constructing Differential Co-expression Network (DCN).

P Phenotypic trait data.

p Number of phenotypic traits.

P_c Phenotypic traits measured under control conditions.

P_s Phenotypic traits measured under stress conditions.

Q Sample condition information.

R Random noise data.

r Number of replicates

S Similarity matrix in Pearson-based Differential Co-expression Network (DCN) construction.

V Set of vertices within a graph.

W Weighted and scale-free adjacency matrix in Pearson-based Differential Co-expression Network (DCN) construction.

X Log Fold Change (LFC) matrix for expression data after removing genes with low variance.

Y Log Fold Change (LFC) matrix for phenotypic data.

z Index representing a specific phenotypic trait.

Bibliography

Abbassi-Daloii, T., Kan, H. E., Raz, V., et al. (2020). Recommendations for the analysis of gene expression data to identify intrinsic differences between similar tissues. *Genomics*, 112(5):3157–3165.

Acosta-Motos, J. R., Ortuño, M. F., Bernal-Vicente, A., Diaz-Vivancos, P., Sanchez-Blanco, M. J., and Hernandez, J. A. (2017). Plant responses to salt stress: adaptive mechanisms. *Agronomy*, 7(1):18.

Aggarwal, A., Mittal, M., and Battineni, G. (2021). Generative adversarial network: An overview of theory and applications. *International Journal of Information Management Data Insights*, 1(1):100004.

Ahn, Y.-Y., Bagrow, J. P., and Lehmann, S. (2010). Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764.

Altenhoff, A. M., Glover, N. M., Train, C.-M., Kaleb, K., Warwick Vesztrocy, A., Dylus, D., de Farias, T. M., Zile, K., Stevenson, C., Long, J., Redestig, H., Gonnet, G. H., and Dessimoz, C. (2018). The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic acids research*, 46(D1):D477–D485.

Aoki, K., Ogata, Y., and Shibata, D. (2007). Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant & cell physiology*, 48(3):381–390.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

Asocaña (2023). Informe anual de asocaña con aspectos generales del sector agroindustrial de la caña 2022-2023 y anexo estadístico. Technical report, Asocaña.

Avgeropoulos, S., Sammut-Bonnici, T., and McGee, J. (2015). Complementary products. *Wiley Encyclopedia of Management*, pages 1–2.

Bai, Y. (2022). Relu-function and derived function review. In *SHS Web of Conferences*, volume 144, page 02006. EDP Sciences.

Bassard, J.-E., Ullmann, P., Bernier, F., and Werck-Reichhart, D. (2010). Phenolamides: bridging polyamines to the phenolic metabolism. *Phytochemistry*, 71(16):1808–1824.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Biau, G. and Scornet, E. (2016). A random forest guided tour. *Test*, 25:197–227.

Björck, Å. (1990). Least squares methods. In *Handbook of Numerical Analysis*, volume 1, pages 465–652. Elsevier.

Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., Richardson, L., Salazar, G. A., Williams, L., Bork, P., Bridge, A., Gough, J., Haft, D. H., Letunic, I., Marchler-Bauer, A., Mi, H., Natale, D. A., Necci, M., Orengo, C. A., Pandurangan, A. P., Rivoire, C., Sigrist, C. J. A., Sillitoe, I., Thanki, N., Thomas, P. D., Tosatto, S. C. E., Wu, C. H., Bateman, A., and Finn, R. D. (2021). The InterPro protein families and domains database: 20 years on. *Nucleic acids research*, 49(D1):D344–D354.

Bujang, M. A. and Baharum, N. (2016). Sample size guideline for correlation analysis. *World journal of social science research*, 3(1):37.

Campbell, M. T., Bandillo, N., Al Shiblawi, F. R. A., Sharma, S., Liu, K., Du, Q., Schmitz, A. J., Zhang, C., Véry, A.-A., Lorenz, A. J., et al. (2017). Allelic variants of OsHKT1; 1 underlie the divergence between indica and japonica subspecies of rice (*Oryza sativa*) for root sodium content. *PLoS Genetics*, 13(6):e1006823.

Capps Jr, O. and Dharmasena, S. (2019). Enhancing the teaching of product substitutes/complements: A pedagogical note on diversion ratios. *Applied Economics Teaching Resources (AETR)*, 1(2226-2019-3953):32–45.

Chang, J., Cheong, B. E., Natera, S., and Roessner, U. (2019). Morphological and metabolic responses to salt stress of rice (*Oryza sativa L.*) cultivars which differ in salinity tolerance. *Plant Physiology and Biochemistry*, 144:427–435.

Chen, C., Norton, G. J., and Price, A. H. (2020). Genome-wide association mapping for salt tolerance of rice seedlings grown in hydroponic and soil systems using the bengal and assam aus panel. *Frontiers in plant science*, 11:1633.

Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F., and Mahmood, F. (2021). Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497.

Choi, J. K., Yu, U., Yoo, O. J., and Kim, S. (2005). Differential coexpression analysis using microarray data and its application to human cancer. *Bioinformatics*, 21(24):4348–4355.

Clough, E. and Barrett, T. (2016). The gene expression omnibus database. In *Statistical Genomics*, volume 1418 of *Methods in Molecular Biology*, pages 93–110. Humana Press, New York, NY.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., et al. (2016). A survey of best practices for rna-seq data analysis. *Genome biology*, 17(1):1–19.

Consortium, T. G. O. (2021). The gene ontology resource: enriching a gold mine. *Nucleic acids research*, 49(D1):D325–D334.

Contiliani, D. F., de Oliveira Nebó, J. F. C., Ribeiro, R. V., Andrade, L. M., Peixoto Júnior, R. F., Lembke, C. G., Machado, R. S., Silva, D. N., Belloti, M., de Souza, G. M., Perecin, D., Pereira, T. C., de Matos Pires, R. C., Fontoura, P. R., Landell, M. G. A., Figueira, A., and Creste, S. (2022). Leaf transcriptome profiling of contrasting sugarcane genotypes for drought tolerance under field conditions. *Scientific reports*, 12(1):9153.

Contiliani, D. F., Nebó, J. F. C. d. O., Ribeiro, R. V., Landell, M. G. d. A., Pereira, T. C., Ming, R., Figueira, A., and Creste, S. (2023). Drought-triggered leaf transcriptional responses disclose

key molecular pathways underlying leaf water use efficiency in sugarcane (spp.). *Frontiers in plant science*, 14:1182461.

Corporación Favorita, Julia Elliott, M. M. (2017). Corporación favorita grocery sales forecasting.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65.

Cui, D.-L., Meng, J.-Y., Ren, X.-Y., Yue, J.-J., Fu, H.-Y., Huang, M.-T., Zhang, Q.-Q., and Gao, S.-J. (2020). Genome-wide identification and characterization of DCL, AGO and RDR gene families in Saccharum spontaneum. *Scientific reports*, 10(1):13202.

D'amico, S., Dall'Olio, D., Sala, C., Dall'Olio, L., Sauta, E., Zampini, M., Asti, G., Lanino, L., Maggioni, G., Campagna, A., et al. (2023). Synthetic data generation by artificial intelligence to accelerate research and precision medicine in hematology. *JCO Clinical Cancer Informatics*, 7:e2300021.

Das, R., Mukherjee, A., Basak, S., and Kundu, P. (2021). Plant miRNA responses under temperature stress. *Plant Gene*, 28:100317.

Devnarain, N., Crampton, B. G., Olivier, N., van der Westhuyzen, C., Becker, J. V. W., and O'Kennedy, M. M. (2019). Transcriptomic analysis of a Sorghum bicolor landrace identifies a role for beta-alanine betaine biosynthesis in drought tolerance. *South African journal of botany: official journal of the South African Association of Botanists = Suid-Afrikaanse tydskrif vir plantkunde: amptelike tydskrif van die Suid-Afrikaanse Genootskap van Plantkundiges*, 127:244–255.

DNDA (2023). Dirección Nacional de Derecho de Autor. `www.derechodeautor.gov.co`. [Accessed 29-10-2023].

Du, Q., Campbell, M., Yu, H., Liu, K., Walia, H., Zhang, Q., and Zhang, C. (2019). Network-based feature selection reveals substructures of gene modules responding to salt stress in rice. *Plant Direct*, 3(8):e00154.

Du, Y., Zhao, Q., Chen, L., Yao, X., Zhang, W., Zhang, B., and Xie, F. (2020). Effect of drought stress on sugar metabolism in leaves and roots of soybean seedlings. *Plant physiology and biochemistry: PPB / Societe francaise de physiologie vegetale*, 146:1–12.

Eizenga, G. C., Ali, M. L., Bryant, R. J., Yeater, K. M., McClung, A. M., and McCouch, S. R. (2014). Registration of the rice diversity panel 1 for genomewide association studies. *Journal of Plant Registrations*, 8(1):109–116.

Fionda, V. (2019). Networks in biology. In Ranganathan, S., Gribskov, M., Nakai, K., and Schönbach, C., editors, *Encyclopedia of Bioinformatics and Computational Biology*, volume 1, pages 915–921. Academic Press, Oxford.

Fonti, V. and Belitser, E. (2017). Feature selection using lasso. *VU Amsterdam research paper in business analytics*, 30:1–25.

Fukushima, A. (2013). Diffcorr: an r package to analyze and visualize differential correlations in biological networks. *Gene*, 518(1):209–214.

Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A.-M., Cruciat, C.-M., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147.

Ghoname, A. A., AbdelMotlb, N. A., Abdel-Al, F. S., Abu El-Azm, N. A., Abd Elhady, S. A., Merah, O., and Abdelhamid, M. T. (2023). Brassinosteroids or proline can alleviate yield inhibition under salt stress via modulating physio-biochemical activities and antioxidant systems in snap bean. *The Journal of Horticultural Science and Biotechnology*, 98(4):526–539.

Gui, J., Sun, Z., Wen, Y., Tao, D., and Ye, J. (2021). A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE transactions on knowledge and data engineering*, 35(4):3313–3332.

Gupta, A., Rico-Medina, A., and Caño-Delgado, A. I. (2020). The physiology of plant responses to drought. *Science*, 368(6488):266–269.

Gupta, K., Dey, A., and Gupta, B. (2013). Plant polyamines in abiotic stress responses. *Acta Physiologiae Plantarum*, 35(7):2015–2036.

Hannan, A., Hoque, M. N., Hassan, L., and Robin, A. H. K. (2020). Adaptive mechanisms of root system of rice for withstanding osmotic stress. *Recent advances in rice research*.

Holmes Finch, W. and Hernandez Finch, M. E. (2019). Regularization Methods for Fitting Linear

Models with Small Sample Sizes: Fitting the Lasso Estimator using R. *Practical Assessment, Research, and Evaluation*, 21(1):7.

Hou, Y., Hu, J., Zhou, L., Liu, L., Chen, K., and Yang, X. (2021). Integrative analysis of methylation and copy number variations of prostate adenocarcinoma based on weighted gene co-expression network analysis. *Frontiers in Oncology*, 11:584.

Hou, Y., Meng, K., Han, Y., Ban, Q., Wang, B., Suo, J., Lv, J., and Rao, J. (2015). The persimmon 9-lipoxygenase gene DkLOX3 plays positive roles in both promoting senescence and enhancing tolerance to abiotic stress. *Frontiers in Plant Science*, 6:1073.

Hsiao, T. C. and Xu, L.-K. (2000). Sensitivity of growth of roots versus leaves to water stress: biophysical analysis and relation to water transport. *Journal of experimental botany*, 51(350):1595–1616.

Hsieh, P.-H., Kan, C.-C., Wu, H.-Y., Yang, H.-C., and Hsieh, M.-H. (2018). Early molecular events associated with nitrogen deficiency in rice seedling roots. *Scientific reports*, 8(1):1–23.

Hussain, K., Majeed, A., Nawaz, K., Nisar, M. F., et al. (2009). Effect of different levels of salinity on growth and ion contents of black seeds (nigella sativa l.). *Current Research Journal of Biological Sciences*, 1(3):135–138.

Hussain, M., Farooq, S., Hasan, W., Ul-Allah, S., Tanveer, M., Farooq, M., and Nawaz, A. (2018). Drought stress in sunflower: Physiological effects and its management through breeding and agronomic alternatives. *Agricultural water management*, 201:152–166.

Hussain, M., Farooq, S., Jabran, K., Ijaz, M., Sattar, A., and Hassan, W. (2016). Wheat Sown with Narrow Spacing Results in Higher Yield and Water Use Efficiency under Deficit Supplemental Irrigation at the Vegetative and Reproductive Stage. *Agronomy*, 6(2):22.

James, G., Witten, D., Hastie, T., Tibshirani, R., James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). Linear model selection and regularization. *An introduction to statistical learning: with applications in R*, pages 225–288.

James, G., Witten, D., Hastie, T., Tibshirani, R., and Taylor, J. (2023). Linear regression. In *An Introduction to Statistical Learning: With Applications in Python*, pages 69–134. Springer.

Jaramillo-Botero, A., Colorado, J., Quimbaya, M., Rebolledo, M. C., Lorieux, M., Ghneim-Herrera, T., Arango, C. A., Tobón, L. E., Finke, J., Rocha, C., et al. (2022). The ómicas alliance, an international research program on multi-omics for crop breeding optimization. *Frontiers in plant science*, 13:992663.

Jin, D., Gabrys, B., and Dang, J. (2015). Combined node and link partitions method for finding overlapping communities in complex networks. *Scientific reports*, 5(1):1–8.

Jin, J., Tian, F., Yang, D.-C., Meng, Y.-Q., Kong, L., Luo, J., and Gao, G. (2016). Planttfdb 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic acids research*, page gkw982.

Kaloev, M. and Krastev, G. (2021). Comparative analysis of activation functions used in the hidden layers of deep neural networks. In *2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, pages 1–5. IEEE.

Kang, J., Peng, Y., and Xu, W. (2022). Crop Root Responses to Drought Stress: Molecular Mechanisms, Nutrient Regulations, and Interactions with Microorganisms in the Rhizosphere. *International journal of molecular sciences*, 23(16).

Kanonidis, E. I., Roy, M. M., Deighton, R. F., and Le Bihan, T. (2016). Protein co-expression analysis as a strategy to complement a standard quantitative proteomics approach: Case of a glioblastoma multiforme study. *PLoS One*, 11(8):e0161828.

Kapilan, R., Vaziri, M., and Zwiazek, J. J. (2018). Regulation of aquaporins in plants under stress. *Biological research*, 51(1):1–11.

Karami, S., Shiran, B., Ravash, R., and Fallahi, H. (2023). A comprehensive analysis of transcriptomic data for comparison of plants with different photosynthetic pathways in response to drought stress. *PloS one*, 18(6):e0287761.

Kaura, V., Malhotra, P. K., Mittal, A., Sanghera, G. S., Kaur, N., Bhardwaj, R. D., Cheema, R. S., and Kaur, G. (2022). Physiological, biochemical, and gene expression responses of sugarcane under cold, drought and salt stresses. *Journal of plant growth regulation*.

Khan, S.-A., Li, M.-Z., Wang, S.-M., and Yin, H.-J. (2018). Revisiting the role of plant transcription factors in the battle against abiotic stress. *International Journal of Molecular Sciences*, 19(6):1634.

Korthauer, K., Kimes, P. K., Duvallet, C., Reyes, A., Subramanian, A., Teng, M., Shukla, C., Alm, E. J., and Hicks, S. C. (2019). A practical guide to methods controlling false discoveries in computational biology. *Genome biology*, 20(1):1–21.

Kour, D., Khan, S. S., Kaur, T., Kour, H., Singh, G., Yadav, A., and Yadav, A. N. (2022). Drought adaptive microbes as bioinoculants for the horticultural crops. *Heliyon*, 8(5).

Kul, R., Ekinci, M., Turan, M., Ors, S., and Yildirim, E. (2021). How Abiotic Stress Conditions Affects Plant Roots. In *Plant Roots*. IntechOpen.

Lacambra, C., Molloy, D., Lacambra, J., Leroux, I., Klossner, L., Talari, M., Cabrera, M. M., Persson, S., Downing, T. E., Downing, E., et al. (2020). *Factsheet Resilience Solutions for the Rice Sector in Colombia.* IADB: Inter-American Development Bank.

Lahiri, A., Rastogi, K., Datta, A., and Septiningsih, E. M. (2021). Bayesian network analysis of lysine biosynthesis pathway in rice. *Inventions*, 6(2):37.

Lakshmanan, P. and Robinson, N. (2013). Stress Physiology: Abiotic Stresses. In *Sugarcane: Physiology, Biochemistry, and Functional Biology*, pages 411–434. John Wiley & Sons Ltd, Chichester, UK.

Langfelder, P. and Horvath, S. (2008). Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):1–13.

Li, A.-M., Liao, F., Wang, M., Chen, Z.-L., Qin, C.-X., Huang, R.-Q., Verma, K. K., Li, Y.-R., Que, Y.-X., Pan, Y.-Q., and Huang, D.-L. (2023). Transcriptomic and Proteomic Landscape of Sugarcane Response to Biotic and Abiotic Stressors. *International journal of molecular sciences*, 24(10).

Li, S., Liu, X., Liu, T., Meng, X., Yin, X., Fang, C., Huang, D., Cao, Y., Weng, H., Zeng, X., et al. (2017). Identification of biomarkers correlated with the tnm staging and overall survival of patients with bladder cancer. *Frontiers in physiology*, 8:947.

Li, W., Fu, L., Geng, Z., Zhao, X., Liu, Q., and Jiang, X. (2021). Physiological Characteristic Changes and Full-Length Transcriptome of Rose (Rosa chinensis) Roots and Leaves in Response to Drought Stress. *Plant & cell physiology*, 61(12):2153–2166.

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930.

Liesecke, F., De Craene, J.-O., Besseau, S., Courdavault, V., Clastre, M., Vergès, V., Papon, N., Giglioli-Guivarc'h, N., Glévarec, G., Pichon, O., and Dugé de Bernonville, T. (2019). Improved gene co-expression network quality through expression dataset down-sampling and network aggregation. *Scientific reports*, 9(1):14431.

Liu, B.-H. (2018). Differential coexpression network analysis for gene expression data. In *Computational Systems Biology*, pages 155–165. Springer.

Liu, C., Chen, K., Zhao, X., Wang, X., Shen, C., Zhu, Y., Dai, M., Qiu, X., Yang, R., Xing, D., et al. (2019). Identification of genes for salt tolerance and yield-related traits in rice plants grown hydroponically and under saline field conditions by genome-wide association study. *Rice*, 12(1):1–13.

Lopes, M. S., Araus, J. L., van Heerden, P. D. R., and Foyer, C. H. (2011). Enhancing drought tolerance in C(4) crops. *Journal of experimental botany*, 62(9):3135–3153.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1–21.

Maier, R., Zimmer, R., and Küffner, R. (2013). A turing test for artificial expression data. *Bioinformatics*, 29(20):2603–2609.

Mateos, G., Segarra, S., Marques, A. G., and Ribeiro, A. (2019). Connecting the dots: Identifying network structure via graph signal processing. *IEEE Signal Processing Magazine*, 36(3):16–43.

McGowan, M. T., Zhang, Z., and Ficklin, S. P. (2021). Chromosomal characteristics of salt stress heritable gene expression in the rice genome. *BMC genomic data*, 22(1):1–13.

Mesterházy, Á., Oláh, J., and Popp, J. (2020). Losses in the grain supply chain: Causes and solutions. *Sustainability*, 12(6):2342.

Min, X., Lin, X., Ndayambaza, B., Wang, Y., and Liu, W. (2020). Coordinated mechanisms of leaves and roots in response to drought stress underlying full-length transcriptome profiling in Vicia sativa L. *BMC plant biology*, 20(1):1–21.

Mittova, V., Tal, M., Volokita, M., and Guy, M. (2002). Salt stress induces up-regulation of an efficient chloroplast antioxidant system in the salt-tolerant wild tomato species *Lycopersicon pennellii* but not in the cultivated species. *Physiologia Plantarum*, 115(3):393–400.

Mostofa, M. G., Hossain, M. A., and Fujita, M. (2015). Trehalose pretreatment induces salt tolerance in rice (oryza sativa l.) seedlings: oxidative damage and co-induction of antioxidant defense and glyoxalase systems. *Protoplasma*, 252(2):461–475.

Munns, R. (2005). Genes and salt tolerance: bringing them together. *New phytologist*, 167(3):645–663.

Muthukrishnan, R. and Rohini, R. (2016). Lasso: A feature selection technique in predictive modeling for machine learning. In *2016 IEEE international conference on advances in computer applications (ICACA)*, pages 18–20. IEEE.

Muthuramalingam, P., Jeyasri, R., Selvaraj, A., Kalaiyarasi, D., Aruni, W., Pandian, S. T. K., and Ramesh, M. (2021). Global transcriptome analysis of novel stress associated protein (SAP) genes expression dynamism of combined abiotic stresses in Oryza sativa (L.). *Journal of biomolecular structure & dynamics*, 39(6):2106–2117.

Nguyen, D. Q., Nguyen, N. L., Nguyen, V. T., Tran, T. H. G., Nguyen, T. H., Nguyen, T. K. L., and Nguyen, H. H. (2023). Comparative analysis of microrna expression profiles in shoot and root tissues of contrasting rice cultivars (oryza sativa l.) with different salt stress tolerance. *Plos one*, 18(5):e0286140.

Oeckinghaus, A., Hayden, M. S., and Ghosh, S. (2011). Crosstalk in nf-κb signaling pathways. *Nature immunology*, 12(8):695.

Önay, E. and Demirbas, S. (2023). Strigolactones affect growth parameters and some antioxidant enzyme activities in wheat (triticum aestivum l.) under salt stress. *International Journal of Innovative Approaches in Agricultural Research*.

Ovens, K., Eames, B. F., and McQuillan, I. (2020). The impact of sample size and tissue type on the reproducibility of gene co-expression networks. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, number Article 12 in BCB '20, pages 1–10, New York, NY, USA. Association for Computing Machinery.

Palla, G., Derényi, I., Farkas, I., and Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818.

Pan, W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 18(4):546–554.

Pearson, K. (1920). Notes on the history of correlation. *Biometrika*, 13(1):25–45.

Peng, H., Meyer, R. S., Yang, T., Whitaker, B. D., Trouth, F., Shangguan, L., Huang, J., Litt, A., Little, D. P., Ke, H., et al. (2019). A novel hydroxycinnamoyl transferase for synthesis of hydroxycinnamoyl spermine conjugates in plants. *BMC Plant Biology*, 19(1):1–13.

Qian, Y., Zhang, T., Yu, Y., Gou, L., Yang, J., Xu, J., and Pi, E. (2021). Regulatory mechanisms of bHLH transcription factors in plant adaptive responses to various abiotic stresses. *Frontiers in Plant Science*, 12:1143.

Rao, X. and Dixon, R. A. (2019). Co-expression networks for plant biology: why and how. *Acta biochimica et biophysica Sinica*, 51(10):981–988.

Rasamoelina, A. D., Adjailia, F., and Sinčák, P. (2020). A review of activation function for artificial neural network. In *2020 IEEE 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pages 281–286. IEEE.

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., and Vilo, J. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic acids research*, 47(W1):W191–W198.

Razzaque, S., Elias, S. M., Haque, T., Biswas, S., Jewel, G. N. A., Rahman, S., Weng, X., Ismail, A. M., Walia, H., Juenger, T. E., et al. (2019). Gene expression analysis associated with salt stress in a reciprocally crossed rice population. *Scientific reports*, 9(1):1–17.

Reddy, I. N. B. L., Kim, B.-K., Yoon, I.-S., Kim, K.-H., and Kwon, T.-R. (2017). Salt tolerance in rice: focus on mechanisms and approaches. *Rice Science*, 24(3):123–144.

Reiner, A., Yekutieli, D., and Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19(3):368–375.

Renkawitz, R. (2006). *Transcription factors and regulation of gene expression*, pages 1886–1890. Springer Berlin Heidelberg, Berlin, Heidelberg.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., and Sabeti, P. C. (2011). Detecting novel associations in large data sets. *science*, 334(6062):1518–1524.

Riascos-Arcos, J. J., Navarro, H. F. E., and Gerena, J. L. (2015). Evaluación de las herramientas de secuenciación masiva (NGS) para identificar genes asociados con tolerancia al estrés hídrico en caña de azúcar. *Acta agronomica*, 64(4):355–362.

Riccio-Rengifo, C., Finke, J., and Rocha, C. (2021a). Cosynthex. `https://github.com/criccio35/workflow_stress`.

Riccio-Rengifo, C., Finke, J., and Rocha, C. (2021b). Identifying stress responsive genes using overlapping communities in co-expression networks. *BMC bioinformatics*, 22:1–17.

Riccio-Rengifo, C., Finke, J., and Rocha, C. (2023a). Cosynthex. `https://github.com/criccio35/CoSynthEx`.

Riccio-Rengifo, C., Ramirez-Castrillon, M., Sosa, C. C., Trujillo-Montenegro, J. H., Aguilar, F. S., Riascos, J. J., Finke, J., and Rocha, C. (2023b). Assessing drought stress in sugarcane with gene expression and phenomic data using CSI-OC. *Industrial Crops and Products*. In revision.

Riyazuddin, R., Nisha, N., Singh, K., Verma, R., and Gupta, R. (2022). Involvement of dehydrin proteins in mitigating the negative effects of drought stress in plants. *Plant Cell Reports*, 41(3):519–533.

Roychoudhury, A., Basu, S., and Sengupta, D. N. (2011). Amelioration of salinity stress by exogenously applied spermidine or spermine in three varieties of indica rice differing in their level of salt tolerance. *Journal of Plant Physiology*, 168(4):317–328.

Saavedra-Díaz, C., Trujillo-Montenegro, J. H., Jaimes, H. A., Londoño, A., Salazar Villareal, F. A., López, L. O., Viveros Valens, C. A., López, J., Riascos, J. J., Quevedo, Y. M., and Aguilar, F. S. (2023). Genetic association analysis in sugarcane (saccharum spp.) for sucrose accumulation in humid environments in colombia. *BMC Plant Biology*. in revision.

Saelens, W., Cannoodt, R., and Saeys, Y. (2018). A comprehensive evaluation of module detection methods for gene expression data. *Nature Communications*, 9(1):1–12.

Savage, N. (2023). Synthetic data could be better than real data. *Nature*.

Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., and Waterston, R. H. (2017). Dna sequencing at 40: past, present and future. *Nature*, 550(7676):345–353.

Shrestha, A., Thapa, B., and Dulal, G. (2023). Sugarcane response and its related gene expression under water stress condition. In *Sugarcane - Its Products and Sustainability*. IntechOpen.

Shrivastava, P. and Kumar, R. (2015). Soil salinity: a serious environmental issue and plant growth promoting bacteria as one of the tools for its alleviation. *Saudi Journal of Biological Sciences*, 22(2):123–131.

Song, L., Langfelder, P., and Horvath, S. (2012). Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC bioinformatics*, 13(1):1–21.

Sprent, P. (2011). Fisher exact test. *International encyclopedia of statistical science*, pages 524–525.

Sundarrajan, S. and Arumugam, M. (2016). Weighted gene co-expression based biomarker discovery for psoriasis detection. *Gene*, 593(1):225–234.

Surya Krishna, S., Harish Chandar, S. R., Ravi, M., Valarmathi, R., Lakshmi, K., Prathima, P. T., Manimekalai, R., Viswanathan, R., Hemaprabha, G., and Appunu, C. (2023). Transgene-Free Genome Editing for Biotic and Abiotic Stress Resistance in Sugarcane: Prospects and Challenges. *Agronomy*, 13(4):1000.

Tahmasebi, A. and Niazi, A. (2021). Comparison of Transcriptional Response of C3 and C4 Plants to Drought Stress Using Meta-Analysis and Systems Biology Approach. *Frontiers in plant science*, 12:668736.

Tang, Y., Li, J., Song, Q., Cheng, Q., Tan, Q., Zhou, Q., Nong, Z., and Lv, P. (2023). Transcriptome and wgcna reveal hub genes in sugarcane tiller seedlings in response to drought stress. *Scientific Reports*, 13(1):12823.

Taratima, W., Chomarsa, T., and Maneerattanarungroj, P. (2022). Salinity stress response of rice (oryza sativa l. cv. luem pua) calli and seedlings. *Scientifica*, 2022.

Tavormina, P., De Coninck, B., Nikonorova, N., De Smet, I., and Cammue, B. P. A. (2015). The Plant Peptidome: An Expanding Repertoire of Structural Features and Biological Functions. *The Plant cell*, 27(8):2095–2118.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Trujillo-Montenegro, J. H., Rodríguez Cubillos, M. J., Loaiza, C. D., Quintero, M., Espitia-Navarro, H. F., Salazar Villareal, F. A., Viveros Valens, C. A., González Barrios, A. F., De Vega, J., Duitama, J., and Riascos, J. J. (2021). Unraveling the Genome of a High Yielding Colombian Sugarcane Hybrid. *Frontiers in plant science*, 12:694859.

Upadhyay, P. (2019). Climate change and adaptation strategies: A study of agriculture and livelihood adaptation by farmers in bardiya district, nepal. *Adv. Agr. Environ. Sci*, 2:47–52.

Van Bel, M., Silvestri, F., Weitz, E. M., Kreft, L., Botzki, A., Coppens, F., and Vandepoele, K. (2022). PLAZA 5.0: extending the scope and power of comparative and functional genomics in plants. *Nucleic acids research*, 50(D1):D1468–D1474.

van de Peppel, J. and Holstege, F. C. (2005). Multifunctional genes. *Molecular systems biology*, 1(1):2005–0003.

Vandereyken, K., Van Leene, J., De Coninck, B., and Cammue, B. P. A. (2018). Hub Protein Controversy: Taking a Closer Look at Plant Stress Response Hubs. *Frontiers in plant science*, 9:694.

Vázquez-Glaría, A., Eichler-Löbermann, B., Loiret, F., Ortega, E., and Kavka, M. (2021). Root-system architectures of two cuban rice cultivars with salt stress at early development stages. *Plants*, 10(6):1194.

Verma, K. K., Song, X.-P., Budeguer, F., Nikpay, A., Enrique, R., Singh, M., Zhang, B.-Q., Wu, J.-M., and Li, Y.-R. (2022a). Genetic engineering: an efficient approach to mitigating biotic and abiotic stresses in sugarcane cultivation. *Plant signaling & behavior*, 17(1):2108253.

Verma, K. K., Song, X.-P., Rajput, V. D., Boldyreva, V., Zhang, B.-Q., Minkina, T., and Li, Y.-R. (2022b). Morpho-physiological, biochemical, and ultrastructural modifications on sugarcane to prolonged water deficit. In *Agro-industrial Perspectives on Sugarcane Production under Environmental Stress*, pages 139–158. Springer Nature Singapore, Singapore.

Vives-Peris, V., Lopez-Climent, M. F., Perez-Clemente, R. M., and Gomez-Cadenas, A. (2020). Root involvement in plant responses to adverse environmental conditions. *Agronomy*, 10(7):942.

Woolson, R. F. (2007). Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials*, pages 1–3.

Yang, E.-W., Girke, T., and Jiang, T. (2013). Differential gene expression analysis using coexpression and rna-seq data. *Bioinformatics*, 29(17):2153–2161.

Yu, H., Du, Q., Campbell, M., Yu, B., Walia, H., and Zhang, C. (2021). Genome-wide discovery of natural variation in pre-mrna splicing and prioritising causal alternative splicing to salt stress response in rice. *New Phytologist*, 230(3):1273–1287.

Yuan, L., Qian, G., Chen, L., Wu, C.-L., Dan, H. C., Xiao, Y., and Wang, X. (2018). Co-expression network analysis of biomarkers for adrenocortical carcinoma. *Frontiers in genetics*, 9:328.

Zahoor, Z. and Babar, B. H. (2023). Computational Identification and Functional Characterization of Novel Genes Involved in Sugarcane Drought Tolerance. *Pakistan Sugar Journal*, 38(1):7–11.

Zeng, L. and Shannon, M. C. (2000). Salinity effects on seedling growth and yield components of rice. *Crop science*, 40(4):996–1003.

Zhang, Y.-M., Ye, D.-X., Liu, Y., Zhang, X.-Y., Zhou, Y.-L., Zhang, L., and Yang, X.-L. (2023). Peptides, new tools for plant protection in eco-agriculture. *Advanced Agrochem*, 2(1):58–78.

Zhao, C., Zhang, H., Song, C., Zhu, J.-K., and Shabala, S. (2020). Mechanisms of plant responses and adaptation to soil salinity. *The innovation*, 1(1).