Using Overlapping Communities and Network Structure for Identifying Reduced Groups of Stress Responsive Genes

Camila Riccio, Jorge Finke, and Camilo Rocha
Pontificia Universidad Javeriana, Cali, Colombia
{camila.riccio,jfinke,camilo.rocha}@javerianacali.edu.co

Abstract. This paper proposes a workflow to identify genes responding to a specific treatment in an organism, such as abiotic stresses, a main cause of extensive agricultural production losses worldwide. On input RNA sequencing read counts (measured for genotypes under control and treatment conditions) and biological replicates, it outputs a collection of characterized genes, potentially relevant to treatment. Technically, the proposed approach is both a generalization and an extension of WGCNA; its main goal is to identify specific modules in a network of genes after a sequence of normalization and filtering steps. In this work, module detection is achieved by using Hierarchical Link Clustering, which can recognize overlapping communities and thus have more biological meaning given the overlapping regulatory domains of systems that generate co-expression. Additional steps and information are also added to the workflow, where some networks in the intermediate steps are forced to be scale-free and LASSO regression is employed to select the most significant modules of phenotypical responses to stress. Finally, the workflow is showcased with a systematic study on rice (Oryza sativa), a major food source that is known to be highly sensitive to salt stress: a total of 6 modules are detected as relevant in the response to salt stress in rice; these genes may act as potential targets for the improvement of salinity tolerance in rice cultivars. The proposed workflow has the potential to ultimately reduce the search-space for candidate genes responding to a specific treatment, which can considerably optimize the effort, time, and money invested by researchers in the experimental validation of stress responsive genes.

1 Introduction

Abiotic stresses are key factors that can negatively influence plant development and productivity. They are a main cause of extensive agricultural production losses worldwide [16]. Soil salinity is one of the most devastating abiotic stresses, causing reduction in the cultivable land, crop quality, and productivity. It has been estimated that 20% of total cultivated and 33% of irrigated agricultural lands worldwide are already affected by high salinity [23]. Moreover, due to human activities and natural causes, salinized areas are gradually increasing every
year and are expected to reach 50% by the end of the year 2050 [23]. Salinity tolerance and susceptibility in plants is known to be the result of elaborated interactions between morphological, physiological, and biochemical processes that are regulated, in the end, by multiple genes in different parts of a genome [20]. Therefore, identifying groups of stress responsive genes may lead to crop improvement in terms of salinity tolerance and, ultimately, contribute solutions to the general problem of food sustainability in the years to come.

This paper proposes a workflow to identify stress responsive genes in organisms, which is known to be a complex quantitative trait. In order to discover key genes and their interaction with phenotypes related to treatment tolerance, the approach requires a collection of phenotypic traits, measured for the given genotypes. Specifically, as input, it takes RNA sequencing read counts (measured for genotypes under control and treatment conditions, and representing gene expression profiles of the target organism) and biological replicates. The output of the workflow is a collection of characterized genes, potentially relevant to treatment, yielding insight on the possible behavior of specific genes and the role they may play in functional pathways in response to the studied treatment of the organism of interest. The proposed workflow can thus take advantage of transcriptomic data for different organisms and conditions, based on the current availability of high-throughput technologies that include microarrays and RNA sequencing, to study the reaction of organisms under different environmental stimuli, such as abiotic stresses.

Technically, the proposed approach is both a generalization and an extension of Weighted Gene Co-expression Network Analysis (WGCNA) [14], a widely applied workflow that has been successfully used for identifying target genes related to diseases and cancer in several organisms [25]. The general idea behind each approach is to identify specific modules in a network of genes after a sequence of normalization and filtering steps. The proposed approach is considered a generalization of WGCNA because module detection can now recognize overlapping communities, which may have more biological meaning given the overlapping regulatory domains of systems that generate co-expression [11]. This is achieved by using Hierarchical Link Clustering (HLC) [2]. It is also an extension of WGCNA because additional steps and information are added to the workflow: namely, some networks in the intermediate steps are forced to be scale-free [4] and LASSO regression [26] is employed to select the most significant modules of phenotypical responses to stress. The advantage of using HLC as clustering method is its ability to detect overlapping modules, since biological components are involved in multiple functions and therefore biological communities tend to be highly overlapping. On the other hand, LASSO is a regularized regression technique widely used in variable selection, thanks to its ability to obtain zero regression coefficients for the less relevant variables [9]. Moreover, LASSO is especially useful in problems where the number of variables is much larger than the number of samples, which may be the case more often than desired. The proposed workflow is also modular, since other module detection and selection techniques could be used, instead HLC and LASSO, respectively.
The approach is showcased with a systematic study on rice (Oryza sativa), a major food source that is known to be highly sensitive to salt stress [7]. RNA-seq data was accessed from the GEO database [1] (accession number GSE98455). It corresponds to 57845 gene expression profiles of shoot tissues measured for both control and salt condition in 92 accessions of the Rice Diversity Panel 1. As output, 6 modules are detected as relevant in the response to salt stress in rice: 3 modules of 3 genes each, all associated with shoot K content, 2 modules of 3 genes associated with shoot biomass, and 1 module of 4 genes associated with root biomass. These genes may act as potential targets for the improvement of salinity tolerance in rice cultivars. From the 19 genes, all but 3 genes (associated with K content), were also identified as deferentially expressed for at least one of the 92 accessions, suggesting that those genes are strong candidates as stress responsive genes. Only 2 of the 16 differentially expressed genes, both from the module related to shoot biomass, are named and have an associated protein product: Spermidine hydroxycinnamoyltransferase 2 (SHT2) and Lipoygenase. In other words, further studies are needed to elucidate the detailed biological function of the remaining 14 genes that have not been named so far, which may have potential relevance in stress responsive mechanisms to salt conditions in rice. The goal is that the results reported in this paper may allow biologist to develop new rice cultivars with higher resistance to salinity.

Paper Outline. The rest of the paper is organized as follows. Section 2 gathers preliminaries on co-expression networks, HLC, and LASSO. The proposed workflow is presented in Section 3 with especial focus on the logical steps of the process and the internal structures supporting the approach. Section 4 presents the case study on the identification of potential genes highly sensitive to salt stress. Section 5 concludes the paper.

2 Preliminaries

This section presents preliminaries on networks, hierarchical link clustering, and the LASSO linear regression technique.

2.1 Co-expression network

A network is an undirected graph $G = (V,E)$ where $V = \{v_1, v_2, \ldots, v_n\}$ is a set of vertices or nodes and $E = \{e_1, e_2, \ldots, e_q\}$ is a set of edges or links that connect vertices. In a gene co-expression network, each node corresponds to a gene. A pair of genes is connected if they show similar expression patterns. A simple and unweighted network can be represented by an adjacency matrix $A \in \{0,1\}^{n \times n}$ that is symmetric with a positive one in the positions $(v_i, v_j)$ and $(v_j, v_i)$ whenever there is an edge connecting vertices $v_i$ and $v_j$, and zeros elsewhere. Co-expression networks are of biological interest because adjacent nodes in the network represent co-expressed genes that are usually controlled by the same transcriptional regulatory pathway, functionally related, or members of the same pathway or metabolic complex.
2.2 Hierarchical Link Clustering

The Hierarchical Link Clustering (HLC) algorithm was proposed by Ahn et al. [2]. The HLC approach represents communities as groups of links (rather than nodes), and each node inherits all memberships of its links and can thus belong to multiple, overlapping communities. It maps links to nodes and connects them if a pair of links shares a node. The similarity between two links \( e_{ik} \) and \( e_{jk} \) is computed using the Jaccard index

\[
S(e_{ik}, e_{jk}) = \frac{|n(i) \cap n(j)|}{|n(i) \cup n(j)|},
\]

where \( n(i) \) denotes the set containing exactly node \( i \) and its neighbors. The algorithm uses single-linkage hierarchical clustering to build a dendrogram in which each leaf is a link from the original network and branches represent link communities. Hierarchical clustering algorithms repeatedly merge groups until all elements are members of a single cluster.

For the purpose of finding meaningful communities, it is crucial to know where to partition the dendrogram. In this case, the most relevant communities are established at the maximal partition density \( D \), a function based on link density inside communities measuring the quality of a link partition. The partition density \( D \) has a single global maximum along the dendrogram in almost all cases, because its value is the average density at the top of the dendrogram (a single giant community with every link and node) and it is very small at the bottom of the dendrogram (most communities consists of a single link). In particular, it is the case that \( D = 1 \) when every community is a fully connected clique and \( D = 0 \) when each community is a tree. If a community is less dense than a tree (i.e., when the community subgraph has disconnected components), then such a community contribute negatively to \( D \), which can take negative values. The minimum density inside a community is \(-2/3\), given by one community of two disconnected edges. Since \( D \) is the average of the intra-community density, there is a lower bound of \(-2/3\) for \( D \). Computing \( D \) at each level of the link dendrogram can help the purpose of picking the best level to cut, although meaningful structure could exist above or below the threshold. The output of cutting is a set of node clusters, where each node can participate in multiple communities.

2.3 Least Absolute Shrinkage Selector Operator

The Least Absolute Shrinkage Selector Operator (LASSO) is a regularized linear regression technique. It combines a regression model with a procedure of contraction of some parameters towards zero and selection of variables, imposing a restriction or a penalty on the regression coefficients. In other words, LASSO solves the least squares problem with restriction on the \( L_1 \)-norm of the coefficient vector. It can be especially useful to solve problems where the number of variables (e.g., genes) \( n \) is much greater than the number of samples \( m \) (i.e., \( n \gg m \)).
Consider a dataset consisting of \( m \) samples, each of which consists of \( n \) covariates and a single outcome. Let \( y_i \) be the outcome and \( x_i := (x_{i1}, ..., x_{in}) \) be the covariate vector for the \( i \)-th sample. The objective of LASSO is to solve

\[
\min \left\{ \sum_{i=1}^{m} \left( y_i - \sum_{j=1}^{n} \beta_j x_{ij} \right)^2 \right\}, \quad \text{subject to} \quad \sum_{j=1}^{n} |\beta_j| \leq s.
\]

(2)

Equivalently, in the Lagrangian form, it minimizes

\[
\sum_{i=1}^{p} \left( y_i - \sum_{j=1}^{n} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{n} |\beta_j|
\]

(3)

where \( s \) is the regularization penalty and \( \lambda \geq 0 \) is the corresponding Lagrange multiplier. Since the \( \lambda \) value determines the degree of penalty, the accuracy of the model depends on its choice. Cross-validation is often used to select the regularization parameter, choosing the one that minimizes the mean-squared error.

3 The Workflow

The proposed workflow is depicted in Figure 1. This section explains the five macro-processes (A)-(E) of the proposed workflow. In comparison with WGCNA, it adds the macro-step (D) and generalizes macro-steps (A)-(C).

The workflow uses RNA-seq read counts, representing gene expression levels, as input data. More precisely, it uses \( n_0 \) gene expression profiles of an organism, measured for \( m \) different genotypes under control and treatment conditions, and \( r \) biological replicates. This raw data is represented as a matrix \( D_0 \in \mathbb{N}_0^{n_0 \times 2mr} \). In order to discover key genes and their interaction with phenotypes related to treatment tolerance, the approach also requires a set of \( p \) phenotypic traits, measured for the \( m \) genotypes. The phenotypic data is seen as a matrix \( P \in \mathbb{R}^{2m \times p} \) containing two phenotypic values per genotype, one under control condition and a second one under treatment condition.

3.1 Data Pre-processing

The goal of the data pre-processing stage is to build matrices \( P_t \) and \( L_1 \) representing, respectively, the changes in phenotypic values and expression levels between control and treatment condition, from RNA-seq and phenotypic data found in matrices \( D_0 \) and \( P \), respectively.

The RNA-seq data cannot be directly interpreted. Therefore, a normalization process is applied to deal with the problem of possible biases affecting the quantification of results. The suggested normalization technique for correcting library size and RNA composition bias is DESeq2 \cite{15}. The normalized data is represented as a matrix \( D_1 \in \mathbb{R}_0^{n_0 \times 2mr} \), and the biological replicates of each
Fig. 1. The proposed workflow comprising five macro-steps: A. Data pre-processing, B. Co-expression network construction, C. Co-expression module identification, D. Modules association to phenotypic traits, and E. Genes enrichment.
genotype are averaged and represented as a matrix \( D_2 \in \mathbb{R}^{n_0 \times 2m} \). The genes exhibiting low variance or low expression are removed from \( D_2 \), thus identifying a subset of size \( n_1 \leq n_0 \) of the original genes. The control and treatment data is separated into the matrices \( C \in \mathbb{R}^{n_1 \times m} \) and \( T \in \mathbb{R}^{n_1 \times m} \), respectively. The matrix entries \( c_{ij} \) in \( C \) and \( t_{ij} \) in \( T \) represent the normalized expression level of gene \( i \) in accession \( j \) under control and treatment condition, respectively. Control and treatment data is also separated from phenotypic data \( P \), obtaining the \( P_c \) and \( P_t \) matrices of dimensions \( m \times p \).

In the above configuration, the changes in expression levels and phenotypic values, between control and treatment conditions, are measured in terms of logarithmic ratios. In the case of expression levels, the log ratios are represented in the Log Fold Change matrix \( L_0 \in \mathbb{R}^{n_1 \times m} \), where \( \ell_{ij} = \log_2(t_{ij}/c_{ij}) \). Similarly, the log ratios of the phenotypic data are computed and represented in the \( P_\ell \in \mathbb{R}^{m \times p} \) matrix.

The final step of the data pre-processing is to filter \( L_0 \) by removing rows (e.g., genes) with low variance in the differential expression patterns, obtaining a new matrix \( L_1 \) of dimensions \( n_2 \times m \), with \( n_2 \leq n_1 \).

### 3.2 Co-expression Network Construction

A gene co-expression network connects genes with similar expression patterns across biological conditions. The purpose of this stage is to describe how to build the co-expression network \( A \) from the Log Fold Change matrix \( L_1 \), capturing the relationship between genes according to the change in expression levels between the two studied conditions. These co-expression patterns are meaningful for the identification of genes not yet associated with the response to the treatment condition.

The Log Fold Change matrix \( L_1 \) is used to build the co-expression network following the first two steps of the WGCNA methodology [14]. 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 is used as the similarity measure between genes and the resulting values are stored in the similarity matrix \( S \in \mathbb{R}_+^{n_2 \times n_2} \). Second, the matrix \( S \) is transformed into an adjacency matrix \( A \in \mathbb{R}_+^{n_2 \times n_2} \) where each entry \( a_{ij} = (s_{ij})^\beta \) 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. These networks contain many nodes with very few connections and a small number of hubs with high connections. 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 \( \beta \) is chosen to be the smallest power such that the \( R^2 \) of the linear regression between \( \log_{10}(p(k)) \) and \( \log_{10}(k) \) is close to 1 (e.g., \( R^2 > 0.85 \)).
3.3 Co-expression Module Identification

The next step in the workflow is to identifying communities (also called modules) from the co-expression network structure and dynamics represented in \( A \). The idea is to cluster genes with similar patterns of differential expression 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. Thus, unlike WGCNA, the adjacency matrix \( A \) is used to detect overlapping (rather than non-overlapping) communities, using the Hierarchical Link Clustering (HLC) algorithm (see Section 2).

As a preliminary step, matrix \( A \) is transformed into an unweighted network \( \hat{A} \in \{0, 1\}^{n_2 \times n_2} \) before using the clustering algorithm. To this end, the Pearson Correlation Coefficient (PCC) cutoff is determined using the approach described in [3]. The number of nodes, edges, and the network density is determined for different PCC cutoffs. Near the most biological relevant 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 which have a correlation score higher than the threshold are considered to have important co-expression relationship. Above the cutoff, the entries of matrix \( A \) become 1 and below the cutoff \( A \) values become 0. The HLC algorithm organizes the \( n_2 \) genes of matrix \( \hat{A} \) into \( c \) modules, where each gene can belong to one, multiple, or no module at all. This information is represented as an affiliation matrix \( F \in \{0, 1\}^{n_2 \times c} \), where \( f_{iu} = 1 \) iff node \( i \) is member of module \( u \) (it is 0, otherwise).

3.4 Module Association to Phenotypic Traits

To identify the most relevant groups (modules) of genes, associated with the phenotypic response to a specific treatment in an organism, the proposed workflow uses a LASSO based approach. Each module is represented by an eigengene, which is defined as the first principal component of such module. An eigengene can be thought of as an average differential expression profile for each community: it is computed from the Log Fold Change Matrix \( L_1 \) and the affiliation matrix \( F \). Given a module \( u \), the affiliation matrix is used to identify the genes belonging to \( u \) and then the corresponding rows of the matrix \( L_1 \) are 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 the least absolute shrinkage and selection operator (LASSO). In this context, the eigengenes (i.e., the columns of \( M \)) act as regressor variables and each phenotypic trait (i.e., each column of \( P_z \)) is used as an outcome variable.

The output after applying LASSO is a set \( W_z \) of modules for each phenotypic trait \( z \), where \( W_z \subset \{u \mid 1 \leq u \leq c\} \) for \( z = 1, 2, \ldots, p \). The target genes \( I \) for downstream analysis, which may be important in the treatment response, are the union of genes belonging to the selected modules; that is \( I = \bigcup_{z=1}^{p} W_z \), where \( I \subset \{i \mid 1 \leq i \leq n_2\} \).
3.5 Gene Enrichment

The goal of this final stage of the process is to characterize with additional information the genes identified in previous stages, helping to elucidate their possible behavior and role in the response to the studied treatment.

A crucial step is to identify the differentially expressed genes in set \( I \). That is, to select those genes in \( I \) having an absolute value of the Log Fold Change of at least 2 (\(|\ell_{ij}| \geq 2\)) for at least one sample. This represents genes whose expression level is quadrupled (up or down) from control to treatment condition; they are strong candidates for treatment responsive genes.

Also, functional category enrichment can be done by, e.g., searching for gene ontology annotations in databases such as QuickGO \([5]\). Such annotations can provide evidence of biological implications of the target genes in the treatment-tolerance mechanisms. Furthermore, QuickGO can be used to identify genes with reported protein products, which can be used to perform additional relevant analysis reviewing their reported protein-protein interactions in other databases, such as STRING \([24]\). The interactions include direct (physical) and indirect (functional) associations, and they stem from computational prediction, knowledge transfer between organisms, and interactions aggregated from other (primary) databases. This information can give new insights on how the selected genes are involved in functional pathways that can be related to the treatment of interest.

4 Identifying Potential Saline Stress Responsive Genes in Rice

This section presents a case study on the identification of genes in \( Oryza sativa \) that may respond to saline stress following the approach presented in Section 3.

The RNA-seq data used for the experiments summarized in this section was obtained from GEO database \([1]\) (accession number GSE98455). This data corresponds to \( n_0 = 57 845 \) gene expression profiles of shoot tissues measured for both control and salt condition in \( m = 92 \) accessions of the Rice Diversity Panel 1, with \( r = 2 \) biological replicates. A total of \( p = 3 \) phenotypic traits are used: shoot \( K^+ \) content, root biomass, and shoot biomass. These traits were measured for the same 92 genotypes, under control and salt stress conditions, and can be found in the supplementary information of \([6]\).

4.1 Data Pre-processing

DESeq2 normalization is applied to the raw data and the biological replicates are averaged. Genes exhibiting low variance are identified as those with ratio of upper quantile to lower quantile smaller than 1.5 and are removed from the normalized data. Genes with low expression, corresponding to those having more than 80% samples with values smaller than 10, are also removed. A total of \( n_1 = 9 414 \) genes are kept after this filtering process.
From the Log Fold Change matrix $L_0$, genes whose difference between upper quantile and lower quantile greater than 0.25 are removed. Therefore, the resulting matrix $L_1$ contains the log ratios of $n_2 = 8928$ genes. The logarithmic ratios of the phenotypic data, for the 92 accessions and the 3 traits, are also computed.

### 4.2 Co-expression Network Construction

The Log Fold Change matrix $L_1$ is used to compute the corresponding similarity matrix. For this network, it is observed that $\beta = 3$ is the smallest integer such that $R^2 \geq 0.8$. Figure 2 depicts the degree distribution of the similarity matrix (left) and the degree distribution of the adjacency matrix (right), which is the degree distribution of a scale-free network with $R^2 = 0.8$ with $\beta = 3$.

![Fig. 2. Degree distribution of $A$ (left) and $\hat{A}$ (right)](image)

The resulting adjacency matrix $A$ represents a complete graph $G = (V, E)$, with $|V| = 8928$ genes and $|E| = 39850128$ edges.

### 4.3 Co-expression Module Identification

The adjacency matrix $A$ is transformed into an unweighted network $\hat{A}$ applying the approach described in [3], based on the density of the network combined with the decreasing number of nodes and edges with higher PCC values. The cutoff value is set to 0.2, thus keeping only the connections above this threshold and removing the isolated nodes. The resulting adjacency matrix $\hat{A}$ has 5810 connected genes and accounts for 16875145 edges.

After applying the HLC algorithm, a total of 4131 genes are distributed in $c = 5143$ overlapping modules of at least 3 genes. Figure 3 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.
4.4 Module Association to Phenotypic Traits

The phenotypic traits under study are shoot $K^+$ content, root biomass, and shoot biomass. Figure 4 suggests that there are significant differences in the values of these phenotypic traits between stress and control conditions. This supports the working hypothesis that these three variables represent tolerance-associated traits in rice under salt stress.

Using the affiliation matrix $F$ derived from the HLC output and the Log Fold Change matrix $L_1$, a matrix $M$ is built by computing the eigengene for each of the $c = 5143$ modules. The LASSO technique is applied by using each of the phenotypic traits as the outcome variable, one at a time. As shown in Figure 5, cross-validation is performed for each phenotypical trait in order to select the corresponding regularization parameter $\lambda$ that minimizes the mean-squared error.

Finally, three LASSO models are adjusted by using the corresponding $\lambda$ and phenotypical data with the eigengenes of matrix $M$. As result, 6 modules are detected as relevant in the response to salt stress in rice: 3 modules of 3 genes, each associated with shoot $K$ content; 2 modules of 3 genes associated with shoot biomass; and 1 module of 4 genes associated with root biomass (see Figure 6).
4.5 Gene Enrichment

From the 19 genes selected by LASSO, all but 3 genes (the ones associated to $K$ content), are also identified as deferentially expressed ($|f_{ij}| \geq 2$) for at least one of the 92 accessions. This suggests that those genes are strong candidates as stress responsive genes to salt conditions in rice.

Figure 5 summarizes how from the initial $n_0 = 57,845$ genes, the proposed workflow selects a reduced set of 19 genes. First, 48,431 genes are discarded after filtering the normalized expression data $D_2$ and then 486 additional genes are discarded when filtering the Log Fold Change matrix $L_0$, to finally arrive at 19 genes, of which 16 are differentially expressed.

According to the Quickgo database, only 2 of the 16 differentially expressed genes (both from the module related to shoot biomass) are named and have an associated protein product: spermidine hydroxycinnamoyltransferase 2 (SHT2) and lipoxygenase. Figure 7 shows their corresponding 3D protein structures.

The Uniprot database [8] reports, on the one hand, that SHT2 contributes to the natural variation of spermidine-based phenolamides in rice cultivars. On the other hand, it is reported in [8] that plant lipoxygenase may be involved in a number of diverse aspects of plant physiology including growth and development, pest resistance, and senescence or responses to wounding. This protein is involved in the pathway oxylipin biosynthesis, which is part of Lipid metabolism. Additionally, previous studies in [12, 13, 17, 19, 21] provide evidence of biological implications of spermidine and lipoxygenase in tolerance to salt stress in other plants or even in rice cultivars.

As a conclusion, the results presented in this section suggest that further studies are needed to elucidate the detailed biological function of the remaining 14 genes that have not been named so far in the literature. They may have the potential to intervene in stress responsive mechanisms to salt conditions in rice.
Fig. 6. Venn diagram representing the number of genes selected at different stages of the proposed workflow for the case study in rice.

Fig. 7. 3D protein structure of named genes selected by LASSO, available from [24].
5 Concluding Remarks

This manuscript provides a detailed description of a network-based analysis workflow for the discovery of key genes responding to a specific treatment in an organism. It links transcriptomic with phenotypic data, and identifies overlapping gene modules.

The proposed approach is inspired by the workflow suggested in the WGCNA [14]. Its main steps are the preprocessing of the gene expression data, the construction of a co-expression network, the detection of modules within the network, the relation of modules with external information (e.g. phenotypic data), and the enrichment of the identified key genes with additional information. Both approaches are structured in a modular way, which allows modifying and exploring different techniques in each step of the workflow.

The proposed workflow is designed to integrate expression data measured under two different conditions (namely, control and treatment), unlike the usually co-expression-based approaches which work with both conditions independently or consider only a single condition. For this purpose, an approach similar to that proposed in [10] is used, where the control and treatment data are compiled in a single matrix using the Log Fold Change measure. Thus, the input to construct the co-expression network is not the expression data, but instead the changes in the expression levels from one condition to the other, making room for capturing the signal of changes caused by the treatment.

Another important feature in the proposed workflow is the module detection technique. The co-expression network is computed, as in WGCNA, until a scale-free network is obtained. In the proposed approach, this network is then used to apply the HLC algorithm, a clustering technique capable of detecting overlapping communities. Several approaches of module detection from gene expression have been proposed and were evaluated in [22]. Most of them focus mainly on disjoint (non-overlapping) communities; the described techniques dealing with overlaps are not clustering, but bi-clustering and decomposition methods. It is well known that communities in real networks, including biological ones, are overlapping [18]. Thus, the approach presented in this work can be seen as a generalization of the previous approaches such as WGCNA with the potential to deal with genes associated to multiple biological processes.

The approach was applied in a case study with rice under salt stress. The results show a group of 14 genes, of which only 2 of them have been previously related to saline stress response in other studies. As future work, other overlapping module detection and selection techniques should be used instead HLC and LASSO, respectively. The combination of these techniques would allow finding target genes for future biological studies that evaluate their potential as genes that respond to salt stress in rice, and other crops and stresses. In-vivo laboratory experimentation needs to be conducted to validate the findings of this paper in relation to salinity stress.

Finally, the workflow is presented as a protocol capable of considerably reducing the number of genes detected as relevant in the response to stress of choice. Other traditionally used methods for this purpose tend to generate a large list of
candidate genes, thus limiting subsequent efforts in experimental validation. In this sense, the proposed workflow can help in reducing such efforts in time and money invested by researchers in the experimental validation of stress-responsive genes.

References

1. Geo accession viewer. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE98455 (Accessed on 10/16/2019)
2. Ahn, Y.Y., Bagrow, J.P., Lehmann, S.: Link communities reveal multiscale complexity in networks. nature 466(7307), 761–764 (2010)
3. Aoki, K., Ogata, Y., Shibata, D.: Approaches for extracting practical information from gene co-expression networks in plant biology. Plant and Cell Physiology 48(3), 381–390 (2007)
4. Barabási, A.L., Bonabeau, E.: Scale-free networks. Scientific american 288(5), 60–69 (2003)
5. Binns, D., Dimmer, E., Huntley, R., Barrell, D., O’donovan, C., Apweiler, R.: Quickgo: a web-based tool for gene ontology searching. Bioinformatics 25(22), 3045–3046 (2009)
6. 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.: 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 (2017)
7. Chang, J., Cheong, B.E., Natera, S., Roessner, U.: 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 (2019)
8. Consortium, U., et al.: Uniprot: the universal protein knowledgebase. Nucleic acids research 46(5), 2699 (2018)
9. Desboulets, L.D.D.: A review on variable selection in regression analysis. Econometrics 6(4), 45 (2018)
10. Du, Q., Campbell, M., Yu, H., Liu, K., Walia, H., Zhang, Q., Zhang, C.: Network-based feature selection reveals substructures of gene modules responding to salt stress in rice. Plant direct 3(8), e00154 (2019)
11. Gaiteri, C., Ding, Y., French, B., Tseng, G.C., Sibille, E.: Beyond modules and hubs: the potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders. Genes, brain and behavior 13(1), 13–24 (2014)
12. Gupta, K., Dey, A., Gupta, B.: Plant polyamines in abiotic stress responses. Acta physiologicae plantarum 35(7), 2015–2036 (2013)
13. Hou, Y., Meng, K., Han, Y., Ban, Q., Wang, B., Suo, J., Lv, J., Rao, J.: The perismon 9-lipoxygenase gene dkllox3 plays positive roles in both promoting senescence and enhancing tolerance to abiotic stress. Frontiers in plant science 6, 1073 (2015)
14. Langfelder, P., Horvath, S.: Wgncna: an r package for weighted correlation network analysis. BMC bioinformatics 9(1), 559 (2008)
15. Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome biology 15(12), 550 (2014)
16. Mesterházy, Á., Oláh, J., Popp, J.: Losses in the grain supply chain: Causes and solutions. Sustainability 12(6), 2342 (2020)
17. Mittova, V., Tal, M., Volokita, M., Guy, M.: 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 (2002)
18. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. nature 435(7043), 814–818 (2005)
19. 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.: A novel hydroxycinnamoyl transferase for synthesis of hydroxycinnamoyl spermine conjugates in plants. BMC plant biology 19(1), 1–13 (2019)
20. Reddy, I.N.B.L., Kim, B.K., Yoon, I.S., Kim, K.H., Kwon, T.R.: Salt tolerance in rice: focus on mechanisms and approaches. Rice Science 24(3), 123–144 (2017)
21. Roychoudhury, A., Basu, S., Sengupta, D.N.: 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 (2011)
22. Saclens, W., Cannoodt, R., Saeyts, Y.: A comprehensive evaluation of module detection methods for gene expression data. Nature communications 9(1), 1–12 (2018)
23. Shrivastava, P., Kumar, R.: 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 (2015)
24. Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., et al.: The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic acids research p. gw937 (2016)
25. Tian, H., Guan, D., Li, J.: Identifying osteosarcoma metastasis associated genes by weighted gene co-expression network analysis (wgcna). Medicine 97(24) (2018)
26. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58(1), 267–288 (1996)