METHODOLOGY

An Analysis of Overlapping Communities for Identifying Stress Responsive Genes

Camila Riccio[1*], Jorge Finke[2] and Camilo Rocha[2]

[*]Correspondence:
camila.riccio@javerianacali.edu.co
[1]Department of Natural Sciences and Mathematics, Pontificia Universidad Javeriana, Cali, Colombia
Full list of author information is available at the end of the article

Abstract

Background: This paper proposes a workflow to identify which genes respond to specific treatments in plants. The workflow takes as input the RNA sequencing read counts and phenotypical data of different genotypes, measured under control and treatment conditions. It outputs a reduced group of genes marked as relevant for treatment response. Technically, the proposed approach is both a generalization and an extension of WGCNA. It aims to identify specific modules of overlapping communities underlying the co-expression network of genes. Module detection is achieved by using Hierarchical Link Clustering. Identifying such modules enables us to take into account the overlapping nature of the regulatory domains of the systems that generate co-expression. LASSO regression is employed to analyze phenotypic responses of modules to treatment.

Results: The workflow is applied to rice (*Oryza sativa*), a major food source known to be highly sensitive to salt stress. The workflow identifies 19 rice genes that seem relevant in the response to salt stress. They are distributed across 6 modules: 3 modules, each grouping together 3 genes, are associated to shoot K content; 2 modules of 3 genes, are associated to shoot biomass; and 1 module of 4 genes is associated to root biomass. These genes represent target genes for the improvement of salinity tolerance in rice.

Conclusions: A more effective framework to reduce the search-space for target genes that respond to a specific treatment is introduced. It facilitates experimental validation by restraining efforts to a smaller subset of genes of high potential relevance.

Keywords: Stress-responsive genes; co-expression network; overlapping communities; phenotypic traits; LASSO; salinity; rice; *Oryza sativa*

Introduction

Stresses are key factors that influence plant development, often associated to extensive losses in agricultural production [1, 2]. Soil salinity is one of the most devastating abiotic stresses. According to [2], 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 2050 areas of high salinity are expected to reach 50% of the cultivated land [2].

Salinity tolerance and susceptibility 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 [3]. Consequently, identifying

groups of responsive genes is an important step in efforts to improve crop varieties in terms of salinity tolerance. This paper proposes a workflow to identify stress responsive genes associated with a complex quantitative trait.

To discover which genes are associated with a phenotypic response to treatment the workflow takes as input the gene expression profiles of the target organism, specifically, the RNA sequencing read counts (measured under control and treatment conditions) of at least two biological replicates per genotype. It also receives phenotypic data, specifically, observable traits, measured for each genotype under the two conditions. The output of the workflow is a set of genes which are characterized as potentially relevant to treatment.

Broadly speaking, the workflow provides a framework that yields insight into the possible behavior of specific genes and the role they play in functional pathways in response to treatment. It takes advantage of the current availability of high-throughput technologies, which enables the access to transcriptomic data of organisms under different conditions, and a better understanding of their reaction under different environmental stimuli.

The proposed approach is both a generalization and an extension of the widely applied workflow for identifying target genes called Weighted Gene Co-expression Network Analysis (WGCNA) [4, 5]. Like WGCNA, the general idea behind the proposed approach is to identify, after a sequence of normalization and filtering steps, specific modules of overlapping communities underlying the co-expression network of genes. The proposed approach is considered a *generalization* of WGCNA because module detection recognizes overlapping communities using Hierarchical Link Clustering (HLC) [7] algorithm. Conceptually, the generalization takes into account the overlapping nature of the regulatory domains of the systems that generate the co-expression network [6]. More specifically, overlapping modules allow for scenarios where biological components are involved in multiple functions.

The workflow is also an *extension* of WGCNA because two additional constraints are considered: networks in the intermediate steps are forced to be scale-free [8], and LASSO regression [9] selects the most relevant modules of responsive genes. The regularized regression technique of LASSO forces the less relevant variables to be associated to regression coefficients of value zero [10], which is of interest in scenarios where the number of variables is much larger than the number of samples. This condition is satisfied when the target variables represent the overlapping communities (obtained with HLC) and the samples represent genotype data, which is usually a small set due to the high cost of the RNA sequencing process. Finally, the proposed workflow is also modular, since other module detection and selection techniques could be explored instead HLC and LASSO.

The approach is showcased with a systematic study on rice (*Oryza sativa*), a food source that is known to be highly sensitive to salt stress [11]. RNA-seq data was accessed from the GEO database [12] (accession number GSE98455). It represents 57845 gene expression profiles of shoot tissues measured under control and

stress conditions in 92 accessions of the Rice Diversity Panel 1 [13]. A total of 6 modules are detected as relevant in the response to salt stress in rice: 3 modules, each grouping together 3 genes, are associated to shoot K content; 2 modules of 3 genes, are associated to shoot biomass; and 1 module of 4 genes is associated to root biomass. These genes are potential targets for experimental validation of salinity tolerance. From the 19 genes, 16 are also identified as deferentially expressed for at least one of the 92 accessions, which re-enforces the labelling of the genes as stress responsive. Only 2 of the 16 differentially expressed genes, associated to shoot biomass, are named and known to produce protein products: spermidine hydroxycinnamoyltransferase 2 (SHT2) and lipoxygenase. Further studies are needed to elucidate the detailed biological functions of the remaining 14 genes and their role in the mechanisms that respond to salt conditions.

Paper Outline. The remainder of the paper is organized as follows. The Preliminaries section gathers foundations on gene co-expression networks, HLC, and LASSO. The proposed workflow is presented in the Workflow section, which enphasizes on the logical steps of the data analysis process and the internal structures supporting the approach. The Case Study section presents an application of the workflow for the identification of rice genes that are sensitive to salt stress. Finally, the Concluding Remarks section draws some conclusions and future research directions.

Preliminaries

This section presents preliminaries on networks, the clustering algorithm HLC, and the linear regression technique LASSO.

Co-expression network

Consider a network as an undirected graph $G = (V, E)$ where $V = \{v_1, v_2, \ldots, v_n\}$ is a set of n *vertices* (or *nodes*) and $E = \{e_1, e_2, \ldots, e_q\}$ is a set of q *edges* (or *links*) that connect vertices. In a co-expression network of genes, each node corresponds to a gene, and a link indicates similar expression pattern between two genes. The network can be represented by an adjacency matrix $A \in \{0,1\}^{n\times n}$ that is symmetric. A matrix entry in positions (v_i, v_j) and (v_j, v_i) is equal to 1 whenever there is an edge connecting vertices v_i and v_j, and equal to 0 otherwise. 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 [14].

Hierarchical Link Clustering

The Hierarchical Link Clustering (HLC) algorithm partitions groups of links (rather than nodes), each node inherits all memberships of its links and can belong to multiple, overlapping communities [7]. More specifically, HLC evaluates the similarity between links if they share a particular node. Consider a pair of links e_{ik} and e_{jk}, which are adjacent to node k. The similarity between e_{ik} and e_{jk} is defined based on the Jaccard index as

$$S(e_{ik}, e_{jk}) = \frac{|\ \eta(i) \cap \eta(j)\ |}{|\ \eta(i) \cup \eta(j)\ |}, \tag{1}$$

where $\eta(i)$ denotes the set containing 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 network, and branches represent link communities.

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

$$D = \frac{2}{|E|} \sum_c |E_c| \frac{|E_c| - |V_c| + 1}{(|V_c| - 1)(|V_c| - 2)} \tag{2}$$

Note that in most cases, the partition density D has a single global maximum along the dendrogram. If the dendogram is cut at the top, then D represents the average link density of a single giant community. If the dendogram is cut at the bottom, then most communities consists of a single link. In other words, note that $D = 1$ when every community is a 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 contributes 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. 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).

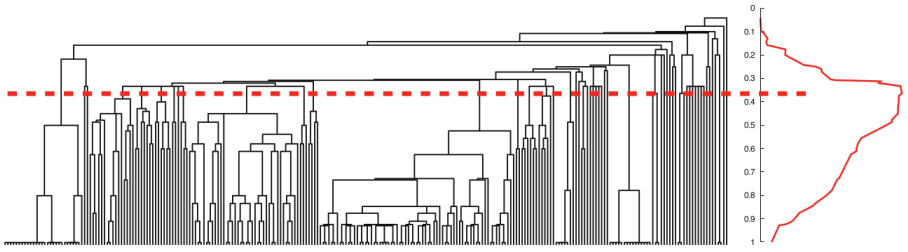

Figure 1 Example of a full link dendrogram (left) and partition density (right), borrowed from [7].

The output of cutting is a set of node clusters, where each node can participate in multiple communities.

Least Absolute Shrinkage Selector Operator (LASSO)

LASSO is a regularized linear regression technique. By combining a regression 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

$$\min \left\{ \sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{c} \beta_j x_{ij} \right)^2 \right\} \quad , \quad \text{subject to} \quad \sum_{j=1}^{c} |\beta_j| \leq s. \tag{3}$$

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

$$\sum_{i=1}^{p} \left(y_i - \sum_{j=1}^{c} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{c} |\beta_j| \tag{4}$$

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.

The Workflow

Figure 2 introduces the proposed workflow. It can be broken down into five macro-processes (A)-(E). Compared to WGCNA, the workflow adds the macro-step (D) and generalizes macro-steps (A)-(C).

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

A. Data Pre-processing

The goal of the data pre-processing stage is to build matrices P_ℓ and L_1 which represent, respectively, the changes in phenotypic values and expression levels between control and treatment condition. In other words, P_ℓ and L_1 are constructed from RNA-seq and phenotypic data found in matrices D_0 and P.

A normalization process is applied to interpret RNA-seq data and handle possible biases affecting the quantification of results. Here, DESeq2 [15] is used to correct for library size and RNA composition bias. The normalized data is represented as a matrix $D_1 \in \mathbb{R}^{n_0 \times 2mr}$, and the biological replicates of each 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. Consequently, this stage of the approach reduces

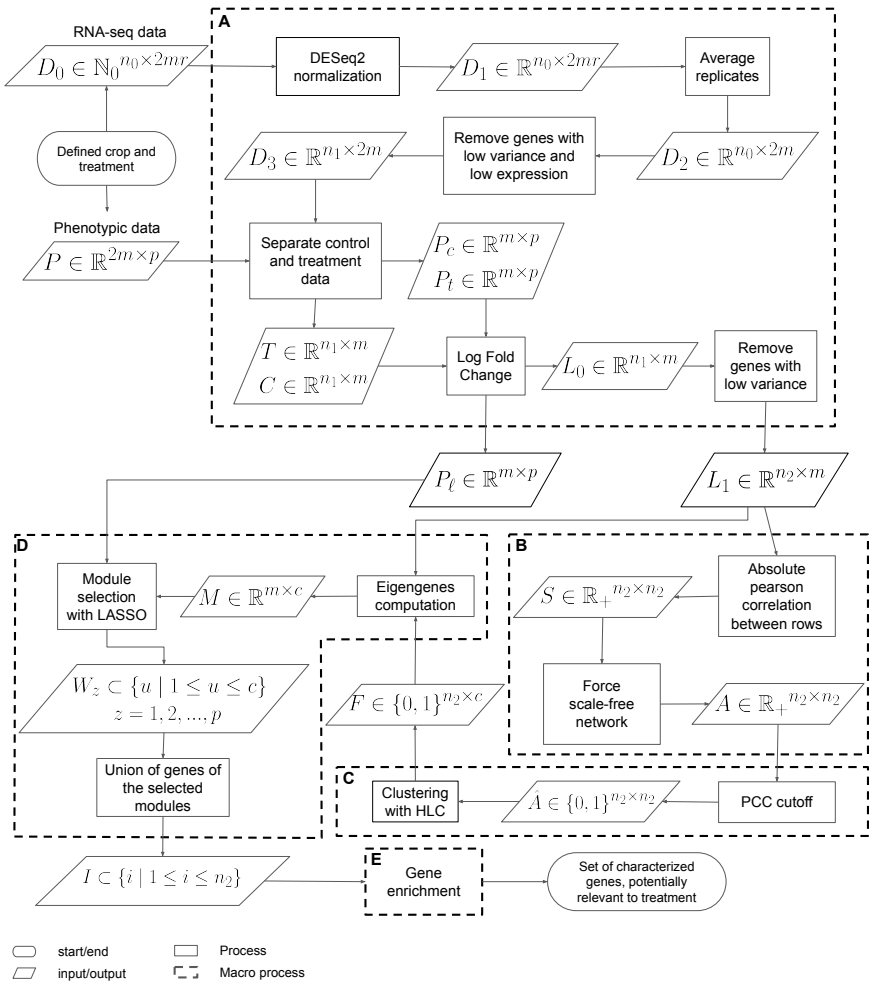

Figure 2 The proposed workflow is broken down into five macro-steps: A. Data pre-processing, B. Co-expression network contruction, C. Identification of co-expression modules, D. Detection of modules association to phenotypic traits, and E. Gene enrichment.

the set of genes from a pool of size n_0 to n_1. 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 stage of 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 $n_2 \leq n_1$.

B. Construction of the Co-expression Network

A gene co-expression network connects genes with similar expression patterns across biological conditions. The purpose of this step 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 that are not yet associated to treatment response.

The Log Fold Change matrix L_1 is used to build the co-expression network following the first two steps of WGCNA [4]. 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 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 (Here, $R^2 > 0.8$).

C. Identification of Co-expression Modules

The next step in the workflow is to identifying modules of overlapping communities from the co-expression network represented by 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. Unlike WGCNA, the workflow applies the Hierarchical Link Clustering (HLC) algorithm (overviewed in the Preliminaries section) to detect overlapping rather than non-overlapping communities.

First, the adjacency matrix A is transformed into an unweighted network $\hat{A} \in \{0,1\}^{n_2 \times n_2}$. To this end, the Pearson Correlation Coefficient (PCC) cutoff is determined using the approach described in [16]. 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 which have a correlation score higher than the threshold are considered to have a significant level of co-expression. 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 zero or multiple modules. This information is represented as an affiliation matrix $F \in \{0,1\}^{n_2 \times c}$, where $f_{iu} = 1$ if node i is a member of module u (and $f_{iu} = 0$ otherwise).

D. Detection of Modules Association to Phenotypic Traits

To identify the most relevant modules, associated with the phenotypic response to a specific treatment, the proposed workflow uses LASSO. Specifically, each module is represented by an eigengene, which is defined as the first principal component of such module. An eigengene can be seen 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 F 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 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_ℓ) 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 $z = 1, 2, .., p$. The target genes I for downstream analysis are the union of genes belonging to the selected modules; that is $I = \cup_{z=1}^{p} W_z$, where $I \subset \{i \mid 1 \leq i \leq n_2\}$.

E. Gene Enrichment

The goal of this final step of the process is to annotate 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 the genes in I that have 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 the control to treatment condition; they are target genes.

Furthermore, functional category enrichment can be carried out by, e.g., searching for gene ontology annotations in databases such as QuickGO [17]. Such annotations can provide evidence of biological implications of the target genes in the treatment-tolerance mechanisms. Furthermore, QuickGO can be used to identify the protein products of genes, which can be used to perform additional analysis that provides new insights into how target genes are involved in functional pathways that can be related to treatment. Such analysis includes a review of reported protein-protein interactions in databases such as STRING [18]. The protein interactions include direct (physical) and indirect (functional) associations. They stem from computational prediction, knowledge transfer between organisms, and interactions aggregated from other (primary) databases. The search for unknown interactions would extend the workflow with additional steps.

Identifying Potential Saline Stress Responsive Genes in Rice

This section presents a case study, applying the approach introduced in the Workflow section, to identify genes in *Oryza sativa* that respond to saline stress.

The RNA-seq data are available from the GEO database [12] (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 [13], 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 treatment conditions, and can be found in the supplementary information for [19].

A. 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. After this filtering process a total of $n_1 = 9\,414$ genes are kept for further analysis.

From the Log Fold Change matrix L_0, genes whose difference between upper quantile and lower quantile is greater than 0.25 are removed. Therefore, the resulting matrix L_1 contains the log ratios of $n_2 = 8\,928$ genes. The logarithmic ratios of the phenotypic data, for the 92 accessions and the 3 traits, are also computed.

B. Construction of the Co-expression Network

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 the $R^2 \geq 0.8$. Figure 3 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 $\beta = 3$.

The resulting adjacency matrix A represents a complete graph $G = (V, E)$, with $|V| = 8\,928$ genes ($|E| = 39\,850\,128$ edges).

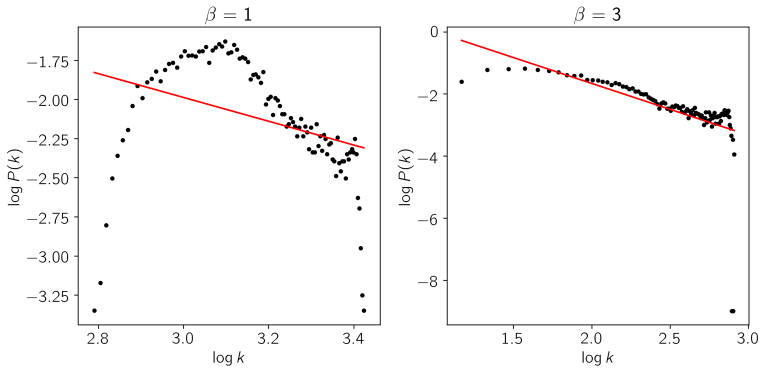

Figure 3 Degree distribution of the network represented by S (left) and A (right).

C. Identification of Co-expression Modules

The adjacency matrix A is transformed into an unweighted network $\hat{A}$ applying the approach described in [16]. The cutoff value is set to 0.2, based on the density of the network combined with the decreasing number of nodes and edges with higher PCC values. Thus keeping only the connections above this threshold and removing any isolated nodes. The resulting adjacency matrix $\hat{A}$ has 5 810 connected genes and accounts for 614 501 edges.

After applying the HLC algorithm, a total of 4 131 genes are distributed in $c =$ 5 143 overlapping modules of at least 3 genes. Figure 4 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.

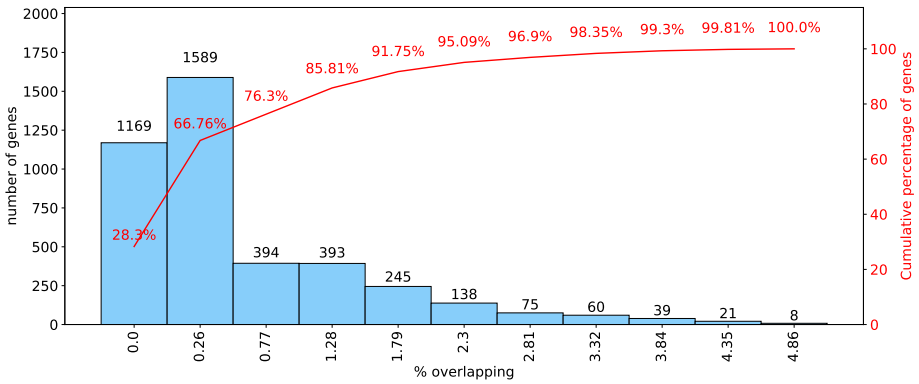

Figure 4 Overlapping percentage of genes after applying HLC.

D. Detection of Modules Association to Phenotypic Traits

The phenotypic traits under study are shoot K content, root biomass, and shoot biomass. Figure 5 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.

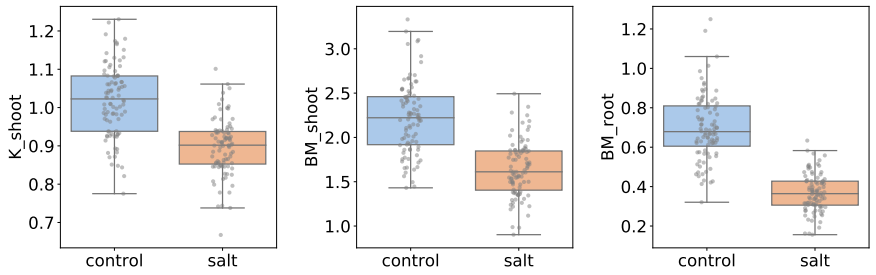

Figure 5 Phenotipic traits distribution under control and salt stress.

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 $c = 5\,143$ 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 6, cross-validation is performed for each phenotypical trait in order to select the corresponding regularization parameter λ that minimizes the mean-squared error.

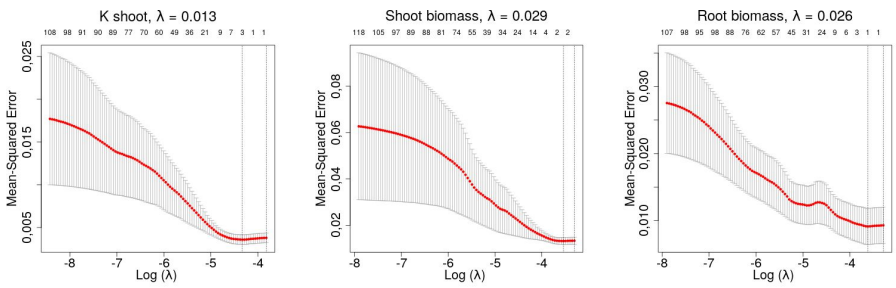

Figure 6 Cross-validation of the LASSO regularization parameter λ, for each phenotypic trait.

Finally, three LASSO models are adjusted by using the corresponding λ 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 7).

E. 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 ($|\ell_{ij}| \geq 2$) for at least one of the 92 accessions. These genes are strong candidates target genes in rice.

Figure 7 summarizes how from the initial $n_0 = 57\,845$ genes, the proposed workflow identifies 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.

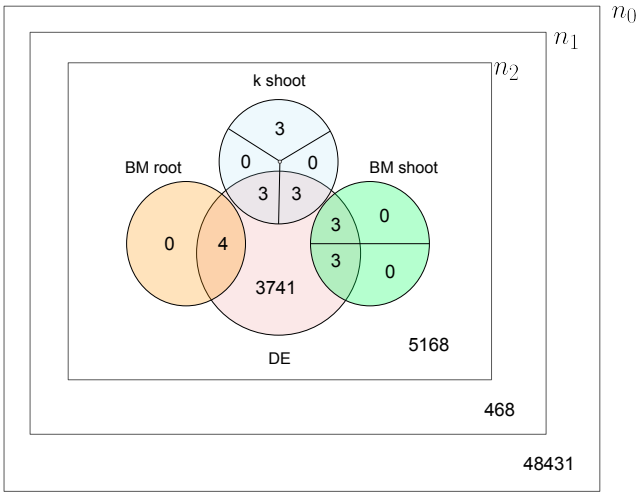

Figure 7 Venn diagram representing the number of genes selected at different stages of the proposed workflow for the case study in rice.

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 8 shows their corresponding 3D protein structures.

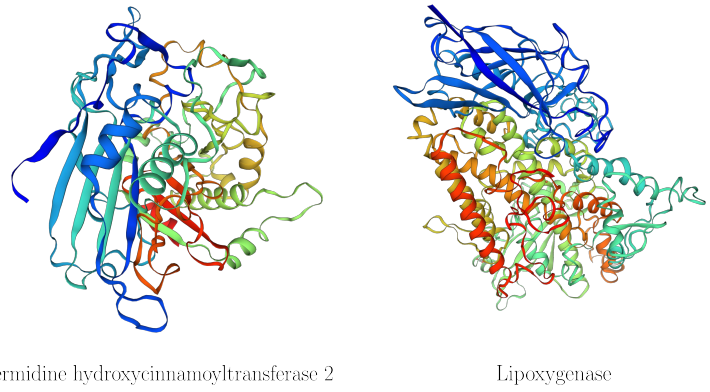

Figure 8 3D protein structure of named genes selected by LASSO, borrowed from [18].

The Uniprot database [20] 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 [20] 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 [21, 22, 23, 24, 25] provide evidence of biological implications of sperimidine 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.

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 [4]. 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 [26] 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.

An 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 [27]. Most of them focus mainly on disjoint (non-overlapping) communities; the techniques described 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 likely overlap [28]. 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.

Acknowledgements
Not applicable

Funding
This work was funded by the OMICAS program: Optimización Multiescala In-silico de Cultivos Agrícolas Sostenibles (Infraestructura y Validación en Arroz y Caña de Azúcar), anchored at the Pontificia Universidad Javeriana in Cali and funded within the Colombian Scientific Ecosystem by The World Bank, the Colombian Ministry of Science, Technology and Innovation, the Colombian Ministry of Education and the Colombian Ministry of Industry and Turism, and ICETEX, under GRANT ID: FP44842-217-2018.

Abbreviations
HLC: Hierarchical Link Clustering
LASSO: Least Absolute Shrinkage Selector Operator
PCC: Pearson Correlation Coefficient
RNA-seq: RNA sequencing
SHT2: Spermidine hydroxycinnamoyltransferase 2
WGCNA: Weighted Gene Co-expression Network Analysis

Availability of data and materials
The datasets analysed during the current study are publicy available. They can be found in the following locations:

- RNA-seq data of salt stress in rice is available on the GEO (GSE98455).
- Phenotypic data of salt stress in rice is a subset of the supplementary file 1 included in [19].

Ethics approval and consent to participate
Not applicable

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable

Authors' contributions

J.F. and C.R. proposed the original idea. J.F. provide advice on algorithms concepts and implementation. C.R. structured the methodology and performed the analysis. C.R., J.F., and H.C.R. wrote the manuscript. All authors read and approved the final manuscript.

Author details
[1]Department of Natural Sciences and Mathematics, Pontificia Universidad Javeriana, Cali, Colombia. [2]Department of Electronics and Computer Science, Pontificia Universidad Javeriana, Cali, Colombia.

References

1. Mesterházy, Á., Oláh, J., Popp, J.: Losses in the grain supply chain: Causes and solutions. Sustainability **12**(6), 2342 (2020)
2. 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)
3. 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)
4. Langfelder, P., Horvath, S.: WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics **9**(1), 559 (2008)
5. Tian, H., Guan, D., Li, J.: Identifying osteosarcoma metastasis associated genes by weighted gene co-expression network analysis (WGCNA). Medicine **97**(24) (2018)
6. 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)
7. Ahn, Y.-Y., Bagrow, J.P., Lehmann, S.: Link communities reveal multiscale complexity in networks. Nature **466**(7307), 761–764 (2010)
8. Barabási, A.-L., Bonabeau, E.: Scale-free networks. Scientific American **288**(5), 60–69 (2003)
9. Tibshirani, R.: Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society: Series B (Methodological) **58**(1), 267–288 (1996)
10. Desboulets, L.D.D.: A review on variable selection in regression analysis. Econometrics **6**(4), 45 (2018)
11. 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)
12. Clough, E., Barrett, T.: The gene expression omnibus database. In: Statistical Genomics. Methods in Molecular Biology, vol. 1418, pp. 93–110. Humana Press, New York, NY (2016)
13. Eizenga, G.C., Ali, M.L., Bryant, R.J., Yeater, K.M., McClung, A.M., McCouch, S.R.: Registration of the rice diversity panel 1 for genomewide association studies. Journal of Plant Registrations **8**(1), 109–116 (2014)
14. Fionda, V.: Networks in biology. In: Ranganathan, S., Gribskov, M., Nakai, K., Schönbach, C. (eds.) Encyclopedia of Bioinformatics and Computational Biology vol. 1, pp. 915–921. Academic Press, Oxford (2019)
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. 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)
17. 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)
18. 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, 937 (2016)
19. 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), 1006823 (2017)
20. Consortium, U., *et al.*: UniProt: the universal protein knowledgebase. Nucleic Acids Research **46**(5), 2699 (2018)
21. Gupta, K., Dey, A., Gupta, B.: Plant polyamines in abiotic stress responses. Acta Physiologiae Plantarum **35**(7), 2015–2036 (2013)
22. Hou, Y., Meng, K., Han, Y., Ban, Q., Wang, B., Suo, J., Lv, J., Rao, J.: 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 (2015)
23. 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)
24. 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)
25. 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)
26. 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), 00154 (2019)
27. Saelens, W., Cannoodt, R., Saeys, Y.: A comprehensive evaluation of module detection methods for gene expression data. Nature Communications **9**(1), 1–12 (2018)
28. 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)