

Identification of rice genes which respond to¹ saline stress from co-expression networks analysis

Camila Riccio Rengifo
Pontificia Universidad Javeriana Cali

INTRODUCTION

Abiotic stresses are the key factors which negatively influence plant development and productivity. They are the main cause of extensive agricultural production losses worldwide. One of the most devastating abiotic stresses, causing reduction in the cultivable land, crop quality and productivity is soil salinity. It has been estimated that 20% of total cultivated and 33% of irrigated agricultural lands worldwide are already affected by high salinity. Due to the human activities and natural causes, salinized areas are gradually increasing every year and are expected to reach 50% by the end of year 2050 [18]. Salinity effects are the result of elaborated interactions among morphological, physiological, and biochemical processes. Those processes are regulated by multiple genes and determine the salt tolerance or susceptibility of the crop [16]. Thus, identifying this group of stress responsive genes may lead to crop improvement in salt tolerance, which is known as a complex quantitative trait. Find this target genes is a complex task, because the function of many genes are still not understood and many novel non-coding genes have been discovered. Particularly rice (*Oryza sativa*), the major food source around the world, is highly sensitive to salt stress [6]. Therefore, identification of target genes in rice may allow biologist use them as a genetic resource to develop new cultivars with resistance to salinity.

We propose a methodology to identify stress responsive genes to salt conditions in rice. The methodology is based on Weighted Gene Co-expression Network Analysis (WGCNA). This is considered an effective and accurate bioinformatic method, using co-expression networks, that has been widely applied in identifying target genes for disease and cancer fields [20]. We follow the WGCNA workflow but with a new approach in the module detection step. Our modules are detected using the Hierarchical Link Clustering (HLC) technique [2] that allows the recognition of overlapping communities, which may have more biological meaning given the overlapping regulatory domains of systems that generate co-expression [9]. We conduct a systematic study with a large set of rice data using the proposed methodology. RNA-seq data was accessed through GEO database [1] (Accession number GSE98455), corresponding 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 the analysis result, 6 modules are detected as relevant in the response to salt stress in rice: 3 modules of 3 genes each one 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 those 19 genes, all but 3 genes (associated with *K* content), were also

identified as differentially expressed (with a log fold change value greater than 2) 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 with shoot biomass, are named and have an associated protein product: Spermidine hydroxycinnamoyltransferase 2 (SHT2) and Lipoxygenase. 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 a potential relevance in stress responsive mechanisms to salt conditions in rice.

With the development of high-throughput technologies, including microarrays and RNA sequencing (RNA-seq), genome-wide gene expression can be studied under different environmental stimuli (e.g. salt stress). Our methodology uses this kind of transcriptomic data measured for two different conditions (control and stress). After a process of normalization and filtering of the raw data, a differential expression profile of the genes is built calculating the log fold change (LFC) from control to stress condition. The LFC matrix will be the input for the co-expression network construction through the WGCNA method. A similarity matrix is calculated using the absolute value of Pearson's correlation coefficient between pairs of genes. Then, the similarity matrix is forced to be a scale-free network, finding a beta exponent such that by raising each entry of the matrix to that value, the probability distribution follows a power law. Next, unlike WGCNA, the scale-free network is used to detect overlapping rather than non-overlapping communities, using the HLC technique. We also implement a LASSO regression [21] to select the most significant modules associated with rice phenotypical responses to salt stress. Finally, for the genes found, we look for previous evidence of important biological implications in tolerance to salt stress. That is, the genes differentially expressed within the selected modules are enriched with gene ontology annotations from QuikGO database [4] and their interaction networks reported in STRING database [19] are reviewed.

The proposed methodology is modular, since other module detection and selection techniques could be used, instead HLC and LASSO respectively. 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 [7]. Additionally, LASSO is especially useful in problems where the number of variables is much larger than the number of samples, which is our case having more than 5000 modules (variables) and 92 accession (samples). The combinations 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. Furthermore, this study can be extended to other stresses and even to other crops.

I. PRELIMINARIES

A. Co-expression network

A network is an undirected graph $G = (V, E)$ where $V = \{v_1, v_2, \dots, v_n\}$ is a set of vertices or nodes and $E = \{e_1, e_2, \dots, e_q\}$ is a set of edges or links that connect the vertices. In a gene co-expression network each node corresponds to a gene and a pair of genes is connected if

they show similar differential expression patterns. These co-expression networks are of biological interest since the co-expressed genes are usually controlled by the same transcriptional regulatory pathway, functionally related or members of the same pathway or metabolic complex.

When the network is simple and unweighted, it can be represented by an adjacency matrix $A \in \{0, 1\}^{n \times n}$. This adjacency matrix 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.

B. Hierarchical Link Clustering

Hierarchical Link Clustering (HLC) algorithm was proposed by Ahn et al. [2]. HLC approach reinvent communities as groups of links rather than nodes, each node inherits all memberships of its links and can thus belong to multiple, overlapping communities. The algorithm 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 of node i and its neighbors.

With this similarity, the method uses single-linkage hierarchical clustering to build a dendrogram where each leaf is a link from the original network and branches represent link communities. Hierarchical clustering methods repeatedly merge groups until all elements are members of a single cluster. To find meaningful communities it is crucial to know where to partition the dendrogram. Thus, the most relevant communities are established at the maximal partition density D , a function based on link density inside communities that measures the quality of a link partition. The partition density has a single global maximum along the dendrogram in almost all cases, because the value is just 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). $D = 1$ when every community is a fully connected clique and $D = 0$ when each community is a tree. If a link community is less dense than a tree (when the community subgraph has disconnected components), then that community will give a negative contribution to D . Computing D at each level of the link dendrogram allows to pick the best level to cut (although meaningful structure exists above and below that threshold). This way, the final output corresponds to node clusters, where each node can participate in multiple communities.

C. Least Absolute Shrinkage Selector Operator

Least Absolute Shrinkage Selector Operator (LASSO) is a regularized linear regression technique, a method that 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. Very useful in problems where the number of variables (genes) n is much

greater than the number of samples m ($n \gg m$). Lasso solves the least squares problem with restriction on the L_1 -norm of the coefficient vector:

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

Or equivalently minimizing:

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

being $s, \lambda \geq 0$ the respective penalty parameters for complexity.

Since the λ 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. With that selected λ value, the model is adjusted again, using all the observations.

II. METHODOLOGY

This methodology uses as input data RNA-seq read counts, representing gene expression levels. More precisely, 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 can be 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 methodology also requires a set of p phenotypic traits, measured for the m genotypes. The phenotypic data can be seen as a matrix $P \in \mathbb{R}^{2m \times p}$ containing two phenotypic values per genotype, one under control condition and the second under treatment condition.

A. Data pre-processing

The RNA-seq data cannot be directly interpreted, therefore a normalization process has to be done to deal with the various biases that affect quantification results. In order to correct library size and RNA composition bias, the suggested normalization technique is DESeq2 [13]. From the normalized data $D_1 \in \mathbb{R}^{n_0 \times 2mr}$, average the biological replicates of each genotype, so that the data matrix becomes $D_2 \in \mathbb{R}^{n_0 \times 2m}$. Next, remove genes exhibiting low variance or low expression, getting a new number of $n_1 \leq n_0$ genes. Separate control and treatment data into two matrices $C \in \mathbb{R}^{n_1 \times m}$ and $T \in \mathbb{R}^{n_1 \times m}$, respectively. Each matrix entry, c_{ij} for control and t_{ij} for treatment, represent the normalized expression level of gene i in accession j .

Now, compile the control and treatment data by measuring the changes in expression levels in terms of logarithmic ratios. Matrix $L_0 \in \mathbb{R}^{n_1 \times m}$, known as the Log Fold Change matrix, is computed by setting $\ell_{ij} = \log_2(t_{ij}/c_{ij})$. Repeat this process on the phenotypic data separating the control and treatment data into matrices P_c and P_t of dimensions $m \times p$. Calculate the

corresponding logarithmic ratios, obtaining the matrix $P_\ell \in \mathbb{R}^{m \times p}$. Finally, filter L_0 matrix by removing rows (genes) with low variance in the differential expression patterns, getting a new matrix L_1 of dimensions $n_2 \times m$, with $n_2 \leq n_1$.

B. Co-expression network construction

The Log fold change matrix L_1 is used to build the co-expression network following the first steps of the WGCNA methodology. First, measure the level of concordance between gene differential expression profiles across samples. Use the absolute value of the Pearson correlation coefficient as the similarity measure between genes and store the values in the similarity matrix $S \in \mathbb{R}_+^{n_2 \times n_2}$.

Then, transform S 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 the degree distribution will fit a scale-free network. This kind of 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 $P(k)$ (the probability of a node to have degree k) is approximately inversely proportional to the logarithm of k (the degree of a node). So the parameter β is chosen as the smallest value of β such that the R^2 of the linear regression between $\log_{10}(p(k))$ and $\log_{10}(k)$ is close to 1 (e.g. $R^2 > 0.85$).

Finally, prepare matrix A to apply the clustering algorithm, transforming it into an unweighted network $\hat{A} \in \{0, 1\}^{n_2 \times n_2}$. To determine the Pearson Correlation Coefficient (PCC) cutoff for finding biologically relevant co-expressed modules, we use the approach described by [3] based on density of the network combined with decreasing number of nodes and edges with higher PCC values. Above the cutoff, the entries of matrix A become 1 and when the value is below the threshold it becomes zero.

C. Co-expression module identification

Next step in the methodology is to study the co-expression network structure and dynamics identifying communities also called modules. The idea is to cluster genes with similar differential expression change patterns. Membership in these modules may overlap in biological contexts, where modules may be related to specific molecular, cellular or tissue functions and the biological components (i.e. genes) are involved in multiple functions. Thus, unlike WGCNA, the adjacency matrix \hat{A} , from the co-expression network G , is used to detect overlapping rather than non-overlapping communities, using the Hierarchical Link Clustering (HLC) algorithm proposed in [2].

The HLC algorithm organizes the n_2 genes of matrix \hat{A} into c communities, where each gene can belong to one, multiple, or no community. This information is represented as an affiliation matrix $F \in \{0, 1\}^{n_2 \times c}$, where $f_{iu} = 1$ if node i is member of cluster u .

D. Modules association to phenotypic traits

To identify the most relevant gene groups (modules), associated with the phenotypic response to a specific treatment in an organism, we use a LASSO based approach. Each module can be represented by a eigengene, which is defined as the first principal component of such module. A eigengene can be thought of as an average differential expression profile for each community and is computed from the Log Fold Change Matrix L_1 and the affiliation matrix F . Given a module u , use the affiliation matrix to identify the genes belonging to u and then select the corresponding rows of the matrix L_1 to compute the first principal component of u . Each principal component would be a column of the matrix $M \in \mathbb{R}^{m \times c}$.

These profiles are then associated with each phenotypic trait using the least absolute shrinkage and selection operator (LASSO). LASSO combines a regression model with a procedure of contraction of some parameters towards zero, imposing a restriction or a penalty on the regression coefficients. This technique allows selection of the most relevant variables. In our context, the eigengenes (the columns of M) act as regressor variables and each phenotypic trait (each column of P_ℓ) is used as an outcome variable.

The output after applying LASSO is a set of modules for each phenotypic trait. The union of genes belonging to the selected modules are the target genes for downstream analysis.

E. Genes enrichment

For the genes found, first identify the differential expressed ones, that is genes showing an absolute value of the log fold change greater than 2 ($|\ell_{ij}| \geq 2$) for at least one sample. This represents genes whose level of expression is quadrupled (up or down) from control to treatment condition, suggesting that those genes are strong candidates as treatment responsive genes.

You also can perform a functional category enrichment searching for the Gene Ontology (GO) annotations from databases like QuikGO [4]. This annotations can provide evidence of biological implications of the target genes in the treatment-tolerance mechanisms.

Those named genes (with GO annotations), can be used to performe another relevant analysis reviewing their reported protein-protein interaction networks using the STRING database [19]. The interactions include direct (physical) and indirect (functional) associations; they stem from computational prediction, from knowledge transfer between organisms, and from interactions aggregated from other (primary) databases. This information elucidates how the selected genes are involved in functional pathways that can be related with the treatment of interest.

III. CASE STUDY

RNA-seq data was accessed through GEO database [1] (Accession number GSE98455), corresponding to $n_0 = 57845$ 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. On the other hand, we use $p = 3$ phenotypic traits: shoot K^+ content, root biomass

and shoot biomass. This traits were measured for the same 92 genotypes, under control and salt stress conditions, and can be found in the supplementary information of [5].

A. Data pre-processing

The DESeq normalization was applied to the raw data and the biological replicates were averaged. Genes exhibiting low variance were identified as those with ratio of upper quantile to lower quantile smaller than 1.5, and were removed. Genes with low expression were also removed, corresponding to those having more than 80% samples with values smaller than 10. After this filtering process, $n_1 = 8928$ genes were kept.

From the compiled matrix L_0 we remove genes whose difference between upper quantile and lower quantile is greather than 0.25. So the resulting matrix L_1 contains the log ratios of $n_2 = 8928$ genes. We also compute the logarithmic ratios of the phenotypic data, for the 92 accessions and the 3 traits.

B. Co-expression network construction

We used the Log Fold Change matrix L_1 to compute the corresponding similarity matrix. We found that for this network, the smallest integer such that the $R^2 \geq 0.8$ is $\beta = 3$. Figure 1 shows 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 forced network with $R^2 = 0.8$ corresponding to $\beta = 3$.

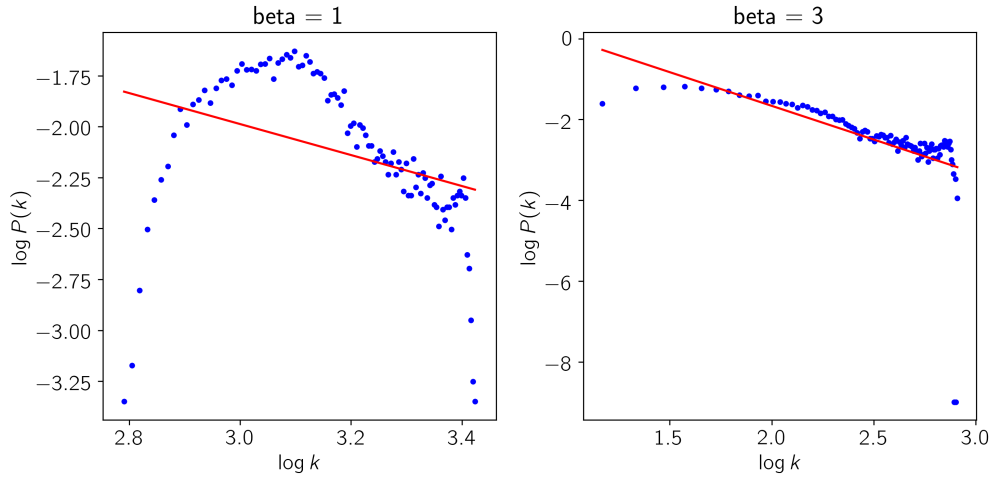


Figure 1: Degree distributions

The resulting adjacency matrix A represents a complete graph $G = (V, E)$, with $|V| = 8928$ genes (nodes) and $|E| = 39850128$ edges. We choose 0.2 as cutoff value, keeping only the connections above this threshold and removing the isolated nodes, so the new adjacency matrix \hat{A} has $n_3 = 5810$ genes and accounts for 16875145 edges.

C. Co-expression module identification

After applying the HLC algorithm, a total of 4131 genes were distributed in $c = 5143$ overlapping modules of 3 or more genes. Figure 2 shows a histogram of the overlapping percentage of these genes, measured as the proportion of modules to which each gene belongs.

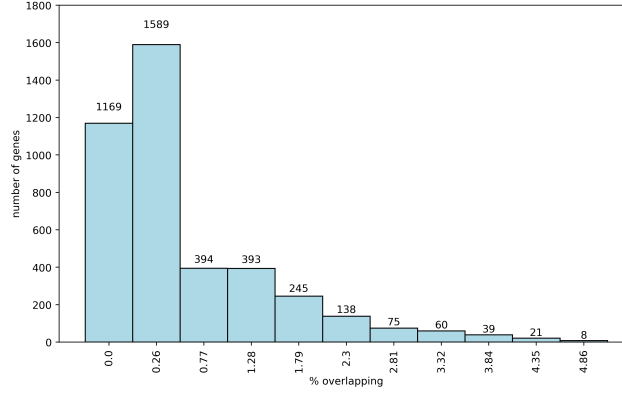


Figure 2: Overlapping percentage

D. Modules association to phenotypic traits

The phenotypic traits under study are shoot K^+ content, root biomass and shoot biomass. Figure 3 shows that there are significant differences in the values of these phenotypic traits between stress and control conditions. This supports the idea that these variables represent tolerance-associated traits in rice under salt stress.

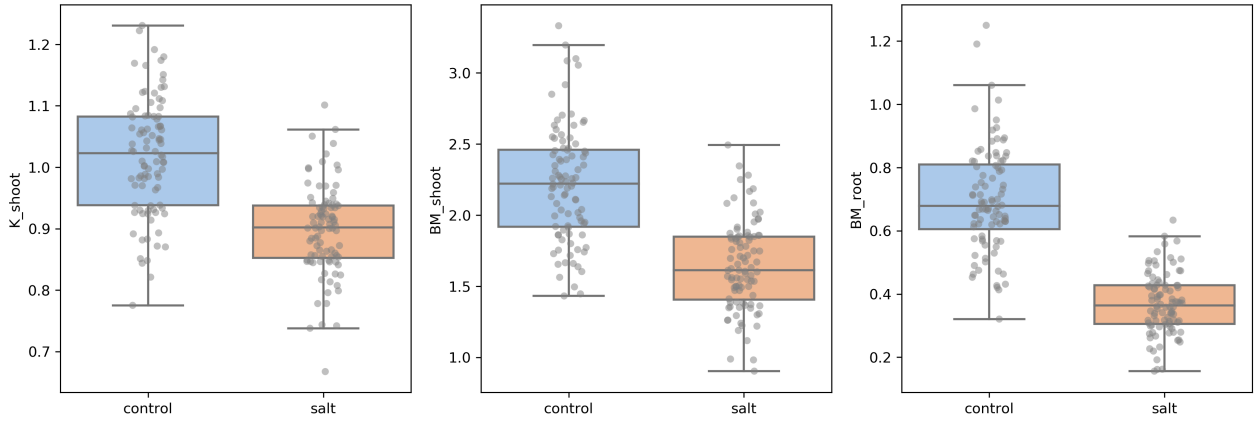


Figure 3: Phenotypic traits distribution under control and salt stress

Using the affiliation matrix F derived from HLC output and the Log Fold Change matrix L_1 , we build the matrix M computing the eigengene for each of the $c = 5143$ modules. We apply the LASSO technique using each of the phenotypic traits as the outcome variable, one at a time. As shown in Figure 4 we perform cross-validation for each phenotypic trait, in order to select the corresponding regularization parameter λ that minimizes the mean-squared error.

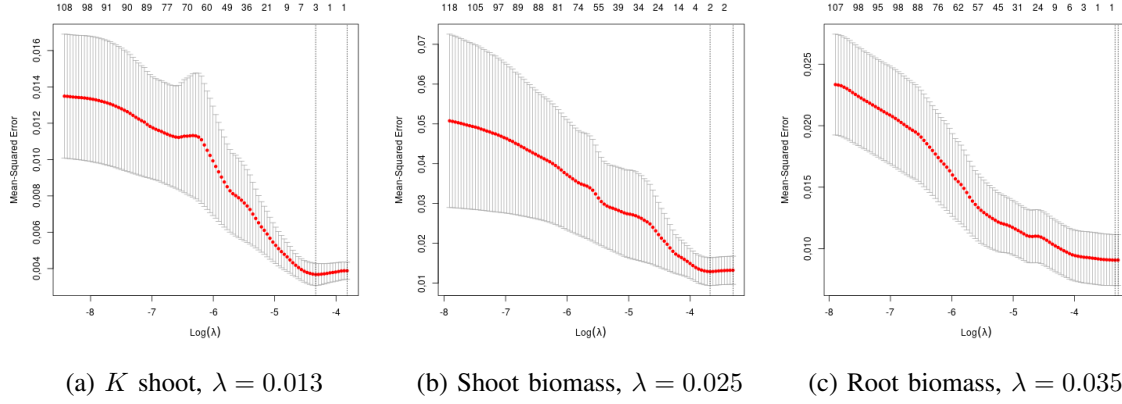


Figure 4: Cross-validation of the LASSO regularization parameter λ , for each phenotypic trait.

We finally adjust the three models using all observations and the corresponding λ values. As result, 6 modules were detected as relevant in the response to salt stress in rice: 3 modules of 3 genes each one associated with shoot K content, 2 modules of 3 genes associated with shoot biomass, and 1 module of 4 genes associated with root biomass.

E. Genes enrichment

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

According to the Quickgo database, only 2 of the 16 differentially expressed genes, both from the module related with shoot biomass, are named and have an associated protein product: Spermidine hydroxycinnamoyltransferase 2 (SHT2) and Lipoxygenase. Figure 5 shows their corresponding 3D protein structure.

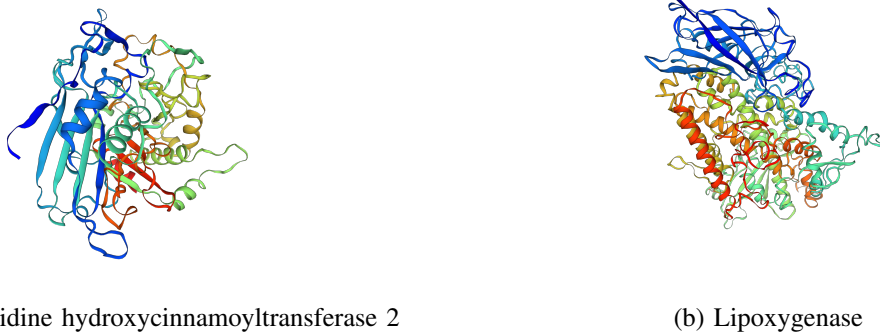


Figure 5: 3D protein structure of named genes selected by LASSO.

On the one hand, SHT2 contributes to the natural variation of spermidine-based phenolamides in rice cultivars. On the other hand 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.

Previous studies ([10], [11], [14], [15], [17]) provide evidence of biological implications of spermidine and lipoxygenase in tolerance to salt stress in other plants or even in rice cultivars. However, 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 a potential relevance in stress responsive mechanisms to salt conditions in rice.

IV. DISCUSSION

The methodology described in this manuscript provide an approach to discover key genes responding to a specific treatment in an organism, linking transcriptomic with phenotypic data, and identifying overlapping gene modules. Similar approximations have been proposed ([12], [8]), but they only focus on disjoint (non-overlapping) communities. However, it is well known that communities in real networks are overlapping.

The methodology was applied in a case study with rice under salt stress. The results show a group of 14 genes where two of them are related to the response to saline stress according to previous studies, validating the ability of the method to detect this kind of key genes.

An advantage of the proposed methodology is its modularity. As future work, other module detection and selection techniques could be used, instead HLC and LASSO respectively. The combinations 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. Furthermore, this study can be extended to other stresses and even to other crops.

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., AND LEHMANN, S. Link communities reveal multiscale complexity in networks. *nature* 466, 7307 (2010), 761–764.
- [3] AOKI, K., OGATA, Y., AND SHIBATA, D. Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant and Cell Physiology* 48, 3 (2007), 381–390.
- [4] BINNS, D., DIMMER, E., HUNTLEY, R., BARRELL, D., O'DONOVAN, C., AND APWEILER, R. Quickgo: a web-based tool for gene ontology searching. *Bioinformatics* 25, 22 (2009), 3045–3046.
- [5] 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 (2017), e1006823.
- [6] CHANG, J., CHEONG, B. E., NATERA, S., AND 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 (2019), 427–435.
- [7] DESBOULETS, L. D. D. A review on variable selection in regression analysis. *Econometrics* 6, 4 (2018), 45.
- [8] DU, Q., CAMPBELL, M., YU, H., LIU, K., WALIA, H., ZHANG, Q., AND ZHANG, C. Network-based feature selection reveals substructures of gene modules responding to salt stress in rice. *Plant direct* 3, 8 (2019), e00154.
- [9] GAITERI, C., DING, Y., FRENCH, B., TSENG, G. C., AND 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 (2014), 13–24.
- [10] GUPTA, K., DEY, A., AND GUPTA, B. Plant polyamines in abiotic stress responses. *Acta physiologiae plantarum* 35, 7 (2013), 2015–2036.

- [11] HOU, Y., MENG, K., HAN, Y., BAN, Q., WANG, B., SUO, J., LV, J., AND 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 (2015), 1073.
- [12] LANGFELDER, P., AND HORVATH, S. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics* 9, 1 (2008), 559.
- [13] LOVE, M. I., HUBER, W., AND ANDERS, S. Moderated estimation of fold change and dispersion for rna-seq data with *deseq2*. *Genome biology* 15, 12 (2014), 550.
- [14] MITTOVA, V., TAL, M., VOLOKITA, M., AND 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 (2002), 393–400.
- [15] 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 (2019), 1–13.
- [16] REDDY, I. N. B. L., KIM, B.-K., YOON, I.-S., KIM, K.-H., AND KWON, T.-R. Salt tolerance in rice: focus on mechanisms and approaches. *Rice Science* 24, 3 (2017), 123–144.
- [17] ROYCHOUDHURY, A., BASU, S., AND 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 (2011), 317–328.
- [18] SHRIVASTAVA, P., AND 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 (2015), 123–131.
- [19] 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* (2016), gkw937.
- [20] TIAN, H., GUAN, D., AND LI, J. Identifying osteosarcoma metastasis associated genes by weighted gene co-expression network analysis (wgcna). *Medicine* 97, 24 (2018).
- [21] TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58, 1 (1996), 267–288.