Identification of rice genes which respond to saline stress from co-expression networks analysis [1]

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 [11]. 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 [10]. 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 [13]. 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 [8]. 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 deferentially expressed ($|LFC| > 2$) for at least one of the 92 accessions, suggesting that those genes are strong candidates as stress responsive genes. Only 2 of the 16 diferentially 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 trough 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 [14] 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 deferentially expressed within the selected modules are enriched with gene ontology annotations from QuikGO database [4] and their interaction networks reported in STRING database [12] 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. Methodology

This methodology uses as input data RNA-seq read counts, representing gene expression levels. More precisely, n gene expression profiles of an organism, measured for m different genotypes under control and treatment conditions, and r biological replicates. This data can be represented as a matrix $D = [d_{ij}]$, where $d_{ij} \in \mathbb{N}_0$ for all $i = 1, 2, ..n$ and $j = 1, 2, .., 2mr$. To discover

key genes and their interaction for treatment-tolerance related phenotypes, the methodology also requires a set of p phenotypic traits, measured for the m genotypes. The phenotypic data can be seen as a matrix $P = [p_{jk}]$, where $p_{jk} \in \mathbb{R}$ for all $k = 1, 2, ..., p$.

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 [9]. Then, from the normalized data, average the biological replicates of each accession and remove genes exhibiting low variance or low expression. Next, separate control and salinity stress treatment data into two matrices $C = [c_{ij}]_{n\times m}$ and $T = [t_{ij}]_{n\times m}$, respectively, where c_{ij} and t_{ij} represent normalized expression level of the gene i in the accession j.

Next step is to compile the control and treatment data by measuring the changes in expression levels in terms of logarithmic ratios. Matrix $L = [\ell_{ij}]_{n\times p}$, known as the Log Fold Change matrix, is computed by setting $\ell_{ij} = \log_2(t_{ij}/c_{ij})$. Finally, filter L matrix by removing rows (genes) with low variance in the differential expression patterns.

B. Co-expression network construction

The Log fold change matrix L 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 $S = [s_{ij}]_{n\times n}$.

Then, transform S into an adjacency matrix $A = [a_{ij}]_{n\times n}$ where $a_{ij} = (s_{ij})^\beta$ 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 $\hat{A}$. 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.

The matrix $\hat{A}$ can be seen as an undirected graph $G = (V, E)$ where $V = \{v_1, v_2, \ldots, v_n\}$ is a set of vertices or nodes and $E = \{e_1, e_2, \ldots, e_p\}$ is a set of edges or links that connect the vertices. In this genomic context, the graph G represents a co-expression network, where each node corresponds to a gene and a pair of genes is connected if they show similar differential expression patterns.

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 co-expression network G is used to detect overlapping rather than non-overlapping communities, using the Hierarchical Link Clustering (HLC) algorithm proposed in [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. They compute the similarity between links using the Jaccard index for unweighted networks. With this similarity, they use single-linkage hierarchical clustering to build a dendrogram where each leaf is a link from the original network and branches represent link communities. Finally, the most relevant communities are established at the maximal partition density, a function based on link density inside communities.

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 rows corresponding to the genes belonging to a specific module.

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 eigenges act as regressor variables and each phenotypic trait its 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 looking 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 [12]. 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.

II. Case study

RNA-seq data was accessed through GEO database [1] (Accession number GSE98455), corresponding to $n = 57845$ gene expression profiles of shoot tissues measured for both control and salt condition in $m = 92$ accessions of the Rice Diversity Panel 1, with two biological repetitions ($r = 2$).

Genes exhibiting low variance (the ratio of upper quantile to lower quantile smaller than 1.5) or low expression (more than 80% samples with values smaller than 10) were removed. The final network contains 8928 genes of the initial 57845.

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 $\beta = 3$.

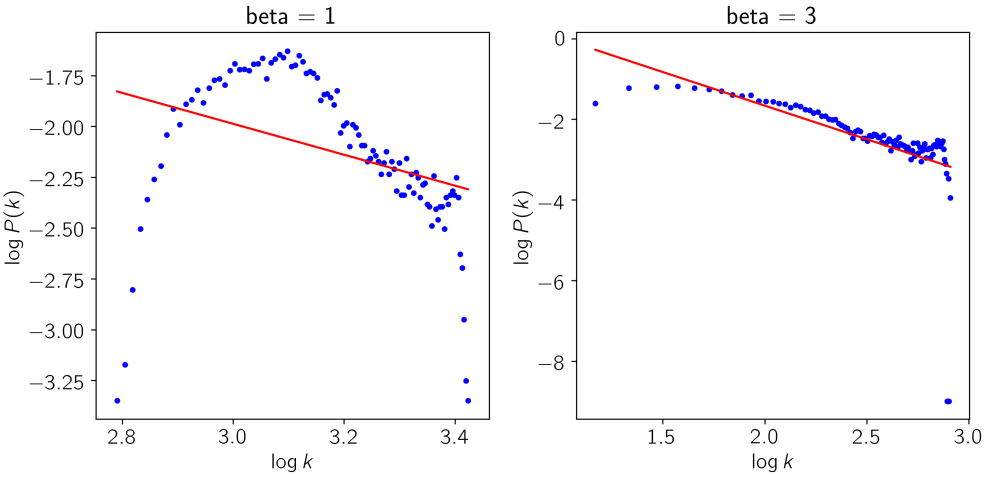

Figure 1. Degree distributions

The current network represented by the adjacency matrix A, corresponds to a complete and weighted network of 8928 genes (nodes) and 39850128 edges. For computational reasons, this network was transformed into an unweighted one $\hat{A}$, keeping only the connections above the cutoff value of 0.2. The resulting unweighted network has a total of 5810 nodes and 16875145 edges. After applying the HLC algorithm, a total of 4131 genes were distributed in 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.

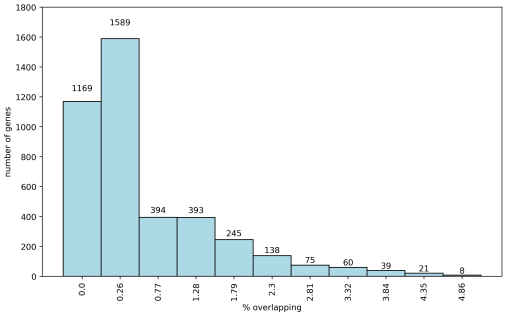

Figure 2. Overlapping percentage

As shown in Figure 3, there are significant differences in the values of these phenotypic traits between both stress and control conditions. This supports the idea that these variables represent tolerance-associated traits in rice under salt stress.

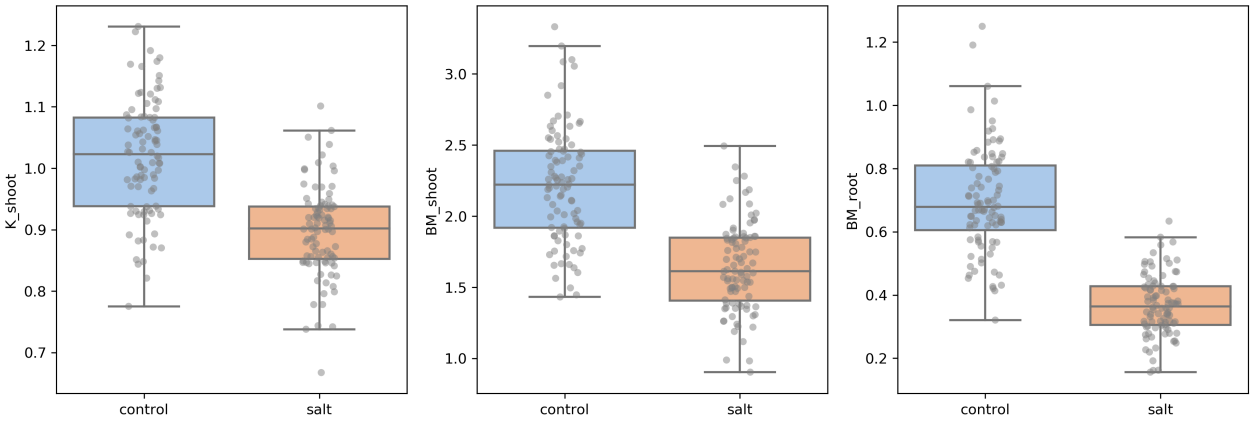

Figure 3. Phenotipic traits distribution under control and salt stress

We use 3 phenotypic traits: shoot K^+ content, root biomass and shoot biomass. These were measured for the 92 genotypes studied, under control and salt stress conditions. These data can be found in the supplementary information of [5].

Finally, 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. From those 19 genes, all but 3 genes (associated with K content), were also identified as deferentially expressed ($|LFC| > 2$) for at least one of the 92 accessions. Only 2 of the 16 diferentially expressed genes, both from the module related with shoot biomass, are named and have an associated protein product: Spermidine hydroxycinnamoyltransferase 2 (SHT2) and Lipoxygenase.

III. Discussion

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] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] Tian, H., Guan, D., and Li, J. Identifying osteosarcoma metastasis associated genes by weighted gene co-expression network analysis (wgcna). *Medicine 97*, 24 (2018).

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

APPENDIX

A. A: Variable selection with LASSO

LASSO (Least Absolute Shrinkage Selector Operator) 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 p ($n \gg p$). Lasso solves the least squares problem with restriction on the L_1-norm of the coefficient vector:

$$\min \left\{ \sum_{i=1}^{p} \left(y_i - \sum_{j=1}^{n} \beta_j x_{ij} \right)^2 \right\}, \text{sujeto a} \sum_{j=1}^{n} |\beta_j| \leq s \tag{1}$$

Or equivalently minimizing:

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

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, this time using all the observations.

In the module gene selection context, the outcome variable correspond to the phenotypic trait, whereas the predictors are the modules, detected by the hierarchical clustering, represented by the first principal component of the module. After running the Lasso regression will select the most significant modules associated with phenotypic data.