Building Differential Co-expression Networks with Variable Selection and Regularization

Camila Riccio(✉), Jorge Finke, and Camilo Rocha

Pontificia Universidad Javeriana, Cali, Colombia
criccio35@gmail.com

Abstract. This work introduces a technique for the inference of differential co-expression networks. The approach takes as input a matrix of differential expression profiles, where each entry corresponds to the Log Fold Change of a gene expression between control and stress conditions for a specific sample. It outputs a matrix of coefficients, where each non-zero entry represents a pairwise connection between genes. The proposed approach builds on Lasso, and is applied to differential expression profiles of rice between control and salt-stress conditions. A total of 25 genes were identified to respond to salt stress and as differentially expressed. About half of these genes (11) were reported with a statistically significant number of different GO annotations relevant to salt stress response.

Keywords: Network inference · Lasso-based inference · Overlapping clustering · Salt-stress · Rice · *Oryza sativa*

1 Introduction

Gene regulatory networks specify how biological systems respond to perturbations through the rewiring of molecular interactions. Co-expression networks provide a framework to better understand molecular mechanisms and gene regulation. Thanks to the increasingly high availability of transcriptomic data, robust gene co-expression networks are becoming more widely available. A differential co-expression network represents a particular type of network, which is used to identify changes in response to external stimuli (e.g., changes in activity of gene expression regulators or signaling [3,7,28]). Differential co-expression network analysis is an approach for identifying modules of genes with meaningful variation between different experimental conditions (e.g., control and stress). Such an analysis uses as input gene expression data, representing gene expression between control and stress conditions, and output a set of genes that are likely to be involved in the biological response to the specific stress.

Technically, differential co-expression analysis builds a network from the relationships between genes differential expression profiles, that is, the log fold

H. Cherifi et al. (Eds.): COMPLEX NETWORKS 2016 2022, SCI 1077, pp. 277–288, 2023.
https://doi.org/10.1007/978-3-031-21127-0_23

change (LFC) between control and stress expression across multiple observations for each gene. LFC represents a logarithm of the ratio between the control expression and the expression under stress and involves two key steps. First, setting a correlation measure between the genes. Second, filtering the list of pairs using a threshold value for the correlation score [29]. The Pearson correlation coefficient is the most popular measure used for step 1. It assumes linear correlation, normally distributed values, and is sensitive to outliers [17]. A major limitation of using this approach, for building a differential co-expression network, is that it demands large enough sample sizes for statistically reliable results [6,20,22]. However, large samples sizes are often prohibitive due to time and computational constraints. For instance, a differential co-expression network with N genes across M control samples and M stressed samples is built by first building a matrix $X \in \mathbb{R}^{N \times M}$, where each entry $X(n, m)$ corresponds to the LFC value of gene n for sample m. Then, the Pearson's coefficient is computed for each pair of distinct genes n_1 and n_2 with the input vectors $X(n_1, \_)$ and $X(n_2, \_)$, each of length M. That is, assuming as input X, building the differential co-expression network takes $O\left(\binom{N}{2} M\right) = O(N^2 M)$ time, where $O(M)$ is the usual time for computing the Pearson's coefficient on M samples.

This paper introduces a method based on the penalized least absolute shrinkage and selection operator (Lasso) [33] for building differential co-expression networks. It overcomes the above-mentioned computational limitation by computing a regression of the differential expression profile of one gene against all others, instead of using a correlation metric to identify significant edges between genes. The resulting coefficients represent the strength of the relationship between the corresponding pair of genes, where zero strength indicates no edge between them. Additionally, Lasso has the advantage of yielding accurate parameter estimates even with small sample sizes [13]. For N genes and M samples (both under control and stress), building the differential co-expression network with Lasso takes $O(NM^2)$ time. However, for the case where there are many more variables than observations (i.e., $N \gg M$), as the usual biological expression datasets, it takes $O(NM)$ time [11]. While the proposed approach is used in this work for building *differential* co-expression networks, it can also more generally be used for building co-expression networks.

Lasso simultaneously performs variable selection and regularization by forcing the least significant coefficients to be zero, which naturally favors the inference of a sparse network. It performs ℓ_1 regularization by forcing the sum of the absolute value of the regression coefficients to be less than a fixed value. Lasso iteratively searches for a degree of penalty λ that minimizes the mean square error of the regression. At the optimal value of λ, it performs variable selection, which results in a reduced number of non-zero coefficients. The variables with a zero value coefficient are excluded from impacting the regression, which prevents the model from over-fitting. Lasso properties are particularly useful in the construction of co-expression networks since these type of networks are expected to be sparse [36]. Moreover, it avoids the need to define a threshold for selecting meaningful edges from the pairwise relationships between genes.

Finally, since Lasso supports small sample sizes, it is well suited for most typical transcriptomic data sets.

The proposed approach is used in combination with a slightly modified version of the workflow proposed in [26] (but with a different module detection algorithm to identify genes that respond to salt stress, namely, the ANGEL algorithm [27]). RNA-seq data was accessed from the GEO database [8], accession number GSE98455. It represents 57,845 gene expression profiles of shoot tissues measured under control and stress conditions in 92 accessions of the Rice Diversity Panel 1 [12]. A total of 25 genes are identified to respond to salt stress and as differentially expressed genes (DEG). About half of these genes (11) are reported with a statistically significant number of different GO annotations relevant to salt stress response.

2 Methodology

This section presents a description of differential co-expression networks, introduces the Lasso-based approach to build differential co-expression networks, summarize the Angel algorithm to detect overlapping modules in the network, and explains an enrichment technique to evaluate the biological significance of the detected modules.

2.1 Differential Co-expression Network

A *network* is an undirected graph $G = (V, E)$ where $V = \{v_1, v_2, \ldots, v_N\}$ is a set of N *vertices* (or *nodes*) and $E = \{e_1, e_2, \ldots, e_Q\}$ is a set of Q *edges* (or *links*) between vertices. $G = (V, E)$ 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. *Differential co-expression* is the altered co-expression patterns of genes between two particular conditions (e.g., control and stress). In a differential co-expression network, each vertex corresponds to a gene. A link indicates a common alteration in the expression pattern between two genes when changing from one condition to the other. Differential co-expression networks are of biological interest because adjacent nodes in the network represent genes that jointly respond to similar stress conditions.

2.2 Co-expression Network Construction with Lasso

Lasso regressions [33] can be seen an advantageous approach for the Constructing co-expression networks using Lasso [33] offers key advantages compared to other methods that rely on the Pearson correlation coefficient (or any other correlation metric). Several assumptions behind computing Pearson limit effectiveness [17] and statistical significance especially if sample sizes are small [6]. Such condition is common across numerous efforts to build co-expression networks [20,22]. The majority of typical transcriptome datasets tend to be small in terms of the

number of samples. Differential co-expression networks are built using differential expression profiles instead of the expression profiles themselves as in co-expression networks. This section explains build differential co-expression networks using Lasso can lead to robust networks even in the presence of small sample sizes.

Furthermore, note that building a network $G = (V, E)$, that is, a representation of pairwise relationships E over a set of vertices V, is equivalent to inferring a neighborhood for each vertex (i.e., the set of vertices to which it is connected). Given M observations on N genes (vertices) represented in a data matrix $X \in \mathbb{R}^{N \times M}$, the set of neighbors of vertex $v_i \in V$, denoted,

$$V(v_i) := \{v_j : (v_i, v_j) \in E\}$$

is inferred by regressing x_i against all other variables

$$x_{\setminus i} := [x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N]^T \in \mathbb{R}^{N-1}.$$

The result is a matrix $B \in \mathbb{R}^{N \times N}$ whose diagonal is zero and the remaining $N-1$ entries of a row i correspond to the coefficients of the regression of x_i against $x_{\setminus i}$. Each entry $B(i, j)$ represents the strength of the relationship between vertices v_i and v_j, where zero strength indicates no connection.

For each variable x_i the regression problem has the form:

$$\underset{\beta_i}{\text{minimize}} \left\| X_i - X_{\setminus i}\beta_i \right\|_2^2 + \lambda \left\| \beta_i \right\|_1, \tag{1}$$

where X_i and $X_{\setminus i}$ represent the observations on x_i (i.e., the transpose of the first row of X) and the rest of the variables, respectively. The vector $\beta_i \in \mathbb{R}^{N-1}$ is a vector of coefficients for x_i. In Eq. 1, the first term can be interpreted as a local log-likelihood of β_i and the ℓ_1 penalty is added to enforce sparsity. The regularization parameter λ balancing the two terms. Lasso is repeated for all the variables leading to a set of $N \times N$ coefficients that are computed from $\beta_1, \ldots, \beta_N$. Note that there is no guarantee that $B(i, j) \neq 0$ implies $B(j, i) \neq 0$. Therefore, the information in $V(v_i)$ and $V(v_j)$ is combined to enforce symmetry: an edge (v_i, v_j) is meaningful, if $B(i, j)$ and $B(j, i)$ are both non-zero.

Note also that including the ℓ_1 penalty allows Lasso to identify the variables that are strongly associated with the response variable (i.e., variable selection). Since the value of the regularization parameter λ determines the degree of penalty and the accuracy of the model, cross-validation is used to select a regularization parameter that minimizes the mean-squared error. If the degree of penalty λ is equal to zero, the solution is the same as least-squares (LS) linear regression [5]. For larger values of λ, larger number of coefficients are shrunk towards zero. Compared to LS, Lasso offers the following advantage. Unlike LS, Lasso does not yield non-zero estimates, which would results in a fully connected network, and giving rise to the problem of setting a threshold above which and edge is considered significant. Lasso avoids this additional step as it simultaneously performs parameter estimation and variable selection by forcing the least significant coefficients to zero through the ℓ_1 penalty.

2.3 Overlapping Clustering with ANGEL

ANGEL [27] is a static node-centric algorithm for detecting potentially overlapping modules in networks. It takes as input a graph G, a merging threshold ϕ, and an empty set of communities C. The algorithm's main loop cycles over each node, extracts the corresponding ego-minus-ego network, and computes the local communities it contains using Label Propagation (LP) [23]. During LP, every node is initialized with a unique label. In following steps, each node adopts the current label of the majority of its neighbors. In case of bow-tie situations, the classic LP formulation randomly selects a single label for the contended node. Here however soft community memberships are allowed, that is, each node can belong to multiple communities for the case of a bow-tie configuration. Once the outer loop on the network nodes is completed, the algorithm compacts the community set to avoid the presence of fully contained communities.

Finally, note that compared to HLC [1], the computational cost of ANGEL is significant less and can be approximated by $O(|V|)$.

2.4 Functional Enrichment

Our analysis of differential co-expression networks relies on the detection of gene modules. Such modules are used to investigate relationships occurring between genes performing similar biological functions [34]. Functional enrichment of each module is a critical step to understand the underlying processes contributing to phenotype or stress responses. This section describes how to evaluates the quality of the modules using Gene Ontology (GO) [2,9] enrichment.

Given a gene module, an enrichment analysis finds which GO terms are over-represented or under-represented by using annotations for that gene set. Gene module enrichment analysis is performed using the Fisher's Exact Test [32] in combination with a robust False Discovery Rate (FDR) [19] correction for multiple testing. Fisher's exact test is a statistical significance test used in the analysis of contingency tables. The FDR control is a statistical method used in multiple hypothesis testing to correct for multiple comparisons. In a list of statistically significant findings, FDR is used to control the expected proportion of incorrectly rejected null hypotheses ("false discoveries"). Here, a Benjamini-Hochberg correction is used [4]. The result for each module is a list of statistically significant GO terms ranked by their adjusted p-values.

For each GO term in each module, a contingency table is built (see Table 1). The hypothesis statement is the following for each module:

H_0: The module is a random sample from network.
H_1: The module has more genes with the GO term than expected by chance.

Following the configuration in Table 1, the p-value is calculated:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{N}{a+c}} = \frac{\binom{a+b}{b}\binom{c+d}{d}}{\binom{N}{b+d}} \tag{2}$$

Table 1. Contingency table configuration for each module.

	Genes in module	**Genes out the module**	**Total**
Annotated genes	a	b	$a+b$
Not annotated genes	c	d	$c+d$
Total	$a+c$	$b+d$	N

The p-value represents the probability (or chance) of seeing at least a genes out of the total $a+c$ genes in the module annotated with a particular GO term, given the proportion $(a+b)/N$ of genes in the whole genome that are annotated with that GO term. That is, the GO terms shared by the genes in each module are compared to the background distribution of the annotations. The closer the p-value is to zero, the more significant is the association of the particular GO term with the module of genes (i.e., the less likely that the observed annotation of the GO term to the module occurs by chance). In other words, if all of the genes in a module were associated with, say "DNA repair", this term would be optimally significant. However, since all genes in the genome (with GO annotations) are indirectly associated with the top level term "biological process", this would not be significant if all the genes in a module were associated with this high-level term.

If a module has at least one GO term with a significant p-value, the module is said to be enriched. This binary classification of a module between enriched and non-enriched allows us to evaluate, in a general way, the biological significance of the modules. The higher the proportion of enriched modules in a differential co-expression network, the better they capture the biological interactions of genes that jointly respond to a specific stress condition.

3 Case Study

This section presents a case study in the identification of genes that respond to saline stress in rice. The differential co-expression network is built using the approach presented in Section 2.2 and the framework in [26], with the ANGEL module detecting algorithm. The goal of this case study is to evaluate whether the proposed approach, in addition to being less computationally costly, finds a number of differentially expressed genes (DEG) and genes with GO annotations relevant to salt stress that is statistically significant.

3.1 Association Network Construction with Lasso

Consider the input data $X \in \mathbb{R}^{N \times M}$, which represents the matrix of differential expression profiles (corresponds to L_1 in [26]). The matrix X results from pre-processing RNA-seq data of rice both under control and salt stress conditions (GEO database [8] accession number GSE98455). Therefore, X contains the LFC of $N = 8,929$ genes under control and salt stress conditions in $M = 92$ samples.

The differential co-expression network, inferred using Lasso regressions, is composed of $|V| = 7,474$ vertices and $|E| = 67,061$ edges. All the genes in this network are part of the network previously constructed in [26] and preserves 21,123 out of 39,850,128 of the original connections. IN other words, the resulting network is a subnetwork of the one constructed in [26].

3.2 Identification of Co-expression Modules

In [26], the approach for module detection requires finding a threshold for Pearson correlations to define the adjacency matrix. Here the proposed Lasso-based approach bypasses this step since it is able to directly infer network connections without additional parameters.

The ANGEL algorithm distributes a total of 5,577 genes across 1,462 modules with at least 3 genes each. Using the module enrichment technique described in Section 2, a total of 65% of all modules are identified as enriched, meaning that they have some over-represented GO annotations. In other words, the modules identified by ANGEL are biologically relevant. Figure 1 compares the threshold Pearson network (thP) in [26] with Lasso-based network (nbL), both in terms of the proportion of enriched modules and gene overlaps. Regarding module enrichment, note that the proposed approach surpasses the approach of [26].

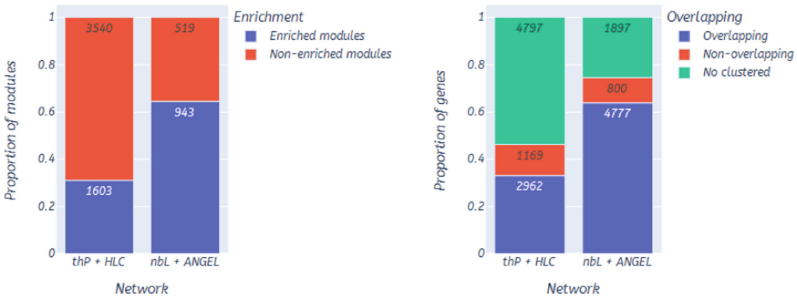

Fig. 1. Modules enrichment proportion and overlapping proportion of genes for thP and nbL networks.

Regarding the overlapping modules of genes for the nbL network, notice that the amount of transcription factors (TF) in the gene set belonging to multiple modules is statistically significant (p-value less than 0.05 for the Fisher's Exact Test). This supports the biological relevance of the overlapping modules. TFs regulate the expression of multiple genes and hence affect multiple pathways of varying functions [25]. Since TFs control different functions, they are expected to be found in overlapping modules. Another interesting finding is that, according to an enrichment analysis in ShynyGO [14], one of the pathways with the highest over-representation in the set of overlapping genes that corresponds to "response to stress" (GO:0006950).

3.3 Gene Selection

Based on the modules detected with ANGEL in the nbL network (following the workflow proposed in [26]), a total of 25 genes are identified as responsive to salt stress. All 25 genes are also identified as DEG. Genes LOC_Os07g39390, LOC_Os04g35010, and LOC_Os01g33450 are selected by both approaches, in the thP and nbL network.

From the 25 identified genes, after individual gene enrichment with the RGAP [18] and UniProt [10] databases, 11 genes report 17 different GO annotations relevant to salt stress response (which is statistically significant based on Fisher's exact test, p-value< 0.05). Table 2 lists these genes and the corresponding GO annotations relevant to salt stress response. The remaining 14 selected genes are:

- LOC_Os01g25920,
- LOC_Os01g33450,
- LOC_Os01g35789,
- LOC_Os01g35930,
- LOC_Os02g05790,
- LOC_Os05g36994,
- LOC_Os06g16050,
- LOC_Os06g22394,
- LOC_Os06g38210,
- LOC_Os07g22494,
- LOC_Os07g39390,
- LOC_Os09g06634,
- LOC_Os10g04050,
- LOC_Os10g24094.

The apoplast (GO:0048046) is the first subcellular compartment confronted with stress conditions when plants are subjected to salt stress [31]. Stress is first sensed by the receptors in membranes (GO:0016020), which then generates secondary signal messengers like calcium, reactive oxygen species, kinases (GO:0004672, GO:0016301, GO:0016740), and phosphates followed by the activation of transcription factor genes (GO:0003700) that eventually coordinates the plant's adaptive biochemical and physiological responses [16] (GO:0006950, GO:0009628, GO:0006952). Protein kinases regulate the phosphorylation and dephosphorylation of other proteins, and play a crucial role in stress signal transduction. In addition, serine/threonine protein kinases (GO:0004674) have also been known to be involved in multi-stress tolerance in plants [16].

Salt-induced toxicity negatively affects CO_2 fixation and thylakoid reactions of photosynthesis, which take place in thylakoids (GO:0009579) and the stroma of the chloroplast, resulting in poor plant growth and reduction in yield [15]. An essential process for growth, development, and homeostasis of organisms is the dynamic balance between ubiquitination and deubiquitination (GO:0071108, GO:1990380, GO:0004843, GO:1990380) [30]. In particular, inhibition of shoot and root development (GO:2000280) is the primary response to salt stress [35]. Other, independent studies confirm that the enhanced catalytic and transferase activities (GO:0016740) in salt-stressed rice plants, as well as the transport (GO:0006810) of salt and all related ions through the plant, reinforce salt stress tolerance [24].

Table 2. Selected genes with associated GO terms relevant to salt stress response.

LOC ID	GO term	GO name
LOC_Os03g44880	GO:2000280	Regulation of root development
LOC_Os03g63870	GO:0006950	Response to stress
	GO:0009628	Response to abiotic stimulus
LOC_Os04g35010	GO:0003700	DNA-binding transcription factor activity
LOC_Os06g09688	GO:0016020	Membrane
	GO:0009579	Thylakoid
LOC_Os07g15440	GO:0071108	Protein K48-linked deubiquitination
	GO:1990380	Lys48-specific deubiquitinase activity
	GO:0004843	Cysteine-type deubiquitinase activity
	GO:1990380	Lys48-specific deubiquitinase activity
LOC_Os07g37385	GO:0006810	Transport
LOC_Os07g43570	GO:0004674	Protein serine/threonine kinase activity
	GO:0004672	Protein kinase activity
	GO:0016301	Kinase activity
	GO:0016740	Transferase activity
	GO:0016020	Membrane
LOC_Os10g40520	GO:0006810	Transport
LOC_Os12g01290	GO:0006952	Defense response
LOC_Os12g14440	GO:0048046	Apoplast
LOC_Os12g27220	GO:0016740	Transferase activity

4 Concluding Remarks

This work proposes a novel approach for constructing co-expression networks based on the penalized least absolute shrinkage and selection operator (Lasso) [33]. Edges between genes are stablished based on the Lasso regression coefficients between the differential expression profile of one gene against all others. The approach extends the workflow described in [26] for identifying genes related to salt stress in rice. In particular, it uses the ANGEL algorithm, a static node-centric algorithm for detecting modules with overlaps, which improves effectiveness and time complexity. Note that the proposed approach can be used for building differential and non-differential co-expression networks.

The Lasso-based approach is computationally appealing as each of the N Lasso problems can be solved independently. This makes the proposed approach a good candidate to exploit parallelization [21]. Moreover, this method avoids the additional problem that often arises in constructing co-expression networks based on correlations: the definition of a threshold to identify the strongest connections that define the edges of the network. Using Lasso is especially convenient in inferring differential co-expression networks because it yields accurate param-

eter estimates even with small sample sizes [13], a common condition studying expression data under control and stress conditions.

The modified workflow was applied to a case study of rice under salt stress. The resulting network, inferred with the Lasso approach contained 7,474 vertices and 67,061 edges. 65% of the identified modules were enriched. The amount of transcription factors in the set of genes belonging to multiple modules (overlapping genes) was statistically significant. Finally, a total of 25 genes were selected as genes that respond to salt stress in rice. All 25 genes were also identified as DEG. Of these identified genes, 11 reported a statistically significant number of different GO annotations relevant to the salt stress response.

As future work, the proposed Lasso-based approach can be applied to infer co-expression networks of other organisms and other types of stresses. Moreover, the inferred networks can be used in comparisons and downstream analysis of co-expression networks. Developing a parallel implementation of the current workflow is an important research direction to further reduce time complexity. Further exploration of co-expression networks should provide valuable insights into the gene interactions and their joint response to stresses.

Acknowledgments. 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 Tourism, and ICETEX, under GRANT ID: FP44842-217-2018.

References

1. Ahn, Y.Y., Bagrow, J.P., Lehmann, S.: Link communities reveal multiscale complexity in networks. Nature **466**(7307), 761–764 (2010)
2. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.: Gene ontology: tool for the unification of biology. Nature Genet. **25**(1), 25–29 (2000)
3. Aydin, B., Arga, K.Y.: Co-expression network analysis elucidated a core module in association with prognosis of non-functioning non-invasive human pituitary adenoma. Front. Endocrinol. **10**, 361 (2019)
4. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. Ser. B (Methodol.) **57**(1), 289–300 (1995)
5. Björck, Å.: Least squares methods. Handbook Num. Anal. **1**, 465–652 (1990)
6. Bujang, M.A., Baharum, N.: Sample size guideline for correlation analysis. World **3**(1), 37–46 (2016)
7. Chowdhury, H.A., Bhattacharyya, D.K., Kalita, J.K.: (Differential) Co-expression analysis of gene expression: a survey of best practices. IEEE/ACM Trans. Comput. Biol. Bioinform. **17**(4), 1154–1173 (2019)
8. Clough, E., Barrett, T.: The gene expression omnibus database. In: Statistical Genomics, pp. 93–110. Springer (2016)

9. The Gene Ontology Consortium: The gene ontology resource: enriching a gold mine. Nucl. Acids Res. **49**(D1), D325–D334 (2021)
10. Consortium, U.: Uniprot: a worldwide hub of protein knowledge. Nucleic Acids Res. **47**(D1), D506–D515 (2019)
11. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Ann. Stat. **32**(2), 407–499 (2004)
12. 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. J. Plant Regist. **8**(1), 109–116 (2014)
13. Finch, W.H., Finch, M.E.H.: Regularization methods for fitting linear models with small sample sizes: Fitting the Lasso estimator using R. Pract. Assess. Res. Eval. **21**(1), 7 (2016)
14. Ge, S.X., Jung, D., Yao, R.: Shinygo: a graphical gene-set enrichment tool for animals and plants. Bioinformatics **36**(8), 2628–2629 (2020)
15. Hameed, A., Ahmed, M.Z., Hussain, T., Aziz, I., Ahmad, N., Gul, B., Nielsen, B.L.: Effects of salinity stress on chloroplast structure and function. Cells **10**(8), 2023 (2021)
16. Hossain, M.R., Bassel, G.W., Pritchard, J., Sharma, G.P., Ford-Lloyd, B.V.: Trait specific expression profiling of salt stress responsive genes in diverse rice genotypes as determined by modified significance analysis of microarrays. Front. Plant Sci. **7**, 567 (2016)
17. Hou, J., Ye, X., Feng, W., Zhang, Q., Han, Y., Liu, Y., Li, Y., Wei, Y.: Distance correlation application to gene co-expression network analysis. BMC Bioinform. **23**(1), 1–24 (2022)
18. Kawahara, Y., de la Bastide, M., Hamilton, J.P., Kanamori, H., McCombie, W.R., Ouyang, S., Schwartz, D.C., Tanaka, T., Wu, J., Zhou, S., et al.: Improvement of the oryza sativa nipponbare reference genome using next generation sequence and optical map data. Rice **6**(1), 1–10 (2013)
19. Korthauer, K., Kimes, P.K., Duvallet, C., Reyes, A., Subramanian, A., Teng, M., Shukla, C., Alm, E.J., Hicks, S.C.: A practical guide to methods controlling false discoveries in computational biology. Genome Biol. **20**(1), 1–21 (2019)
20. Liesecke, F., De Craene, J.O., Besseau, S., Courdavault, V., Clastre, M., Vergès, V., Papon, N., Giglioli-Guivarc'h, N., Glévarec, G., Pichon, O., et al.: Improved gene co-expression network quality through expression dataset down-sampling and network aggregation. Sci. Rep. **9**(1), 1–16 (2019)
21. Osborne, M.R., Presnell, B., Turlach, B.A.: A new approach to variable selection in least squares problems. IMA J. Num. Anal. **20**(3), 389–403 (2000)
22. Ovens, K., Eames, B.F., McQuillan, I.: The impact of sample size and tissue type on the reproducibility of gene co-expression networks. In: Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. pp. 1–10 (2020)
23. Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Phys. Rev. E **76**(3), 036106 (2007)
24. Rahman, A., Nahar, K., Al Mahmud, J., Hasanuzzaman, M., Hossain, M.S., Fujita, M.: Salt stress tolerance in rice: emerging role of exogenous phytoprotectants. Adv. Int. Rice Res. **9**(3), 139–174 (2017)
25. Renkawitz, R.: Transcription Factors and Regulation of Gene Expression, pp. 1886–1890. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)
26. Riccio-Rengifo, C., Finke, J., Rocha, C.: Identifying stress responsive genes using overlapping communities in co-expression networks. BMC Bioinform. **22**(1), 1–17 (2021)

27. Rossetti, G.: Angel: efficient, and effective, node-centric community discovery in static and dynamic networks. Appl. Netw. Sci. **5**(1), 1–23 (2020)
28. Savino, A., Provero, P., Poli, V.: Differential co-expression analyses allow the identification of critical signalling pathways altered during tumour transformation and progression. Int. J. Molecul. Sci. **21**(24), 9461 (2020)
29. Serin, E.A., Nijveen, H., Hilhorst, H.W., Ligterink, W.: Learning from co-expression networks: possibilities and challenges. Front. Plant Sci. **7**, 444 (2016)
30. Snyder, N.A., Silva, G.M.: Deubiquitinating enzymes (dubs): Regulation, homeostasis, and oxidative stress response. J. Biol. Chem. **297**(3) (2021)
31. Song, Y., Zhang, C., Ge, W., Zhang, Y., Burlingame, A.L., Guo, Y.: Identification of NACL stress-responsive apoplastic proteins in rice shoot stems by 2d-dige. J. Proteom. **74**(7), 1045–1067 (2011)
32. Sprent, P.: Fisher Exact Test (2011)
33. Tibshirani, R.: Regression shrinkage and selection via the Lasso. J. Roy. Stat. Soc. Ser. B (Methodol.) **58**(1), 267–288 (1996)
34. Van Dam, S., Vosa, U., van der Graaf, A., Franke, L., de Magalhaes, J.P.: Gene co-expression analysis for functional classification and gene-disease predictions. Brief. Bioinform. **19**(4), 575–592 (2018)
35. Wang, Y., Zhang, W., Li, K., Sun, F., Han, C., Wang, Y., Li, X.: Salt-induced plasticity of root hair development is caused by ion disequilibrium in arabidopsis thaliana. J. Plant Res. **121**(1), 87–96 (2008)
36. Yeung, M.S., Tegnér, J., Collins, J.J.: Reverse engineering gene networks using singular value decomposition and robust regression. Proc. Nat. Acad. Sci. **99**(9), 6163–6168 (2002)