J Appl Biotechnol Rep. 2023 September;10(3):1079-1090

Journal of
Applied Biotechnology Reports

doi 10.30491/JABR.2023.388345.1611

Original Article

Identification of Key Responsive Genes to some Abiotic Stresses in *Arabidopsis thaliana* at the Seedling Stage based on Coupling Computational Biology Methods and Machine Learning

Abbas Karimifard [1], **Abbas Saidi** [1*], **Masoud Tohidfar** [1*], **Aditya Saxena** [2]

[1] Department of Plant Sciences and Biotechnology, Faculty of Life Sciences and Biotechnology, Shahid Beheshti University, Tehran, Iran
[2] Department of Biotechnology, Institute of Applied Sciences and Humanities, GLA University, Mathura, India

Corresponding Authors: Abbas Saidi, PhD, Professor; Masoud Tohidfar, PhD, Professor. Department of Plant Sciences and Biotechnology, Faculty of Life Sciences and Biotechnology, Shahid Beheshti University, Tehran, Iran. Tel: +98-2122431664, E-mail: abbas.saidi@gmail.com (A. Saidi); gtohidfar@yahoo.com (M. Tohidfar)

Received March 4, 2023; **Accepted** July 19, 2023; **Online Published** September 10, 2023

Abstract

Introduction: Abiotic limitations, like water deficit, high temperature, salinity, and cold are some of the main barrier agents to plant growth throughout the world. To obtain a comprehensive view of a plant's response to abiotic stresses, we applied the robust bioinformatics approaches that including the integration of meta-analysis, weighted gene co-expression network analysis (WGCNA), and machine learning.

Materials and Methods: In this paper, 32 samples from four different stresses were chosen for analysis. Cross-platform combination method was used to conduct meta-analysis. To find gene co-expression modules related to stress conditions WGCNA analysis was performed. Machine learning methods were applied to validate the most important hub genes.

Results: Meta-analysis detected 275 differential expression genes (DEGs) and WGCNA showed 28 distinct modules under those stresses. Seven potential hub genes (At1g07430 (HAI2), At5g52300 (LTI65), At1g60190 (PUB19), At5g50360, At1g77120 (ADH1), At1g56600 (GolS2), and At5g57050 were detected by network analysis and validated by machine learning methods. These genes are involved in different pathways of cellular response to abiotic stresses.

Conclusions: Analysis indicates that among the hub genes, At5g50360 was identified as a novel candidate gene. As such, the At5g50360 can be used in plant breeding programs for the development of abiotic stress-tolerant crops.

Keywords: Abiotic Stress, Machine Learning, Meta-analysis, Weighted Correlation Network Analysis, Gene Expression

Citation: Karimifard A, Saidi A, Tohidfar M, Saxena A. Identification of Key Responsive Genes to some Abiotic Stresses in *Arabidopsis Thaliana* at the Seedling Stage based on Coupling Computational Biology Methods and Machine Learning. J Appl Biotechnol Rep. 2023;10(3):1079-1090. doi:10.30491/JABR.2023.388345.1611

Introduction

Various environmental agents have negative effects on the growth and development of the plants and ultimately lead to decrease in final yield performance in plants.[1] Salinity, heat, cold, and drought stresses are the most important abiotic stresses affecting plant growth.[2] Annually, environmental stresses can reduce crop yield production from 50 to 70 percent.[3] During growth and differential stages of a plant's life, the seedling stage is one of the most vital phases influencing many plants population features such as size and genetic variations.[4] The seedling stage is one of the most sensitive growth stages to abiotic stresses and the occurrence of abiotic stresses at this stage will lead to decreased plant yield.[5] Therefore, it is important for scientists to acquire more information about the response of plants to abiotic stresses in the seedling stage.

Generally, response to abiotic stresses at the cellular, molecular, and whole plant levels are so complicated.[6] These complexities come from the nature of interactions between stress agents and internal factors of plants.[7] Moreover, tolerance mechanisms to abiotic stresses are a complicated phenomenon because many of the physiological and molecular signaling pathways are active in sensing and responding to these stresses.[8]

One of the most important methods of molecular genetics helping to better comprehend the tolerance mechanisms of the plants in response to abiotic stresses is the use of transcriptome analysis and next-generation sequencing (NGS) technologies.[9] These molecular genetics methods provide valuable insights into the gene expression patterns and molecular processes. involved in plant responses to stress conditions. By using computational biology approaches, we can identify hub genes and important pathways associated with abiotic stresses.[10] These days, two computational biology approaches, namely meta-analysis and WGCNA, are

widely used by researchers to get information at the level of transcriptome of plants facing abiotic stresses.[11] Over the past decade, many scientists have been taking advantage of microarray technology. Merging of several microarray datasets (meta-analysis) is a typical way to improve gene selection.[12] Increasing the number of samples boosts the statistical power for acquiring a more meticulous estimate of gene expression collection.[13] Meta-analysis of microarray data is an efficient bioinformatics method to integrate multiple gene expression datasets.[14] By employing a meta-analysis approach, we can identify key responsive genes involved in metabolic and molecular pathways of abiotic stresses, so that these hub genes may be utilized in plant breeding programs for improving tolerance to stresses.[15] Meta-analysis have been employed by several research teams for identification of DEGs in response to abiotic stresses in plants.[16,17]

Despite the importance of identifying DEGs in plant responses to environmental stresses through meta-analysis, the relationships among these genes are yet to be identified. Therefore, it is necessary to employ a method to bridge this gap and explore the interplay among these genes. WGCNA is a considerable system biology method that recognizes gene or protein functions and discovers correlation pattern among genes.[18] This is a helpful method for the recognition of the module's genes and the distinguishing of potential key genes. Nowadays, WGCNA is widely used to identify abiotic-responsive genes in *Arabidopsis*, rice, soybean, maize, tomato and many other plants.[19,20]

Machine learning (ML), as a distinct branch of artificial intelligence, empowers systems to learn and significantly boost their predictive capabilities by leveraging training data and accumulated experiences. Recently, the application of machine learning in biology is increasing.[21] Lately, researchers are extensively utilizing machine learning algorithms at the level of genomics (for DNA sequence analysis),[22] transcriptomics (gene expression profiling analysis by XGBoost algorithm),[23] and proteomics (i.e., sequence-based prediction of protein-protein interaction analysis).[24] Machine learning plays a significant role in plant biology, particularly in the identification of hub genes under abiotic stresses. Feature selection is recognized as an effective machine learning algorithm, and extensively utilized by researchers to identify significant genes from a given set of genes. Among them, three different algorithms of feature selection methods, namely chi-square test, random forest, and SVM-RFE (support vector Machine-Recursive Feature Elimination), are highly recommended for reducing dimensionality and performing feature selection.[25] SVM-RFE, an effective feature selection method, relies on the power of support vector machines to accurately identify and prioritize relevant features. The SVM-RFE technique was utilized to train an SVM model, enabling the determination of the weight for each gene. Subsequently, genes with the lowest weight were iteratively identified and removed from the feature set. This algorithm's exceptional performance has led to its extensive utilization across numerous domains within the field of biology. The random forest algorithm has been widely adopted by researchers for feature selection and classification purposes.[26] To reduce dimensionality and select the most relevant genes, the chi-square algorithm is another important feature selection method that has been employed by researchers.[27]

In this paper, we integrated meta-analysis of transcriptome data from four diverse abiotic stresses (salinity, heat, cold, and drought) with WGCNA to identify important genes and modules involved in plant response to these stresses. To enhance the efficiency of hub gene selection, we applied three different feature selection algorithms.

Materials and Methods

Selection of Expression Data and Preprocessing

We used GEO (Gene Expression Omnibus) database for choosing expression profiles in *Arabidopsis* (Table 1). The dataset consisted of 32 samples, with 16 samples each collected under stress and normal conditions. These samples were obtained from seven studies investigating four different types of stresses. For each individual dataset, data was normalized by quantile normalization algorithm and then the log2 transformation was carried out. For constructing an expression matrix, we merged all gene expression profiles by Gene Symbol. Batch effects were eliminated by using ComBat function in SVA package.[28]

Table 1. Characterization of the Individual Samples Used in this Study

GEO number	Sample groups (stress: normal)	Platform	Plant	Tissue	Type of Stress	References
GSE39236	3:3	Affymetrix (GPL198)	*Arabidopsis*	Seedling	Salt stress	[29]
GSE41963	2:2	Affymetrix (GPL198)	*Arabidopsis*	Seedling	Salt stress	[30]
GSE44053	2:2	Affymetrix (GPL198)	*Arabidopsis*	Seedling	Heat stress	[31]
GSE56642	3:3	Affymetrix (GPL198)	*Arabidopsis*	Seedling	Drought stress	[32]
GSE106635	2:2	Affymetrix (GPL198)	*Arabidopsis*	Seedling	Cold stress	[33]
GSE109283	2:2	Affymetrix (GPL198)	*Arabidopsis*	Seedling	Salt stress	[34]
GSE112389	2:2	Affymetrix (GPL198)	*Arabidopsis*	Seedling	Salt stress	[35]

Identification of Differential Genes Expression

The limma package[36] in R was utilized for the identification of differentially expressed genes (DEGs).

Construction of WGCNA Network

"WGCNA" package was used for constructing co-expression network[37] to find the relationships among genes. To decrease

background noise, we only chose genes that were expressed differentially in the samples, and genes with similar expression were deleted. A median absolute deviation (MAD) index for every gene, as a potent measure of variability, was calculated, allowing to rank the genes with 4500 genes being selected for further analysis.

To detect missing values and outliers, we used goodSamplesGenes function in WGCNA package and flashClust. To construct the network and adjacency matrix following parameters (cor = bicor, and type = signed hybrid) were used. The pickSoftThreshold function was used to determine the β parameter, an important step in WGCNA analysis, where it accentuates strong correlation between genes and also penalizes weak correlation. Other important WGCNA indexes such as module eigengene and module-trait relationships were performed. In candidate modules, we considered potential hub genes with two parameters: gene significance (GS)>0.2 and module membership (MM)>0.8.

Functional Analysis

Common genes between DEGs and WGCNA were chosen for gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses. The GO and KEGG analyses were conducted using the ShinyGo tool[38] (http://bioinformatics.sdstate.edu/go/). For both analyses, enriched terms and pathways were considered based on false discovery rate (FDR) adjusted *p*-values<0.05.

Protein-protein Interaction (PPI) Analysis

To investigate the protein-protein interactions, common proteins between DEGs and WGCNA were selected as an input data for string11.0 website (https://string-db.org/). To construct protein-protein interaction (PPI) networks, interactions with a score greater than 0.4 were considered for inclusion. To visualize the PPI network, Cytoscape (3.8.2) software was used. For ranking nodes in a network, Cytoscape plugin cytoHubba (Degree and MCC methods) was used. Common genes obtained from each of the methods of cytoHubba plugin were considered as hub genes.

Validation of Potential Hub Genes by Using Feature Selection Algorithms

Feature selection algorithms offer several advantages when applied to biological datasets. These include dimensional reduction, enhanced efficiency in selecting hub genes, improved interpretability of results, and identification of a subset of genes associated with specific biological processes and pathways. Due to these advantages, these algorithms have been widely utilized to leverage their benefits in biological research. Three different feature selection methods were applied to validate the potential hub genes (i.e., common genes between DEGs from meta-analysis and hub genes from the candidate module from WGCNA). To apply feature selection methods, we categorized our data into two distinct groups, consisting of independent variables (common genes) and a dependent variable representing binary classes for stress and normal conditions. Considering that the common genes originate from two robust bioinformatics approaches, it has been observed that a substantial portion of these genes demonstrate the potential to be classified as hub genes. In order to identify and validate these hub genes, the top 30 genes were selected from each analysis, including MCC, Degree, SVM-RFE, Random Forest, and Chi-square. This selection comprised both the hub genes and the genes playing a significant role in the response pathways to abiotic stresses. We utilized the SelectKBest function from scikit-learn, which employs the chi2 scoring function, to identify the K features with the highest chi-square values. The k parameter was set to 30 to select top 30 features with the highest scores. Given the nonlinear nature of non-biological data, the SVM-RFE algorithm was performed with radial basis function kernel. The SVM-RFE and chi-square methods were performed by sklearn library in Python. Random forest algorithm was conducted by using RandomForest package in R. All the data and codes used for implementing machine learning methods on the Figshear website have been uploaded (https://doi.org/10.6084/m9.figshare.23498246).

Results

Data Quality Control

Primary transcriptome data analysis, including quality control (QC) and normalization, was performed on each dataset, so all samples were normalized. To ensure batch effect removal, principal components analysis (PCA) and boxplot were carried out (Figure 1). Figure 1a clearly demonstrates the presence of batch effects (BE) in our datasets, which can introduce confounding factors and impact the interpretation of results. To address this issue, we applied the ComBat algorithm from the SVA package[38] to remove the batch effects. The effectiveness of this correction method is evident in Figure 1b, where the two distinctive groups are clearly distinguishable after the removal of batch effects.

Furthermore, after the batch effects were successfully removed, the boxplot analysis (Figure 1c, d) revealed that the median lines of gene expression levels are positioned closely together across all the samples. This indicates that the removal of batch effects has improved the consistency and comparability of gene expression measurements, reducing the potential bias introduced by batch effects.

Identification of DEGs by Meta-analysis

Cross platform normalization method was used to perform meta-analysis. By using this approach, 167 up-regulated genes ("LogFC≥1, Adjusted *p*.value≤0.05") and 108 down-regulated

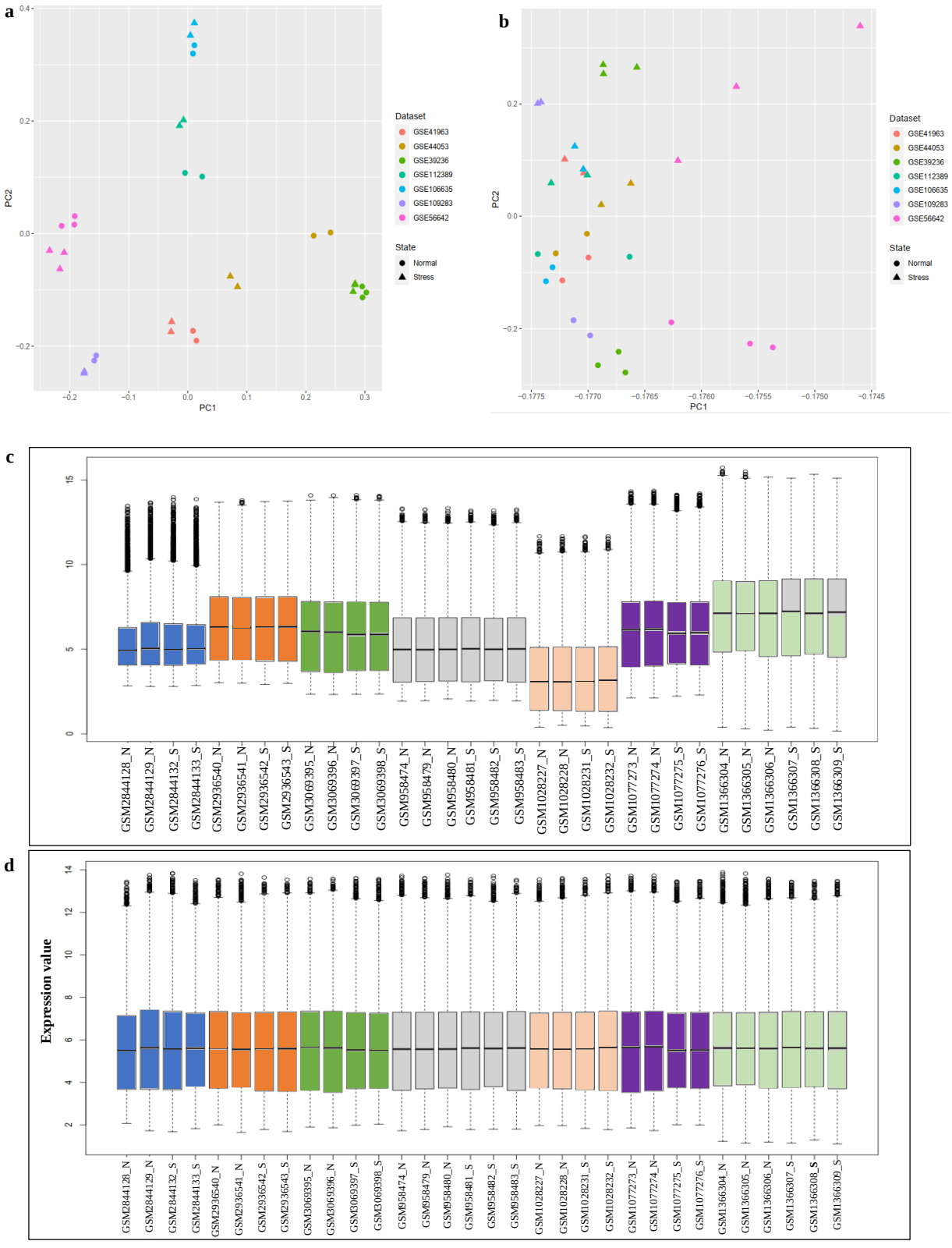

Figure 1. a) Plot of first two principal components: (left panel) visualizing data before; **b)** after BE removal (right panel); **c)** Boxplot of samples before BE removal, and **d)** boxplot of samples after BE removal.

genes ("LogFC≤-1, Adjusted *p*.value≤0.05") were determined (Supplementary File 1). The At1g16850 and At5g20630 were defined as genes with the highest and the lowest values, respectively, (Figure 2).

Constructing of Weighted Co-expression Network, Recognition Candidate Modules and Potential Hub Genes

WGCNA analysis was performed on a merged gene expression of seven datasets that were normalized with its batch effect

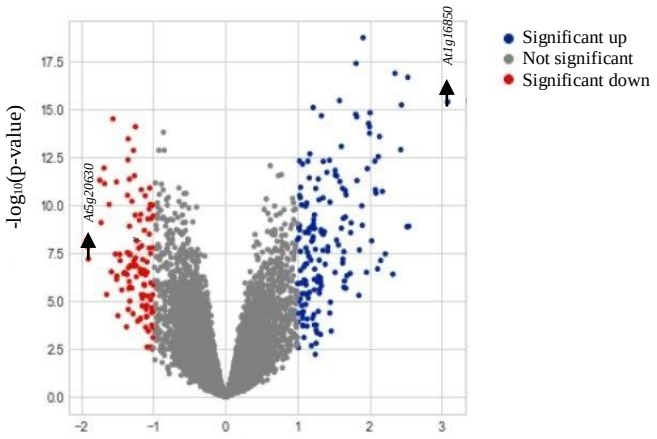

Figure 2. Volcano Plot of DEGs under Normal and Stress Conditions at Seedling Stage in *Arabidopsis*.

removed. To identify outliers, we clustered the samples. As shown in figure 3a, there is no outlier in our data. Moreover, all the samples under normal and stress conditions were classified into separate groups**.**

One of the critical stages for performing WGCNA analysis is determining the soft threshold (power β). In order to fulfill the scale-free topology feature of the WGCNA network, determining the power β is vital. Therefore, we considered 9 as the soft threshold index (β parameter) by fit index greater than 0.8 (Figure 3b). Based on the findings in Figure 3c, opting for a threshold of 9 for beta power results in the least decline in mean connectivity.

By using the parameter β and gene expression matrix, adjacency matrix was generated. Lastly, two indexes namely average hierarchical clustering and dynamic tree were utilized to identify co-expression modules (Figure 3e). One of the WGCNA framework's advantages is the detection of the association between considered trait (normal and stress conditions) and gene expression profile. The relationship between considered trait and module eigengenes was obtained through the pearson's correlation coefficient. By analyzing correlation of module eigengenes with external traits (stress and normal conditions), two modules (black and blue) were determined as modules with the most significant associations with the trait under abiotic stress condition (Figure 3f).

As Shown in Figures 3g and 3h, blue (r = 0.96, *p* = [3e-18]) and black (r = 0.72, *p* = [3e-06]) modules have a highly positive correlation with the external trait (a stress condition), thus potential hub genes based on gene significance greater than 0.2 and module membership greater than 0.8 from these two modules were chosen (Figure 3g, h).

Potential hub genes in key modules (blue and black modules) were selected based on the two criteria mentioned earlier. A total of 225 genes, 119 genes in the blue modules and 106 genes in the black modules, were identified as potential hub genes by WGCNA (Supplementary File 2).

Identification of Common Genes between Meta-analysis and WGCNA

We intersected 275 DEGs from meta-analysis with 225 potential hub genes from two key modules, which yielded 98 genes (Figure 4) (Supplementary File 3).

Functional Enrichment Analysis of Common Genes between Meta-analysis and WGCNA

After conducting a comprehensive analysis, it was concluded that no significant enriched terms were discovered in relation to the cellular components. Therefore, gene ontology, including biological process (BP), molecular function (MF), and KEGG analysis were just performed on the above-mentioned 98 common genes (Figure 5). Totally, in gene ontology annotation, 18 BP terms, 11 MF terms, and 4 KEGG terms were detected.

Protein-protein Interaction (PPI) Analysis of Common Genes between Meta-analysis and WGCNA

The 98 common genes were selected for constructing PPI network. The string database was used for generating the network (Figure 6). The cytoscape (3.7.2) was used to visualize the network. By using the Degree and MCC indexes of cytoHubba (plugin Cytoscape), we restricted our hub genes into two separate categories of 30 genes (Supplementary File 4). These genes were used to find key genes along with other genes obtained from machine learning methods.

Identification of Key Genes by Machine Learning Methods

To identify the most significant genes among 98 common genes, we applied three different algorithms of machine learning. Firstly, random forest (RF) algorithms were applied, followed by calculation of variable significance inside the RF analysis based on the Gini index, which is a measurement of variance for a given variable.[40] In this algorithm, Mean Decrease Gini (MDG) is one of the crucial criteria for scoring the genes. We used it for selecting the genes with top-level scores (Figure 7). The top 30 genes obtained by this method are listed in the Supplementary File 4. As shown in Figure 7, the most remarkable gene (At1g60190) is located at the highest point of the graph.

Another algorithm of machine learning that we applied for choosing the most notable genes was SVM-RFE. Based on this method, the top 30 genes are listed in the Supplementary File 4.

Chi-square method was another important algorithm of machine learning that was used for selecting significant genes. The top 30 significant genes were defined by this method and are listed in the Supplementary File 4.

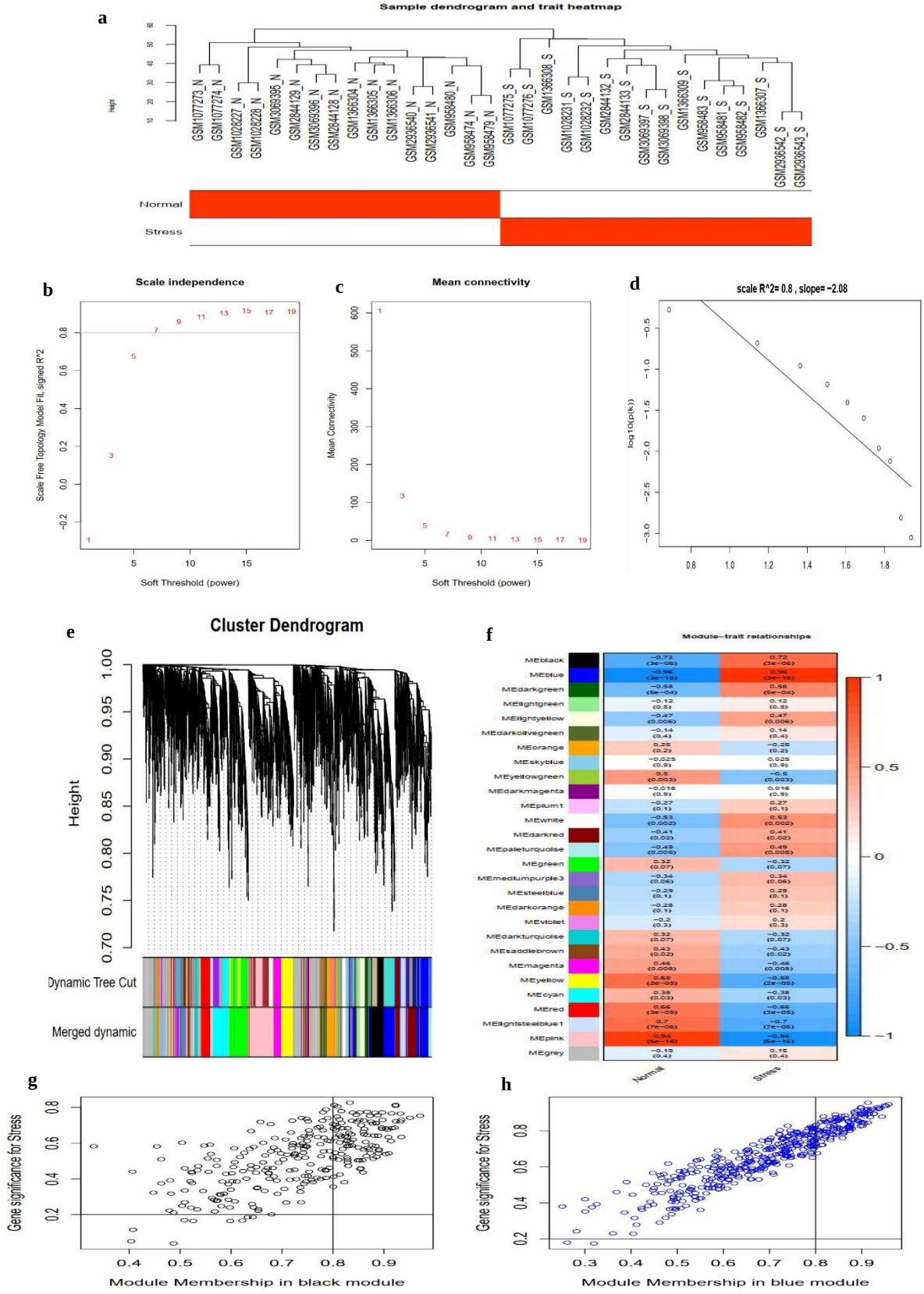

Figure 3. WGCNA Analysis of Abiotic Stress Samples and Normal Samples. **a)** Sample clustering with their external traits (normal and stress condition) and heatmap for 32 samples; **b)** Analysis of the scale-free for different soft thresholding powers, **(c)** Analysis of the mean connectivity at different soft thresholding power; **d)** Check scale free topology with β = 9; **e)** A cluster dendrogram of genes based on the measurement of dissimilarity (1-TOM); **f)** Module-trait relationship, this plot shows correlation between each module with external traits. Module eigengene were listed on y-axis and the two different conditions were placed in each column. Degree of correlation and p-value were also displayed in each cell; **g, h)** Scatter plots between gene significance and module membership in black and blue modules.

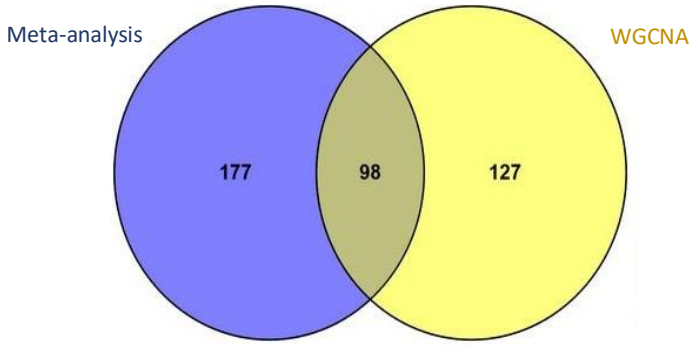

Figure 4. Venny Diagram between DEGs from Meta-analysis and WGCNA.

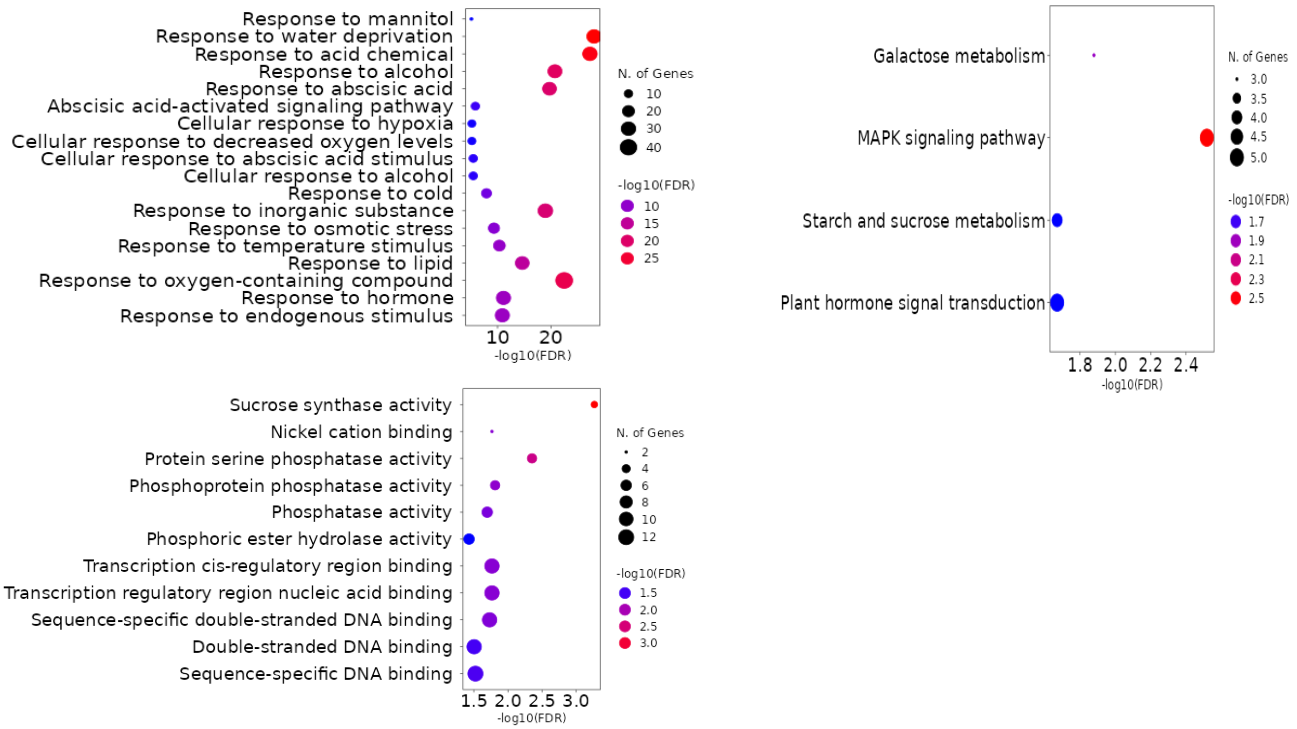

Figure 5. Gene Ontology (GO) Analysis and KEGG Pathway of 98 Common Genes. **a)** GO of biological process; **b)** GO of molecular function, and **c)** KEGG pathways. In this plot, the color of circles represents the degree of enrichment, Reder colors show higher enrichment and the size of circles represent of the number of genes.

To detect the final key genes, we intersected the outcome of the three machine learning algorithms (SVM-RFE, RF, and Chi-square) with the Degree and MCC methods using venny diagram (Figure 8). Seven key genes including At1g07430 (HAI2), At5g52300 (LTI65), At1g60190 (PUB19), At5g50360, At1g77120 (ADH1), At1g56600 (GolS2), and At5g57050 (ABI2) were detected. We also showed these genes in the PPI network (Figure 8).

Gene ontology and KEGG analysis of these seven genes (Figure 9) indicate that they are the most important biological processes involved in responses to abscisic acid, alcohol, water deprivation, lipid, acid chemical, osmotic stress, and hormones. Moreover, MAPK signaling pathway and plant hormones signal transduction were detected as the most enriched KEGG terms.

Discussion

In this study, we implemented an integrated bioinformatic approach to identify important genes responsive to abiotic stresses. To accomplish this goal, we performed a meta-analysis of microarray data, which is a robust computational method widely used in this context. To perform meta-analysis, there are two separate approaches for analysis of multiple microarray datasets obtained from independent studies namely "integrative analysis"[40] and cross-platform normalization (also named "merging"). In this paper, we used a "merging" approach for analyzing the data. Meta-analysis was performed on seven datasets which included 32 samples. We identified 275 DEGs (between stress and normal condition) in our study that among, where the 167 (60 percent) upregulated and 108 (40 percent) downregulated

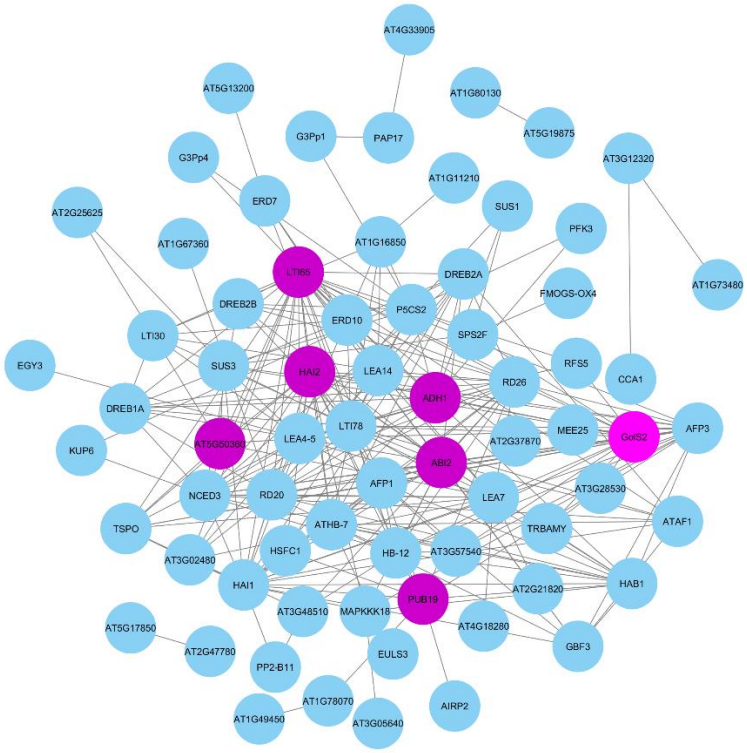

Figure 6. The PPI Network of 98 Common Genes. The red node represents key genes [At1g07430 (HAI2), At5g52300 (LTI65), At1g60190 (PUB19), At5g50360, At1g77120 (ADH1), At1g56600 (GolS2), and At5g57050 (ABI2)] that were validated by 3 different ML methods (SVM-RFE, Random Forest and Chi square) and two criteria cytoHubba (Degree and MCC).

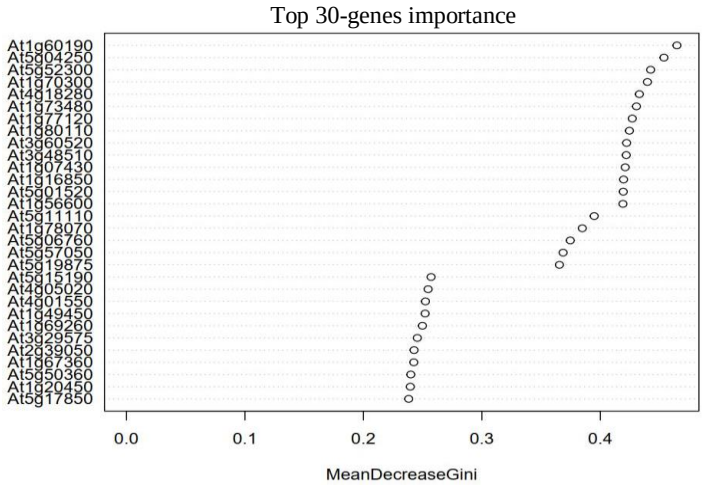

Figure 7. Top 30 Ranked Genes for 98 Common Genes.

genes were selected. The At1g16850, a transmembrane protein having an important role in response to salt and cold stresses, had the highest logFC at the seedling stage.[42] The lowest value (LogFC = -1.91) of down-regulated gene belonged to At5g20630 (GER3) (Figure 2).

To find module eigengene and genes related to the considered trait, WGCNA workflow was applied on the gene expression matrix of 32 samples. In this analysis, the genes with the same expression pattern were placed in the similar modules. As shown in Figure 5a, after combining modules with the same co-expression patterns, 28 various modules were diagnosed. The genes belonging to the gray module are not co-expressed. The blue and black modules have highly positive correlation with the considered trait (stress condition). Respectively, there are 277 and 440 genes in black and blue modules (supplementary 2). After applying modules screening based on gene significance and module membership, the number of genes in black and blue modules were reduced to 106 and 119, respectively.

To better understand tolerance mechanism in plants under

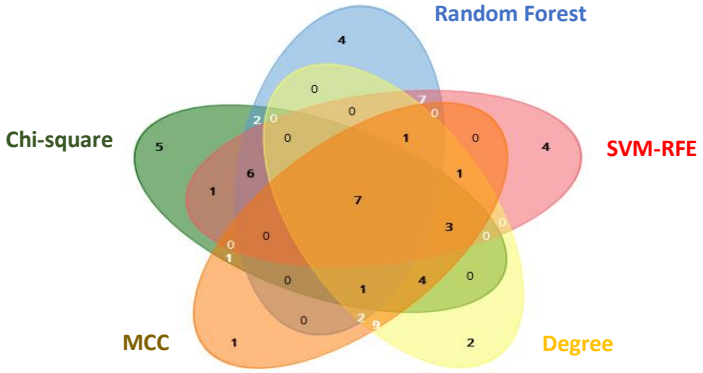

Figure 8. Venny Diagram between three Different Machine Learning Methods with Two Criteria of cytoHubba.

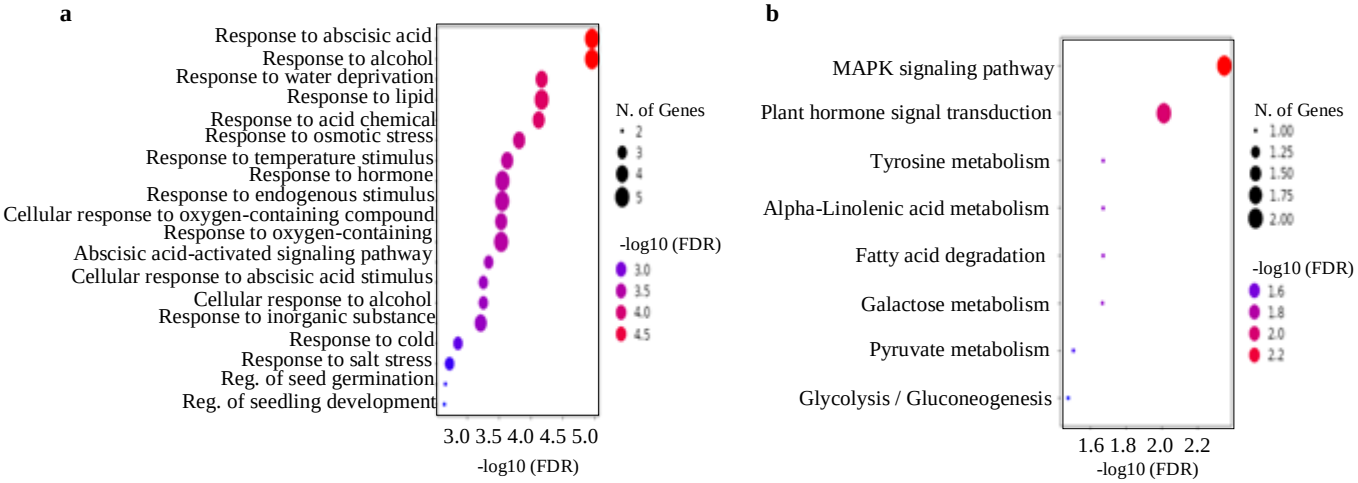

Figure 9. Functional Analysis on Seven Hub Genes. **a)** Biological process; **b)** KEGG pathway.

abiotic stresses at seedling stage different types of analysis such as GO, KEGG, and PPI were conducted on the common genes. A total of 98 genes were selected from meta-analysis of DEGs and candidate modules.

In particular, the most important BP term groups was related to the response of water deprivation with 30 genes.

Another important BP term was a response to acid chemical. Responses to oxygen−containing compound, abscisic acid, and alcohol were other important BP processes (Figure 8a). As expected, the most BP gene groups were involved in the cellular and molecular response to salt and water stress conditions and also in signaling pathways such as response to hormone and abscisic acid. In addition, the analysis of the molecular function (MF) group showed that the genes were chiefly associated with sucrose synthase activity (Figure 8b). Among the four identified groups in KEGG pathways, the only significant term was "MAPK signaling pathway" (Figure 8c). The MAPK pathway is one of the most important components in plant response to water, cold, salt, and heat stresses.[43]

In MF categories, the main term was sucrose synthase activity. Sucrose synthase is the key enzyme involved in biosynthesis of sucrose. The activity of sucrose synthase is increased under osmotic and water stresses.[44] Based on previous research, the level of soluble sugars such as sucrose is increased in cells under abiotic stresses.[45] Sucrose has a dual function, where not only acts as an energy substrate but also as a signaling molecule required for growth and development of plants.[46] Sucrose is synthesized by sucrose phosphate synthase and has a critical role in abiotic stresses.[47]

We have identified two sets of 30 top genes by using degree and MCC methods in cytoscape. To validate the results of PPI analysis, we applied machine learning algorithms on common genes. Machine learning has a special place in almost all sciences, especially in biological sciences. In 2011, SVM-RFE was carried out in *Arabidopsis* to predict drought-resistant genes.[48] Tahmasbi et al.,[49] integrated meta-analysis and machine learning methods for investigating the transcriptomic response to water stress in Populus. The role of novel water deficit specific genes was identified by using machine learning methods in Oryza sativa.[50]

By using three methods of machine learning algorithms, three sets of 10 top genes were obtained. Seven key responsive genes including At1g07430 (HAI2), At5g52300

(LTI65), At1g60190 (PUB19), At5g50360, At1g77120 (ADH1), At1g56600 (GolS2), and At5g57050 (ABI2) were obtained by intersection of these five different methods (Figure 12). It was reported that the level of HAI2 expression during drought stress was increased and it is one of the positive regulators of abscisic acid (ABA), involved in cell signaling transduction.[51] The LTI65 is an ABA-dependent protein induced by low-temperature and drought stress.[52] The At5g52300 encodes desiccation-induced protein involved in protecting cells from stress damages.[53] The At1g60190 acts as a negative regulator of abscisic acid during drought stress.[54] The At1g77120 has important role during abiotic and biotic stresses.[55] The At1g56600 (GolS2) encodes galactinol synthase 2 protein involved in the biosynthesis of raffinose family oligosaccharides from UDP-galactose. The GolS2 has critical role in plant cell's response to water and salt stresses.[56] The At5g57050 encodes protein phosphatase 2C family protein (ABI2) and has a crucial role in ABA-signal transduction during abiotic stresses.[57] Another key gene that we detected in our study was the At5g50360 (Von Willebrand factor A domain protein) which belongs to the blue modules. Although, a positive correlation between abscisic acid treatment and the expression level of the At5g50360 has been established however, the function of this gene is yet to be identified.[58] Our findings provide valuable insights into the fundamental mechanisms associated with abiotic stress responses, which can be leveraged for future genetic enhancement and breeding programs in plants.

Conclusion

In conclusion, based on our findings, we are suggesting a new potential hub gene At5g50360 which may play a critical role in many of the abiotic stress tolerance mechanisms in *Arabidopsis*. and it is highly recommended to conduct additional genetic studies to characterize its function. As such, this gene can be considered as an appropriate candidate for increasing abiotic tolerance in crop improvement programs.

Authors' Contributions

All authors have equal contribution.

Conflict of Interest Disclosures

The authors declare that they have no conflicts of interest.

References

1. Fard AK, Sedaghat S. Evaluation of drought tolerance indices in bread wheat recombinant inbred lines. Eur J Exp Bio. 2013;3(2):201-4.
2. Hajibarat Z, Saidi A, Zeinalabedini M. Evaluation of Drought Tolerance of Potato (Solanum Tuberosum L.) Under Water Deficit. J Crop Breed. 2020;12(35):102-12. doi:10.52547/jcb.12.35.102
3. Daryanto S, Wang L, Jacinthe PA. Global synthesis of drought effects on maize and wheat production. PloS One. 2016;11(5):e0156362. doi:10.1371/journal.pone.0156362
4. Grime JP. Plant strategies, vegetation processes, and ecosystem properties. John Wiley & Sons; 2006.
5. Peleg Z, Blumwald E. Hormone balance and abiotic stress tolerance in crop plants. Curr Opin Plant Biol. 2011;14(3):290-5. doi:10.1016/j.pbi.2011.02.001
6. Shinozaki K, Yamaguchi-Shinozaki K. Functional genomics in plant abiotic stress responses and tolerance: From gene discovery to complex regulatory networks and their application in breeding. PJA Series B. 2022;98(8):470-92. doi:10.2183/pjab.98.024
7. Hirayama T, Shinozaki K. Research on plant abiotic stress responses in the post-genome era: past, present and future. The Plant J. 2010;61(6):1041-52. doi:10.1111/j.1365-313X.2010.04124.x
8. Golldack D, Li C, Mohan H, Probst N. Tolerance to drought and salt stress in plants: unraveling the signaling networks. Front Plant Sci. 2014;5:151. doi:10.3389/fpls.2014.00151
9. Urano K, Kurihara Y, Seki M, Shinozaki K. 'Omics' analyses of regulatory networks in plant abiotic stress responses. Curr Opin Plant Biol. 2010;13(2):132-8. doi:10.1016/j.pbi.2009.12.006
10. Mohanta TK, Bashir T, Hashem A, Abd_Allah EF. Systems biology approach in plant abiotic stresses. Plant Physiol Biochem. 2017;121:58-73. doi:10.1016/j.plaphy.2017.10.019
11. Hajibarat Z, Saidi A, Hajibarat Z. Genome-wide identification of 14-3-3 gene family and characterization of their expression in developmental stages of Solanum tuberosum under multiple biotic and abiotic stress conditions. Funct Integr Genomics. 2022;22(6):1377-90. doi:10.1007/s10142-022-00895-z
12. Hamid JS, Hu P, Roslin NM, Ling V, Greenwood CM, Beyene J. Data integration in genetics and genomics: methods and challenges. Hum Genomics Proteomics. 2009;2009:869093. doi:10.4061/2009/869093
13. Ramasamy A, Mondry A, Holmes CC, Altman DG. Key issues in conducting a meta-analysis of gene expression microarray datasets. PLoS Med. 2008;5(9):e184. doi:10.1371/journal.pmed.0050184
14. Tseng GC, Ghosh D, Feingold E. Comprehensive literature review and statistical considerations for microarray meta-analysis. Nucleic Acids Res. 2012;40(9):3785-99. doi:10.1093/nar/gkr1265
15. Sharma R, Singh G, Bhattacharya S, Singh A. Comparative transcriptome meta-analysis of *Arabidopsis thaliana* under drought and cold stress. PloS One. 2018;13(9):e0203266. doi:10.1371/journal.pone.0203266
16. Ashrafi-Dehkordi E, Alemzadeh A, Tanaka N, Razi H. Meta-analysis of transcriptomic responses to biotic and abiotic stress in tomato. PeerJ. 2018;6:e4631. doi:10.7717/peerj.4631
17. Kong W, Zhong H, Gong Z, Fang X, Sun T, Deng X, et al. Meta-analysis of salt stress transcriptome responses in different rice genotypes at the seedling stage. Plants. 2019;8(3):64. doi:10.3390/plants8030064
18. Zhao W, Langfelder P, Fuller T, Dong J, Li A, Hovarth S. Weighted gene coexpression network analysis: state of the art. Journal of biopharmaceutical statistics. 2010;20(2):281-300. doi:10.1080/10543400903572753
19. Zeng Z, Zhang S, Li W, Chen B, Li W. Gene-coexpression network analysis identifies specific modules and hub genes related to cold stress in rice. BMC Genomics. 2022;23(1):251. doi:10.1186/s12864-022-08438-3

20. Li Y, Zhang Y, Luo H, Lv D, Yi Z, Duan M, et al. WGCNA Analysis Revealed the Hub Genes Related to Soil Cadmium Stress in Maize Kernel (*Zea mays* L.). Genes. 2022;13(11):2130. doi:10.3390/genes13112130
21. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. Nat Rev Mol Cell Biol. 2022;23(1):40-55. doi:10.1038/s41580-021-00407-0
22. Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. 2016;26(7) :990-9. doi:10.1101/gr.200535.115
23. Li W, Yin Y, Quan X, Zhang H. Gene expression value prediction based on XGBoost algorithm. Front Genet. 2019;10:1077. doi:10.3389/fgene.2019.01077
24. Sun T, Zhou B, Lai L, Pei J. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. BMC Bioinform. 2017;18:277. doi:10.1186/s12859-017-1700-2
25. Pirooznia M, Yang JY, Yang MQ, Deng Y. A comparative study of different machine learning methods on microarray gene expression data. BMC Genomics. 2008;9:S13. doi:10.1186/1471-2164-9-S1-S13
26. Xia H, Akay YM, Akay M. Selecting relevant genes from microarray datasets using a random forest model. IEEE Access. 2021;9:97813-21. doi:10.1109/ACCESS.2021.3092368
27. Yan C, Wu B, Ma J, Zhang G, Luo J, Wang J, et al. A novel hybrid filter/wrapper feature selection approach based on improved fruit fly optimization algorithm and chi-square test for high dimensional microarray data. Curr Bioinform. 2021;16(1):63-79. doi:10.2174/1574893615666200324125535
28. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28(6):882-3. doi:10.1093/bioinformatics/bts034
29. Guan Q, Wu J, Yue X, Zhang Y, Zhu J. A nuclear calcium-sensing pathway is critical for gene regulation and salt stress tolerance in *Arabidopsis*. PLoS Genetics. 2013;9(8):e1003755. doi:10.1371/journal.pgen.1003755
30. Sharma R, Priya P, Jain M. Modified expression of an auxin-responsive rice CC-type glutaredoxin gene affects multiple abiotic stress responses. Planta. 2013;238:871-84. doi:10.1007/s00425-013-1940-y
31. Yangueez E, Castro-Sanz AB, Fernandez-Bautista N, Oliveros JC, Castellano MM. Analysis of genome-wide changes in the translatome of *Arabidopsis* seedlings subjected to heat stress. PloS One. 2013;8(8):e71425. doi:10.1371/journal.pone.0071425
32. Noman M, Jameel A, Qiang WD, Ahmad N, Liu WC, Wang FW, et al. Overexpression of GmCAMTA12 enhanced drought tolerance in *Arabidopsis* and soybean. Int J Mol Sci. 2019;20(19):4849. doi:10.3390/ijms20194849
33. Cai W, Yang Y, Wang W, Guo G, Liu W, Bi C. Overexpression of a wheat (*Triticum aestivum* L.) bZIP transcription factor gene, TabZIP6, decreased the freezing tolerance of transgenic *Arabidopsis* seedlings by down-regulating the expression of CBFs. Plant Physiol Biochem. 2018;124:100-11. doi:10.1016/j.plaphy.2018.01.008
34. Mishra P, Jain A, Takabe T, Tanaka Y, Negi M, Singh N, et al. Heterologous expression of serine hydroxymethy ltransferase-3 from rice confers tolerance to salinity stress in *E. coli* and *Arabidopsis*. Front Plant Sci. 2019;10:217. doi:10.3389/fpls.2019.00217
35. Fang L, Wang Z, Su L, Gong L, Xin H. Vitis Myb14 confer cold and drought tolerance by activating Lipid transfer protein genes expression and Reactive oxygen species scavenge. Gene. 2023:147792. doi:10.1016/j.gene.2023.147792
36. Ritchie ME, Phipson B, Wu DI, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47. doi:10.1093/nar/gkv007
37. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics. 2008;9(1):559. doi:10.1186/1471-2105-9-559
38. Ge SX, Jung D, Yao R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. Bioinformatics. 2020;36(8):2628-9. doi:10.1093/bioinformatics/btz931
39. Lazar C, Meganck S, Taminau J, Steenhoff D, Coletta A, Molter C, et al. Batch effect removal methods for microarray gene expression data integration: a survey. Brief Bioinformatics. 2013;14(4):469-90. doi:10.1093/bib/bbs037
40. Feng Y, Wang Z, Yang N, Liu S, Yan J, Song J, et al. Identification of biomarkers for cervical cancer radiotherapy resistance based on RNA sequencing data. Front Cell Dev Biol. 2021;9:724172. doi:10.3389/fcell.2021.724172
41. Taminau J, Lazar C, Meganck S, Nowй A. Comparison of merging and meta-analysis as alternative approaches for integrative gene expression analysis. Int Sch Res Notices. 2014;2014:345106. doi:10.1155/2014/345106
42. McBride Z, Chen D, Reick C, Xie J, Szymanski DB. Global analysis of membrane-associated protein oligomerization using protein correlation profiling. Mol Cell Proteomics. 2017;16(11):1972-89. doi:10.1074/mcp.RA117.000276
43. De Zélicourt A, Colcombet J, Hirt H. The role of MAPK modules and ABA during abiotic stress signaling. Trends Plant Sci. 2016;21(8):677-85. doi:10.1016/j.tplants.2016.04.004
44. Chen L, Cai C, Chen V, Lu X. Learning a hierarchical repre sentation of the yeast transcriptomic machinery using an autoencoder model. BMC Bioinform. 2016;17:S9. doi:10.1186/s12859-015-0852-1
45. Du Y, Zhao Q, Chen L, Yao X, Zhang W, Zhang B, et al. Effect of drought stress on sugar metabolism in leaves and roots of soybean seedlings. Plant Physiol Biochem. 2020;146:1-12. doi:10.1016/j.plaphy.2019.11.003
46. Sami F, Yusuf M, Faizan M, Faraz A, Hayat S. Role of sugars under abiotic stress. Plant Physiol Biochem. 2016;109:54-61. doi:10.1016/j.plaphy.2016.09.005
47. Zhang Y, Zeng D, Liu Y, Zhu W. *SlSPS*, a sucrose phosphate synthase gene, mediates plant growth and thermotolerance in tomato. Horticulturae. 2022;8(6):491. doi:10.3390/horticulturae8060491
48. Liang Y, Zhang F, Wang J, Joshi T, Wang Y, Xu D. Prediction of drought-resistant genes in *Arabidopsis thaliana* using SVM-RFE. PLoS One. 2011;6(7):e21750. doi:10.1371/journal.pone.0021750
49. Tahmasebi A, Niazi A, Akrami S. Integration of meta-analysis, machine learning and systems biology approach for investigating the transcriptomic response to drought stress in Populus species. Sci Rep. 2023;13(1):847. doi:10.1038/s41598-023-27746-6
50. Thanmalagan RR, Roy A, Jayaprakash A, Lakshmi PT. Comprehensive meta-analysis and machine learning approaches identified the role of novel drought specific genes in *Oryza sativa*. Plant Gene. 2022;32:100382. doi:10.1016/j.plgene.2022.100382
51. Lim CW, Kim JH, Baek W, Kim BS, Lee SC. Functional roles of the protein phosphatase 2C, *AtAIP1*, in abscisic acid signaling and sugar tolerance in *Arabidopsis*. Plant Sci. 2012;187:83-8. doi:10.1016/j.plantsci.2012.01.013
52. Nordin K, Vahala T, Palva ET. Differential expression of

two related, low-temperature-induced genes in *Arabidopsis thaliana* (L.) Heynh. Plant Mol Biol. 1993;21:641-53. doi:10.1007/BF00014547

53. Yang SD, Seo PJ, Yoon HK, Park CM. The *Arabidopsis* NAC transcription factor VNI2 integrates abscisic acid signals into leaf senescence via the COR/RD genes. Plant Cell. 2011;23(6):2155-68. doi:10.1105/tpc.111.084913
54. Bergler J, Hoth S. Plant U-box armadillo repeat proteins AtPUB18 and AtPUB19 are involved in salt inhibition of germination in *Arabidopsis*. Plant Biol. 2011;13(5):725-30. doi:10.1111/j.1438-8677.2010.00431.x
55. Shi H, Liu W, Yao Y, Wei Y, Chan Z. *Alcohol dehydrogenase 1* (ADH1) confers both abiotic and biotic stress resistance in *Arabidopsis*. Plant Sci. 2017;262:24-31. doi:10.1016/j.plantsci.2017.05.013
56. Taji T, Ohsumi C, Iuchi S, Seki M, Kasuga M, Kobayashi M, et al. Important roles of drought-and cold-inducible genes for galactinol synthase in stress tolerance in *Arabidopsis thaliana*. Plant J. 2002;29(4):417-26. doi:10.1046/j.0960-7412.2001.01227.x
57. Rodriguez PL, Benning G, Grill E. ABI2, a second protein phosphatase 2C involved in abscisic acid signal transduction in *Arabidopsis*. FEBS Lett. 1998;421(3):185-90. doi:10.1016/S0014-5793(97)01558-5
58. Tian H, Chen S, Yang W, Wang T, Zheng K, Wang Y, et al. A novel family of transcription factors conserved in angiosperms is required for ABA signalling. Plant Cell Environ. 2017;40(12):2958-71. doi:10.1111/pce.13058