MJFAS
Malaysian Journal of Fundamental and Applied Sciences

Zainal-Abidin *et al.* | Malaysian Journal of Fundamental and Applied Sciences, Vol. 19 (2023) 6-18

RESEARCH ARTICLE

# Mining of Stress-related Genes in Pigmented and Non-pigmented Rice using Gene Co-expression Network and Clustering Approaches

**Rabiatul-Adawiah Zainal-Abidin[a]*, Nor Afiqah-Aleng[b], Yun Shin Sew[a], Norliza Abu Bakar[a] and Sanimah Simoh[a]**

[a]Biotechnology & Nanotechnology Research Centre, Malaysian Agricultural Research & Development Institute (MARDI), 43400 Serdang, Selangor, Malaysia; [b]Institute of Marine Biotechnology, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia

Abstract Pigmented rice has been associated with stress-resistant traits, such as resistance to abiotic and biotic stresses, but the genes network that is responsible for such traits remains limited. Hence, this study aims to identify stress-related genes in the pigmented rice using computational approaches. The gene co-expression network was constructed using Pearson Correlation Coefficient (PCC) ≥ 0.9 among differentially expressed genes (DEGs) of trancriptomes between pigmented and non-pigmented rice. The gene co-expression network was clustered using Markov Cluster algorithm (MCL) to identify the functional modules and the hub genes for each module were determined. The functional analyses were performed to each module to determine the related gene ontology (GO) and pathway. Protein-protein interaction (PPI) from STRING database was used to validate the functional analyses. A total of 721 DEGs were used to construct the gene co-expression network of pigmented and non-pigmented rice varieties. Of these, 614 DEGs with 15,259 edges were identified by PCC. Using MCL, 10 clusters were identified in the gene co-expression network of pigmented and non-pigmented rice varieties. Three clusters were enriched with seven GO terms (i.e., response to stress, response to stimulus, transcription factor activity) that are related to stress-resistant traits, indicating the highly correlated genes were the stress-related genes. Interestingly, nine hub genes were found to be related to drought tolerance, disease resistance and hormone biosynthesis. Validation of hub genes using STRING database revealed that 48 hub genes were also connected in the PPI network, suggesting their potential as candidate proteins in stress-related traits. This study demonstrated that the molecular interaction network and the network clustering approach are efficient in identifying the stress-related genes, which could provide new insights in understanding plant responses to abiotic and biotic stresses.

**Keywords**: Clustering gene co-expression network, MCL, pigmented rice, stress-resistant.

***For correspondence:**
rabiatul@mardi.gov.my

**Received:** 24 Dec. 2021
**Accepted:** 4 Dec. 2022



## Introduction

Pigmented rice (i.e., black and red rice) is highly nutritious than white rice. In addition, the secondary metabolites (i.e. flavonoid, anthocyanin, proanthocyanidin) are more abundant in black and red rice than in white rice [1]. Previous study has identified black and red rice with resistance or tolerance against abiotic and biotic stress traits [2]. This finding was due to the accumulation of secondary metabolites (i.e. flavonoid, anthocyanin, proanthocyanidin) that have been reported to respond to abiotic and biotic stress

conditions [3].

In plants, responses to abiotic and biotic stresses are complex traits that are regulated by multiple genes that interact with each other in a network [4]. Abiotic stresses in rice include drought, salinity, heat tolerance and submergence, while biotic stresses in rice are pest (i.e., brown planthopper) and pathogen (i.e., bacterial leaf blight) infections. To date, abiotic and biotic stresses have contributed to the decline in crops production [5]. Hence, many crop improvement programmes focus on developing stress-resistant and tolerant varieties [4]. Many genes with various functions have been identified to be involved in response to abiotic and biotic stresses. For instance, abiotic stress is regulated by transcription factor families, such as abscisic acid (ABA) [6], WRKY [7] and bHLH [8]. In biotic stress, a crop needs salicylic acid and secondary metabolites to protect them from pathogens attack [9], [10]. However, the stress-related genes in pigmented rice are not well-studied. In addition, the stress-related gene has never been fully characterised in terms of biological functions and molecular mechanisms, particularly in black and red rice. A previous study has performed the gene co-expression network analysis of different pigmented rice cultivars. However, their goal focuses on transcriptional factors that regulate flavonoid biosynthesis by interacting with photosynthesis, sugars synthesis and peroxidase pathways [2]. Hence, we performed the gene co-expression network analysis in this study to search for candidate stress-related genes in Malaysian pigmented rice varieties.

The development of a high-throughput sequencing platform has resulted in a massive increase in transcriptomic data. The transcriptomic data provide essential biological information to understand the biological functions and molecular mechanisms of complex traits in crops [11]. The gene co-expression network analysis based on large scale transcriptomics data has been widely performed for the screening and identification of genes that might be involved in stress-resistant trait [12]–[14]. This approach reveals the correlations between gene expression levels across different samples that show similar expression patterns tend to cluster together and are likely to be involved in the same biological process, molecular function, and regulatory process [15], [16]. Previous studies have performed the gene co-expression network analysis to annotate the uncharacterised gene functions in rice [17], to prioritise candidate genes for a wide variety of traits [18], [19], to correlate genes and phenotypic expression [20] and to construct regulatory network [21], [22]. In addition, the gene co-expression network analysis considers all samples together and establish connections among genes, which is the number of connections of a node in a network is known as the degree of connectivity. The highly connected nodes or genes is known as hub gene. The hub genes might be potential or informative [23] and might serve as a valuable biomarker for plant diseases [24].

Several studies demonstrated the utilisation of gene co-expression networks to mine the stress-related genes in heat, cold, and drought rice varieties [25]–[28]. For instance, three modules of tightly co-expressed genes were associated with signalling and heat stress response in rice [25]. Furthermore, the gene co-expression network helps in unravelling the regulatory mechanism of cold stress in specific modules and hub genes related to cold stress in rice [28]. While [27] showed that the interplay of pathways between the metabolism of chlorophyll and flavonoid and the signalling pathways of MAPK, IAA and SA is involved in stress tolerance response. Interestingly, integration of the gene co-expression networks with protein-protein interactions and pathway-level data has been performed to understand better the drought-responsive processes in drought-tolerance rice genotypes [26]. The above studies emphasised that the gene co-expression network approach has been widely used to identify and prioritise stress-related genes in rice.

Clustering is a technique that has been used to divide or partition the networks into several clusters or modules. The functional clusters or modules will be associated with biological processes, molecular functions and cellular components and pathways using gene ontology (GO) and biological pathway enrichment analyses. The clustering approach has been widely applied to gene co-expression networks for extracting densely connected genes [17], to classify metabolites that belong to metabolic pathways [17] and to characterise the genes into specific biological functions [29]. In general, clustering helps to summarise the information by reducing the dimensionality from thousands of genes to a small cluster. Several clustering algorithms have been performed in biological analysis, such as MCODE [29], MCL [12], dpClus, clusterMaker2 [30] and CytoCluster [31].

This study highlights the computational approach to identify and prioritise potential genes that are related to stress-related traits. The gene co-expression network was constructed with 721 significantly differentially expressed genes (DEGs) that were obtained from transcriptomics analysis of pigmented and non-pigmented rice varieties. A total of 10 clusters were detected using the MCL algorithm. The biological information of each cluster was assessed using GO and KEGG pathway enrichment analyses. In addition, predicted novel genes were annotated with putative functions. The aims of our study were to

characterise co-expression clusters in rice using the clustering approach and GO enrichment analysis and to identify significant genes that are potentially involved in stress-related traits. This effort will accelerate the discovery of stress-related genes in rice. Hence, it will contribute to improving the resistance and tolerance of rice using genome-editing and molecular breeding approaches.

# Materials and Methods

## Expression data

The transcriptomes data of four pigmented (Bali, Pulut Hitam 9, MRM 16, MRQ 100) and two non-pigmented (MR 297, MRQ 76) rice varieties were obtained from previous studies [32], [33]. Pulut Hitam 9 and Bali are black rice, while MRM 16 and MRG 100 are red rice. In addition, Pulut Hitam 9, MRQ 100 and MRQ 76 are sticky rice. The uniqueness of MRQ 76, when compared to these five varieties, is its aromatic characteristic. MR 297 is resistant to blast disease and high yield when compared to these five varieties. All the cleaned-reads in the FASTQ format can be retrieved from the European Nucleotide Archive (http://www.ebi.ac.uk/ena/browse/home) under accession number PRJEB34340.

## Identification of expressed genes and gene co-expression network analysis

DEGs (p-value < 0.05) were used as input data for the gene co-expression network analysis. R package (corr) was used to calculate the Pearson Correlation Coefficient (PCC) value for 721 DEGs. PCC is widely used to measure the strength of the linear relationship between the expression profiles of two genes [34]. In addition, PCC is one of the most convenient measures for biologist because it is easy to calculate [35] and is among the most performant for RNA-seq based co-expression studies [36].

In this study, 0.90 was chosen as the PCC cut-off by following the relationship between network density and the correlation coefficient as described by [15]. According to [15], the network density decreased as the cut-off value increased. The Network Analyzer plugin [37] in Cytoscape version 3.8 [38] was used to calculate the network density of gene co-expression network. The density of a gene co-expression network *D* was defined as a ratio of the actual number of links to all possible links of the non-singleton nodes [39]. We can compute the network density as undirected graphs:

$$D = \frac{2*E}{V(V-1)}$$

E: number of edges/number of actual links
V: number of vertices/number of possible links of the non singletone nodes

## Clustering and hub gene analysis

The Markov Cluster (MCL) algorithm was used to cluster the functional modules in the gene networks with an inflation (I) value of 1.5 (command line: mcl rice -I 1.5 -abc). We evaluated the cluster performance using varied I from 1.0 to. 2.0. In this study, I=1.5 produced clusters with gene ontology terms that biologically meaningful information. To identify the hub gene, the degree of connection among the co-expressed genes was calculated using NetworkAnalyzer plugin in Cytoscape version 3.8. The top 20% of co-expressed genes in the 10 clusters were identified as hub genes. The clusters were visualised using Cytoscape version 3.8. The functional category of hub genes was performed using STRING functional enrichment plugin in Cytoscape.

## Gene ontology and pathway analysis

The gene ontology (GO) and KEGG pathway enrichment analysis was performed using Fisher's exact test and FDR less than 0.05 (p-value < 0.05 & FDR < 0.05) in the OmicsBox programme version 2.1 (http://www.biobam.com/omicsbox). Fisher's exact test was applied to identify the significantly enriched GO categories and KEGG pathway in the network clusters.

## Validation of interaction using STRING database

STRING [40] database was used to validate the interaction between hub genes at the protein level. The parameters used were maximum interactor > 2 and confidence score cut-off = 0.4.

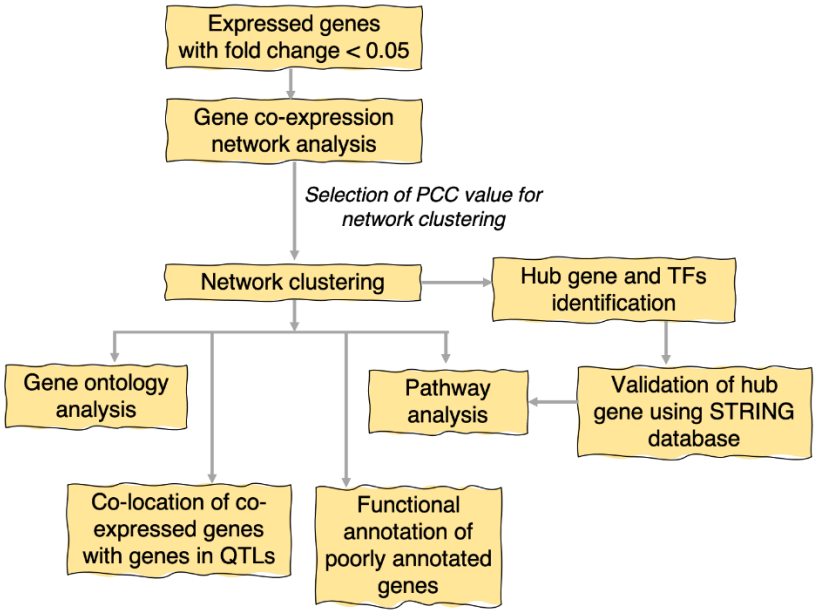



**Figure 1.** Flow chart of workflow for gene co-expression network analysis of pigmented and non-pigmented rice varieties

## Results and Discussion

### Selection of differentially expressed genes (DEGs) and construction of gene co-expression network

More than 10 million reads per sample were generated in this study [33]. A cut-off of 10 million reads per sample for RNA-seq has been suggested for sufficient sample size in the construction of gene co-expression network [41]. To reduce the noise in the gene co-expression network, we removed the low expressed genes [42] and select the DEGs (p-value < 0.05 and FDR < 0.05) for further analysis. A total of 721 (DEGs) (p-value < 0.05 and FDR < 0.05) from existing transcriptome analysis of pigmented and non-pigmented rice varieties were selected for this study (Supplementary Table 1). The EggNog analysis was performed to assign the 721 DEGs into EuKaryotic Orthologous Genes (KOG) categories. Out of 721 DEGs, 511 DEGs were annotated into 20 KOG categories (Supplementary Table 2). The highest number of DEGs was in the categories of 'Post-translational modification, protein turnover, chaperones' (66), followed by 'Transcription' (48) and 'Signal transduction mechanism' (44) categories (Figure 2).

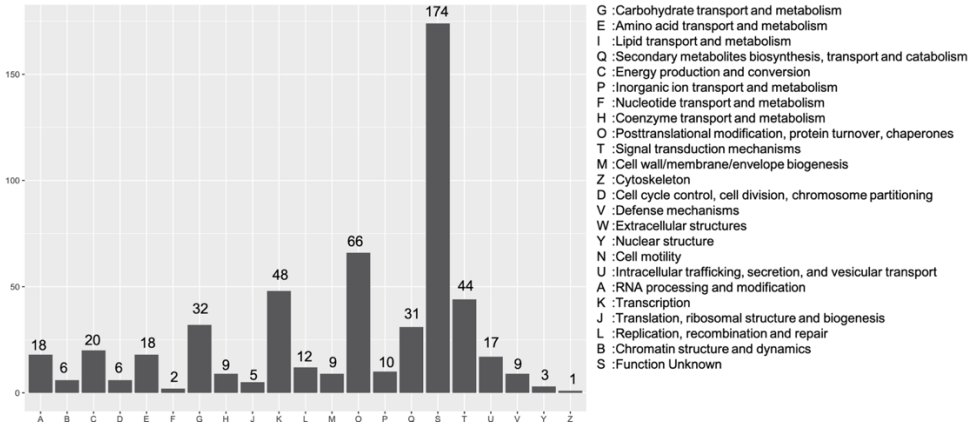



**Figure 2.** Distribution of 511 differentially expressed genes (DEGs) in EuKaryotic Orthologous Genes (KOG) categories

The DEGs in the 'signal transduction mechanisms' category indicate that the DEGs under this category are likely to be involved in stress-related traits, such as blast disease [43] and salinity tolerance [13]. The signal transduction mechanism also plays role in the phytohormone pathway, which positively induces various abiotic and biotic stresses [44]. For instance, in abiotic stress conditions, the up-regulated genes are responsive to abscisic acid (ABA), auxin, jasmonic acid (JA), and salicylic acid (SA). While in the biotics stress condition, the up-regulated genes are responsive to the same hormones, including cytokinin and ethylene. ABA, JA, and SA signallings regulate response to abiotic stresses [6]. JA and SA are positive regulators, and ABA tends to be a negative regulator of resistance to the pathogens [6].

The Pearson correlation coefficient (PCC) was used to calculate the correlation value (r) for all pair-wise combinations of 721 DEGs. A total of 614 DEGs were identified in the gene co-expression network of pigmented and non-pigmented rice. The remaining DEGs (98) were not included in the co-expressed genes, possibly due to the correlation value less than 0.1 (PCC < 0.1). The range of PCC values was from 0.1 to 1.0. As the PCC cut-off increased, the number of edges decreased and increased again as the PCC value larger than 0.70 (PCC > 0.70) (Table 1).

**Table 1.** Summary of nodes and edges in different range of PCC values

| PCC value | Nodes | Edges |
|---|---|---|
| 0.10-0.19 | 614 | 12,257 |
| 0.20-0.29 | 614 | 6608 |
| 0.30-0.39 | 614 | 6867 |
| 0.40-0.49 | 614 | 7237 |
| 0.50-0.59 | 614 | 7972 |
| 0.60-0.69 | 614 | 9519 |
| 0.70-0.79 | 614 | 11,610 |
| 0.80-0.89 | 614 | 14,158 |
| 0.90-0.99 | 613 | 15,259 |

Figure 3 shows the network density was at its minimum cut-off value of 0.2 and then increased as the cut-off value was greater than 0.5 and the highest at 0.9. Gene network properties, such as high density will be adopted as general biological network criterions for the rational selection of a cut-off in the network construction [45]. Higher density indicates higher associations in the network, which implies lower resilience to changes. We found that the network density decreased as the cut-off value increased, but the network density was shown to increase as the cut-off was greater than 0.90 (Figure 3). Hence, 0.9 was used as the cut-off in this study. The increase in network density was attributed to a high correlation value that connected to an increasing number of edges, indicating that biologically significant co-expression is expected to be found above the correlation value. The network contained 613 nodes and 15,259 edges at the PCC value cut-off of 0.9 (PCC > 0.9) (Supplementary Table 3).

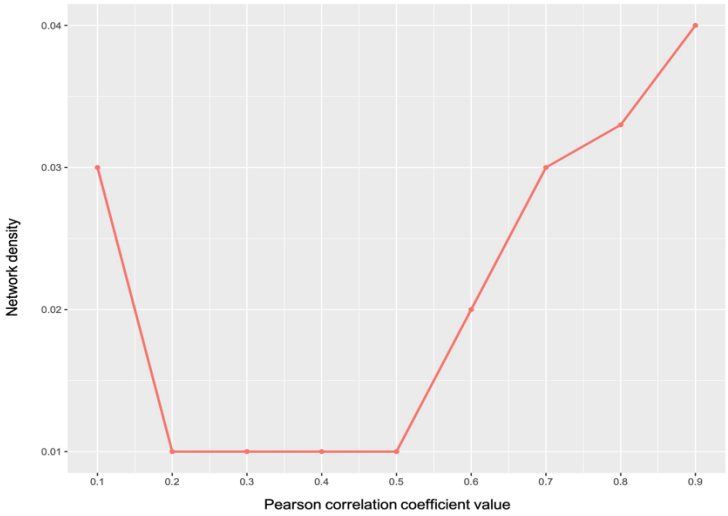



**Figure 3.** Network density at different correlation cut-off values

## Clusters identification and clusters related to biological functions

Using the MCL algorithm (inflation=1.5), 10 co-expression clusters were identified in the gene co-expression network of pigmented and non-pigmented rice (Figure 4 and Supplementary Table 4). We observed that the MCL I parameter of 1.5 produced the best clustering clusters with significance gene ontology terms. Previous studies have performed the MCL to evaluate the cluster performance with varied I parameter and select the the significant values based on biological enrichment analysis [12], [46]. Clusters with less than ten genes were removed because they are often biologically meaningless [47]. The MCL algorithm finds cluster structure in graphs by a mathematical bootstrapping procedure [48].

The 10 co-expression clusters ranged from two to 238 genes in the pigmented rice co-expressed network (Table 2). Genes belonging in the same cluster are likely to encode proteins with related functions or pathways [47]. The potential biological function of the identified clusters was assessed using GO terms. Out of 10 clusters, only four clusters showed significant GO terms with p-value < 0.05 (Figure 3 and Supplementary Table 5). None of the significant GO terms was found in Cluster 6, Cluster 7, Cluster 8, Cluster 9 and Cluster 10 (Figure 4).

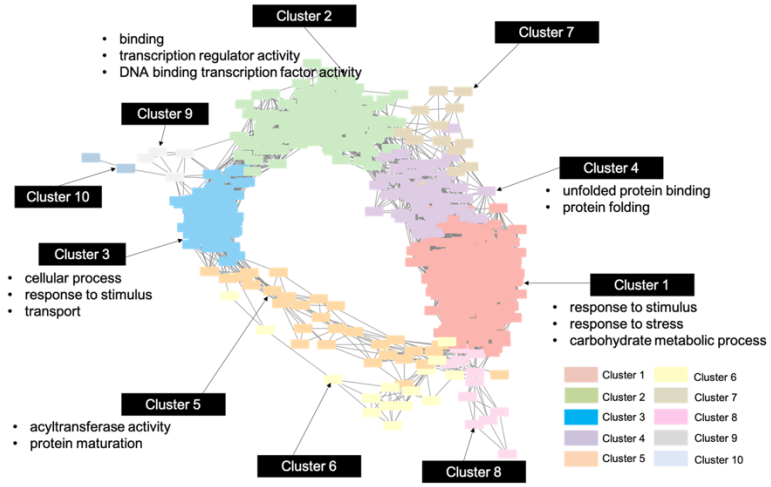



**Figure 4.** Ten clusters were identified using MCL algorithm. Different nodes colours represent different clusters

All the clusters were assessed using GO and KEGG pathway enrichment analysis. The primary goal of this analysis was to provide an insight into the biological processes that are related to pigmented and non-pigmented rice varieties and to prioritise the potential stress-related genes using the gene co-expression network and clustering approaches.

Cluster 1 is the largest in the co-expression network of pigmented and non-pigmented rice varieties, consisting of 238 genes (Table 2). The associated GO terms were response to stress (GO:0006950), response to stimulus (GO:0050896) and carbohydrate metabolic process (GO:0005975). The GO terms indicate that the represented genes are likely to be involved in stress-related traits. However, only one gene was up-regulated while 234 genes were down-regulated, indicating that these genes may serve as a negative regulator in rice response to stress conditions. A similar GO term (stress response) also was observed in Cluster 3. Out of 53 genes in Cluster 3, 90 genes were up-regulated while one of them was down-regulated. These two clusters suggest that genes in Cluster 1 and Cluster 3 may serve as key positive and negative players in rice response to stress conditions. The link of GO terms between Cluster 1 and Cluster 3 may represent robust interaction between these biological processes.

**Table 2.** Summary of 10 clusters in the gene co-expression network of Malaysian rice varieties, associated genes, as well as up and down-regulated genes in three treatments (BR, BW and RW)

| Clusters | Number of genes | Up-regulated genes | Down-regulated genes | Up & Down - regulated |
|---|---|---|---|---|
| C1 | 238 | 1 | 234 | 3 |
| C2 | 155 | 137 | 5 | 13 |
| C3 | 93 | 90 | 1 | 2 |
| C4 | 53 | 3 | 44 | 6 |
| C5 | 29 | 0 | 22 | 7 |
| C6 | 15 | 1 | 14 | 0 |
| C7 | 13 | 0 | 13 | 0 |
| C8 | 11 | 0 | 11 | 0 |
| C9 | 4 | 3 | 1 | 0 |
| C10 | 2 | 0 | 0 | 2 |

Genes in Cluster 2 are enriched for GO terms related to binding (GO:0005488), transcription regulator activity (GO:0140110) and DNA binding transcription factor activity (GO:0003700). Cluster 2 contains 155 genes, of which 137 were up-regulated, and five were down-regulated (Table 2). This finding indicates that these genes were positive regulators in transcription factor activity. GO terms such as protein binding (GO:0005515), and folding (GO:0006457) was enriched in Cluster 4, while GO terms such as acetyltransferase activity (GO:0016407) and protein maturation (GO:0051604) were enriched in Cluster 5 (Figure 4).

The GO enrichment analysis of 10 co-expression clusters suggests these genes are mainly involved in response to stress, metabolic process, transcription factor activity and protein binding. Cluster 4 and Cluster 5 appeared to be involved in distinctive functions. In contrast, Cluster 1, Cluster 2 and Cluster 3 appeared to be involved in similar functions. These findings may reflect the interconnectedness among the clusters due to the presence of similar gene functions in the clusters. Our clustering results suggest transcriptional coordination between genes in Cluster 2, which is involved in transcription regulator and DNA binding transcription factor activities. Genes that were associated with response to stress and response to stimulus were not consistently up- and down-regulated in Cluster 1 and Cluster 3. This pattern was observed in Cluster 1 and Cluster 3, in which more up-regulated genes were associated with response to stress in Cluster 3, while less up-regulated genes in Cluster 1. This finding suggests that Cluster 1 and Cluster 3 were a negative and positive players, respectively, in response to stress.

Pathway enrichment analysis showed DEGs in Cluster 1 were associated with the KEGG pathways, such as amino sugar and nucleotide sugar metabolism (ko00520), flavonoid biosynthesis (ko00941), biosynthesis of siderophore group (ko01053) and benzoate degradation (ko00362). Cluster 4 shows DEGs were enriched in linoleic acid metabolism (ko00591) and folate biosynthesis (ko00790), while DEGs in Cluster 7 were enriched in glyoxylate and dicarboxylate metabolism (ko00630), carbon fixation pathways in prokaryotes (ko00720) and citrate cycle (ko00020). The list of KEGG pathways enriched in Cluster 1, Cluster 4 and Cluster 7 are listed in Supplementary Table 6. Pathway enrichment analysis in Cluster 1, Cluster 4 and Cluster 7 showed pathways related to the biosynthesis of secondary metabolites (flavonoid biosynthesis) and cofactor and vitamin metabolism (folate biosynthesis). The secondary metabolites, cofactor and vitamin metabolism have been associated with resistance to pathogen and defence mechanisms [49]–[51].

## Hub genes identification

The highly connected gene is known as the hub gene. The hub gene within the networks plays an essential role in regulating and organising the biological function and mechanism. In this study, the top 20% degrees of connectivity were classified as hub genes. Previous study has reported that a gene was selected as a hub if it is connected to 10%, 20% or 30% of other genes in the whole network [52]. In addition, the 20% of hub genes in this study showed more genes with interesting biological information associated with the stress-related traits (Supplementary Table 7). A total of 119 hub genes with a degree of connectivity of more than 160 were identified in the gene co-expression network of pigmented rice varieties (Supplementary Table 7). Out of 119 hub genes, 111 were in Cluster 1, while eight were in Cluster 2.

A total of nine hub genes were associated with drought tolerance (Os01g065100, Os01g0727500, Os03g0820500, Os12g0150200, Os12g0555000), disease resistance (Os05g0476700), chalkiness

(Os05g0156900), hormone biosynthesis (Os01g0944700) and cold tolerance (Os05g0476700) traits. A total of six hub genes overlapped with quantitative trait loci (QTL) associated with physiological trait (*CSA*), morphological trait (*gh1, chalk5*) and resistance to abiotic and biotic stresses (*OsWRKY28, YK1, OsHI-XIP*). Eight (Os01g065100, Os01g0727500, Os03g0820500, Os12g0150200, Os12g0555000, Os05g0476700 and Os01g0944700,) candidate hub genes were associated with stress-related traits, which was identified in drought tolerance mechanism, disease resistance, hormone biosynthesis and cold tolerance. Interestingly, co-localisation of hub genes with QTLs revealed that five genes (*gh1, chalk5, OsWRKY28, YK1, OsHI-XIP*) are likely to have a functional impact on phenotypic expression. The *chalk5* or chalkiness 5 was found to control rice grain chalkiness by increasing chalkiness by disturbing pH homeostasis [53]. While *gh1* or *OsCHI* plays an essential role in flavonoid metabolism during the colouration of rice hulls [54]. Both *gh1* and *chalk5* were highly expressed in pigmented rice, indicating their essential roles in controlling pigmented rice's chalkiness and pigmentation process, respectively. In addition, DEG analysis of hub genes in three distinct rice groups, namely, BR, BW, and RW, mainly found that the top 20 hub genes were highly expressed in BR and BW groups (Supplementary Figure 1). These findings revealed that most hub genes showed stronger expressed patterns in pigmented rice compared to non-pigmented rice.

Ten transcription factors (TFs) were found as hub genes in Cluster 1 and Cluster 2, in which the degree of connectivity range was from 162 to 252. The TF families were bHLH, C2H2, CH3, GRAS, MYB, NF-YC and WRKY (Figure 5). The highly connected genes were observed on Cluster 1 and Cluster 2. Interestingly, *OsWRKY24* has been known as a blast disease responsive transcription factor, which positively regulates rice disease resistance [43]. This finding is consistent with the MR 297 (non-pigmented) variety which is resistant to blast disease caused by *Pyricularia oryzae* [55].

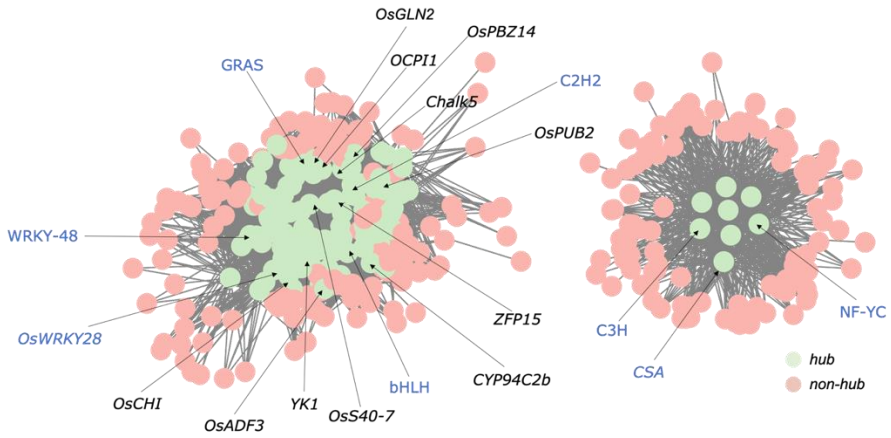



**Figure 5.** Hub gene of gene co-expression network in rice. Green colour represents hub gene. Pink colour represents non-hub genes but interacts with hub genes. The blue font denotes transcription factor families which found as hub genes. The hub genes were from Cluster 1 and Cluster 2

We also investigated the DEG encoding transcription factors in three distinct rice groups, namely, BR (black rice and red rice), BW (black rice and white rice), RW (red rice and white rice). Analysis of hub genes highlights ten TFs from different TF families, such as bHLH, C2H2, CH3, GRAS, MYB, NF-YC and WRKY. The involvement of these hub genes in Cluster 1 and Cluster 2 may increase their chance of master regulatory in each cluster. TFs in the WRKY family have been identified in abiotic stress, such as drought tolerance [7]. TFs in the MYB family have been known involved in salt and cold tolerance in rice [56]. In addition, TFs such as WRKY, MYB and bHLH regulate the genes that are responsible for secondary metabolite biosynthesis in plants during stress conditions [50].

Three TFs (*CSA*, Os01g0580400 and Os08g0491700) were highly expressed in BR and BW groups (Supplementary Figure 2). *CSA* is a TF from the MYB family, which was predicted as a regulator for genes involved in sugar partitioning in flowers, such as starch synthase *WAXY*. Os01g0580400 is a TF from the NF-YC family. Previous study has reported the role of NF-YC as DNA binding in the regulation of embryogenesis, flowering time, seed germination and stress tolerance [57]. Os08g0491700 is a TF from family C3H, and as putative zinc finger has been involved in growth, development and stress response [56]. In the gene co-expression network, TF hub genes are likely candidates for regulatory function. TF provides a clue for the co-regulation of gene subsets in similar biological processes.

## Validation of hub genes using protein-protein interaction (PPI) network in STRING database

The potential hub genes were evaluated with a protein-protein interaction (PPI) network, obtained from STRING database. The STRING database uses six major sources of interaction, including neighbourhood, fusion, co-occurance, co-expression, experimental database and text mining, to define the interaction between proteins using a probabilistic confidence score [40]. The combined score of all these available resources was used to estimate the interaction strength between proteins. It provides assessment and integration of protein-protein interactions. All the hub genes and their interactors were mapped to STRING, and only genes with a high interaction score of more than 0.4 (I > 0.4) were selected. A total of 47 hub proteins and 78 non-hub proteins denote as nodes 202 edges were found in STRING database (Figure 6). Pathway enrichment analysis revealed that flavonoid biosynthesis, phenylpropanoid biosynthesis, terpenoid biosynthesis and MAPK signalling pathway-plant were enriched in the PPI network (Figure 6). Four metabolism pathways (i.e. glutathione metabolism, amino sugar and nucleotide metabolism, ascorbate and alderate metabolism) were enriched in the PPI network (Figure 6).

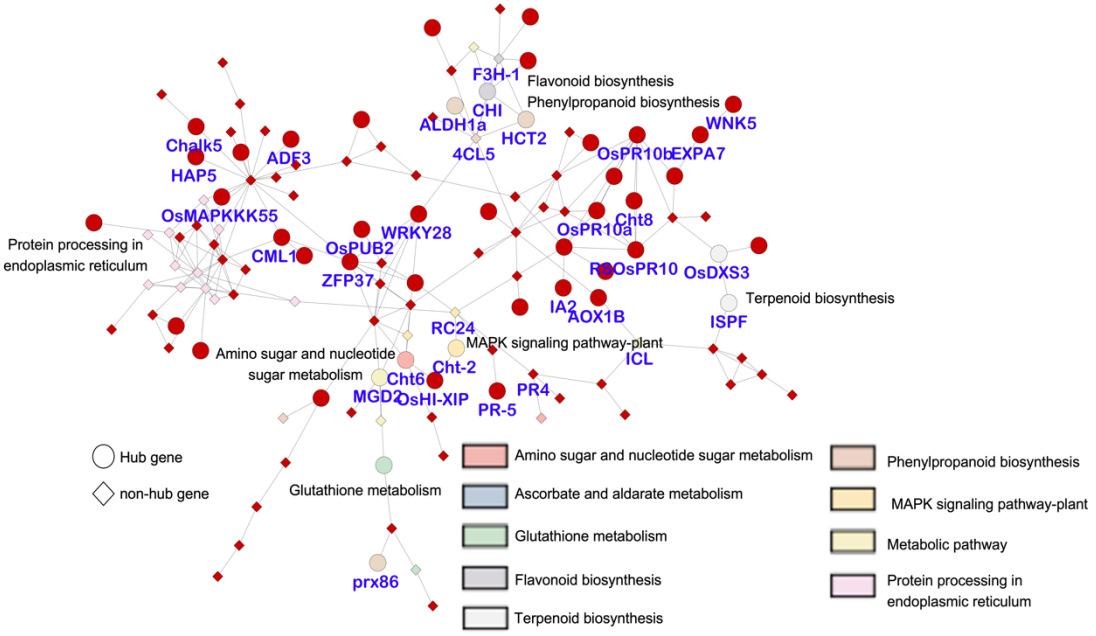



**Figure 6**. Pathway analysis of hub genes and their interactors. Each node represents a gene. Red nodes in a big circle indicate hub genes, and red diamonds indicate connections. Different colors for each node represents pathway name

Predicted PPI networks from the STRING database support the interaction network for 47 hub genes and 78 non-hub genes, suggesting the potential hub proteins as the regulator in flavonoid biosynthesis, terpenoid biosynthesis and mitogen-activated protein kinase (MAPK) signalling pathway. The secondary metabolites such as flavonoid and terpenoid have been associated with stress and defence response mechanisms in plants [50]. Flavonoid and terpenoid provide colour and scent properties to plants, which has repellent and attractive effects on insects and herbivores [50]. The MAPK signalling pathway serves a central role in the intracelluar signal transduction pathways and regulates infection and virulence in plants [51]. Several hub proteins related to the biosynthesis of secondary metabolites (i.e. flavonoid and terpenoid biosynthesis) also have shown that they were highly correlated in the interaction network, indicating these proteins are represented by pigmented rice.

Although the breeding programme of Malaysian rice varieties has been extensively performed, the molecular networks underlying pigmented and non-pigmented rice varieties have been lacking. Hence, this study is valuable for unravelling the hidden molecular information and providing significant information resources for further application in rice breeding programme. In addition, the potential genes highlighted in this study can also be used as a reference to identify the essential genes in response to stress traits, such as abiotic and biotic stresses.

## Conclusions

Gene co-expression network and MCL have been shown to be essential approaches for prioritising of candidate stress-related genes. Through this approach, we screened DEGs for identifying and prioritising the stress-related genes using gene co-expression network and clustering approaches, as well as protein-protein interaction network from the STRING database. Fifteen potential stress-related genes and transcription factors (i.e., Os01g065100, Os01g0727500, Os03g0820500, Os12g0150200, Os12g0555000, Os05g0476700, *OsWRKY28, YK1, OsHI-XIP*) from the three significant clusters that were enriched with GO terms and pathways related to stress-related traits could be prioritised for further research, such as functional genomics and gene editing. The stress-related genes also were highly expressed in pigmented rice varieties, indicating pigmented rice can be candidate varieties to be introduced in rice breeding for stress-resistant or tolerant traits. Ultimately, this work has enriched molecular information of Malaysian pigmented and non-pigmented rice varieties, which will benefit future work in areas of improving rice stress-resistant or tolerant traits.

## Conflicts of Interest

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

## Acknowledgment

This research was carried out in the Malaysian Agricultural Research & Development Institute (MARDI). This work was supported by Malaysian Agricultural Research & Development Institute (MARDI) [PRB-401]. This work has been undertaken as part of the Pembangunan Research Project on Rice Specialty under The Eleventh Malaysia Plan (RMK-11).

## References

[1] Y. Shao, F. Xu, X. Sun, J. Bao, and T. Beta. (2014). Phenolic acids, anthocyanins, and antioxidant capacity in rice (*Oryza sativa* l.) grains at four stages of development after flowering. *Food Chem.*, *143*, 90–96. DOI: 10.1016/j.foodchem.2013.07.042.

[2] X. Chen *et al.*(2019). Transcriptome and Proteome Profiling of Different Colored Rice Reveals Physiological Dynamics Involved in the Flavonoid Pathway. *Int. J. Mol. Sci.*, *20*(2463), 1–23. DOI: 10.3390/ijms20102463.

[3] N. Ithal and A. R. Reddy. (2004). Rice flavonoid pathway genes, *OsDfr* and *OsAns*, are induced by dehydration, high salt and ABA, and contain stress responsive promoter elements that interact with the transcription activator, OsC1-MYB. *Plant Sci.*, *166*(6), 1505–1513. DOI: 10.1016/j.plantsci.2004.02.002.

[4] L. Yang, L. Lei, H. L. Liu, J. Wang, H. Zheng, and D. Zou. (2020). Whole-genome mining of abiotic stress gene loci in rice. *Planta*, *252*(5), 1–20. DOI: 10.1007/s00425-020-03488-x.

[5] A. Raza, J. Tabassum, H. Kudapa, and R. K. Varshney. (2021). Can omics deliver temperature resilient ready-to-grow crops? *Crit. Rev. Biotechnol.*, 1–33. DOI: 10.1080/07388551.2021.1898332.

[6] K. Vishwakarma *et al.* (2017). Abscisic acid signaling and abiotic stress tolerance in plants: A review on current knowledge and future prospects. *Front. Plant Sci., 8*, 1–12. DOI: 10.3389/fpls.2017.00161.

[7] D. L. Rushton *et al.* (2012). WRKY transcription factors: Key components in abscisic acid signalling, *Plant Biotechnol. J.*, *10*(1), 2–11. DOI: 10.1111/j.1467-7652.2011.00634.x.

[8] N. Li *et al.* (2018). Transcriptome analysis of two contrasting rice cultivars during alkaline stress. *Sci. Rep.*, *8*(1), 1–16. DOI: 10.1038/s41598-018-27940-x.

[9] W. Wang, Y. Li, P. Dang, S. Zhao, D. Lai, and L. Zhou. (2018). Rice Secondary Metabolites: Structures, Roles, Biosynthesis, and Metabolic Regulation. *Molecules*, *23*, 1-50. DOI: 10.3390/molecules23123098.

[10] E. Petrussa *et al.* (2013). Plant flavonoids-biosynthesis, transport and involvement in stress responses. *Int. J. Mol. Sci.*, *14*(7), 14950–14973. DOI: 10.3390/ijms140714950.

[11] M. Jain. (2012). Next-generation sequencing technologies for gene expression profiling in plants. *Brief. Funct. Genomics*, *11*(1), 63–70. DOI: 10.1093/bfgp/elr038.

[12] L. Zhang, S. Yu, K. Zuo, L. Luo, and K. Tang. (2012). Identification of gene modules associated with drought response in rice by network-based analysis. *PLoS One*, *7*(5), 1–12. DOI: 10.1371/journal.pone.0033748.

[13] L. Zhang *et al.* (2019). Comprehensive meta-analysis and co-expression network analysis identify candidate genes for salt stress response in Arabidopsis. *Plant Biosyst.*, *153*(3), 367–377. DOI: 10.1080/11263504.2018.1492989.

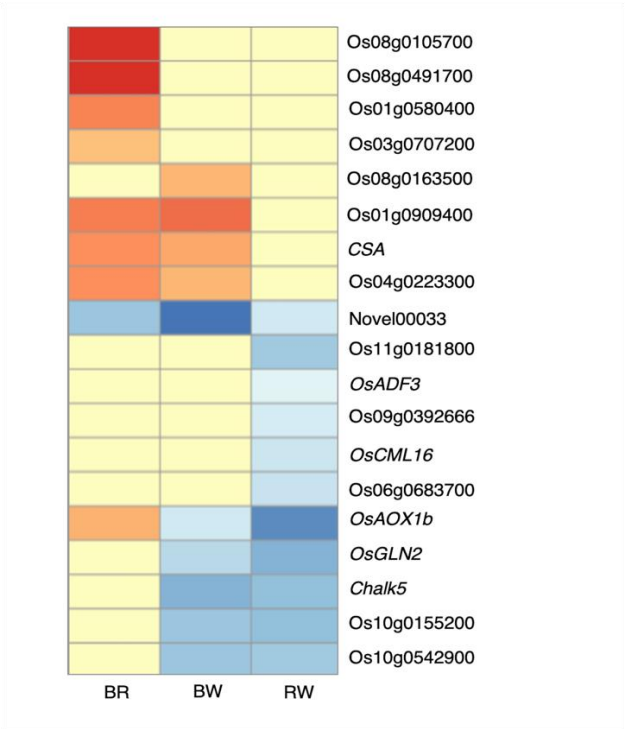
[14] S. Smita, A. Katiyar, D. M. Pandey, V. Chinnusamy, S. Archak, and K. C. Bansal. (2013). Identification of conserved drought stress responsive gene-network across tissues and developmental stages in rice. *Bioinformation*, *9*(2), 72–78. DOI: 10.6026/97320630009072.

[15] K. Aoki, Y. Ogata, and D. Shibata. (2007). Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant Cell Physiol.*, *48*(3), 381–390. DOI: 10.1093/pcp/pcm013.

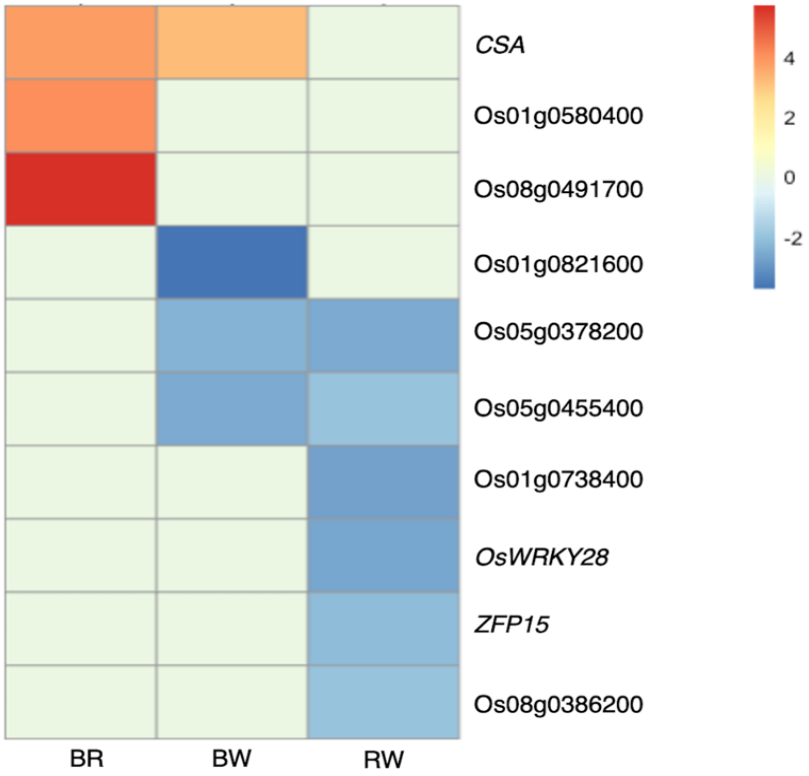
[16] E. A. R. Serin, H. Nijveen, H. W. M. Hilhorst, and W. Ligterink. (2016). Learning from co-expression networks:

possibilities and challenges, *Front. Plant Sci.*, *7*, 1–18. DOI: 10.3389/fpls.2016.00444.
[17] A. Fukushima, S. Kanaya, and M. Arita. (2009). Characterizing gene coexpression modules in Oryza sativa based on a graph-clustering approach, *Plant Biotechnol.*, *26*, 485-493.
[18] Q. You *et al.* (2016). Co-expression network analyses identify functional modules associated with development and stress response in *Gossypium arboreum*. *Sci. Rep.*, *6*, 1–15. DOI: 10.1038/srep38436.
[19] R. J. Schaefer *et al.* (2018). Integrating coexpression networks with GWAS to prioritize causal genes in Maize, *Plant Cell*, *30*(12), 2922–2942. DOI: 10.1105/tpc.18.00299.
[20] A. Suratanee *et al.* (2018). Two-state co-expression network analysis to identify genes related to salt tolerance in Thai rice. *Genes (Basel).*, *9*(12), 1–21. DOI: 10.3390/genes9120594.
[21] J. Du, S. Wang, C. He, B. Zhou, Y. L. Ruan, and H. Shou. (2017). Identifcation of regulatory networks and hub genes controlling soybean seed set and size using RNA sequencing analysis. *J. Exp. Bot.*, *68*(8), 1955–1972. DOI: 10.1093/jxb/erw460.
[22] Z. Lei *et al.* (2018). Transcriptome Analysis Reveals genes involved in flavonoid biosynthesis and accumulation in *Dendrobium catenatum* from different locations, *Sci. Rep.*, *8*(1), 1–16. DOI: 10.1038/s41598-018-24751-y.
[23] F. He and S. Maslov. (2016). Pan- and core- network analysis of co-expression genes in a model plant, *Sci. Rep.*, *6*(8), 1–11. DOI: 10.1038/srep38956.
[24] W. Hu *et al.* (2020). Gene co-expression network analysis provides a novel insight into the dynamic response of wheat to powdery mildew stress. *J. Genet.*, *99*(44), 1–12. DOI: 10.1007/s12041-020-01206-w.
[25] N. K. Sarkar, Y. K. Kim, and A. Grover. (2014). Coexpression network analysis associated with call of rice seedlings for encountering heat stress, *Plant Mol. Biol.*, *84*(1–2), 125–143. DOI: 10.1007/s11103-013-0123-3.
[26] S. Sircar and N. Parekh. (2019). *Meta-analysis of drought-tolerant genotypes in Oryza sativa: A network-based approach*, *14*(5), 1-7. DOI: 10.1371/journal.pone.0216068.
[27] M. R. Abdullah-Zawawi, L. W. Tan, Z. Ab Rahman, I. Ismail, and Z. Zainal. (2022). An Integration of transcriptomic data and modular gene co-expression network analysis uncovers drought stress-related hub genes in transgenic rice overexpressing OsAbp57. *Agronomy*, *12*(8), 1-20. DOI: 10.3390/agronomy12081959.
[28] Z. Zeng, S. Zhang, W. Li, B. Chen, and W. Li. (2022). Gene-coexpression network analysis identifies specific modules and hub genes related to cold stress in rice. *BMC Genomics*, 23(1), 1–18. DOI: 10.1186/s12864-022-08438-3.
[29] H. Takehisa, Y. Sato, B. Antonio, and Y. Nagamura. (2015). Coexpression network analysis of macronutrient deficiency response genes in rice. *Rice*, 8(24), 1-7. DOI: 10.1186/s12284-015-0059-0.
[30] O. Contreras-López, T. C. Moyano, D. C. Soto, and R. A. Gutiérrez. (2018). Step-by-step construction of gene co-expression networks from high-throughput Arabidopsis RNA sequencing data. *Methods Mol. Biol.*, *1761*, 275–301. DOI: 10.1007/978-1-4939-7747-5_21.
[31] M. Li, D. Li, Y. Tang, F. Wu, and J. Wang. (2017). Cytocluster: A cytoscape plugin for cluster analysis and visualization of biological networks. *Int. J. Mol. Sci.*, *18*(9), 1-6. DOI: 10.3390/ijms18091880.
[32] S. Yun-Shin *et al.* (2018). Transcriptome analysis of pigmented and non-pigmented rice grain-focusing on antioxidant and micronutrient perspectives. *Proc. International Conference on Biochemistry, Molecular Biology and Biotechnology 11, 15-16 August 2018, Kuala Lumpu*r, 11.
[33] R. A. Zainal-Abidin *et al.* (2020). RNA-seq data from whole rice grains of pigmented and non-pigmented Malaysian rice varieties. *Data Br.*, *30*, 105432. DOI: 10.1016/j.dib.2020.105432.
[34] X. Rao and R. A. Dixon. (2019). Co-expression networks for plant biology: Why and how. *Acta Biochim. Biophys. Sin. (Shanghai).*, *51*(10), 981–988. DOI: 10.1093/abbs/gmz080.
[35] T. Obayashi and K. Kinoshita. (2009). Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. *DNA Res.*, *16*(5), 249–260. DOI: 10.1093/dnares/dsp016.
[36] S. Ballouz, W. Verleyen, and J. Gillis. (2015). Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. *Bioinformatics*, *31*(13), 2123–2130. DOI: 10.1093/bioinformatics/btv118.
[37] Y. Assenov, F. Ramírez, S. E. S. E. Schelhorn, T. Lengauer, and M. Albrecht. (2008). Computing topological parameters of biological networks. *Bioinformatics*, *24*(2), 282–284. DOI: 10.1093/bioinformatics/btm554.
[38] P. Shannon *et al.* (2008). Cytoscape: A software environment for integrated models. *Genome Res.*, *13*(22), 2498–2504. DOI: 10.1101/gr.1239303.metabolite.
[39] L. Hua, Z. Yang, and J. Shao. (2022). Impact of network density on the efficiency of innovation networks: An agent-based simulation study. *PLoS One*, *17*(6), 1–22. DOI: 10.1371/journal.pone.0270087.
[40] D. Szklarczyk *et al.* (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, *47*(D1), D607–D613. DOI: 10.1093/nar/gky1131.
[41] S. van Dam, U. Võsa, A. van der Graaf, L. Franke, and J. P. de Magalhães. (2018). Gene co-expression analysis for functional classification and gene-disease predictions. *Brief. Bioinform.*, *194*, 575–592. DOI: 10.1093/bib/bbw139.
[42] W. Chen *et al.* (2022). GCEN: An easy-to-use toolkit for gene co-expression network analysis and lncRNAs annotation. *Curr. Issues Mol. Biol.*, *44*(4), 1479–1487. DOI: 10.3390/cimb44040100.
[43] F. Wang *et al.* (2016). Enhanced rice blast resistance by CRISPR/ Cas9-targeted mutagenesis of the ERF transcription factor gene OsERF922. *PLoS One*, *11*(4), 1–18. DOI: 10.1371/journal.pone.0154027.
[44] S. Simm, K. D. Scharf, S. Jegadeesan, M. L. Chiusano, N. Firon, and E. Schleiff. (2016). Survey of genes involved in biosynthesis, transport, and signaling of phytohormones with focus on solanum lycopersicum. *Bioinform. Biol. Insights*, *10*,185–207. DOI: 10.4137/BBI.S38425.
[45] R. Zhang *et al.* (2020). TeaCoN: A database of gene co-expression network for tea plant (Camellia sinensis). *BMC Genomics*, *21*(1), 1–9. DOI: 10.1186/s12864-020-06839-w.
[46] S. Smita *et al.* (2020). Gene network modules associated with abiotic stress response in tolerant rice genotypes identified by transcriptome meta-analysis. *Funct. Integr. Genomics*, *20*(1), 29–49. DOI:

10.1007/s10142-019-00697-w.
[47] M. Mutwil *et al.* (2011). PlaNet: Combined sequence and expression comparisons across plant networks derived from seven species. *Plant Cell*, *23*(3), 895–910. DOI: 10.1105/tpc.111.083667.
[48] S. van Dongen and C. Abreu-Goodger. (2012). Using MCL to extract clusters from networks. *Methods Mol. Biol.*, *804*, 281–295. DOI: 10.1007/978-1-61779-361-5_15.
[49] L. F. De Filippis, in: Azooz, M. M., and Ahmad, P. (Eds.). (2016). Plant secondary metabolites: From molecular biology to health products. *Plant-Environment Interaction: Responses and Approaches to Mitigate Stress*, 263–300.
[50] R. Jan, S. Asaf, M. Numan, Lubna, and K. M. Kim. (2021). Plant secondary metabolite biosynthesis and transcriptional regulation in response to biotic and abiotic stress conditions. *Agronomy*, *11*(5), 1–31. DOI: 10.3390/agronomy11050968.
[51] T. Isah. (2019). Stress and defense responses in plant secondary metabolites production. *Biol. Res.*, *52*(1), 1-32. DOI: 10.1186/s40659-019-0246-3.
[52] S. Mahapatra, B. Mandal, and T. Swarnkar. (2018). Biological networks integration based on dense module identification for gene prioritization from microarray data. *Gene Reports*, *12*: 276–288. DOI: 10.1016/j.genrep.2018.07.008.
[53] Y. Li *et al.* (2014). Chalk5 encodes a vacuolar H(+)-translocating pyrophosphatase influencing grain chalkiness in rice. *Nat. Genet.*, *46*(4), 398–404. DOI: 10.1038/ng.2923.
[54] L. Hong, Q. Qian, D. Tang, K. Wang, M. Li, and Z. Cheng. (2012). A mutation in the rice chalcone isomerase gene causes the golden hull and internode 1 phenotype. *Planta*, *236*(1), 141–151. DOI: 10.1007/s00425-012-1598-x.
[55] L. Zakaria and N. Misman. (2018). The pathogen and control management of rice blast disease. *Malays. J. Microbiol.*, *14*(7), 705–714. DOI: 10.21161/mjm.113717.
[56] Q. Ma *et al.* (2009). Enhanced tolerance to chilling stress in OsMYB3R-2 transgenic rice is mediated by alteration in cell cycle and ectopic expression of stress genes. *Plant Physiol.*, *150*(1), 244–256. DOI: 10.1104/pp.108.133454.
[57] W. Yang, Z. Lu, Y. Xiong, and J. Yao. (2017). Genome-wide identification and co-expression network analysis of the OsNF-Y gene family in rice. *Crop J.*, *5*(1),. 21–31. DOI: 10.1016/j.cj.2016.06.014.




**Supplementary Figure 1**. Heat map of top hub genes that were differentially expressed in black vs red, black vs white and red vs white rice varieties




**Supplementary Figure 2**. Heat map of differentially expressed transcription factors in BR (black rice and red rice), BW (black rice and white rice), RW (red rice and white rice)