

## Article

# Drought Stress-Related Gene Identification in Rice by Random Walk with Restart on Multiplex Biological Networks

Liu Zhu <sup>1</sup> , Hongyan Zhang <sup>1,2,\*</sup>, Dan Cao <sup>3</sup>, Yalan Xu <sup>1</sup>, Lanzhi Li <sup>2</sup>, Zilan Ning <sup>1</sup> and Lei Zhu <sup>1</sup> 

<sup>1</sup> College of Information and Intelligence, Hunan Agricultural University, Changsha 410128, China

<sup>2</sup> Hunan Engineering and Technology Research Center for Agricultural Big Data Analysis and Decision-Making, Hunan Agricultural University, Changsha 410128, China

<sup>3</sup> College of Science, Central South University of Forestry and Technology, Changsha 410004, China

\* Correspondence: hongyan\_zhang@hunau.edu.cn; Tel.: +86-139-7313-6716

**Abstract:** Drought stress-related gene identification is vital in revealing the drought resistance mechanisms underlying rice and for cultivating rice-resistant varieties. Traditional methods, such as Genome-Wide Association Studies (GWAS), usually identify hundreds of candidate stress genes, and further validation by biological experiments is then time-consuming and laborious. However, computational and prioritization methods can effectively reduce the number of candidate stress genes. This study introduces a random walk with restart algorithm (RWR), a state-of-the-art guilt-by-association method, to operate on rice multiplex biological networks. It explores the physical and functional interactions between biological molecules at different levels and prioritizes a set of potential genes. Firstly, we integrated a Protein–Protein Interaction (PPI) network, constructed by multiple protein interaction data, with a gene coexpression network into a multiplex network. Then, we implemented the RWR on multiplex networks (RWR-M) with known drought stress genes as seed nodes to identify potential drought stress-related genes. Finally, we conducted association analysis between the potential genes and the known drought stress genes. Thirteen genes were identified as rice drought stress-related genes, five of which have been reported in the recent literature to be involved in drought stress resistance mechanisms.

**Keywords:** rice; protein–protein interaction; coexpression network; drought stress gene; random walk with restart



**Citation:** Zhu, L.; Zhang, H.; Cao, D.; Xu, Y.; Ning, Z.; Zhu, L.; Li, L. Drought Stress-Related Gene Identification in Rice by Random Walk with Restart on Multiplex Biological Networks. *Agriculture* **2023**, *13*, 53. <https://doi.org/10.3390/agriculture13010053>

Academic Editors: Zhenyu Gao, Qingpo Liu and Youlin Peng

Received: 17 November 2022

Revised: 14 December 2022

Accepted: 20 December 2022

Published: 24 December 2022



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

As an essential food crop, rice has a wide cultivation area worldwide. According to data from the Food and Agriculture Organization of the United Nations, the global rice planting area is about 167 million hectares, and the total production is about 750 million tons [1]. Rice demand is expected to increase as the population increases [2]. However, the yield of rice is severely affected by drought stress. Improving the drought stress resistance of rice will expand the planting area and increase rice production. Therefore, growing rice varieties with drought stress resistance is essential to ensure food security.

Usually, rice drought tolerance is affected by multiple genes. With the development of high-throughput sequencing technology, many biological omics data and biomolecular interactions have been explored. Utilizing these technologies sheds light on the drought tolerance mechanism and the development of drought-tolerant varieties. Previous research mainly focused on Genome-Wide Association Studies (GWAS) [3], weighted gene coexpression network analysis (WGCNA) [4], and gene-by-gene association analysis [5,6]. The WGCNA clusters genes with similar expression patterns and analyzes the relationship between modules and specific traits to find hub genes in each module. The GWAS carries out overall association analysis on genetic variation genes within the whole genome, which can provide an overview of abiotic stress and discover new abiotic stress mechanisms [7]. However, GWAS usually generates hundreds of candidate genes, which need a lot of

biological experiments for verification. The WGCNA-obtained coexpression module is unstable if the sample data are too small. As a state-of-the-art association-by-guilt algorithm, the random walk with restart algorithm (RWR) [8,9] can quickly and effectively predict the degree of association between nodes and the starting seed points in the network. In this manner, the RWR effectively reduces the number of candidate genes and predicts the candidate genes' functional role based on the seed nodes. The RWR has become a powerful approach to identify human disease-related genes [10]; however, its ability to identify rice resistance-related genes has rarely been studied. Given its excellent performance in predicting pathogenic genes, we believe that the RWR is promising for rice drought stress-related gene identifications.

In this study, we fused data from multiple rice public PPI databases, which expanded the information in the PPI network and reduced the sample bias and noise. At the same time, the rice coexpression network was constructed using RNA-seq data. With each node corresponding to the genes or the derived proteins, we merged the PPI and coexpression networks into a two-layer network, allowing multiple edges between a pair of nodes. Then, the RWR on multiplex networks (RWR-M) [8] was applied to this multiplex network to identify rice drought stress-related potential genes. Next, potential genes were analyzed in enrichment analysis and association with known drought stress genes to identify stress-related candidate genes. To validate the findings, a support vector machine (SVM) [11] model was constructed using the candidate genes to predict rice phenotypes. As external validation, we reviewed the recent literature to understand the regulatory mechanisms of the candidate genes.

## 2. Materials and Methods

### 2.1. Dataset

We extracted the rice protein interaction information from the STRING database (<https://cn.string-db.org>) [12], RicePPINet (<http://netbio.sjtu.edu.cn>) [13], and PRIN database (<http://bis.zju.edu.cn/prin/>) [14]. Then, we matched the protein names to gene names using the RAP-DB [15] database, which provides a comprehensive set of gene annotations for the rice genome sequences. In addition, genes with mismatched names were removed. Finally, the numbers of PPIs obtained from STRING, RicePPINet, and PRIN were 8949049, 673489, and 55211, respectively.

We downloaded the rice RNA-seq data from NCBI's Sequence Read Archive (SRA) database [16] to construct a gene coexpression network. The data included ten drought stress treatment samples and ten normal samples (SRR7054176-83, SRR3051740-45, and SRR30517527-57), for a total of 20 samples of about 33,688 genes.

The China Rice Data Center (<https://www.ricedata.cn>) is a rice-themed database sponsored by the China National Rice Research Institute. In total, 218 genes known to be related to drought stress were in the China Rice Data Center as of June 2022, each denoted as  $GK_i (i = 1, 2, \dots, 218)$ . These 218 known genes were considered as the seed nodes in this study.

### 2.2. The Networks' Construction

#### 2.2.1. Protein–Protein Interaction Network

We constructed a rice PPI network based on the STRING database, named SinglePPI. In addition, we obtained the PPI from the PRIN and RicePPINet databases to provide complementary information to the STRING database. Each of the PPIs provides pairwise protein interaction scores. We normalized the interaction scores and then connected a pair of nodes when their interaction score in any database was higher than 0.3. An edge connecting two nodes meant there was an interaction between the corresponding proteins. Finally, a multisource network was constructed by these PPI, named MetaPPI, which consisted of 23,860 nodes and 3,096,015 edges.

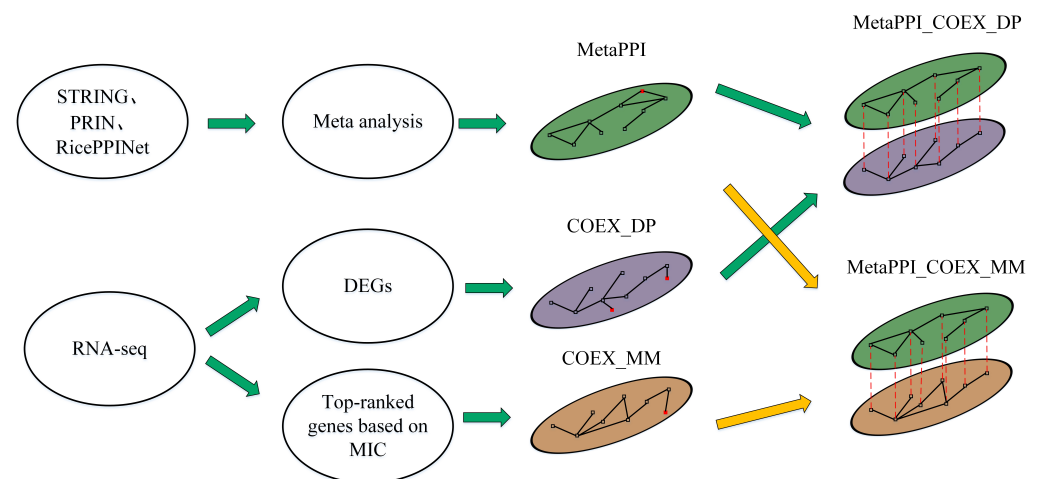
### 2.2.2. Gene Coexpression Network

We constructed rice gene coexpression networks using the RNA-seq data. Differentially expressed genes (DEGs) were identified if  $\text{padj} < 0.05$  and  $|\log_2\text{FoldChange}| > 1$  [17]. Using the Pearson coefficient [18] between DEGs, we constructed a coexpression network named COEX\_DP, which was composed of 8928 nodes and 1,0267,919 edges.

The maximum information coefficient (MIC) [19,20] captures a wide range of correlations, including linear, nonlinear, and nonfunctional correlations [21]. The higher the MIC value, the stronger the correlation between two features. In this study, we calculated the MIC values between each DEG and the phenotype and selected those genes ranked in the top 30% of the MIC values. Then, based on the MIC between any pair of the selected genes, we constructed another coexpression network, named COEX\_MM. It consisted of 8949 nodes and 24,352,943 edges.

### 2.2.3. Multiplex Network

A PPI network and a gene coexpression network were merged into a multiplex network, sharing the same set of nodes and different edges [22]. In this study, the MetaPPI\_COEX\_DP comprised the MetaPPI and COEX\_DP, which consisted of 14972 nodes and 7682248 edges; the MetaPPI\_COEX\_MM comprised the MetaPPI and COEX\_MM, which consisted of 14298 nodes and 16491940 edges. The process of the network construction is illustrated in Figure 1.



**Figure 1.** The network's construction process. In the process of network convergence, we retained the intersection of nodes across the PPI and coexpression networks. A node removed was color-coded red.

### 2.3. Random Walk with Restart on Multiplex Networks

The random walk is an effective approach to calculate the proximity between nodes and extract the topology structure of the network [23]. It describes a path that involves a series of random steps in a mathematical space. Its extension, random walk with restart (RWR) (Pseudocode is shown in Algorithm 1) [24,25], allows random walk to return to the starting point and continues to execute during the random walk process within a certain probability.

However, the existing algorithm was not applicable to multi-omics data. The RWR with multiplex networks (RWR-M) effectively accommodated multi-omics data and improved the prediction accuracy of a gene's functional role. The walk traveled between different nodes in the same layer or jumped to a different layer on the same node, which reduced the prediction error caused by the network structure.

**Algorithm 1** Random Walk with Restart (RWR).

**Input:** a restart probability  $\lambda$ , a transition matrix  $M$ , an initial probability vector  $\vec{P}_0^T$ , and a threshold  $Threshold$

**Output:** A vector indicating the importance of nodes in the network:  $\vec{P}_{t+1}^T$

```

1: function RWR( $M, \lambda, \vec{P}_0^T$ )
2:    $Threshold \leftarrow 10^{-5}$ 
3:    $t \leftarrow 1$ 
4:    $\vec{P}_1^T \leftarrow (1 - \lambda)M\vec{P}_0^T + \lambda\vec{P}_0^T$ 
5:    $Thre \leftarrow \sqrt{\sum (\vec{P}_1^T - \vec{P}_0^T)^2}$ 
6:   while  $Thre \geq Threshold$  do
7:      $\vec{P}_{t+1}^T \leftarrow (1 - \lambda)M\vec{P}_t^T + \lambda\vec{P}_0^T$ 
8:      $Thre \leftarrow \sqrt{\sum (\vec{P}_{t+1}^T - \vec{P}_t^T)^2}$ 
9:      $t \leftarrow t + 1$ 
10:  end while
11:  return  $\vec{P}_{t+1}^T$ 
12: end function

```

A multiplex network is an  $L$ -layer network. The edges in different layers belong to different categories or represent different properties. Each layer  $\alpha = 1, \dots, L$ , is represented by its adjacency matrix,  $A^{[\alpha]}(i, j) = 0, \forall i = 1, \dots, n$ ; if there is an interaction between nodes  $i$  and  $j$  in  $\alpha$  layer, then  $A^{[\alpha]}(i, j) = 1$ , otherwise 0. The automatic interactions were not considered in this study; then,  $A^{[\alpha]}(i, i) = 0, \forall i = 1, \dots, n$ , and a multiplex graph is characterized by its adjacency matrix:

$$A = A^{[1]}, \dots, A^{[L]}. \quad (1)$$

The random walk with restart on a multiplex network randomly walked within the same layers and had a certain probability to jump to another layer on the same node. Each layer of a network is regarded as an  $n * n$  matrix, and an  $nL * nL$  transfer matrix was constructed in the multiplex networks:

$$A = \begin{pmatrix} (1 - \delta)A^{[1]} & \frac{\delta}{L-1}\mathbf{I} & \cdots & \frac{\delta}{(L-1)}\mathbf{I} \\ \frac{\delta}{(L-1)}\mathbf{I} & (1 - \delta)A^{[2]} & \cdots & \frac{\delta}{(L-1)}\mathbf{I} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\delta}{(L-1)}\mathbf{I} & \frac{\delta}{(L-1)}\mathbf{I} & \cdots & (1 - \delta)A^{[L]} \end{pmatrix}, \quad (2)$$

where  $\mathbf{I}$  is the  $n * n$  identity matrix,  $A^{[\alpha]}$  is the adjacency matrix of the  $\alpha$ th layer, the parameter  $\delta \in [0, 1]$  quantifies the probability of staying at each layer or jumping to another layer, and  $\delta = 0$  indicates that the particle always stays in the same layer without switching layers [26].

When using  $M$  to represent the column-normalized transition matrix obtained by  $A$ , the RWR-M is expressed as:

$$\vec{P}_{t+1}^T = (1 - \lambda)M\vec{P}_t^T + \lambda\vec{P}_{RS}^T. \quad (3)$$

The  $\vec{P}_{t+1} = [\vec{P}_{t+1}^1, \dots, \vec{P}_{t+1}^L]$  and  $\vec{P}_t = [\vec{P}_t^1, \dots, \vec{P}_t^L], t \in \mathbb{N}$  are vectors of length  $n * L$ , representing the probability distribution of particles in the multiplex network, and  $\vec{P}_{RS} = [\vec{P}_{RS}^1, \dots, \vec{P}_{RS}^L]$  is the initial probability vector of a restart.  $\vec{P}_{RS} = \tau\vec{P}_0$ , and the vector parameter  $\tau = [\tau_1, \dots, \tau_L]$  indicates the restart probability of each layer of the particle in the multiplex network. We can change the weights between different layers by changing the parameter  $\tau$  to promote the important layers. In this study, all layers were considered equally important. Finally, the RWR-M outputs a score for each gene, reflecting

the global similarity of the gene to the known stress genes. In this study, we chose the top  $n$  genes as potential genes for further analysis.

#### 2.4. Leave-One-Out Cross-Validation Strategy

We used the leave-one-out cross-validation (LOOCV) strategy (Pseudocode is shown in Algorithm 2) [5,27] to evaluate the predictive performance of the RWR on the single and multiplex networks. We first created a test set, where the genes were the known drought stress genes from the China Rice Data Center. We removed the genes in the test set one at a time, named the left-out gene, and considered the rest of the genes as seed nodes in the RWR algorithms. According to their proximity to the seeds, the network nodes were scored and ranked. Using the same model, we recorded the rank of the left-out genes.

---

#### Algorithm 2 Leave-One-Out Cross-Validation Strategy.

---

**Input:** drought stress-related genes:  $test\_set$

**Output:** left-out gene rank sets

```

1: function LOOCV_S( $test\_set$ )
2:    $left\_out\_gene\_result \leftarrow None$ 
3:   for  $i = 1$  to  $length(test\_set)$  do
4:      $left\_out\_gene \leftarrow test\_set[i]$ 
5:      $test\_genes \leftarrow test\_set[1, 2, \dots, i - 1, i + 1, \dots, n]$ 
6:      $all\_genes\_rank \leftarrow RWR(test\_genes)$ 
7:      $left\_out\_gene$  rank
8:      $left\_out\_gene\_result.append(left\_out\_gene$  rank)
9:   end for
10:  return  $left\_out\_gene\_result$ 
11: end function

```

---

#### 2.5. Association Analysis between Potential Genes and Known Drought Stress Genes

Correlated genes are likely to have similar functions [27]. Therefore, a gene interacting with several known drought stress genes was more likely to be a drought stress-related gene. For each potential gene  $GP_j (j = 1, 2, 3, \dots, n)$ , we obtained its interaction scores with  $GK_i (i = 1, 2, 3, \dots, 218)$  from the STRING database. We defined it as  $S(GP_j, GK_i)$ , with  $GK_i$  representing the drought stress gene obtained from the China Rice Data Center. Any potential gene with more than one  $S(GP_j, GK_i) \geq 0.9$  was selected as a candidate gene.

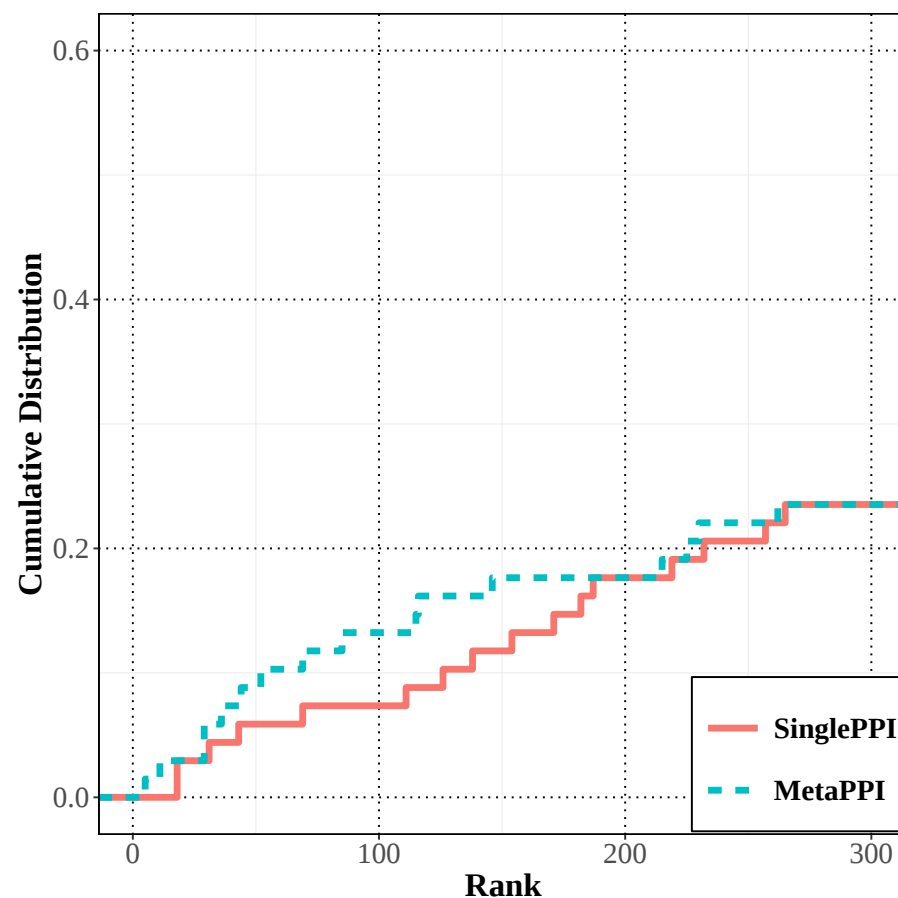
### 3. Results

#### 3.1. Prediction Performance Analysis of the RWR on the PPI Network

We compared the performance of the RWR on the MetaPPI with the SinglePPI by the LOOCV strategy. As shown in Figure 2, The ranking of the RWR on the MetaPPI was more accurate than on the SinglePPI. The result indicated that for the MetaPPI, the relation between proteins from the multisource data was more reliable than that from STRING. Therefore, we used the MetaPPI in the construction of the multiplex network.

#### 3.2. Prediction Performance Analysis of the RWR on the Multiplex Network

In this step, we compared the performance of the RWR on the multiplex network (RWR-M) with the RWR on the PPI network. In the multiplex network, two proteins can be linked by up to two edges, corresponding to the two layers, and the particle chooses between these different edges to move from a node to one of its neighbors. As shown in Figure 3, the RWR on the multiplex network outperformed the PPI networks. Moreover, the prediction performance of the RWR-M on the MetaPPI\_COEX\_DP and MetaPPI\_COEX\_MM was similar, and about 50% of the left-out genes were in the top 300 genes. Therefore, we utilized the multiplex networks to predict the drought stress-related genes.



**Figure 2.** The prediction performance of the RWR algorithm on the PPI network. The cumulative distribution functions represent the ranks of the left-out genes in the LOOCV with different network construction. (The X-axis is the rank of the nodes in the network, and the Y-axis is the ratio of the number of left-out genes ranked at the top to the number of genes in the test set).

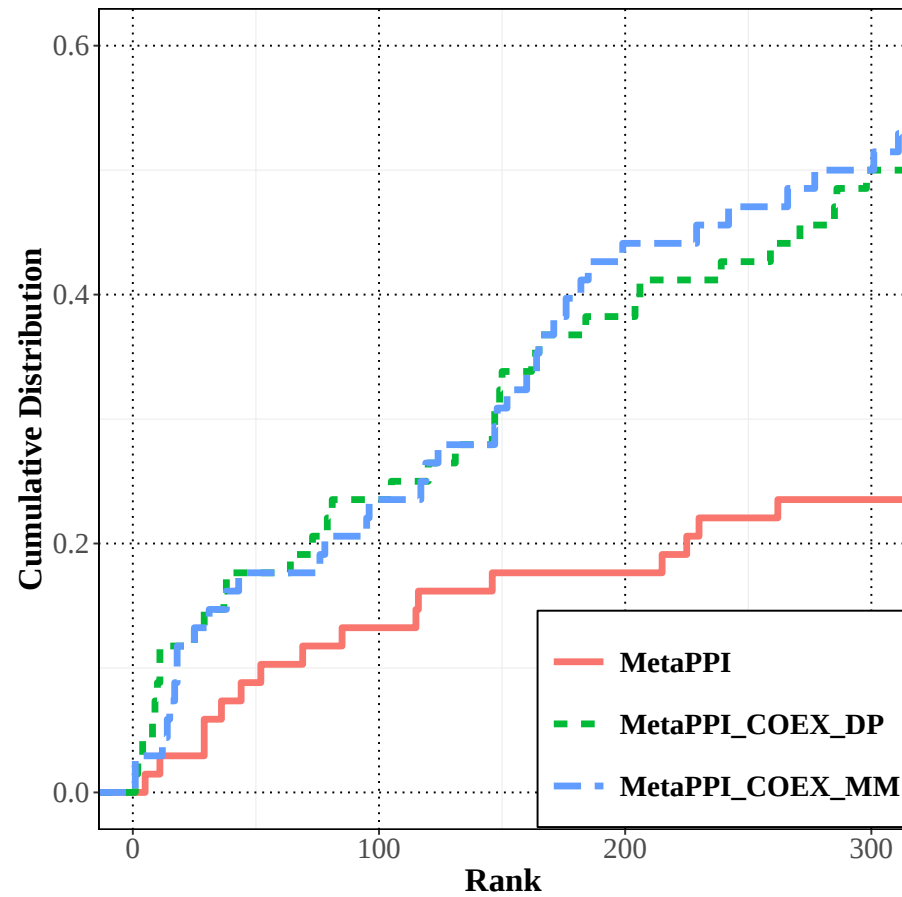
### 3.3. Obtaining Potential Genes Based on the RWR-M

The RWR-M displayed a remarkable performance in prioritizing genes; over 50% of the left-out genes ranked in the top 300. We screened out the top 300 genes by the RWR on MetaPPI\_COEX\_DP or MetaPPI\_COEX\_MM and retained the overlapped findings from the two methods. The consensus hits by both approaches were more likely to be drought stress-related genes, resulting in 174 potential genes for further analysis.

### 3.4. Enrichment Analysis

AmiGO is a web application that allows users to query and visualize ontologies and their related gene products' annotations [28]. In this study, we used AmiGO to analyze the 174 potential genes. These potential genes were enriched in multiple related GO pathways in the three categories of Biological Process (BP), Molecular Function (MF), and Cellular Component (CC) (as shown in Table 1). The GO pathway with a p value less than 0.01 was selected and recorded. The enrichment results showed that the pathways were significantly enriched in biological processes, including the response pathways to stimuli: the response to temperature stimuli, response to endogenous stimuli, and the response to abiotic stimuli; the pathways to metabolic regulation: the regulation of metabolic process; and a pathway that may affect drought tolerance: the response to water deprivation. In molecular functions, the potential genes were significantly enriched in protein threonine phosphatase activity, protein serine phosphatase activity, phosphoric ester hydrolase activity, nucleoside-triphosphatase activity, and heat shock protein binding. In terms of cellular components, the significantly enriched pathways included the cytoplasm and the

intracellular membrane-bounded organelle. Therefore, the potential genes were enriched in response to temperature stimulus, heat, water deprivation, and other pathways related to drought stress.



**Figure 3.** Prediction performance of the RWR on the multiplex network and the PPI network.

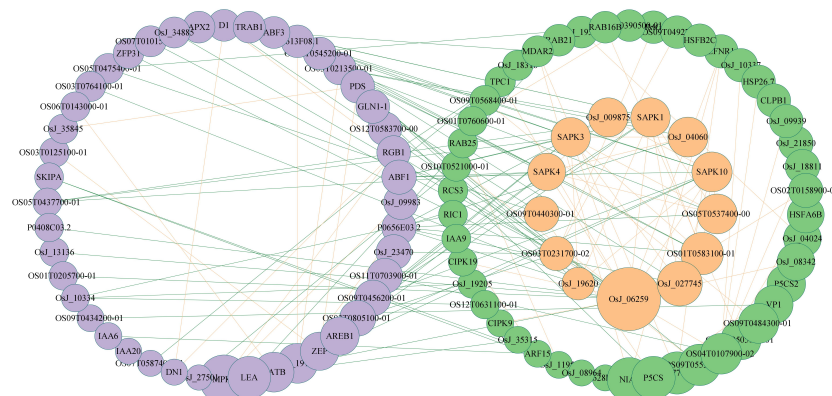
**Table 1.** Results of the enrichment analysis.

| GO Term    | Number of Genes | Ontology | Description                              | <i>p</i> -Value        |
|------------|-----------------|----------|------------------------------------------|------------------------|
| GO:0009266 | 18              | BP       | Response to temperature stimulus         | $2.25 \times 10^{-17}$ |
| GO:0009628 | 22              | BP       | Response to abiotic stimulus             | $1.24 \times 10^{-14}$ |
| GO:0009408 | 11              | BP       | Response to heat                         | $2.17 \times 10^{-10}$ |
| GO:0009631 | 6               | BP       | Cold acclimation                         | $2.37 \times 10^{-08}$ |
| GO:0009414 | 11              | BP       | Response to water deprivation            | $4.27 \times 10^{-12}$ |
| GO:0009719 | 22              | BP       | Response to endogenous stimulus          | $1.51 \times 10^{-13}$ |
| GO:0019222 | 44              | BP       | Regulation of metabolic process          | $3.76 \times 10^{-15}$ |
| GO:0031072 | 5               | MF       | Heat shock protein binding               | $3.27 \times 10^{-03}$ |
| GO:0106307 | 8               | MF       | Protein threonine phosphatase activity   | $2.08 \times 10^{-05}$ |
| GO:0042578 | 10              | MF       | Phosphoric ester hydrolase activity      | $2.28 \times 10^{-03}$ |
| GO:0106306 | 8               | MF       | Protein serine phosphatase activity      | $2.08 \times 10^{-05}$ |
| GO:0017111 | 11              | MF       | Nucleoside-triphosphatase activity       | $1.68 \times 10^{-04}$ |
| GO:0005737 | 60              | CC       | Cytoplasm                                | $6.59 \times 10^{-11}$ |
| GO:0043231 | 80              | CC       | Intracellular membrane-bounded organelle | $2.04 \times 10^{-19}$ |
| GO:0043229 | 83              | CC       | Intracellular organelle                  | $1.08 \times 10^{-19}$ |

### 3.5. Obtaining Candidate Genes Based on Association Analysis

The genes  $GP_j$  with two or more  $S(GP_j, GK_i) \geq 0.9$  were selected as our candidate genes. In our study, 13 related genes were identified. The annotation of the candidate genes

in STRING is shown in Table 2. Cytoscape [29] was used to map the connection between the potential genes and the known drought stress-related genes (as shown in Figure 4).



**Figure 4.** The interaction network between the potential genes and the known drought stress genes. The purple, orange, and light green nodes represent the known drought stress genes, the candidate genes, and other potential genes except the candidate genes, respectively. Each green line represents the interaction between a potential gene and a known drought stress gene.

**Table 2.** Annotation of candidate genes related to drought stress.

| Candidate Genes | Name in STRING          | Annotation                                                                                                                                                                                                        | The Chromosome Location |
|-----------------|-------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|
| Os10g0564500    | SAPK3                   | Serine/threonine-protein kinase <i>SAPK3</i> ; may play a role in the signal transduction of the hyperosmotic response                                                                                            | 10                      |
| Os02g0281000    | OsJ_06259               | Os02g0281000 protein                                                                                                                                                                                              | 2                       |
| Os03g0268600    | PP2C30; OsJ_009875      | Probable protein phosphatase 2C 30; belongs to the PP2C family                                                                                                                                                    | 3                       |
| Os05g0537400    | PP2C50                  | Probable protein phosphatase 2c 50; protein phosphatase involved in abscisic acid (ABA) signaling. Together with <i>PYL3</i> and <i>SAPK10</i> , may form an ABA signaling module involved in the stress response | 5                       |
| Os01g0846300    | PP2C09; OsJ_04060       | Probable protein phosphatase 2c 9; belongs to the PP2C family                                                                                                                                                     | 1                       |
| Os01g0869900    | SAPK4                   | Serine/threonine-protein kinase <i>sapk4</i> ; may play a role in the signal transduction of the hyperosmotic response                                                                                            | 1                       |
| Os03g0390200    | SAPK1                   | Serine/threonine-protein kinase <i>SAPK1</i> ; may play a role in the signal transduction of the hyperosmotic response                                                                                            | 3                       |
| Os03g0610900    | SAPK10                  | Serine/threonine-protein kinase <i>SAPK10</i> ; may play a role in the signal transduction of the hyperosmotic response                                                                                           | 3                       |
| Os03g0231700    | OS03T0231700-02         | Os03g0231700 protein; squalene monooxygenase, putative, expressed                                                                                                                                                 | 3                       |
| Os05g0572700    | OsJ_19620               | Probable protein phosphatase 2C 51; belongs to the PP2C family                                                                                                                                                    | 5                       |
| Os09g0325700    | OsJ_027745; PP108       | Probable protein phosphatase 2C 68; belongs to the PP2C family                                                                                                                                                    | 9                       |
| Os01g0583100    | OS01T0583100-01; PP2C06 | Probable protein phosphatase 2C 6; belongs to the PP2C family                                                                                                                                                     | 1                       |
| Os09g0440300    | OS09T0440300-01         | Aldehyde dehydrogenase-like protein; belongs to the aldehyde dehydrogenase family                                                                                                                                 | 9                       |

### 3.6. Receiver Operating Characteristic Curve Analysis

Support vector machines (SVM) is a powerful method to classify two or more classes of data. The candidate genes were subsequently utilized to construct the SVM classifier. The RNA-seq data was used to validate the classifier model. We created the receiver operating characteristic curve (ROC) [30] using pROC tools [31]. The area under the curve (AUC) value was 1. An area under the curve (AUC) value ranges from 0 to 1, with 1 indicating the candidate genes could distinguish the phenotypes of samples.

## 4. Discussion

When rice suffers from drought stress, its drought stress response genes' expression level changes rapidly and releases rice stress signal molecules to regulate rice plant resistance [32,33]. Often, the resistance is regulated by multiple genes. In this article, we introduced the RWR-M to predict rice drought stress genes, which outperformed the existing methods. To our knowledge, this was the first study in rice that applied the RWR with integrative genomics. To construct the rice biological network, we compared our methods against several methods applied to a platform. Extensive data analyses suggested that multisource data fusion reduced the noise of the single-source data and improved the prediction performance of the algorithm. In addition, we incorporated various similarity matrices in the network construction, such as Pearson's correlation coefficient and the MIC, to capture different aspects of associations. We expected the consensus findings across different similarity matrices to be more reliable. We identified 13 genes, of which five were reported in previous studies. This confirmed the utility of our approach.

Among the candidate genes, *SAPK10* [34], *OsPP108* [35], *PP2C06*, *OsJ\_009875(PP2C30)* [36], and *OsJ\_04060(PP2C09)* [36,37] have been reported to be related to drought stress; the other eight genes were likely to regulate drought resistance in rice. SnRK2s (*SAPK1-10*) are released by ABA-PYR/PYL/RCAR complexes, competitively interacting with the PP2C family [38]. *SAPK1* and *SAPK3* are activated by osmotic stress [39,40]. *SAPK4* is weakly activated by ABA and osmotic stress, and *SAPK10* is strongly activated by ABA and osmotic stress [41]. Absciscic acid (ABA) plays an important role in the drought resistance of plants. In addition, a new "*SAPK10-WRKY87-ABF1*" module revealed that *SAPK10* participated in rice's drought and salt tolerance [34]. Moreover, the type 2C protein phosphate (PP2C) regulates the ABA response [42,43]. Similarly, *OsPP108* [35] is involved in the regulation of ABA, and its overexpression enhances drought resistance and salt tolerance. *PP2C06*, *PP2C09*, and *PP2C30* can be used as positive regulatory factors of stress signals; *PP2C09* regulates the drought response regulators and activates the ABA-independent signaling pathways by activating the DRE promoters [36,37]. Liu et al. [16] predicted six genes (*HSF11*, *HSF5*, *HsFB2C*, *PP2C06*, *OS03T0231700-02*, and *OS03T0376100-01*) to be associated with drought stress using the WGCNA with multiplatform genomic data. *PP2C06* and *OS03T0231700-02* were also identified in our study.

In sum, the RWR on the multiplex network has the potential to identify key genes related to drought stress. Our findings were consistent with previous studies. This has value as a reference for the gene mining of a crop's abiotic stress response. In associate analysis, the number of candidate genes obtained will vary with the setting of the parameter thresholds. In this study, we selected  $S(GP_j, GK_i) \geq 0.9$ , and the threshold can be adjusted appropriately according to the actual situation.

## 5. Conclusions

In this article, we constructed a rice multiplex biological network; the RWR was applied to this multiplex network to predict genes that were highly related to drought stress, and we conducted a series of analyses to further screen the genes. This approach yielded 13 candidates, five of which were involved in drought stress resistance mechanisms, according to the supporting experimental evidence available, and the other eight candidate genes may be involved in the drought stress regulation of rice.

The construction of multiplex molecular networks of other abiotic stresses in rice is similar to drought stress, so this method can be extended to other abiotic stress-related gene mining. This study provides a new idea for fully utilizing multisource data to mine abiotic stress-related genes in rice and also provides a reference for further research on stress-resistant rice varieties.

**Author Contributions:** Conceptualization, H.Z. and L.Z. (Liu Zhu); methodology, H.Z., L.Z. (Liu Zhu) and Y.X.; software, L.Z. (Liu Zhu) and Y.X.; validation, H.Z., D.C., Z.N. and L.L.; formal analysis, L.Z. (Liu Zhu) and L.Z. (Lei Zhu); investigation, H.Z. and L.Z. (Liu Zhu); resources, H.Z. and D.C.; data curation, L.Z. (Liu Zhu) and Y.X.; writing—original draft preparation, H.Z., L.Z. (Liu Zhu) and Y.X.; writing—review and editing, H.Z., D.C., L.Z. (Lei Zhu), Z.N. and L.L.; visualization, H.Z. and L.Z. (Liu Zhu); supervision, H.Z.; project administration, H.Z.; funding acquisition, H.Z., D.C., L.Z. (Lei Zhu) and L.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Natural Science Foundation of Hunan Province, grant numbers 2021JJ30351, 2020JJ4039, and 2022JJ40190 and the Scientific Research Foundation of Education Office of Hunan Province, grant number 21C0133.

**Data Availability Statement:** The data and code used in the experiment can be downloaded at [https://github.com/Fisherzl/Rice\\_Genes\\_Prediction.git](https://github.com/Fisherzl/Rice_Genes_Prediction.git) (accessed on 16 November 2022).

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

|       |                                               |
|-------|-----------------------------------------------|
| GWAS  | Genome-Wide Association Studies               |
| PPI   | Protein-Protein Interaction                   |
| RWR   | Random Walk with Restart                      |
| RWR-M | Random Walk with Restart on multiplex network |
| WGCNA | Weighted gene co-expression network analysis  |
| MIC   | Maximal Information Coefficient               |
| ROC   | Receiver operating characteristics            |
| SVM   | Support vector machines                       |
| LOOCV | Leave-One-Out Cross-Validation                |
| DEGs  | Differentially expressed genes                |
| SRA   | Sequence Read Archive                         |
| CC    | Cellular Component                            |
| BP    | Biological Process                            |
| MF    | Molecular Function                            |

## References

1. Zhang, Q. Successful Experience and Development Trend of China Hybrid Rice Seed Export and Enterprise Development Abroad. *Chin. Rice*. **2021**, *27*, 104–106. (In Chinese)
2. Gupta, A.; Rico-Medina, A.; Caño-Delgado, A.I. The physiology of plant responses to drought. *Science* **2020**, *368*, 266–269. [CrossRef]
3. Bodily, P.M.; Fujimoto, M.S.; Page, J.T.; Clement, M.J.; Ebbert, M.T.; Ridge, P.G.; Alzheimer’s Disease Neuroimaging Initiative. A novel approach for multi-SNP GWAS and its application in Alzheimer’s disease. *BMC Bioinform.* **2016**, *17*, 268. [CrossRef] [PubMed]
4. Peter, L.; Steve, H. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinform.* **2008**, *9*, 559.
5. Yo-Han, Y.; Minjae, K.; Nalini, C.A.K.; Hong, W.J.; Ahn, H.R.; Lee, G.T.; Kang, S.; Suh, D.; Kim, J.O.; Kim, Y.J.; et al. Genome-Wide Transcriptome Analysis of Rice Seedlings after Seed Dressing with *Paenibacillus yonginensis* DCY84T and Silicon. *Int. J. Mol. Sci.* **2019**, *20*, 5883.
6. Xiaobo, Z.; Chunjuan, L.; Shubo, W.; Tingting, Z.; Caixia, Y.; Shihua, S. Transcriptomic analysis and discovery of genes in the response of *Arachis hypogaea* to drought stress. *Mol. Biol. Rep.* **2021**, *45*, 119–131.
7. Junxia, Y.; Tsutomu, T.; Toshihiro, O.; Hiroyuki, A.; Ikuko, T.; Eishin, O.; Hiroko, O.; Hatasu, K.; Toshiaki, H.; Wanyang, L.; et al. Combined linkage analysis and exome sequencing identifies novel genes for familial goiter. *J. Hum. Genet.* **2013**, *58*, 104–106.
8. Alberto, V.; Laurent, T.; Claire, N.; Perrin, S.; Odeline, G.; Levy, N.; Cau, P.; Remy, E.; Baudot, A. Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics* **2018**, *35*, 497–505.

9. Li, Y.; Patra, J.C. Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network. *Bioinformatics* **2010**, *26*, 1219–1224. [[CrossRef](#)]
10. Qu, J.; Wang, C.; Cai, S.; Zhao, W.D.; Cheng, X.L.; Ming, Z. Biased Random Walk With Restart on Multilayer Heterogeneous Networks for MiRNA–Disease Association Prediction. *Front. Genet.* **2021**, *12*, 1427. [[CrossRef](#)]
11. Chauhan, V.K.; Dahiya, K. Problem formulations and solvers in linear SVM: A review. *Artificial Intell. Rev.* **2019**, *52*, 803–855. [[CrossRef](#)]
12. Szklarczyk, D.; Gable, A.L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N.T.; Morris, J.H.; Bork, P.; et al. STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **2018**, *47*, D607–D613. [[CrossRef](#)]
13. Liu, S.; Liu, Y.; Zhao, J.; Cai, S.; Qian, H.M.; Zuo, K.; Zhao, L.; Zhang, L. A computational interactome for prioritizing genes associated with complex agronomic traits in rice (*Oryza sativa*). *Plant J.* **2017**, *90*, 177–188. [[CrossRef](#)]
14. Gu, H.; Zhu, P.; Jiao, Y.; Meng, Y.; Chen, M. PRIN: A predicted rice interactome network. *BMC Bioinform.* **2011**, *12*, 161. [[CrossRef](#)]
15. Hiroaki, S.; Shin, L.S.; Tsuyoshi, T.; Hisataka, N.; Jungsok, K.; Yoshihiro, K.; Hironobu, W.; Ching-chia, Y.; Masao, I.; Takashi, A.; et al. Rice Annotation Project Database (RAP-DB): An Integrative and Interactive Database for Rice Genomics. *Plant Cell Physiol.* **2012**, *54*, e6.
16. Liu, Y.; Zhang, H.; Cao, D.; Li, L. Prediction of drought and salt stress-related genes in rice based on multi-platform gene expression data. *Plant J.* **2021**, *47*, 2423–2439. (In Chinese)
17. Liu, S.; Wang, Z.; Zhu, R.; Wang, F.; Cheng, Y.; Liu, Y. Three Differential Expression Analysis Methods for RNA Sequencing: Limma, EdgeR, DESeq2. *J. Vis. Exp. JoVE.* **2021**, *175*, 177–188. [[CrossRef](#)]
18. Xu, H.; Deng, Y. Dependent Evidence Combination Based on Shearman Coefficient and Pearson Coefficient. *IEEE Access* **2018**, *6*, 11634–11640. [[CrossRef](#)]
19. Guo, Z.; Yu, B.; Hao, M.; Wang, W.; Zong, F. A novel hybrid method for flight departure delay prediction using Random Forest Regression and Maximal Information Coefficient. *Aerosp. Sci. Technol.* **2021**, *116*, 106822. [[CrossRef](#)]
20. Cao, D.; Xu, N.; Chen, Y.; Zhang, H.Y.; Li, Y.; Yuan, Z.M. Construction of a Pearson- and MIC-Based Co-expression Network to Identify Potential Cancer Genes. *Interdiscip. Sci. Comput. Life Sci.* **2022**, *14*, 245–257. [[CrossRef](#)]
21. Sun, G.; Li, J.; Dai, J.; Song, Z.; Lang, F. Feature selection for IoT based on maximal information coefficient. *Future Gener. Comput. Syst.* **2018**, *89*, 606–616. [[CrossRef](#)]
22. Battiston, F.; Nicosia, V.; Latora, V. Structural measures for multiplex networks. *Phys. Rev. E* **2014**, *89*, 14. [[CrossRef](#)] [[PubMed](#)]
23. Xia, F.; Liu, J.; Nie, H.; Fu, Y.; Wan, L.; Kong, X. Random Walks: A Review of Algorithms and Applications. *IEEE Trans. Emerg. Top. Comput. Intell.* **2020**, *4*, 95–107. [[CrossRef](#)]
24. Lei, X.; Bian, C. Integrating random walk with restart and k-Nearest Neighbor to identify novel circRNA-disease associatio. *Sci. Rep.* **2020**, *10*, 1943. [[CrossRef](#)] [[PubMed](#)]
25. Li, J.; Chen, L.; Wang, S.; Zhang, Y.H.; Kong, X.Y.; Huang, T.; Cai, Y.D. A computational method using the random walk with restart algorithm for identifying novel epigenetic factors. *Mol. Genet. Genom.* **2018**, *293*, 293–301. [[CrossRef](#)]
26. Liu, L.-L.; Zhang, S.-W. Advances in Predicting The Risk Pathogenic Genes With Random Walk. *Prog. Biochem. Biophys.* **2021**, *48*, 1184–1195.
27. Belotti, F.; Peracchi, F. Fast leave-one-out methods for inference, model selection, and diagnostic checking. *Stata J.* **2020**, *20*, 785–804. [[CrossRef](#)]
28. Carbon, S.; Ireland, A.; Mungall, C.J.; Shu, S.Q.; Marshall, B.; Lewis, S. AmiGO: Online access to ontology and annotation data. *Bioinformatics* **2008**, *25*, 288–289. [[CrossRef](#)]
29. Otasek, D.; Morris, J.H.; Bouças, J.; Pico, A.R.; Demchak, B. Cytoscape Automation: Empowering workflow-based network analysis. *Genome Biol.* **2019**, *20*, 185. [[CrossRef](#)]
30. Martínez-Camblor, P.; Pérez-Fernández, S.; Díaz-Coto, S. The area under the generalized receiver-operating characteristic curve. *Int. J. Biostat.* **2021**, *18*, 293–306. [[CrossRef](#)]
31. Robin, X.; Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J.C.; Muller, M. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform.* **2011**, *12*, 77. [[CrossRef](#)]
32. Kamiyama, Y.; Hirotani, M.; Ishikawa, S.; Minegishi, F.; Katagiri, S.; Rogan, C.; Takahashi, F.; Nomoto, M.; Ishikawa, K.; Kodama, Y.; et al. Arabidopsis group C Raf-like protein kinases negatively regulate abscisic acid signaling and are direct substrates of SnRK2. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2100073118. [[CrossRef](#)]
33. Yoshiaki, K.; Sotaro, K.; Taishi, U. Growth Promotion or Osmotic Stress Response: How SNF1-Related Protein Kinase 2 (SnRK2) Kinases Are Activated and Manage Intracellular Signaling in Plants. *Plants* **2021**, *10*, 1443.
34. Liu, Y.; Wang, B.; Li, J.; Sun, Z.; Chi, M.; Xing, Y.; Xu, B.; Yang, B.; Li, J.; Liu, J.; et al. A novel SAPK10-WRKY87-ABF1 biological pathway synergistically enhance abiotic stress tolerance in transgenic rice (*Oryza sativa*). *Plant Physiol. Biochem.* **2021**, *168*, 252–262.
35. Singh, A.; Jha, S.K.; Bagri, J.; Pandey, G.K. ABA Inducible Rice Protein Phosphatase 2C Confers ABA Insensitivity and Abiotic Stress Tolerance in Arabidopsis. *PLoS ONE* **2015**, *10*, e0125168. [[CrossRef](#)]
36. Min, M.K.; Kim, R.; Hong, W.J.; Jung, K.H.; Lee, J.Y.; Kim, B.G. OsPP2C09 Is a Bifunctional Regulator in Both ABA-Dependent and Independent Abiotic Stress Signaling Pathways. *Int. J. Mol. Sci.* **2021**, *22*, 393. [[CrossRef](#)]

37. Miao, J.; Li, X.; Li, X.; Tan, W.; You, A.; Wu, S.; Tao, Y.; Chen, C.; Wang, J.; Zhang, D.; et al. OsPP2C09, a negative regulatory factor in abscisic acid signalling, plays an essential role in balancing plant growth and drought tolerance in rice. *New Phytol.* **2020**, *227*, 1417–1433. [[CrossRef](#)]
38. Soma, F.; Takahashi, F.; Yamaguchi-Shinozaki, K.; Shinozaki, K. Cellular Phosphorylation Signaling and Gene Expression in Drought Stress Responses: ABA-Dependent and ABA-Independent Regulatory Systems. *Plants* **2021**, *10*, 756. [[CrossRef](#)]
39. Vavrdová, T.; Samaj, J.; Komis, G. Phosphorylation of Plant Microtubule-Associated Proteins During Cell Division. *Front. Plant Sci.* **2019**, *10*, 238. [[CrossRef](#)]
40. Anna, K.; Izabela, W.; Ewa, K.; Maria, B.; Grazyna, D. SnRK2 Protein Kinases—Key Regulators of Plant Response to Abiotic Stresses. *OMICS: J. Integr. Biol.* **2011**, *15*, 859–872.
41. Holappa, L.D.; Ronald, P.C.; Kramer, E.M. Evolutionary Analysis of Snf1-Related Protein Kinase2 (SnRK2) and Calcium Sensor (SCS) Gene Lineages, and Dimerization of Rice Homologs, Suggest Deep Biochemical Conservation across Angiosperms. *Front. Plant Sci.* **2017**, *8*, 395. [[CrossRef](#)] [[PubMed](#)]
42. Chen, Y.; Zhang, J.-B.; Wei, N.; Liu, Z.-H.; Li, Y.; Zheng, Y.; Li, X.-B. A type-2C protein phosphatase (GhDRP1) participates in cotton (*Gossypium hirsutum*) response to drought stress. *Plant Mol. Biol.* **2021**, *107*, 499–517. [[CrossRef](#)] [[PubMed](#)]
43. Han, S.; Lee, J.Y.; Lee, Y.; Kim, T.; Lee, S. Comprehensive survey of the VxG phi L motif of PP2Cs from *Oryza sativa* reveals the critical role of the fourth position in regulation of ABA responsiveness. *Plant Mol. Biol.* **2019**, *101*, 455–469. [[CrossRef](#)] [[PubMed](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.