frontiers in Plant Science

ORIGINAL RESEARCH
published: 20 September 2017
doi: 10.3389/fpls.2017.01640

RECoN: Rice Environment Coexpression Network for Systems Level Analysis of Abiotic-Stress Response

*Arjun Krishnan[1†], Chirag Gupta[2], Madana M. R. Ambavaram[1†] and Andy Pereira[1,2]**

[1] Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA, United States, [2] Crop, Soil, and Environmental Sciences, University of Arkansas, Fayetteville, AR, United States

OPEN ACCESS

Edited by:
Sneh Lata Singla-Pareek, International Centre for Genetic Engineering and Biotechnology, India

Reviewed by:
Anil Kumar Singh, Indian Institute of Agricultural Biotechnology (ICAR), India
Saurabh Raghuvanshi, University of Delhi, India

***Correspondence:**
Andy Pereira
apereira@uark.edu

†Present address:
Arjun Krishnan, Computational Mathematics, Science, and Engineering, Michigan State University, East Lansing, MI, United States; Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, United States
Madana M. R. Ambavaram, Yield10 Bioscience, Inc., Woburn, MA, United States

Specialty section:
This article was submitted to Plant Abiotic Stress, a section of the journal Frontiers in Plant Science

Received: 21 May 2017
Accepted: 06 September 2017
Published: 20 September 2017

Citation:
Krishnan A, Gupta C, Ambavaram MMR and Pereira A (2017) RECoN: Rice Environment Coexpression Network for Systems Level Analysis of Abiotic-Stress Response. Front. Plant Sci. 8:1640. doi: 10.3389/fpls.2017.01640

Transcriptional profiling is a prevalent and powerful approach for capturing the response of crop plants to environmental stresses, e.g., response of rice to drought. However, functionally interpreting the resulting genome-wide gene expression changes is severely hampered by the large gaps in our genomic knowledge about which genes work together in cellular pathways/processes in rice. Here, we present a new web resource – RECoN – that relies on a network-based approach to go beyond currently limited annotations in delineating functional and regulatory perturbations in new rice transcriptome datasets generated by a researcher. To build RECoN, we first enumerated 1,744 abiotic stress-specific gene modules covering 28,421 rice genes (>72% of the genes in the genome). Each module contains a group of genes tightly coexpressed across a large number of environmental conditions and, thus, is likely to be functionally coherent. When a user provides a new differential expression profile, RECoN identifies modules substantially perturbed in their experiment and further suggests deregulated functional and regulatory mechanisms based on the enrichment of current annotations within the predefined modules. We demonstrate the utility of this resource by analyzing new drought transcriptomes of rice in three developmental stages, which revealed large-scale insights into the cellular processes and regulatory mechanisms involved in common and stage-specific drought responses. RECoN enables biologists to functionally explore new data from all abiotic stresses on a genome-scale and to uncover gene candidates, including those that are currently functionally uncharacterized, for engineering stress tolerance.

Keywords: rice, coexpression network, drought, abiotic stress, webserver, database

INTRODUCTION

The complex response of plants to abiotic stress spans several orders of magnitude in time and space, causing system-wide adverse reactions and protective responses. Gene expression profiling has been used successfully to capture the system-wide molecular programs that underlie the cellular response to abiotic stresses (Deyholos, 2010). Analyses of the drought-stress inducible transcriptome in Arabidopsis, for example, reveal a plethora of responses including the induction of transcription factors, phospholipases C and D, protein kinases (MAPK, CDPK), proteinases,

water channel proteins, antioxidant enzymes and molecules (GSTs, thioredoxins, peroxiredoxins), factors such as chaperones that afford protection for macromolecules (LEA proteins, HSPs) and osmoprotectant synthases (for proline, betaine, sugar) (Seki et al., 2002; Shinozaki et al., 2003; Harb et al., 2010). Making such analytical inferences from the transcriptome hinges on the availability of prior functional and regulatory knowledge about a large number of genes in the genome, which can then be used to meaningfully summarize genome-wide gene-expression changes. Although far from complete in Arabidopsis, such functional/regulatory information about genes – what they do, how they work together, and how they are regulated – is severely lacking in rice both in quality and genomic coverage. This paucity has led to a significant lag in the number of characterized drought-responsive genes and cellular processes in rice compared to Arabidopsis. Therefore, we need alternative approaches that can go beyond currently available gene annotations for fully extracting knowledge from rice transcriptomes and help toward gaining a comprehensive understanding of stress response in this crop plant.

A promising starting point for such an alternative approach is the large amount of currently publicly available gene expression data in plants. A powerful analysis framework that has emerged in recent years involves estimating the similarity of expression patterns between all pairs of genes across diverse conditions to build gene coexpression networks, representing the genome-wide transcriptional organization of the cell (Usadel et al., 2009; Pearce et al., 2015). Particularly, in the crop model rice, recent studies have used gene coexpression to gain biological insights into general (Wang et al., 2009; Shaik and Ramakrishna, 2013) and case-specific (Fu and Xue, 2010; Ambavaram et al., 2014) gene regulation. Coexpression networks have also been used extensively in plants to organize genes into transcriptional modules and explore their functions (Mentzen and Wurtele, 2008; Mao et al., 2009). These modules represent groups of genes/proteins are likely to work together to perform a coherent biological function inside the cell (Hartwell et al., 1999), essentially expanding upon the available functional annotations. Hence, coexpressed modules can be used as sets of functionally coherent genes to see their enrichment in new expression data, especially in genomes where the functional annotations are sparse and incomplete. However, we still lack such a resource that allows enrichment analysis of coexpressed clusters/modules in new expression data of rice.

Therefore, it would be valuable to reconstruct a rice coexpression network that integrates information across a large number of datasets specifically in the context of abiotic stress. In addition, it would be highly beneficial if researchers could bring this coexpression network to bear on their new gene-expression profiles (in one or more conditions related to abiotic stress) for functional resolution and comparison. Interpreting the long lists of responsive genes, a typical result of a gene-expression study, will become amply tractable by identifying subsets of responsive genes that are likely to be functionally coherent. Likewise, comparison between gene responses in different growth stages or conditions is likely to be more meaningful and robust at the level of cellular functions/pathways than at the level of individual genes (capturing the perturbation of different subsets of the same cellular apparatus and overcoming the various sources of noise in high-throughput assays).

To meet all these critical needs, here we present a genomic resource for comprehensive analysis of abiotic stress response in rice based on a modular coexpression network specific to response to environmental conditions, and apply this resource to perform a detailed analysis of stage-specific drought response in rice. First, we carried out a genome-scale analysis integrating publicly available rice gene expression datasets generated in the context of response to a range of environmental conditions. Next, using this integrated data, we constructed, what is termed, the Rice Environment Coexpression Network (RECoN), based on gene expression correlation across environmental conditions. Finally, we partitioned RECoN into densely connected modules using a graph-clustering algorithm. As a pertinent test case for our approach, we performed gene expression profiling of rice plants subjected to drought at three developmental stages. We used this data to perform both a traditional analysis – functional analysis [using Gene Ontology (GO)] – and a new analysis using RECoN, teasing out drought-related modules within the drought-response genes identified from our experiments. The new RECoN-based analysis of new experimental data helped highlight pathways, processes, regulatory genes, and potential transcriptional regulatory mechanisms critical for drought response in rice. We have made RECoN available for rice stress biologists through an interactive network browser at https://plantstress-pereira.uark.edu/RECoN/. Biologists can use this resource to explore coexpression clusters within their stress transcriptome and systematically guide follow-up experimental studies for constructing the underlying gene network.

MATERIALS AND METHODS

Coexpression Network Analysis

A total of 29 publicly available gene expression datasets comprising of 414 samples of the Affymetrix rice GeneChip from were collected from NCBI GEO (Barrett et al., 2009) and ArrayExpress (Parkinson et al., 2009). From these, 129 samples (45 groups) with a unifying biological theme, i.e., response to some environmental condition, were used for coexpression analysis (see Supplementary Table S1).

We previously reported the re-annotation of Rice GeneChip to increase the reliability of expression quantification (Ambavaram et al., 2011). Briefly, the chip definition file (CDF GPL11322) was created by mapping probes to target genes that have perfect sequence similarity, and regrouping probesets such that each represents a single corresponding gene. We used this custom CDF to background correct, normalize and summarize the raw data using justRMA (Irizarry et al., 2003), with values in replicate samples averaged. To estimate coexpression, the Pearson correlations between every pair of genes were first calculated (Huttenhower et al., 2008) and then normalized using Fisher's Z-transform (David, 1949). Then, the standardization of these scores resulted in coexpression score (z_{cs}) indicating the number of standard deviations it lies from the mean, and

follows a normal distribution to be interpretable by the level of significance, as $|z_{cs}|$ values greater than 1.96 allowed a 95% confidence interval to work with.

A coexpression network was then constructed connecting pairs of genes that have a $z_{cs} > 1.96$ (top 2.5% of all pairs of genes ordered in decreasing order of correlation). This cutoff corresponded to a Pearson correlation coefficient of 0.632. This network that contains 34,792 genes connected by ∼18.5 million edges was then clustered using SPICi (Jiang and Singh, 2010). Since SPICi requires a density parameter T_d as input, a range of values of the parameter from 0.1 to 0.9 was tested. Clusters obtained using each T_d value were evaluated using several criteria including the number of clusters formed, fraction of genes in clusters of size three or more, average segregation (modularity), and extent of overlap between clusters and GO BP gene sets (termed 'functions'). In order to calculate average segregation, as desired property of dense interaction networks, the coexpression network is modeled as an undirected graph $G = (V, E)$, consisting of a set V of nodes (i.e., genes) and a set E of edges (i.e., coexpressing gene pairs). Let w_{uv} denote the weight of the edge $(u, v) \in E$, denoting the Pearson correlation coefficient of gene pairs (u, v). The graph $G_c = (V_c, E_c)$ is defined as the graph induced by the genes that are part of cluster c, and average segregation is computed as:

$$\frac{\sum_{(u,v)\in E_c} w_{uv} \Big/ |E_c|}{\sum_{u\in V_c; y\in V; (u,y)\in E} w_{uy} \Big/ |E_{c'}|}$$

where $E_{c'}$ is the set of edges in G that are incident on V_c. For functional enrichment analysis the overlap between genes within a cluster and genes annotated to a given GO BP term using the cumulative hypergeometric test. Using only GO BP terms that annotate < 500 genes (to ensure a certain level of specificity in definition), for a pair of gene sets (cluster and GO BP term) i and j, if N is the total number of genes, n_i and n_j are the number of genes in gene set i and j, and m is the number of genes common to the gene sets, the probability (p-value) of an overlap (enrichment) of size equal to or greater than observed is given by the formula below.

$$P(X = x \geq m) = \sum_{x=m}^{\min(n_i, n_j)} \frac{\binom{n_i}{x}\binom{N-n_i}{n_j-x}}{\binom{N}{n_j}}$$

p-Values from the test were converted to q-values to correct for multiple hypothesis testing using Benjamini–Hochberg method (Benjamini and Hochberg, 1995) and cluster-GO_BP pairs with q-value < 0.1 were considered for analysis. The level of functional enrichment in a cluster is quantified using $-log_{10}(q\text{-value})$.

After clustering the network using SPICi with T_d value of 0.65, from all the clusters, those relevant to drought were determined by testing which clusters contained a significantly high number of drought-regulated genes up- or down-regulated in any one of the stages (again using a cumulative hypergeometric test). Then, four sets of genes were extracted from each 'drought' cluster – all the genes in the cluster, and seedling, vegetative and reproductive drought-regulated genes – for discovery of putative *cis*-regulatory elements (CREs) and enrichment analysis of GO biological processes (BPs) in the rice genome.

Rice Plant Material and Drought-Stress Treatments

Rice (*Oryza sativa* L. ssp. *Japonica* cv. Nipponbare) seeds were germinated in hydroponic half-strength Hoagland solution and seedlings were grown about a week in an environmentally controlled growth chambers maintained at 28 ± 1°C temperature, 65% relative humidity with a daily photoperiodic cycle of 14 h light and 10 h dark, and then plants allowed to reach the reproductive stages were grown in soil under greenhouse conditions. Samples of well-watered and drought stressed were collected at various developmental stages which include 7-day old seedlings, vegetative (V4) and reproductive (R4) stages based on discrete morphological criteria as described by Counce et al. (2000).

For drought treatment, plants were gradually subjected to field drought stress in order to reach 50% field capacity (FC) by regulating the water supply, whereas control plants were maintained at 100% FC. During the stress period the pots were weighed daily and the difference in weight on subsequent days was corrected by adding water to maintain the required FC. The physiological condition of plants at 50% FC was monitored by chlorophyll fluorescence, quantum yield (Fv/Fm) and the relative water content (RWC) (Supplementary Table S1). For dry down drought treatment, rice plants after transplanting were separated, with five pots maintained at well-watered condition serving as control while another set of five pots were used for drought experiments. For drought stress treatment water was withheld until the moisture level progressively dropped down to 6%. Drought stress symptoms were monitored for leaf rolling and measurement of soil moisture content everyday using soil a moisture meter (Rapitest). For all the stages (seedlings, vegetative, and reproductive), three biological replicates were harvested from independent populations of plants, when leaves were completely rolled and RWC was around 65–70%. RWC was measured in the leaves used from where photosynthesis was measured. Leaf fragments of same size were cut and fresh weight was measured and hydrated immediately to full turgidity in deionized water for 6 h. After 6 h the leaf fragments were blotted on paper towels and the fully turgid weight was taken. Turgid leaf samples were then oven dried at 80°C for 72 h and weighed to determine dry weight. RWC percentage was measured as: RWC (%) = (fresh weight - dry weight)/(turgid weight - dry weight) × 100. The drought stress symptoms such as leaf rolling and basal leaf senescence were apparent in stress-induced plants, while control plants growing at 100% FC were observed to grow well-showing 95% RWC.

Measurement of Chlorophyll Fluorescence and Quantum Yield

Chlorophyll fluorescence and the quantum yield was measured by using the Modulated Chlorophyll Fluorometer OS1-FL (Opti-Sciences, Inc., United States). During and after stress treatments, flag leaf from stressed and unstressed wild-type was placed

in close contact with the Photosynthetically Active Radiation (PAR) clip, which provides basic data to the OS1-FL system on ambient conditions. The PAR sensor is designed to measure leaf temperature and the light intensity. The ratio of variable fluorescence (Fv/Fm) and the yield of quantum efficiency (Y) are indicative of photosystems I and II performance of the plants under stress.

RNA Isolation, Probe Labeling, and Hybridization

Total RNA was isolated from the rice seedlings, vegetative (V4) and reproductive (R4) tissues of both control and stress treated plants using the RNeasy plant kit (Qiagen, United States) according to manufactures protocol. For each stage/treatment, three independent biological replicates were used for RNA isolation.

RNA quantity, quality, and purity were assessed with the use of the RNA 6000 Nano assay on the Agilent 2100 Bioanalyzer (Agilent Technologies, United States). Total RNA (∼4 μg) from each sample was used to generate first-strand cDNA with a T7-Oligo(dT) primer. Following second-strand synthesis, *in vitro* transcription was performed using the GeneChip® IVT Labeling Kit according to the manufacturer's instructions. The preparation and processing of labeled and fragmented cRNA targets, as well as hybridization to arrays, washing, staining, and scanning were carried out according to manufacturer's instructions[1]. The Affymetrix Rice GeneChips (which contain ∼43,000 probe sets or genes), washing and scanning were carried out in Gene chip fluidics Station 450 (Affymetrix) and the Gene chip Scanner 3000 by Affymetrix (Santa Clara, CA, United States), respectively.

Analysis of Differential Gene Expression

The custom CDF file was used to background correct, normalize and summarize all the raw expression data using RMA in R (Ihaka and Gentleman, 1996; Irizarry et al., 2003; Gentleman et al., 2004). Genes that had the interquartile range (IQR) less than the median were detected as lowly varying, and were removed from further analysis of differential expression. To estimate differential expression among the remaining genes, a linear model was used (Smyth, 2004). The resulting p-values of the t-tests were corrected for multiple hypothesis testing and reported as q-values (Storey and Tibshirani, 2003). A threshold of q-value < 0.01 was set to select significantly differentially expressed genes in response to drought.

Functional and Regulatory Annotations of Clusters

Functional annotations of rice genes in GO BP and KEGG pathways categories were downloaded from the PlantGSEA website (Yi et al., 2013). Gene sets that annotated more than 1500 genes and less than 10 genes were removed to gain resolution in the BPs presented by enrichment analysis. The statistical significance of overlap between a given gene set and a cluster was tested using a cumulative hypergeometric test and expressed in terms of false discovery rates (q-values). The enrichment score (ES) was reported as $-\log_{10}$ q-value. The Clusters were then annotated with functional categories (GO BP and KEGG) that had an overlap ES greater than 1.3 (q-value < 0.05). CREs enriched in the clusters were identified *de novo* using FIRE (Elemento et al., 2007). FIRE uses mutual information (MI) to find an association between the given expression profile of a gene (cluster participation) and their motif profile, and uses randomization tests to score for statistical significance. Motifs reported by FIRE were matched to known plant motifs (Higo et al., 1999) using STAMP (Mahony and Benos, 2007). Motifs that were present in more than 50% of all the genes in a given cluster were considered for cluster annotations.

Perl scripts were used to parse all the data. Plots were generated using R (Ihaka and Gentleman, 1996) and gene expression matrices were visualized using MeV (Saeed et al., 2006).

The gene expression data reported here are available from the NCBI GEO database with the accession number GSE81253.

RESULTS

The Rice Environment Coexpression Network

To determine biologically meaningful stress transcriptional modules in rice, we have designed an extensive pipeline that uses data from publicly available gene expression profiles in parallel with our in-house generated datasets measuring drought response in three developmental stages (**Figure 1**). We obtained 129 publicly available rice Affymetrix microarrays related to response of the rice plant to some environmental condition and worked with the raw data (**Figure 1**, step 1a). The data was normalized and summarized into a gene expression matrix based on a custom probe-gene reannotation of the rice GeneChip. The reannotation increases the accuracy of the gene expression quantification process by assigning only specific probes to genes, and increases coverage of the array. The gene expression data was then converted into a matrix of 34,792 genes and 45 distinct conditions/groups and used to construct a coexpression network connecting pairs of genes that have a significantly high correlation between their expression profiles across the conditions (top 2.5% of all pairs of genes ordered in decreasing order of correlation; see section "Materials and Methods"). This network, termed Rice Environment Coexpression Network or RECoN, contains 34,792 genes connected by ∼18.5 million edges.

There are several clustering algorithms that work with weighted networks and find groups of densely connected nodes (Enright et al., 2002; Bader and Hogue, 2003). SPICi, a clustering tool was selected due to its ability to cluster large networks extremely fast (Jiang and Singh, 2010), and used to cluster our extremely large network. However, like every clustering algorithm, amongst a few, there is a single user defined parameter T_d that determines the density of the resultant clusters and heavily influences the clustering process. To avoid an *ad hoc* or even a wrong choice of this parameter, we performed exhaustive data-driven tests on the network clustered using a range of T_d values to identify the best parameter for the network at hand (**Figure 2**).

[1]http://www.affymetrix.com

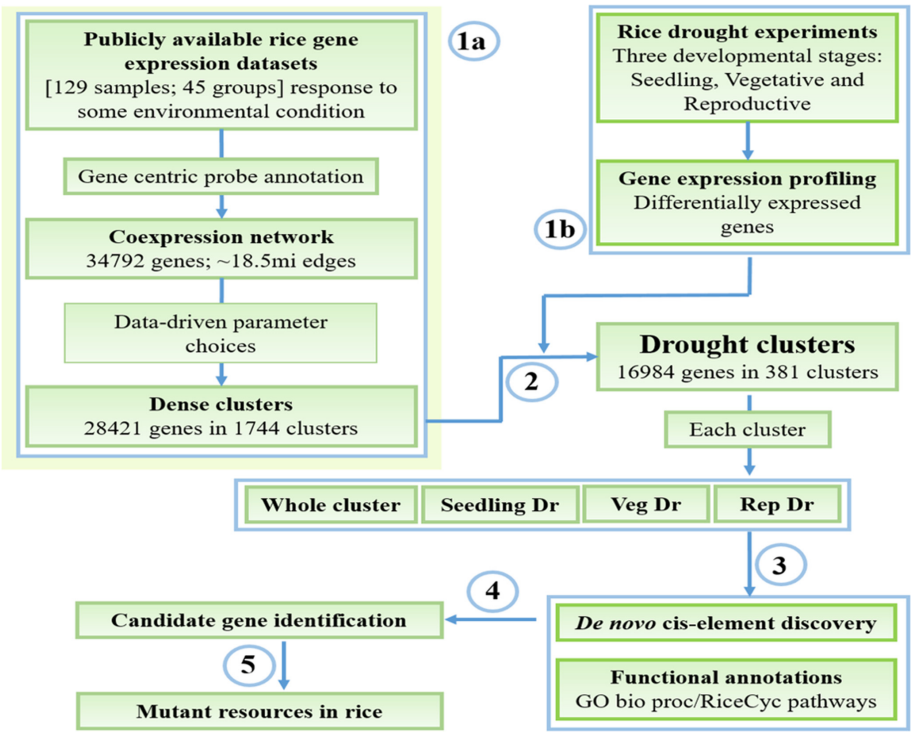

FIGURE 1 | Workflow for mining and characterization of drought transcriptional modules. (1a) Reconstruction and clustering of the rice environmental coexpression network from publicly available gene expression datasets. (1b) Identification of drought-responsive genes in the three developmental stages. (2) Determination of 'drought' clusters based on the combination of results from the steps 1a and 1b, and extraction of whole cluster and specific drought gene sets (3) Functional enrichment analysis and *cis*-regulatory motif discovery. (4) Presentation of these data to the user where (s)he explores the results to identify candidate genes for functional validation. (5) Availability of mutants in genes of interest that can be used to study gene function.

First, for different values of T_d, we tracked the number of clusters obtained and the fraction of genes in the original network that were in clusters of three or more genes (**Figure 2A**). At small values of T_d, there are very few clusters and only a few broken links. As T_d increases, the number of clusters increases, but, however, very high T_d will break the network so much that the clusters with three or more genes will again become rare. Similarly, as T_d increases, the number of genes that are part of clusters will steadily decrease until a critical value beyond which a large portion of genes will get disconnected and fall out of good-sized clusters. By testing for the value of T_d after which there is the first significant drop in the number of clusters and fraction of genes in clusters, we found that this is at $T_d = 0.65$. Second, we calculated a measure of modularity called average segregation that quantified how well genes within a cluster are connected to each other compared to their connection to all the genes in the network (**Figure 2B**) (Yook et al., 2004). Since we are interested in finding coherent biological modules, finding a T_d that preserves segregation is sought after. It was surprising that the network showed the highest values of segregation for the smallest values of T_d, indicating that even the original network with ~18.5 million edges is highly modular. Therefore, in the context of this network, at least, it was only important to look out for partitioning the network as much as possible without a significant drop in the inherent modularity. The first significant drop in average segregation (measured more qualitatively than quantitatively using the notches in the box plots; see **Figure 2** legend) occurs when the T_d value is increased from 0.65 to 0.70, suggesting that setting $T_d = 0.65$ ensures the maximum modularity-preserving partitioning of the network.

Third, as we are interested in the functional consistency of genes within a cluster in addition to topological cohesiveness, we characterized the functional enrichment of all the clusters for a given T_d value using GO BP enrichment analysis (**Figure 2C**). Since this approach will suffer from the very sparse functional annotation of rice genes, we used this analysis only as a rough guide. Following the number of clusters that were significantly enriched with at least one specific GO BP ('function'), we observed that the maximum enrichment again occurs at $T_d = 0.65$ (slightly better than $T_d = 0.70$). However, contrary to what is expected, the number of distinct enriched functions dropped steadily with increasing T_d. Finally, using data from the enrichment analysis, we plotted the distribution of ESs of all the clusters for different T_d values and found that T_d values in the range of 0.65 to 0.80 were giving overall more significant overlap between clusters and functions (**Figure 2D**). Therefore, based on all the four analyses, we decided on a $T_d = 0.65$ to be the best choice for clustering RECoN.

We subsequently clustered RECoN using SPICi with $T_d = 0.65$ to uncover 1744 dense clusters with three or more genes. 28,421

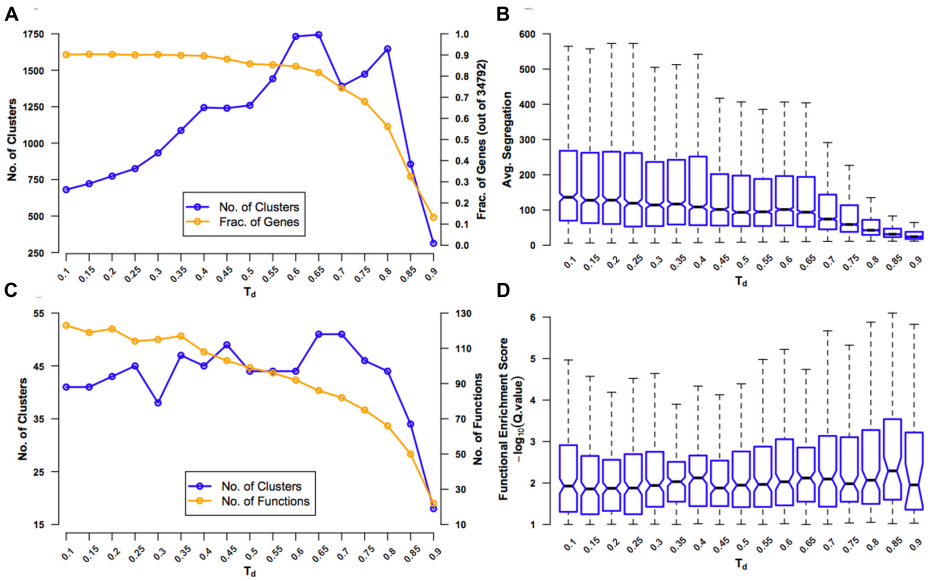

FIGURE 2 | Evaluation of coexpression network clustering. The rice 'environment' coexpression network was clustered using SPICi, for a range of values – 0.1-0.9 – of the density parameter T_d that determines how dense the final clusters are. The clusters obtained using each T_d value were evaluated using several criteria: **(A)** Number of clusters that were formed (left y-axis) and the fraction of 34,792 genes in the original network present in one of the clusters (right y-axis) are plotted. These numbers were calculating by considering only clusters containing three or more genes. As T_d increases, more and more genes are left out of clusters. **(B)** Average segregation of a cluster is a measure of how well genes in that cluster interact with other genes belonging to the same cluster compared to interactions with genes belonging to other clusters. Hence, average segregation measures cluster modularity. The overall modularity at a given T_d value is plotted a box plot, leaving out outlier values above the whiskers for clarity. The center of the box corresponds to the median (2nd quartile; Q_2) of the distribution of average segregation values of all the clusters, and the extremes of the box correspond to the 1st (Q_1) and 3rd (Q_3) quartiles. The whiskers denote $Q_2 \pm 1.5*IQR$, where IQR is the interquartile range ($Q_3 - Q_1$). The notches in each box extend to $\pm 1.58\ IQR/\sqrt{n}$ (n being the sample size) (McGill et al., 1978). They are based on asymptotic normality of the median and roughly equal sample sizes for the medians being compared, and are said to be rather insensitive to the underlying distributions of the samples. The notches give roughly a 95% confidence interval for the difference in two medians. **(C)** The extent of overlap between clusters (defined based on a particular T_d value) and GO BP gene sets (termed 'functions') is measured using the hypergeometric test. The number of clusters with significant overlap (FDR q-value < 0.1) (left y-axis) and number of distinct functions significantly overlapping with the clusters (right y-axis) are plotted. **(D)** Functional enrichment of the clusters is quantified using $-log_{10}$(q-value) and plotted using a box plot representing the distribution of the enrichment scores (ESs) for all the clusters at a given T_d value. Here again, outliers beyond the whiskers have been left out for clarity.

genes (∼81.7% of all the genes in the original network) fell within one of the clusters. Clustering the conditions based on their expression profiles also yields an expected grouping, especially with the drought-, salt-, and cold stress samples clustering together (Supplementary Figure S1). We linked these clusters to BP categories from the GO ontology and to CREs identified using a *de novo* motif discovery pipeline (Harb et al., 2010) (see section "Materials and Methods"). The clusters thus identified can be used in a geneset enrichment analysis framework of new stress transcriptomes of rice. To demonstrate this analytical pipeline, we generated rice drought transcriptomes at three developmental stages, and used RECoN to identify clusters that are significantly perturbed in at least one stage.

Gene Expression Profiling of Drought in Rice

We profiled RNA samples from rice plants treated to drought at the seedling, vegetative and reproductive stages using the rice Affymetrix GeneChips (**Figure 1**, step 1b). In addition, we measured phenotypic and physiological responses of the plants to drought stress (see Supplementary Note and Supplementary Table S2). Statistical analysis of differential expression showed that a large number of genes are perturbed, given a stringent q-value cut-off of < 0.01 (Supplementary Table S3). The largest shift in expression compared to well-watered controls happened at the seedling stage with ∼12,300 genes showing differential expression, compared to only ∼2,500 genes in the reproductive stage and ∼9000 genes at the vegetative stage. A comparison of the differentially expressed genes at the three developmental stages showed that ∼33% of the genes were shared with the genes in the other stages in the case of both up- and down-regulated genes (**Figures 3A,B**).

To see the level of functional enrichment using genesets from the GO, we took a union of all the drought-regulated genes and split them into sets of genes that show identical pattern of regulation across the stages. We then determined the processes defined by GO BP annotations that were enriched in each of these gene sets (**Figures 3C,D**). As expected, the most significant GO term among the set of genes up-regulated in all stages was 'response to water.' Similarly, different combinations of genes involved in protein dephosphorylation and small GTPase-mediated signaling are up-regulated in all stages. Among the genes down-regulated in all stages, photosynthesis and related processes are clearly enriched. Genes involved in translation

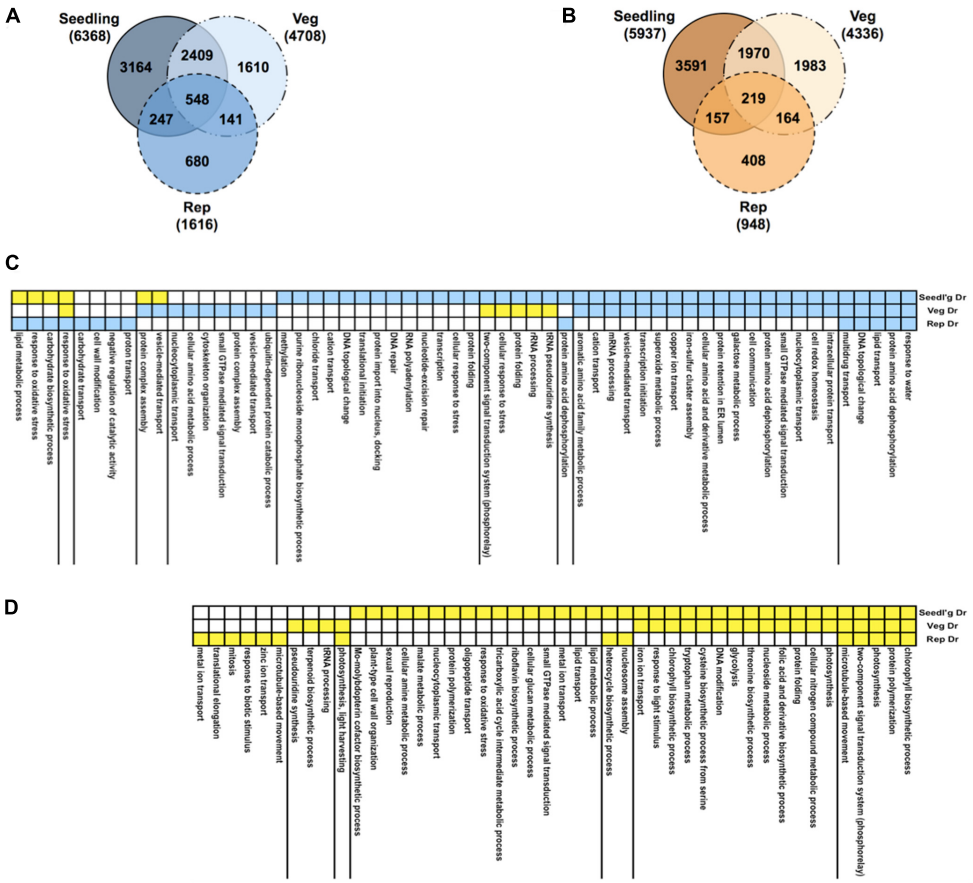

FIGURE 3 | Gene expression profiles under drought. Venn diagrams comparing up-regulated **(A)** and down-regulated **(B)** genes in response to drought in three growth stages: seedling, vegetative, and reproductive. Total numbers of genes for all gene sets are indicated in brackets. Functions, processes and pathways common and specific to various drought stress treatments and time-points. These are defined broadly based on Gene Ontology (GO) biological process (BP) annotations of rice genes. First, the total of all drought-regulated genes from all stages were pooled together and were then partitioned based on the combination of their regulation in the three stages (e.g., up-up-up, or down-up-down). Then, GO BP terms of interest (rows) were identified by analysis of enrichment of the set of genes annotated with a given GO BP term in each regulation-combination defined by the yellow-blue color-coding along the rows where blue means up-regulation and yellow means down-regulation. Statistical significance of enrichment was calculated using the hypergeometric test and terms with q-value < 0.1 in at least one of the treatments were retained. **(C)** GO BP terms enriched in gene sets up-regulated in at least one stage. **(D)** GO BP terms enriched in gene sets only down-regulated in one or more stage.

are induced and repressed in the seedling and reproductive stages, respectively. Cell wall modification genes that are usually repressed by drought (Moore et al., 2008) are also repressed at the seedling stage but specifically up-regulated in the reproductive stage. Comparisons of the GO BP category revealed the most obvious differences between up- and down-regulated genes (Supplementary Table S4). In this category, response to water (GO:0009415), lipid transport (GO:0006869), cellular response to stress (GO:0033554), transcription (GO:0006350), response to oxidative stress and carbohydrate biosynthetic process (GO:0016051) were found at higher proportions in up- than in down-regulated genes. In contrast, the processes photosynthesis (GO: 0015979), chlorophyll biosynthesis (GO:0015995), and glycolysis (GO:0006096) were specifically represented in the down-regulated set of genes.

Although this analysis gave us a few insights into drought-regulated gene expression, apart from the *bona fide* stress response themes, it is hard to pinpoint biological functions that are specifically affected in the different stages. The most important reason for this shortfall is the fact that rice genes are extremely poorly characterized and very few genes have been annotated well. This scenario becomes evident when we look at the small number of genes common between any GO term and the set of drought genes. Therefore, we need to pursue other approaches that will give us a better picture of the underlying changes during drought.

Identification of Drought-Related Clusters from RECoN

The next operation was to interface the information gained from drought expression profiling to identify drought-related modules from the coexpression data (**Figure 1**, step 2). Each of the 1744 clusters from RECoN were tested for enrichment of drought-responsive (up- or down-regulated) genes from any one of the

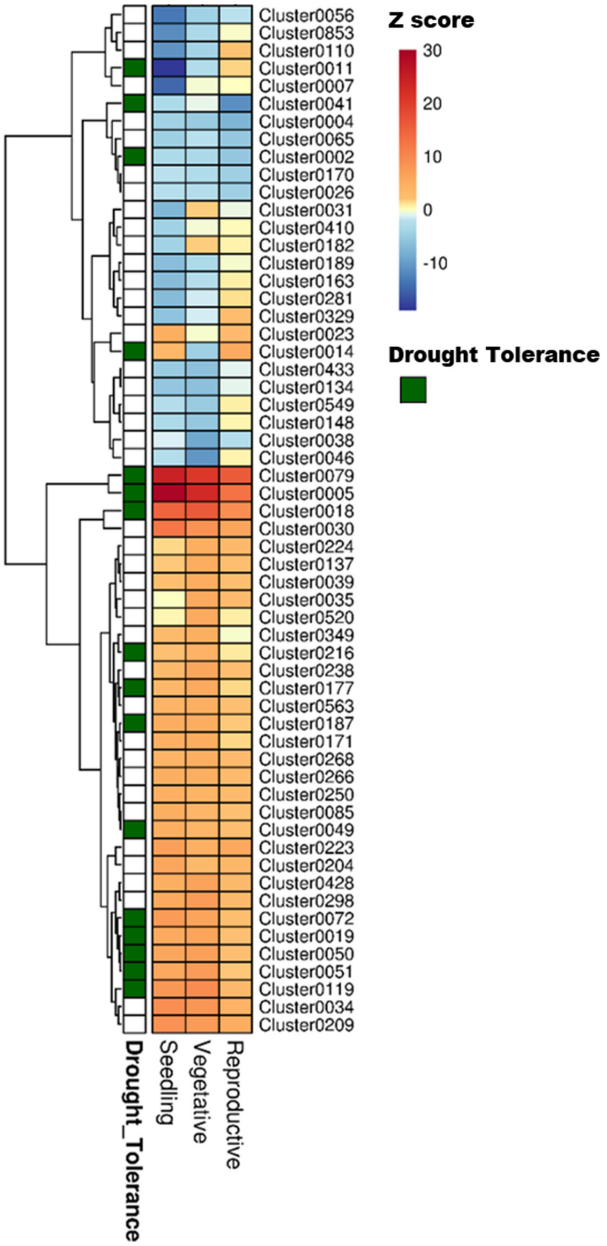

FIGURE 4 | Heat map showing the different clusters enriched in drought-responsive genes in three developmental stages of rice. The fold change values obtained from the differential expression tests was used as a parameter for the parametric analysis of geneset enrichment algorithm. The heat map shows Z scores obtained from the enrichment analysis, color coded with a red and blue gradient for positive and negative enrichment, respectively. The green grids along the rows indicate the 'drought tolerance clusters,' identified by mapping known drought tolerance genes to clusters. A *q*-value cut-off of 0.001 was set as a threshold to select the clusters.

stages (seedling, vegetative, or reproductive) (**Figure 4**). Drought clusters provide a handle on putative functional interactions between genes transcriptionally regulated by drought that were otherwise unassociated parts lists. This makes gene-by-gene interpretation a much easier and constructive process. Moreover, we reasoned that since a cluster is a coherent group of genes, all the genes in a 'drought' cluster might have a role in mediating drought-response, not necessarily by responding to drought through gene expression changes. This is possible by either being ubiquitously present as support machinery (between well-watered and drought conditions) or being conditionally active under drought due to non-transcriptional modes of regulation including post-translational modification. These clusters, hence, provide a means for functionally associating post-transcriptionally modified regulatory/signaling genes to transcriptionally regulated genes.

Examples of Drought Transcriptional Modules

We present here some drought transcriptional modules as examples to showcase the usefulness of this approach in understanding developmental stage-specific drought response. All the genes in drought clusters, their relative expression across the stages and their 'drought' cluster membership are provided in Supplementary Table S5.

Cluster0013 contains 294 genes enriched with genes up-regulated in the seedling stage and down-regulated in the reproductive stage. Genes in this cluster are involved in ribosome biogenesis pathway (q-value $< 10^{-16}$), process related to protein import (q-value $< 10^{-4}$), mitochondrial protein localization (q-value $< 10^{-3}$) (which concerns transporting of mitochondrial oxidative phosphorylation proteins to the mitochondrion), and contain the GCC-core, Telo-box and the Site II motifs in their upstream sequences. This combination of BPs and CREs represents a well-known regulatory program: the site II motifs are recognized by TFs of the TCP family and have been confirmed to be important in the regulation of ribosome protein (RP) genes in combination with the telo-box motif (Tremousaygue et al., 2003). These motifs are co-located in the promoters of about 70% of 216 ribosomal protein genes in Arabidopsis. In addition, there is evidence that the site II motifs also possibly coordinate the expression of nuclear genes encoding components of the mitochondrial oxidative phosphorylation machinery in both Arabidopsis and rice (Welchen and Gonzalez, 2006). Therefore, this program involving site II and telo-box motifs could mediate the down-regulation of major processes that affect protein production under drought stress in the reproductive tissue. The GCC-core motif is known to be bound by AP2-ERF TFs (Ohme-Takagi and Shinshi, 1995), which are involved in gene regulation under a variety of abiotic stresses conserved between Arabidopsis and rice (Nakashima et al., 2009).

Cluster0010 contains 635 genes including genes involved in glucose metabolism (q-value $< 10^{-3}$), terms related to amino acid transport and metabolism (q-value < 0.05), glycolysis (q-value < 0.05), and two-component signal transduction system (phosphorelay) (q-value < 0.05). Genes in this cluster are down-regulated in the seedling and vegetative stages, but up-regulated in reproductive stage. Of particular interest in this cluster is the *OsVIN1* gene (LOC_Os04g45290) coding for a vacuolar invertase gene. OsVIN1 has high fructan exohydrolase activity and is known to play an important role in carbon allocation to developing organs like the reproductive

tissue. The expression of *OsVIN1* is not induced by our drought treatment, and this is in agreement with previous observation that *OsVIN1* is expressed in flag leaves, panicles (the reproductive tissue) and anthers in an essentially drought-insensitive manner (Ji et al., 2005; Parent et al., 2009). It is therefore a case where a gene involved in mediating a process (resource allocation) relevant to drought is not transcriptionally affected, but is associated with other drought-regulated genes in clusters defined by us. Another important gene observed in this cluster is *SNAC3* (LOC_Os01g09550) that has been previously shown to confer tolerance to multiple stresses like salinity, drought, and oxidative stresses (Fang et al., 2015). *OsCPK9* (LOC_Os03g48270), a calcium dependent protein kinase is another key member in this cluster that has previously been shown to impart drought tolerance in transgenic rice plants by enhancing stomatal closure and stomatal adjustment (Wei et al., 2014). Taken together these results suggest that the genes in this cluster contribute to drought tolerance by regulating osmotic adjustment and ROS scavenging processes and can also be putative candidates for increasing yield under drought.

The 193 genes in Cluster0041 are enriched primarily in almost all processes involved in cell cycle, a process integral to panicle development and elongation, and these genes are specifically down-regulated by drought at the reproductive stage (the most drought sensitive stage of rice). Upstream regions of these genes contain the *SEF3* binding site/ACII element, *MYB* recognition site found in *rd22* and other genes, and E2F consensus, potential binding sites of TFs that have been implicated to be important in regulating cell cycle in the reproductive tissue of Arabidopsis (Hennig et al., 2004).

The other aspect of using this approach is in discovery of drought tolerance genes. A variety of gene families with regulatory function have been shown to have a role in drought tolerance by overexpression/knockout experiments, and that regulate a battery of downstream genes (Umezawa et al., 2006). Therefore, to evaluate this aspect, we first cataloged a number of genes that confer drought tolerance in rice on overexpression or knockout, and then mapped them to RECoN clusters (**Figure 5**). The primary observation is that almost all the drought tolerance genes were part of drought clusters. However, this observation could be trivial if all those genes are indeed regulated by drought in the first place. Out of the 54 genes presented here, 45 are indeed regulated by drought in stage-specific or independent manner while 9 of these are not drought-regulated, but are associated with a drought module. Therefore, we reaffirm that the approach lends itself to identification of genes that are not necessarily transcriptionally perturbed by drought, if at all regulated by it. Some examples for the drought-tolerance clusters follow.

Cluster0079 contains 71 genes. The genes in this cluster include a receptor-like cytoplasmic kinase *OsRLCK253* (LOC_Os08g28710) and a phosphatase *OsPP108* (LOC_Os09g15670), both of which have been shown to improve drought tolerance in transgenic Arabidopsis plants (Giri et al., 2011; Singh et al., 2015). Along with the aforementioned genes, this cluster also comprises of other known drought tolerance genes comprising of dehydrins like *OsLea3-1* (LOC_Os05g46480) (Brohee and van Helden, 2006), enzymes like *OsUGE-1* (LOC_Os05g51670) (Nardini et al., 2011) and TFs like *OsDREB2A* (LOC_Os01g07120) (Skirycz and Inze, 2010), *SNAC2*/*OsNAC6* (LOC_Os01g66120) (Bergmann et al., 2004), *CMYB1*(LOC_Os02g46030) (Kitano, 2002) and *ZFP182* (LOC_Os03g60560) (**Figure 6A**). Most genes in the cluster are up-regulated by drought in all three developmental stages, which appears to indicate the diverse roles that these proteins play including detoxification, osmotic adjustment, and signaling pathways. Since our data revealed many putative stress inducible genes, a few of these genes are likely to have a dual role as developmentally regulated and stress responsive. Nevertheless, the functional role of these genes needs to be characterized to further enhance our understanding of the mechanisms that impart drought/abiotic stress tolerance to rice.

Cluster0424 contains 20 genes enriched specifically with reproductive drought and these genes too contain an ABRE-like motif – HACGYGTNS – in their upstream sequence. The drought tolerance genes part of this cluster are *OsDREB1F* (LOC_Os01g73770) (Barrero et al., 2007; Choi et al., 2007), *OsNPKL2* (LOC_Os01g50400) and *OsNPKL3* (LOC_Os01g50410) (**Figure 6B**). The tandem duplicate genes *NPKL2* and *NPKL3* are previously known to be strongly induced by drought at the reproductive stage (Ning et al., 2008). Highly induced expression of these genes under drought stress indicates that these two genes can be candidates for drought tolerance and increased yield under drought, for their potential role as a kinase as they are found to be located in the genomic region with three QTLs: RSN (relative number of spikelets per panicle under drought stress) and LDS (leaf drying score), which are mainly related to drought tolerance, and DIDRV (deep root rate in volume induced by drought conditions) (data not shown). Although expression of *OsDERF1* is induced by drought and phyothormone treatments, its overexpression in rice negatively regulates drought tolerance by repressing ethylene biosynthesis by binding to ERF repressors *OsERF3* and *OsAP2-39* (Wan et al., 2011). *OsDREB1F* was induced by abiotic stresses including osmotic stress (using PEG) as well as ABA treatment and overexpression in rice and Arabidopsis gave drought tolerance that can be mediated by ABA dependent pathway (Choi et al., 2007). However, our progressive drought treatment does not perturb this gene (at least not at the stringent level of significance chosen). Developmental stage-specific drought-regulation of *OsDREB1F* is not clear except that the gene by itself is expressed differently in different stages and tissues. We therefore implicate *OsDREB1F* as being important in progressive drought response at the reproductive stage.

Cluster0177 contains 33 genes involved in the regulation of innate immune/defense/stress response as well as response to jasmonic acid and salicylic acid. Drought-regulated genes in this cluster are up-regulated specifically in the seedling and vegetative stages. This cluster again contains the drought-tolerance genes *OsMAPK5* (LOC_Os03g17700) and *OsNPKL4* (LOC_Os01g50420). *OsMAPK5* is known to be induced by drought, other abiotic stresses and ABA, as well as pathogen infection and that the overexpressiwon gives abiotic stress

Cluster	No. Genes	No. Dr. Tolerance genes	GeneID:GeneName	Rice Dr		
				Seedl'g	Veg	Rep
Cluster0079	71	6	Os08g28710:OsRLCK25 | Os09g15670:OsPP108 | Os01g66120:OsNAC6/SNAC2 | Os01g07120:OsDREB2A | Os02g46030:OsCMYB1 | Os03g60560:ZPF182	24.829	28.811	13.374
Cluster0001	1713	3	Os05g28350:OsABI4 | Os07g01770:OsSMPC1 | Os08g06050:OsTOP6A2/SPO11-2	196.385	4.990	0.000
Cluster0424	20	4	Os01g73770:OsDREB1F | Os01g50400:OsNPKL2 | Os01g50410:OsNPKL3 | Os08g35240:OsDERF1	0.000	0.000	1.509
Cluster0005	972	3	Os08g43334:OsSIPP2C1 | Os03g27280:SAPK1 | Os01g64000:OsABI5 | Os08g45110:OsDREB2C | Os01g39020: OsHSFA7	134.490	93.440	13.338
Cluster0006	1545	4	Os01g73250:OsMYB4 | Os01g62200:OsERF4a | Os03g57240:OsDST | Os01g03570:OXHS2	-296.778	-197.347	-85.746
Cluster0018	252	3	Os02g52780:OsbZIP23 | Os05g10670:OsTZF1 Os09g28310:OsbZIP72	82.263	79.406	8.123
Cluster0050	34	2	Os01g58420:AP37 | Os05g38660:OsSIK1	7.315	6.867	1.816
Cluster0051	62	2	Os11g03300:OsNAC10 | Os11g08210:OsNAC5	6.671	9.003	0.000
Cluster0119	52	2	Os08g39450:OsSAP11 | Os09g35030:OsDREB1A | Os09g28440:OsEATB | Os09g35010:OsDREB1B	12.484	6.681	1.584
Cluster0177	33	2	Os03g17700:OsMAPK5 | Os01g50420:OsNPKL4	2.211	2.375	0.000
Cluster0216	15	2	Os02g43790:OsAP59 | Os12g39400:ZFP252	1.937	3.514	0.000
Cluster0003	1601	1	Os01g50370:OsNPKL1	-63.469	4.752	12.491
Cluster0011	507	1	Os02g50970:OsDSM1	-112.341	-4.211	-10.141
Cluster0014	392	2	Os01g55450:OsCIPK12 | Os03g41460:SAPK10	8.685	-18.297	0.000
Cluster0016	290	1	Os07g48760:OsCIPK03	-1.248	2.201	0.000
Cluster0019	160	1	Os06g41010:OsSAP8	25.531	22.722	0.000
Cluster0049	75	1	Os05g27930:OsDREB2B	21.794	11.431	0.000
Cluster0062	65	1	Os01g01420:OsCOIN	2.680	-4.680	0.000
Cluster0072	38	1	Os05g45020:OsMT1a	5.377	6.687	1.198
Cluster0108	37	3	Os04g43680:OsMYB4 | Os03g08310:OsJAZ9 | Os03g53020:OsbHLH148	1.179	0.000	1.700
Cluster0111	29	1	Os02g50930:OsBIRF1	0.000	2.342	0.000
Cluster0187	23	1	Os01g62410:OsMYB3R-2	10.406	9.990	0.000
Cluster0218	13	1	Os03g17610:OsTOP6A3/SPO11-3	1.249	0.000	0.000
Cluster0352	17	1	Os09g10770:OsTOP6B	5.451	0.000	0.000
Cluster0396	9	1	Os02g52250:OsSKIPa	2.868	1.853	0.000
Cluster0508	14	1	Os04g49510:OsCDPK7	2.161	0.000	1.878
Cluster0728	7	1	Os03g60080:SNAC1	1.828	1.584	0.000
Cluster1207	4	1	Os02g08440:OsABF2	2.212	0.000	0.000
Cluster1478	4	1	Os02g43970:OsARAG1	-1.285	0.000	0.000

FIGURE 5 | Drought clusters containing known drought tolerance genes. Genes in black are regulated by drought at one or more of the growth stages while genes in red are not drought-regulated. The values in the color-coded columns correspond to the level of significance (measured as score equal to the $-log_{10}[q\text{-}value]$) of drought-regulated genes. For convenience the scores themselves are signed and colored based on the direction of their regulation (+/blue – up-regulation; –/yellow -down). Since only enrichments with q-value < 0.1 were considered, all the other values were set to 1 (because of which, their negative logarithms are 0 s).

tolerance but disease susceptibility (Xiong and Yang, 2003). It is hence considered to be a key link in the cross talk between disease resistance and abiotic stress tolerance. We propose that other genes in this cluster are putative links of crosstalk between the stresses. Previous research has shown that *OsNPKL4* is very strongly induced at the seedling stage, but has a moderate to low level of induction at the anthesis stage (Ning et al., 2008), consistent with the drought-pattern of this cluster.

Cluster 0108 contains 37 genes with three drought tolerance genes. *OsMYB4* (LOC_Os04g43680), characterized as a universal stress response gene induced under a variety of biotic and abiotic stresses gives abiotic stress tolerance when overexpressed in apples (Yu et al., 2006; Narsai et al., 2013) by modulating osmolytic balance, *OsbHLH148* (LOC_Os03g53020) and its interacting partner *OsJAZ1* (LOC_Os03g08310) mediates drought response via the jasmonic acid pathway (Seo et al.,

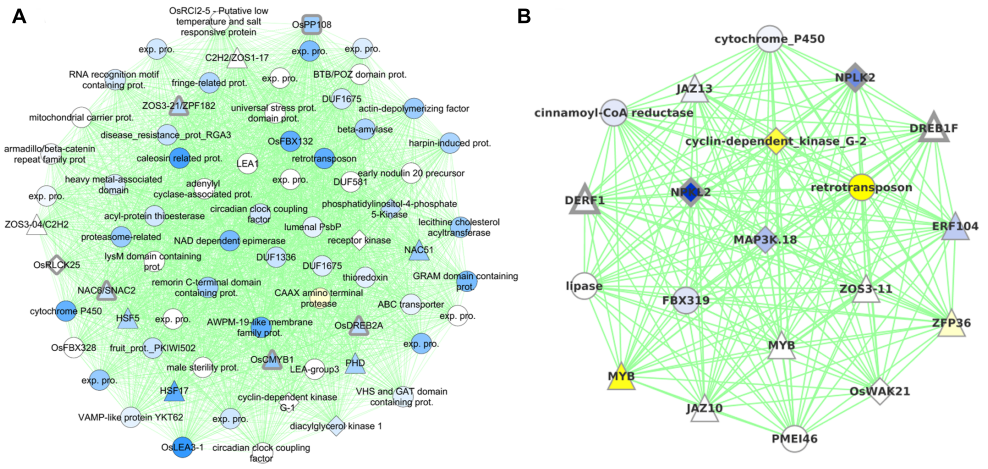

FIGURE 6 | Graphical visualization of **(A)** 71 genes in Cluster0079 that contains six drought tolerance genes (with thick gray borders) and **(B)** 20 genes in Cluster0424 that contains four drought tolerance genes. All the coexpression edges are colored green. Node shapes correspond to type of gene: triangles are TFs, diamonds are protein kinases, rounded squares are protein phosphatases and circles are other genes. Node color corresponds to the level of differential expression under drought in the vegetative stage for Cluster0079 and reproductive stage for Cluster0424 (where the clusters have maximum enrichment): blue for up-regulation and yellow for down-regulation. Uncharacterized gene are labeled 'exp. pro.' (for 'expressed protein').

2011). Genes in this cluster also could potentially be involved in jasmonic acid mediated hormonal signal transduction.

A Web Interface for Further Exploration of Rice Abiotic Stress Response

We have made RECoN available at https://plantstress-pereira.uark.edu/RECoN/. The platform provides an interface to perform an exploratory analysis of abiotic stress response in rice in three different modes (**Figure 7**). The cluster enrichment tool allows users to upload a genome-wide differential expression profile and test the enrichment of abiotic stress coexpression clusters defined by us in this study. As results, all the clusters enriched (Kim and Volsky, 2005) within a user-defined significance threshold (Z-scores of clusters with mean change significantly larger than the background), along with the cluster size and a representative GO BP term within the cluster (the term with the lowest q-value) are displayed. Users can click on clusters of interest to view various attributes of the cluster. The functional information of the cluster is displayed in separate 'Process' and 'KEGG Pathways' tabs, as well as all the regulatory sequences that were found enriched in each cluster under the 'Motifs' tab. The 'Genes' tab shows all the genes within that cluster that were also present in the user data, along with other attributes linked to annotations from the MSU database (Kawahara et al., 2013), their matched locus ids in the RAP-DB (Sakai et al., 2013), Arabidopsis homologs and links to Manually Curated Database of Rice Proteins (Gour et al., 2014) to explore ontologies the gene is annotated to. The page also displays the differential expression value of each gene in all the three drought transcriptomes generated by us and reported in this study, as well as in the data uploaded by the user. The 'Graph' tab displays the cluster in a Cytoscape-web enabled page (Lopes et al., 2010). The interactive graphical display sets node (gene) attributes to highlight changes in the user provided transcriptome.

To test this utility, we queried RECoN with an independent dataset listed as GSE14275 in GEO. This dataset quantified mRNA from rice seedlings exposed to heat shock treatments (Hu et al., 2009). We calculated and uploaded the text file with $\log_2$ transformed fold change values of all the genes using the cluster enrichment tool. The results page shows 82 clusters enriched with a threshold of q-value < 0.01. Of these, Cluster0223 has the highest Z-score (22.01) (Supplementary Figure S2). From the 25 genes in this cluster, 8 (32%) are heat shock proteins, with a single TF listed as Heat Stress Transcription Factor A2A (HSFA2A) (Supplementary Figure S3). The homolog of HSFA2A is known to modulate heat-stress response in Arabidopsis (Lämke et al., 2016), making HSFA2A an interesting candidate for heat stress study in rice. As expected of the queried dataset, the GO term 'response to heat stress' is most significantly enriched within this cluster ($-\log_{10}$ q-value 4.186), along with 'response to abiotic stimulus' ($-\log_{10}$ q-value 4.186), and 'protein folding' ($-\log_{10}$ q-value 1.845) (Supplementary Figure S4). This corresponds well with the enriched KEGG pathway 'protein processing in endoplasmic reticulum' ($-\log_{10}$ q-value 7.705). The Cytoscape graph page shows the gene peptidyl-prolyl isomerase is most highly upregulated gene in response to heat amongst all genes in the cluster (Supplementary Figure S5), and its homolog in Arabidopsis interacts with heat shock proteins to modulate thermotolerance (Meiri and Breiman, 2009).

The platform also allows users to perform searches for a BP of interest and retrieve clusters linked to the search term (**Figure 7**). For example, RECoN can be directly searched with the term 'response to heat,' which retrieves two clusters- Cluster0223 as described above and the large Cluster0005 with 972 genes (Supplementary Figure S6). According to GO terms annotated to Cluster0005, these genes appear to be involved in RNA processing

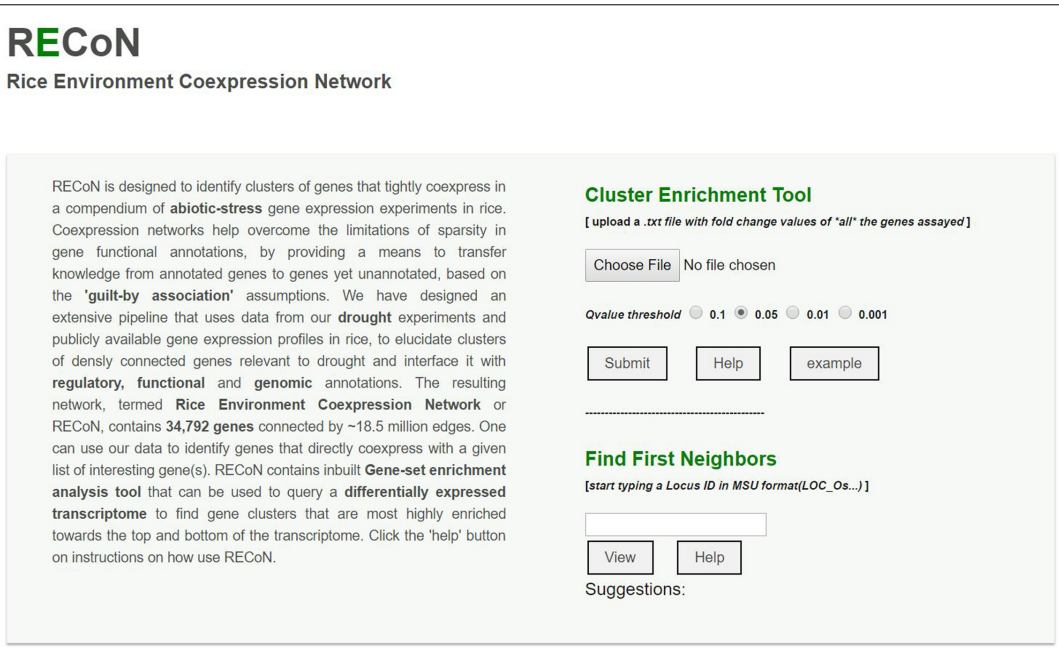

FIGURE 7 | A screenshot of the RECoN webserver available at the link provided in the main text. The online platform allows two types of analyses. The user can upload a genome-wide differential expression profile using the 'choose file' option, which will be used by the cluster enrichment tool to identify clusters that are significantly perturbed in the uploaded transcriptome, within the selected *q*-value threshold. The uploaded file should contain two columns (with headers) with MSU formatted rice gene locus IDs in the first column and their respective fold change values determined from the differential expression tests in the second column. The results will be displayed in a new page with enriched clusters listed and links to display each cluster using Cytoscape-web, as well as BPs and *cis*-regulatory elements (CREs) enriched in the clusters (see Supplementary Figures S2–S4). In cases where a single gene is of interest (rather than a genome-wide analysis), its locus ID can be entered in the input box under the 'Find First Neighbors' section. This analysis will report the genes within one path length of the query gene and a default coexpression score of 0.80 (which can be changed from the results page, see Supplementary Figure S5).

and alternative splicing during general abiotic stress responses (Supplementary Figure S7), and this conclusion is reinforced by enrichment of the KEGG pathway 'Spliceosome' ($-\log_{10}$ *q*-value 3.468) within this cluster.

Apart from this, RECoN also allows querying a single guide gene and retrieve a coexpression neighborhood to explore its functional context. By default, this service will display all the coexpressed genes with edge scores above 0.80, which can be altered from the results page to increase or decrease the subnetwork size (Supplementary Figure S8). Together, all these functionalities in RECoN will enable biologists gain a network-based understanding of the abiotic-stress response in rice and prioritize candidates for studying experimental phenotypes.

DISCUSSION

Plant responses to environmental stress span across several layers of organization including signaling, transcription, and metabolism, making it vital to understand stress response at the systems-level. For less studied models like rice, the current scope for systems analysis is mostly restricted to transcriptional profiling under various conditions. Therefore, to make the best use of currently available data in rice, we have created a resource for exploration of transcriptional, developmental, functional, and regulatory aspects of abiotic-stress response in rice.

We sought to organize genes into coherent groups and work further from there. To this end, we designed and implemented a pipeline for automatic mining of condition-specific gene expression datasets intended for analysis of coexpression. At a practical level, accurate quantification of gene expression using technologies like Affymetrix GeneChips has been hard due to the problem of cross-hybridization. This has been noted to affect calculation of coexpression (Casneuf et al., 2007) and the proposed solution is a remapping of microarray probes to genes to ensure unique hybridization (Dai et al., 2005). We hence used a custom probe-gene mapping and used this reannotation to make reliable estimation of gene expression across 45 conditions. Then, a coexpression network was built (RECoN) and clustered to obtain tightly coexpressed groups of genes that revealed the modular organization of genes.

We demonstrated the use of RECoN by analyzing new stage-specific drought transcriptomes. In order to both understand drought response and discover novel drought tolerance genes, we combined drought-responsive genes from our experiments with the transcriptional modules to uncover drought clusters, where each cluster, by design, contains several genes in addition to genes transcriptionally regulated by drought. Drought modules thus present an opportunity to discover regulatory genes that do not change in gene expression but can affect the response mediated by that module. In this process, we are basically imputing uncharacterized genes within a cluster with the function/role of

characterized genes (even at the level of transcriptional response). In species with very little annotation, such as rice, cluster-level function prediction has been shown to be useful (Song and Singh, 2009). We have validated this approach by inspecting the cluster membership of known drought tolerance genes that are not drought-responsive but are associated with a cluster that is enriched in genes following a drought expression pattern expected from what is known about the tolerance gene.

With the enormous amount of data generated in this work that can be used for inference of gene function and pathway analysis, all these results are summarized and presented in a flexible visual interface for dynamic exploration. The RECoN online platform, to the best of our knowledge, is the first of a kind that allows users to upload their own transcriptomic data (e.g., output of an RNA-seq assay) and find clusters that are significantly enriched. The clusters are linked to GO BPs and KEGG pathways in the current version, which will be regularly updated with other newer ontologies as they become available (e.g., Trait Ontology, Stress Ontology, etc.). RECoN has the potential to enhance the functional interpretation of high-throughput expression data, with diverse set of information made available within a single platform that can be searched quickly.

The approach presented here is widely applicable: as shown with an independent test dataset from heat stress, genome-wide transcriptional modules recovered here on the basis of gene expression under different environmental conditions can be similarly extended to study other abiotic stresses including salt and cold to find common stress-specific modules. This approach lends itself to identification of abiotic stress related genes that are usually hidden in a typical transcriptome assay.

AUTHOR CONTRIBUTIONS

AK: designed the network, conducted statistical analysis, drafted the manuscript; CG: contributed text, updated figures, and created the webserver; MA: conducted rice drought and gene expression experiments; AP: designed experiments and coordinated research; all authors contributed to writing the manuscript.

FUNDING

The study was supported by National Science Foundation (DBI-0922747).

ACKNOWLEDGMENTS

We thank members of the Pereira lab- Supratim Basu, Venkategowda Ramegowda, Ritu Mihani, and Julie Thomas for testing and validating the network clusters for gene function analysis, some of which are in publications. We also thank Pawel Wolinski from the Arkansas High Performance Computing Center for his contribution to creation of webserver.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpls.2017.01640/full#supplementary-material

REFERENCES

Ambavaram, M. M., Krishnan, A., Trijatmiko, K. R., and Pereira, A. (2011). Coordinated activation of cellulose and repression of lignin biosynthesis pathways in rice. *Plant Physiol.* 155, 916–931. doi: 10.1104/pp.110.168641

Ambavaram, M. M. R., Basu, S., Krishnan, A., Ramegowda, V., Batlang, U., Rahman, L., et al. (2014). Coordinated regulation of photosynthesis in rice increases yield and tolerance to environmental stress. *Nat. Commun.* 5, 5302. doi: 10.1038/ncomms6302

Bader, G. D., and Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4:2. doi: 10.1186/1471-2105-4-2

Barrero, J. M., Gonzalez-Bayon, R., del Pozo, J. C., Ponce, M. R., and Micol, J. L. (2007). INCURVATA2 encodes the catalytic subunit of DNA polymerase α and interacts with genes involved in chromatin-mediated cellular memory in *Arabidopsis thaliana*. *Plant Cell* 19, 2822–2838. doi: 10.1105/tpc.107.054130

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., et al. (2009). NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res.* 37, D885–D890. doi: 10.1093/nar/gkn764

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300.

Bergmann, S., Ihmels, J., and Barkai, N. (2004). Similarities and differences in genome-wide expression data of six organisms. *PLOS Biol.* 2:e9. doi: 10.1371/journal.pbio.0020009

Brohee, S., and van Helden, J. (2006). Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* 7:488. doi: 10.1186/1471-2105-7-488

Casneuf, T., Van de Peer, Y., and Huber, W. (2007). In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation. *BMC Bioinformatics* 8:461. doi: 10.1186/1471-2105-8-461

Choi, K., Park, C., Lee, J., Oh, M., Noh, B., and Lee, I. (2007). Arabidopsis homologs of components of the SWR1 complex regulate flowering and plant development. *Development* 134, 1931–1941. doi: 10.1242/dev.001891

Counce, P. A., Keisling, T. C., and Mitchell, A. J. (2000). A uniform, objective, and adaptive system for expressing rice development. *Crop Sci.* 40, 436–443. doi: 10.2135/cropsci2000.402436x

Dai, M., Wang, P., Boyd, A. D., Kostov, G., Athey, B., Jones, E. G., et al. (2005). Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* 33, e175. doi: 10.1093/nar/gni179

David, F. N. (1949). The moments of the Z and F distributions. *Biometrika* 36, 394–403. doi: 10.1093/biomet/36.3-4.394

Deyholos, M. K. (2010). Making the most of drought and salinity transcriptomics. *Plant Cell Environ.* 33, 648–654. doi: 10.1111/j.1365-3040.2009.02092.x

Elemento, O., Slonim, N., and Tavazoie, S. (2007). A universal framework for regulatory element discovery across all genomes and data types. *Mol. Cell* 28, 337–350. doi: 10.1016/j.molcel.2007.09.027

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30, 1575–1584. doi: 10.1093/nar/30.7.1575

Fang, Y., Liao, K., Du, H., Xu, Y., Song, H., Li, X., et al. (2015). A stress-responsive NAC transcription factor SNAC3 confers heat and drought tolerance through modulation of reactive oxygen species in rice. *J. Exp. Bot.* 66, 6803–6817. doi: 10.1093/jxb/erv386

Fu, F. F., and Xue, H. W. (2010). Co-expression analysis identifies Rice Starch Regulator1 (RSR1), a rice AP2/EREBP family transcription factor, as a novel

rice starch biosynthesis regulator. *Plant Physiol.* 154, 927–938. doi: 10.1104/pp.110.159517

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* 5:R80. doi: 10.1186/gb-2004-5-10-r80

Giri, J., Vij, S., Dansana, P. K., and Tyagi, A. K. (2011). Rice A20/AN1 zinc-finger containing stress-associated proteins (SAP1/11) and a receptor-like cytoplasmic kinase (OsRLCK253) interact via A20 zinc-finger and confer abiotic stress tolerance in transgenic Arabidopsis plants. *New Phytol.* 191, 721–732. doi: 10.1111/j.1469-8137.2011.03740.x

Gour, P., Garg, P., Jain, R., Joseph, S. V., Tyagi, A. K., and Raghuvanshi, S. (2014). Manually curated database of rice proteins. *Nucleic Acids Res.* 42, D1214–D1221. doi: 10.1093/nar/gkt1072

Harb, A., Krishnan, A., Ambavaram, M. M. R., and Pereira, A. (2010). Molecular and physiological analysis of drought stress in Arabidopsis reveals early responses leading to acclimation in plant growth. *Plant Physiol.* 154, 1254–1271. doi: 10.1104/pp.110.161752

Hartwell, L., Hopfield, J., Leibler, S., and Murray, A. (1999). From molecular to modular cell biology. *Nature* 402, C47–C52.

Hennig, L., Gruissem, W., Grossniklaus, U., and Kohler, C. (2004). Transcriptional programs of early reproductive stages in Arabidopsis. *Plant Physiol.* 135, 1765–1775. doi: 10.1104/pp.104.043182

Higo, K., Ugawa, Y., Iwamoto, M., and Korenaga, T. (1999). Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res.* 27, 297–300. doi: 10.1093/nar/27.1.297

Hu, W., Hu, G., and Han, B. (2009). Genome-wide survey and expression profiling of heat shock proteins and heat shock factors revealed overlapped and stress specific response under abiotic stresses in rice. *Plant Sci.* 176, 583–590. doi: 10.1016/j.plantsci.2009.01.016

Huttenhower, C., Schroeder, M., Chikina, M. D., and Troyanskaya, O. G. (2008). The Sleipnir library for computational functional genomics. *Bioinformatics* 24, 1559–1561. doi: 10.1093/bioinformatics/btn237

Ihaka, R., and Gentleman, R. (1996). R: a language for data analysis and graphics. *J. Comput. Graph Stat.* 5, 299–314.

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003). Summaries of affymetrix GeneChip probe level data. *Nucleic Acids Res.* 31, e15. doi: 10.1093/nar/gng015

Ji, X. M., Raveendran, M., Oane, R., Ismail, A., Lafitte, R., Bruskiewich, R., et al. (2005). Tissue-specific expression and drought responsiveness of cell-wall invertase genes of rice at flowering. *Plant Mol. Biol.* 59, 945–964. doi: 10.1007/s11103-005-2415-8

Jiang, P., and Singh, M. (2010). SPICi: a fast clustering algorithm for large biological networks. *Bioinformatics* 26, 1105–1111. doi: 10.1093/bioinformatics/btq078

Kawahara, Y., de la Bastide, M., Hamilton, J., Kanamori, H., McCombie, W., Ouyang, S., et al. (2013). Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice* 6, 1–10. doi: 10.1186/1939-8433-6-4

Kim, S.-Y., and Volsky, D. J. (2005). PAGE: Parametric Analysis of Gene Set Enrichment. *BMC Bioinformatics* 6:144. doi: 10.1186/1471-2105-6-144

Kitano, H. (2002). Systems biology: a brief overview. *Science* 295, 1662–1664. doi: 10.1126/science.1069492

Lämke, J., Brzezinka, K., and Bäurle, I. (2016). HSFA2 orchestrates transcriptional dynamics after heat stress in *Arabidopsis thaliana*. *Transcription* 7, 111–114. doi: 10.1080/21541264.2016.1187550

Lopes, C. T., Franz, M., Kazi, F., Donaldson, S. L., Morris, Q., and Bader, G. D. (2010). Cytoscape Web: an interactive web-based network browser. *Bioinformatics* 26, 2347–2348. doi: 10.1093/bioinformatics/btq430

Mahony, S., and Benos, P. V. (2007). STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.* 35, W253–W258. doi: 10.1093/nar/gkm272

Mao, L., Van Hemert, J., Dash, S., and Dickerson, J. (2009). Arabidopsis gene co-expression network and its functional modules. *BMC Bioinformatics* 10:346. doi: 10.1186/1471-2105-10-346

McGill, R., Tukey, J. W., and Larsen, W. A. (1978). Variations of box plots. *Am. Stat.* 32, 12–16. doi: 10.2307/2683468

Meiri, D., and Breiman, A. (2009). Arabidopsis ROF1 (FKBP62) modulates thermotolerance by interacting with HSP90.1 and affecting the accumulation of HsfA2-regulated sHSPs. *Plant J.* 59, 387–399. doi: 10.1111/j.1365-313X.2009.03878.x

Mentzen, W. I., and Wurtele, E. S. (2008). Regulon organization of Arabidopsis. *BMC Plant Biol.* 8:99. doi: 10.1186/1471-2229-8-99

Moore, J. P., Vicre-Gibouin, M., Farrant, J. M., and Driouich, A. (2008). Adaptations of higher plant cell walls to water loss: drought vs desiccation. *Physiol. Plant.* 134, 237–245. doi: 10.1111/j.1399-3054.2008.01134.x

Nakashima, K., Ito, Y., and Yamaguchi-Shinozaki, K. (2009). Transcriptional regulatory networks in response to abiotic stresses in Arabidopsis and grasses. *Plant Physiol.* 149, 88–95. doi: 10.1104/pp.108.129791

Nardini, A., Lo, G. M. A., and Salleo, S. (2011). Refilling embolized xylem conduits: is it a matter of phloem unloading? *Plant Sci.* 180, 604–611. doi: 10.1016/j.plantsci.2010.12.011

Narsai, R., Wang, C., Chen, J., Wu, J., Shou, H., and Whelan, J. (2013). Antagonistic, overlapping and distinct responses to biotic stress in rice (*Oryza sativa*) and interactions with abiotic stress. *BMC Genomics* 14:93. doi: 10.1186/1471-2164-14-93

Ning, J., Liu, S., Hu, H., and Xiong, L. (2008). Systematic analysis of NPK1-like genes in rice reveals a stress-inducible gene cluster co-localized with a quantitative trait locus of drought resistance. *Mol. Genet. Genomics* 280, 535–546. doi: 10.1007/s00438-008-0385-7

Ohme-Takagi, M., and Shinshi, H. (1995). Ethylene-inducible DNA binding proteins that interact with an ethylene-responsive element. *Plant Cell* 7, 173–182. doi: 10.1105/tpc.7.2.173

Parent, B., Hachez, C., Redondo, E., Simonneau, T., Chaumont, F., and Tardieu, F. (2009). Drought and abscisic acid effects on aquaporin content translate into changes in hydraulic conductivity and leaf growth rate: a trans-scale approach. *Plant Physiol.* 149, 2000–2012. doi: 10.1104/pp.108.130682

Parkinson, H., Kapushesky, M., Kolesnikov, N., Rustici, G., Shojatalab, M., Abeygunawardena, N., et al. (2009). ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res.* 37, D868–D872. doi: 10.1093/nar/gkn889

Pearce, S., Ferguson, A., King, J., and Wilson, Z. A. (2015). FlowerNet: a gene expression correlation network for anther and pollen development. *Plant Physiol.* 167, 1717–1730. doi: 10.1104/pp.114.253807

Saeed, A. I., Bhagabati, N. K., Braisted, J. C., Liang, W., Sharov, V., Howe, E. A., et al. (2006). TM4 microarray software suite. *Methods Enzymol.* 411, 134–193. doi: 10.1016/S0076-6879(06)11009-5

Sakai, H., Lee, S. S., Tanaka, T., Numa, H., Kim, J., Kawahara, Y., et al. (2013). Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. *Plant Cell Physiol.* 54, e6. doi: 10.1093/pcp/pcs183

Seki, M., Narusaka, M., Ishida, J., Nanjo, T., Fujita, M., Oono, Y., et al. (2002). Monitoring the expression profiles of 7000 *Arabidopsis* genes under drought, cold and high-salinity stresses using a full-length cDNA microarray. *Plant J.* 31, 279–292. doi: 10.1046/j.1365-313X.2002.01359.x

Seo, J. S., Joo, J., Kim, M. J., Kim, Y. K., Nahm, B. H., Song, S. I., et al. (2011). OsbHLH148, a basic helix-loop-helix protein, interacts with OsJAZ proteins in a jasmonate signaling pathway leading to drought tolerance in rice. *Plant J.* 65, 907–921. doi: 10.1111/j.1365-313X.2010.04477.x

Shaik, R., and Ramakrishna, W. (2013). Genes and co-expression modules common to drought and bacterial stress responses in *Arabidopsis* and rice. *PLOS ONE* 8:e77261. doi: 10.1371/journal.pone.0077261

Shinozaki, K., Yamaguchi-Shinozaki, K., and Seki, M. (2003). Regulatory network of gene expression in the drought and cold stress responses. *Curr. Opin. Plant Biol.* 6, 410–417. doi: 10.1016/S1369-5266(03)00092-X

Singh, A., Jha, S. K., Bagri, J., and Pandey, G. K. (2015). ABA inducible rice protein phosphatase 2C confers ABA insensitivity and abiotic stress tolerance in Arabidopsis. *PLOS ONE* 10:e0125168. doi: 10.1371/journal.pone.0125168

Skirycz, A., and Inze, D. (2010). More from less: plant growth under limited water. *Curr. Opin. Biotechnol.* 21, 197–203. doi: 10.1016/j.copbio.2010.03.002

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* 3, Article3. doi: 10.2202/1544-6115.1027

Song, J., and Singh, M. (2009). How and when should interactome-derived clusters be used to predict functional modules and protein function? *Bioinformatics* 25, 3143–3150. doi: 10.1093/bioinformatics/btp551

Storey, J. D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.* 100, 9440–9445. doi: 10.1073/pnas.1530509100

Tremousaygue, D., Garnier, L., Bardet, C., Dabos, P., Herve, C., and Lescure, B. (2003). Internal telomeric repeats and 'TCP domain' protein-binding sites co-operate to regulate gene expression in *Arabidopsis thaliana* cycling cells. *Plant J.* 33, 957–966. doi: 10.1046/j.1365-313X.2003.01682.x

Umezawa, T., Fujita, M., Fujita, Y., Yamaguchi-Shinozaki, K., and Shinozaki, K. (2006). Engineering drought tolerance in plants: discovering and tailoring genes to unlock the future. *Curr. Opin. Biotechnol.* 17, 113–122. doi: 10.1016/j.copbio.2006.02.002

Usadel, B., Obayashi, T., Mutwil, M., Giorgi, F. M., Bassel, G. W., Tanimoto, M., et al. (2009). Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant Cell Environ.* 32, 1633–1651. doi: 10.1111/j.1365-3040.2009.02040.x

Wan, L., Zhang, J., Zhang, H., Zhang, Z., Quan, R., Zhou, S., et al. (2011). Transcriptional activation of *OsDERF1* in *OsERF3* and *OsAP2*-39 negatively modulates ethylene synthesis and drought tolerance in rice. *PLOS ONE* 6:e25216. doi: 10.1371/journal.pone.0025216

Wang, X., Haberer, G., and Mayer, K. F. (2009). Discovery of cis-elements between sorghum and rice using co-expression and evolutionary conservation. *BMC Genomics* 10:284. doi: 10.1186/1471-2164-10-284

Wei, S., Hu, W., Deng, X., Zhang, Y., Liu, X., Zhao, X., et al. (2014). A rice calcium-dependent protein kinase OsCPK9 positively regulates drought stress tolerance and spikelet fertility. *BMC Plant Biol.* 14:133. doi: 10.1186/1471-2229-14-133

Welchen, E., and Gonzalez, D. H. (2006). Overrepresentation of elements recognized by TCP-domain transcription factors in the upstream regions of nuclear genes encoding components of the mitochondrial oxidative phosphorylation Machinery. *Plant Physiol.* 141, 540–545. doi: 10.1104/pp.105.075366

Xiong, L., and Yang, Y. (2003). Disease resistance and abiotic stress tolerance in rice are inversely modulated by an abscisic acid-inducible mitogen-activated protein kinase. *Plant Cell* 15, 745–759. doi: 10.1105/tpc.008714

Yi, X., Du, Z., and Su, Z. (2013). PlantGSEA: a gene set enrichment analysis toolkit for plant community. *Nucleic Acids Res.* 41, W98–W103. doi: 10.1093/nar/gkt281

Yook, S. H., Oltvai, Z. N., and Barabasi, A. L. (2004). Functional and topological characterization of protein interaction networks. *Proteomics* 4, 928–942. doi: 10.1002/pmic.200300636

Yu, M., Yuan, M., and Ren, H. (2006). Visualization of actin cytoskeletal dynamics during the cell cycle in tobacco (*Nicotiana tabacum* L. cv Bright Yellow) cells. *Biol. Cell* 98, 295–306.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Krishnan, Gupta, Ambavaram and Pereira. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.