frontiers in Genetics

ORIGINAL RESEARCH
published: 24 June 2021
doi: 10.3389/fgene.2021.652189

Using Network-Based Machine Learning to Predict Transcription Factors Involved in Drought Resistance

*Chirag Gupta†, Venkategowda Ramegowda†‡, Supratim Basu‡ and Andy Pereira**

Department of Crop, Soil, and Environmental Sciences, University of Arkansas, Fayetteville, AR, United States

OPEN ACCESS

Edited by:
Valentino Ruggieri, Sequentia Biotech, Spain

Reviewed by:
Hui Song, Qingdao Agricultural University, China
Raju Datla, Global Institute for Food Security (GIFS), Canada
Fan Lin, Brightseed Inc., United States
Ki-Hong Jung, Kyung Hee University, South Korea

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

†These authors have contributed equally to this work

‡Present address:
Venkategowda Ramegowda, Department of Crop Physiology, University of Agricultural Sciences, Bengaluru, India
Supratim Basu, New Mexico Consortium, Los Alamos, NM, United States

Specialty section:
This article was submitted to Plant Genomics, a section of the journal Frontiers in Genetics

Received: *11 January 2021*
Accepted: *13 May 2021*
Published: *24 June 2021*

Citation:
Gupta C, Ramegowda V, Basu S and Pereira A (2021) Using Network-Based Machine Learning to Predict Transcription Factors Involved in Drought Resistance. Front. Genet. 12:652189. doi: 10.3389/fgene.2021.652189

Gene regulatory networks underpin stress response pathways in plants. However, parsing these networks to prioritize key genes underlying a particular trait is challenging. Here, we have built the Gene Regulation and Association Network (GRAiN) of rice (*Oryza sativa*). GRAiN is an interactive query-based web-platform that allows users to study functional relationships between transcription factors (TFs) and genetic modules underlying abiotic-stress responses. We built GRAiN by applying a combination of different network inference algorithms to publicly available gene expression data. We propose a supervised machine learning framework that complements GRAiN in prioritizing genes that regulate stress signal transduction and modulate gene expression under drought conditions. Our framework converts intricate network connectivity patterns of 2160 TFs into a single drought score. We observed that TFs with the highest drought scores define the functional, structural, and evolutionary characteristics of drought resistance in rice. Our approach accurately predicted the function of OsbHLH148 TF, which we validated using *in vitro* protein-DNA binding assays and mRNA sequencing loss-of-function mutants grown under control and drought stress conditions. Our network and the complementary machine learning strategy lends itself to predicting key regulatory genes underlying other agricultural traits and will assist in the genetic engineering of desirable rice varieties.

Keywords: rice, oryza, drought, transcription factor, gene regulatory network, machine learning, abiotic stress, R shiny

INTRODUCTION

The occurrence of environmental stressors, such as extreme drought, heat, cold, and salinity, negatively regulates the growth and development of crop plants, causing a substantial loss in yield and quality (Boyer, 1982; Bray, 1997; Yamaguchi-Shinozaki and Shinozaki, 2006; Palanog et al., 2014). Plants and specific genotypes within a plant species that can withstand sub-optimal growth conditions would be identified as 'stress-tolerant' and offer examples to study the mechanisms involved in their survival and productivity in terms of yield. While conventional breeding has been the preferred method of improving stress tolerance in rice and other crops, modern genomics, and genetic engineering strategies have become an integral part of trait enhancement programs (Umezawa et al., 2006; Ashraf, 2010; Gaj et al., 2013). However, a prerequisite for the effective use of genetic engineering tools in trait improvement is the prior knowledge about candidate genes that are likely to produce a desirable phenotype when genetically intervened. Although transcriptome analysis of rice under water-limited conditions, for example, has identified thousands

of differentially expressed genes, it is difficult to narrow down the selection of candidate genes for testing function and genetic modification of drought resistance (DR). This lack of candidate genes will be a significant bottleneck in the future, as it impedes our ability to upscale targeted genetic screens in order to select leads for further crop improvement (Gutterson and Zhang, 2004; Century et al., 2008; Jansing et al., 2019; Baxter, 2020). Therefore, new data-driven approaches capable of discovering critical genes regulating complex traits like DR are needed.

Gene regulatory networks (GRN) play a central role in mediating plant responses to environmental changes (Chen and Zhu, 2004; Clauw et al., 2016; Lovell et al., 2018). Transcription Factors (TFs) are vital nodes (genes) in these networks as they regulate the expression of several downstream genes involved in many stress-responsive pathways and biological processes. TFs act as 'switches' in genetic networks and can be exploited to engineer stress-resistant crop varieties (Tran et al., 2010; Rabara et al., 2014; Krannich et al., 2015; Wang et al., 2016; Hoang et al., 2017). Such gene activity can be monitored dynamically under varying experimental conditions using genome-scale technologies such as microarrays and RNA-sequencing. Integration of such transcriptome-level datasets for inference of GRNs remains a feasible approach (Razaghi-Moghadam and Nikoloski, 2020). Transcriptome-based network inference techniques have also shown great promise in accelerating *in silico* gene discovery for *in planta* gene validation in plants (Li et al., 2015; Gupta and Pereira, 2019; Haque et al., 2019).

There are several caveats to GRN inference using expression data, which mainly stem from co-expression used as a proxy for co-regulation. A physical interaction (e.g., TF-promoter and TF-TF protein complex) cannot be guaranteed with an observed TF-gene pair that co-express. Incorporating TF-DNA binding data (e.g., ChIP-seq datasets, predictably conserved TF-DNA binding motif relationships) into the network inference workflow can overcome some of these limitations. However, careful methodological considerations can also circumvent some of these limitations. An increasing corpus of network inference algorithms aims to eliminate likely indirect interactions between TFs and other genes, i.e., correlations arising from transcriptional regulation cascades. These algorithms provide an advantage of inferring GRNs using expression data to cover those TFs for which DNA-binding sites have not been found or confirmed as yet, which remains the case for rice (Wilkins et al., 2016), and mostly all crops. We believe that removing TFs with no DNA-binding data from network inference essentially leads to the loss of regulatory signals that can be measured by analyzing expression patterns.

The outcomes of network inference considerably differ between different algorithms because they adopt different statistical assumptions and filtering schemes to detect regulatory interactions in expression patterns. Therefore, different network inference strategies have their strengths and weaknesses, making it difficult to narrow down on a single best approach (Stolovitzky et al., 2009; Marbach et al., 2010). Previously, large-scale evaluations showed that the advantages of combining predictions from different algorithms complement each other, and their limitations tend to cancel out (Michoel et al., 2009; De Smet and Marchal, 2010; Marbach et al., 2012; Hase et al., 2013). Rather than relying on only one approach, an ensemble-centric approach of combining predictions from multiple algorithms appears to be an excellent strategy to infer GRNs even in plants (Vermeirssen et al., 2014; Taylor-Teeples et al., 2015; Redekar et al., 2017; Foo et al., 2018).

Post the inference of a GRN, mining relevant signals that may lead to actionable hypotheses is not straightforward. For example, a typical network analysis workflow aims to find modules (communities of densely connected genes) in the network and assign a biological meaning to these modules using statistical enrichment of gene ontologies and pathway annotations. Biological interpretation using enrichment analysis typically require modules with a considerable number of genes for reliable overlap statistics with the already sparse and incomplete functional annotations. Therefore, modules containing many genes are readily interpreted in functional contexts, while smaller modules typically remain less interpretable.

Large modules of densely connected genes can be un-inviting for experimental biologists who wish to apply network models in the wet-lab. Biologists should have a protocol that converts complex 'hairballs' of connected genes into a single score for each gene, allowing non-subjective candidate prioritization before validation. Gene prioritization before experimental testing is vital for reducing associated costs, especially when one intends to work on more than one node in a sub-network (or module) of interest. The popular concept of 'hub' genes (genes with a relatively large number of connections in the network) is contextual (Langfelder et al., 2013; Walley et al., 2016; Vandereyken et al., 2018), as hubs in a protein coexpression network can be very different from hubs in a protein coexpression network (Walley et al., 2016). In terms of regulatory networks, studies in yeast have shown that hierarchy, rather than connectivity, better reflects regulators' importance (Bhardwaj et al., 2010). Therefore, new computational approaches beyond the estimation of 'hubbiness' or other network parameters for gene prioritization are required.

Gene prioritization is an essential technique for selecting lead candidates before experimental testing. One might assume that a simple test of differential expression can be used for gene prioritization based on the magnitude of fold change under certain experimental conditions. However, we argue that this method is not the most logical approach for gene prioritization, especially for TFs. Given their regulatory nature, subtle changes in the expression of TFs could have profound effects on the expression of downstream genes. Therefore, technically speaking, such TFs might not naturally qualify to find a position toward the top of the sorted list of genes based on fold changes.

Recently, supervised machine learning has been useful in generating predictive models for various research aspects in plant and crop biology (Ma et al., 2014; Sperschneider, 2019, 2020). Supervised machine learning algorithms leverage experimentally validated gold-standard example genes from the literature to make new predictions on genes with similar attributes. For example, thousands of genomic and evolutionary features that characterize known essential genes were used to train models predictive of other untested lethal-phenotype genes

(Lloyd et al., 2015). Similarly, several distinguishing features of genes currently annotated in secondary or primary metabolism pathways were used to train models capable of predicting new specialized metabolism genes (Moore et al., 2019). Putative *cis*-regulatory elements (CREs) involved in general abiotic and biotic stress responses (Zou et al., 2011), and CREs involved in the regulation of root cell type responses to high salinity stress (Uygun et al., 2019) have also been identified by the application of supervised machine learning models.

We are particularly interested in studies that used a genome-scale network, instead of heterogeneous genomic features, as input to the learning algorithm. Such frameworks aim to capture the network connectivity patterns that characterize a set of gold standard (or marker) genes. Network-based machine learning has been used to make reliable predictions on disease-gene associations in humans (Guan et al., 2010, 2012; Krishnan et al., 2016; Liu et al., 2019). Such predictive systems have immense potential in the development of decision systems in clinical diagnostics. However, whether this network-based supervised machine learning approach can be applied to predicting regulatory genes associated with specific agricultural biology traits remains to be tested.

In this study, we developed the Gene Regulation and Association Network (GRAiN) of rice. We built GRAiN using a collection of publicly available gene expression datasets and an ensemble of five different network prediction algorithms (**Figure 1A**). GRAiN links 2160 rice TFs to 740 modules of co-regulated genes that manifest under abiotic-stress conditions. We utilized GRAiN to develop a model predictive of TFs involved in the regulation of DR. We used a training set of TFs that are already known regulators of DR as input to a learning algorithm (support vector machine). The learning algorithm used this training data to learn general network patterns that characterize DR. We then used the trained model to identify other TFs that resemble TFs in the training set. Our strategy scored 2160 rice TFs according to their potential association with DR (**Figure 1B**). Leveraging these scores, we described the functional, evolutionary, and structural characteristics of drought regulation (**Figure 1C**). We also developed a web application to browse GRAiN[1] easily. Furthermore, we experimentally validated GRAiN's predictions on the OsbHLH148 TF using *in vitro* protein-DNA binding assays and mRNA sequencing loss-of-function mutants grown under control and drought stress conditions. Our study will provide a valuable resource for generating new testable hypotheses on the genetic basis of stress tolerance in rice.

RESULTS AND DISCUSSION

We obtained 35 independently published publicly available gene expression datasets. These datasets comprise samples from 50 different genotypes and cultivars, three developmental stages of rice growth, five different tissues, and nine different environmental stress conditions. We normalized and integrated the datasets to create a single gene expression matrix representing 35,151 rice genes' intrinsic expression in 265 individual samples. Our objective was to utilize the correlated and mutually informative expression patterns in this matrix to predict potential regulatory interactions between TFs and target genes.

The Outcome of Network Inference Varies Between Different Algorithms

Rather than using a single algorithm for the inference of the rice GRN, we created an ensemble of five diverse methods that use different edge-scoring and filtering strategies. We included Context Likelihood of Relatedness (CLR) and Algorithm for Reconstruction of Accurate Cellular Networks (ARACNe) in the first category of algorithms that use mutual information (MI) to estimate similarity in expression patterns. We included Pearson's Correlation Coefficient (PCC) and Spearman's Correlation Coefficient (SCC) as the second category's two correlation-based methods. In the third category, we used GEne Network Inference by an Ensemble of trees (GENIE3) algorithm as the regression-based method that infers edges with directionality. We then supplied each of these five algorithms with the gene expression matrix to predict regulatory interactions (edges) between TFs and target genes (see section "Materials and Methods").

We retained only the top 500,000 high confidence edges from each algorithm's outcome to reduce the computational burden in the network analysis workflow (**Supplementary Data 1**). These 500,000 edges represented less than 1% of all theoretically possible edges between TFs and their target genes in the input gene expression matrix (see section "Materials and Methods"). We asked if these high confidence edges predicted by the five algorithms are similar. We observed a minimal overlap (less than 1%) between all five algorithms' outcomes (**Figure 2**). The most considerable fraction of unique edges came from the CLR algorithm, followed by GENIE3 and PCC. We observed a relatively more extensive overlap between the algorithms in different categories than algorithms in the same category. For example, the overlap between SCC and ARACNe eclipses the overlap between SCC and PCC. This is probably because SCC and ARACNe, unlike PCC, are not constrained to detecting only linear correlations between TF and target genes. Similarly, we observed a relatively more generous overlap between GENIE3 and CLR than between CLR and ARACNe. This could be because, for filtering edges, both CLR and GENIE3 account for each gene's local distribution of background values separately. On the other hand, ARACNe examines triplets of connected genes and relies on a global threshold to eliminate the edge with the lowest score in each triplet as an indirect relationship.

Overall, our analysis suggests that the outputs of different network inference algorithms vary greatly and depend mainly on the filtering schemes used to eliminate low confidence edges.

The Performance of Network Inference Can Be Improved by Combining Networks Inferred by Multiple Algorithms

To test the performance of each algorithm in predicting known targets of TFs, we obtained experimentally identified

[1]http://rrn.uark.edu/shiny/apps/GRAiN/

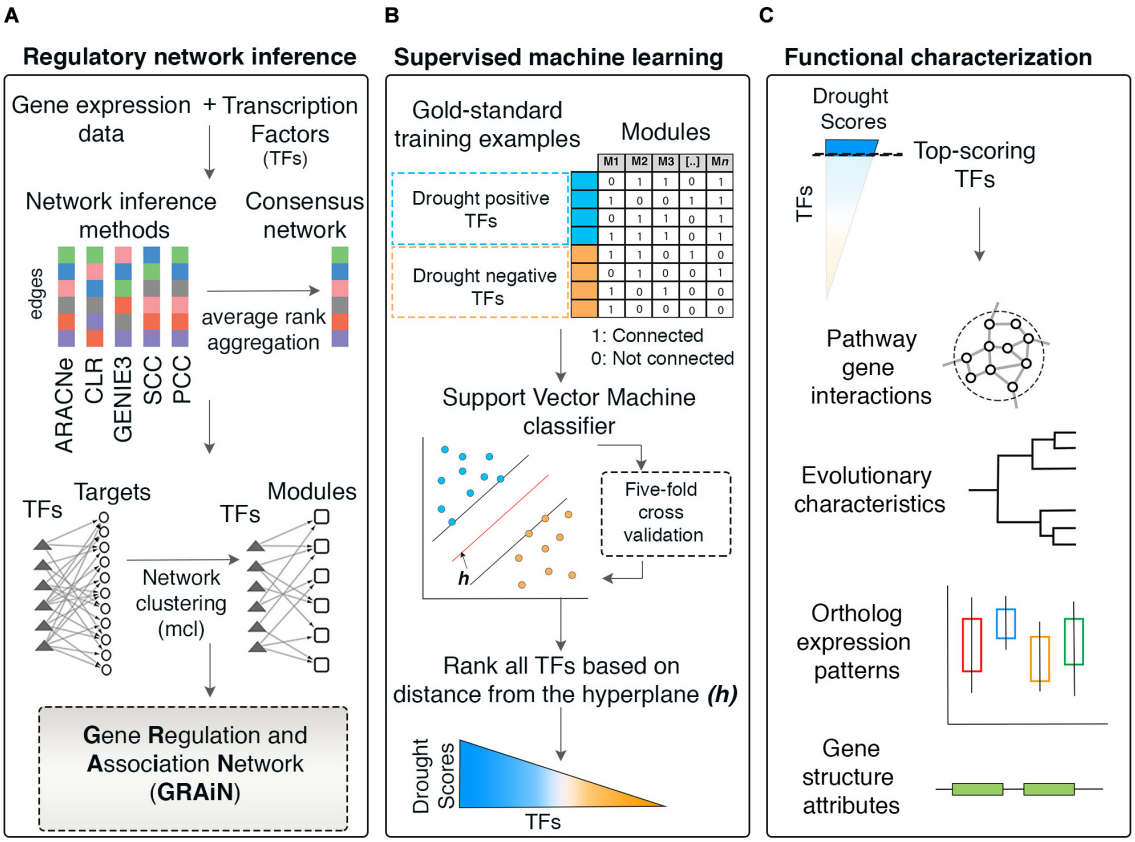

FIGURE 1 | Workflow of the network-based machine learning framework developed in this study. **(A)** A gene regulatory network (GRN) depicting regulatory relationships between transcription factors (TFs) and potential target genes was inferred from large-scale expression data (microarrays) of rice. An ensemble of network prediction algorithms was applied to the data and the networks inferred by different algorithms were statistically combined using the average rank aggregation method, resulting in a single consensus gene regulatory network. This network then was clustered using a network clustering algorithm to identify modules of coregulated, and, therefore, functionally associated genes. This modular core of the network was then reconciled with the GRN core, resulting in weighted assignments of TFs as the potential regulators of the modules. The network was named Gene Regulation and Association Network (GRAiN). mcl, Markov clustering algorithm. **(B)** Several rice knowledgebases and the literature were mined to obtain a list of TFs that are experimentally validated and reported as regulators of drought response phenotypes in rice. We found 165 such TFs reported to date. We regarded these TFs as the 'gold standard' examples of drought resistance (DR). All DR TFs were labeled as the 'drought positive' class. The group of TFs that did not differentially express in our reanalysis of several published drought experiments were labeled as the drought negative class. These benchmark drought TFs (positive and negative class), along with their network connectivity patterns in GRAiN, were used as input to train a binary classification algorithm, the support vector machine (SVM). The SVM learnt unique network patterns that can discriminate between the two classes of benchmark TFs. These patterns were fivefold cross-validated and subsequently used to predict the class label (positive or negative) of the remaining unlabeled TFs (ones that are neither in the positive nor the negative class). The final model's output was used to represent each TF in GRAiN (2160 total) along a continuous spectrum (called drought scores), representing its potential association with drought resistance. **(C)** The functional, evolutionary, and genomic features unique to most TFs at the top end of the drought score spectrum were identified and described. GRAiN and predictions on regulators of DR can be freely accessed online at http://rrn.uark.edu/shiny/apps/GRAiN/.

targets of 9 TFs in published ChIP-seq experiments (Lu et al., 2013; Tsuda et al., 2014; Zong et al., 2016; Chung et al., 2018; Li et al., 2019). Using these 9 TFs as the benchmark, we asked what fraction of their ChIP targets each algorithm could correctly predict. We observed that GENIE3 recovered ChIP targets of 8 out of the 9 TFs in the benchmark, CLR recovered targets of 6 TFs, while ARACNe, PCC, and SCC recovered targets of only 1 TF each (**Figure 3A** and **Supplementary Table 1**). To quantify each algorithm's overall performance as a single measure, we calculated the *F1* score as the harmonic mean of precision (the fraction of predicted targets that are also ChIP targets) and recall (the fraction of ChIP targets amongst all predicted targets). We observed that the CLR algorithm consistently achieved the highest *F1* score in more cases than the next best performer, GENIE3 (**Figure 3A**).

Note that the TFs used in the ChIP-seq benchmark represents only a fraction (less than 1%) of all TFs for which targets were predicted. Therefore, we could not regard the ChIP-seq dataset as a comprehensive benchmark for evaluating different network inference methods we used in our study. We built additional *ad hoc* 'reference networks' to gauge the algorithms' performance. We sought to create reference networks that reflect putative targets of TFs that can be predicted independently of

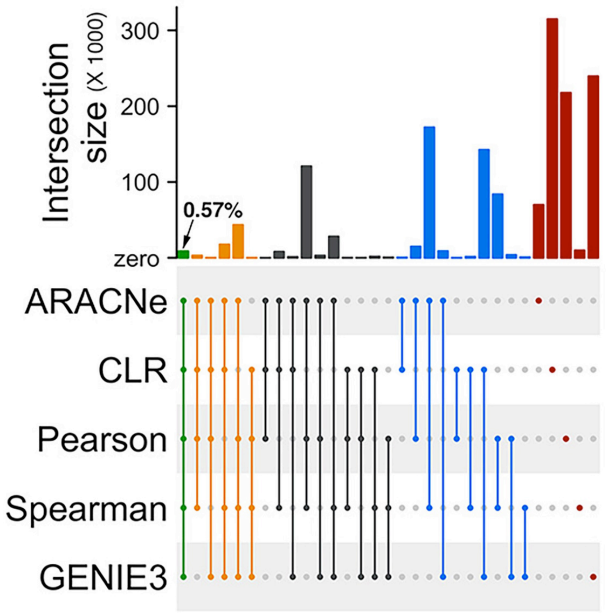

FIGURE 2 | Low overlaps between edges inferred by different network prediction algorithms. An Upset plot showing overlaps between the two mutual-information based algorithms (ARACNe and CLR), two correlation-based algorithms (PCC and SCC), and one decision-tree based algorithm (GENIE3) used for prediction the rice gene regulatory network. ARACNe, Algorithm for Reconstruction of Accurate Cellular Networks; CLR, Context Likelihood of Relatedness; PCC, Pearson's Correlation Coefficient; SCC, Spearman's Correlation Coefficient; and GENIE3, Gene Network Inference by an Ensemble of trees. The filled dots in the canter matrix indicate association between the respective sets and the bars on the top show size of the intersection. Green, orange, black, and blue bars indicate intersection size between five, four, three, and two algorithms. Red bars indicate unique edges identified by the corresponding algorithm.

their expression profiles, since we built the network using only expression data.

For the first reference network, we obtained experimentally verified protein interaction partners from the protein interaction network of rice (PRIN) database (Gu et al., 2011). For the second reference network, we used promoters of genes with known DNA-binding sites of TFs by analyzing the CIS-BP database (Weirauch et al., 2014). We created the third reference network by linking TFs and non-TFs if they are co-annotated in carefully selected, non-redundant Gene Ontology (GO) Biological Process (BP) terms. For the GO BP reference network, we assumed the TF and non-TF genes co-annotated to the same BP terms are more likely to have a biological relationship, relative to genes annotated to distant or unrelated GO BP categories. Although the second and the third reference networks do not guarantee real biological relationships between TFs and target genes, they provided us with a valuable resource to include more TFs in the evaluation and gauge the agreement between different data types in predicting targets of TFs.

We asked what fraction of edges present in the three independent reference networks could be predicted by each algorithm in our ensemble. We observed that the CLR algorithm consistently attained the best recall rate in all three reference networks and outperformed other methods in reconstructing the CIS-BP and GO BP reference networks in terms of the *F1* score (**Supplementary Table 1**). The ARACNe algorithm performed with the best precision in the CIS-BP reference network and outperformed others in reconstructing the protein interaction network in terms of the *F1* score. SCC's precision in predicting CIS-BP edges was lower than that of ARACNe but better than the other three methods. We observed that PCC outperformed SCC in CIS-BP and GO BP reference networks (**Figure 3B**).

Based on these evaluations, we could not establish any single algorithm as the best performer in reconstructing the three reference networks or correctly predicting ChIP targets. Therefore, we asked whether combining the networks inferred by different algorithms into a single network improves the overall performance. Rather than taking a union or an intersection, we chose the 'average rank aggregation' approach to combine different networks (Marbach et al., 2012). The underlying idea behind the average rank aggregation approach is that a biologically meaningful edge tends to occur consistently at high ranks (or confidence scores) across different networks predicted using different approaches. Hence, averaging the ranks of individual edges essentially reinforces likely real edges in the final aggregate network. This approach has been previously shown to efficiently integrate different edge-weighted GRNs into a single consensus network, even in plants (Vermeirssen et al., 2014).

Following the average rank aggregation method, we combined the networks inferred by all five algorithms in our ensemble into a single GRN (see section "Materials and Methods"). Then, we asked whether this aggregation improved the accuracy by re-evaluating the ChIP-seq benchmark and the three reference networks described above. We observed that the aggregate network could not outperform CLR and GENIE3 in most cases in the Chip-seq set but consistently outperformed ARACNe, SCC, and PCC (**Figure 3A**). Interestingly, removing the two correlation-based methods from the aggregate almost always improved the performance, compared to the aggregate that included the correlation-based methods (**Figures 3A,B**). The aggregate-sans-correlation network achieved, on average, 49 and 20% increase in *F1* score when tested on the ChIP-seq benchmark and the three reference networks, respectively, relative to the aggregate that included PCC and SCC. Therefore, the aggregate of CLR, ARACNe, and GENIE3 was chosen as the final 'consensus' GRN of rice and used in further analysis.

Clustering the GRN Identifies Modules of Functionally Related and Co-regulated Genes

Our next objective was to find clusters of co-regulated genes, i.e., groups of genes regulated by the same set of TFs. Assuming a guilt-by-association, we expected network clustering to identify modules of co-regulated, and therefore functionally related genes. Such modules thereby provide pointers on pathways and biological processes that could be under the regulatory control of specific TFs (Hartwell et al., 1999; Segal et al., 2003; Ma et al., 2004; Joshi et al., 2009). To achieve such a network clustering,

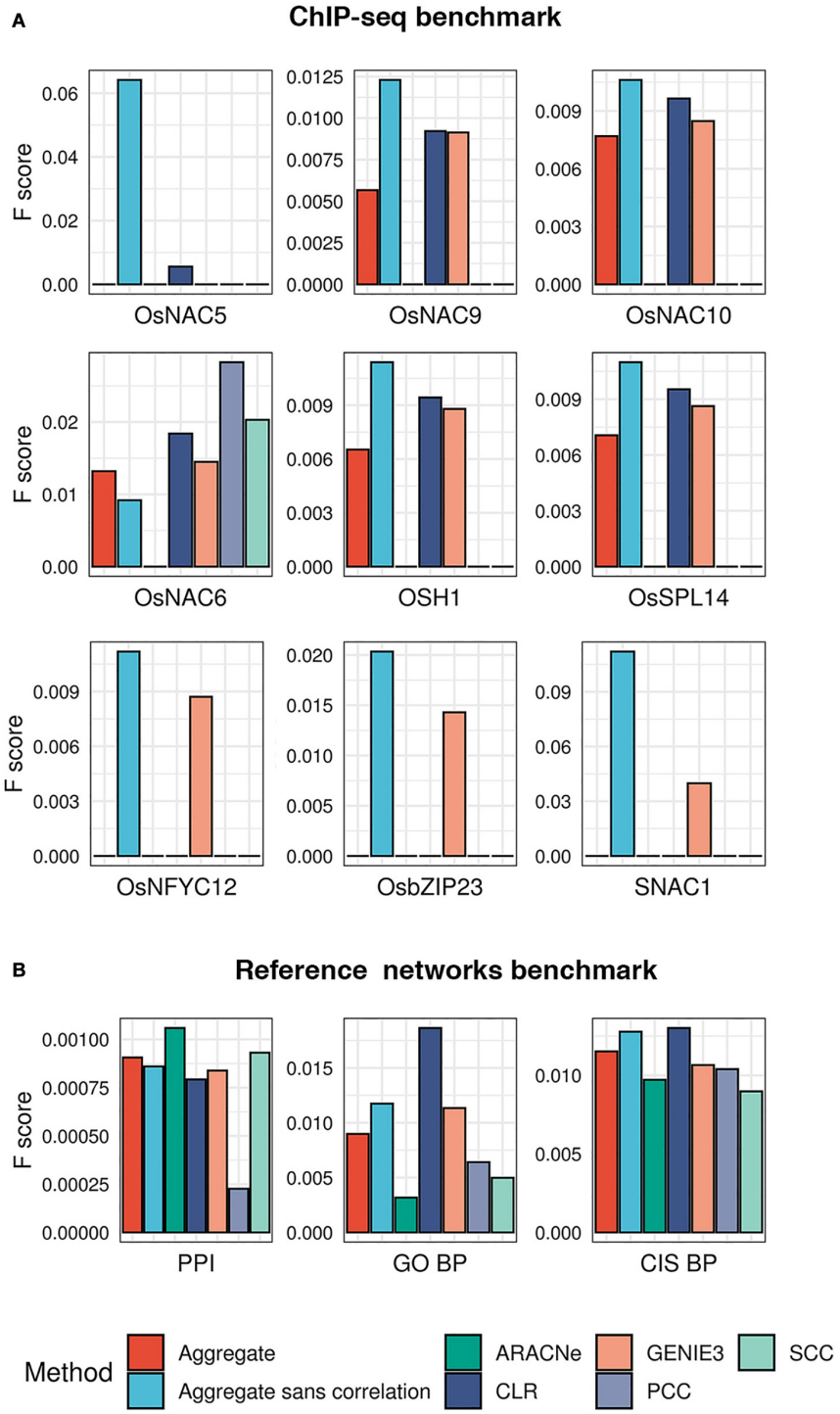

FIGURE 3 | Evaluation of the five network prediction algorithms and their aggregate. **(A)** A ChIP-seq benchmark for 9 TFs was created from publicly available datasets. For each of these 9 TFs, we checked the overlap between experimentally validated targets (ChIP-bound genes) and network-predicted targets (genes predicted by each of the five algorithms in our ensemble). This evaluation was also made for the consensus network obtained by statistically aggregating the predictions from the five algorithms. Each bar plot shows *F1* scores (*y* axis; a measure of performance, the higher the better) of each algorithm (*x* axis) in correctly predicting ChIP-targets of TFs. **(B)** Due to the unavailability of experimentally validated targets of a large number of TFs in our network, we created additional 'reference networks' to gauge the quality of the inferred networks. PPI, reference network derived from the predicted protein–protein interaction network of rice (PRIN database); GO BP, reference network derived from co-annotations in select gene ontology biological process terms; and CIS BP, reference network obtained by utilizing the available putative DNA-binding sites of TFs in the CIS BP database (see "Materials and Methods" for details). The bar plots of *F1* scores shows the performance of each algorithm in reconstructing the reference networks.

we first linked target genes that had high overlaps between their predicted regulators in the consensus GRN, as done previously with the Arabidopsis stress GRN (Vermeirssen et al., 2014). We then applied the Markov clustering algorithm to this co-regulated gene network (van Dongen and Abreu-Goodger, 2012). We identified a total of 740 modules, with an average of 45 genes each (**Supplementary Data 2**).

To confirm the regulatory association of genes within each module, we analyzed their 1000 bp upstream promoter regions to check whether putative DNA-binding sites were over-represented. We employed the FIRE (finding informative regulatory elements) algorithm (Elemento et al., 2007). FIRE uses a *de novo* approach to find short stretches of DNA-sequence motifs that explain promoters' module-membership (Elemento et al., 2007). Application of the FIRE algorithm on our network data detected 84 DNA-binding motifs within the co-regulated modules. We observed that more than 50% of all coregulated modules harbor between five and ten motifs each (**Supplementary Figure 1A**), indicating a high level of coordination between TFs. We observed that eighty-one of the FIRE-detected motifs are identical to known plant CREs listed in multiple plant databases and other sources, whereas three are novel DNA motifs (**Supplementary Figure 1B** and **Supplementary Data 3**). Network analysis of the genes with the three novel motifs suggests that two distinct groups of TF families target them (**Supplementary Figure 2**). Overall, the over-representation of common plant CREs in the promoters of module genes testified that the observed modules are non-random gene groupings and represent sets of co-regulated genes.

To further test whether the observed gene modules also represent a joint biological function, we used function annotations from the rice GO BP category and pathway-level annotations from Mapman, KEGG, and CYC databases. We found statistically significant associations of these functional annotations in 31% of all observed modules (hypergeometric test FDR corrected *p-values* < 0.05). We also found that ∼41% of all modules we detected in this study were preserved in an independent coexpression network we built earlier with a different dataset and the cluster detection algorithm (Krishnan et al., 2017). Interestingly, 22% of these preserved modules are the ones that could not be annotated by gleaning function annotation databases, highlighting significant gaps still exist in the current state of function annotations of rice genes (**Supplementary Data 4**).

We linked TFs to the co-regulated modules, and set the edge-weight according to the Jaccard's Index (JI) of overlap between the predicted targets of TFs in the consensus GRN and module genes. The JI ranges between 0 and 1, where 0 indicates no regulatory association between the corresponding TF-module pair and a JI of 1 indicates a certainly likely regulatory association. Therefore, these operations generated a modular GRN of rice, where TFs are directly linked to target genes and indirectly but quantifiably associated with functional processes. We refer to this network as GRAiN. GRAiN can be searched through an online portal (demonstrated in the last section of this manuscript).

Sorting Genes Based on the Magnitude of Differential Expression Is Not a Suitable Approach for Gene Prioritization

Past genetic research in rice has revealed several examples of 'gold-standard' drought genes identified by reverse-genetics. This documented knowledge about the genetic basis of drought is most comprehensive among other abiotic-stresses. It presents us with a unique opportunity to test whether differential expression measures can be used as a proxy for gene prioritization. We scanned the functional rice gene database (Yao et al., 2018), the rice mutant database (Zhang et al., 2006), and the Oryzabase (Kurata and Yamazaki, 2006). We found 732 genes with genetic evidence of association with drought listed in these knowledgebases. This list of 'drought associated' genes did not represent any particular physiological, morphological, or biochemical phenotype typically measured in the analysis of drought response. For the sake of convention, we use 'DR' as a broad term to encapsulate various molecular mechanisms by which plants adapt, escape, or otherwise tolerate water limiting conditions (Levitt, 1980; Basu et al., 2016). There are currently 165 DR TFs in this list of known drought genes. We regarded these 165 TFs as the gold-standard examples of DR, and refer to them as DR TFs.

Next, we asked how the DR TFs respond to drought in terms of differential expression. We re-analyzed data from seven independently published drought experiments performed on multiple varieties, growth stages, and tissues of rice plants (Wang et al., 2011; Ding et al., 2013; Pabuayon et al., 2016; Mishra et al., 2018). We estimated genome-wide fold change values in each of these datasets, and ranked all TFs based on the absolute values of these fold changes. We observed that, in each experiment, the majority of the DR TFs showed minimal changes (fold change values < 1) regardless of the tissue, growth stage, or the variety of rice plant (**Figures 4A,B**). This suggests that gene prioritization based on differential expression values is constrained by experimental factors and will downplay those that show subtle changes in expression but have relevant biological effects. Therefore, we need a more sophisticated technique for prioritizing TFs that likely regulate drought responses in rice.

Network-Based Supervised Machine Learning Enables Classification and Scoring of Drought Resistance Regulators

Our next objective was to develop a network-based gene prioritization framework that can be objectively tested using independent data. We utilized the two good pieces of information at hand; a high-quality modular GRN (GRAiN) and a list of literature curated gold-standard DR TFs. We posited that advanced machine learning models could be trained to recognize network patterns in GRAiN that characterize the gold standard DR TFs. These patterns could then be matched with the patterns of other yet untested TFs and estimate whether they resemble the DR TFs.

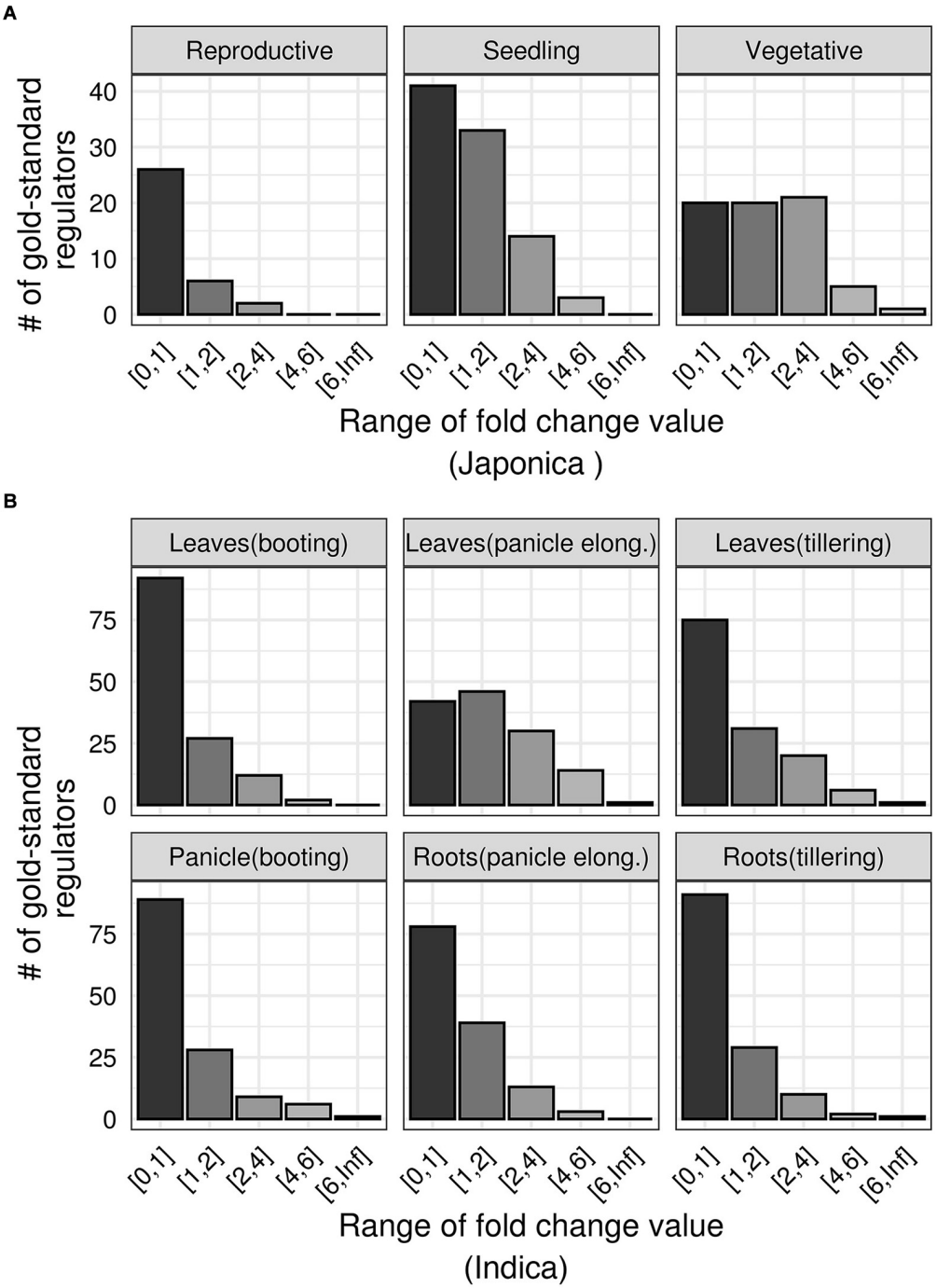

FIGURE 4 | Differential expression patterns of gold standard drought regulators. **(A)** The range of absolute fold-change values (*x-axis*) of gold standard TFs (*y-axis*) in three growth stages of *Japonica* rice variety exposed to drought (data from GSE81253). **(B)** The range of absolute fold-change values (*x-axis*) of gold standard TFs (*y-axis*) in multiple tissues of *indica* rice variety exposed to drought (data from GSE26280).

We chose the support vector machine (SVM), a popular binary classification algorithm (Cortes and Vapnik, 1995), to develop the DR classifier. We supplied the SVM with a training set of TFs and their connectivity patterns in GRAiN, along with binary labels indicating whether each TF is a DR TF or not (see section "Materials and Methods"). We evaluated the SVM's accuracy using fivefold cross-validation tests and the area under the precision-recall curve (AUC-PR) statistics. The AUC-PR ranges between 0 and 1, with values closer to 1 indicating the model's superior performance. Our DR classifier achieved

an average AUC-PR of 0.81 in 10 independent runs of five-fold cross-validation tests. We asked if this AUC-PR could be achieved by randomly picking TFs from the rice genome instead of using the DR TFs for training. We found the DR classifier's AUC-PR to be significantly larger than the AUC-PR of the classifier trained using randomly picked TFs. Because family membership could play an essential role in TF function, we also tested the AUC-PR of the classifier trained by randomly picking TFs while maintaining the family distribution as that of the DR TFs. We observed that the AUC-PR of this classifier was not different than the random classifier, indicating that family memberships of TFs is not indicative of their roles under drought (**Figure 5A**).

We applied the cross-validated SVM model to the whole network of 2160 TFs. We used the model's output – which represented the model's confidence in its classification of a TF as a DR TF – to rank each TF. We then scaled the ranks within a range of 0 and 1 to make the ranks more interpretable, and referred to the resulting scores as drought scores (DS). The TFs with DS close to 1 have the strongest predicted association with DR, while TFs with relatively smaller DS values are less likely to be associated with DR (**Supplementary Data 5**).

To evaluate this scoring scheme objectively, we asked if the occurrence of drought can be inferred by the transcript abundance of TFs with the largest DS. In other words, we wanted to check if the intrinsic expression levels of TFs with high DS can indicate if a plant has sensed drought or not. Operationally, this technique is similar to the ones used in developing clinical diagnostic models that seek to classify human patient samples as disease or healthy based on the expression levels of marker genes. To perform such an evaluation of our model, we downloaded and reanalyzed the recently published RNA-seq dataset of 214 seedling samples (71 drought samples and 143 control samples) from four different rice varieties (Wilkins et al., 2016). Assuming the first decile TFs in our predictions as 'drought markers', we asked whether the intrinsic expression levels (measured as transcripts per million units) of these drought markers can predict a sample in the Wilkins dataset as control or drought.

We observed that RNA-seq sample classification accuracy was almost perfect when we used the intrinsic expression of top decile TFs as features. However, this accuracy gradually decreased as we moved toward lower decile TFs (**Figure 5B**). We observed that the expression levels of TFs in the last decile was least accurate in classifying a sample as drought or control (**Figure 5B**). This analysis suggests that the top-scoring TFs are likely responsible for causing the transcriptional-level changes that occur under drought, and therefore validates our ranking approach.

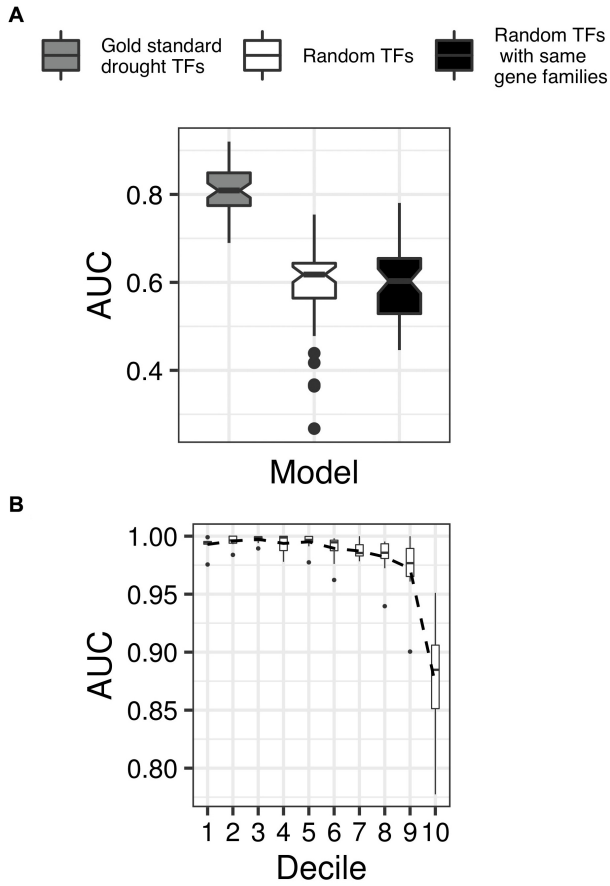

FIGURE 5 | Cross-validation of the network-based classifier. **(A)** Boxplots showing the distribution of the area under the precision-recall curve (AUC-PR; *y-axis*) in ten independent runs of fivefold cross-validation tests of the classifier trained using gold standard drought TFs (shaded gray), the classifier trained using randomly picked TFs instead of gold standard TFs (shaded white), and the classifier trained using randomly chosen TFs but from the same families like that of the gold standard examples (shaded black). The non-overlapping notches in the boxplots indicate significant differences in the median AUC-PR for all three classifiers. **(B)** TFs were sorted according to their decreasing order of drought scores assigned by the final classifier and grouped into 100 equal-sized bins. Expression levels (transcript per million units) of TFs in each bin were then used as features to classify a set of labeled RNA-seq samples as drought or control (data from GSE74793). Each boxplot shows the distribution of AUC-ROC *(x-axis)* from threefold cross-validation tests in groups of ten bins, with lower-numbered bins (*y-axis*) indicating TFs with higher drought scores. The black dotted line connects the mean of each decile's AUC-ROC scores, indicating decreasing AUC-ROC with lower drought scores.

Predicted Regulators of Drought Resistance Are Involved in Hormone-Mediated Responses

It is important to note that the DS we predicted and the out-degree of TFs in the network are not correlated (**Supplementary Figure 3**), indicating that the predicted DS do not merely reflect on the 'hubbiness' of TFs. We investigated the few modules (features) that served as the best predictors for the classification of DR TFs in our model. We selected top 'drought modules' using the 'feature importance' scores from the model output (**Supplementary Data 6**). We extracted all TF and CREs linked with these drought modules and explored the interconnected network in Cytoscape (Shannon et al., 2003; **Supplementary Figure 4**).

Exploring this network, we found that the drought modules comprise a total of 6968 genes that form core communities enriched in several stress response pathways and biological

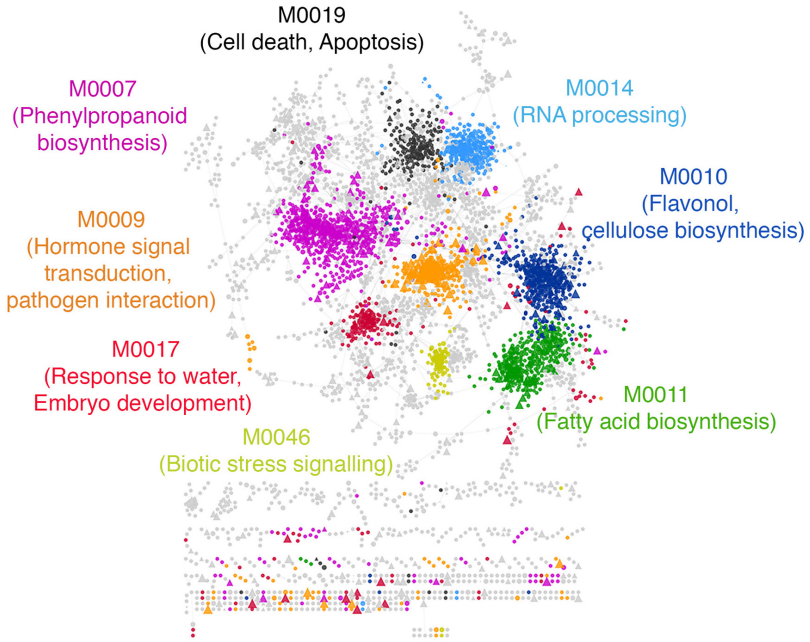

FIGURE 6 | Functional characterization of predicted drought resistance transcription factors. A subset of modules with the highest feature importance scores from the drought classifier were extracted and labeled as 'drought modules.' The drought modules consist of a total of ~6000 genes. The network shows the top 5% edges induced between them. Every circle is a functional gene, and triangles are TFs. Genes within a module are similarly colored, and the GO BP enriched within each module is labeled with the same color in the text. Modules with no statistically enriched GO BP terms are colored gray.

processes (**Figure 6**). Interestingly, we found that the drought module are enriched with genes annotated to secondary metabolism pathways broadly related to hormonal signal transduction, such as phenylpropanoid biosynthesis and jasmonic acid biosynthesis. These are traits specific to land plants and is believed to have played an essential role in the adaption of plants to water limiting environments (Kenrick and Crane, 1997; Emiliani et al., 2009; Wang et al., 2015; Ahammed et al., 2016; Verma et al., 2016), given its role in lignin biosynthesis (Fraser and Chapple, 2011). Other relevant GO biological process terms such as 'response to water,' 'response to abscisic acid stimulus,' 'cellulose biosynthesis,' 'flavonol biosynthesis,' and 'trehalose biosynthesis' were also recovered within the drought modules.

We found that the most prominent *de novo* predicted CREs within the drought modules are related to the abscisic acid response complex ABRE3HVA22 (Shen et al., 1996) and the vascular-specific motif ACIIPVPAL2 (Hatton et al., 1995), along with the light-responsive GT-1 motif (Lam and Chua, 1990), the anaerobic-responsive motif GCBP2ZMGAPC4 (Geffers et al., 2000) and the dehydration responsive DREB1A motif (Maruyama et al., 2004; **Supplementary Data 3**).

Predicted Drought Scores Are Associated With Evolutionary Features

The enrichment of genes related to the abscisic acid and salicylic acid pathways, along with jasmonate signaling pathways, as well as some of the observed CREs (e.g., vascular-specific ACIIPVPAL2) within the drought modules indicated a drought response machinery in rice ubiquitous and specific to land plants (Wang et al., 2015). Therefore, we pursued this lead and examined if the orthologs of rice TFs with high DS in our study have conserved responses to drought exposure.

We created three sets of Arabidopsis drought TFs with known orthologs in rice. The first set was differentially expressed TFs in response to mild and severe drought stress we reported previously (Harb et al., 2010). The second set comprised experimentally verified drought TFs in the Arabidopsis phenotype database (Lloyd and Meinke, 2012). The third list of TFs was previously predicted to be involved in mild drought responses (Clauw et al., 2016). We asked if the rice orthologs of these three sets of Arabidopsis TFs have higher DS than the background of all remaining TFs that did not become a part of the three sets (either due to biological variability or due to lack of ortholog identity). In all three sets, we observed a significantly larger mean DS of orthologous TFs compared to the background (**Figure 7A**). Similarly, we observed that rice TFs with orthologs that differentially expressed in response to the application of drought stress in cobs and leaves of maize (Kakumanu et al., 2012), leaves of barley (Cantalapiedra et al., 2017), and leaves of sorghum, have significantly larger mean DS than the mean DS of the background (**Figure 7B**).

We also asked if the predicted DS and evolutionary age of a TF are related. We first ordered all rice genes in 13 age groups (phylostrata) starting from the oldest (i.e., genes conserved across all cellular life) to the youngest (i.e., genes that appeared in the terminal clade *Oryza*) (Wang et al., 2018). Plotting the distribution of DS of TFs within each phylostrata (PS) showed

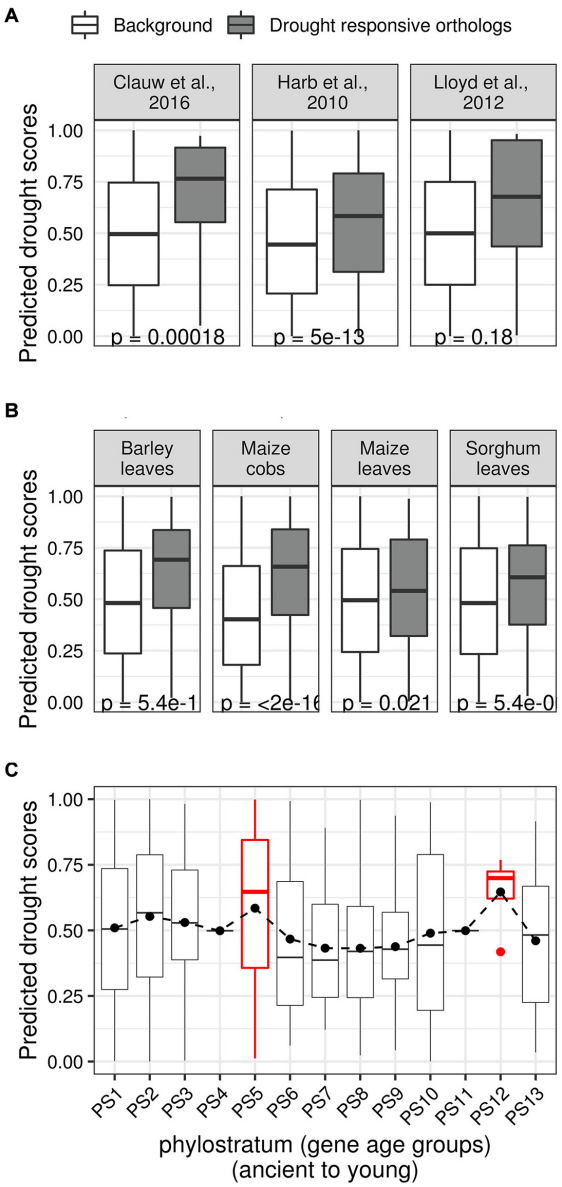

FIGURE 7 | Relationships between predicted drought scores and evolutionary features. **(A)** Three sets of putative drought regulators in Arabidopsis were curated from the literature, and their rice orthologs were identified. The three sets represent rice TFs with orthologs in Arabidopsis genes that were predicted as drought regulators (Clauw et al., 2016), that differentially expressed upon different drought treatment regimes (Harb et al., 2010), have been experimentally characterized as drought regulators (Lloyd and Meinke, 2012). The boxplots show the distribution of the predicted drought scores of these ortholog sets (gray) along with the drought scores of the background of remaining TFs (white; rice TFs that did not become part of the three ortholog sets). In each case, the median predicted drought scores of orthologous rice TFs was found to be significantly higher than the drought scores of the background. **(B)** Similarly, boxplots showing the distribution of drought scores of rice TFs with orthologs in genes that are differentially expressed in different crop datasets. **(C)** Box plot showing the distribution of drought scores in different age groups (ancient to young) according to NCBI taxonomic classification. The distribution of drought scores stays relatively flat, except for two peaks that correspond to the Embryophytes clade (PS5) and the Oryza clade (PS12).

two prominent peaks. The first peak in PS5, which corresponds with the Embryophytes (land plants) clade, and the second peak in PS12, which coincides with the Oryza clade, both mirror significant events in the evolutionary history of rice (**Figure 7C**). We also examined the available pan-genome of rice (Sun et al., 2017) to investigate the distribution of DS of TFs that arose in the terminal clade (*O. sativa*, closely related rice varieties). However, we did not find any significant differences in DS between core and distributed TFs, or TFs that are *Indica*- or *Japonica*-dominant (**Supplementary Figure 5**).

Overall, our analysis suggests that a large fraction of high-scoring DR TFs possibly played a crucial role in driving critical adaptations of land plants. A few high-scoring TFs that emerged specifically in rice might be involved in recent morphological adaptions that contribute to DR (e.g., panicle architecture, pollen and seed development). Therefore, it would be interesting to analyze the drought phenotypes of mutants lacking these high-scoring TFs specific to rice.

Predicted Drought Scores and Structural Characteristics Are Related

Recent studies in rice and other organisms suggest that younger genes have relatively simple exon/intron and protein structure (Neme and Tautz, 2013; Cui et al., 2015; Wang et al., 2018). Other studies have shown that simple genes, for example, those that lack introns, are rapidly regulated (Jeffares et al., 2008; Speth et al., 2018). Such genes represent an essential component of the possibly conserved stress response machinery in land plants (Jeffares et al., 2008; Zhu et al., 2016; Morozov and Solovyev, 2019).

Following this lead, we next investigated if the predicted DS of TFs and their structural attributes are related since a large fraction of high-scoring TFs our analysis also appear to have first emerged in land plants. We started with examining the family memberships of TFs. We found a statistically significant enrichment of WRKY, Tify, NAC, MYB, and AP2/ERF families among the top 10% TFs with the largest DS (Top decile; FDR corrected hypergeometric test p values < 0.1) (**Figure 8A**). These gene families are well-known to associate with drought stress in multiple crops (Yu et al., 2012; Gahlaut et al., 2016; Hoang et al., 2017). In contrast, we found that TFs with the smallest DS (bottom decile) are enriched in growth and development associated gene families such as the MADS, FAR1, and TRAF (Smaczniak et al., 2012; Tedeschi et al., 2017; Ma and Li, 2018). We observed the top decile TFs (top 10% TFs with highest DS) have relatively fewer InterPro protein domain annotations than the background of all TFs in the remaining deciles (remaining 90% TFs with relatively smaller DS) (**Figure 8B**). We also observed that the top decile TFs have significantly smaller average gene length, average CDS length, and average intron length (**Figure 8C**) compared to the background of remaining deciles. This indicates that TFs with high DS are small genes with simple structures.

Overall, our analysis suggests that TFs that likely regulate DR in rice have peculiar functional, structural, and evolutionary characteristics. It is interesting to see such grouping in our data,

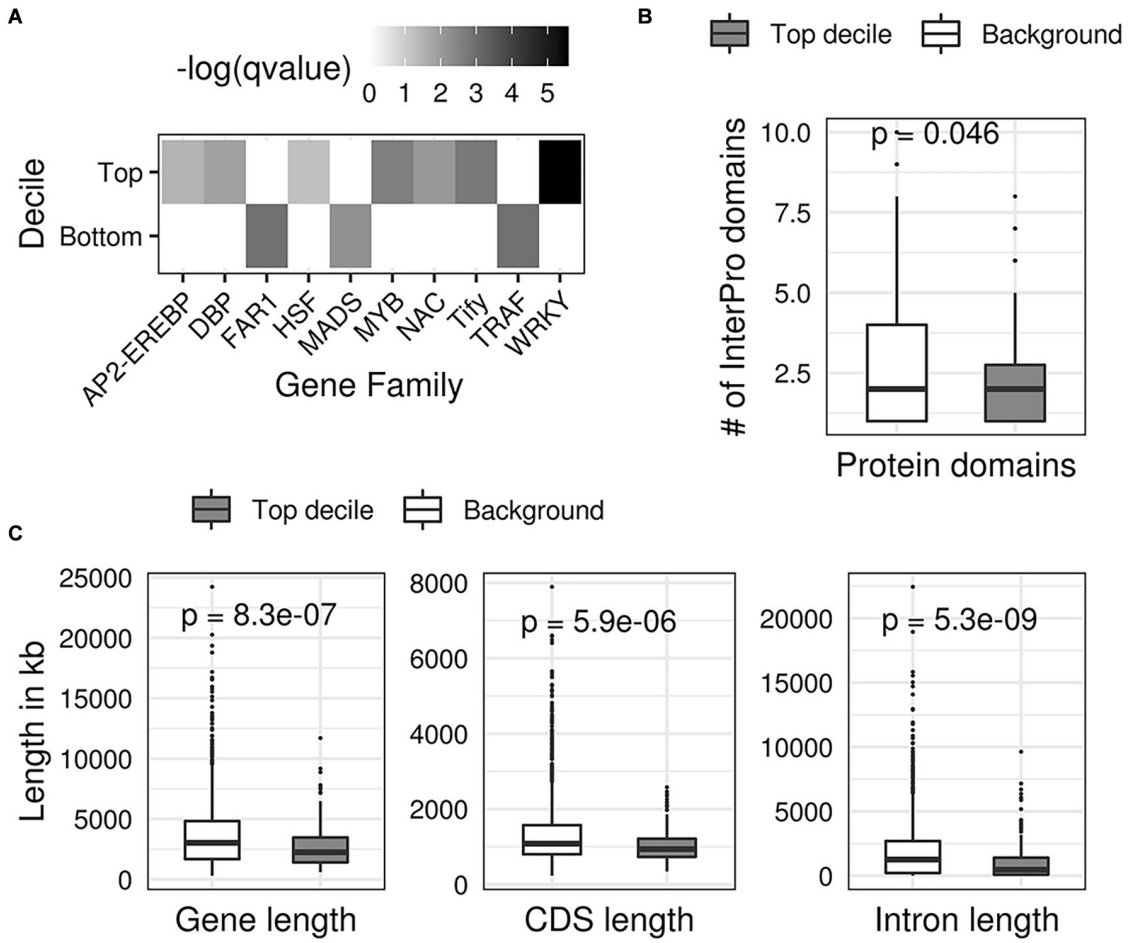

FIGURE 8 | Structural features of transcription factors with the highest drought scores. **(A)** A heatmap showing the enrichment of gene families in the first and last decile TFs (top 10% and bottom 10% drought scores). Each grid in the heatmap shows the FDR corrected -log (*p*-value, Fisher's exact test) of the gene family on the *x-axis* for the decile on the *y* axis. **(B)** Boxplots showing that top decile TFs contain a significantly different number of protein domains compared to the background of TFs in the remaining deciles. **(C)** Boxplots showing that the top decile TFs have significantly smaller average gene length, coding sequence length (CDS), and intron length compared with the background of remaining TFs.

given that the underlying network using which we made our predictions started with unclassified gene expression data.

The GRAiN Web Application Is for Experimental Rice Biologists; Using OsbHLH148 as an Example

We used the R Shiny framework to develop a user-friendly web application that allows users to interact with GRAiN and predictions on DR TFs. There are currently two main features of the GRAiN web application active at http://rrn.uark.edu/shiny/apps/GRAiN/. It allows users to search for a single TF gene of interest. In this case, the GRAiN algorithm first retrieves all the genes predicted as targets of the query TF and then uses the inbuilt enrichment analysis tool to find pathways and biological processes over-represented in the predicted targets. The second feature of the GRAiN application allows users to query a set of genes instead of a single TF. In this case, the enrichment analysis tool is used to find co-regulated modules (defined in this study) over-represented in the query genes. Significantly enriched modules are presented back to the user, along with functional (GO BP and Mapman annotations) and cis-regulatory annotations (FIRE-identified CREs and weighted links to TFs).

We chose the rice transcription factor OsbHLH148 (LOC_Os03g53020) to demonstrate the GRAiN web application features. OsbHLH148 was initially present in our list of gold standard drought regulators, as it was earlier reported to be involved in the regulation of drought response via the jasmonic acid pathway (Seo et al., 2011). However, instead of using it as a DR TF in the training set, we kept it a hidden example and treated it as an unlabeled TF throughout model training and evaluation. Since OsbHLH148 was already being studied in our laboratory, our intention behind removing it from the training data was to repurpose its phenotypic and RNA-seq data for experimental validations of the GRAiN web application and the DR classifier.

Our model strongly predicted the association of OsbHLH148 to DR, assigning it a DS of 0.99 and placing it at rank # 4 among all rice TFs. We asked if the GRAiN web application can recover the known functional associations of OsbHLH148. The GRAiN query shows 385 genes predicted as targets of OsbHLH148 (**Supplementary Data 7a**). Enrichment results show that these predicted target genes participate in the jasmonic acid-mediated signaling pathway and response to salt and osmotic stresses (**Supplementary Data 7b**), in agreement with its previously validated function by Seo et al., 2011. We observed that OsbHLH148 is potentially involved in the regulation of ∼54% of genes in the module it is a part of (M0009; jasmonic-acid biosynthesis genes), indicating it acts as a hub in the local subnetwork. Additionally, we found 81 TFs among the predicted targets of OsbHLH148 and ∼82% of these TFs have more than one known bHLH binding site (5′-CANNTG-3′) within the 1000 bp upstream promoters (**Supplementary Data 7c**). This indicated that most predicted targets of OsbHLH148 are more likely to be downstream targets. Other TFs with no bHLH DNA-binding sites could be potentially be components of a larger co-activator complex. Among the predicted targets, we found three of the five TFs previously shown to interact with OsbHLH148 using Y2H assays (Seo et al., 2011). Among other predicted targets, OsRAP2.6 (also known as ERF101) and DREB1B TFs were most interesting because both these TFs are well-known to be critical regulators of stress responses in rice. DREB1B is a well-known TF previously shown to function in abiotic stress-responsive gene expression (Dubouzet et al., 2003). The OsRAP2.6 TF has been recently shown to regulate drought responses during rice's reproductive development (Jin et al., 2018).

Therefore, the prediction of OsRAP2.6 within the OsbHLH148 regulation raised a hypothesis that OsbHLH148 could also act as a regulator of drought responses during rice's reproductive development. We acquired the homozygous loss-of-function knockout mutant line designated as 'bhlh148' to pursue this hypothesis. We performed extensive testing of this mutant's phenotypes under controlled drought stress at the vegetative and reproductive stages. Under a well-watered (WW) condition, we found no significant phenotypic difference between the mutant and WT plants. However, under controlled drought stress treatment at 40% field capacity (FC), the mutant plants showed higher sensitivity with leaves rolled and collapsed than the WT plants (**Figure 9A**). Under drought, the *bhlh148* mutant plants showed a significant reduction in net photosynthetic rate, instantaneous water use efficiency (WUEi), the efficiency of Photosystem II measured in light-adapted leaves (Fv'/Fm'), above-ground biomass, and the relative water content (RWC) compared to WT (**Figures 9B–F**).

We applied drought stress to WT plants at the reproductive (R3) stage and observed a very high induction of OsbHLH148 in the inflorescence (3.1-fold) compared to flag leaf (0.55-fold) under drought stress relative to WW plants. The yield parameters for drought stress response, quantified by the number of spikelets/panicles (**Figure 9G**), spikelet sterility (**Figure 9H**), grain yield (**Figure 9I**), and the number of panicle/plant (**Figure 9J**), testify that OsbHLH148 is involved in grain yield under drought stress.

Next, we tested whether GRAiN correctly predicted the interaction of OsbHLH148 with OsRAP2.6 and DREB1B TFs. An electrophoretic mobility shift assay (EMSA) confirmed that bHLH148 binds to the promoters of OsRAP2.6 (LOC_Os08g36920) and OsDREB1B (LOC_Os09g35010) genes. We then used the steroid receptor-based inducible system to confirm that OsbHLH148 directly activates the expression of OsRAP2.6, while activation of OsDREB1B by OsbHLH148 requires additional factors (**Figures 9K,L**).

We also wanted to check if the other remaining genes predicted by GRAiN as targets of OsbHLH148 are correct. To confirm this, we performed gene expression profiling of bhlh148 and WT plants under WW and controlled drought stress conditions using mRNA sequencing (see **Supplementary Methods**). We used leaf tissue from plants maintained at 100 and 40% FC for 10 days as WW and controlled drought stress samples, respectively. We estimated the differential expression of genes that (1) responded to the knockout, and (2) responded specifically to the interaction of mutant with drought (subtracting the WT effect of drought from the mutant) (**Supplementary Data 8**). We found a relatively low overlap (<2%) between GRAiN predicted targets of OsbHLH148 and those significantly differentially expressed in the knockout. However, more than 32% of GRAiN predicted genes differentially expressed specifically due to bhlh148's interaction with drought (**Figure 9M**). These observations testify that GRAiN naturally captures regulatory relationships that manifest specifically under stress rather than normal growth conditions.

Experimental Support of the Predicted Drought Scores in the Literature

We re-scanned the literature to collect new TFs reported to be involved in DR phenotypes but published after we concluded our study. We found three such new DR TFs not included in our training data; OsHSFA3 (Zhu et al., 2020), OsMYB6 (Tang et al., 2019), and ONAC66 (Yuan et al., 2019). Our DR prediction model placed OsHSFA3 at rank 96 (decile 1), OsMYB6 at rank 376 (decile 2), and ONAC66 at rank 615 (decile 3), indicating that our model performs with great accuracy in the real-world.

CONCLUSION

We integrated publicly available gene expression data of rice to infer an abiotic-stress response GRN. Because we used only microarray samples to create the gene expression matrix, our workflow was primed to be missing a considerable fraction of known rice genes not represented on the Affymetrix chip. This limitation could have been overcome by using RNA-seq datasets to assay a larger fraction of the genome. However, the number of RNA-seq datasets currently available to cover the broad spectrum of rice's abiotic stress responses is limited. Using microarray chips allowed us to achieve a relatively larger sample-size while covering gene expression dynamics under various abiotic-stress treatments, growth stages, and cultivars.

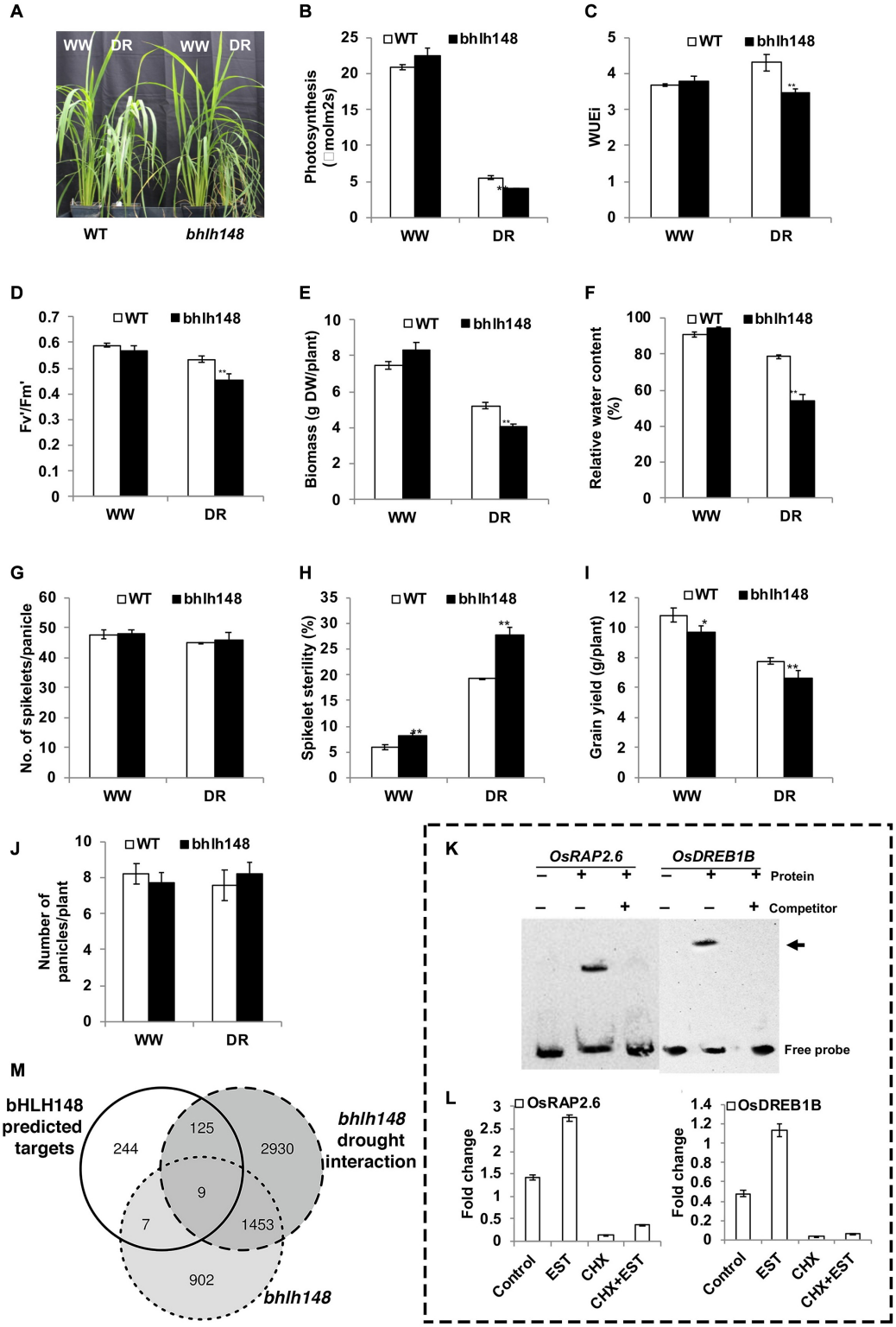

FIGURE 9 | Phenotyping *bHLH148* as experimental validation of the GRAiN web application. OsbHLH148 was strongly predicted for association with drought by our network-based SVM classifier. We queried OsbHLH148 in the GRAiN web application and sought to test the predictions experimentally. **(A)** Increased sensitivity of *bhlh148* mutant plants under controlled drought stress conditions. Forty-five-day old plants were maintained at 100% (well-watered – WW) and 40% (drought – DR) FC (field capacity) for 10 days by a gravimetric approach, and performance was measured at the end of the stress period. **(B–F)** The phenotype of the WT

(Continued)

FIGURE 9 | Continued
and *bhlh148* mutant plants under drought stress. **(B)** Assimilation rate, **(C)** instantaneous water use efficiency (WUEi), **(D)** the efficiency of Photosystem II in light-adapted leaves, **(E)** above-ground biomass (dry weight), and **(F)** relative water content (RWC). Gas exchange measurements were taken using a portable photosynthesis system LI-6400XT at a CO_2 concentration of 370 μmol/mol and light intensity of 1000 $\mu mol/m^2/s$. The data are the means ± SE (n = 10) and significance using the t-test (**$P \leq 0.01$). **(G–J)** Reduced grain yield of *bhlh148* plants under well-watered as well as drought stress conditions. Drought stress was applied by withholding irrigation at the R3 stage for 4–8 days until the leaves roll and wilt, followed by re-watering and maintaining under well-watered conditions until physiological maturity. Yield components were measured under well-watered and drought stress conditions at physiological maturity. **(G)** the number of spikelets, **(H)** percent spikelet sterility, **(I)** grain yield, and **(J)** the total number of panicles. The data are means ± SE (n = 6) and significance using t-test (*$P \leq 0.05$ and **$P \leq 0.01$). **(K,L)** Experimental validation of predicted OsbHLH148 targets predicted from the GRAiN web application. **(K)** Electrophoretic mobility shift assay (EMSA) was performed with bHLH148 protein and biotin-labeled promoter elements of potential bHLH148 regulated genes. bHLH148-6xHis recombinant protein was incubated with promoter elements at room temperature for 20 min. For competition analysis, the binding reaction was incubated for 10 min on ice before adding 100-fold excess of unlabeled promoter elements, followed by incubation at room temperature for 20 min. The samples were subjected to EMSA by PAGE and subsequent chemiluminescence detection. + and - indicate the presence and absence of the respective component in the binding reaction. Arrows indicate the labeled "free probe" and DNA-protein complex "bound probe" positions. **(L)** Direct activation of *OsRAP2.6* and *OsDREB1B* by bHLH148. Rice protoplasts were transfected with a bHLH148-HER fusion construct driven by the CaMV35S promoter. Transfected protoplasts were treated with estradiol (EST), cycloheximide (CHX), or EST and CHX together. The expression levels of *OsRAP2.6* and *OsDREB1B* in control and treated protoplast was analyzed by qPCR and shown for *RAP2.6* and *OsDREB1B*. Each data point is mean values ± SE of three biological replicates. **(M)** Venn diagram showing overlaps between GRAiN predicted targets of OsbHLH148 and genes that are differentially expressed in the mutant as well the mutant treated with drought.

Our study agrees with the previous reports which claimed that rather than using a single algorithm, an ensemble-centric approach improves the GRN inference performance. Adding diverse methods to an ensemble of network prediction methods should, theoretically, stabilize biologically relevant relationships between TFs and target genes (Marbach et al., 2012). We observed this phenomenon in our study, as the consensus predictions from the five network prediction algorithms outperformed individual methods (**Figure 3**). Interested researchers who wish to apply this consensus approach should also note that having more ensemble algorithms does not guarantee superior network inference performance. The correct combination of methods will depend on several factors, including the dimensions and the nature of the underlying dataset. In our study, removing the two correlation-based methods from the ensemble seems to have improved the final network's performance in the experimental benchmark (ChIP-seq data) and not the two secondary reference networks (**Figure 3**). This could be explained by the fact that simple correlation-based methods are prone to a high accumulation of false positives arising from indirect correlations. The other three algorithms (CLR, GENIE3, and ARACNe) are specially designed to attenuate this problem. Using a consensus of only these three were better able to detect direct regulatory edges represented by ChIP-seq data. The *ad hoc* reference networks, on the other hand, might contain several false positives because they were built from derived data rather than direct experimental evidence. Therefore, removing the correlation-based methods from the ensemble barely made any difference when tested on this benchmark.

Several other algorithms not used in our study can potentially further improve the ensemble's diversity. For example, module-based algorithms first apply clustering algorithms to the expression data and then assign regulators to the identified modules. While such an approach can potentially retrieve targets of TFs with less correlated expression profiles (De Smet and Marchal, 2010), there are several places in the module-based inference workflows where subjective biases can be introduced (e.g., the choice of the number of clusters to extract, which should ideally be chosen by thorough testing a range of clustering parameters). We found that module-based network inference algorithms generally have a more considerable computational burden (data not shown), especially on the relatively larger rice gene expression matrices. Other algorithms that use an integrative or supervised approach could also not be used in our study (Bonneau et al., 2006; Banf and Rhee, 2017; Zarayeneh et al., 2017). This is because the only other mutually exclusive datatype available for integration is the sequence-based DNA motif data. However, unlike expression patterns, most DNA binding motifs of rice TFs are not experimentally determined but predicted based on homologies. Also, a large fraction of rice TFs do not even have their corresponding binding sites predicted. Therefore, using DNA-motif data only to gauge the quality of the networks we predicted, but not the network inference itself, kept us in line with our goal of including as many genes as possible.

We named our network GRAiN. GRAiN is essentially a bipartite network as it has two types of nodes (TFs and modules). Our final goal was to develop an algorithm that uses machine learning to identify GRAiN patterns that characterize a particular set of nodes with verifiable attributes (gold standard TFs). To select our gold standard, we surveyed various phenotype databases. Our survey shows that while currently ∼2% (1098 at the time of this study) of all known rice genes have been linked to various abiotic stresses experimentally, more than 15% of these stress genes are TFs linked with drought or water deficit related responses. Our survey suggests that the genetic selection of favorable alleles of the stress-inducible TFs has been widely and inadvertently used as a tool to improve/select for drought tolerance. We listed 165 TFs linked with drought to train the machine learning algorithm. Our observations that most of these gold-standard drought regulators do not show sizeable differential expression patterns under drought further motivated us to develop such a computational model (**Figure 4**).

Our framework funnels an inferred modular GRN into the SVM that learned to discriminate between real drought TFs from those that are likely not regulators of drought, based on their network connectivity patterns. Our model's application ranked

every TF in the network according to their predicted association with DR. Therefore, the selection of regulatory genes using our approach remains less prone to subjective bias.

GRAiN and the subsequent network-based machine learning approach we presented in this study can also be applied to transcriptome collections within other biological contexts for which enough training labels are also available. Furthermore, the vertical integration of different data types could allow the development of more mechanistically informed models. Integrating GRAiN with other diverse sources of information (representing different layers of gene regulation) into a single prediction model will allow candidate gene selection in a truly holistic manner. For example, datasets featuring paired measurements of transcriptome networks (tissue or cell-type specific) and post-transcriptional regulation (e.g., small RNA-sequencing). Integration of other data types such as epigenetic profiles and post-translational modifications (PTM) such as phosphorylation also seems feasible with the SVM approach. Some excellent resources, such as the Plant PTM Viewer (Willems et al., 2019) and the database of phospho-sites in plants (Cheng et al., 2014) currently allow such data mining for a few plant TFs. Perhaps, vertical integration of heterogeneous data types could also help achieve a better classification of functional alleles in *indica* and *japonica* sub-types of rice, which remains a limitation of our study. However, the prediction models' generalizability will depend upon the quality of training examples, the standard of validation data, and feature engineering.

In a nutshell, our study developed a novel computational framework for network-based prioritization of regulatory genes. Application of this pipeline accomplished three main challenges in rice: (i) identification of genes that participate in similar biological processes and pathways on the occurrence of abiotic-stresses, (ii) identification of genes co-regulated by a group of TFs, and (iii) prioritization of regulatory genes and modules associated with DR. We expect our drought prediction model to have superior performance in the real-world scenario, evident by the fact that three recently reported drought TFs, which we did not include in training our models, were correctly predicted. The network-based machine learning approach presented here, in conjunction with resources like the KitaakeX Mutant Database (Li et al., 2017), can support targeted screens to narrow down the search for TFs involved in specific physiological, morphological, and biochemical phenotypes to delineate specific DR mechanisms further. We anticipate that our study will be valuable for exploring the transcriptional regulatory code of stress responses in rice.

MATERIALS AND METHODS

Development of the Consensus Gene Regulatory Network

A set of 35 Affymetrix microarray datasets comprising 265 individual gene expression samples under the context of abiotic stress were obtained from the gene expression omnibus (**Supplementary Data 9**). Datasets with at least four samples and two groups were retained, normalized, and processed into an integrated expression matrix. A comprehensive list of 2304 known rice genes annotated as TFs in several public databases was obtained (Yilmaz et al., 2009; Jung et al., 2010; Priya and Jain, 2013; Jin et al., 2014). This list of TFs and the gene expression matrix was supplied to five reverse-engineering algorithms. ARACNE was downloaded from the web link in the original publication. GENIE3 (Huynh-Thu et al., 2010) and CLR (Faith et al., 2007) runs were performed using the R package minet (Meyer et al., 2008). PCC and SCC were calculated using the Sleipnir library of functional genomics (Huttenhower et al., 2008). Note that all these algorithms are non-integrative and un-supervised. Meaning they aim to infer relationships between TFs and target genes solely from gene expression data (non-integrative), and without leveraging any prior knowledge of known interactions in the prediction process (un-supervised). Then, assuming no combinatorial regulation and feedback loops, ∼80 million regulatory links could have been predicted (35,151 genes X 2304 TFs) from the gene expression matrix. However, GRN inference remains an underdetermined problem, and knowing the exact number of true edges in a network is impossible (De Smet and Marchal, 2010). Therefore, to reduce the runtime of our workflow's subsequent steps, we selected only the top 500,000 edges from each algorithm's output (sorted and ranked based on the confidence metric given by the individual algorithm). Then, the union of selected edges from all algorithms was used to create an edge matrix E, with edges i in rows of E and algorithms j in columns of E. Each cell in the E_{ij} was populated by the rank given to i by j. Missing edges were substituted with the lowest rank of that column plus one (Marbach et al., 2012). The average rank of each row (edge) was then computed, and these averages were re-ranked to generate the final rankings. Hence, edges with small final ranks indicated greater confidence in all five methods. Edges with a final rank value of more than 500,000 were removed and the rest retained in the final consensus network.

Creation of the ChIP-Seq Benchmark and Other Reference Networks

The ChIP-targets of 9 TFs were extracted from published data files with the original studies. The Position Weight Matrices (PWM) of ∼588 rice TFs listed in the CIS-BP database were obtained (Weirauch et al., 2014). PWMs indicate DNA sequence preferences of TFs and can infer DNA motifs in the promoter regions of functional genes. The 1000 bp upstream promoters were scanned for at least one or more PWM motifs using the FIMO tool in the MEME suite (Bailey et al., 2015). Motifs found in more than 50% of all the genes were treated as 'constitutive elements' and removed. Genes harboring all the remaining motifs with a p-value $<$ 1E-10 were linked to the corresponding TFs. The GO BP reference was created by using evidence of functional relationships between TFs and non-TF genes co-annotated in the rice biological process (BP) ontologies. Only those annotation labels consisting of less than 200 genes were chosen for this. Excluding large BPs ensured that minimally related genes (in processes such as 'translation,' 'DNA repair,'

'signal transduction' etc.) did not become part of the reference network. The PPI reference network of rice was obtained from the PRIN database (Gu et al., 2011) hosted at http://bis.zju.edu.cn/prin/download.do. Only experimentally verified interactions were used, and edges with at least one TF as a corresponding node were identified.

Finding Modules in the Gene Regulatory Network

Note that the network structure obtained by taking the consensus of the ensemble was essentially a mixed bipartite graph. One set of nodes represented TFs, and the other set of nodes represent target genes (functional genes plus TFs). To detect modules in such a network, the graph's biadjacency matrix was first converted to a unipartite network following the approach used to build the Arabidopsis stress network (Vermeirssen et al., 2014). Then, the similarity in predicted regulators of every pair of genes was estimated using the JI of overlap. Operationally, this technique accounts for the dogma of co-regulation instead of co-expression, thus preserving the network's regulatory nature. Then, we applied the Markov clustering algorithm (mcl) on this network to find modules of co-regulated genes. The inflation parameter of the mcl algorithm was set to a value of 2 after tuning (data not shown).

Functional Annotations of Network Modules

Annotations in the rice GO, the KEGG, and CYC pathways were obtained from the plant GSEA server (Yi et al., 2013). The mapman annotation file was obtained from the MapManStore[2]. GO annotations were propagated from parent terms to child terms using the true-path rule, as described before (Ambavaram et al., 2014). From all these databases, categories that annotate more than 500 genes and less than three genes were removed. Statistical significance of overlaps between remaining gene-sets and co-regulated gene modules was calculated using hypergeometric tests. The resulting p-values were corrected for multiple testing using the Benjamini–Hochberg method (Benjamini and Hochberg, 1995). A q-value threshold of 0.05 was set to declare an observed enrichment as significant.

De novo Analysis of CREs

The FIRE tool was supplied with 1000 bp upstream promoter sequences of genes within each module identified by the mcl algorithm. The parameter k, which defines the length of seed sequences, was kept between 3 and 12, and only DNA motifs were inferred. Only those motifs that occur within at least 50% of each module's genes were selected for all FIRE runs. These motifs were then matched against plant CREs listed in various public databases using the STAMP server (Mahony and Benos, 2007). Matching CREs with a p-value less than 0.001 were tagged as associated with the corresponding module. FIRE detected motifs were converted to meme format and matched with DAP-seq motifs from Arabidopsis using the tomtom tool in the meme suite (Bailey et al., 2015). CIS-BP motifs were similarly compared.

[2]https://mapman.gabipd.org/mapmanstore

Development of the GRAiN Application

The GRAiN web application is open source[3]. The code is written in the R programming language (R Core Development Team, 2012) and runs on the Shiny platform[4]. The application uses the piano package (Väremo et al., 2013) to estimate the statistical significance of overlaps between a query geneset (user entered list of genes or retrieved TF targets) and a collection of genesets (modules or GO BP sets). All genesets with the Fisher's exact test FDR corrected p-value < 0.05 are returned to the user. Network data is processed using the igraph library (Csardi and Nepusz, 2006) and visualized using the visNetwork package[5].

Generating Examples for Machine Learning

The gene keyword file from the funRiceGenes server was obtained https://funricegenes.github.io/ in May 2019. Gene lists available in the Oryzabase database were obtained from https://shigen.nig.ac.jp/rice/oryzabase/download/gene on the same day. The rice mutant database was obtained from the published article (Zhang et al., 2006). Using a word cloud analysis (not shown), most prominent keywords in these databases were visualized. Genes linked with keywords related to abiotic stress such as "drought," "water-deficit," "salt," "cold," "heat," and "temperature" were then extracted. The retrieved locus IDs and publication records of genes were manually scanned for consistency by expert stress biologists, and TFs linked with drought (and related keywords) were labeled as positives. Note that OsbHLH148 was initially present in our dataset as a drought positive TF (Seo et al., 2011). However, we removed it from the positive list before training the models as a hidden example on which wet-lab experiments were performed later. We listed negative examples from the remaining TFs as those that were not positive for any abiotic stress (salt, cold, and heat), since many genes are multi-stress responsive. We also reanalyzed seven published gene expression datasets covering drought stress responses in various organs and tissues of rice plants across multiple genotypes. TFs that did not differentially express in these datasets were also counted as drought negatives. In addition to this, the rice stress TF database was downloaded (Priya and Jain, 2013) from http://www.nipgr.ac.in/RiceSRTFDB.html and TFs not listed as responsive to drought and salt in this database was also included as negative TFs. Altogether, we created a pool of 752 TFs that are most likely not regulators of drought stress responses.

Network-Based Classifier

GRAiN is structured as a matrix G, with each entry in G_{ij} corresponding to the JI of TF i in a row with module j in the column. G was supplied as the feature-set to the linear kernel SVM classifier. The vector of JI values of each labeled TF across all modules in G represented its feature vector. The objective of an SVM function is to identify the best

[3]https://github.com/cngupta/GRAiN

[4]https://shiny.rstudio.com/

[5]https://datastorm-open.github.io/visNetwork/

hyperplane that separates the two classes of the training data (drought positive and negative TFs) using their feature vectors. The width of the margin that separates the two classes was controlled by optimizing the classification trade-off parameter (C; a penalty for a miss-classified example). An optimal $C = 1$ was chosen by testing a range of values from 0.001 to 10 in increments of 0.1 and five-fold cross-validation tests. This test split all training examples (drought positive and negative TFs) into five equal parts. The model was trained on four of the five splits and tested on the remaining split kept hidden in training, ensuring that each split was used as the test-set only once. Model accuracy was evaluated using ROC statistics. Classifier training and learning were performed using the SVMperf implementation in the Sleipnir library. Cross-validation splits and performance evaluation was performed using the ROCR (Sing et al., 2005) and PRROC packages in R. The SVM returned distance from the hyperplane D for each TF in GRAiN. The values of D were averaged over four values from the final five-fold cross-validation test and then scaled over a range from 0 to 1. The resulting value of each TF was referred to as its DS.

Analysis of RNA-Seq Data and Estimation of Differential Expression

Raw fastq files of individual samples from all external datasets were downloaded from the SRA. The Nipponbare RefSeq (MSU version 7) was obtained from the rice genome annotation project website (Kawahara et al., 2013). The barley and sorghum genomes and annotations were downloaded from the Phytozome web portal (Goodstein et al., 2012). The following procedure was uniformly applied across all RNA-seq samples, including samples from mutant experiments generated in the study described here. Reads were mapped to the respective reference genomes using STAR version 2.7 (Dobin et al., 2013). The bam files obtained from STAR runs we sorted using samtools and used as input to the HTseq software version 0.11.2 (Anders et al., 2015) with its default parameters for counting reads per gene per sample. Count of reads obtained from HTseq runs were then integrated as a count matrix (one for each experiment) with columns representing individual samples and rows representing genes. Each cell of the matrix presented raw counts of the gene in the corresponding sample. Each gene's count was first scaled by its length to give reads per kilobase (RPK). The sum of all RPK values per sample divided by 1 million gave us a scaling factor. Then, dividing each RPK value by this scaling factor computed gene expression as transcripts per million (TPM) units. The effective gene length to be used in RPK values calculations was calculated as the sum of non-overlapping exon lengths using the genomic features package in R (Lawrence et al., 2013). The GFF3 files of all genomes were converted to GTF format using GFF utilities (gffread) of the cufflinks software (Trapnell et al., 2010). The resulting GTF file was used as input to genomic features for effective gene length calculation. Note that the rice GFF3 file on rice MSU reference has mis-annotations of ∼1000 gene isoforms, which hampered gene length calculations. Conversion of GFF3 to GTF ensured proper grouping of individual transcripts to parent gene ID. For the test of differential expression, the raw count data was normalized using edgeR (Robinson et al., 2009) and transformed using voom (Law et al., 2014). The voom-transformed values were used for linear modeling using the limma package in R (Ritchie et al., 2015). Differential expression of genes between control and treatment samples was estimated from the coefficients of the linear models. The interaction between *bhlh148* and drought was estimated by subtracting the baseline effect of drought on the WT sample from the effect of drought on the mutant sample (on a log scale). Differential expression from each microarray dataset was calculated as follows. Each dataset was background corrected, normalized, and summarized using the RMA algorithm (Irizarry et al., 2003). Then, genes with interquartile range across samples less than the median interquartile range were filtered. A linear model was then used to detect the remaining genes' differential expression using limma, as described above. In all cases, *p-values* were converted to *q-values* using the qvalue package in R to account for multiple hypothesis testing.

Controlled Drought Stress at Vegetative Stage and Physiological Measurements in Rice

To test the drought stress response of mutant plants at the vegetative stage, we applied controlled drought stress on 45-day-old plants using a gravimetric approach. One-week old equal-sized individual seedlings were transplanted into 4 square inch plastic pots filled with Redi-earth potting mix of a known weight and water holding capacity. Thirty-five days after transplanting, controlled drought stress (DR) was initiated on ten pots and monitored gravimetrically. The soil water content was brought down to 40% FC for 3–4 days, and plants were maintained at that level for 10 days by weighing the pots daily at a fixed time of the day and replenishing the water lost through evapotranspiration. Another ten pots were maintained at 100% FC and treated as WW conditions (Ramegowda et al., 2014). At the end of the stress period, gas exchange and light-adapted fluorescence measurements (Fv'/Fm') were taken on the 2nd fully expanded leaves from the top, using a portable photosynthesis meter, LI-6400XT (LI-COR Inc., NE, United States) at a CO_2 concentration of 370 $\mu molmol^{-1}$, the light intensity of 1000 $\mu molm^{-}2s^{-1}$ and RH of 55–60%. Instantaneous WUEi was calculated using the net photosynthetic rate (A) and transpiration rate (T) as WUEi = (A/T). Leaf RWC was measured as described (Barr and Weatherley, 1962) in the leaves used for gas exchange measurements. The leaf fragments of the same length were excised, and fresh weight (FW) was measured immediately. Leaf fragments were hydrated to full turgidity by floating them on deionized water for six h, then blotted on a paper towel and the fully turgid weight (TW) taken. The leaf samples were then oven-dried at 80°C for 72 h and weighed to determine the dry weight (DW). The percent RWC was calculated as RWC (%) = (FW – DW)/(TW – DW) × 100. Shoots were harvested, oven-dried at 80°C for 72 h, and weighed to determine biomass.

Grain Yield Analysis Under Reproductive Drought in Rice

The effect of drought stress on grain yield of the rice genotypes was tested by applying drought stress to plants at the R3 stage (Counce et al., 2000). Individual plants in 4 square inch plastic pots were grown at WW conditions until the R3 stage. Drought stress was applied by withholding water at the R3 stage for 4–8 days until all of the leaves wilted, followed by re-watering. Panicles exposed to drought stress during the 4–8 days window were marked and used for yield component analysis. A set of WW plants were also maintained as controls. Plants were further grown in WW conditions until physiological maturity. Drought exposed panicles were harvested, and the number of filled and unfilled spikelets counted to determine spikelet sterility (%). The filled spikelets were dried at 37°C for 5 days and weighed to determine grain yield/plant.

Electrophoretic Mobility Shift Assay (EMSA)

The total RNA isolated from drought-stressed rice plants was used to amplify full-length cDNA encoding bHLH148 and cloned into pET28(a) vector at *Bam*HI and *Eco*RI sites. The bHLH148-6xHis recombinant fusion protein expression was induced with 1 mM IPTG for 4 h and purified using Ni-NTA resin. The identity of the purified protein was confirmed by western blotting (data not shown) using the His-tag antibody. The binding reaction and EMSA were carried out using a standard protocol according to the manufacturer's instructions (LightShift Chemiluminescent EMSA Kit). Promoter sequences (2 kb upstream of transcription start site) of AP2/ERF TFs were identified using the PlantPAN database[6] (Chang et al., 2008) and searched for the presence of E-box elements in the PLACE database[7] (Higo et al., 1999). Specific sets of primers were used to amplify 200 bp E-box flanking regions of each of the putative bHLH148-regulated gene promoters using rice genomic DNA as a template. The amplified promoter fragments were biotin-labeled at the 3′ end using the Biotin 3′ End DNA Labelling Kit (Pierce). The binding reactions were carried out in a buffer containing 10 mM Tris (pH 7.5), 50 mM KCl, 1 mM dithiothreitol, 2.5% glycerol, 5 mM MgCl, 0.05% Nonidet P-40, and 50 ng/μl of poly(dI-dC). For competition analysis, the binding reactions were incubated for 10 min on ice before adding 100-fold excess of unlabeled competitor DNA, and the reaction mixture was further incubated for 20 min at room temperature before loading onto a 5% native polyacrylamide gel. The resolved DNA-protein complexes were electro-blotted onto nylon membranes and subsequently detected using the chemiluminescence detection kit.

Steroid-Inducible System for Testing Direct Activation of Genes by bHLH148

The bHLH148-HER expression construct was generated by ligating the PCR-amplified full-length cDNA of bHLH148 at the *Kpn*I site fused with the regulatory region of HER at the C terminus between the CaMV 35S promoter and the NOS terminator in pUC19 vector. The construct was transfected into rice protoplasts by electroporation and incubated with 2 μM estradiol for 6 h to release cytoplasmic bound bHLH148. For the control reactions, the same concentration of ethanol used to dissolve estradiol was used. Protoplasts were treated with cycloheximide (2 μM) for 30 min before the addition of estradiol to inhibit new protein synthesis. Total RNA was isolated from the treated protoplasts and used for qPCR analysis. The data presented are the averages of three biological replicates.

All primers used in this study can be found in **Supplementary Table 2**.

[6]http://plantpan.itps.ncku.edu.tw/

[7]http://www.dna.affrc.go.jp/PLACE/

DATA AVAILABILITY STATEMENT

New RNA-seq datasets generated in this study can be found in NCBI GEO online repository (https://www.ncbi.nlm.nih.gov/geo/) under the accession GSE65024.

AUTHOR CONTRIBUTIONS

CG and AP conceived the idea. CG executed it, developed the web application, and drafted the manuscript. CG, VR, and AP designed the experiments. VR performed the drought assays and physiological analysis. SB performed the interaction experiments. AP acquired the funding and supervised the research. All the authors contributed to writing the manuscript.

FUNDING

This study was supported by the National Science Foundation NSF-MCB award 1716844: 'Systems genetics analysis of photosynthetic carbon metabolism in rice'; the NSF-EPSCoR award 1826836: RII Track-2 FEC: 'Systems genetics studies on rice genomes for analysis of grain yield and quality under heat stress,' and the Arkansas High Performance Computing Center which is funded through multiple National Science Foundation grants and the Arkansas Economic Development Commission.

ACKNOWLEDGMENTS

The authors would like to thank Arjun Krishnan of the University of Michigan, and members of the Pereira lab for relevant discussions on the topic. The authors would also like to thank the original contributors of all publicly available datasets used in this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.652189/full#supplementary-material

Supplementary Figure 1 | Analysis of *cis*-regulatory elements within coregulated modules. **(A)** A bar plot showing the proportion of co-regulated modules (*x*-axis) containing **a** differing number of *cis*-regulatory elements (CREs) (*y*-axis). **(B)** A waffle plot (alternative to a circular pie chart) showing the fraction of *de novo* identified CREs that matched different sources (colored uniquely) of putative plant CREs.

Supplementary Figure 2 | A network analysis of novel *cis*-regulatory elements. Besides recovering several known plant CREs, FIRE identified three novel motifs that did not match to any known plant CRE. The heatmap shows that these novel motifs could potentially be direct or 'associative' binding sites of members from seven TF families, based on significant overlaps of the predicted targets of TFs from the families on the *x*-axis within the genes that harbor the three novel CREs on the *y* axis (FDR-corrected hypergeometric tests p-value < 0.01). The color gradient indicates the network score, calculated as the average ranks of edges from the consensus gene regulatory network. Darker color indicates a stronger association between the CRE and the TF family, as indicated in the key.

Supplementary Figure 3 | A line plot showing relationships between the predicted drought score and network degrees of TFs.

Supplementary Figure 4 | A subset of modules with the highest feature importance scores from the drought classifier were connected to *cis*-regulatory elements (CREs; predicted by *de novo* analysis) found enriched within them, as well as to their predicted regulators (TFs). The regulators were, in turn connected to the CREs based on enrichment analysis (FDR corrected hypergeometric test p-value < 0.01). This interconnected network with three node types (modules, CREs, TFs) was visualized in Cytoscape (version 3.0). Modules are indicated in rounded rectangles, CREs in ellipses and TFs in triangles colored according to the family membership indicated in the key on the right. The network shows hubs of different node-types.

Supplementary Figure 5 | Predicted drought scores in relation to the pan-genome of rice. Boxplots showing drought score distributions of **(A)** core TF, **(B)** indica dependent TFs, and **(C)** japonica dependent TFs.

Supplementary Table 1 | Evaluation of different network inference methods. Algorithms were tested for their ability in recovering putative cis elements listed in CIS-BP database in the promoters of their predicted target genes.

Supplementary Table 2 | Primers used in the study.

Supplementary Data 1 | Top 500,000 edges inferred by the ensemble and their aggregate.

Supplementary Data 2 | Gene-module memberships.

Supplementary Data 3 | Module CRE annotations.

Supplementary Data 4 | Module pathway/process annotations.

Supplementary Data 5 | Drought scores.

Supplementary Data 6 | Feature importance scores.

Supplementary Data 7 | OsbHLH148 GRAiN prediction results.

Supplementary Data 8 | OsbHLH148 differential expression test results.

Supplementary Data 9 | GEO datasets used to build GRAiN.

REFERENCES

Ahammed, G. J., Li, X., Zhou, J., Zhou, Y.-H., and Yu, J.-Q. (2016). "Role of hormones in plant adaptation to heat stress," in *Plant Hormones under Challenging Environmental Factors*, eds G. J. Ahammed and J.-Q. Yu (Dordrecht: Springer), 1–21. doi: 10.1007/978-94-017-7758-2_1

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.

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq - a python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169. doi: 10.1093/bioinformatics/btu638

Ashraf, M. (2010). Inducing drought tolerance in plants: recent advances. *Biotechnol. Adv.* 28, 169–183. doi: 10.1016/j.biotechadv.2009.11.005

Bailey, T. L., Johnson, J., Grant, C. E., and Noble, W. S. (2015). The MEME suite. *Nucleic Acids Res.* 43, W39–W49.

Banf, M., and Rhee, S. Y. (2017). Enhancing gene regulatory network inference through data integration with markov random fields. *Sci. Rep.* 7:41174.

Barr, H. D., and Weatherley, P. E. (1962). A re-examination of the relative turgidity technique for estimating water deficit in leaves. *Aust. J. Biol. Sci.* 15, 413–428. doi: 10.1071/bi9620413

Basu, S., Ramegowda, V., Kumar, A., and Pereira, A. A. (2016). Plant adaptation to drought stress. *F1000Res.* 5:F1000FacultyRev-1554.

Baxter, I. (2020). We aren't good at picking candidate genes, and it's slowing us down. *Curr. Opin. Plant Biol.* 54, 57–60. doi: 10.1016/j.pbi.2020.01.006

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 Methodol.* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Bhardwaj, N., Kim, P. M., and Gerstein, M. B. (2010). Rewiring of transcriptional regulatory networks: hierarchy, rather than connectivity, better reflects the importance of regulators. *Sci. Signal.* 3:ra79. doi: 10.1126/scisignal.2001014

Bonneau, R., Reiss, D. J., Shannon, P., Facciotti, M., Hood, L., Baliga, N. S., et al. (2006). The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol.* 7:R36.

Boyer, J. S. (1982). Plant productivity and environment. *Science* 218, 443–448. doi: 10.1126/science.218.4571.443

Bray, E. A. (1997). Plant responses to water deficit. *Trends Plant Sci.* 2, 48–54. doi: 10.1016/s1360-1385(97)82562-9

Cantalapiedra, C. P., García-Pereira, M. J., Gracia, M. P., Igartua, E., Casas, A. M., and Contreras-Moreira, B. (2017). Large differences in gene expression responses to drought and heat stress between elite barley cultivar scarlett and a spanish landrace. *Front. Plant Sci.* 8:647. doi: 10.3389/fpls.2017.00647

Century, K., Reuber, T. L., and Ratcliffe, O. J. (2008). Regulating the regulators: the future prospects for transcription-factor-based agricultural biotechnology products. *Plant Physiol.* 147, 20–29. doi: 10.1104/pp.108.117887

Chang, W.-C., Lee, T.-Y., Huang, H.-D., Huang, H.-Y., and Pan, R.-L. (2008). PlantPAN: plant promoter analysis navigator, for identifying combinatorial cis-regulatory elements with distance constraint in plant gene groups. *BMC Genomics* 9:561. doi: 10.1186/1471-2164-9-561

Chen, W. J., and Zhu, T. (2004). Networks of transcription factors with roles in environmental stress response. *Trends Plant Sci.* 9, 591–596. doi: 10.1016/j.tplants.2004.10.007

Cheng, H., Deng, W., Wang, Y., Ren, J., Liu, Z., and Xue, Y. (2014). dbPPT: a comprehensive database of protein phosphorylation in plants. *Database* 2014:bau121.

Chung, P. J., Jung, H., Choi, Y. D., and Kim, J.-K. (2018). Genome-wide analyses of direct target genes of four rice NAC-domain transcription factors involved in drought tolerance. *BMC Genomics* 19:40. doi: 10.1186/s12864-017-4367-1

Clauw, P., Coppens, F., Korte, A., Herman, D., Slabbinck, B., Dhondt, S., et al. (2016). Leaf growth response to mild drought: natural variation in *Arabidopsis* sheds light on trait architecture. *Plant Cell* 28:2417. doi: 10.1105/tpc.16.00483

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297.

Counce, P. A., Keisling, T. C., and Mitchell, A. J. (2000). A uniform, objective, and adaptive system for expressing rice development Paper no. 99001 published with the approval of the director, agricultural experiment station, university of arkansas, fayetteville, AR 72701. This research was supported by a grant from the arkansas rice research and promotion board. *Crop Sci.* 40, 436–443.

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal Complex Syst.* 1695, 1–9.

Cui, X., Lv, Y., Chen, M., Nikoloski, Z., Twell, D., and Zhang, D. (2015). Young genes out of the male: an insight from evolutionary age analysis of the pollen transcriptome. *Mol. Plant* 8, 935–945. doi: 10.1016/j.molp.2014.12.008

De Smet, R., and Marchal, K. (2010). Advantages and limitations of current network inference methods. *Nat. Rev. Microbiol.* 8, 717–729. doi: 10.1038/nrmicro2419

Ding, X., Li, X., and Xiong, L. (2013). Insight into differential responses of upland and paddy rice to drought stress by comparative expression profiling analysis. *Int. J. Mol. Sci.* 14, 5214–5238. doi: 10.3390/ijms14035214

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Dubouzet, J. G., Sakuma, Y., Ito, Y., Kasuga, M., Dubouzet, E. G., Miura, S., et al. (2003). OsDREB genes in rice, *Oryza sativa* L., encode transcription activators that function in drought-, high-salt- and cold-responsive gene expression. *Plant J.* 33, 751–763. doi: 10.1046/j.1365-313x.2003.01661.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

Emiliani, G., Fondi, M., Fani, R., and Gribaldo, S. (2009). A horizontal gene transfer at the origin of phenylpropanoid metabolism: a key adaptation of plants to land. *Biol. Direct* 4:7. doi: 10.1186/1745-6150-4-7

Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., et al. (2007). Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* 5:e8. doi: 10.1371/journal.pbio.0050008

Foo, M., Gherman, I., Zhang, P., Bates, D. G., and Denby, K. J. (2018). A framework for engineering stress resilient plants using genetic feedback control and regulatory network rewiring. *ACS Synth. Biol.* 7, 1553–1564. doi: 10.1021/acssynbio.8b00037

Fraser, C. M., and Chapple, C. (2011). The phenylpropanoid pathway in Arabidopsis. *Arab. B.* 9:e0152. doi: 10.1199/tab.0152

Gahlaut, V., Jaiswal, V., Kumar, A., and Gupta, P. K. (2016). Transcription factors involved in drought tolerance and their possible role in developing drought tolerant cultivars with emphasis on wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* 129, 2019–2042. doi: 10.1007/s00122-016-2794-z

Gaj, T., Gersbach, C. A., and Barbas, C. F. III (2013). ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol.* 31, 397–405. doi: 10.1016/j.tibtech.2013.04.004

Geffers, R., Cerff, R., and Hehl, R. (2000). Anaerobiosis-specific interaction of tobacco nuclear factors with cis-regulatory sequences in the maize GapC4 promoter. *Plant Mol. Biol.* 43, 11–21.

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., et al. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40, D1178–D1186.

Gu, H., Zhu, P., Jiao, Y., Meng, Y., and Chen, M. (2011). PRIN: a predicted rice interactome network. *BMC Bioinformatics* 12:161. doi: 10.1186/1471-2105-12-161

Guan, Y., Ackert-Bicknell, C. L., Kell, B., Troyanskaya, O. G., and Hibbs, M. A. (2010). Functional genomics complements quantitative genetics in identifying disease-gene associations. *PLoS Comput. Biol.* 6:e1000991. doi: 10.1371/journal.pcbi.1000991

Guan, Y., Gorenshteyn, D., Burmeister, M., Wong, A. K., Schimenti, J. C., Handel, M. A., et al. (2012). Tissue-specific functional networks for prioritizing phenotype and disease genes. *PLoS Comput. Biol.* 8:e1002694. doi: 10.1371/journal.pcbi.1002694

Gupta, C., and Pereira, A. (2019). Recent advances in gene function prediction using context-specific coexpression networks in plants. *F1000Res* 8:F1000FacultyRev-153.

Gutterson, N., and Zhang, J. Z. (2004). Genomics applications to biotech traits: a revolution in progress? *Curr. Opin. Plant Biol.* 7, 226–230. doi: 10.1016/j.pbi.2003.12.002

Haque, S., Ahmad, J. S., Clark, N. M., Williams, C. M., and Sozzani, R. (2019). Computational prediction of gene regulatory networks in plant growth and development. *Curr. Opin. Plant Biol.* 47, 96–105. doi: 10.1016/j.pbi.2018.10.005

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

Hase, T., Ghosh, S., Yamanaka, R., and Kitano, H. (2013). Harnessing diversity towards the reconstructing of large scale gene regulatory networks. *PLoS Comput. Biol.* 9:e1003361. doi: 10.1371/journal.pcbi.1003361

Hatton, D., Sablowski, R., Yung, M. H., Smith, C., Schuch, W., and Bevan, M. (1995). Two classes of cis sequences contribute to tissue-specific expression of a PAL2 promoter in transgenic tobacco. *Plant J.* 7, 859–876. doi: 10.1046/j.1365-313x.1995.07060859.x

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

Hoang, X. L. T., Nhi, D. N. H., Thu, N. B. A., Thao, N. P., and Tran, L. P. (2017). Transcription factors and their roles in signal transduction in plants under abiotic stresses. *Curr. Genomics* 18, 483–497.

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

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS One* 5:e12776. doi: 10.1371/journal.pone.0012776

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4, 249–264. doi: 10.1093/biostatistics/4.2.249

Jansing, J., Schiermeyer, A., Schillberg, S., Fischer, R., and Bortesi, L. (2019). Genome editing in agriculture: technical and practical considerations. *Int. J. Mol. Sci.* 20:2888. doi: 10.3390/ijms20122888

Jeffares, D. C., Penkett, C. J., and Bahler, J. (2008). Rapidly regulated genes are intron poor. *Trends Genet.* 24, 375–378. doi: 10.1016/j.tig.2008.05.006

Jin, J., Zhang, H., Kong, L., Gao, G., and Luo, J. (2014). PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Res.* 42, D1182–D1187.

Jin, Y., Pan, W., Zheng, X., Cheng, X., Liu, M., Ma, H., et al. (2018). OsERF101, an ERF family transcription factor, regulates drought stress response in reproductive tissues. *Plant Mol. Biol.* 98, 51–65. doi: 10.1007/s11103-018-0762-5

Joshi, A., De Smet, R., Marchal, K., Van de Peer, Y., and Michoel, T. (2009). Module networks revisited: computational assessment and prioritization of model predictions. *Bioinformatics* 25, 490–496. doi: 10.1093/bioinformatics/btn658

Jung, K. H., Cao, P., Seo, Y. S., Dardick, C., and Ronald, P. C. (2010). The rice kinase phylogenomics database: a guide for systematic analysis of the rice kinase super-family. *Trends Plant Sci.* 15, 595–599. doi: 10.1016/j.tplants.2010.08.004

Kakumanu, A., Ambavaram, M. M. R., Klumas, C., Krishnan, A., Batlang, U., Myers, E., et al. (2012). Effects of drought on gene expression in maize reproductive and leaf meristem tissue revealed by RNA-Seq. *Plant Physiol.* 160:846. doi: 10.1104/pp.112.200444

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

Kenrick, P., and Crane, P. R. (1997). The origin and early evolution of plants on land. *Nature* 389, 33–39. doi: 10.1038/37918

Krannich, C. T., Maletzki, L., Kurowsky, C., and Horn, R. (2015). Network candidate genes in breeding for drought tolerant crops. *Int. J. Mol. Sci.* 16, 16378–16400. doi: 10.3390/ijms160716378

Krishnan, A., Gupta, C., Ambavaram, M. M. R., and Pereira, A. (2017). RECoN: rice environment coexpression network for systems level analysis of abiotic-stress response. *Front. Plant Sci.* 8:1640.

Krishnan, A., Zhang, R., Yao, V., Theesfeld, C. L., Wong, A. K., Tadych, A., et al. (2016). Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat. Neurosci.* 19, 1454–1462. doi: 10.1038/nn.4353

Kurata, N., and Yamazaki, Y. (2006). Oryzabase. An integrated biological and genome information database for rice. *Plant Physiol.* 140, 12–17. doi: 10.1104/pp.105.063008

Lam, E., and Chua, N. H. (1990). GT-1 binding site confers light responsive expression in transgenic tobacco. *Science* 248, 471–474. doi: 10.1126/science.2330508

Langfelder, P., Mischel, P. S., and Horvath, S. (2013). When is hub gene selection better than standard meta-analysis? *PLoS One* 8:e61505. doi: 10.1371/journal.pone.0061505

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* 15:R29.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., et al. (2013). Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* 9:e1003118. doi: 10.1371/journal.pcbi.1003118

Levitt, J. (1980). *Responses of Plants to Environmental Stresses. Volume II Water Radiation, Salt, and other Stresses*. London: Academic Press.

Li, G., Jain, R., Chern, M., Pham, N. T., Martin, J. A., Wei, T., et al. (2017). The sequences of 1504 mutants in the model rice variety kitaake facilitate rapid functional genomic studies. *Plant Cell* 29, 1218–1231. doi: 10.1105/tpc.17.00154

Li, X., Chang, Y., Ma, S., Shen, J., Hu, H., and Xiong, L. (2019). Genome-wide identification of SNAC1-targeted genes involved in drought response in rice. *Front. Plant Sci.* 10:982. doi: 10.3389/fpls.2019.00982

Li, Y., Pearl, S. A., and Jackson, S. A. (2015). Gene networks in plant biology: approaches in reconstruction and analysis. *Trends Plant Sci.* 20, 664–675. doi: 10.1016/j.tplants.2015.06.013

Liu, R., Mancuso, C. A., Yannakopoulos, A., Johnson, K. A., and Krishnan, A. (2019). Supervised-learning is an accurate method for network-based gene classification. *bioRxiv [Preprint]* doi: 10.1101/721423

Lloyd, J., and Meinke, D. (2012). A comprehensive dataset of genes with a loss-of-function mutant phenotype in *Arabidopsis*. *Plant Physiol.* 158, 1115–1129. doi: 10.1104/pp.111.192393

Lloyd, J. P., Seddon, A. E., Moghe, G. D., Simenc, M. C., and Shiu, S.-H. (2015). Characteristics of plant essential genes allow for within- and between-species prediction of lethal mutant phenotypes. *Plant Cell* 27, 2133–2147. doi: 10.1105/tpc.15.00051

Lovell, J. T., Jenkins, J., Lowry, D. B., Mamidi, S., Sreedasyam, A., Weng, X., et al. (2018). The genomic landscape of molecular responses to natural drought stress in *Panicum hallii*. *Nat. Commun.* 9:5213.

Lu, Z., Yu, H., Xiong, G., Wang, J., Jiao, Y., Liu, G., et al. (2013). Genome-wide binding analysis of the transcription activator ideal plant architecture1 reveals a complex network regulating rice plant architecture. *Plant Cell* 25, 3743–3759. doi: 10.1105/tpc.113.113639

Ma, C., Zhang, H. H., and Wang, X. (2014). Machine learning for big data analytics in plants. *Trends Plant Sci.* 19, 798–808.

Ma, H.-W., Buer, J., and Zeng, A.-P. (2004). Hierarchical structure and modules in the *Escherichia coli* transcriptional regulatory network revealed by a new top-down approach. *BMC Bioinformatics* 5:199. doi: 10.1186/1471-2105-5-199

Ma, L., and Li, G. (2018). FAR1-RELATED SEQUENCE (FRS) and FRS-RELATED FACTOR (FRF) family proteins in *Arabidopsis* growth and development. *Front. Plant Sci.* 9:692. doi: 10.3389/fpls.2018.00692

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

Marbach, D., Costello, J. C., Kuffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., et al. (2012). Wisdom of crowds for robust gene network inference. *Nat. Meth.* 9, 796–804. doi: 10.1038/nmeth.2016

Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D., and Stolovitzky, G. (2010). Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl. Acad. Sci. U.S.A.* 107, 6286–6291. doi: 10.1073/pnas.0913357107

Maruyama, K., Sakuma, Y., Kasuga, M., Ito, Y., Seki, M., Goda, H., et al. (2004). Identification of cold-inducible downstream genes of the *Arabidopsis* DREB1A/CBF3 transcriptional factor using two microarray systems. *Plant J.* 38, 982–993. doi: 10.1111/j.1365-313x.2004.02100.x

Meyer, P. E., Lafitte, F., and Bontempi, G. (2008). minet: a R/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics* 9:461. doi: 10.1186/1471-2105-9-461

Michoel, T., Smet, R., Joshi, A., Peer, Y., and Marchal, K. (2009). Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks. *BMC Syst. Biol.* 3:49. doi: 10.1186/1752-0509-3-49

Mishra, P., Singh, N., Jain, A., Jain, N., Mishra, V., G, P., et al. (2018). Identification of cis-regulatory elements associated with salinity and drought stress tolerance in rice from co-expressed gene interaction networks. *Bioinformation* 14, 123–131.

Moore, B. M., Wang, P., Fan, P., Leong, B., Schenck, C. A., Lloyd, J. P., et al. (2019). Robust predictions of specialized metabolism genes through machine learning. *Proc. Natl. Acad. Sci. U.S.A.* 116:2344. doi: 10.1073/pnas.1817074116

Morozov, S. Y., and Solovyev, A. G. (2019). Emergence of intronless evolutionary forms of stress response genes: possible relation to terrestrial adaptation of green plants. *Front. Plant Sci.* 10:83. doi: 10.3389/fpls.2019.00083

Neme, R., and Tautz, D. (2013). Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics* 14:117. doi: 10.1186/1471-2164-14-117

Pabuayon, I. M., Yamamoto, N., Trinidad, J. L., Longkumer, T., Raorane, M. L., and Kohli, A. (2016). Reference genes for accurate gene expression analyses across different tissues, developmental stages and genotypes in rice for drought tolerance. *Rice* 9:32.

Palanog, A. D., Swamy, B. P. M., Shamsudin, N. A. A., Dixit, S., Hernandez, J. E., Boromeo, T. H., et al. (2014). Grain yield QTLs with consistent-effect under reproductive-stage drought stress in rice. *Field Crops Res.* 161, 46–54. doi: 10.1016/j.fcr.2014.01.004

Priya, P., and Jain, M. (2013). RiceSRTFDB: a database of rice transcription factors containing comprehensive expression, cis-regulatory element and mutant information to facilitate gene function analysis. *Database* 2013:bat027.

R Core Development Team (2012). *R: A Language and Environment for Statistical Computing. the R Foundation for Statistical Computing*. Vienna: R Core Development Team.

Rabara, R. C., Tripathi, P., and Rushton, P. J. (2014). The potential of transcription factor-based genetic engineering in improving crop tolerance to drought. *OMICS* 18, 601–614. doi: 10.1089/omi.2013.0177

Ramegowda, V., Basu, S., Krishnan, A., and Pereira, A. (2014). Rice growth under drought kinase is required for drought tolerance and grain yield under normal and drought stress conditions. *Plant Physiol.* 166, 1634–1645. doi: 10.1104/pp.114.248203

Razaghi-Moghadam, Z., and Nikoloski, Z. (2020). Supervised learning of gene regulatory networks. *Curr. Protoc. Plant Biol.* 5:e20106.

Redekar, N., Pilot, G., Raboy, V., Li, S., and Saghai Maroof, M. A. (2017). Inference of transcription regulatory network in low phytic acid soybean seeds. *Front. Plant Sci.* 8:2029. doi: 10.3389/fpls.2017.02029

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2009). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., et al. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* 34, 166–176. doi: 10.1038/ng1165

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

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., and Ramage, D. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Shen, Q., Zhang, P., and Ho, T. H. (1996). Modular nature of abscisic acid (ABA) response complexes: composite promoter units that are necessary and sufficient

for ABA induction of gene expression in barley. *Plant Cell* 8:1107. doi: 10.2307/3870355

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics* 21, 3940–3941. doi: 10.1093/bioinformatics/bti623

Smaczniak, C., Immink, R. G. H., Angenent, G. C., and Kaufmann, K. (2012). Developmental and evolutionary diversity of plant MADS-domain factors: insights from recent studies. *Development* 139:3081. doi: 10.1242/dev.074674

Sperschneider, J. (2019). Machine learning in plant–pathogen interactions: empowering biological predictions from field scale to genome scale. *New Phytol.* 228, 35–41. doi: 10.1111/nph.15771

Sperschneider, J. (2020). Machine learning in plant–pathogen interactions: empowering biological predictions from field scale to genome scale. *New Phytol.* 228, 35–41.

Speth, C., Szabo, E. X., Martinho, C., Collani, S., zur Oven-Krockhaus, S., Richter, S., et al. (2018). *Arabidopsis* RNA processing factor SERRATE regulates the transcription of intronless genes. *eLife* 7:e37078.

Stolovitzky, G., Prill, R. J., and Califano, A. (2009). Lessons from the DREAM2 Challenges. *Ann. N. Y. Acad. Sci.* 1158, 159–195. doi: 10.1111/j.1749-6632.2009.04497.x

Sun, C., Hu, Z., Zheng, T., Lu, K., Zhao, Y., Wang, W., et al. (2017). RPAN: rice pan-genome browser for 3000 rice genomes. *Nucleic Acids Res.* 45, 597–605. doi: 10.1093/nar/gkw958

Tang, Y., Bao, X., Zhi, Y., Wu, Q., Guo, Y., Yin, X., et al. (2019). Overexpression of a MYB family gene, OsMYB6, increases drought and salinity stress tolerance in transgenic rice. *Front. Plant Sci.* 10:168. doi: 10.3389/fpls.2019.00168

Taylor-Teeples, M., Lin, L., de Lucas, M., Turco, G., Toal, T. W., Gaudinier, A., et al. (2015). An *Arabidopsis* gene regulatory network for secondary cell wall synthesis. *Nature* 517, 571–575.

Tedeschi, F., Rizzo, P., Rutten, T., Altschmied, L., and Bäumlein, H. (2017). RWP-RK domain-containing transcription factors control cell differentiation during female gametophyte development in *Arabidopsis*. *New Phytol.* 213, 1909–1924. doi: 10.1111/nph.14293

Tran, L.-S. P., Nishiyama, R., Yamaguchi-Shinozaki, K., and Shinozaki, K. (2010). Potential utilization of NAC transcription factors to enhance abiotic stress tolerance in plants by biotechnological approach. *GM Crops* 1, 32–39. doi: 10.4161/gmcr.1.1.10569

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515. doi: 10.1038/nbt.1621

Tsuda, K., Kurata, N., Ohyanagi, H., and Hake, S. (2014). Genome-wide study of KNOX regulatory network reveals brassinosteroid catabolic genes important for shoot meristem function in rice. *Plant Cell* 26, 3488–3500. doi: 10.1105/tpc.114.129122

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

Uygun, S., Azodi, C. B., and Shiu, S.-H. (2019). Cis-regulatory code for predicting plant cell-type transcriptional response to high salinity. *Plant Physiol.* 181:1739. doi: 10.1104/pp.19.00653

van Dongen, S., and Abreu-Goodger, C. (2012). Using MCL to extract clusters from networks. *Methods Mol. Biol.* 804, 281–295. doi: 10.1007/978-1-61779-361-5-_15

Vandereyken, K., Van Leene, J., De Coninck, B., and Cammue, B. P. A. (2018). Hub protein controversy: taking a closer look at plant stress response Hubs. *Front. Plant Sci.* 9:694. doi: 10.3389/fpls.2018.00694

Väremo, L., Nielsen, J., and Nookaew, I. (2013). Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Res.* 41, 4378–4391. doi: 10.1093/nar/gkt111

Verma, V., Ravindran, P., and Kumar, P. P. (2016). Plant hormone-mediated regulation of stress responses. *BMC Plant Biol.* 16:86.

Vermeirssen, V., De Clercq, I., Van Parys, T., Van Breusegem, F., and Van de Peer, Y. (2014). *Arabidopsis* ensemble reverse-engineered gene regulatory network discloses interconnected transcription factors in oxidative stress. *Plant Cell* 26, 4656–4679. doi: 10.1105/tpc.114.131417

Walley, J. W., Sartor, R. C., Shen, Z., Schmitz, R. J., Wu, K. J., Urich, M. A., et al. (2016). Integration of omic networks in a developmental atlas of maize. *Science* 353, 814–818. doi: 10.1126/science.aag1125

Wang, C., Liu, Y., Li, S.-S., and Han, G.-Z. (2015). Insights into the origin and evolution of the plant hormone signaling machinery. *Plant Physiol.* 167:872. doi: 10.1104/pp.114.247403

Wang, D., Pan, Y., Zhao, X., Zhu, L., Fu, B., and Li, Z. (2011). Genome-wide temporal-spatial gene expression profiling of drought responsiveness in rice. *BMC Genomics* 12:149. doi: 10.1186/1471-2164-12-149

Wang, H., Wang, H., Shao, H., and Tang, X. (2016). Recent advances in utilizing transcription factors to improve plant abiotic stress tolerance by transgenic technology. *Front. Plant Sci.* 7:67. doi: 10.3389/fpls.2016.00067

Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., et al. (2018). Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* 557, 43–49. doi: 10.3923/ajcs.2011.43.48

Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* 158, 1431–1443. doi: 10.1016/j.cell.2014.08.009

Wilkins, O., Hafemeister, C., Plessis, A., Holloway-Phillips, M.-M., Pham, G. M., Nicotra, A. B., et al. (2016). EGRINs (Environmental Gene Regulatory Influence Networks) in rice that function in the response to water deficit, high temperature, and agricultural environments. *Plant Cell* 28, 2365–2384. doi: 10.1105/tpc.16.00158

Willems, P., Horne, A., Van Parys, T., Goormachtig, S., De Smet, I., Botzki, A., et al. (2019). The Plant PTM Viewer, a central resource for exploring plant protein modifications. *Plant J.* 99, 752–762. doi: 10.1111/tpj.14345

Yamaguchi-Shinozaki, K., and Shinozaki, K. (2006). Transcriptional regulatory networks in cellular responses and tolerance to dehydration and cold stresses. *Annu. Rev. Plant Biol.* 57, 781–803. doi: 10.1146/annurev.arplant.57.032905.105444

Yao, W., Li, G., Yu, Y., and Ouyang, Y. (2018). funRiceGenes dataset for comprehensive understanding and application of rice functional genes. *Gigascience* 7, 1–9. doi: 10.1080/87559129.2020.1733596

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

Yilmaz, A., Nishiyama, M. Y., Fuentes, B. G., Souza, G. M., Janies, D., Gray, J., et al. (2009). GRASSIUS: a platform for comparative regulatory genomics across the grasses. *Plant Physiol.* 149, 171–180. doi: 10.1104/pp.108.128579

Yu, S., Liao, F., Wang, F., Wen, W., Li, J., Mei, H., et al. (2012). Identification of rice transcription factors associated with drought tolerance using the Ecotilling method. *PLoS One* 7:e30765. doi: 10.1371/journal.pone.0030765

Yuan, X., Wang, H., Cai, J., Bi, Y., Li, D., and Song, F. (2019). Rice NAC transcription factor ONAC066 functions as a positive regulator of drought and oxidative stress response. *BMC Plant Biol.* 19:278. doi: 10.1186/s12870-019-1883-y

Zarayeneh, N., Ko, E., Oh, J. H., Suh, S., Liu, C., Gao, J., et al. (2017). Integration of multi-omics data for integrative gene regulatory network inference. *Int. J. Data Min. Bioinform.* 18, 223–239. doi: 10.1504/ijdmb.2017.087178

Zhang, J., Li, C., Wu, C., Xiong, L., Chen, G., Zhang, Q., et al. (2006). RMD: a rice mutant database for functional analysis of the rice genome. *Nucleic Acids Res.* 34, D745–D748.

Zhu, K., Chen, F., Liu, J., Chen, X., Hewezi, T., and Cheng, Z.-M. (2016). Evolution of an intron-poor cluster of the CIPK gene family and expression in response to drought stress in soybean. *Sci. Rep.* 6:28225.

Zhu, M.-D., Zhang, M., Gao, D.-J., Zhou, K., Tang, S.-J., Zhou, B., et al. (2020). Rice OsHSFA3 gene improves drought tolerance by modulating polyamine biosynthesis depending on abscisic acid and ROS levels. *Int. J. Mol. Sci.* 21:1857. doi: 10.3390/ijms21051857

Zong, W., Tang, N., Yang, J., Peng, L., Ma, S., Xu, Y., et al. (2016). Feedback regulation of ABA signaling and biosynthesis by a bZIP transcription factor targets drought-resistance-related genes. *Plant Physiol.* 171:2810. doi: 10.1104/pp.16.00469

Zou, C., Sun, K., Mackaluso, J. D., Seddon, A. E., Jin, R., Thomashow, M. F., et al. (2011). Cis-regulatory code of stress-responsive transcription in *Arabidopsis*

thaliana. *Proc. Natl. Acad. Sci. U.S.A.* 108, 14992–14997. doi: 10.1073/pnas.1103202108

Conflict of Interest: 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.