Environmental gene regulatory influence networks in rice (*Oryza sativa*): response to water deficit, high temperature and agricultural environments

Olivia Wilkins[1,2,10], Christoph Hafemeister[1,10], Anne Plessis[1,3], Meisha-Marika Holloway-Phillips[4], Gina Pham[1], Adrienne B. Nicotra[4], Glenn B. Gregorio[5,6], S.V. Krishna Jagadish[5,7], Endang M. Septiningsih[5,8], Richard Bonneau[1,9,11], Michael Purugganan[1,11]

[1] Department of Biology and Center for Genomics and Systems Biology, New York University, New York, New York, USA, 11225

[2] Present address: Department of Plant Science, McGill University, Montréal, Québec, Canada, H9X 3V9

[3] Present address: School of Biological Sciences, Plymouth University, Drake Circus, Plymouth, UK

[4] Research School of Biology, Australian National University, Canberra, Australian Capital Territory, Australia

[5] International Rice Research Institute, DAPO Box 7777, Metro Manila, Philippines

[6]Present address: East-West Seed Company, Sampaloc, San Rafael, Bulacan, Philippines

[7]Present address: Department of Agronomy, 3706 Throckmorton Plant Sciences Center, Kansas State University, Manhattan, Kansas, 66506

[8]Present address: Department of Soil and Crop Sciences, Texas A&M University, College Station, Texas, USA

[9]Simons Center for Data Analysis, Simons Foundation, New York, New York, USA

Contact Information

Richard Bonneau: rb133@nyu.edu @RichBonneauNYU

Michael Purugganan: mp132@nyu.edu

Additional Footnotes

[10] these authors contributed equally to this work

[11] these authors are the senior and corresponding authors

SUMMARY

We inferred an environmental gene regulatory influence network (EGRIN) of the response of tropical Asian rice (*Oryza sativa*) to high temperatures, water deficit and agricultural environments. This network integrates transcriptome data (RNA-seq) and chromatin accessibility measurements (ATAC-seq) from five rice cultivars that were grown in controlled experiments and in agricultural fields. We identified open chromatin regions covering ~2% of the genome. These regions were highly overrepresented proximal to the transcriptional start sites of genes and were used to define the promoters for all genes. We used the occurrences of known *cis*-regulatory motifs in the promoters to generate a network prior comprising 77,071 interactions. We then estimated the regulatory activity of each TF (TFA;143 TFs) based on the expression of its target genes in the network prior across 360 experimental conditions. We inferred an EGRIN using the estimated TFA, rather than the TF expression, as the regulator. The EGRIN identified hypotheses for 4,052 genes regulated by 113 TFs; of these, 18% were in the network prior. We resolved distinct regulatory roles for members of a large TF family, including a putative regulatory connection between abiotic stress and the circadian clock, as well as specific regulatory functions for TFs in the drought response. We find that TFA estimation is an effective way of incorporating multiple genome-scale measurements into network inference and that supplementing data from controlled experimental conditions with data from outdoor field conditions increases the resolution of EGRIN inference.

INTRODUCTION

Plants alter the expression levels of different sets of genes to coordinate physiological and developmental responses to environmental change (Nagano et al., 2012; Plessis et al., 2015). The ability to respond to environmental signals is the hallmark of adaptation and underlies tolerance to biotic and abiotic stresses. For domesticated crop species, like Asian rice (*Oryza sativa*) adaptive gene expression patterns associated with environmental changes can ensure high yields under diverse climatic regimes (Mickelbart et al., 2015; Olsen and Wendel, 2013).

Rice is a staple food for more than half of the world's population (Khush, 2005). Changes in rice yield caused by climate change have major implications for global food security (Pachauri et al., 2014). An estimated 45% of rice growing lands are at risk of drought because they are not irrigated and depend entirely on rainfall for water (Tuong and Bouman, 2003). Moreover, many rice-growing regions have temperatures bordering critical limits for optimal grain production (Peng et al., 2004; Prasad et al., 2006; Wassmann et al., 2009). Current climate models predict marked reductions in rice yield due to changes in the frequency and intensity of extreme climate events (Pachauri et al., 2014; Redfern et al., 2012). Understanding the mechanisms that permit growth in fluctuating environments is a critical step towards identifying the molecular processes that could be targeted through traditional breeding or genetic engineering for developing stress tolerant plants.

Plants rely on gene regulatory networks to orchestrate dynamic adaptive changes that enable them to survive in growth limiting environments. Gene regulatory networks are the core information processing mechanism of the cell; they coordinate the timing and rate of genome-wide gene expression in response to environmental and developmental signals (Bonneau, 2008; Huynh-Thu and Sanguinetti, 2015; Imam et al., 2015; Roy et al., 2013; Schulz et al., 2012). Environmental gene regulatory influence networks (EGRINs) are defined by the environmentally-modulated interactions of protein transcription factors (TFs) with conserved regulatory elements in genomic DNA (Sullivan et al., 2014; Zhang et al., 2015) to effect the organization of the chromosome and the transcription of RNAs, including miRNAs and lncRNAs, that in turn have regulatory potential (Mercer and Mattick, 2013). Plant genomes encode an expanded repertoire of TFs relative to other organisms (Shiu et al., 2005) that is consistent with their sessile lifestyles and their dependence on gene response mechanisms to cope with variable environments.

The need to map out and dissect gene regulatory networks has led to the development of various experimental and computational approaches to infer their structure and composition (Bonneau, 2008; Greenfield et al., 2013; Koryachko et al., 2015; Roy et al., 2013). The simplest large scale EGRINs are based on transcriptome data, measured by high-throughput sequencing or array-based technology, without regard

for other post-transcriptional and translational regulatory events that are known to influence transcriptional regulation (Koryachko et al., 2015). These methods assume that the expression of genes across environmental conditions, perturbations and genotypes can be used to predict regulatory relationships; however, many TF proteins exist in an inactive form in the cytosol or nucleus until they are activated by environmental or developmental signals (Fu et al., 2011; Ohama et al., 2016). It is not feasible to measure all of the complex and varied factors contributing to the regulation of gene expression; as such, network inference algorithms have been developed to predict regulatory interactions in the absence of complete data by incorporating additional complementary data types or prior knowledge of the network structure to estimate the effects of unmeasured regulatory layers (Arrieta-Ortiz et al., 2015; Bonneau, 2008; Fu et al., 2011; Greenfield et al., 2013; Misra and Sriram, 2013; Roy et al., 2013).

In model prokaryotic and eukaryotic systems, where much of the true architecture of many regulatory networks is known, methods that combined expression data and additional data types that define structure priors were able to infer more accurate regulatory networks that methods based on gene expression data alone (Arrieta-Ortiz et al., 2015; Fu et al., 2011; Greenfield et al., 2013). For rice, where only a small minority of regulatory relationships are known, several EGRINs have been published that use additional knowledge of network regulators to inform transcriptome-based network inference models. For example, Obertello et al., 2015 included knowledge of regulatory interactions in other species, and Nigam et al., 2015 consider microRNA-mediated TF activity. Additionally, Sullivan *et al*. (Sullivan et al., 2014) used changes in chromatin accessibility upstream of the coding regions of genes, a hallmark of gene regulatory activity (Buenrostro et al., 2013; Zhang et al., 2015), and knowledge of *cis*-regulatory motifs, rather than changes in gene expression, to construct a *cis*-regulatory network of the response of *Arabidopsis* to high temperature and during photomorphogenesis. By mapping TFs to target genes in this manner, they were able to overcome some of the inherent limitations of co-expression networks, namely the use of co-expression as a proxy for regulation. However, this type of method does not capture the output of the regulatory network, specifically, the regulated changes in transcript abundance that expression based networks measure.

Our method combines chromatin accessibility with knowledge of *cis*-regulatory motifs to connect TFs with regulatory targets to construct a network prior. These connections are used as prior knowledge of the regulatory network structure and are a starting point for network inference based on gene expression data in response to elevated temperatures and water deficits. To uncover a greater proportion of the gene regulatory network, we incorporated transcriptome data from agricultural fields collected over multiple growing seasons with experimentally induced stress responses. Our approach overcomes some of the shortcomings implicit in transcriptome based network prediction (e.g. co-expression as a proxy for regulation), by leveraging a multi-factor experimental design. Using this approach, we have assembled the first high-resolution view of global environmental gene regulation in rice.

RESULTS

For this study, we used multiple genome scale measurements – transcriptome, nucleosome-free chromatin, and *cis*-regulatory motif occurrence – to learn the gene regulatory networks associated with response to environmental change. To this end we carried out two major types of experiments (Figure 1A) in which we exposed rice plants to a wide range of environmental conditions and monitored their functional and transcriptome responses over time. In the first experiment, we grew rice plants in climate controlled growth chambers for 14 days in hydroponic culture before initiating heat shock (a transfer from 30°C to 40°C) or water deficit treatments (removal of all available water). We measured gene expression every 15 minutes for up to 4.5h after the onset of the stress. In the second, we grew three rice cultivars in two agricultural fields, one irrigated and one rain fed, during two seasons, rainy and dry, and we monitored weather variables throughout the experimental period. This experiment is described in detail in Plessis et al. (2015).

We included five tropical rice cultivars in these experiments, all of which were traditional land races including representatives of two of the major rice subspecies – *indica* (cultivars Kinandang Puti, and Tadukan), and *japonica* (cultivars Azuenca, Pandan Wangi, and Palawan, Figure 1A). These varieties were traditionally used in either

irrigated culture (Pandan Wangi and Tadukan) or in rain fed fields (Azucena, Kinandang Puti, Palawan). Our method systematically incorporates diverse genome-scale measurements to generate predictions of gene regulatory interactions in rice leaves, including chromatin accessibility, TF binding motifs, gene expression. The regulatory network is interpreted in the context of the plant functional measurements and weather data (Figure 1B).

Photosynthetic rate is decreased in response to environmental stress

To provide the functional context in which gene expression was measured, we monitored plant physiological status during the heat shock and water deficit treatments, including carbon assimilation rate and stomatal conductance. We observed distinct functional responses for each stress type, with similar responses observed for all four cultivars (Supplemental Figure S1). In response to the heat treatment, we observed an initial steep and transient decrease in the carbon assimilation rate (~80% control) followed by a recovery to a sustained rate of ~90% control for the length of the heat stress treatment (Figure 2A, Supplemental Figure S1). Upon return to the lower temperature, the carbon assimilation rates remained stable, but below the rate in control conditions for the duration of the measured recovery period. The degree of change is variable across the cultivars, but that the trend is consistent across all.

In response to the water deficit treatment, we observed a continuous decrease in rate of carbon assimilation over the period of the treatment. For example, Azucena's carbon assimilation rate declines to around 90% of its well-watered rate after only 15 minutes of water deficit stress and to around 15% after 90 minutes, the last measurement before the plants were returned to the water-unlimited treatment. The other three cultivars similarly have decreasing rates of carbon assimilation over the period of the water-deficit treatment. In all four rice cultivars, we observe a gradual increase in the carbon assimilation rate when the plants are given ample water for recovery, though none of them returned to pre-treatment rates in the monitored recovery period.

For neither treatment did we observe functional differences that could be associated with either the subspecies (*indica v. japonica*) or the type of field (irrigated *v.* rain fed) in which the rice was traditionally grown. Because the functional measures of

the sampled leaves showed partial recovery upon return to the initial experimental conditions, and after four days post-treatment we did not observe leaf mortality on the stress treated plants, we are confident that the treatments did not cause catastrophic damage to the sampled leaves.

Fast responses to heat, slow responses to water deficit

A comparison of the four cultivars indicated that they had similar transcriptomes as shown by the high correlation of the gene expression between cultivars across the genome (Figure 2B). The correlation of gene expression was highest within each subspecies (cor=0.98 for *japonica* cultivars; cor=0.97 for the *indica* cultivars); correlation between the transcriptomes of different subspecies was also high (cor > 0.95). Given this similarity, there should only be very low bias introduced by having aligned all cultivars to the same Nipponbare reference genome. We identified differentially expressed genes in the controlled chamber experiments in response to single-factor environmental perturbations. Results indicate that the two types of treatments (heat and water deficit) lead to responses with distinct temporal patterns of expression regulation (Figure 2C) that paralleled the leaf functional responses to stress. During the heat treatment, there is rapid increase in the expression of ~300 genes, but after 90 minutes only few genes remain perturbed (Figure 2C, Supplementary File S1). In contrast, the response to water deficit is much slower, with almost no differentially expressed genes detected before 60 minutes of treatment after which a large number of differentially expressed genes were observed for the remainder of the stress period. Even 90 minutes after the plants were returned to water-unlimited conditions many genes remain perturbed. The dynamics of the stress responses are summarized in a multi dimensional scaling plot based on Euclidean distance of log-fold-change of 2,097 genes that are differentially expressed in at least one condition (Figure 2D, Supplemental Figure S2). We note that, in contrast to the response to heat shock where the response of all four genotypes was similar, the response to the water deficit stress is stronger (as measured by the number of differentially expressed genes) in the Japonica cultivars (Azucena, Pandan Wangi) than in the Indica cultivars (Kinandang Puti, Tadukan); we did not observe these differences in the plant functional measurements.

ATAC-seq interrogation of rice leaves under multiple conditions reveals stable accessible regulatory regions

Nucleosome-free regions of the genome are strongly associated with active sites of transcription. We used Assay of Transposase Accessible Chromatin (ATAC)-seq to identify nucleosome-free regions of the genome (Figure 3A). To call chromosomal regions "open", we count the number of ATAC cut sites (first base of an aligned fragment and first base after the fragment) in its proximity. We called a base open if more than half of the libraries contained at least one cut site in the 72 bp window centered on the base (Supplemental File S3). If two open bases are less than 72 bp apart, we call all intermediate bases open. Based on this definition, the average open region length was 268 bp, median 206 bp (Figure 3B). In rice, the distance between nucleosome centers (dyads) ranges from ~175bp in promoter proximal regions to ~191 bp in intergenic regions (Wu et al., 2014). In the human lymphoblastoid cell line in which the ATAC-seq method was developed, the fragment length distribution had a clear periodicity of ~200 bp, with the largest peak corresponding to the length of a single nucleosome, and multiple smaller peaks corresponding to integer multiples of up to six nucleosomes (Buenrostro et al., 2013). We did not detect peaks with lengths corresponding to multiple nucleosomes in rice which may reflect the more compact structure of plant promoters. In Arabidopsis, more than half of DNase I hypersensitivity sites near transcriptional start sites (ng y) were located within the first 200 bp upstream of the TSS, indicating that the displacement of a single nucleosome may be sufficient to permit regulated gene expression in both rice and Arabidopsis.

Designating open regions in this manner resulted in 29,978 open regions covering ca. 8M bp (~2% of the genome). The open regions were distributed through out the genome, in genic and intergenic regions, with a frequency similar to those identified using DNase I hypersensitivity sites (Zhang et al., 2012) (Supplementary Figure S3). As expected nucleosome-free regions were non-randomly distributed in the genome with respect to genomic features. We examined their distribution and calculated the enrichment of bases belonging to open regions for the following features: 500 bp upstream of TSSs, 5' UTR, coding sequence, exon, intron, 3' UTR, 500 bp downstream

of gene (Figure 3C). We observed an almost 5.6 fold enrichment of bases belonging to open regions in the 500 bp upstream of the TSS of genes and an 8.5 fold enrichment in the 5' UTRs of genes with 15% of all bases occurring in open regions (Figure 3C). Introns (~0.3 fold) and coding sequences (~0.2 fold) were depleted of open regions.

To get a better idea of where in the promoter region open chromatin was located for rice, we aligned all 56k genes in the genome with respect to their TSS. For every base from 1000 bp before the TSS to 500 bp after the TSS, we calculated the fraction of genes that were covered by an open region as described above (Figure 3D). The resulting distribution shows a sharp peak around 50 bp before the TSS where ~15% of all genes have a region of open chromatin. The distribution quickly falls off to both sides almost reaching the background level of 2.1% at -1000 and +500 bp. We used this curve to define the promoter boundaries of -453 and +137 bp based on the coverage threshold of 4% (more than 4% of genes have open chromatin between -453 and +137 bp). To test the effects of promoter openness on gene expression, we compared the expression of genes with the number of ATAC cut sites that fall within each gene's promoter (Figure 3E). For this analysis we considered only Azucena growth-chamber samples and summed the ATAC results of all libraries. Of the ~33K genes whose transcripts were not detected (FPKM of 0), 27K had 30 or fewer cut sites in their promoter. In contrast, 74% of the expressed genes had more than 30 cuts (13-fold enrichment, $p < 1e-15$) and we observed a significant correlation between the number of cuts and expression (Pearson correlation of 0.42).

TF binding motif occurrence in open promoters used to derive priors on network structure

We constructed a network prior - a collection of hypotheses connecting TFs and target genes - by mapping known TF binding motifs to the open chromatin in the promoter regions of expressed genes (Figure 4A). The *cis*-regulatory binding motifs have been determined for 666 TFs (Matys et al., 2003; Weirauch et al., 2014) of the more than 1800 TFs in the rice genome (Jin et al., 2014). We searched for these motifs in open promoter regions defined by the ATAC-seq analysis (Figure 4B), and found significant motif matches for 445 TFs. By limiting the search to the open promoter regions (~2.3

Mbp in the promoter regions of ~9K genes) we greatly reduced the search space for finding motif occurrences compared to the promoters of all expressed genes (14.7M bp); ultimately, this reduced the number of tests during motif matching. In this way, we mapped 445 TFs to 5,447 target genes via 77,071 interactions. The median out-degree (the number of regulatory edges initiating from a TF) of TFs with targets was 75; the median in-degree of targets with TFs was 8. For example, the known binding site of Heat Shock Factor A2a (OsHSFA2a - LOC_Os03g53340) is the Heat Shock Element (HSE). Fifty-three expressed genes had the HSE in an open region of their promoter, and were mapped as targets of OsHSFA2a in the open-chromatin network prior (Figure 4C).

Given this ATAC-seq and motif derived network prior, we used our expression data to select interactions that were likely to be regulatory. This step was critical as a substantial fraction of interactions in the network prior was not expected to be related to regulatory TF binding in our experimental conditions. To identify and remove non-functional interactions from the network prior, we assumed that for a given TF, true prior targets should show coordinated expression across at least some of our experimental conditions, and that the targets of the TF should be enriched for that particular expression pattern with respect to background. An example output of this step for the OsHSFA2a is shown in Figure 4C. Each target gene was given a score indicating our confidence that it was a true target of the TF based on the criteria described above; genes with low scores were removed from the network prior. The coherence-filtered network prior had 38,137 interactions; 357 TFs were connected with 3,240 target genes (Supplemental File S4). Of the TFs with targets, the median out-degree was 44; of the targets with TFs, the median in-degree was 5. By de-noising the targets, we reduced the number of edges in the network prior by 51% (Figure 4D). Some TFs had identical target gene sets in the network prior. This occurred in cases where the TFs were members of large TF families with identical DNA binding motifs. For example, there are 25 HSFs in the rice genome, all of which are predicted to bind to the same HSE. As a consequence, all HSFs have identical targets in the prior network; we group these TFs together as a TF predictor group. A total of 276 TFs formed 62 TF predictor groups. This generated a network prior of 13,937 interactions, connecting 143 regulators (individual TFs and predictor groups) with 3,240 target genes (Figure 4D).

Transcription factor activity estimation from known regulatory targets

The regulatory activity of a TF can be estimated by monitoring changes in the expression of its target genes (if known) across experimental conditions. Our method estimates TF activities (TFAs) based on partial knowledge of TF-target relationships with the aim of using this TFA estimate to then learn better estimates of TF-target relationships. Based on the prior network and the expression data, we used network component analysis (NCA) (Liao et al., 2003) to estimate TFAs. The principal idea of NCA is to use a simplified model of transcriptional regulation and treat the known or putative targets of a TF as reporters of its activity. This approach requires at least some known TF targets, hence we estimated the activities only for the 143 regulators with targets in our network prior. We found that for the majority of TFs activities and expression profiles were poorly correlated with 85% of correlation values between -0.5 and 0.5 (Figure 4E), consistent with the fact that posttranscriptional and posttranslational mechanisms are important for TF activity and localization. Many TFs had similar TFAs in our experimental conditions that could be associated with the different experimental treatments (Figure 4F). The activities indicate that some TFs primarily regulate their target genes in response to only one of the experimental treatments (i.e. heat or water deficit) while others regulate the expression of their targets in response to multiple experimental treatments (e.g. heat and water deficit). Moreover, we identify TFs with similar activities in the controlled experiments, but with divergent responses in the agricultural setting.

Because the network prior was constructed from predicted TF-target interactions based on *cis*-regulatory motifs determined by *in vitro* TF binding to protein binding microarrays we anticipated that it could contain many incorrect or irrelevant interactions. Therefore, we investigated the impact that simulated changes in the network prior had on the estimated TFAs. To do so, we subsampled the prior matrix with replacement (keeping only 63% of the interactions on average) 201 times. In this way we obtained 201 TFA estimates for every TF; we called the mean pairwise Pearson correlation 'TFA stability'. TFA stability ranged from 0.15 (Figure 4G) to greater than 0.96 (Figure 4H). As expected, predictors (TFs and TF predictor groups) with few targets in the prior (fewer

than eight) show low stability. In the remaining set of predictors, we see 64 (52%) with very stable activities (> 0.75), which shows that at least parts of the prior network are self-consistent and estimated activities for those TFs are systematic rather random. For TFs with greater than four targets, there does not appear to be a relationship between the number of targets and stability score (Supplementary Figure S4).

EGRIN: a dynamic model of transcriptional regulation in response to environmental changes

In a second step, we use the observed gene expression and the estimated TF activities to infer the EGRIN. For every gene, we assumed that we could explain its expression as a linear combination of the activities of a small number of TFs regulating it. For a given target gene, we constructed all models that corresponded to the inclusion and exclusion of prior regulators and those regulators that show a high mutual information with the target. From this large set of models, we select the one with the lowest Bayesian Information Criterion while slightly favoring models that also agree with the network prior (Greenfield et al., 2013) (see Method section). We have evaluated this method in *Bacillus subtilis* and yeast and have shown that it is robust to false interactions in the prior network and that the inferred network was consistent genome-wide expression levels following TF knock-out (Arrieta-Ortiz et al., 2015).

To estimate the error incurred by our EGRIN inference and to better rank regulatory interactions, we bootstrapped the expression data by subsampling from the conditions with replacement. Additionally, to control for the observed variability in TFAs with respect to small changes in our prior network (i.e. TFA stability), we also subsampled the network prior at every bootstrap as described above. Our approach then only chose interactions for the final consensus network if the predictor had a stable activity that predicted the target expression across a broad range of conditions. Given multiple bootstraps we kept those interactions that were present in more than 50% of the bootstrap networks. TFs with low stability were less likely to recall the same regulatory interaction in a large number of the bootstraps and as a consequence were less likely to be assigned target genes in the consensus network. As we increased the number of bootstraps, the network converges to a core set of interactions quickly — after 150

bootstraps more than 98% of the interactions remain stable after adding more bootstraps (Supplemental Figure S5). To obtain our final rice EGRIN, we performed 201 bootstraps. The final EGRIN contains 4,151 nodes (TFs, TF predictor groups and target genes) and 4,498 interactions (Figure 5A, Supplemental Files S5 and S6). Of all predicted interactions, 18% (796) were also in the network prior. TFs that were grouped in the network prior because they had identical targets were included individually in the network inference as potential target genes, but only as one potential regulator.

Known and novel regulatory control of coordinated biological processes

The predicted rice regulatory network consists of 113 TFs, around half of those (62) are TF predictor groups that were created by grouping TFs with identical targets in the prior. In total, 4,498 interactions connect the TFs to 4,052 genes. The number of targets for each regulator, and the number of regulators per target genes are shown in Figure 5B and 5C, respectively. Fifteen predictors (11 TFs and 4 TF predictor groups) had more than 100 targets in the EGRIN (Table 1). For these 15 predictors, 62% to 96% of the predicted targets were novel, i.e. not in the network prior. Many of the genes in the network were differentially expressed in response to some of our experimental treatments. We marked five groups of genes on the inferred network (Figure 5A) based on the treatments in which they were differentially expressed, including heat shock and water deficit. We also marked genes as 'circadian' if time of day was a good predictor of their expression in the chamber experiments or as 'field' if the field conditions contributed more to the overall variance than the chamber data. This notation reveals that most TFs and TF groups regulate target genes primarily in response to one treatment, while a small number of TFs regulate targets in multiple conditions. For example, TF group 5, which includes two MYB TFs (LOC_Os01g09640 and LOC_Os05g10690), had the largest number of targets of any regulator in the inferred network (355 target genes). These targets are enriched for genes involved in photosynthesis, and many of them are differentially expressed in response to both the heat and water deficit treatments – stresses that altered the photosynthetic rate in our experimental conditions (Figure 2A).

We performed Gene Ontology (GO) term enrichment analyses on the target genes of each regulator and found that the targets of 27 regulators in the network were

functionally enriched for a biological process or molecular function (each $p < 0.0001$, Supplemental File S7). Based on these analyses, we identified regions of the network whose genes were enriched for particular functions and which respond to particular environmental signals. For example, we identified a region of the network that was enriched for genes involved in RNA binding and for ribosomal structural components whose expression varied primarily as a function of the time of day. We also identified regions of the network enriched for genes with functions associated with kinases and transporter functions whose expression varied primarily in response to water deficit. Finally, we also identified regions of the network that were enriched for genes associated with photosynthesis, cell wall biosynthesis, and cellulose synthase functions whose expression vary in response to diverse environmental stimuli.

Recapitulating and extending the known functions of HSFs

Resolving large families of transcription factors with similar binding sites is a critical problem in genome-wide regulatory studies. Here we initially grouped TFs with identical binding sites, and thus identical targets in the network prior, because their estimated activities were identical following the first step in our procedure (TFA estimation). Although we learned a regulatory model for the control of each TF in large TF groups separately, we had limited resolution to distinguish different outgoing edges for these large TF-family members (the in-degree in our network is an individual TF property and the out degree is instead an aggregate property of the members of the TF family). This limitation suggests that future experiments aimed at investigating TFs in large families with nearly identical binding sites are needed. Thus for the largest TF protein-families we lack the resolution to determine which of the TFs in the predictor group are regulating the expression of the target genes.

For the most well studied predictor group in our network, the 25 Heat Shock Factors (HSFs), we wanted to determine if *post hoc* we could identify which of the HSFs were the most likely regulators of the inferred target genes. All HSFs were connected to 46 potential target genes in the network prior via the canonical Heat Shock Element (HSE) TTCnnGAAnnTTC. The estimated TFA for the predictor group increases rapidly after the onset of the heat shock treatment, peaking after 30 minutes and then slowly

decreasing over the course of the stress period; upon return to the pre-stress temperature, there is a rapid decrease in TFA – lower than the activity in control conditions – that persists for the remainder of the measured time points (Figure 6A top panel). In the EGRIN, the HSF predictor group regulates the expression of 240 target genes (Figure 5A). The targets include a number of genes involved in the unfolded protein response, including Heat Shock Proteins (HSPs) – HSP70, HSP110, DnaJ, and DnaK - and other classes of chaperone proteins (T-complex proteins, chaperonins etc.).

While our method is based on the assumption that the expression of a TF will often not be indicative of its activity, TFs whose expression is highly correlated with the group's activity are prime candidates for being the true regulators of the group's targets. We therefore examined the expression profiles of the HSFs and compared them to the TFA for the HSF predictor group (Figure 6A bottom panel). The transcripts of four HSFs were undetectable in all of our samples; the remaining 21 formed three distinct subgroups based on their expression profiles: HSFs in subgroup one were characterized by increased expression in response to heat and water deficit; HSFs in subgroup two were primarily induced in response to water deficit; and, the expression of the HSFs in subgroup three was not clearly associated with any experimental treatment. The TFA for the HSF predictor group was embedded in subgroup one, suggesting that the true predictor could be found amongst them.

Seven of the eight HSFs in subgroup one are targets of the HSF predictor group in the EGRIN. Three of them have the canonical HSE in their promoter regions, as defined above. *Os*HSFA2d and *Os*HSFA6 each have one upstream HSE, and *Os*HSFA2a has three, indicating that they are potentially regulatory targets of HSFs in addition to their potential role as regulators. The other five HSFs in this group (*Os*HSFA2c, *Os*HSFA4b, *Os*HSFB2b, *Os*HSFB2c, *Os*HSFA9) do not have HSE in their promoters, and so are likely not direct targets of HSFs. We reasoned that these HSFs might be the regulators of the inferred target genes of the HSF predictor group. *Os*HSFA2c is the TF with the highest correlation between transcript abundance and TFA in our entire data set (cor=0.92). Moreover, the binding of *Os*HSFA2c is temperature regulated *in vitro* (Mittal

et al., 2011) supporting its role as a heat responsive transcriptional regulator acting through the HSE.

The over-expression of *Os*HSFA7, a member of HSF subgroup two, enhances growth and survival in response to salt stress and water deficit by unknown mechanisms (Liu et al., 2013). Consistent with a role in the water deficit response, we find that the expression of all HSFs in subgroup two were induced in response to water deficit. However, the TFA, which is based on the occurrence of the HSE in open promoters, appears to be exclusively responsive to the high temperature treatment in our data set. We therefore hypothesize that the role of HSFs as regulators of the water deficit response are mediated by an unknown regulatory element other than the canonical HSE.

Dissimilar activity and expression hint at novel functional roles of TFs

The estimated activity of a TF is not typically correlated with the expression of that TF itself; it depends only on the expression of its targets in the network prior and which other regulators are assigned to those targets. As a result, we observe a wide range of relationships between TF expression and TFA specific to each TF. One interesting example is Early Phytochrome Responsive 1 (EPR1, LOC_Os06g51260), a MYB gene that is a core component of the circadian oscillator (Filichkin et al., 2011) that binds to the canonical Evening Element, AAAATATCT, *in vitro* (Weirauch et al., 2014). In our data, we estimate that the activity for EPR1 increases over the experimental period in the growth chamber – a 4.5h period – regardless of the experimental treatments across all four genotypes, suggesting that most of its 114 target genes in the network prior were under circadian control as expected (Figure 6B top panel). We examined the expression of these target genes in an independent data set that was designed to monitor gene expression at many time points throughout the day in rice (Nagano et al., 2012) and we found that their expression was indeed oscillating daily (Supplemental Figure S6). In our inferred network, EPR1 is the regulator of 107 target genes, of which 44 were also in the network prior (Figure 6B bottom panel).

The expression of EPR1 is poorly correlated with its estimated TFA (cor=-0.52). In the control and water deficit treatments, the levels of EPR1 expression are relatively unchanging. However, at the onset of the heat stress treatment, there is a rapid and

transient increase in EPR1 expression, which returns to the levels of the control condition for the remainder of the heat stress treatment; upon return to the pre-stress temperature, there is a steep and transient repression of EPR1 expression which returns to the level of the control conditions over the course of the recovery period. This expression profile is consistent with the activity of the HSF predictor group, and in fact, EPR1 is a target of the HSF predictor group both in the network prior and in the inferred regulatory network. In Arabidopsis, *At*HSFB2b binds to the promoter of another circadian clock component, Pseudoresponse Regulator 7 at a canonical HSE and is required for maintaining accurate circadian clock rhythms in high temperature and salt stress conditions (Kolmos et al., 2014). Notably, the better-characterized EPR1 ortholog in *Arabidopsis thaliana* also responds to heat stress (Kilian et al., 2007), though its role as a regulator of response to heat has not been investigated. These findings suggest that EPR1 may be an entry point for integrating the heat stress response with the circadian clock. The coordination of stress responses with the circadian clock is thought to be an adaptive strategy to maintain plant growth during periods of stress (Seo and Mas, 2015; Wilkins et al., 2009, 2010)

Dynamic agricultural environments increase resolution of EGRIN

Upon closer examination of the network prior and the inferred network, we found evidence of the additional value of combining growth chamber and field experiments. Because these experiments perturbed different parts of the regulatory network (e.g. different time scales and treatment duration, complexity of environments, age of plants) we were able to expand the network beyond what simply increasing sample size for any one of the two experimental designs would have afforded. By combining them in a single analysis we were able to resolve parts of the network where target genes expression was similar in response to some environmental conditions, but quite different in response others. The bZIP predictor group (TF predictor group 32 in supplementary material) is one example of a predictor where targets in the network prior showed mostly correlated expression in growth chamber data but could be divided into finer groups based on the field data (Figure 6C). Prior target groups 1, 2, and 3 show distinct expression patterns only in the field; the expression of target group 3 is most similar to the estimated activity of this predictor group. The group is composed of three bZIP TFs (LOC_Os02g49560,

LOC_Os08g38020, LOC_Os09g29820) that bind to a five-nucleotide motif CACGT. The estimated TFA for this predictor group shows increasing activity in response to the water deficit treatment in the growth chambers and then decreasing activity in the drought recovery period; moreover, based on the data collected in field, the predictor group's activity increases over the course of the dry season only in the rain fed fields (Figure 6C). We infer 66 targets of this predictor group, including a number of genes involved in cellular protection in response to water deficit: trehalose-6-phospate, three Late Embryogenesis Abundant proteins involved in osmotic stress response, five dehydrin proteins, and three protein phosphatase 2C genes.

DISCUSSION

The aim of this work was to reconstruct an environmental gene regulatory network in rice leaves in response to two of the most important environmental stresses that impact agricultural productivity – high temperature and water deficit. For this purpose, we generated two genome wide data sets: 1) 720 RNA-seq libraries generated from plants grown in heat and water deficit stress experiments in controlled environments and from plants grown in agricultural field conditions, and 2) the first ATAC-seq data set that identifies open chromatin sites in rice. We combined these data types with the largest available collection of TF binding motifs to generate a network of regulatory hypotheses for 4052 genes regulated by 113 TFs and TF groups.

This experimental design allowed us to query the transcriptomic response to heat and water deficit in an agricultural setting, as well as in a controlled environment. A greater proportion of the underlying regulatory network can be examined by exposing the plants to a broader range of environmental perturbations, such as those that exist in an agricultural setting. However, the attribution of the regulatory responses to a particular environmental signal in these conditions is challenging. By including both controlled and field experiments in our analysis, we were able to leverage the complexity of the agricultural field conditions to distinguish between the activities of TFs that responded with similar profiles to known environmental signals in the chamber. This was particularly the case for TFs whose activities responded to the water deficit treatment in

the controlled experiment. We had somewhat parallel conditions in the agricultural fields: the rice cultivars were grown in adjacent fields, where one field was irrigated and the other was rain-fed. The rain-fed field became drier over the course of the sampling period, particularly during the dry season. Some of the TFs (e.g., several bZIPs) whose activity responded to the water deficit treatment in the controlled experiments responded similarly in the dry fields while others did not, thereby distinguishing between the activities of TFs that were similar in more controlled conditions.

We used previously reported TF binding motifs and the results of our ATAC-seq analysis to construct a network prior. This prior can be interpreted as a *cis*-regulatory network connects a TF to a target gene if a TF-associated motif can be found in the accessible part of a gene's promoter. By restricting the motif search to accessible promoter regions, we reduced the search space to ~0.6% of the genome, which lowered the number of random motif matches. This approach was similar to the work of Sullivan et al. (2014), who mapped DNase I hypersensitive sites in *A. thaliana* seedlings followed by TF footprinting to derive a TF regulatory network. However, we also integrate the expression data by means of Network Component Analysis. This step allowed us to estimate TF activities that we in turn use to predict TF targets. With our approach we are limited to TFs that have known binding motifs or otherwise proposed targets. While this will have led to gaps and mis-assignments in our network, we can use new data as it becomes available to update our network prior. For example, ATAC-seq may be combined with data of ChIP-seq, DNA methylation patterns, and novel TF binding motifs in more sophisticated ways than shown here (Hoffman et al., 2012). This would increase our resolution and help break up predictor groups, as well as reduce the number of indirect interactions that we predict.

The network and source data provided with this work provides a vital starting point for studying gene regulation in rice. Our predictions testable hypotheses that may help to identify regulatory relationships for a wide range of TFs. In the future we plan to verify parts of the network through targeted assays, for example chromatin immunoprecipitation, and by reprogramming cell circuits using gene lesions and manipulations at the transcriptional level (Chavez et al., 2015; Konermann et al., 2014;

Maeder et al., 2013). The ability to examine relationships across the network will also be a useful tool in future crop design, as well as provide opportunities to study the evolution of gene networks and their relation to adaptive diversification of plant species in different environments.

Methods

Plant materials and growth conditions

Seeds of the five rice landraces used in this experiment were obtained from the International Rice Research Institute (IRRI): Azucena (AZ; IRGC#328, *Japonica*), Kinandang Puti (KP; IRGC#44513, *Indica*), Palawan (PA; IRGC#4020, *Japonica*), Pandan Wangi (PW; IRGC#35834, *Japonica*), and *Tadukan* (TD; IRGC#9804, *Indica*). AZ, KP, PW and TD were used in the controlled growth chamber experiments; AZ, PW, and PA were used in the field experiments. All experiments were conducted at the International Rice Research Institute in Los Baños, Philippines.

Single factor perturbation experiments were conducted in walk-in growth rooms. Seed dormancy was broken by incubating seeds for five days at 50°C in a dry convection oven. Seeds were germinated in water in the dark for 48h at 30°C and were then sown on hydroponic rafts suspended in 1X Peters solution (J.R. Peters Inc., Allentown, PA). The pH of the growth media was maintained at 5-5.5 throughout. Plants were grown for 14 days in climate-controlled growth chambers (12h days; 30°C/20°C day/night, 300-500 µmol quanta m^{-2} s^{-1} at the leaf surface). Relative humidity was maintained between 50-70%. Seeds and plants for the field experiments were treated as described previously (Plessis *et al.* 2015).

Experimental treatments

Single factor perturbation experiments were conducted on 14-day-old seedlings. Treatments began precisely 2h after the chamber lights were turned on. Samples were collected every 15 minutes for up to 4.5h for each of five treatments: control, heat shock, recovery from of heat shock, water deficit, recovery from water deficit. Heat shock was

initiated by moving the hydroponic rafts to a 40°C climate controlled growth chamber (RH and light intensity as above). After 2h, some plants were returned to the 30°C chamber (heat shock recovery); the remainder were kept in the 40°C chamber for the duration (heat shock). Water deficit was initiated by removing the rafts from the hydroponic media and allowing the roots to air-dry. After 1.5h, some of the plants were returned to hydroponic tanks containing Peters solution (recovery from water deficit); the remainder were kept in tanks without water (water deficit). Each sample comprised the leaves - from 10 cm above the seed - of 16 juvenile plants. Samples were harvested and flash frozen in liquid nitrogen for each treatment, time point, landrace, and replicate. The entire experiment was replicated on two consecutive days yielding two biological replicates for each condition. Samples from field experiments were harvested as described previously (Plessis *et al.*, 2015). Aliquots of these samples were used for the RNA-seq analyses.

Gas exchange

Instantaneous photosynthetic rate and stomatal conductance of the youngest fully expanded leaf was measured using a portable gas analyzer (Li-6400; Licor, Lincoln, NE, USA). Conditions in the leaf cuvette were set at a saturating light intensity of 1000 µmol quanta $m^{-1}s^{-1}$ with a target air temperature of 30°C (40°C in the case of the heat shock measurements), reference CO_2 concentration of 400µmol mol^{-1} and humidity of 70%. The measurements were corrected for leaf area. Time 0 for the heat treatment coincides with minimum functional status observed within 15 min of the start of the stress imposition. Further, the break within the trace coincides with the transition from 40 to 30°C where there was instrument instability

RNA extraction and library preparation

All protocols were conducted as per the manufacturers' instructions. Total RNA was extracted using RNeasy Mini Kits (Qiagen). RNA quality was determined by gel electrophoresis. Contaminating DNA was removed from the total RNA samples by treatment with Baseline-Zero DNase (Epicentre, Madison, WI, USA), and ribosomal RNA was removed using the Ribo-Zero rRNA Depletion Kit (Epicentre, Madison, WI, USA). Strand-specific RNA-seq libraries were synthesized using the Plant Leaf ScriptSeq

Complete Kit (Epicentre, Madison, WI, USA). The libraries were sequenced – either six or eight libraries per lane – using standard methods for paired-end 50 base pair reads, on an Illumina HiSeq 2000 at the New York University GenCore facility.

RNAseq processing

The reads were aligned to the *O. sativa* Nipponbare release 7 of the MSU Rice Genome Annotation Project reference (Kawahara et al., 2013) which consists of 373,245,519 base pairs of non-overlapping rice genome sequence from the 12 rice chromosomes. Also included are the sequences for chloroplast (134,525 bp), mitochondrion (490,520 bp), Syngenta pseudomolecule (592,136 bp), and the unanchored BAC pseudomolecule (633,585 bp). The annotation contains 56,143 genes (loci), of which 6,457 had additional alternative splicing isoforms resulting in a total of 66,495 transcripts.

We used Tophat (Kim et al., 2013; Trapnell et al., 2009) version 2.0.6 to align the reads, discarding low-quality alignments (quality score below 1). To count the number of reads that uniquely mapped to genes we used HTSeq (Anders et al., 2014) version 0.6.1. We compensated for variable sequencing depth between samples using the median-of-ratios method of DESeq2 (Love et al., 2014) version 1.6.3, and further performed a variance stabilizing transformation provided by the same package. We used the normalized count data for downstream analysis. Replicates are averaged. We removed all genes that have zero counts in more than 90% of the conditions. We also removed all genes that have a coefficient of variation smaller or equal to 0.05. This left us with 25,499 genes.

Differential Expression

For the chamber experiments, we determined differentially expressed (DE) genes using DESeq2 and the raw read counts as reported by HTSeq. For every gene and cultivar-condition combination this gave us a log-fold-change value and an adjusted p-value. To visualize the similarities and differences between the conditions and the cultivars with respect to differentially expressed genes, we applied Kruskal's non-metric multidimensional scaling to the fold change matrix. Only genes with at least one cultivar-

condition where the adjusted p-value was below 0.0001 and the absolute fold-change was greater than 2 were considered (2097 genes). The distance metric was Euclidean.

ATAC-seq library preparation

Plants were grown in the conditions described above. The second leaves of 2-week old Azucena rice seedlings were harvested and flash frozen in liquid nitrogen. Intact nuclei were isolated using a protocol generously shared with us by Drs. Wenli Zhang and Jiming Jiang (personal communication). Briefly, ground tissue was suspended in nuclear isolation buffer and washed repeatedly using nuclear wash buffer following a standard nuclear isolation protocol. Chromatin was fragmented and tagged following the standard ATAC-seq protocol (Buenrostro et al., 2015). Libraries were prepared from control, water deficit and heat-treated plants after 0.5, 2, and 4h. Libraries were sequenced on an Illumina HiSeq 2500 instrument in the RapidRun mode at the New York University GenCore facility.

ATAC-seq data processing

We prepared ATAC-seq libraries from leaves at control (30min, 2h, 4h), heat (30min, 2h, 4h), heat-recovery (4h), water deficit (2h), and water deficit-recovery (4h) conditions in an effort to identify the maximum number of open chromatin regions relevant to our experimental conditions. Two biological replicates were prepared for each condition. The average number of reads was 14.8M and the total for all libraries was 266,839,208 reads (Supplementary file S2). We used Bowtie version 2.2.3 to align the reads to the reference genome. The alignment rate was around 92%. Only 12,249,328 reads (4.6%) aligned exactly one time, as 70% of the reads aligned at least once to the chloroplast genome and to nuclear encoded chloroplast genes; only reads that aligned exactly once to the genome were used in downstream analyses.

To compare the 18 ATAC-seq samples to each other with respect to location and number of ATAC-seq cut sites (first base of an aligned fragment and first base after the fragment), we counted the number of cuts in all overlapping windows of 72 bp in each library. For each pair of libraries, we then calculated Pearson correlations of number of cuts (in $\log_{10}$) of all windows that have more than one cut in both libraries.

In order to define an atlas of accessible regions to be used in network inference we combined the ATAC-seq results from all libraries to maximize the number of identified nucleosome-free regions in the genome relevant to our experimental conditions. To define "open" regions we counted the number of ATAC cut sites that fell into the 72 bp window centered on each base. We considered a base open if its window contained at least one cut site in more than half of the libraries. If two open bases were less than 72 bp apart, we called all intermediate bases open.

Combining ATAC data and TF motifs as network prior

We used published TF binding motifs and knowledge of open chromatin based on our ATACseq data to generate a prior gene regulatory network for rice. For this, we obtained rice TF binding motifs from the CIS-BP database (Weirauch et al., 2014) dated 2015/05/18, the TRANSFAC Professional database (Matys et al., 2003) version 140805, and manual curation of literature. Since the 5' UTR upstream gene regions showed a high enrichment for open chromatin, we assumed that *cis*-regulatory elements have the largest effect. To find relevant motif occurrences, we scanned only the open regions of the rice genome (as determined by ATAC-seq) that were also in the promoter region of a gene (-453 to +137 bp with respect to transcription start site). We used FIMO (Grant et al., 2011) to find motif matches with a p-value below 1e-4, keeping only the best (lowest p-value) match per motif-gene pair. For every TF we filtered the associated motif matches by adjusting the p-values using Holm's method (controlling the Family-wise error rate) and only keeping matches with an adjusted p-value of less than 0.01. Any gene with a motif match in the open part of its promoter was then recorded as a target of the current TF.

We observed that different TFs can be associated with different motifs but yet can have almost identical sets of targets in the network prior. To prevent amplification of these miniscule differences (which stem from redundant motifs) in the downstream analysis, we unified the targets of TFs that were less than 5% different (binary distance) and assigned the union of the targets to both.

Removing uninformative edges from the network prior

To remove uninformative edges from the network prior, we grouped all genes into 160 clusters (square root of 25,499, the number of genes in the data set) to define the set of expression patterns in our data. We used hierarchical clustering with average linkage and 1 minus Pearson correlation as distance function. Then, for each TF and each expression cluster, we tested whether the TF prior targets were enriched for the members of the cluster. If the Fisher's exact test p-value was below 0.05 we marked all TF prior targets that were also members of the cluster as high-confidence targets. The above procedure was repeated for 64 bootstraps of the expression data (conditions were sampled with replacement), and only interactions with a high-confidence frequency of at least 50% were kept for the de-noised network prior.

Estimating transcription factor activity

Let X be the matrix of gene expression values, where rows are genes and columns represent experiments/samples. Let P be a matrix of regulatory relationships between TFs (columns) and target genes (rows). The entries in the prior matrix (P, the network prior) are members of the set $\{0, 1\}$. We set $P_{i,j}$ to one if and only if there is a regulatory interaction between TF j and gene i in the network prior. Auto-regulatory interactions are always set to zero in P. Estimation of TF activities is then based on the following model (Liao *et al*, 2003; Fu *et al*, 2011):

$$X_{\mathrm{i,j}} = \sum_{k \in \mathrm{TFs}} P_{\mathrm{i,k}} A_{\mathrm{k,j}},$$

where the expression of gene i in sample j can be written as the weighted sum of connected TF activities A. In matrix notation this can be written as $X = PA$, which we solve for the unknown TF activities A. This is an overdetermined system, but we can find $\hat{A}$ which minimizes $||P\,\hat{A} - \mathrm{X}||^2$ using the pseudoinverse of P. Special treatment is given to time-series experiments, with the modified model:

$$X_{\mathrm{i,t}_n + \frac{\tau}{2}} = \sum_{k \in \mathrm{TFs}} P_{\mathrm{i,k}} A_{\mathrm{k,t}_n},$$

where the expression of gene i at time $t_n + \tau/2$ is used to inform the TF activities at time t_n. Here τ is the time shift between TF expression and target expression used when inferring regulatory relationships (see next section). Here we use a smaller time shift of $\tau/2$, because changes in TF activities should be temporarily closer to target gene expression changes. If there is no expression measurement at time $t_n + \tau/2$, we use linear

interpolation to fit the values. In cases where there are no known targets for a TF, we cannot estimate its activity profile, and remove the TF from the set of potential regulators.

Network inference

The main input to the network inference procedure is the expression data X, the estimated TF activity $\hat{A}$, and the known regulatory relationships encoded in the matrix P. The core model is based on the assumption that the expression of a gene i at condition j can be written as linear combination of the activities of the TFs regulating it. Specifically, in the case of steady-state measurements, we assume

$X_{\mathrm{i,j}} = \sum_{k \in \mathrm{TFs}} \beta_{\mathrm{i,k}} \hat{\mathrm{A}}_{\mathrm{k,j}}$(1.1).

For time-series data, we explicitly model a time-shift between the target gene expression response and the TF activities:

$X_{\mathrm{i,t}_n} = \sum_{k \in \mathrm{TFs}} \beta_{\mathrm{i,k}} \hat{\mathrm{A}}_{\mathrm{k,t}_n - \tau}$(1.2).

Here, we are modeling the expression of gene i at time t_n as the sum of activities at time $t_n - \tau$, where t_n is the time of the nth measurement in the time-series and τ= 15 minutes is the desired time-shift. In cases where we do not have measurements for $\hat{\mathrm{A}}_{\mathrm{k,t}_n - \tau}$ we use linear interpolation to add missing data points.

The goal of our inference procedure is to find a sparse solution to β, i.e. a solution where most entries are zero. The left hand sides of Equation (1.1) and (1.2) are concatenated as response, while the right hand sides are concatenated as design variables. We use our previously described method Bayesian Best Subset Regression (BBSR) (Greenfield *et al*, 2013) to solve for β. With BBSR we compute all possible regression models for a given gene corresponding to the inclusion and exclusion of each potential predictor. For a given target gene i, potential predictors are those TFs that have a known regulatory effect on i, and the ten TFs with highest mutual information as measured by time-lagged context likelihood of relatedness (CLR) (Greenfield *et al*, 2010; Madar *et al*, 2010). Prior knowledge is incorporated by using a modification of Zellner's g-prior (Zellner, 1983) to include subjective information on the regression parameters. A g-prior equal to 1.1 was used for the combined network described in this study. Sparsity of our

solution is enforced by a model selection step based on the Bayesian Information Criterion (BIC) (Schwarz, 1978).

After model selection is carried out, the output is a matrix of dynamical parameters β, where each entry corresponds to the direction (i.e. correlated or anti-correlated) and strength (i.e. magnitude) of a regulatory interaction. To further improve inference and become more robust against over-fitting and sampling errors, we employ a bootstrapping strategy. We resample the input conditions with replacement, as well as the prior network, and run model selection on the new input. This procedure is repeated 201 times and the resulting lists of interactions are filtered to only keep those observed in more than 50% of the bootstraps.

Go enrichment analysis

The reference protein of Oryza Sativa was obtained from the Uniprot Database (http://www.uniprot.org/proteomes/UP000000763). The April, 2013 release of the Gene Ontology (http://archive.geneontology.org/full/2013-04-01/) was then queried by these ids, returning all annotations attributed to genes in the Oryza Sativa reference proteome. These annotations were propagated via the true path rule, whereby any protein with an annotation to a GO term also gains annotations for all terms that are parents of the given term, as specified by the GO hierarchy. Lastly, a separate proteome for Oryza Sativa was obtained from the Rice Genome Annotation Project (RGAP), available at: ftp://ftp.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_7.0/all.dir/. A blastp was performed, using default parameters, and for each Locus ID from the RGAP, the best-matching Uniprot ID was chosen, and the annotations transferred from that Uniprot ID to the Locus ID. Enrichment analysis of predictor targets was performed using the GOstats R package where all genes present in the network were used as background universe.

Data are available from the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/) under accession numbers GSE73609 (RNA-seq field data), GSE74793 (RNA-seq controlled chamber data), and GSE75794 (ATAC-seq data).

AUTHOR CONTRIBUTIONS

Conceptualization: OW, CH, RB, MP; Methodology: OW, CH; Software: CH, RB; Formal Analysis: CH, RB; Investigations: OW, MMHP, ABN, AP, GP; Resources: EMS, SVKJ, GBG; Data curation: CH, OW; Writing – Original Draft: OW, CH; Writing – Reviewing & Editing: RB, MP, and all authors; Visualization: CH, OW; Supervision: RB, MP; Project Administration: OW, CH, MP, RB; Funding Acquisition: EMS, RB, MP. The authors declare that they have no competing interests.

ACKNOWLEDGEMENTS

We thank Ms. Maria Pokrovskii and Dr. Emily Miraldi for their guidance on the ATAC-seq experimental protocol and analysis, Dr. Noah Youngs for sharing his working Gene Ontology annotations with us, Drs. Wenli Zhang and Jiming Jiang for sharing their nuclear isolation protocol with us, Zennia Jean Gonzaga, John Carlos Ignacio, Josefina Mendoza, Eloisa Suiton, and Junrey Amas for their help in preparation of planting, collecting and preparation of leaf samples, and Drs. Zoé Joly-Lopez and Simon Cornelis Groen for thoughtful comments on the manuscript.

REFERENCES

Anders, S., Pyl, P.T., and Huber, W. (2014). HTSeq - A Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

Arrieta-Ortiz, M.L., Hafemeister, C., Bate, A.R., Chu, T., Greenfield, A., Shuster, B., Barry, S.N., Gallitto, M., Liu, B., Kacmarczyk, T., et al. (2015). An Experimentally Supported Model of the Bacillus subtilis Global Transcriptional Regulatory Network. Mol. Syst. Biol.

Bonneau, R. (2008). Learning biological networks: from modules to dynamics. Nat. Chem. Biol. *4*, 658–664.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods *10*, 1213–

1218.

Buenrostro, J.D., Wu, B., Chang, H.Y., and Greenleaf, W.J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr. Protoc. Mol. Biol. *109*, 21.29.1–9.

Chavez, A., Scheiman, J., Vora, S., Pruitt, B.W., Tuttle, M., P R Iyer, E., Lin, S., Kiani, S., Guzman, C.D., Wiegand, D.J., et al. (2015). Highly efficient Cas9-mediated transcriptional programming. Nat. Methods *12*, 326–328.

Filichkin, S.A., Breton, G., Priest, H.D., Dharmawardhana, P., Jaiswal, P., Fox, S.E., Michael, T.P., Chory, J., Kay, S.A., and Mockler, T.C. (2011). Global profiling of rice and poplar transcriptomes highlights key conserved circadian-controlled pathways and cis-regulatory modules. PLoS One *6*, e16907.

Fu, Y., Jarboe, L.R., and Dickerson, J.A. (2011). Reconstructing genome-wide regulatory network of E. coli using transcriptome data and predicted transcription factor activities. BMC Bioinformatics *12*, 233.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017–1018.

Greenfield, A., Hafemeister, C., and Bonneau, R. (2013). Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. Bioinformatics *29*, 1060–1067.

Hoffman, M.M., Buske, O.J., Wang, J., Weng, Z., Bilmes, J.A., and Noble, W.S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat. Methods *9*, 473–476.

Huynh-Thu, V.A., and Sanguinetti, G. (2015). Combining tree-based and dynamical systems for the inference of gene regulatory networks. Bioinformatics *31*, 1614–1622.

Imam, S., Schäuble, S., Brooks, A.N., Baliga, N.S., and Price, N.D. (2015). Data-driven integration of genome-scale regulatory and metabolic network models. Front. Microbiol. *6*, 409.

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.

Kawahara, Y., de la Bastide, M., Hamilton, J.P., Kanamori, H., McCombie, W.R., Ouyang, S., Schwartz, D.C., Tanaka, T., Wu, J., Zhou, S., et al. (2013). Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice *6*, 4.

Khush, G.S. (2005). What it will take to feed 5.0 billion rice consumers in 2030. Plant Mol. Biol. *59*, 1–6.

Kilian, J., Whitehead, D., Horak, J., Wanke, D., Weinl, S., Batistic, O., D'Angelo, C., Bornberg-Bauer, E., Kudla, J., and Harter, K. (2007). The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. Plant J. *50*, 347–363.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. *14*, R36.

Kolmos, E., Chow, B.Y., Pruneda-Paz, J.L., and Kay, S.A. (2014). HsfB2b-mediated repression of PRR7 directs abiotic stress responses of the circadian clock. Proc. Natl. Acad. Sci. U. S. A. *111*, 16172–16177.

Konermann, S., Brigham, M.D., Trevino, A.E., Joung, J., Abudayyeh, O.O., Barcena, C., Hsu, P.D., Habib, N., Gootenberg, J.S., Nishimasu, H., et al. (2014). Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature *517*, 583–588.

Koryachko, A., Matthiadis, A., Ducoste, J.J., Tuck, J., Long, T.A., and Williams, C. (2015). Computational approaches to identify regulators of plant stress response using high-throughput gene expression data. Curr. Plant Biol.

Liao, J.C., Boscolo, R., Yang, Y.-L., Tran, L.M., Sabatti, C., and Roychowdhury, V.P. (2003). Network component analysis: reconstruction of regulatory signals in biological systems. Proc. Natl. Acad. Sci. U. S. A. *100*, 15522–15527.

Liu, A.-L., Zou, J., Liu, C.-F., Zhou, X.-Y., Zhang, X.-W., Luo, G.-Y., and Chen, X.-B. (2013). Over-expression of OsHsfA7 enhanced salt and drought tolerance in transgenic rice. BMB Rep. *46*, 31–36.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Maeder, M.L., Linder, S.J., Cascio, V.M., Fu, Y., Ho, Q.H., and Joung, J.K. (2013). CRISPR RNA-guided activation of endogenous human genes. Nat. Methods *10*, 977–979.

Matys, V., Fricke, E., Geffers, R., Gössling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O. V, et al. (2003). TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res. *31*, 374–378.

Mercer, T.R., and Mattick, J.S. (2013). Structure and function of long noncoding RNAs in epigenetic regulation. Nat. Struct. Mol. Biol. *20*, 300–307.

Mickelbart, M. V., Hasegawa, P.M., and Bailey-Serres, J. (2015). Genetic mechanisms of abiotic stress tolerance that translate to crop yield stability. Nat. Rev. Genet. *16*, 237–251.

Misra, A., and Sriram, G. (2013). Network component analysis provides quantitative insights on an Arabidopsis transcription factor-gene regulatory network. BMC Syst. Biol. *7*, 126.

Mittal, D., Enoki, Y., Lavania, D., Singh, A., Sakurai, H., and Grover, A. (2011). Binding affinities and interactions among different heat shock element types and heat shock factors in rice (Oryza sativa L.). FEBS J. *278*, 3076–3085.

Nagano, A.J., Sato, Y., Mihara, M., Antonio, B.A., Motoyama, R., Itoh, H., Nagamura, Y., and Izawa, T. (2012). Deciphering and prediction of transcriptome dynamics under fluctuating field conditions. Cell *151*, 1358–1369.

Ohama, N., Kusakabe, K., Mizoi, J., Zhao, H., Kidokoro, S., Koizumi, S., Takahashi, F., Ishida, T., Yanagisawa, S., Shinozaki, K., et al. (2016). The Transcriptional Cascade in the Heat Stress Response of Arabidopsis Is Strictly Regulated at the Level of Transcription Factor Expression. Plant Cell *28*, 181–201.

Olsen, K.M., and Wendel, J.F. (2013). Crop plants as models for understanding plant adaptation and diversification. Front. Plant Sci. *4*, 290.

Pachauri, R.K., Allen, M.R., Barros, V.R., Broome, J., Cramer, W., Christ, R., Church, J.A., Clarke, L., Dahe, Q., Dasgupta, P., et al. (2014). Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.

Peng, S., Huang, J., Sheehy, J.E., Laza, R.C., Visperas, R.M., Zhong, X., Centeno, G.S., Khush, G.S., and Cassman, K.G. (2004). Rice yields decline with higher night temperature from global warming. Proc Natl Acad Sci U S A *101*, 9971–9975.

Plessis, A., Hafemeister, C., Wilkins, O., Gonzaga, Z.J.Z., Meyer, R.S., Pires, I., Mueller, C., Septiningsih, E.M., Bonneau, R., Purugganan, M.D., et al. (2015). Multiple abiotic stimuli are integrated in the regulation of rice gene expression under field conditions. Elife *4*.

Prasad, P.V.V., Boote, K.J., Allen, L.H., Sheehy, J.E., and Thomas, J.M.G. (2006). Species, ecotype and cultivar differences in spikelet fertility and harvest index of rice in response to high temperature stress. F. Crop. Res. *95*, 398–411.

Redfern, S.K., Azzu, N., Binamira, J.S., Meybeck, A., Lankoski, J., Redfern, S., and Gitz, V. (2012). Rice in Southeast Asia: facing risks and vulnerabilities to respond to climate change. In Building Resilience for Adaptation to Climate Change in the Agriculture Sector. Proceedings of a Joint FAO/OECD Workshop, Rome, Italy, 23-24 April 2012., (Food and Agriculture Organization of the United Nations (FAO)), pp. 295–314.

Roy, S., Lagree, S., Hou, Z., Thomson, J.A., Stewart, R., and Gasch, A.P. (2013). Integrated module and gene-specific regulatory inference implicates upstream signaling networks. PLoS Comput. Biol. *9*, e1003252.

Schulz, M.H., Devanny, W.E., Gitter, A., Zhong, S., Ernst, J., and Bar-Joseph, Z. (2012). DREM 2.0: Improved reconstruction of dynamic regulatory networks from time-series expression data. BMC Syst. Biol. *6*, 104.

Seo, P.J., and Mas, P. (2015). STRESSing the role of the plant circadian clock. Trends

Plant Sci. *20*, 230–237.

Shiu, S.-H., Shih, M.-C., and Li, W.-H. (2005). Transcription factor families have much higher expansion rates in plants than in animals. Plant Physiol. *139*, 18–26.

Sullivan, A.M., Arsovski, A.A., Lempe, J., Bubb, K.L., Weirauch, M.T., Sabo, P.J., Sandstrom, R., Thurman, R.E., Neph, S., Reynolds, A.P., et al. (2014). Mapping and Dynamics of Regulatory DNA and Transcription Factor Networks in A. thaliana. Cell Rep. *8*, 2015–2030.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105–1111.

Tuong, T.P., and Bouman, B.A.M. (2003). Rice Production in Water-scarce Environments. In Water Productivity in Agriculture : Limits and Opportunities for Improvement. Comprehensive Assessment of Water Management in Agriculture Series, Number 1, R.E. Barker, J.W. Kijne, and D. Molden, eds. (Wallingford, Oxon, GBR: CABI Publishing), pp. 53–68.

Wassmann, R., Jagadish, S., Sumfleth, K., Pathak, H., Howell, G., Ismail, A., Serraj, R., Redona, E., Singh, R., and Heuer, S. (2009). Regional vulnerability of climate change impacts on Asian rice production and scope for adaptation. In Advances in Agronomy, D. Sparks, ed. (Elsevier), pp. 91–133.

Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., et al. (2014). Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. Cell *158*, 1431–1443.

Wilkins, O., Waldron, L., Nahal, H., Provart, N.J., and Campbell, M.M. (2009). Genotype and time of day shape the Populus drought response. Plant J. *60*, 703–715.

Wilkins, O., Bräutigam, K., and Campbell, M.M. (2010). Time of day shapes Arabidopsis drought transcriptomes. Plant J. *63*, 715–727.

Wu, Y., Zhang, W., and Jiang, J. (2014). Genome-wide nucleosome positioning is orchestrated by genomic regions associated with DNase I hypersensitivity in rice. PLoS Genet. *10*, e1004378.

Zhang, T., Zhang, W., and Jiang, J. (2015). Genome-wide nucleosome occupancy and positioning and their impact on gene expression and evolution in plants. Plant Physiol. *168*, 1406–1416.

Zhang, W., Wu, Y., Schnable, J.C., Zeng, Z., Freeling, M., Crawford, G.E., and Jiang, J. (2012). High-resolution mapping of open chromatin in the rice genome. Genome Res. *22*, 151–162.

SUPPLEMENTAL INFORMATION

Supplemental Figure S1: Plant functional measurements
Supplemental Figure S2: Multi-dimensional scaling plots of differentially expressed genes
Supplemental Figure S3: Comparison of DNase-seq and ATAC-seq data in rice
Supplemental Figure S4: TFA stability for all TFs in the network prior
Supplemental Figure S5: Convergence on a stable inferred network
Supplemental Figure S6: Expression of the targets of EPR1 in a circadian data set
Supplemental File S1: Differentially expressed genes
Supplemental File S2: RNA-seq summary statistics
Supplemental File S3: Open chromatin regions
Supplemental File S4: The network prior
Supplemental File S5: Cytoscape file of inferred network
Supplemental File S6: Network predictor statistics
Supplemental File S7: Gene Ontology enrichment results

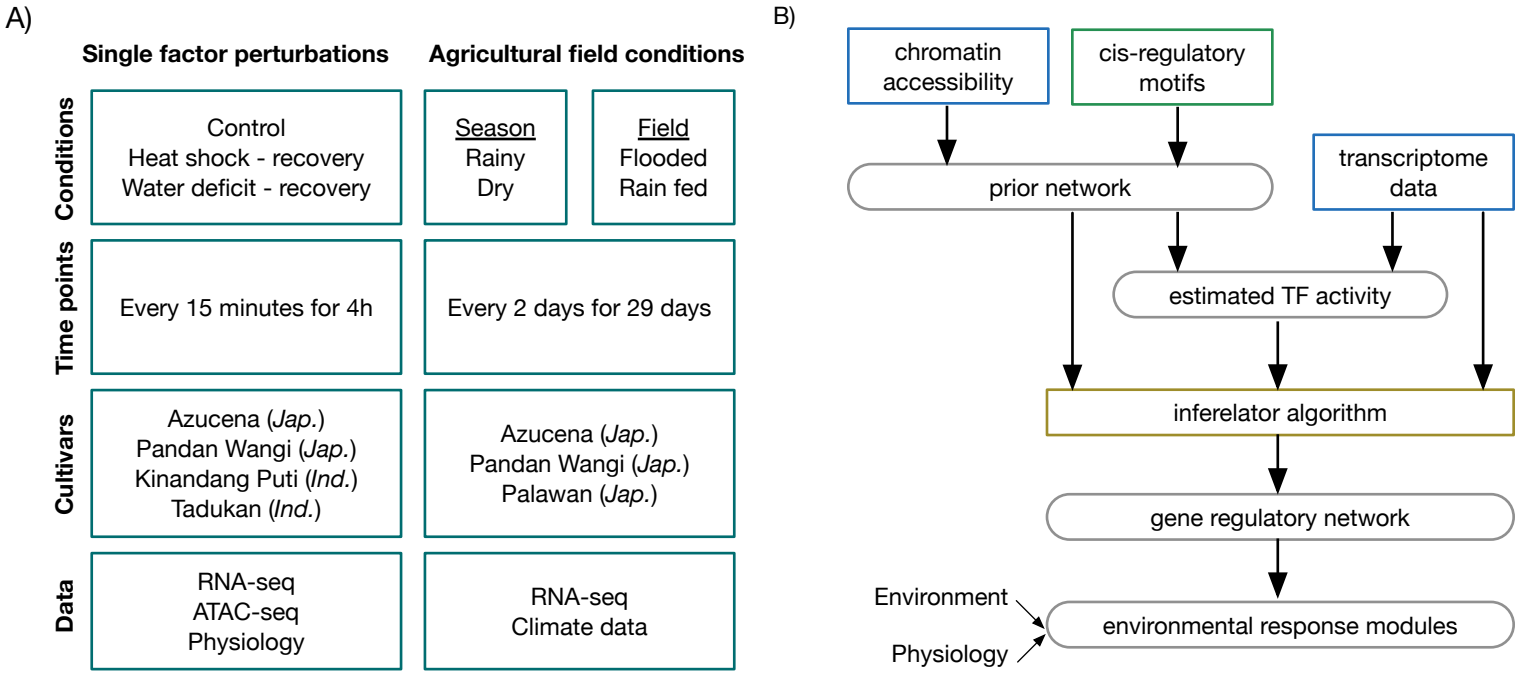


FIGURE 1: Schematic overview A) experimental design and B) data analysis.

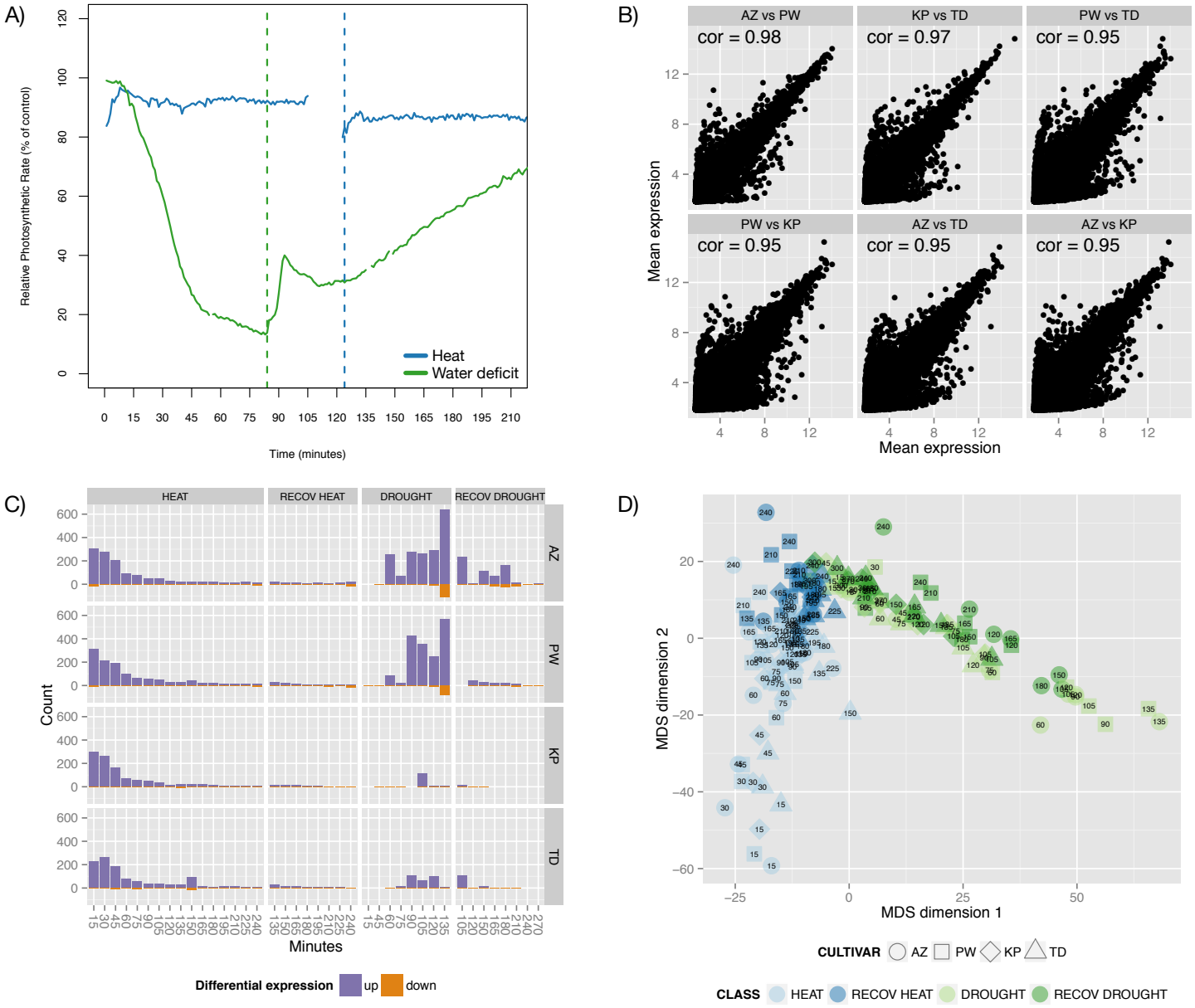


FIGURE 2. Overview of experimental data. A) Functional responses: The relative photosynthetic rate of stress treated Azucena plants is presented (n=3 for each treatment). Data fo other genotypes are in supplementary material. The vertical dashed lines indicate the end of the stress treatment and the start of the recovery treatment. B) For each cultivar we averaged the expression across control conditions for every gene. Pairwise scatter plots and Pearson correlations for Azucena (AZ) and Pandan Wangi (PW), Kinandang Puti (KP), and Tadukan (TD) are shown. C) Number of differentially expressed genes for each genotype, time point and treatment. Genes with negative fold change are shown as negative counts. D) Multi-dimensional scaling plot based on Euclidean distance of log-fold-change of 2097 differentially expressed genes are presented (see method section). The number inside each data point indicates the time in minutes since the onset of the treatment at which the sample was measured.

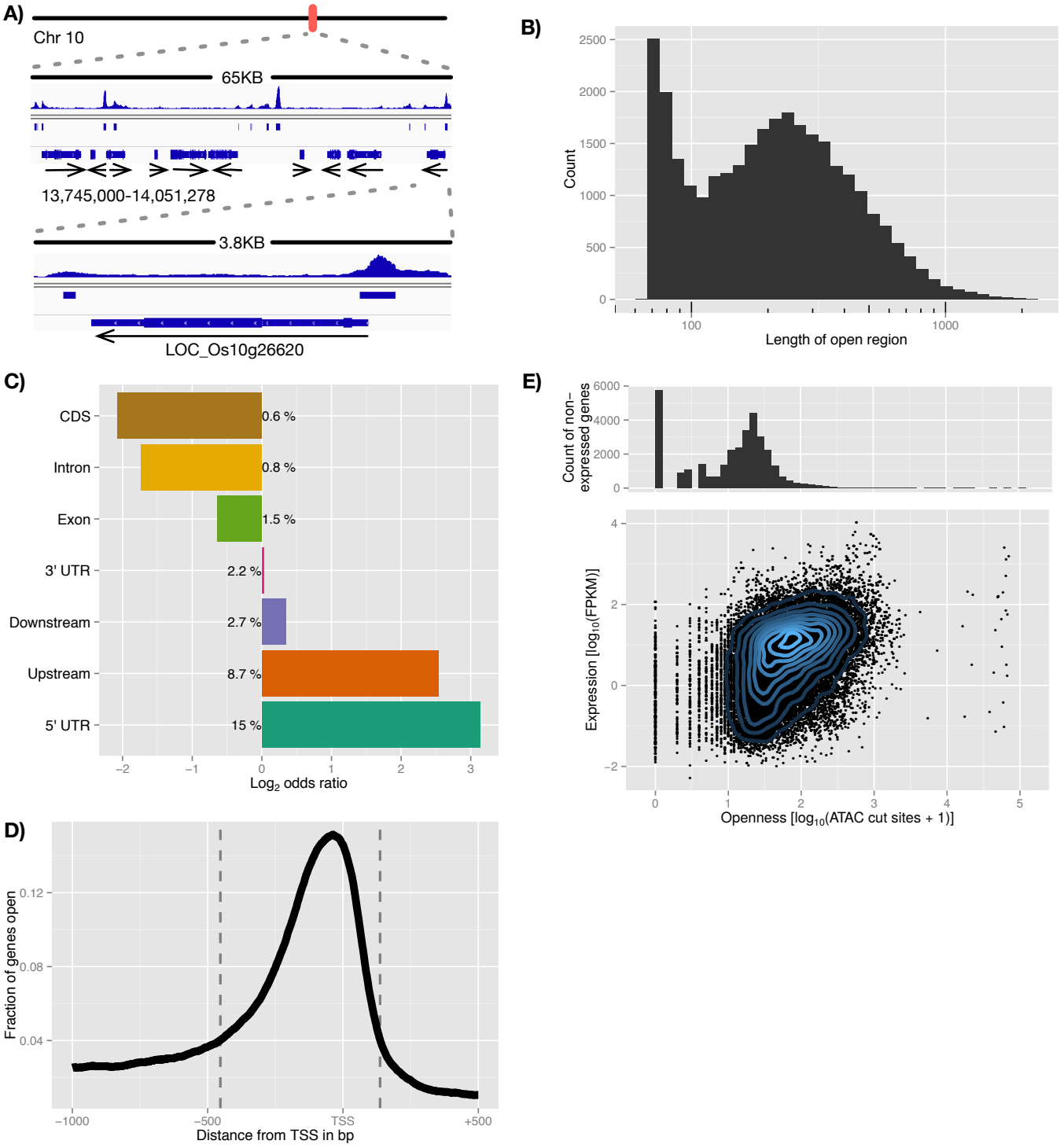


FIGURE 3. Genomic distribution of open chromatin regions identified by ATAC-seq A) The genomic location of LOC_Os10g26620, DOF Zn-finger domain containing protein, is presented as representative ATAC-seq data. Solid black lines represent regions of genomic DNA; the histograms indicate the sequencing reads that align to a given genomic region; the blue bars indicate regions that were determined to be open; the blue bars of variable width indicate the structure of gene models; the black arrows indicate on which strand the gene is encoded. B) Length distribution of open genomic regions identified by ATAC-seq. C) Location of open chromatin regions relative to gene features, including the regions 500 bp upstream of the TSS and downstream of the 3' end of the gene model. Numbers next to bars indicate what percentage of bases that belong to a specific feature fall into open regions. All odds ratios are highly significant ($p < 1e-13$). D) Distribution of open chromatin around TSSs. For all 56k genes (first isoforms only), we determined the bases that are covered by an open region in the -1000 bp to +500 bp window around the TSS. The dashed lines indicate the start (-453 bp and end (+137 bp) of the promoter region where more than 4% of the genes are open. E) Histogram of number of ATAC-seq cuts in promoter for non-expressed genes (top). Gene expression (median FPKM across all AZ samples) vs number of ATAC-seq cuts in promoter for expressed genes; correlation 0.42 (bottom).

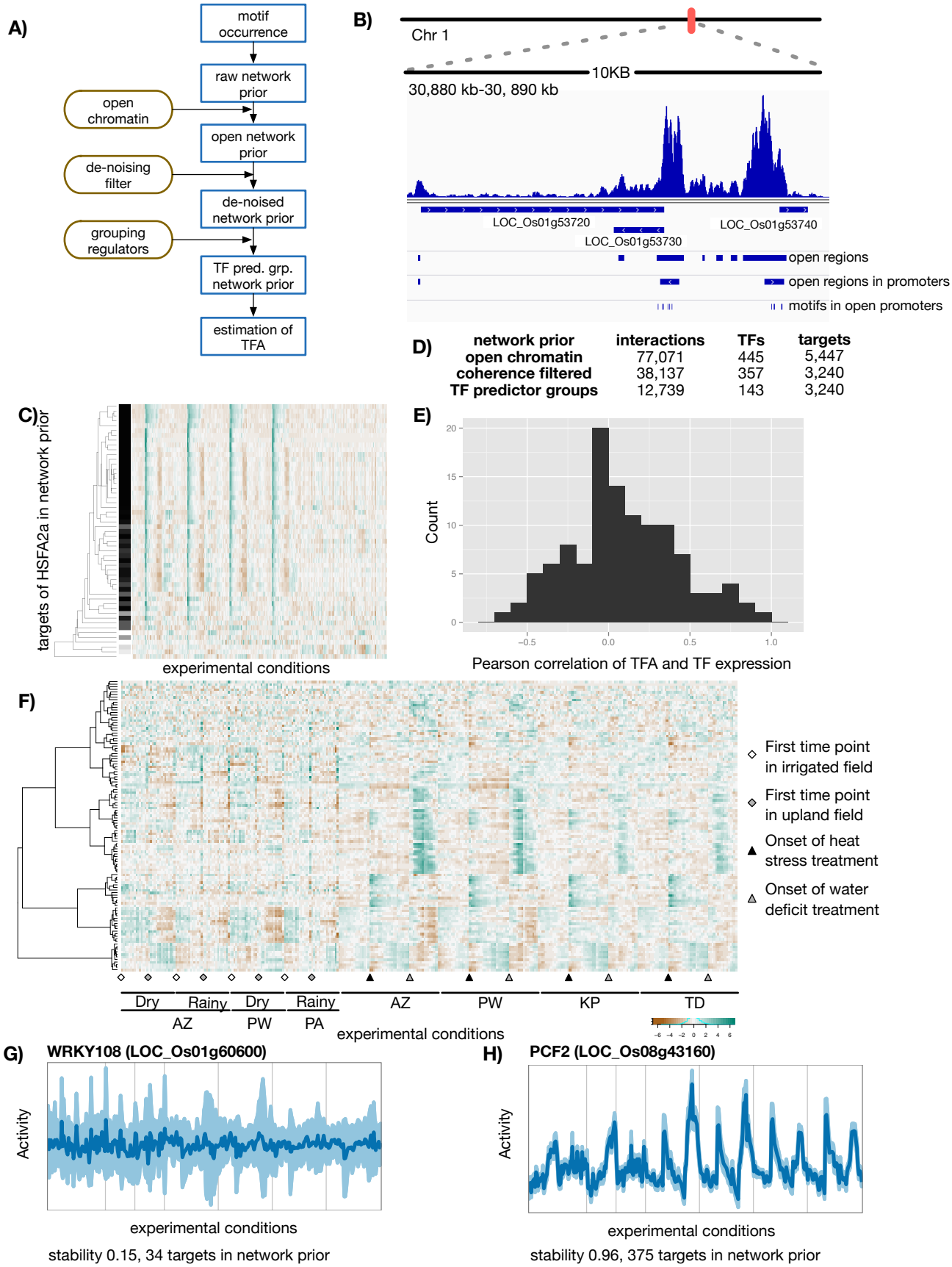
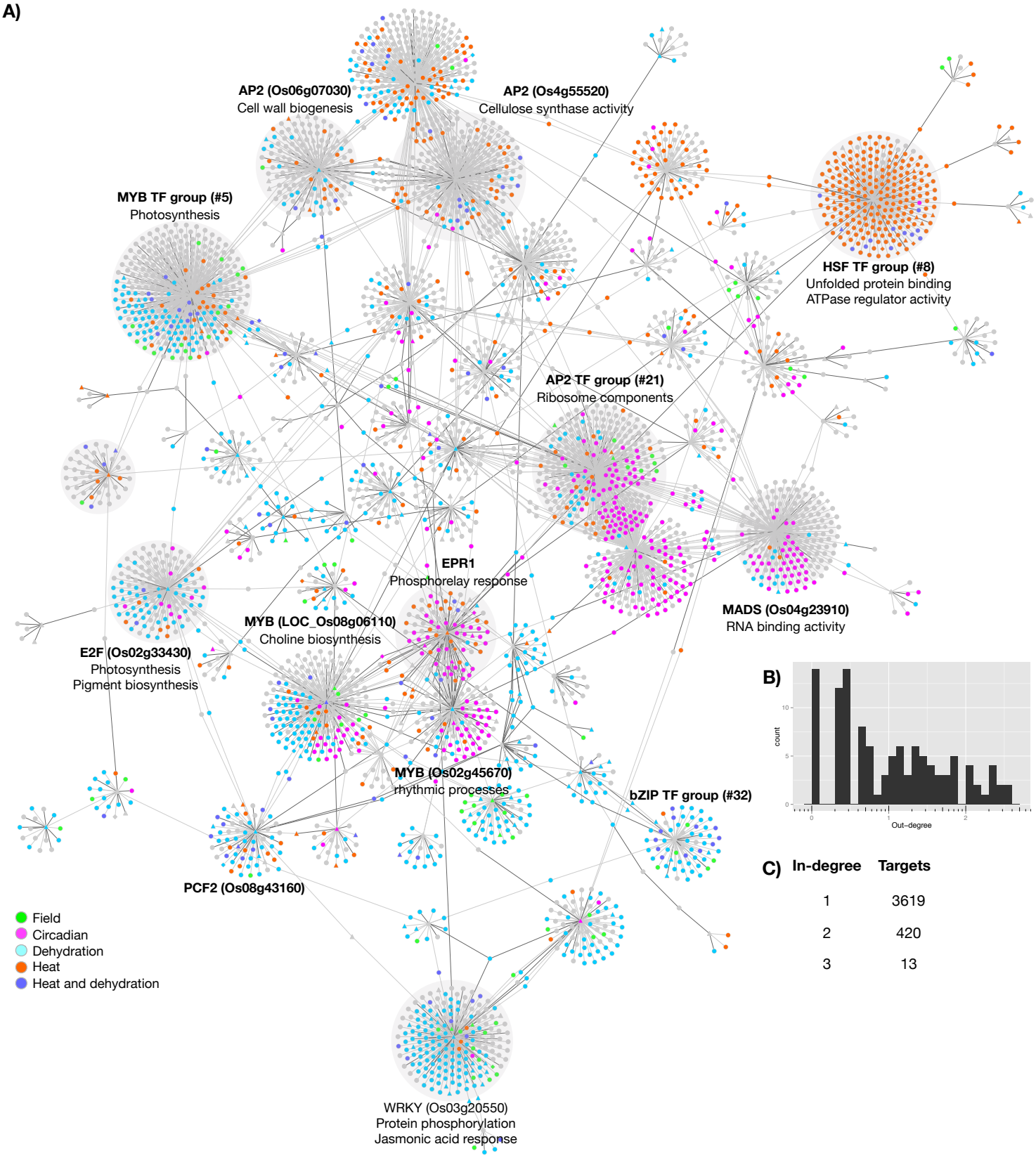


Figure 4. Prior knowledge of network structure is used to calculate TF Activity. A) Overview of the workflow for building a network prior and for estimating TFA. B) Region of chromosome 1 encoding 3 genes is highlighted to indicate how motif locations used in the network prior were identified. C) Heatmap of open chromatin prior for OsHSFA2a. The grey scale bar indicates score in the de-noising analysis - darker tones indicate higher confidence. D) Number of regulators, edges and targets in each stage of the network prior. E) Perason correlation between expression and TFA for each TF. F) Heatmap of TFAs for all TFs and TF predictor groups for which activities were calculated. G) TFA stability (quantified as correlation of TFA estimates between 201 bootstrap subsamples of the network prior) for LOC_Os01g60600 and H)LOC_Os08g41360. In these plots, the light bands shows the region between the 5th and 95th percentile of TFAs; the dark line shows the average TFA.



In-degree	Targets
1	3619
2	420
3	13

FIGURE 5. Gene regulatory network for rice leaves across environmental conditions A) The inferred network. TFs are noted with triangles; other genes are noted with circles. The colours indicate the conditions in which the genes were differentially expressed B) Histogram of network in-degree frequency C) Table of network out-degree frequency.

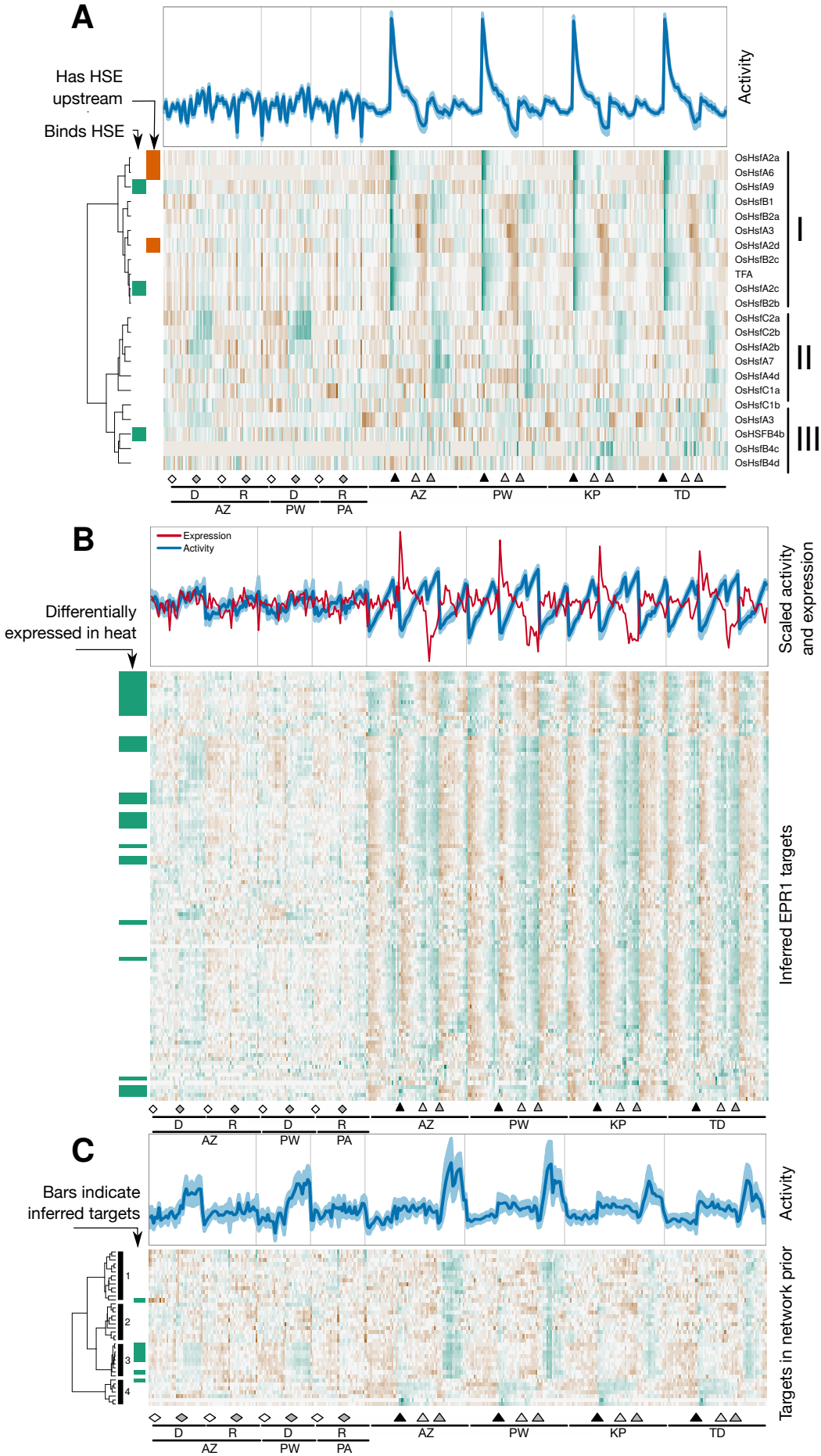


FIGURE 6. Post-hoc interpretation of inferred regulatory relationships A) TFA of the HSF predictor group (top). Heatmap of TF transcript abundance of all group members (bottom). Subgroups are noted to the right of the figure. B) TFA and transcript abundance of EPR1 (top). Heatmap of abundance of inferred EPR1 targets. C) TFA of bZIP predictor group (top). Heatmap of transcript abundance of targets in network prior (bottom).