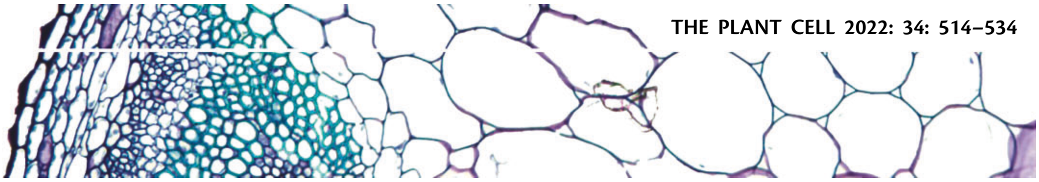
doi:10.1093/plcell/koab267

THE PLANT CELL

THE PLANT CELL 2022: 34: 514–534



Large-Scale Biology

Prediction of conserved and variable heat and cold stress response in maize using cis-regulatory information

Peng Zhou ,[1,†] Tara A. Enders ,[1,‡] Zachary A. Myers ,[1] Erika Magnusson ,[1] Peter A. Crisp ,[1,2] Jaclyn M. Noshay ,[1] Fabio Gomez-Cano ,[3] Zhikai Liang ,[1] Erich Grotewold ,[3] Kathleen Greenham [1] and Nathan M. Springer [1,*,§]

1 Department of Plant and Microbial Biology, University of Minnesota, Saint Paul, Minnesota 55108, USA
2 School of Agriculture and Food Sciences, The University of Queensland, Brisbane, QLD 4072, Australia
3 Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan 48824, USA

*Author for correspondence: springer@umn.edu
†Present address: Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China
‡Present address: Department of Biology, Hofstra University, Hempstead, New York, 11549, USA
§Senior author.

P.Z., T.A.E., P.A.C., K.G., and N.M.S. designed the research and generated the biological materials. P.Z., Z.A.M., and E.M. analyzed data. F.G.C., Z.L., and E.G. contributed analytic tools. P.Z. and N.M.S. wrote the paper.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in Instructions for Authors (https://academic.oup.com/plcell) is: Nathan M. Springer (springer@umn.edu).

Abstract

Changes in gene expression are important for responses to abiotic stress. Transcriptome profiling of heat- or cold-stressed maize genotypes identifies many changes in transcript abundance. We used comparisons of expression responses in multiple genotypes to identify alleles with variable responses to heat or cold stress and to distinguish examples of cis- or trans-regulatory variation for stress-responsive expression changes. We used motifs enriched near the transcription start sites (TSSs) for thermal stress-responsive genes to develop predictive models of gene expression responses. Prediction accuracies can be improved by focusing only on motifs within unmethylated regions near the TSS and vary for genes with different dynamic responses to stress. Models trained on expression responses in a single genotype and promoter sequences provided lower performance when applied to other genotypes but this could be improved by using models trained on data from all three genotypes tested. The analysis of genes with cis-regulatory variation provides evidence for structural variants that result in presence/absence of transcription factor binding sites in creating variable responses. This study provides insights into cis-regulatory motifs for heat- and cold-responsive gene expression and defines a framework for developing models to predict expression responses across multiple genotypes.

Received August 02, 2021. Accepted October 27, 2021. Advance access publication November 4, 2021

IN A NUTSHELL

Background: Plants have developed sophisticated mechanisms to respond and acclimate to changing environmental conditions. Transcriptional regulation can occur in trans (by transcription factors, TFs) and/or in cis (transcription factor binding sites [TFBSs] in promoters). Within a population, studies have also revealed widespread variation both at the trans- (TF abundance) and cis- level (sequence difference at TFBSs). While it is straightforward to link TF abundance with their target gene responsiveness to stress, it remains largely unknown how sequence differences in cis-elements contribute to differences in a gene's responsiveness to stress in general.

Question: Why do some genes respond to thermal stress in one genotype but not another? Is this "variable" response due in trans (abundance of its regulator TF) or in cis (promoter sequence difference)? Can we identify cis-regulatory motifs associated with stress-responsive expression patterns? Can these motifs be used to train a machine learning model to predict a gene responsiveness to stress?

Findings: Through three large-scale RNA-seq experiments, we systematically characterized the transcriptome response of maize seedlings to heat and cold stress. We identified many genes exhibiting altered expression patterns under stress between different maize genotypes, and assigned this expression variation to cis- or trans-regulatory mechanisms with the use of F_1 hybrid data. We identified motifs associated with different stress-responsive patterns and used them to develop models to predict transcriptome responses. We highlight important parameters for motif detection and modeling of expression responses. We found both potential uses, and pitfalls, in how data from one genotype can be used to predict expression responses in other genotypes. Genes with variable response due to cis-regulatory variation highlight the importance of InDels and structural variants creating polymorphisms in key motifs in different haplotypes.

Next steps: The identification of known and novel cis-regulatory elements involved in stress response allows biotechnology applications such as design of stress-responsive promoters. There is also room for substantial improvement in model construction: by synthesizing more training data (larger genotype panel, more data types) and implementing more sophisticated training algorithm (convolutional neural networks), we hope to train models better at capturing the combinatorial logics of multiple cis-regulatory elements.

Introduction

Plants are regularly exposed to variable environmental conditions throughout their life cycle and must be able to respond and acclimate to these conditions to survive and reproduce. Recent and rapid changes in climate have led to an increased frequency of extreme temperature fluctuations (Madani et al., 2018). Plants have developed sophisticated mechanisms at the cellular and metabolic levels that allow them to withstand temperature stress. In recent years, various regulatory mechanisms that involve phytohormone signaling, light signaling, circadian clock regulation and reactive oxygen species homeostasis at the transcriptional, epigenetic, and posttranslational levels have been identified during cold and heat stress (Chinnusamy et al., 2007; Nakashima et al., 2009; Mittler et al., 2012; Ohama et al., 2017; Li et al., 2018; Guo et al., 2018a; Ding et al., 2019, 2020).

Key players in the ability of plants to respond to temperature stress have been identified over the past decades. Members of the C-REPEAT-BINDING FACTOR/DEHYDRATION-RESPONSIVE ELEMENT-BINDING PROTEIN 1 (CBF/DREB1) family of transcription factors (TFs) have been identified as essential regulators of plant responses to cold (Agarwal et al., 2006) by activating both cold-regulated genes and secondary signaling pathways (Fowler and Thomashow, 2002; Shi et al., 2017; Ding et al., 2019). Similarly, the HEAT SHOCK TF (HSF) was identified as master regulators during heat stress by activating heat stress-responsive gene expression (Scharf et al., 2012). HSFs also turn on heat shock proteins that act as molecular chaperones, thus protecting cellular proteins by preventing their denaturation and aggregation, and facilitating the refolding of proteins damaged by heat (Wang et al., 2004; Busch et al., 2005; Charng et al., 2007). In addition, the APETALA 2/ETHYLENE-RESPONSIVE ELEMENT BINDING PROTEIN family and its largest subfamily—ETHYLENE RESPONSE FACTORs—participate in many developmental processes and play pivotal roles in adaptation to biotic or abiotic stresses including cold and heat stress responses (Dietz et al., 2010; Mizoi et al., 2012; Cheng et al., 2013; Hsieh et al., 2013; Licausi et al., 2013; Yao et al., 2017; Huang et al. 2021).

Several prior studies have documented gene expression changes in response to thermal stress in maize (*Zea mays*) seedlings (Li et al., 2017; Waters et al., 2017; Zhang et al., 2017; Avila et al., 2018; He et al., 2019; Hoopes et al., 2019; Frey et al., 2020). These studies have found evidence for transcriptome changes in many of the expected pathways, and identified TF genes whose expression is upregulated in response to heat or cold stress at multiple developmental stages. While substantial progress has been made in understanding some of the key TFs and TF binding sites (TFBSs) that play a role in response to heat and cold stress through studies of single varieties, the use of genetic variation within species can provide insights into the diversity of potential mechanisms by which variable cis-responses arise. Several studies have assessed variable responses to drought stress and revealed widespread cis-regulatory variation (Cubillos et al., 2014; Lovell et al., 2016; Liu et al., 2020). Analyses of maize allele-specific responses to several stress treatments also find evidence for both cis- and trans-regulatory variation (Waters et al., 2017). These studies have shown that, although there are many genes with consistent responses to abiotic stress in multiple genotypes, there are also genes with highly variable responses to abiotic stress within the same species, which often arises due to cis-regulatory variation. Promising results have been made to link cis-regulatory variation with gene expression and even plant phenotypes (Alonge et al., 2020; Kwon et al., 2020). However, it still remains largely unknown how sequence differences contribute to differences in responsiveness of promoters in general.

Machine learning approaches have provided new and powerful ways for understanding and predicting gene expression in plants (Washburn et al., 2019; Azodi et al., 2020b; Wang et al., 2020b). These approaches have been used to predict expression levels (Sartor et al., 2019; Washburn et al., 2019), regulatory architecture (Mejía-Guerra and Buckler, 2019), as well as gene expression responses to abiotic stress (Zou et al., 2011; Uygun et al., 2017, 2019; Schwarz et al., 2020; Azodi et al., 2020a). These studies highlight the potential to develop predictive models that use putative cis-regulatory motifs to predict gene expression responses to stress. However, whether these models can be applied to different genotypes other than the reference to predict consistent or variable expression response is not well understood.

We sought to investigate potential avenues for understanding transcriptome responses to heat and cold stress in several maize genotypes. We identified many genes that exhibited altered expression after a heat or cold stress event. Comparisons of inbred and hybrid genotypes revealed many examples of cis- and trans-regulatory variation that lead to varied expression responses to heat or cold stress. We mined genes with changes in transcript abundance in response to heat or cold stress to identify potential cis-regulatory motifs, which we also used to develop machine learning models to predict responsiveness of gene expression. Our findings highlight which parameters are important in motif detection and modeling of expression responses. We discovered both the potential uses and possible pitfalls in how data from one genotype can be used to predict expression responses in other genotypes. The analysis of genes with variable response due to cis-regulatory variation highlight the importance of insertion/deletion (InDel) polymorphisms or other structural variants that create presence/absence of key motifs in different haplotypes.

Results

Characterization of gene expression responses to heat and cold stress in seedling leaves of several maize inbreds

We studied the changes in the transcriptome in response to heat or cold stress in maize seedlings from three maize inbred genotypes with de novo genome assemblies (B73, Mo17, and W22) and F_1 hybrids representing all three combinations of the parental genotypes (B73xMo17, W22xB73, and W22xMo17). The specific stress treatments used in this study result in slower growth of maize seedlings but do not result in plant death or necrosis. Prior studies have shown differential responses of these inbreds to similar cold stress treatments (Waters et al., 2017; Enders et al., 2019). These differential responses were, however, subtle and likely represented quantitative variation in response to the stress rather than "tolerant" and "sensitive" genotypes. We selected the three parental genotypes due to both subtle variation in response to these stress conditions as well as the availability of high-quality genome resources. The experimental design included three biological replicates for each treatment; the specific growth conditions are described in "Materials and methods" (Figure 1, A and B). We sampled each of the six genotypes a time 0 (prior to application of the stress) as well as at 1 and 25 h into treatment for both stressed and control plants to document early and late responses to the stress at the same circadian point. We collected 126 samples and subjected their RNA to transcriptome deep sequencing (RNA-seq), generating ~30 M reads per sample (Supplemental Data Set S1). We determined per-gene expression levels based on alignments to the B73 reference genome using a variant-aware approach (see "Materials and methods" for details).

The initial analyses focused on comparison of the transcriptome data for the three parental genotypes. We assessed the quality and structure of the data by clustering using principal component analysis (PCA) with count per million (CPM) values for all genes (Figure 1C; Supplemental Figure S1). When all three genotypes were included in the same PCA, we observed a significant influence of genotype on the clustering pattern (Supplemental Figure S1A). The analysis of the different conditions for B73 (Figure 1C), Mo17 or W22 (Supplemental Figure S1, B and C) revealed very similar effects from treatments in all three genotypes. The control samples generally clustered together, suggesting relatively minor changes based on the 1-h difference in circadian sampling or one day of growth (Figure 1C; Supplemental Figure S1, B and C). The cold and heat treatments all resulted in a shift from the control samples but showed different patterns for the 1- and 25-h treatments. The effect of 1 h of cold treatment was not as pronounced as the 25-h cold treatment (Figure 1C; Supplemental Figure S1, B and C). In contrast, a 1-h heat treatment resulted in a shift that was similar to the 25-h treatment.

For each stress treatment, we identified differentially expressed genes (DEGs) by comparing the stress-treated sample with a control sample collected at the matched time point ("matched control," Figure 1D). In addition, we also identified DEGs by comparing each stress-treated sample to a control sample collected prior to the initiation of the stress treatment ("control 0 h," Figure 1D). This approach allowed us to identify genes that exhibit consistent changes

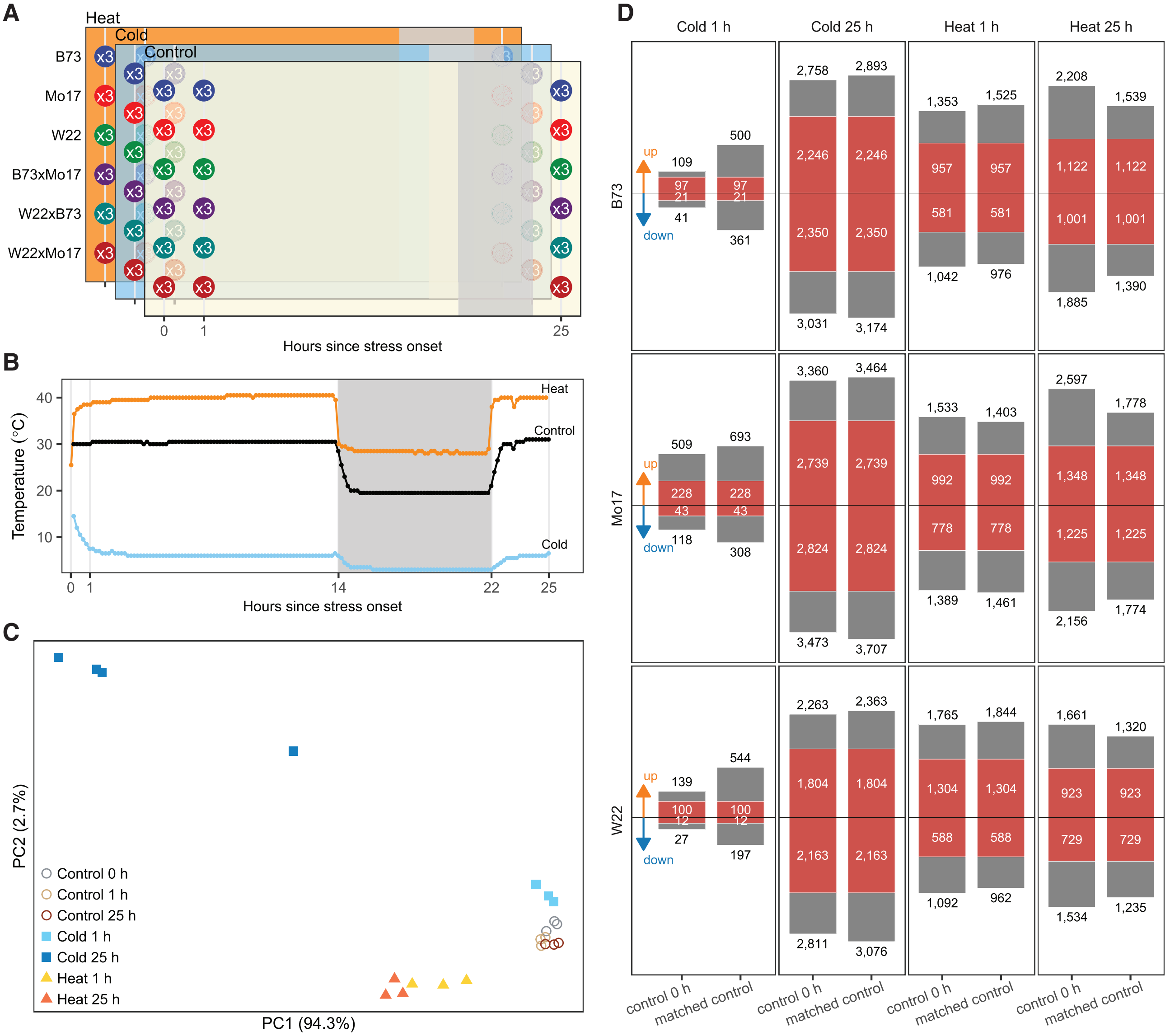


Figure 1 Experimental design and identification of DEGs in response to heat or cold stress. A, Experimental design for the generation of RNA-seq data. Three biological replicates were sampled from three maize inbreds and their F_1 hybrids at time 0 and two time points during stress. B, Temperature readings throughout the experiment, as measured from a sensor that was with the plants. The gray shaded area indicates darkness. C, Principal component clustering of B73 samples from the hybrid experiment under control, cold, and heat conditions. The treatment conditions are indicated by different symbols/colors in the plot. D, Number of DEGs under cold and heat conditions at the 1 h and 25 h time points. For each time point, the number of DEGs relative to the control sample collected at time 0 (onset of stress; control 0 h) and the number of DEGs relative to the control sample collected at the matching time point (i.e. 1 or 25 h; matched control) are shown. Numbers inside the red bars represent DEGs in both comparisons.

in expression relative to both the initial time point and to a circadian-matched sample (shown in red in Figure 1D). This approach was also important, as some genes that are deemed upregulated or downregulated based on a contrast of a 1- or 25-h treatment and a time-matched control may reflect a change in expression in the control relative to time zero and not a change induced by the treatment. We identified DEGs based on transcript abundance, which may reflect differences in transcription rate and/or in transcript stability, as both mechanisms likely affect transcript abundance following abiotic stress treatments. The subsequent analyses in this study focused on genes that show consistent differential expression relative to both time 0 and the matched time point (shown in red in Figure 1D).

The analysis of the DEGs found evidence for the expected transcriptome responses to heat and cold stress. Gene ontology (GO) analysis of all upregulated genes following heat stress identified an enrichment for terms associated with response to heat, RNA modification, protein folding, and heat acclimation (Supplemental Data Set S2). Upregulated genes following cold stress were enriched for terms associated with DNA binding, transcription regulation, calcium-binding,

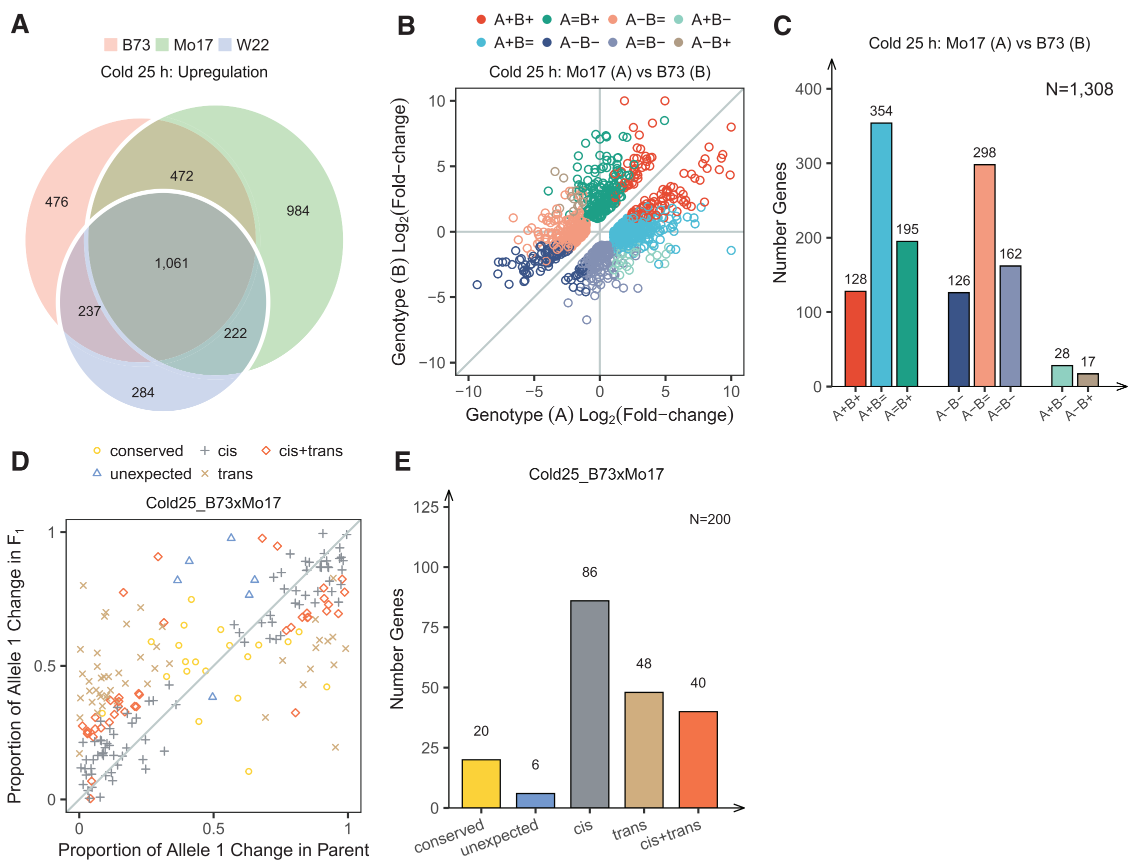


Figure 2 Characterization of genes with variable stress-responsive patterns among inbreds. A, Venn diagram showing the extent of overlap between upregulated genes in response to 25 h of cold stress for B73, Mo17, and W22. The overlap of DEGs at the other time points is shown in Supplemental Figure S1A. B, For genes that show significantly stronger (or weaker) response to cold at 25 h in B73 compared to Mo17, we show the Log_2(Fold-change[cold 25 h/control 25 h]) for both inbreds. The classification of differential responses for other genotype contrasts and timepoints is provided in Supplemental Figure S5. Posthoc tests were used to classify genes with varying differential expression between genotypes, indicated by different colors. C, Number of genes associated with each category shown in (B). For each class, the response in the two genotypes (A and B) is indicated as upregulated (" + "), downregulated ("–") or not DE (=). D, For the subset of genes classified as having a response in only one of the two genotypes that also had SNPs, we assessed allele-specific expression in the F_1 hybrid. The proportion of allele 1 (B73) change in stress versus control of the F_1 hybrid (*x*-axis) was compared to the proportion of the change in expression in the parental genotypes (*y*-axis). A maximum likelihood model was applied to classify cis and trans inheritance patterns; these classifications are shown in different colors. E, Number of genes classified into each type of regulatory pattern for response to abiotic stress shown in (D). Similar analyses for other genotypes, stress, and time points are shown in Supplemental Figure S6.

and serine/threonine kinase activity (Supplemental Data Set S2). These results confirmed that the samples exhibit the expected responses to heat or cold stress. We assessed the responsiveness of several TFs or TF families that have been previously implicated in response to thermal stress using time-course data from a separate RNA-seq experiment. The maize B73 genome encodes 29 HSF TFs (Yilmaz et al., 2009; Zhang et al., 2020), 11 of which were upregulated in response to heat in at least one of the two time points (Supplemental Figure S2, A–C). Many of the HSFs classified as DEGs exhibited very rapid increases in transcript levels (some with >100-fold rise in 30 min) (Supplemental Figure S2). We also noticed that several *HSF* genes (*ZmHSF4*, *ZmHSF12*, *ZmHSF13*, and *ZmHSF20*) not classified here as DEGs showed a very strong activation 30 min into treatment, but had already returned to basal expression levels by 1 h, and were thus not considered as differentially expressed (DE) in our replicated experiment (Supplemental Figure S2D). We also assessed expression patterns for a set of 104 TFs previously reported to play a role in cold stress response in various cereal species (Baillo et al., 2019): we determined that the expression of 40 of these TFs is induced in our replicated B73 dataset and exhibit a variety of patterns in the time-course data (Supplemental Figure S3). Together, the analysis of GO terms and TFs provide evidence that our heat/cold treatments elicited the expected transcriptional responses.

Genetic variation for responses to cold and heat stress

We were interested in exploring the frequency, and underlying causes, of variable expression responses to abiotic stress in maize seedlings based on comparisons of the transcriptome responses in the three inbreds. We used those nonredundant (nr) DEGs (relative to both time 0 and a matched time point) at 1 or 25 h of heat or cold stress in at least one of the inbred lines to classify conserved and variable responses. The results for 25 h cold upregulation are shown in Figure 2; the full set of responses to all treatments is shown in Supplemental Figures S4–S6. While some genes displayed consistent upregulation in all three genotypes, we also observed many examples of upregulation in only one genotype (Figure 2A; Supplemental Table S3; Supplemental Figure S4A). In general, 30%–50% of the nr DEGs from all three genotypes were DE in two or three parental genotypes, with the remaining DEGs being specific for one of the three parental genotypes (Supplemental Table S1). Hierarchical clustering of upregulated genes in one or two of the parental inbreds indicated that many of these genes show minimal expression changes in some genotypes (Supplemental Figure S4B). However, this analysis also revealed many examples of genes that respond in multiple genotypes but are simply not classified as significantly DE in some genotypes (Supplemental Figure S4B). To identify the set of genes with a robust response to stress in a subset of genotypes, we introduced an interaction term (genotype:condition) to model the genotype-specific condition effect under the generalized linear framework of DESeq2 to assess the responses of genes (Figure 2, B and C; Supplemental Figure S5). This modeling approach allowed us to classify genes that exhibit significant upregulation in one genotype but not the others (A + B= or A = B+), as well as a set of genes that respond more strongly in one genotype but do exhibit responses in both sets of genotypes (A + B+) (Figure 2, B and C).

Our experimental design allowed us to characterize the variation in cold- or heat-responsive expression in the three maize inbreds and to separate cis- and trans-acting regulatory variation using their F_1 hybrids. We grew the F_1 hybrids at the same time as the inbred parents; they exhibited very similar responses to the treatments based on clustering (Supplemental Figure S1, D–F). We proceeded to assess allele-specific expression for the sets of genes that have a significant expression response in one genotype but not in another (A + B=, A = B+, A – B=, A = B–). For these genes, we asked whether the responsiveness was due to cis-acting features present in one allele or to trans-acting effects that might influence responsiveness of both alleles in the F_1 hybrid. A subset (11%–37%) of these genes with variable responses contained single-nucleotide polymorphisms (SNPs) and had sufficient allele-specific read-depth to perform allele-specific expression analysis in the F_1 hybrid (Figure 2, D and E; Supplemental Figure S6; Supplemental Table S2). Using a model that incorporates allele-specific expression in the F_1 and relative expression levels in the two parents (see “Materials and methods” for details), we sorted their regulatory patterns as cis only (24%–54%), trans only (0%–29%), or as a mix of cis and trans (14%–43%) (Figure 2, D and E; Supplemental Figure S6; Supplemental Table S2). It is important to note that these classifications of cis/trans regulation refer here to the responsiveness to the stress condition rather than regulatory variation within a specific environment. These cis/trans classifications revealed many examples of genes with different gene expression responses to heat or cold stress in these three hybrid genotypes. The allele-specific data identified many cases of both cis- and trans-regulatory variation that controls the differential response to stress. The examples of cis-regulatory variation suggested that changes in the promoter sequences for one allele alter the ability to respond to abiotic stress, which prompted us to assess whether we could identify sources of this cis-regulatory variation or predict different responses using only the promoter sequence. It is worth noting that the number of genes identified with cis-regulatory variation here is likely an underestimation, since many genes with differential response did not have SNPs within their coding region, which thus precluded us from performing allele-specific expression analyses.

Identification of enriched motifs for each cluster of heat/cold stress-induced genes

One objective of this study was to develop models to predict gene expression responses to heat/cold stress in maize. DNA motifs that provide TF TFBSs are expected to provide a significant portion of the input information that influences gene expression responses. We identified significantly enriched motifs for heat or cold by assessing sequences of all significant DEGs using the STREME algorithm (Bailey, 2020). We used a variety of different potential “promoter” sequence space parameters, including different lengths, directions relative to the transcription start site (TSS), and chromatin filters, to increase the potential to identify enriched motifs (Figure 3, A and B). In all cases, we assessed enrichment of sequence motifs through a comparison to a control set of expressed genes that does not show evidence of differential expression after heat or cold stress.

We collapsed the resulting full set of 1,188 motifs from the B73 reference genome into 419 motif groups based on sequence similarity, 110 of which were assigned to known TFBSs in the catalog of inferred sequence Binding Preferences (cis-BPs) (Weirauch et al., 2014), while the remaining 309 did not correspond to previously described motifs. An assessment of the top 40 most significantly enriched motifs from the search that used ±2-kb sequences for each set of DEGs revealed that a subset of the enriched motifs reflects previously characterized TFBSs, while others represented novel motifs not captured by the current collection of plant TFBSs (Figure 3C). The motifs that represented known TFBSs included some of the expected sites such as HSF binding sites for heat or CBF/DREB binding sites for cold (Figure 3C). For the subsequent application of using enriched motifs as features in a predictive model, it was not critical that all motifs be valid, as the models can be trained to utilize the most predictive features, but it was reassuring to note that we identified many of the expected motifs in the top set of features.

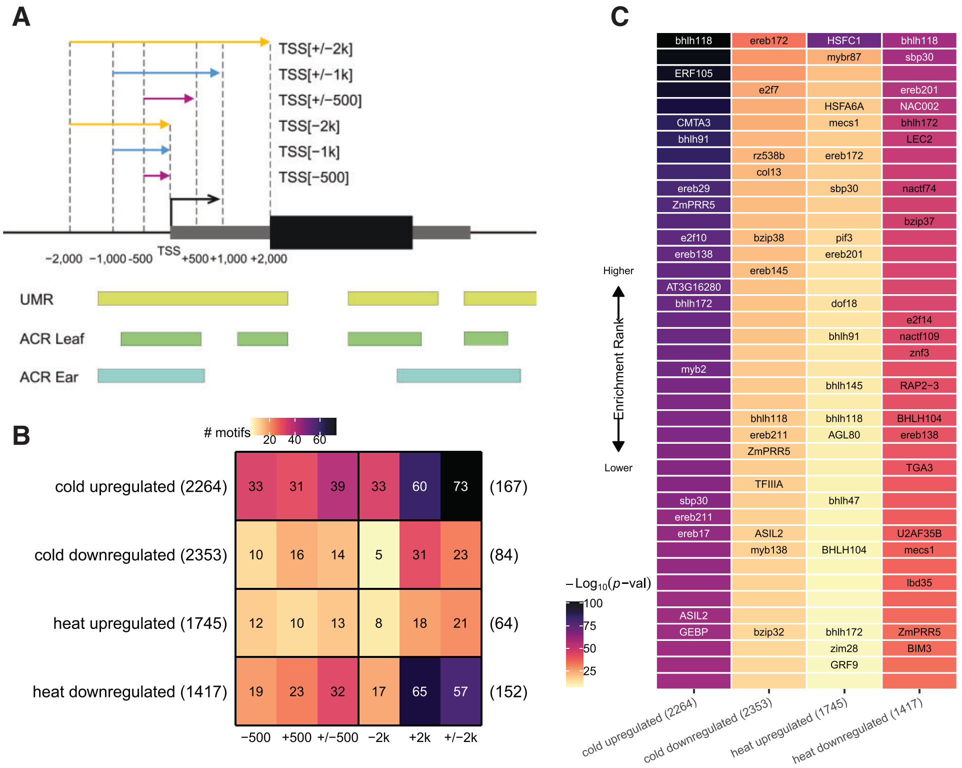
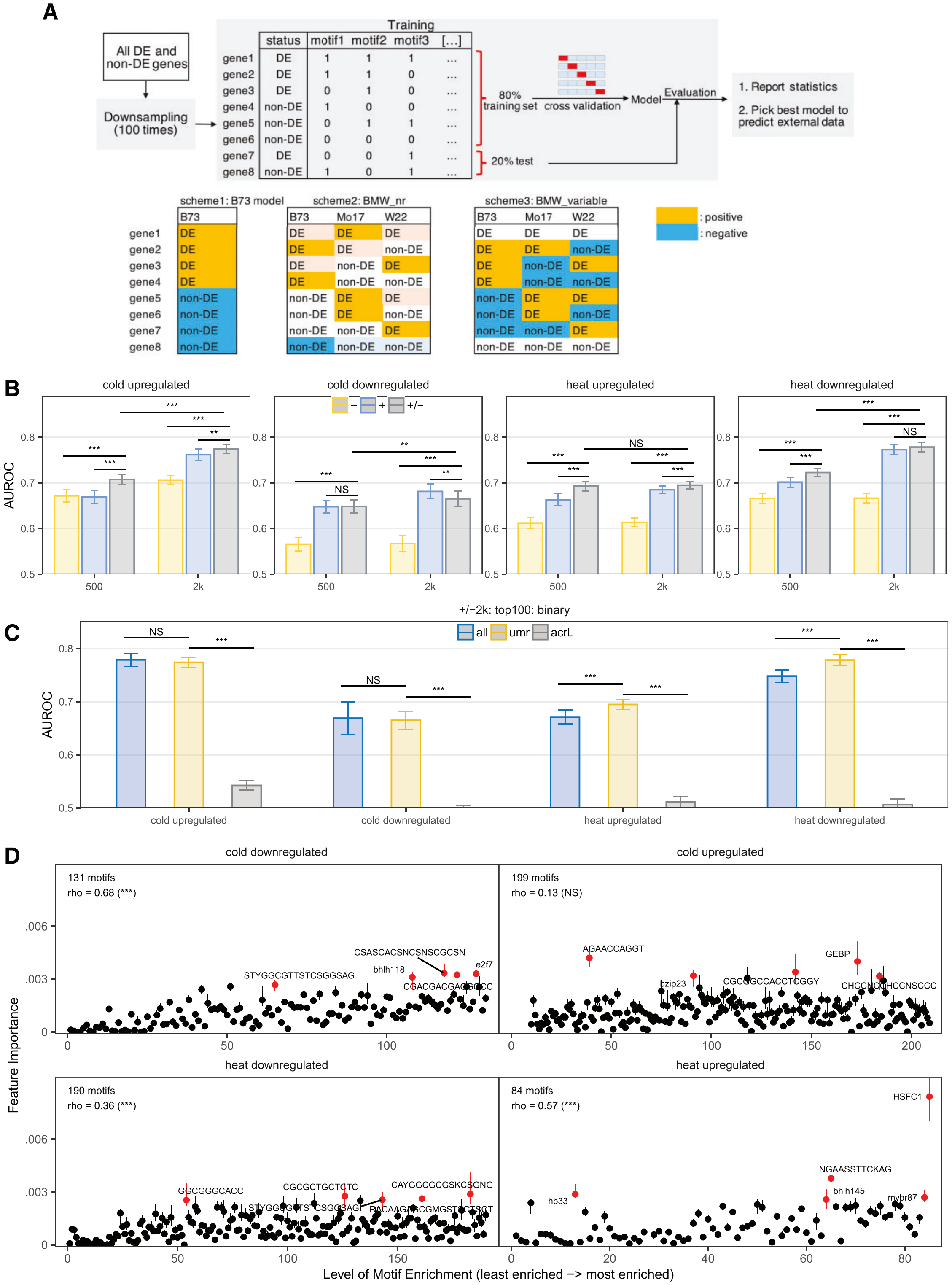


Figure 3 Identification of enriched motifs in cold- and heat-responsive genes. A, Varying potential "promoter" sequence spaces were used to search for motifs enriched in different sets of genes. The schematic diagram indicates a representative gene with the TSS indicated. The potential regions include different lengths of sequences upstream the promoter [Q](−500 bp, −1 kb, −2 kb) as well as sets of sequence that include both upstream and downstream sequence (i.e. ±500 bp). In addition, for each of these potential regions, we also subsetted the sequence to only include regions that are unmethylated (UMRs) or that are classified as accessible based on ATAC-seq analysis (ACR, accessible chromatin region) in leaf or ear tissue (Ricci et al., 2019). B, Number of nr motifs found using different B73 promoter spaces (*x*-axis) and different DEG sets (*y*-axis). Numbers include motifs identified in all sequence contexts ("all genomic," "UMR," "ACR Leaf," and "ACR Ear"). Darker colors indicate more motifs identified, with the exact numbers marked in each cell. Numbers in parentheses to the left of the heatmap indicate the number of genes used for motif mining; numbers to the right indicate the total number of nr motifs found for each set of genes. C, Top 40 enriched motifs identified in each set of DEG calls include known TFBSs as well as novel motifs. For each set of DEGs, up to the top 40 most enriched motifs found using B73 promoter space are shown (*P*-value for enrichment is indicated by color). If the enriched motif matched a previously characterized TFBS (Pearson's correlation coefficient > 0.8), the name of the TF is shown. In each case, there was a mixture of previously characterized motifs and novel motifs.

Generation of models to predict heat- and cold-responsive gene expression

We sought to generate and assess predictive models of B73 stress-responsive expression using the potential cis-regulatory elements (motifs) as "features" to classify a gene's expression response to heat or cold stress and assess whether the presence of sequence motifs might be used to accurately predict stress response. We used presence/absence of the motifs within different sets of search spaces (Figure 3A) to develop random forest models to assess how the use of different potential "promoter" regions would affect model performance. We implemented a previously described approach (Zou et al., 2011; Uygun et al., 2017, 2019; Azodi et al., 2020a) to utilize the motif features to predict whether genes will exhibit cold- or heat-responsive expression (Figure 4A). We developed separate models for the upregulated and downregulated genes for each stress. For each set of nr responsive genes from both time points, we also included an equivalent number of nonresponsive expressed genes that are not classified as DE (Figure 4A—Scheme 1). We performed the sub-sampling of balanced numbers of DE and nonDE genes 100 times and each time we divided the genes as 80% for model training and 20% for model testing (Figure 4A; see "Materials and Methods" for details).

Overall, we were able to achieve moderate accuracies for predicting which genes would respond to heat (0.63–0.75) or cold stress (0.57–0.78) in B73, with higher accuracies for cold upregulated genes (area under the receiver operating characteristic [AUROC]: 0.68–0.78) and heat downregulated (AUROC: 0.69–0.75) genes (Figure 4B). In general, the use of sequences both upstream and downstream of the TSS (±2 kb, ±500 bp) provided higher prediction accuracies than when using the upstream or downstream sequences alone (−2 kb or + 2 kb, −500 bp, or + 500 bp). Similarly, using longer sequences surrounding the TSS (±2 kb) provided higher prediction accuracies than the use of only ±500-bp sequences (Figure 4B). For subsequent analyses, we assessed several



	status	motif1	motif2	motif3	[...]
gene1	DE	1	1	1	...
gene2	DE	1	1	0	...
gene3	DE	0	1	0	...
gene4	non-DE	1	0	0	...
gene5	non-DE	0	1	1	...
gene6	non-DE	0	0	0	...
gene7	DE	0	0	1	...
gene8	non-DE	1	0	1	...

scheme1: B73 model

	B73
gene1	DE
gene2	DE
gene3	DE
gene4	DE
gene5	non-DE
gene6	non-DE
gene7	non-DE
gene8	non-DE

scheme2: BMW_nr

B73	Mo17	W22
DE	DE	DE
DE	DE	non-DE
DE	non-DE	DE
DE	non-DE	non-DE
non-DE	DE	DE
non-DE	DE	non-DE
non-DE	non-DE	DE
non-DE	non-DE	non-DE

scheme3: BMW_variable

B73	Mo17	W22
DE	DE	DE
DE	DE	non-DE
DE	non-DE	DE
DE	non-DE	non-DE
non-DE	DE	DE
non-DE	DE	non-DE
non-DE	non-DE	DE
non-DE	non-DE	non-DE

Figure 4 Training scheme and performance (AUROC) evaluation of different machine learning models predicting cold and heat-responsive expression. A, Training workflow (upper) and three training schemes (below). B73-only training scheme: DE and nonDE genes in B73 are highlighted as positive (orange) and negative (blue), respectively. BMW nr training scheme: a gene showing DE in ⩾1 genotypes is labeled positive, while a DEG in all three genotypes is labeled as negative. Nr, nr, meaning that for each gene triplet, promoter sequence was only picked from one of the responsive (dark orange) or nonresponsive (dark blue) genotypes. BMW variable genes training scheme: only genes showing variable responsive pattern were kept for training, where promoter sequences from the responsive genotypes are labeled positive (orange) and those from the nonresponsive genotypes labeled negative (blue). B, Performance comparison of models trained using small promoter spaces (−500 bp, + 500 bp, ±500 bp) against larger promoter spaces (−2 kb, + 2 kb, ±2 kb). In each case, the same chromatin filters (UMRs) and number of features (top100) were used for training. Within each stress-responsive category (e.g. cold upregulated), performance comparisons were made between the model

additional parameters while using motifs within 2 kb of the TSS. We tested whether only using motifs present within unmethylated regions (UMRs) or accessible chromatin might improve prediction accuracy compared to using all sequences within the ±2-kb window (Figure 4C). Indeed, the prediction accuracies increased significantly for three (cold up, cold down, and heat downregulated genes) of the four groups tested when focusing only on the motifs within UMRs, even though substantially less sequence was used (Figure 4C). In contrast, the use of motifs only within accessible regions, as defined in a prior study of maize seedling tissue (Ricci et al., 2019), exhibited significantly lower performance for nearly all of the groups of genes (Figure 4C).

We assessed whether the most highly enriched motifs were also the most predictive features. We thus compared the motif enrichment ranks (least enriched to the most enriched) with the feature importance of that motif in the predictive models for each of the sets of genes responding to heat or cold stress (Figure 4D). We generally observed a positive correlation between feature importance score and motif enrichment rank, such that on average the motifs with a higher enrichment rank had higher feature importance (Spearman correlation between 0.13 and 0.68, Figure 4D). This observation was particularly true for the genes that are upregulated in response to heat, for which the HEAT SHOCK TF C1 (HSFC1) motif was the most enriched and also had the highest feature importance score in models (Figure 4D). While we obtained an overall positive correlation for the other groups of genes, the most highly enriched motif was not necessarily highly informative and some of the more informative motifs in the model did not rank highly for enrichment (Figure 4D). This observation suggested that while the level of enrichment is generally correlated with predictive power, there are exceptions when the most enriched motifs often are not the most predictive features for determining responsive gene expression.

We developed the models described above to predict genes that exhibit significant upregulation or downregulation based on data from two time points. We were curious if the models would display variable performance based on the dynamic pattern of expression changes during a time course of stress response. We thus generated an unreplicated time-course dataset that sampled gene expression at nine time points in the three inbred lines (Supplemental Figure S7A). Hierarchical or t-distributed stochastic neighbor embedding (t-SNE) clustering based on the expression levels of all genes revealed patterns that are consistent with genotype and type of treatment (Supplemental Figure S7B). Since this data were not replicated, we were not in a position to define DEGs but instead used this data to classify the dynamic pattern of response for the genes that were previously classified as significant DEGs in the initial replicated experiment at 1or 25 h. We then used these significant DEGs from each genotype to define co-expression clusters based on their time-course differential expression profiles (see “Materials and methods” for details) (Figure 5). We identified a set of co-expression clusters that exhibit upregulation or downregulation of their constituent genes in response to heat or cold. These clusters included examples of early or late responses, as well as transient responses and stable responses (Figure 5A). We used the models developed for all upregulated or downregulated genes to predict responses in each of these subsets of co-expression clusters. Co-expression clusters showed better predictions in the model trained for the same directional response, as expected (Supplemental Table S5). There was variation for whether the genes within specific co-expression clusters exhibited higher prediction accuracies than all upregulated or downregulated genes. A subset of co-expression clusters had significantly higher performance than all upregulated or downregulated genes, but in other cases, the performance was similar or lower (Figure 5B; Supplemental Table S3).

Association of variable response to heat or cold stress with model predictions or features

In the sections above, we focused on generating predictive models based on the B73 reference genome and on changes in expression in the B73 background. An important long-term objective is the ability to develop models that can predict conserved or variable responses to abiotic stress in different genotypes. We thus sought to assess how several factors might influence predictions of response to heat or

Figure 4 (Continued)
using promoter sequences both upstream and downstream of the TSS (i.e. “ + /−”) and the models using only upstream or downstream sequences (“−“ and “ + ”) using t test followed by multiple test correction (“Benjamini and Hochberg”) with significance levels indicated (**P* < 0.05; ***P* < 0.01; ****P* < 0.001). Within each group an extra test was made between the “ + /−500 bp” model and the “ + /−2 kb” model. C, Performance comparison of models using all genomic sequence (“all”), UMR regions only (“umr”) or using leaf-accessible regions only (“acrL”) with the same sized promoter spaces (±2 kb) and number of features (top100). In each setting, average AUROC (*N* = 100 downsampling and model trainings) is shown along with the standard deviation. Within each stress-responsive category, performance comparisons were made between the “umr” model and the other two models (“all” and “acrL”) using *t* test followed by multiple test correction with significance levels indicated. D, Relationship between motif enrichment level and feature importance score in different categories of stress-responsive genes. All significantly enriched motifs were assessed for each group of DEGs. Motif enrichment levels were determined by a hypergeometric test, using motif occurrences in positive and negative gene sets and ordered from least significant to most significant (*x*-axis). Permutation-based feature importance scores from 100 random forest models are shown on the *y*-axis with error bar indicating 25%–75% quantiles. The top five feature importance scores are labeled with motif names (known motif) or consensus sequences (novel motif). In each panel, the Spearman’s correlation coefficient (rho) between feature importance score and motif enrichment rank is reported and the level of significance is also indicated. All feature importance score estimates are based on the models trained using “ + /−2 kb,” “UMR,” “top200,” and “binary” parameters.

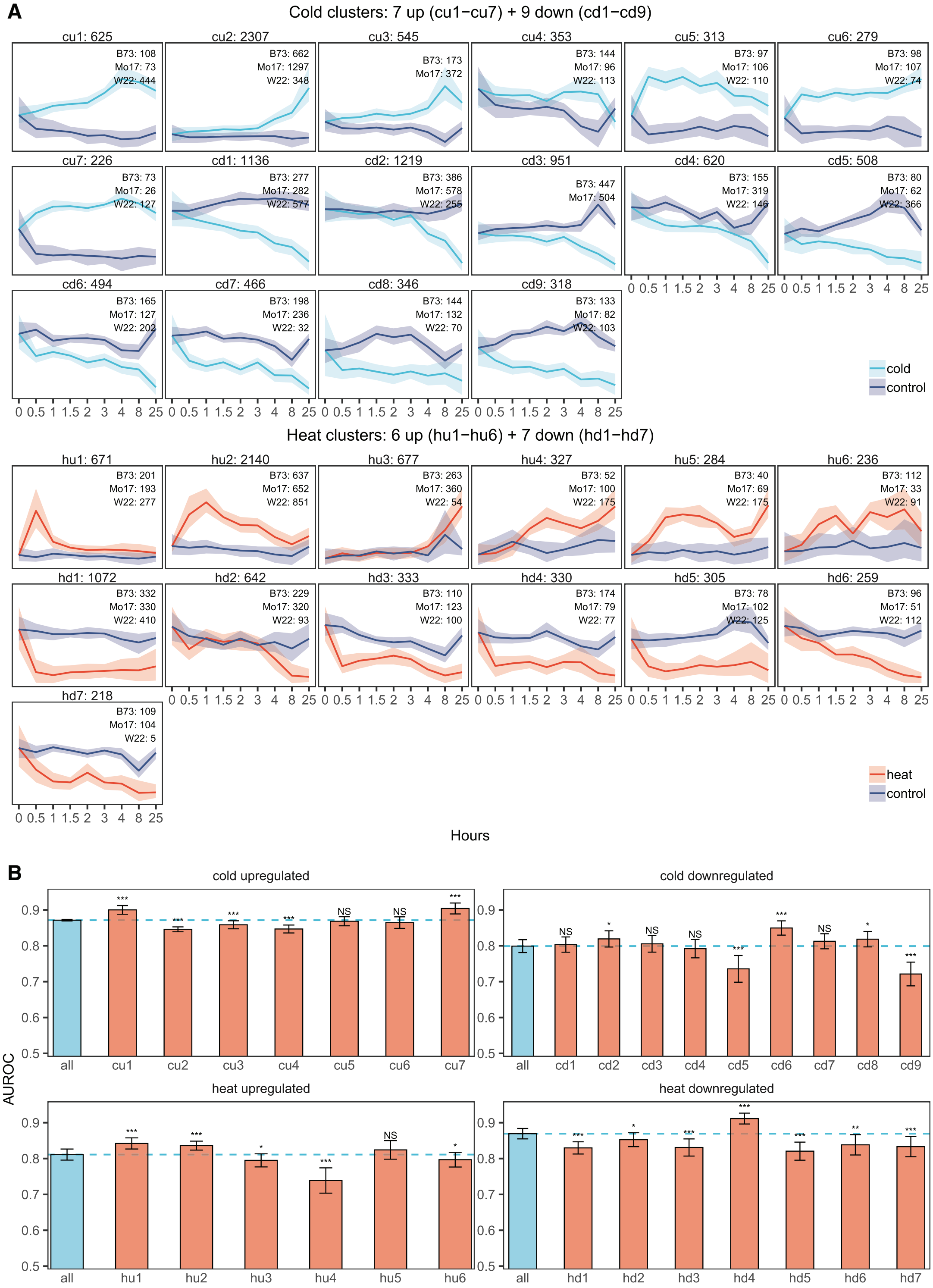


Figure 5 Identification and prediction of cold- and heat-responsive gene clusters. A, Expression profiles of cold- and heat-responsive gene clusters. B73, Mo17, and W22 genes that exhibit significant differential expression 1 or 25 h into stress treatment were used to perform co-expression clustering based on their time-course expression pattern (see "Materials and methods"). The median expression level of control and stress conditions for the genes within each module is shown and the number of B73, Mo17, and W22 genes in each module is listed on the right. The shaded area at each time point represents 25%–75% quantile expression levels. B, Model prediction accuracy (AUROC) on all stress-responsive DEGs and different co-expression clusters (as defined in (A)). Trained models (cold-up, cold-down, heat-up, and heat-down) were used to predict different co-expression clusters using promoter sequences. The best performing model in each stress-responsive gene category was used for prediction. Evaluation datasets were downsampled 100 times to achieve balance; the final mean and standard deviation of AUROC scores are reported. Prediction accuracies (AUROC scores, see Supplemental Data Set S1 for all metrics) on specific co-expression clusters were compared with model accuracy on "all" DEGs in each category with the level of significance reported after multiple test correction ("Benjamini & Hochberg" adjustment; $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$).

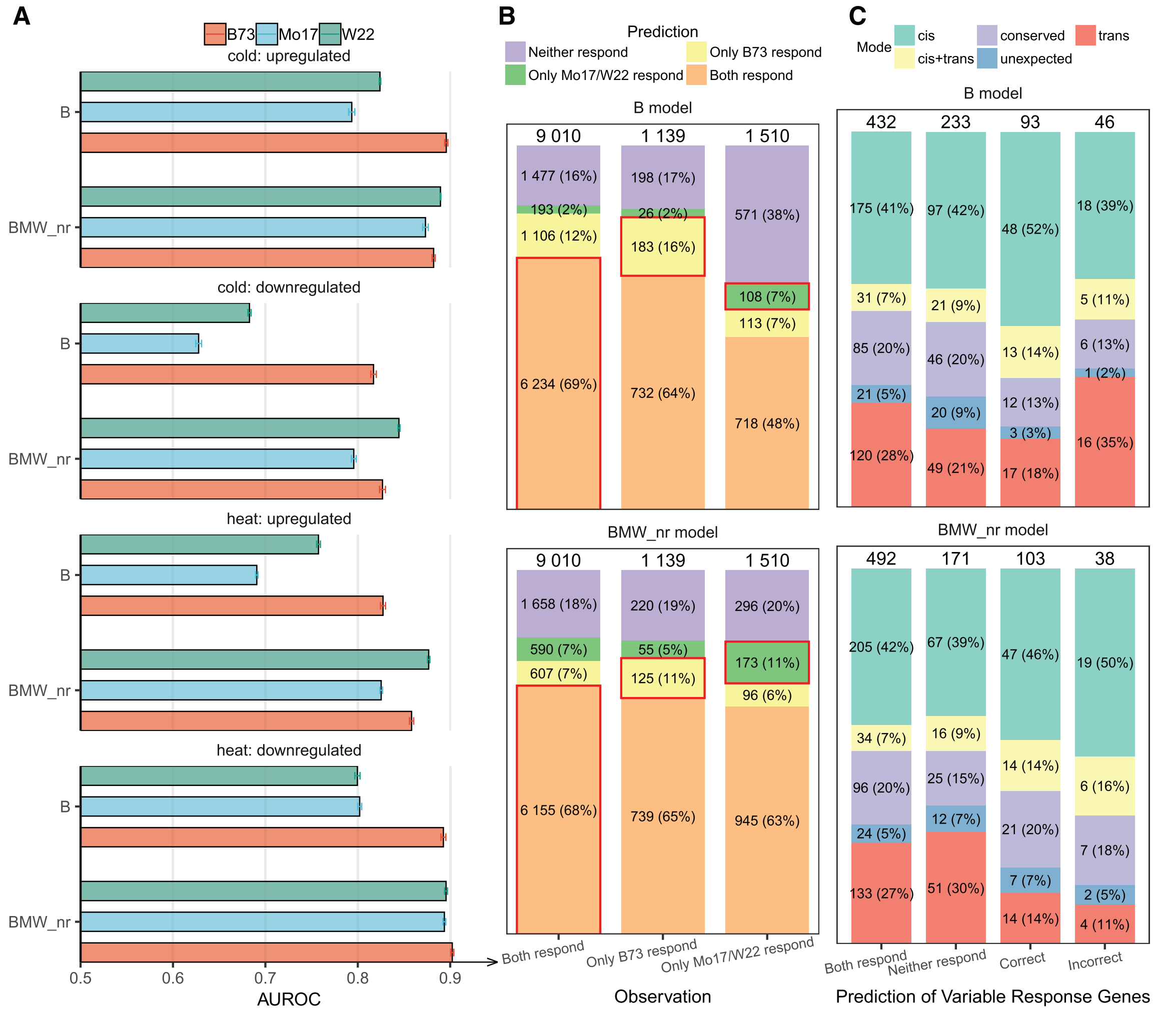


Figure 6 Cross-genotype performance of machine learning models predicting cold or heat-responsive expression. Models were trained only using B73 sequence and DE labels ("B73 model") or data from all three genotypes after redundancy removal ("BMW_nr model"). A, AUROC for models predicting stress-responsive expression in B73, Mo17, and W22. Average AUROC (N = 100 model permutations) is shown along with the standard deviation for both the B73 and BMW_nr models. B, Model prediction accuracy for genes showing consistent ("Both respond") or variable ("Only B73 or Mo17/W22 respond") response patterns among genotypes. In each observed category, the number and proportion of predictions are indicated in the plot with the correct predictions highlighted with red boxes. C, Dissection of regulatory patterns for genes showing variable response patterns among genotypes. Variable response genes were first grouped by whether model prediction agrees with the observed status ("Correct" if the model correctly predicts one genotype responds but the other does not, "Incorrect" if the model predicts oppositely, "Both respond" and "Neither respond" if the model predicts both or neither genotypes respond—although in reality only one genotype responds). Then within each group, the number of proportion of different regulatory patterns ("cis," "trans") were marked.

cold stress across multiple genotypes (Figure 6). The initial models described above were trained using B73 DEGs and motifs (Figure 4A—Scheme 1). These models performed more poorly when applied to the DEGs and using the presence of motifs in the promoter sequence from the other inbreds Mo17 or W22 (Figure 6A). However, even in these cases, we obtained AUROC scores >0.7 for most groups of DEGs in these other two genotypes, suggesting that information from B73-based models can be useful for predicting responses in other genotypes. We then trained models that utilized DEGs and significantly enriched motifs from all three genotypes (Figure 4A—Scheme 2), as well as control genes from each genotype. To remove redundant responses, we randomly selected only one ortholog if two or three genotypes exhibited differential expression for the same gene; we refer to this model as the BMW (B73, Mo17, and W22)-nr model. This approach offered increased prediction accuracies for all Mo17 and W22 responses, such that all three genotypes reached roughly similar accuracy (Figure 6A). These results suggested that while there is some accuracy in untrained predictions from one genotype to others, the accuracies are substantially improved by training using data from all genotypes.

While these overall accuracies were high, we further tested the ability of these models to correctly predict variation for responses to heat or cold stress. We were interested in documenting how accurately we might predict genes with stress-responsive expression in one genotype but not

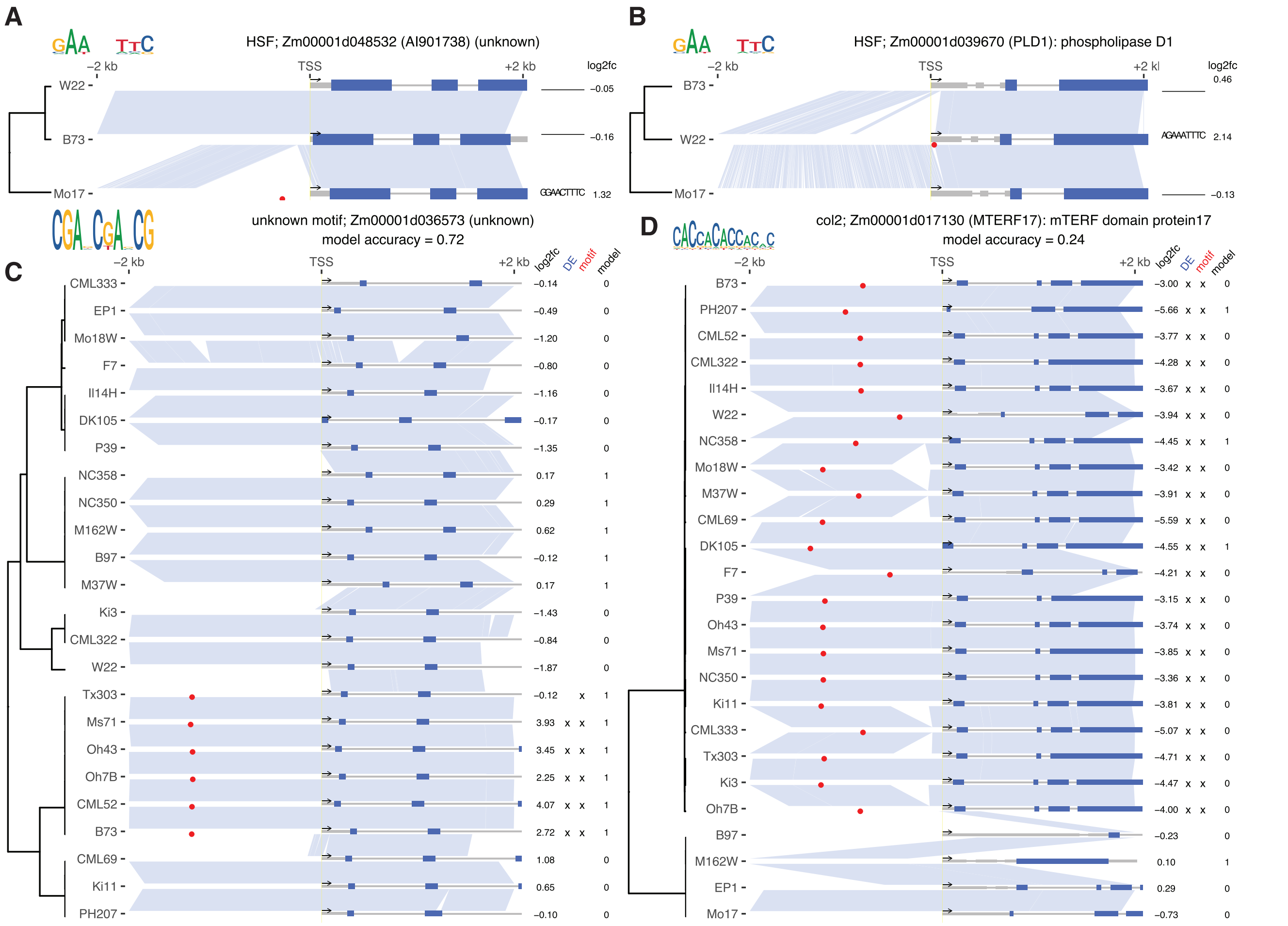


Figure 7 Identification of TFBS variation associated with variable stress-responsive expression patterns. A and B, Presence of an HSF TFBS (GAANNTTC motif) in the promoter region that is associated with activation of the target gene. C and D, Presence/absence of a TFBS motif associated with the Log_2(Fold-change) under cold treatment for a panel of 25 maize genotypes. In each panel, the upstream and downstream 2-kb sequences around TSS were extracted for each genotype, aligned to each other, and plotted as (light blue) synteny blocks. Gene structures were also plotted for each genotype (blue boxes represent exons and gray segments represent introns/untranslated regions). Motif logo is shown on the top left of each panel, and the presence of the motif in the 4-kb regions was determined using find individual motif occurrences (FIMO) and marked as red dots. A phylogenetic tree derived from the multiple sequence alignment is shown on the left. On the right of the alignment are the actual aligned sequences in the motif as well as four additional columns showing the Log_2(Fold-change) of each genotype under cold treatment ("log2fc"), whether differential expression was called for the genotype ("DE"), whether a full motif was detected in this region ("motif") and whether the machine learning model predicts a stress response for this genotype ("model").

another. We used the classification of the relative responses in multiple genotypes (Figure 2C; Supplemental Figure S3) and assessed the prediction for these genes relative to the observed patterns (Figure 6B). We obtained similar results using either a model trained only on B73 or a nr BMW set of DE genes (Figure 6B). The majority of genes that exhibit conserved transcript abundance responses in both genotypes were correctly predicted to respond in both genotypes, with relatively few of these genes predicted to not respond in either genotype. The correct prediction accuracies (highlighted in red boxes) were much lower when focusing on the subset of genes that only responds in B73 or in another genotypes (Figure 6B).

Regardless of which model we used, we determined that most genes with genotype-specific responses are incorrectly predicted to respond in both genotypes (Figure 6B). Since our predictions were largely based on the presence of putative cis-regulatory motifs, it might be expected that prediction accuracies would be higher for genes with cis-regulatory variation for responsiveness when compared to genes with trans-regulatory variation. We thus assessed the classification of cis- and trans-regulatory variations for genes with variable prediction in each group (Figure 6C). Genes with the correct predictions did show slight, but not significant (NS), higher proportions of cis-regulation (cis versus noncis; $P = 0.207$, Fisher's exact test) than incorrectly predicted genes and were depleted for trans-regulation (trans versus nontrans; $P = 0.036$, Fisher's exact test), as expected. However, the enrichments were fairly subtle and many genes with cis-regulatory variation were still predicted incorrectly. We also attempted to develop models that only use genes with variable response and training based on allelic variation

(Figure 4A—Scheme 3). However, we were not able to generate informative models, likely due to the relatively low amount of useful data available for training these models.

TFBS variation associated with differences in stress-responsive gene expression

We sought to utilize the enriched motifs and/or model predictions to understand the sources of variable responses for different haplotypes. There is substantial variation, including both SNPs and structural variants (InDels and transposon element insertions) between maize haplotypes that often make it difficult to identify specific causal variants. The upregulated genes in response to heat stress exhibited a significant enrichment for the presence of an TFBS for an HSF (Figure 3C). We identified 21 genes with cis-regulatory variation for response to heat that we correctly predicted in all three genotypes by the BMW-nr model. Of these, 15 genes had an HSF TFBS within 1 kb of the TSS and we sought to assess whether their observed variation for responsiveness correlated with variation in the presence or position of the HSF TFBS (Supplemental Table S4). For five genes, the presence/absence of an HSF TFBS correlated with response or lack of response. In each of these cases, the HSF TFBS presence/absence was the result of an InDel rather than an SNP changing the sequence (Figure 7, A and B). Another seven genes had at least one HSF TFBS in all three inbred alleles. However, in several cases, the allele that did not respond to heat stress had an InDel resulting in a shift of the HSF TFBS further away from or closer to the TSS.

To assess the sources of variation for response to cold stress, we generated an additional dataset consisting of a single replicate of control and cold-stressed plants at two time points (1 and 25 h) for 23 maize genotypes with genomic resources, including SNP calls relative to B73. We included an additional replicate of B73, Mo17, and W22 in this panel, which all exhibited expression changes that are consistent with the initial replicated transcriptome analysis and supported the classification of variable responses for these genes (Supplemental Figure S8A). We focused on the set of 2,147 (1,088 upregulated and 1,059 downregulated) genes that exhibit variable response to cold in the three initial genotypes. To identify examples of clear classification as “response” or “no response” to cold, we assessed the distribution of the ratio (Log_2) of transcript abundance in cold relative to control to identify genes with clear bi-modal distributions. We identified a subset of 518 genes with significant bi-modal or multimodal distributions for which genes can be classified as responding to cold or not responding (Supplemental Figure S8, B and C). Based on the selection criteria of these genes, at least one of the three core genotypes (B73, Mo17, and W22) was always within both the responding and nonresponding group, as expected.

To identify potential sequence variants associated with response to cold stress, we aligned the genomic sequences of the 25 maize genotypes to B73 to extract bi-allelic variants within 2 kb of each gene to perform a local association test with the cold response pattern (i.e. responding and nonresponding genotypes). We used the PLINK toolset (Purcell et al., 2007) to perform a standard chi-square association test for each gene. Many of the genes (299/518) showed at least one significant association for a local variant and the expression response. We also assessed whether enriched cold-responsive sequence motifs with presence/absence were highly associated with the variable expression response. We identified 9% (47/529) of genes with a motif that has a >90% accuracy for prediction of the response and 25% (130/529) of genes with a motif with a >80% response accuracy (Supplemental Data Set S4). Most (110/130) of these latter genes that can be predicted based on the presence/absence of a motif were also identified by PLINK as having a significant local association. Of the 130 genes with >80% association between a motif and the expression response, only 13 (10%) illustrated a case where motif variation was caused by SNP variants only. The other genes all included InDels, suggesting widespread contribution of InDels to variation for responsiveness. We highlighted two examples of variable responses (Figure 7, C and D). The gene Zm00001d036573 exhibited significant upregulation in B73, but not in W22 or Mo17. We detected an unknown motif (CGANCGANCG) in all alleles of this gene that are classified as responding to cold (Figure 7C). This motif was also present in genotype Tx303, even though the expression of this gene in this genotype did not show responsiveness to cold treatment. The overall accuracy for prediction of response to cold for this gene was 0.72 when using the BMW-nr model. The incorrect predictions were largely due to predicted responses in several genotypes that failed to respond. Another interesting example was a maize mitochondrial transcription termination factor (mTERF17—Zm00001d017130; Figure 7D), which showed significant downregulation upon cold stress in all maize genotypes except Mo17, B97, EP1, and M162W. The presence of a col2 motif (CACCACACCACNC) appeared to be highly associated with the expression response. The four genotypes that failed to respond showed substantial structural variation and did not share homology to the other haplotypes in the region containing the col2 motif. However, the BMW-nr model had quite a limited accuracy in its prediction for this gene (0.24). The model only predicted a response for a small set of the genotypes classified as responding to cold. Several other examples are highlighted in Supplemental Figure S9, A and B and demonstrate the potential role of both SNPs and structural variants in creating presence/absence of motifs associated with expression response changes.

Discussion

Understanding gene expression responses to abiotic stress will contribute to efforts to develop more resilient crop varieties. In this study, we focused on monitoring transcriptome responses to heat and cold stress in the leaves of maize seedling. The use of paired heat and cold stress with the same sampling times and genotypes revealed some key differences in the dynamics of response. Many of the responses to the heat stress event used here occurred very rapidly and were fairly transient. In contrast, the cold stress employed here tended to have less drastic effects in the first hours, but included changes in expression that continued, and often strengthened, over the course of 24 h. Including two different abiotic stresses with variable patterns of response provides an opportunity to compare approaches that use cis-regulatory features to predict response to abiotic stress.

A key long-term goal is to understand how cis-regulatory elements, and variation among them, drives variation for gene expression responses to thermal stress in different genotypes of crop species. This goal has implications for understanding the evolutionary sources of regulatory variation for stress responsiveness. These approaches also may provide the ability to identify genotypes with particular responses to abiotic stress. The analyses performed in this study provide insights into the changes in transcript abundance that occur in response to heat and cold stress, the applications of predictive models of gene expression responses as well as the potential opportunities and associated risks when attempting to predict expression responses across genotypes.

Identification of TFs and TFBSs associated with response to heat and cold stress

Environmental stresses induce significant differential gene regulation, and these large-scale changes likely include examples in which activation of a relatively small set of stress-responsive TFs can activate, or repress, a much larger number of stress-responsive genes (Reményi et al., 2004; Song et al., 2016; Vihervaara et al., 2018). We identified a number of TFs whose encoding genes exhibited increased transcript abundance following heat or cold stress, including a subset activated very early upon exposure to stress (Figure 5; Supplemental Figures S2 and S3). In many cases, we also observed connections between the upregulated TF genes and some of the TFBSs that were enriched in the promoters of upregulated genes. Genes upregulated by heat were significantly enriched for both HSF TF genes and the previously identified HSFC1 binding site (Franco-Zorrilla et al., 2014) in the early activation cluster, as well as in the full set of heat upregulated DEGs. Among cold upregulated genes, we observed a significant enrichment for both WRKY TF genes and several previously identified WRKY binding sites, as well as a significant enrichment for both MYB TF genes and several previously identified MYB binding sites (Franco-Zorrilla et al., 2014). Members of each of these TF families have demonstrated roles in stress responses. HSF TFs are thoroughly characterized regulators of heat stress responses (Scharf et al., 2012), while members of both WRKY and MYB TF families have been implicated in cold stress responses (Chen et al., 2012; Li et al., 2015).

"Promoter" definitions influence ability to identify motifs and predict responses

Our understanding of the architecture of regulatory regions in plants remains limited (Long et al., 2016; Weber et al., 2016). It is tempting to use a simple criterion such as the 1 kb of sequence immediately upstream of the core promoter when searching for potential regulatory elements. However, there is evidence that important regulatory elements can be further upstream or located in regions within the gene or downstream of the gene (Jeong et al., 2006; Laxa, 2016; Weber et al., 2016; Gallegos and Rose, 2019). Models that attempt to predict changes in transcript abundance in response to stress likely should utilize motifs that may predict transcriptional changes as well as motifs within the RNA that influence the stability of transcripts. We used a variety of different parameters to identify or filter the sequences that were used to discover enriched motifs or predict expression responses. We noted several important take-away messages from our attempts to document enriched motifs or predict responses to stress in B73. The use of larger regions and the inclusion of sequences within the transcribed region resulted in the discovery of more motifs, which was expected. More importantly, we also discovered that using these larger regions also improves the prediction accuracy of the models. However, analyses that only utilized UMRs suggested that prediction accuracies are not solely based on the amount of sequence used. Our findings also suggest that Assay for Transposase-Accessible Chromatin followed by sequencing (ATAC-seq) data from control tissues is not particularly useful for finding regions important for predicting responses.

The functional binding of TFs often requires both the presence of a matching motif (TFBS) and a proper chromatin state (Weirauch et al., 2014). Recent studies on accessible chromatin or chromatin modifications have documented many potential cis-regulatory elements in maize (Oka et al., 2017; Ricci et al., 2019); however, surveys of potential regulatory elements in unstressed plants have likely missed key potential stress-induced regulatory elements. The location of regulatory elements can be inferred through analyses of accessible chromatin, especially when focused on regions that are accessible only following stress treatments (Maher et al., 2018; Han et al., 2020; Raxwal et al., 2020). Alternatively, DNA methylation signatures are stable across tissues, developmental stages, and environmental conditions, and can provide effective filters in mining functional regulatory elements (Crisp et al., 2020). Predictive models that only used the presence of motifs within UMRs or accessible regions performed better for some groups of genes, even though this filter removed $>50\%$ of the sequence that is methylated. In contrast, the use of regions defined as accessible chromatin in a previous study (Ricci et al., 2019) did not improve prediction accuracies. Recent work suggested that UMRs may provide a catalog of potential regulatory elements in plants (Crisp et al., 2020). Many regions that contain stress-specific regulatory elements are likely unmethylated even in control conditions, while many of the regions of accessible chromatin may be more dynamic and only become accessible under specific conditions (Zeng et al., 2019; Parvathaneni et al., 2020; Wang et al., 2020a).

Predictions of response to the environment across genotypes

A key goal in this work was to investigate the variation in response to abiotic stress in different maize genotypes and to assess approaches for predicting this variation. Local adaptation of plant populations likely involves changes to the cis-regulatory elements that allow for gene expression responses to environmental challenges. There are open questions about the nature of molecular variants that will create cis-regulatory changes in responsiveness to stress. SNPs within critical motifs may change the response, but are more likely to result in loss of a response rather than a gain of responsiveness. Alternatively, structural variants such as deletions or transposon insertions may provide novel elements or change the spacing between potential cis-regulatory elements and the TSS. We identified hundreds of genes that exhibit a response to abiotic stress in one genotype but not another. The cis-regulatory variation for response to stress indicated that one allele contains the elements necessary for response while the other does not. Further studies of these examples will provide insights into the molecular basis for changes in response to the environment.

The availability of genes with documented variation for response to heat or cold stress offered an opportunity to assess how well we can predict variation within germplasm for these responses. Ideally, we would be able to use data from a single genotype to effectively predict expression responses in other genotypes, as the value of predictive models is partially due to the reduction in which data need to be collected. However, we obtained substantially lower prediction accuracies when models trained solely on B73 expression responses were used to predict Mo17 or W22. If we use expression responses from all three genotypes, we achieved much higher prediction accuracies for the other genotypes. A truer estimate of the accuracy for cross-genotype predictions arose from focusing on the ability to accurately predict genes with variable response. While we were able to predict some of these examples of variable response, the rates were relatively low (~10%–20%). Most genes with variable expression responses between genotypes were predicted to respond in both genotypes. For a subset of the genes, we used the allele-specific expression data to document cis- and trans-acting regulatory variation. We expected to see that the model would make accurate predictions for genes with cis-regulatory variation at a much higher rate than genes with trans-acting regulatory variation. We saw minor enrichments for cis-acting regulatory variation and depletions for trans-acting regulatory variation in the genes with correct predictions, but there were examples of correct predictions for trans-acting regulatory prediction as well. The alleles with trans-regulatory variation likely still contain motifs necessary for the response, even if the trans-acting factor might be absent, which may lead to predicted responses even when the trans-acting factor is missing. We also considered generating models that were trained only using genes that have a variable response to stress in which the negative control set reflected the alleles that did not respond to stress; however, these models would not have included enough examples for proper training of models. Substantial improvements in cross-genotype prediction accuracies for genes with a variable response would be needed before this approach can be valuable in generating information to inform breeding decisions.

A closer examination of variation for motif presence/absence associated with variable responses for heat or cold highlights the potential role of structural variants. For both heat responsiveness and cold responsiveness, we obtained many more examples of InDels that caused presence/absence of the TFBS rather than SNPs or small sequence changes that might influence binding. The high levels of structural diversity among maize haplotypes may create substantial variation for motif presence in different alleles. It is important to understand the molecular sources that drive variation for responses in gene expression as we seek to predict potential variation for response to the environment.

Materials and methods

Plant materials and experimental design

Three experiments were carried out to study the response of maize seedlings to cold or heat stress treatments. The first experiment included three biological replicates (each sample was a pool of tissue from 3 to 4 individual seedlings) for three inbred parents (B73, Mo17, and W22) as well as their F_1 hybrids (B73xMo17, B73xW22, and Mo17xW22) at three time points under control, cold, and heat conditions. A second experiment utilized a time-course to assess cold and heat response at nine different time points (0, 0.5, 1, 1.5, 2, 3, 4, 8, and 25 h) after stress treatment for three maize inbreds (B73, Mo17, and W22—due to low germination, several time points were omitted for W22). This time-course experiment included a single biological replicate of pooled individuals (3–4 individuals per sample). The third experiment included a single biological replicate of three pooled individuals for a diverse panel of 25 maize inbreds at two time points under control and cold conditions. All three experiments were performed from late May to early July in 2019. In each experiment, maize seeds of the selected genotypes were hydrated for 24 h in distilled water and grown in growth chambers at 30°C/20°C under 16-h light/8-h dark cycles to the V2/V3 stage (Day 9) for stress treatments and collection of the V2 leaves. For all three experiments, the stress treatment (cold or heat) was initiated 2 h after dawn on Day 9 (i.e. time zero), and samples were collected from the control, cold, and heat groups simultaneously at the indicated time points. The temperature settings for control, cold, and heat conditions were 30°C/20°C, 6°C/2°C, and 39°C/29°C, respectively. For each replicate, V2 leaves from three maize seedlings were collected and pooled. A total of 292 samples were collected for profiling: 126 samples in the replicated experiment of the three inbreds and their F_1 hybrids, 66 samples in the time-course experiment, and 100 samples for the diversity panel experiment. One sample did not meet the minimum cDNA content requirement during RNA-seq library preparation and was omitted (Supplemental Data Set S1).

RNA-seq data processing

Sequencing libraries were prepared using the standard TruSeq Stranded mRNA library protocol and sequenced on a NovaSeq S4 flow cell as 150-bp paired-end reads to produce at least 20 million reads for each sample (Supplemental Data Set S1). Both library construction and sequencing were done at the University of Minnesota Genomics Center. Sequencing reads were then processed through the nf-core RNA-seq pipeline (Ewels et al., 2020) for initial quality control and raw read counting. In short, reads were trimmed using Trim Galore! and aligned to the B73 maize reference genome (AGPv4, Ensembl Plant release 32) using the variant-aware aligner Hisat2 (Kim et al., 2015) and a graph index incorporating 90 million common maize variants to account for mapping bias. Uniquely aligned reads were then counted per feature by featureCounts (Liao et al., 2014). Raw read counts were then normalized by library size and corrected for library composition bias using the Trimmed Mean of M-values normalization approach (Robinson and Oshlack, 2010) to obtain CPM reads for each gene in each sample, allowing direct comparison across samples. CPM values were then normalized by gene coding sequence length to yield fragments per kilobase of exon per million reads values. Hierarchical clustering, PCA and t-SNE clustering were used to explore sample cluster patterns and to remove questionable samples. The two missing time points in W22 (3 and 8 h) were imputed from the two neighboring time points using linear imputation. Two samples from the replicated experiment at the 25 h time point (HY91 and HY93) were also removed due to poor correlation with other biological replicates and substituted by two from the time-course samples (TC64 and TC66) that had the same treatment and genotype.

Identification of DEGs and characterization of genotypic differences

We used the replicated inbred/hybrid experiment to call DEGs between stress-treated samples and control samples. Since each time point in the experiment has two potential controls (one time 0 unstressed sample; one unstressed sample at the matching time point), two comparisons were made: one comparing the stress-treated samples at 1 or 25 h with unstressed samples at time zero (0 h); and one comparing the treated samples with unstressed samples at the matching time point (1 or 25 h). This scheme led to two sets of DEGs identified for each genotype at each time point, the overlap between which was considered true stress-responsive genes and retained for downstream analysis. All statistical tests were done using the DESeq2 package version 1.32.0 in R (false discover rate [FDR] adjusted $P < 0.05$ and a minimum fold-change of 2) (Love et al., 2014).

To characterize the genotypic effect in each gene response to cold and heat stress, we made use of the Generalized Linear Model fitting framework in DESeq2 version 1.32.0 (Love et al., 2014). An interaction term between treatment condition and genotype was introduced and the model design was formulated as Design = ∼Genotype + Condition + Genotype:Condition. Using B73 as a baseline, we identified genes showing significantly different responses in Mo17 and W22 (compared to the response of B73) as well as those showing different responses between Mo17 and W22. In each comparison, significance was defined as FDR adjusted $P < 0.05$ and a minimum fold-change (between the response of the two genotypes to cold/heat treatment) of 2. Similar to calling DEGs for each genotype, this between-genotype test was also performed twice—once using time 0 unstressed samples; once using unstressed samples at the matched time points. A gene needed to show significance in both comparisons in order to be deemed as having a significant genotype effect.

The list of genes showing genotypic effect in each of the three pairwise comparisons (B73 to Mo17, B73 to W22, and Mo17 to W22) in each condition (cold_1h, cold_25h, heat_1h, and heat_25h) were further grouped into 27 categories based on their response status in the three genotypes. We focused on the categories where stress response (either activation of repression) was lost in one or two genotypes, and used these as candidates to validate the cis-regulatory motifs we discovered.

Identification of cold- and heat-responsive co-expression clusters

The time-course experiment was used to identify co-expression clusters following cold or heat stress. Our first analysis only looked at genes showing differential expression in the replicated inbred F1 hybrid experiment at one of the two time points in B73. Since each time point of the time-course experiment has one sample in the treatment group and one control sample at the matching time point, we explored three ways to construct the raw expression matrix for clustering: (1) CPM values of the treated sample at each time point; (2) CPM differences between treatment and control; and (3) Log_2 ratio of CPMs between treatment and control. We found that option 2 generally leads to both interpretable and an ideal number of clusters. Expression levels for the three missing time points in W22 were imputed linearly using two neighboring time points. Distance-based hierarchical clustering was then used to discover gene co-expression clusters. The matrix of gene expression differences ($\text{CPM}_{\text{stress}} - \text{CPM}_{\text{control}}$) was normalized using variance stabilizing transformation (Anders and Huber, 2010). The Pearson's correlation coefficient-based distance matrix was then obtained and used for hierarchical clustering (method = "ward.D2"). The resulting gene tree was cut using the "cutreeDynamic" function (deepSplit = 3, minGap = 0) to yield 10–30 clusters along with their eigengenes (i.e. first principal component of the standardized expression vectors). Clusters with very similar eigengenes were then merged using different parameters (cutHeight = 0.1/0.15/0.2/0.25/0.3), the results of which were visually inspected to determine the best cutting height.

Identification of enriched motifs in upstream/downstream regions of stress-responsive gene clusters

Motif mining was performed using different sets of genes, including the list of all upregulated and downregulated genes under cold or heat stress, as well as specific co-expression clusters showing distinct time-course expression patterns (e.g. early or late upregulation). STREME (version 5.3.0) was run on each gene set to identify enriched motifs (8–20-bp ungapped k-mers) using genes not showing differential expression (P.adj > 0.05, fold-change < 1.5 in average expression) as negative controls (Bailey, 2011). For each gene set, we explored the effect of different search spaces including different promoter size (0.5, 1, and 2 kb) around the TSS, or a combination of these regions, as well as a methylation- or accessibility-masked promoter space (i.e. only retaining regions classified as unmethylated (Crisp et al., 2020) or accessible in leaves (acrL) or ears, based on (Ricci et al., 2019).

Known TF binding motifs from Arabidopsis (*Arabidopsis thaliana*) and maize were obtained from multiple sources (Weirauch et al., 2014; Ricci et al., 2019; Tu et al., 2020). Position weight matrices (PWMs) were either directly downloaded from the cis-BP website or built from called DNA Affinity Purification and sequencing peaks using GEM (version 3.4; Guo et al., 2018b). All motifs identified by STREME were compared to these public TF PWMs by calculating an all-against-all pairwise similarity matrix using bioconductor package *universalmotif* version 1.8.3 (method = "PCC," min.mean.ic = .0, min.overlap = 5, score.strat = a.mean") (Tremblay). Two rounds of hierarchical clustering (method = "average") were then performed to cluster motifs into motif groups. Any STREME motif grouping with a public TF motif was assigned a "known" status and the corresponding TF label. Motif groups containing only STREME-identified motifs and no public TF motifs were assigned a "novel" status.

Training of machine learning models to predict stress response based on cis-regulatory elements

Different sets of genes (e.g. all upregulated or downregulated genes under cold or heat stress, early or late upregulation under cold or heat stress) were selected as positives, while nonDE genes were selected as negatives to train machine learning random forest models to predict stress responsiveness. For each round of model generation, downsampling of DE and nonDE genes was done prior to training to achieve balance between label groups. To train each model, we first split data into 80% training set and 20% test set. Training was done within the 80% training set using 10-fold cross validation. Model hyperparameters (number of predictors, number of trees, and minimum number of data points in a node) were determined using a grid search algorithm ("tune_grid" function) implemented in the R package *tidymodels* (version 0.1.3). Downsampling and training were repeated 100 times (using a different random seed for downsampling) and separate model training was performed for each of these 100 sets of DE/non-DE genes. The model was then evaluated with the 20% test data. The mean, median, and standard deviation of performance metrics (F_1 score, AUROC, area under the precision-recall curve) for all 100 models were obtained (Supplemental Data Set S1). Permutation-based feature importance scores of each input motif were extracted from trained models using the R package *ranger* (Wright and Ziegler, 2015). The best performing model (highest F_1 score) was picked for each stress-responsive gene set and later used for evaluation of external datasets (e.g. Figures 5, B and 6, A). After assessing the effect (F_1 score) of different model training parameters (Figure 4), we determined a set of optimal parameters: " + /−2 kb TSS and + /−2 kb TTS," "UMR," and "top100" features and using motif presence/absence (0/1) as feature representation.

Characterization of genes showing stress-responsive cis- or trans- regulatory patterns

The relative expression of genes, and alleles, in the parents and F_1 hybrids was used to classify the cis- and trans-regulatory variation that influences variable expression responses to heat or cold stress for the subset of genes exhibiting significant response to stress in one genotype but not in another. Both cis- and trans-regulatory variation would result in differential expression responses in one genotype compared to another. However, for cis-regulatory variation we would expect to see only one of the two alleles responding in the F_1 hybrid, whereas differences in trans-regulatory variation would be expected to result in changes to the expression of both alleles in the F_1. To classify gene expression levels into different regulatory categories, for each gene, we introduced the following notation (using B73 and Mo17 as examples):

pa_i = expression of the gene in the ith B73 parent under control condition.

pb_i = expression of the gene in the ith Mo17 parent under control condition.

ha_j = number of reads mapping to the B73 allele in the jth F_1 hybrid under control condition.

hb_j = number of reads mapping to the Mo17 allele in the jth F_1 hybrid under control condition.

pa_i' = expression of the gene in the ith B73 parent under stress condition.

pb_i' = expression of the gene in the ith Mo17 parent under stress condition.

ha_j' = number of reads mapping to the B73 allele in the jth F_1 hybrid under stress condition.

hb_j' = number of reads mapping to the Mo17 allele in the jth F_1 hybrid under stress condition.

Here i and j take values between 1 and 3. Subsequently, we make the following distributional assumptions:

$$pa_i \sim NB(\mu_1, \gamma),\ pb_i \sim NB(\mu_2, \gamma),\ ha_j \sim NB(\mu_3, \gamma),\ hb_j \sim NB(\mu_4, \gamma)$$

$$pa_i' \sim NB(\mu_1', \gamma),\ pb_i' \sim NB(\mu_2', \gamma),\ ha_j' \sim NB(\mu_3', \gamma),\ hb_j' \sim NB(\mu_4', \gamma)$$

The dispersion parameter (γ) for each gene was estimated using the *estimateDispersions*() function within DESeq2 with *fitType* = "*parametric*" option (Love et al., 2014). The marginal distributions of a_i are negative binomial. Subsequently, different constraints upon the parameters can be imposed to describe the following biological situations:

$$\text{First define}: d\mu_1 = \mu_1' - \mu_1, d\mu_2 = \mu_2' - \mu_2, d\mu_3 = \mu_3' - \mu_3, d\mu_4 = \mu_4' - \mu_4$$

conserved: $d\mu_1 = d\mu_2$ and $d\mu_3 = d\mu_4$,
unexpected: $d\mu_1 = d\mu_2$ and $d\mu_3 \neq d\mu_4$,
cis: $d\mu_1 \neq d\mu_2$ and $d\mu_1/d\mu_2 = d\mu_3/d\mu_4$,
trans: $d\mu_1 \neq d\mu_2$ and $d\mu_3 = d\mu_4$,
cis and trans: $d\mu_1 \neq d\mu_2$ and $d\mu_1/d\mu_2 \neq d\mu_3/d\mu_4$

Each gene was allocated into one of these four categories by first fitting the five models (conserved, cis only, trans only, cis and trans, and unexpected) to the data by maximizing the likelihood function, then calculating the Bayesian Information Criterion using R package *bbmle* (version 1.0.24) to determine which of the five models best fitted the data for each gene.

Identification of putative variants regulating cold-responsive expression in a panel of 25 maize genotypes

Starting from a set of approximately 2,000 genes showing variable response patterns 25 h into cold treatment between B73, Mo17, and W22, we checked their response patterns in a wider set of 25 maize genotypes. The levels of activation/repression (Log_2[Fold-change]) for different maize inbreds were plotted. A Gaussian finite mixture model was fitted to the Log_2(Fold change) values for each gene to distinguish unimodal (one cluster), bimodal (two clusters), or multimodal (more than two clusters) distributions of data points. This was done using the *densityMclust*() function from the R package *mclust* version 5.4.7 (Scrucca et al., 2016). Bimodally distributed genes were further filtered, requiring one cluster centering around 0 suggesting no DE, with the other cluster departing from 0 suggesting significant DE. A total of approximately 500 genes was identified showing significant bimodal distribution of their Log_2(Fold change) values among the 25 diverse genotypes that contain at least one nonDE genotype and one DE genotype. Genomic sequences of the 25 maize genotypes were then aligned to B73 and bi-allelic variants within 2 kb of each gene were extracted to perform a local association test with the cold response pattern (i.e. treating the bi-modal distribution as one group of stress-responsive genotypes and another group of nonstress-responsive genotypes). We used PLINK (version 1.90b6.21) to perform a standard chi-square test for each gene with multiple testing corrections ("–assoc –adjust") (Purcell et al., 2007). Significant associations were determined by requiring the Bonferroni corrected $P < 0.05$. Independently, the presence/absence status of the top 100 enriched motifs was checked for their co-occurrence with the response phenotype in the 25 genotypes assayed, with the proportion of correct motif-response associations reported as the prediction accuracy of that motif. The resulting 130 motifs with higher than 0.8 prediction accuracy were reported.

Accession numbers

Raw RNA-Seq reads have been deposited in NCBI Sequence Read Archive under accession PRJNA747925. All source codes used for quantification, normalization, statistical testing, and machine learning training and evaluation, all processed data sets including gene lists of each stress responsive pattern, lists of genes under cis/trans stress responsive regulation, lists of enriched motifs in each co-expression cluster are available on Github: https://github.com/orionzhou/stress.

Supplemental data

The following materials are available in the online version of this article.

Supplemental Figure S1. Principal component clustering of different set of samples from the hybrid experiment under control, cold, and heat conditions.

Supplemental Figure S2. Expression profile of maize HSFs response to heat stress.

Supplemental Figure S3. Expression profile of TFs that have previously been reported to play a role in maize response to cold.

Supplemental Figure S4. Comparison of heat- and cold-response gene expression in B73, Mo17, and W22.

Supplemental Figure S5. Characterization of genes with variable stress-responsive patterns among inbreds.

Supplemental Figure S6. Cis/trans characterization of genes showing different stress response among inbreds.

Supplemental Figure S7. Hierarchical (A) and t-SNE (B) clustering of all samples from the time course experiment under control, cold, and heat conditions.

Supplemental Figure S8. Identification of cis-regulatory variants associated with variable cold responsive pattern in a panel of 25 maize genotypes.

Supplemental Figure S9. Identification of TFBS variation associated with variable stress responsive expression patterns.

Supplemental Table S1. Overlap of significantly DE genes among genotypes.

Supplemental Table S2. Number of genes assigned to cis or trans categories in different genotype and stress contrasts.

Supplemental Table S3. Model prediction accuracy (AUROC) on different co-expression clusters.

Supplemental Table S4. Genes showing associated HSF motif presence/absence with heat stress responsive expression.

Supplemental Data Set S1. Information about samples used for RNA-Seq.

Supplemental Data Set S2. Enriched GO terms for heat upregulated and cold upregulated genes.

Supplemental Data Set S3. Performance metrics (F1 score, AUROC, area under the precision-recall curve) for all models trained in this study.

Supplemental Data Set S4. Motifs associated with variable cold stress responsive expression.

Acknowledgments

We thank Peter Hermanson for assistance with plant growth, sampling and library preparation. We thank the Minnesota Supercomputing Institute at the University of Minnesota (http://www.msi.umn.edu) for providing resources that contributed to the research results reported within this article.

Funding

This study was funded by the National Science Foundation (grants IOS-IOS-1444456, IOS-1546899, and IOS-1733633).

Conflict of interest statement. None declared.

References

Agarwal PK, Agarwal P, Reddy MK, Sopory SK (2006) Role of DREB transcription factors in abiotic and biotic stress tolerance in plants. Plant Cell Rep **25**: 1263–1274

Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, Suresh H, Ramakrishnan S, Maumus F, Ciren D, et al. (2020) Major impacts of widespread structural variation on gene expression and crop improvement in tomato. Cell **182**: 15–161

Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol **11**: R106

Avila LM, Obeidat W, Earl H, Niu X, Hargreaves W, Lukens L (2018) Shared and genetically distinct *Zea mays* transcriptome responses to ongoing and past low temperature exposure. BMC Genomics **19**: 761

Azodi CB, Lloyd JP, Shiu SH (2020a) The cis-regulatory codes of response to combined heat and drought stress in *Arabidopsis thaliana*. NAR Genom Bioinform **2:** lqaa049

Azodi CB, Tang J, Shiu SH (2020b) Opening the black box: interpretable machine learning for geneticists. Trends Genet **36**: 442–455

Bailey TL (2011) DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics **27**: 1653–1659

Bailey TL (2020) STREME: accurate and versatile sequence motif discovery. Bioinformatics **37**: 2834–2840

Baillo EH, Kimotho RN, Zhang Z, Xu P (2019) Transcription factors associated with abiotic and biotic stress tolerance and their potential for crops improvement. Genes **10**: 771

Busch W, Wunderlich M, Schöffl F (2005) Identification of novel heat shock factor-dependent genes and biochemical pathways in *Arabidopsis thaliana*. Plant J **41**: 1–14

Charng YY, Liu HC, Liu NY, Chi WT, Wang CN, Chang SH, Wang TT (2007) A heat-inducible transcription factor, HsfA2, is required for extension of acquired thermotolerance in Arabidopsis. Plant Physiol **143**: 251–262

Cheng MC, Liao PM, Kuo WW, Lin TP (2013) The Arabidopsis ETHYLENE RESPONSE FACTOR1 regulates abiotic stress-responsive gene expression by binding to different cis-acting elements in response to different stress signals. Plant Physiol **162**: 1566–1582

Chen L, Song Y, Li S, Zhang L, Zou C, Yu D (2012) The role of WRKY transcription factors in plant abiotic stresses. Biochim Biophys Acta **1819**: 120–128

Chinnusamy V, Zhu J, Zhu JK (2007) Cold stress regulation of gene expression in plants. Trends Plant Sci **12**: 444–451

Crisp PA, Marand AP, Noshay JM, Zhou P, Lu Z, Schmitz RJ, Springer NM (2020) Stable unmethylated DNA demarcates expressed genes and their cis-regulatory space in plant genomes. Proc Natl Acad Sci USA **117**: 23991–24000

Cubillos FA, Stegle O, Grondin C, Canut M, Tisné S, Gy I, Loudet O (2014) Extensive cis-regulatory variation robust to environmental perturbation in Arabidopsis. Plant Cell **26**: 4298–4310

Dietz KJ, Vogel MO, Viehhauser A (2010) AP2/EREBP transcription factors are part of gene regulatory networks and integrate metabolic, hormonal and environmental signals in stress acclimation and retrograde signalling. Protoplasma **245**: 3–14

Ding Y, Shi Y, Yang S (2019) Advances and challenges in uncovering cold tolerance regulatory mechanisms in plants. New Phytol **222**: 1690–1704

Ding Y, Shi Y, Yang S (2020) Molecular regulation of plant responses to environmental temperatures. Mol Plant **13**: 544–564

Enders TA, St. Dennis S, Oakland J, Callen ST, Gehan MA, Miller ND, Spalding EP, Springer NM, Hirsch CD (2019) Classifying cold-stress responses of inbred maize seedlings using RGB imaging. Plant Direct **3**: e00104

Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, Garcia MU, Di Tommaso P, Nahnsen S (2020) The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol **38**: 276–278

Fowler S, Thomashow MF (2002) Arabidopsis transcriptome profiling indicates that multiple regulatory pathways are activated during cold acclimation in addition to the CBF cold response pathway. Plant Cell **14**: 1675–1690

Franco-Zorrilla JM, López-Vidriero I, Carrasco JL, Godoy M, Vera P, Solano R (2014) DNA-binding specificities of plant transcription factors and their potential to define target genes. Proc Natl Acad Sci USA **111**: 2367–2372

Frey FP, Pitz M, Schön CC, Hochholdinger F (2020) Transcriptomic diversity in seedling roots of European flint maize in response to cold. BMC Genomics **21**: 300

Gallegos JE, Rose AB (2019) An intron-derived motif strongly increases gene expression from transcribed sequences through a splicing independent mechanism in *Arabidopsis thaliana*. Sci Rep **9**: 13777

Guo X, Liu D, Chong K (2018a) Cold signaling in plants: insights into mechanisms and regulation. J Integr Plant Biol **60**: 745–756

Guo Y, Tian K, Zeng H, Guo X, Gifford DK (2018b) A novel k-mer set memory (KSM) motif representation improves regulatory variant prediction. Genome Res **28**: 891–900

Han J, Wang P, Wang Q, Lin Q, Chen Z, Yu G, Miao C, Dao Y, Wu R, Schnable JC, et al. (2020) Genome-wide characterization of DNase I-hypersensitive sites and cold response regulatory landscapes in grasses. Plant Cell **32**: 2457–2473

He J, Jiang Z, Gao L, You C, Ma X, Wang X, Xu X, Mo B, Chen X, Liu L (2019) Genome-wide transcript and small RNA profiling reveals transcriptomic responses to heat stress. Plant Physiol **181**: 609–629

Hoopes GM, Hamilton JP, Wood JC, Esteban E, Pasha A, Vaillancourt B, Provart NJ, Buell CR (2019) An updated gene atlas for maize reveals organ-specific and stress-induced genes. Plant J **97**: 1154–1167

Hsieh EJ, Cheng MC, Lin TP (2013) Functional characterization of an abiotic stress-inducible transcription factor AtERF53 in *Arabidopsis thaliana*. Plant Mol Biol **82**: 223–237

Huang J, Zhao X, Bürger M, Wang Y, Chory J (2021) Two interacting ethylene response factors regulate heat stress response. Plant Cell **33**: 338–357

Jeong YM, Mun JH, Lee I, Woo JC, Hong CB, Kim SG (2006) Distinct roles of the first introns on the expression of Arabidopsis profilin gene family members. Plant Physiol **140**: 196–209

Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements.. Nat Methods **12**: 357–360

Kwon CT, Heo J, Lemmon ZH, Capua Y, Hutton SF, Van Eck J, Park SJ, Lippman ZB (2020) Rapid customization of Solanaceae fruit crops for urban agriculture. Nat Biotechnol **38**: 182–188

Laxa M (2016) Intron-mediated enhancement: a tool for heterologous gene expression in plants? Front Plant Sci **7**: 1977

Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics **30**: 923–930

Li B, Gao K, Ren H, Tang W (2018) Molecular mechanisms governing plant responses to high temperatures. J Integr Plant Biol **60**: 757–779

Licausi F, Ohme-Takagi M, Perata P (2013) APETALA2/Ethylene responsive factor (AP2/ERF) transcription factors: mediators of stress responses and developmental programs. New Phytol **199**: 639–649

Li C, Ng CKY, Fan LM (2015) MYB transcription factors, active players in abiotic stress signaling. Environ Exp Bot **114**: 80–91

Li P, Cao W, Fang H, Xu S, Yin S, Zhang Y, Lin D, Wang J, Chen Y, Xu C, Yang Z (2017) Transcriptomic profiling of the maize (*Zea mays* L.) leaf response to abiotic stresses at the seedling stage. Front Plant Sci **8**: 290

Liu S, Li C, Wang H, Wang S, Yang S, Liu X, Yan J, Li B, Beatty M, Zastrow-Hayes G, **et al.** (2020) Mapping regulatory variants controlling gene expression in drought response and tolerance in maize. Genome Biol **21**: 163

Long HK, Prescott SL, Wysocka J (2016) Ever-changing landscapes: transcriptional enhancers in development and evolution. Cell **167**: 1170–1187

Lovell JT, Schwartz S, Lowry DB, Shakirov EV, Bonnette JE, Weng X, Wang M, Johnson J, Sreedasyam A, Plott C, **et al.** (2016) Drought responsive gene expression regulatory divergence between upland and lowland ecotypes of a perennial C4 grass. Genome Res **26**: 510–518

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

Madani N, Kimball JS, Ballantyne AP, Affleck DLR, van Bodegom PM, Reich PB, Kattge J, Sala A, Nazeri M, Jones MO, **et al.** (2018) Future global productivity will be affected by plant trait response to climate. Sci Rep **8**: 2870

Maher KA, Bajic M, Kajala K, Reynoso M, Pauluzzi G, West DA, Zumstein K, Woodhouse M, Bubb K, Dorrity MW, **et al.** (2018) Profiling of accessible chromatin regions across multiple plant species and cell types reveals common gene regulatory principles and new control modules. Plant Cell **30**: 15–36

Mejía-Guerra MK, Buckler ES (2019) A k-mer grammar analysis to uncover maize regulatory architecture. BMC Plant Biol **19**: 103

Mittler R, Finka A, Goloubinoff P (2012) How do plants feel the heat? Trends Biochem Sci **37**: 118–125

Mizoi J, Shinozaki K, Yamaguchi-Shinozaki K (2012) AP2/ERF family transcription factors in plant abiotic stress responses. Biochim Biophys Acta **1819**: 86–96

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

Ohama N, Sato H, Shinozaki K, Yamaguchi-Shinozaki K (2017) Transcriptional regulatory network of plant heat stress response. Trends Plant Sci **22**: 53–65

Oka R, Zicola J, Weber B, Anderson SN, Hodgman C, Gent JI, Wesselink JJ, Springer NM, Hoefsloot HCJ, Turck F, **et al.** (2017) Genome-wide mapping of transcriptional enhancer candidates using DNA and chromatin features in maize. Genome Biol **18**: 137

Parvathaneni RK, Kumar I, Braud M, Eveland AL (2020) Regulatory signatures of drought response in stress resilient Sorghum bicolor. bioRxiv https://doi.org/10.1101/2020.08.07.240580 (June 23, 2021)

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, **et al.** (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet **81**: 559–575

Raxwal VK, Ghosh S, Singh S, Katiyar-Agarwal S, Goel S, Jagannath A, Kumar A, Scaria V, Agarwal M (2020) Abiotic stress-mediated modulation of the chromatin landscape in *Arabidopsis thaliana*. J Exp Bot **71**: 5280–5293

Reményi A, Schöler HR, Wilmanns M (2004) Combinatorial control of gene expression. Nat Struct Mol Biol **11**: 812–815

Ricci WA, Lu Z, Ji L, Marand AP, Ethridge CL, Murphy NG, Noshay JM, Galli M, Mejía-Guerra MK, Colomé-Tatché M, **et al.** (2019) Widespread long-range cis-regulatory elements in the maize genome. Nat Plants **5**: 1237–1249

Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol **11**: R25

Sartor RC, Noshay J, Springer NM, Briggs SP (2019) Identification of the expressome by machine learning on omics data. Proc Natl Acad Sci USA **116**: 18119–18125

Scharf KD, Berberich T, Ebersberger I, Nover L (2012) The plant heat stress transcription factor (Hsf) family: structure, function and evolution. Biochim Biophys Acta **1819**: 104–119

Schwarz B, Azodi CB, Shiu SH, Bauer P (2020) Putative cis-regulatory elements predict iron deficiency responses in Arabidopsis roots. Plant Physiol **182**: 1420–1439

Scrucca L, Fop M, Murphy TB, Raftery AE (2016) mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. R J **8**: 289–317

Shi Y, Huang J, Sun T, Wang X, Zhu C, Ai Y, Gu H (2017) The precise regulation of different COR genes by individual CBF transcription factors in *Arabidopsis thaliana*. J Integr Plant Biol **59**: 118–133

Song L, Huang SSC, Wise A, Castanon R, Nery JR, Chen H, Watanabe M, Thomas J, Bar-Joseph Z, Ecker JR (2016) A transcription factor hierarchy defines an environmental stress response network. Science **354**: aag1550

Tremblay BJ (2021) universalmotif: import, modify, and export motifs with R. R package version 1.8.2. https://bioconductor.org/packages/universalmotif/

Tu X, Mejía-Guerra MK, Valdes Franco JA, Tzeng D, Chu PY, Shen W, Wei Y, Dai X, Li P, Buckler ES, **et al.** (2020) Reconstructing the maize leaf regulatory network using ChIP-seq data of 104 transcription factors. Nat Commun **11**: 5089

Uygun S, Azodi CB, Shiu SH (2019) Cis-regulatory code for predicting plant cell-type transcriptional response to high salinity. Plant Physiol **181**: 1739–1751

Uygun S, Seddon AE, Azodi CB, Shiu SH (2017) Predictive models of spatial transcriptional response to high salinity. Plant Physiol **174**: 450–464

Vihervaara A, Duarte FM, Lis JT (2018) Molecular mechanisms driving transcriptional stress responses. Nat Rev Genet **19**: 385–397

Wang FX, Shang GD, Wu LY, Xu ZG, Zhao XY, Wang JW (2020a) Chromatin accessibility dynamics and a hierarchical transcriptional regulatory network structure for plant somatic embryogenesis. Dev Cell **54**: 742–757.e8

Wang H, Cimen E, Singh N, Buckler E (2020b) Deep learning for plant genomics and crop improvement. Curr Opin Plant Biol **54**: 34–41

Wang W, Vinocur B, Shoseyov O, Altman A (2004) Role of plant heat-shock proteins and molecular chaperones in the abiotic stress response. Trends Plant Sci **9**: 244–252

Washburn JD, Mejia-Guerra MK, Ramstein G, Kremling KA, Valluru R, Buckler ES, Wang H (2019) Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. Proc Natl Acad Sci USA **116**: 5542–5549

Waters AJ, Makarevitch I, Noshay J, Burghardt LT, Hirsch CN, Hirsch CD, Springer NM (2017) Natural variation for gene expression responses to abiotic stress in maize. Plant J **89**: 706–717

Weber B, Zicola J, Oka R, Stam M (2016) Plant enhancers: a call for discovery. Trends Plant Sci **21**: 974–987

Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, et al. (2014) Determination and inference of eukaryotic transcription factor sequence specificity. Cell **158**: 1431–1443

Wright MN, Ziegler A (2015) Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. arXiv [stat.ML]

Yao Y, He RJ, Xie QL, Zhao XH, Deng XM, He JB, Song L, He J, Marchant A, Chen XY, et al. (2017) ETHYLENE RESPONSE FACTOR 74 (ERF74) plays an essential role in controlling a respiratory burst oxidase homolog D (RbohD)-dependent mechanism in response to different stresses in Arabidopsis. New Phytol **213**: 1667–1681

Yilmaz A, Nishiyama MY, Jr, Fuentes BG, Souza GM, Janies D, Gray J, Grotewold E (2009) GRASSIUS: a platform for comparative regulatory genomics across the grasses. Plant Physiol **149**: 171–180

Zeng Z, Zhang W, Marand AP, Zhu B, Buell CR, Jiang J (2019) Cold stress induces enhanced chromatin accessibility and bivalent histone modifications H3K4me3 and H3K27me3 of active genes in potato. Genome Biol **20**: 123

Zhang H, Li G, Fu C, Duan S, Hu D, Guo X (2020) Genome-wide identification, transcriptome analysis and alternative splicing events of Hsf family genes in maize. Sci Rep **10**: 8073

Zhang Y, Ngu DW, Carvalho D, Liang Z, Qiu Y, Roston RL, Schnable JC (2017) Differentially regulated orthologs in sorghum and the subgenomes of maize. Plant Cell **29**: 1938–1951

Zou C, Sun K, Mackaluso JD, Seddon AE, Jin R, Thomashow MF, Shiu SH (2011) Cis-regulatory code of stress-responsive transcription in *Arabidopsis thaliana*. Proc Natl Acad Sci USA **108**: 14992–14997