Environmental stress leads to dramatic transcriptional reprogramming, which is central to plant survival. Although substantial knowledge has accumulated on how a few plant cis-regulatory elements (CREs) function in stress regulation, many more CREs remain to be discovered. In addition, the plant stress cis-regulatory code, i.e., how CREs work independently and/or in concert to specify stress-responsive transcription, is mostly unknown. On the basis of gene expression patterns under multiple stresses, we identified a large number of putative CREs (pCREs) in Arabidopsis thaliana with characteristics of authentic cis-elements. Surprisingly, biotic and abiotic responses are mostly mediated by two distinct pCRE superfamilies. In addition, we uncovered cis-regulatory codes specifying how pCRE presence and absence, combinatorial relationships, location, and copy number can be used to predict stress-responsive expression. Expression prediction models based on pCRE combinations perform significantly better than those based on simply pCRE presence and absence, combinatorial relationships, location, and copy number. Furthermore, instead of a few master combinatorial rules for each stress condition, many rules were discovered, and each appears to control only a small subset of stress-responsive genes. Given there are very few documented interactions between plant CREs, the combinatorial rules we have uncovered significantly contribute to a better understanding of the cis-regulatory logic underlying plant stress response and provide prioritized targets for experimentation.

Results and Discussion

Multiple pCREs Implicated in Regulating Abiotic and Biotic Stress-Responsive Transcription Belong to Two Motif Superfamilies. The assumption that genes with similar expression patterns are likely coregulated and have the same CREs has been applied to identify plant CREs (25, 26). Therefore, we set out to identify pCREs involved in regulating stress-responsive transcription on the basis of coexpression patterns of A. thaliana genes under 16 abiotic/biotic stress conditions (SI Methods and Table S1). Each Arabidopsis thaliana gene was categorized as up-regulated, down-regulated, or not changed under each condition, and a motif discovery pipeline was used to identify 1,215 pCREs from putative promoter regions (Fig. 1A, SI Methods, and Dataset S1). The sites where these pCREs were mapped are preferentially found in the promoters of stress-responsive genes (up- and/or down-regulated) compared with promoters of genes without significant changes under stress (Fig. 1B and Dataset S2 A and B). Among these pCREs, 346 are highly similar [Pearson’s correlation coefficient (PCC) ≥ 0.9] (SI Methods) to 52 known CREs (Dataset SI).

Author contributions: C.Z., R.J., M.F.T., and S.-H.S. designed research; C.Z., K.S., J.D.M., A.E.S., and S.-H.S. performed research; R.J. contributed new reagents/analytic tools; C.Z., J.D.M., and A.E.S. analyzed data; and C.Z., M.F.T., and S.-H.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

To whom correspondence should be addressed. E-mail: shius@msu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1103202108/-/DCSupplemental.

Cis-regulatory code of stress-responsive transcription in Arabidopsis thaliana

Cheng Zou, Kelian Sun, Joshua D. Mackalus, Alexander E. Seddon, Rong Jin, Michael F. Thomashow, and Shin-Han Shiu

Departments of Plant Biology, Computer Science and Engineering, Biochemistry and Molecular Biology, and Crop and Soil Sciences, and Department of Energy Plant Research Laboratory, Michigan State University, East Lansing, MI 48824

Edited by Philip Benfey, Duke University, Durham, NC, and approved July 19, 2011 (received for review March 3, 2011)
pCREs Have Properties of Authentic CREs from Plants and Other Model Systems. Several lines of evidence indicate that the pCREs identified in this study are authentic. All pCREs were found on the basis of significant enrichment in the promoters of stress-responsive genes and many are similar to known CREs. In addition, known CREs regulating stress responses are recovered. There is also a significant positional bias of pCREs (Fig. 2) that is similar to experimentally established motifs in yeast (29), humans (30), and plants (31, 32). We found that pCREs are enriched in putative promoter regions of stress-responsive genes particularly within ~300 bp upstream of the transcriptional start site (TSS) (Fig. 2A). In addition, such pCRE positional bias is, in large part, observed only for genes responsive to conditions under which the pCREs were originally identified (Fig. 2B). This location bias also holds for pCRE families (Fig. S2). Another characteristic that supports pCRE authenticity is the significantly higher degree of evolutionary conservation of sites where pCREs were mapped compared with sites mapped by randomized pCREs (Fig. 2C and D). When the degree of conservation was assessed through a comparison of putative orthologous regions between A. thaliana, Arabidopsis lyrata, and poplar genomes (SI Methods and Dataset S2E), 80% of pCREs were found to have significantly different conservation score distributions from randomized
transcription factor that is significantly up-regulated under high salt stress treatment for 3 h (Salt3). The presence of a CRE will be a perfect predictor for Salt3-responsive transcription if the response is controlled by only one CRE and there is no other level of regulation. Contrary to this naive scheme, among Salt3-responsive genes, only 13.9% contain a pCRE (GTCGGTs, reverse complement) for Salt3 response are also low guesses of Salt3 response would have a precision of 0.067 because 6.7% of A. thaliana genes are Salt3 responsive (Fig. 4A). Similarly, the precision and recall for an ABRE-like pCRE (rmSACGTGkm) for Salt3 response are also low (Fig. 4A), but still higher than expected by chance; random guesses of Salt3 response would have a precision of 0.067 because 6.7% of A. thaliana genes are Salt3 responsive (Fig. 4A, dotted line). Although taking into account the presence of either ABRE or DRE significantly improves the precision compared with that of random guesses, other CREs and/or other regulatory mechanisms are apparently necessary to fully explain Salt3 response.

To test whether the inclusion of additional CREs would improve Salt3 response prediction, we first constructed a predictive model for Salt3 responses on the basis of known plant CREs with Support Vector Machine (Methods). As shown in Fig. 4A, considering more known CREs led to marked improvement in precision and recall. In addition, we observed moderate improvement in Salt3 response prediction when pCREs were used (Fig. 4A). This improvement is expected because pCREs include or DRE significantly improves the precision compared with that of random guesses, other CREs and/or other regulatory mechanisms are apparently necessary to fully explain Salt3 response.

To test whether the inclusion of additional CREs would improve Salt3 response prediction, we first constructed a predictive model for Salt3 responses on the basis of known plant CREs with Support Vector Machine (Methods). As shown in Fig. 4A, considering more known CREs led to marked improvement in precision and recall. In addition, we observed moderate improvement in Salt3 response prediction when pCREs were used (Fig. 4A). This improvement is expected because pCREs include or DRE significantly improves the precision compared with that of random guesses, other CREs and/or other regulatory mechanisms are apparently necessary to fully explain Salt3 response.

Presence or Absence of pCREs Can Be Used to Predict Stress-Responsive Expression. Our current understanding is that the expression pattern of a gene is influenced not only by the presence of individual CREs but also by combinatorial controls (13). Computational approaches jointly considering CRE properties and patterns of gene expression allow the identification of a cis-regulatory code that stipulates how CREs control gene expression. To uncover a global plant stress cis-regulatory code, we asked how well the presence and absence of pCREs explains up-regulation of genes under a particular stress condition. Initially we focused on salt stress treatment for 3 h (Salt3). The presence of a CRE will be a perfect predictor for Salt3-responsive transcription if the response is controlled by only one CRE and there is no other level of regulation. Contrary to this naive scheme, among Salt3-responsive genes, only 13.9% contain a pCRE (GTCGGTs, reverse complement) for Salt3 response are also low guesses of Salt3 response would have a precision of 0.067 because 6.7% of A. thaliana genes are Salt3 responsive (Fig. 4A). Similarly, the precision and recall for an ABRE-like pCRE (rmSACGTGkm) for Salt3 response are also low (Fig. 4A), but still higher than expected by chance; random guesses of Salt3 response would have a precision of 0.067 because 6.7% of A. thaliana genes are Salt3 responsive (Fig. 4A, dotted line). Although taking into account the presence of either ABRE or DRE significantly improves the precision compared with that of random guesses, other CREs and/or other regulatory mechanisms are apparently necessary to fully explain Salt3 response.

To test whether the inclusion of additional CREs would improve Salt3 response prediction, we first constructed a predictive model for Salt3 responses on the basis of known plant CREs with Support Vector Machine (Methods). As shown in Fig. 4A, considering more known CREs led to marked improvement in precision and recall. In addition, we observed moderate improvement in Salt3 response prediction when pCREs were used (Fig. 4A). This improvement is expected because pCREs include or DRE significantly improves the precision compared with that of random guesses, other CREs and/or other regulatory mechanisms are apparently necessary to fully explain Salt3 response.

Presence or Absence of pCREs Can Be Used to Predict Stress-Responsive Expression. Our current understanding is that the expression pattern of a gene is influenced not only by the presence of individual CREs but also by combinatorial controls (13). Computational approaches jointly considering CRE properties and patterns of gene expression allow the identification of a cis-regulatory code that stipulates how CREs control gene expression. To uncover a global plant stress cis-regulatory code, we asked how well the presence and absence of pCREs explains up-regulation of genes under a particular stress condition. Initially we focused on salt stress treatment for 3 h (Salt3). The presence of a CRE will be a perfect predictor for Salt3-responsive transcription if the response is controlled by only one CRE and there is no other level of regulation. Contrary to this naive scheme, among Salt3-responsive genes, only 13.9% contain a pCRE (GTCGGTs, reverse complement) for Salt3 response are also low guesses of Salt3 response would have a precision of 0.067 because 6.7% of A. thaliana genes are Salt3 responsive (Fig. 4A). Similarly, the precision and recall for an ABRE-like pCRE (rmSACGTGkm) for Salt3 response are also low (Fig. 4A), but still higher than expected by chance; random guesses of Salt3 response would have a precision of 0.067 because 6.7% of A. thaliana genes are Salt3 responsive (Fig. 4A, dotted line). Although taking into account the presence of either ABRE or DRE significantly improves the precision compared with that of random guesses, other CREs and/or other regulatory mechanisms are apparently necessary to fully explain Salt3 response.

To test whether the inclusion of additional CREs would improve Salt3 response prediction, we first constructed a predictive model for Salt3 responses on the basis of known plant CREs with Support Vector Machine (Methods). As shown in Fig. 4A, considering more known CREs led to marked improvement in precision and recall. In addition, we observed moderate improvement in Salt3 response prediction when pCREs were used (Fig. 4A). This improvement is expected because pCREs include or DRE significantly improves the precision compared with that of random guesses, other CREs and/or other regulatory mechanisms are apparently necessary to fully explain Salt3 response.

Presence or Absence of pCREs Can Be Used to Predict Stress-Responsive Expression. Our current understanding is that the expression pattern of a gene is influenced not only by the presence of individual CREs but also by combinatorial controls (13). Computational approaches jointly considering CRE properties and patterns of gene expression allow the identification of a cis-regulatory code that stipulates how CREs control gene expression. To uncover a global plant stress cis-regulatory code, we asked how well the presence and absence of pCREs explains up-regulation of genes under a particular stress condition. Initially we focused on salt stress treatment for 3 h (Salt3). The presence of a CRE will be a perfect predictor for Salt3-responsive transcription if the response is controlled by only one CRE and there is no other level of regulation. Contrary to this naive scheme, among Salt3-responsive genes, only 13.9% contain a pCRE (GTCGGTs, reverse complement) for Salt3 response are also low guesses of Salt3 response would have a precision of 0.067 because 6.7% of A. thaliana genes are Salt3 responsive (Fig. 4A). Similarly, the precision and recall for an ABRE-like pCRE (rmSACGTGkm) for Salt3 response are also low (Fig. 4A), but still higher than expected by chance; random guesses of Salt3 response would have a precision of 0.067 because 6.7% of A. thaliana genes are Salt3 responsive (Fig. 4A, dotted line). Although taking into account the presence of either ABRE or DRE significantly improves the precision compared with that of random guesses, other CREs and/or other regulatory mechanisms are apparently necessary to fully explain Salt3 response.

To test whether the inclusion of additional CREs would improve Salt3 response prediction, we first constructed a predictive model for Salt3 responses on the basis of known plant CREs with Support Vector Machine (Methods). As shown in Fig. 4A, considering more known CREs led to marked improvement in precision and recall. In addition, we observed moderate improvement in Salt3 response prediction when pCREs were used (Fig. 4A). This improvement is expected because pCREs include or DRE significantly improves the precision compared with that of random guesses, other CREs and/or other regulatory mechanisms are apparently necessary to fully explain Salt3 response.

Presence or Absence of pCREs Can Be Used to Predict Stress-Responsive Expression. Our current understanding is that the expression pattern of a gene is influenced not only by the presence of individual CREs but also by combinatorial controls (13). Computational approaches jointly considering CRE properties and patterns of gene expression allow the identification of a cis-regulatory code that stipulates how CREs control gene expression. To uncover a global plant stress cis-regulatory code, we asked how well the presence and absence of pCREs explains up-regulation of genes under a particular stress condition. Initially we focused on salt stress treatment for 3 h (Salt3). The presence of a CRE will be a perfect predictor for Salt3-responsive transcription if the response is controlled by only one CRE and there is no other level of regulation. Contrary to this naive scheme, among Salt3-responsive genes, only 13.9% contain a pCRE (GTCGGTs, reverse complement) for Salt3 response are also low guesses of Salt3 response would have a precision of 0.067 because 6.7% of A. thaliana genes are Salt3 responsive (Fig. 4A). Similarly, the precision and recall for an ABRE-like pCRE (rmSACGTGkm) for Salt3 response are also low (Fig. 4A), but still higher than expected by chance; random guesses of Salt3 response would have a precision of 0.067 because 6.7% of A. thaliana genes are Salt3 responsive (Fig. 4A, dotted line). Although taking into account the presence of either ABRE or DRE significantly improves the precision compared with that of random guesses, other CREs and/or other regulatory mechanisms are apparently necessary to fully explain Salt3 response.
ditive models on the basis of newly identified pCREs led to significant improvements compared with known motif-based models (Fig. 4 B and C), lending support to the hypothesis that pCREs are authentic components of plant cis-regulation. We also built pCRE family-based models, but they do not perform as well as models based on individual pCREs (Fig. S2D). Therefore, individual pCREs likely include more specificity information and may better resemble authentic cis-elements than pCRE families that are likely more general descriptions of binding sites.

Our findings indicate that presence and absence of motifs are important predictors of stress-responsive transcription, and models based on pCREs in general lead to moderate (Salt3) or significant (UV1, Flg1) improvement over those built with known CREs. The rather modest improvement in Salt3 response prediction with pCRE-based models is likely due to more extensive knowledge of cis-regulatory mechanisms for salt and related cold, drought, and osmotic stress conditions (35–37). Regardless, the models based on motif presence and absence are clearly insufficient because responses of many genes were not correctly predicted. In addition, despite the fact that we found 851 pCREs enriched in the promoters of Salt3 up-regulated genes, only the top 100 ranked pCREs were needed to build predictive models with similar performances to models built with all pCREs (Fig. S3). These findings raise the question whether the lower-ranked pCREs are involved in Salt3 response, if they play a minor role, potentially as low-affinity sites, or if they are important only in combinations.

**Considering Combinatorial Controls Leads to Further Improvement in Stress Cis-Regulatory Models.** Because condition-specific expression is likely controlled by one or more transcription factors (13), we next tested the hypothesis that considering binary pCRE combinations would further improve the performance of expression prediction models. Using a classification algorithm that integrates association rule mining (Methods), 274, 357, and 271 pCRE combinatorial rules with above threshold precision and recall were identified for Salt3, UV1, and Flg1, respectively (Dataset S2A). Similar to earlier studies of yeast combinatorial control (13), some pCREs appear to be hub-like, working in combination with multiple distinct CREs (Fig. S4). In addition, predictive models built from these combinatorial rules led to substantial improvement in stress-response predictions for each of the three conditions (Fig. 4 A–C, red line; and Fig. S5A). For example, when recall is 20%, the combinatorial rule-based models led to 25–31% better precision than models based on pCRE presence or absence.

We emphasize that no single combinatorial rule has >5% recall (Dataset S3), indicating that instead of one or a few master regulatory rules that control the majority of responsive genes, multiple regulatory rules exist, each controlling a small number of genes under a stress condition. Consistent with this, distinct combinations relevant to a stress condition tend to be found in overlapping but mostly distinct sets of genes (Fig. S4D). A minority of combinatorial rules is found in similar sets of genes (Fig. 5 A and B, clusters a–c). Within these small clusters, the member pCREs tend to be similar to each other (Fig. 5 B and C). Nonetheless, there is little overlap in putative regulatory targets between clusters (Fig. S5D). In addition, although combinations with high similarity tend to contain similar pairs of pCREs (e.g., in cluster a, the second, third, and fourth combinations, Fig. S5D), there are substantial differences in the gene sets to which these pCREs are mapped.

Supporting the validity of the computational predictions, the combinatorial rules include the experimentally established ABRE-DRE (18) and ABRE-EVENING (19) CRE combinations (Fig. S5C). Furthermore, our experiments confirmed the computationally predicted combinatorial rule between two ABRE-like pCREs and a previously unknown pCRE (Fig. 3); mutations in any of these three pCREs led to significantly lower levels of salt-induced expression compared with constructs with intact sites (two-sample t tests, all with P values <0.001). Thus, the results demonstrate the necessity of the ABRE-like and the unique pCREs in combination for proper salt-induced expression. Considering the performance of the combinatorial rule-based models, the identification of known combinations, and our proof-of-concept validation experiments, it is likely many of the combinatorial rules identified in this study are relevant to the control of stress-responsive transcription.

**pCRE Copy Number and Location Are Important but Their Incorporation Does Not Significantly Improve Model Accuracy.** In addition to presence or absence of and combinatorial relationships between CREs, motif location (e.g., refs. 13 and 21) as well as copy number does not significantly improve model accuracy.
number (e.g., ref. 20) is important for cis-regulatory control. To assess the importance of motif location, we asked first how well pCRE-mapped sites in each location bin from −1 kb to the TSS predict Salt3 response. Note that here we are interested in finding out how important each location bin is if we considered all pCRE sites collectively. We found that pCRE sites located from −200 bp to the TSS have significantly better power to predict Salt3 response (Fig. 4A and B). We next asked which location bins are more important for predicting Salt3 response for each pCRE. Similar to considering pCREs in a bin jointly, pCREs located in regions proximal to the TSS tend to have higher weights (Fig. S5A). However, despite the importance of motif location, the model incorporating pCRE location performs similarly to the simpler presence/absence model (area under receiver–operating characteristic curves (AUC-ROCs) 0.781 and 0.789, respectively).

We next considered the importance of the number of pCRE sites in predicting Salt3 response (Fig. 4E). Similar to pCRE location, we found that although the number of pCRE sites is important, the model based on pCRE copy number does not outperform the model based solely on pCRE presence or absence (AUC-ROCs are 0.783 and 0.789, respectively). In addition to pCRE location and copy number, we also explored a more complicated model for predicting levels of up-regulation instead of predicting just up-regulation. We found that more highly differentially expressed genes are better predicted (Fig. 4F); however, this model does not perform as well as the model that simply classifies stress responses into up-regulation and no significant change (Fig. 4A).

In this study, we evaluated model performance by cross-validation, dividing our data into training and validation sets. We found that models considering pCRE location, copy number, and level of differential expression were likely overfitted because these models do not lead to further improvement in precision and recall. These more complicated models may explain the training data very well but not the validation data. Thus, there are likely limitations to how much information one can extract from the expression dataset in building cis-regulatory models. Nonetheless, our findings highlight the relative importance of combinatorial regulation compared with other cis-regulatory features; despite the very large parameter space (large number of possible pCRE combinations), it still outperforms the model considering motif numbers, location, or level of expression.

**Conclusion**

Our studies led to the discovery of 1,215 pCREs with multiple properties that resemble experimentally identified cis-elements. In addition, we provided a comprehensive first look at plant stress cis-regulatory codes on the basis of presence or absence of pCREs, their combinatorial relationships, locations in putative promoters, and copy number. Our ability to use pCREs to make reasonable expression predictions provides additional support for the notion that the pCREs identified computationally are likely authentic cis-elements. Furthermore, prediction accuracies of regulatory models based on binary relationships are much higher compared with those of presence/absence-based models. There are very few documented binary interactions between plant CREs. Thus, the combinatorial rules we have uncovered provide prioritized targets for experimentation.

Despite the importance of motif location and copy number in transcriptional regulation, considering these cis-regulatory features does not lead to better performing models. Thus, there is clearly room for improving the stress cis-regulatory model. Aside from optimizing the parameters of the analysis steps, performance would be further improved if additional information is incorporated, including transcription factor–CRE relations, transcription factor level, site affinity, or epigenetic states under stress conditions. For example, it would likely be very informative to consider information about how the expression patterns of transcription factors are affected by stress (38) and which transcription factors bind to pCREs on the basis of genome-wide one-hybrid studies (39). Further iterations of model building considering the above information in conjunction with experimental verification will be necessary for a more detailed global stress cis-regulatory code.

**Methods**

Putative CRE Identification, Assessing Location Bias, and Conservation. The overall procedure for identifying pCREs among stress-responsive genes is shown in Fig. 1A. See SI Methods, section 1 for details on pCRE identification,
pCRE mapping, random sequence generation, and statistical analysis of positional bias. Dataset S1 contains information on the pCREs identified. Information on pCREs identified in this study is also available in The Arabidopsis Information Resource (TAIR). To evaluate the conservation of pCREs, we first identified constrained sites in aligned genomic sequences among 1-to-1-to-1 syntenic orthologs in A. thaliana, A. lyrata, and Populus trichocarpa. A pCRE was defined as conserved between species if it mapped to constrained sites significantly more frequently than its randomly shuffled counterpart. The constrained sites were identified through comparisons of observed substitution patterns in each aligned putative promoter site against a neutral evolution model derived from substitution rates of fourfold degenerate sites in orthologous coding sequences (SI Methods, section 1.4).

Building pCRE-Based Models for Predicting Stress-Responsive Transcription. To determine how well the presence or absence of pCREs explains stress-responsive transcription, we used the Support Vector Machine (SVM) (40) algorithm to generate classification models for predicting expression with twofold cross-validation. SVM was used to generate expression prediction models on the basis of (i) only motif (known or pCRE) presence or absence, (ii) pCRE location, (iii) pCRE copy number, and (iv) levels of differential expression. To assess the importance of pCRE location, a “weight” that reflects the importance of pCRE sites in a given location bin was calculated (SI Methods, section 2.3). The importance of copy number was examined by cross-validation. SVM was used to generate expression prediction models for discussion and critical comments on this manuscript, Yang Zhou and Jinfeng Yi for advice on group lasso, and two anonymous reviewers and the editor for very helpful comments and suggestions. This work was partially funded by the National Science Foundation (Grant MCB-0748634 and DEB-0919452 to S.-H.S.) and DBI-0701709 (to M.F.T.).

Experimental Validation of pCRE and Combinatorial Rules. The β-glucuronidase (GUS) reporter gene constructs were generated by cloning truncated promoters, both wild-type and mutated variants, into the Gateway cloning vector PMDC163 (42). Mesophyll protoplasts were isolated from fresh A. thaliana rosette leaves, and ∼ 5 x 10^6 protoplasts were transformed with the constructs (43). After incubation at room temperature for 13 h under light, the protoplasts were evenly divided into two subsamples and treated for 10 min with water and 250 mM NaCl, respectively. Fluorometric assays of GUS activity were performed using the fluorogenic substrate 4-methylumbelliferyl glucuronide and were normalized with the total protein content (Bio-Rad Laboratories).

Acknowledgments. We thank David Arnosti and Melissa D. Lehti-Shiu for discussion and critical comments on this manuscript, Yang Zhou and Jinfeng Yi for advice on group lasso, and two anonymous reviewers and the editor for very helpful comments and suggestions. This work was partially funded by the National Science Foundation (Grant MCB-0748634 and DEB-0919452 to S.-H.S.) and DBI-0701709 (to M.F.T.).