CIS-REGULATORY CODE CONTROLLING SPATIALLY SPECIFIC HIGH SALINITY RESPONSE IN *ARABIDOPSIS THALIANA*

By

Alexander Seddon

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Plant Biology - Master of Science

2015

ABSTRACT

CIS-REGULATORY CODE CONTROLLING SPATIALLY SPECIFIC HIGH SALINITY RESPONSE IN *ARABIDOPSIS THALIANA*

By

Alexander Seddon

Plants are subjected to a variety of environmental stress, and their ability to respond to stress depends, in a large part, on the proper regulation of gene activities including transcription. Earlier studies show that the regulation of stress transcriptional response has a significant spatial component, namely, each organ, tissue, and cell type may respond to a stress by differentially regulating different sets of genes. Although our knowledge is accumulating on how specific transcription factors (TFs) and their associated *cis*-regulatory elements (CREs) are involved in stress responses, a genome wide model of what plant TFs and CREs are key to the spatial stress response regulation has yet to emerge. In this study, a set of 1,894 putative CREs (pCREs) were identified that are associated with salt stress up-regulated genes in the root and shoot of *Arabidopsis thaliana.* These pCREs led to models that can better predict salt up-regulated genes in root and shoot compared to models based on known TF binding motifs. The full pCRE set could be broken into root, shoot and general subsets that are enriched amongst root, shoot, or both root and shoot salt up-regulated genes, respectively. We also identified pCRE subsets that are enriched amongst genes induced by salt in root cell-types. Most importantly, combinations of the pCRE subsets allowed predictions of genes up-regulated by high salinity in root, shoot, as well as various root cell types. In addition, consideration of pCRE combinatorial rules further improved salt up-regulation prediction. Our results suggest that the organ and cell-type transcriptional response to high salinity is regulated by a core set of pCREs that need to be considered in combinations, and provides a genome-wide view on the *cis*-regulation of spatial transcriptional responses to stress.

To my sister, Kassie Seddon,
my aunt and uncle, Christine and Stephen Seddon,
and my partner, Melissa Shaner.

ACKNOWLEDGMENTS

The work in this thesis would not be possible without the help of many other people. I would like to thank them all in these acknowledgments. My thesis advisor Dr. Shin-Han Shiu for all his help with the development and execution of this project. Sahra Uygun, for performing the analysis on the cell-type specific aspect of this project. Melissa Lehti-Shiu, Ming Jung Liu, Sebastian Stankewicz, Bethany Moore, and Christina Azodi, for working on experimental validation of the computational work presented in this thesis. Nicholas Panchy for his development of the randomized sequence motif scripts. Johnny Lloyd for useful discussions on machine learning and data analysis. My graduate committee members, Dr. David Arnosti and Dr. Yanni Sun for graciously providing their time to read and evaluate my work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

TF	Transcription Factor
CRE	*Cis*-regulatory element
pCRE	Putative *cis*-regulatory element
DNA	Deoxyribonucleic Acid
GO	Gene Ontology
FET	Fisher's Exact Test
TFBM	Transcription Factor Binding Motif

INTRODUCTION

Plants are equipped with an awe inspiring range of mechanisms to respond to environmental stresses such as excess heat, salinity, drought, and pathogen attack. Among the stress response mechanisms that are indispensable for plant survival, many operate through changes in gene expression which ultimately impact physiological and developmental responses to stress [1–4]. For example, plants subjected to abiotic stresses have a change in the expression of a core set of stress response genes [4,5] as well as genes specific to each stress [5]. Stress induced gene expression changes are not uniform across the whole plant. Instead, plant organs display substantial differences in which genes are differentially expressed under stress [1,4,5]. In the case of salt stress, the primary response of the root is excluding sodium from the xylem and sending hormonal signals of stress to the shoot, while the shoot must respond to the effects of ion toxicity and limitation of water [6,7]. In addition to the organ level changes in transcription, the transcriptional responses to stress tends to be more specific at the cell-type level in an organ. For example, more genes were considered differentially expressed during salt or iron-deprivation stress when measured in individual root cell-types than if expression was measured across the whole root [2]. The research highlighted so far illustrates that a plants response to stress involves a significant change in the transcriptome of the plant, and the transcriptome change varies across spatial levels of the plant. This prior research raises the question of how transcription is spatially regulated during stress.

Amongst the many mechanisms of transcriptional regulation, the roles of transcription factors (TFs) and their associated *cis*-regulatory elements (CREs) in gene regulation during stress have received considerable attention [8]. An example is the CBF/DREB1 in *A. thaliana*, a TF that regulates genes during cold and drought stress whose consensus binding CRE is G/ACCGAC. The

CRE for CBF/DREB1 is found on a subset of genes that are induced under drought and cold stress [9]. As the availability of gene expression and TF binding data has increased, a number of studies have focused on identifying CREs on the promoters of co-expressed genes [10–13] or through *in vitro* and *in vivo* determined TF binding preferences [14–16]. The co-expression approach was successfully used to identify putative CREs (pCREs) that were used to generate accurate models of stress expression patterns for genes in yeast (*Saccharomyces cerevisiae*) [10] as well as *A. thaliana* [12]. Another study in yeast utilizing CREs identified through TF binding data uncovered a complex regulatory code involving combinations of multiple CREs [14]. One major conclusion from this line of research is that all the TFs and CREs do not work independently to regulate genes. Instead, the promoter of a gene is composed of multiple different CREs which have an additive effect on that genes expression pattern. For example, 16% of *A. thaliana* genes with the CRE known as the Drought Response Element (DRE) on the promoter were up-regulated by salt in the shoot, but DRE was only present on 13% of the salt up-regulated genes [12]. However, models using 1,215 pCREs simultaneously resulted in more precise predictions of salt stress up-regulated genes [12] then looking at individual CREs. The set of all CREs involved in gene regulation, along with all the relevant combinatorial rules of the CREs will be referred to as the "*Cis*-regulatory Code" (CRC). Understanding the genome wide response of an organisms to stress requires the discovery of the CRC for that organism.

Much like stress responsive gene regulation, there is evidence that spatial gene regulation is also governed by a CRC involving multiple CREs working in combinations. We would like to note that much of this research has focused on spatial gene regulation, but not in the context of stress. In human studies, it has been demonstrated that genes expressed in specific tissues are regulated by particular combinations of TFs and CREs [17], but the TFs themselves are less

specific to expression in an given tissue [18]. This suggests that combinations of TFs and their associated CREs may be more critical to spatial gene regulation than individual CREs. Spatial transcriptional atlases in plants have begun to emerge in recent years. Analysis of a rice transcriptome atlas revealed that there were short sequence motifs that are overrepresented in the 1kb promoters of genes expressed in specific tissues and cell-types [19]. In addition, some of these sequence motifs overrepresented on spatially specific genes are similar to the known binding motifs of TFs involved in regulating genes in response to stress [19], suggesting CREs related to stress response may also be involved in spatially specific transcription regulation. Nonetheless, there has not been a systemic study of spatially specific *cis*-regulation in the context of stress. We expected that a co-expression CRE identification framework could potentially explain the spatial stress response in plants.

Our objective in this study was to uncover the mechanism regulating genes in specific organs and cell-types under stress. Specifically, we were interested in the CREs that are involved in spatially regulating genes under salt stress in *A. thaliana*. Salt stress was chosen because it is well studied both physiologically [6,7] and molecularly [20,21]. There are already known transcriptional machineries for salt stress which we can use to verify some of our results [21–24], and there are known differences in the transcriptional response to high salinity in the root and shoot [1,4] and in root cell types [2,3]. To assess transcriptional changes to stress across different spatial levels in *A. thaliana*, we first looked into the similarities and differences of stress transcriptional response in the root and shoot, and asked how salt up-regulated genes in these organs differed in their functional annotation. Next, to address the question of how well current knowledge of CREs in *A. thaliana* can explain spatial salt stress up-regulation, we used CREs identified through *in vitro* derived TF binding motifs (TFBMs) [16] and co-expression derived pCREs ("Zou pCREs")

[12] to generate models of salt stress up-regulated genes in the root and shoot. We then wanted to see if the TFBMs and Zou pCREs may be missing CREs critical to the salt stress up-regulation of genes in the root and shoot. To identify potentially missing CREs, we used a co-expression approach to identify a new set of pCREs of root and shoot salt stress response. With these pCREs we further tested if they individually or in combinations could be used to establish a more extensive *cis*-regulatory model explaining spatial patterns of up-regulation during salt stress. Finally, to see if CREs working at the level of organs are also regulating genes at the cell-type specific level, we looked into how well pCREs identified on the organ level can predict the salt stress up-regulated genes at the level of individual cell-types in the root.

RESULTS AND DISCUSSION

Transcriptional responses to stress have a strong spatial component

Earlier global gene expression studies have demonstrated that different plant organs and cell types have distinct transcriptional response to stress [1–4]. To assess the extent to which organs have unique expression patterns under different stress conditions, and, particularly, to determine the correlations between whole organ (root vs. shoot) and cell-type stress response, we calculated the correlations between the levels of differential expression across multiple conditions and time points using three types of datasets: (1) root and shoot samples under abiotic stress [4], (2) shoot samples under biotic stress, and (3) root cell type samples under salt stress [2] (see **Methods**). There were several patterns worth noting. First, samples tended to cluster together according to stress condition, and these "stress condition clusters" tended to have root and shoot sub-clusters (**Figure 1A, S1 Figure**). For example, osmotic and salt stress samples formed a cluster, with sub-clusters composed of shoot, root and root cell-type samples (**Figure 1A;** dotted rectangles I, II and III, respectively).

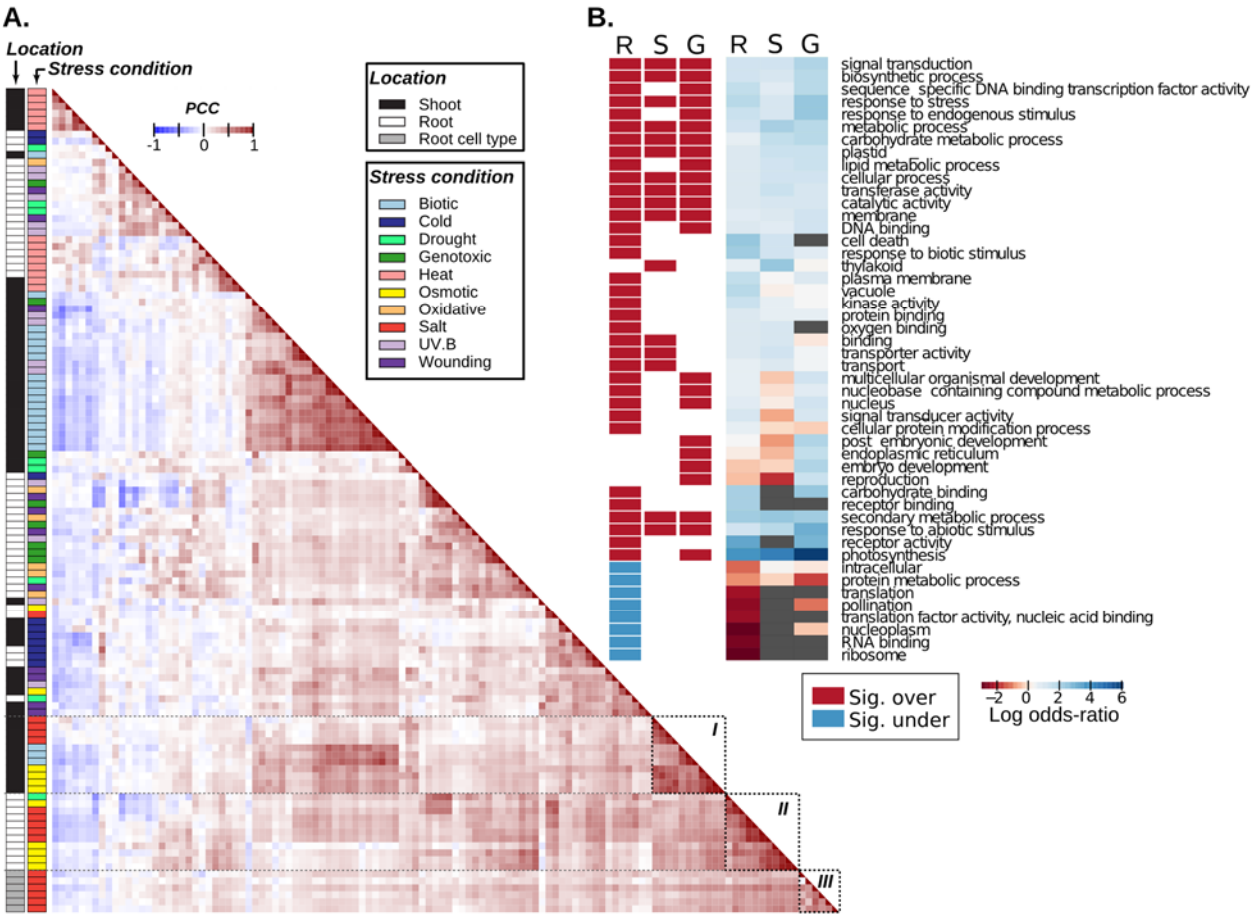

Figure 1. A. thaliana gene expression correlation across stress datasets and Gene Ontology (GO) terms enriched in salt responsive genes.

(A) Between-sample Pearson's Correlation Coefficients (PCC) calculated based on log fold changes ($\log_2$(stress/control)) for genes under each stress condition. The orders of rows and columns are the same, and are sorted based on hierarchical clustering of the pairwise PCC values (see **S1 Figure** for dendrogram and sample details). Side color bars represents the organ/cell-type and stress condition of each row in the heatmap. Dotted rectangles I, II, and III highlight an osmotic and salt stress cluster. **(B)** Heatmap indicating Plant GO Slim terms enriched in genes that are differentially up-regulated during salt stress after 3 hours specifically in roots (R) or shoots (S), or both organs (globally, G). Left heat map summarizes the terms that are significantly over or

Figure 1. (Cont'd)

underrepresented in each gene set (red: significantly overrepresented, $p \leq 0.05$; blue: significantly underrepresented $p \leq 0.05$). Right heatmap summarizes the $\log_2$ odds ratio from the enrichment test (grey: data not available). The terms shown are those significantly enriched in R, S, and/or G gene sets. For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this thesis. All supplemental figures and tables (numbered sequentially as S1, etc.) are available with the online version of this thesis.

The presence of organ sub-clusters does not necessarily mean that organ was the primary determinant of the expression correlations between samples because the median Pearson's Correlation Coefficient (PCC) of the log fold-change values between samples from the same organ (0.17) were significantly lower than those between samples from the same stress condition (0.31) (Mann-Whitney, p=2.2e-16). Instead, the expression correlations were strongly influenced by both the specific stress conditions and the organs, with the stress condition playing a more influential role. Specifically, we found that samples from the same condition and organ (but with different treatment durations) had a median PCC of 0.37, significantly higher than the median PCC within the same condition but between different organs (0.22) (Mann-Whitney, p=2.2e-16). Some of the stress conditions had stronger organ/cell-type specific effects as illustrated in the salt and osmotic stress cluster (dotted rectangles, **Figure 1A**). Nonetheless, the stress response differences between organs were substantial. Salt stress samples from the same organ had a significantly higher correlation (median PCC=0.69) than between organs (median PCC=0.24) (Mann-Whitney, p=2e-16). Next we asked whether the spatial component of stress response is more prominent compared to the treatment duration. We found that expression correlations among samples from the same organ and stress condition (median PCC=0.37) tended to be higher than those among samples with the same treatment duration and stress condition (median PCC=0.32) (Mann-Whitney, p=0.02). The higher correlation within organs suggests that spatial response is more influential than treatment duration when it comes to differential expression under the conditions examined.

Another observation is that root cell-type samples under salt stress were clustered with the whole root expression data under salt stress (**Figure 1A**, dotted rectangles II and III). Consistent with this clustering pattern, differential expression levels of genes in root cell-type salt samples

were significantly more correlated to whole root salt stress samples than to shoot salt stress samples (median PCC=0.36 vs. median PCC=0.22; Mann-Whitney, *p*=2e-09). As expected, this pattern indicates that the whole root data may be capturing the expression patterns observed in the individual root cell types, and it aligns with an organ-tissue-cell type transcriptome clustering hierarchy observed in rice [19]. Our results extend upon this observation, suggesting that each stress condition may elicit a core transcriptional response, but with more refinement at lower spatial levels. Taken together, our findings demonstrate that, while there is a specific transcriptional response to each stress, this response is further influenced by spatial considerations such as the organ and cell-type where genes are expressed.

Roots and shoots differ in the types of genes up-regulated during salt stress

Our observation of related but distinct transcriptional responses to stress in the root and shoot of *A. thaliana* suggest that differences in differentially regulated genes between these organs may contribute to different physiological responses to stress. Thus, we asked what types of genes were differentially regulated within each the root and shoot during salt stress. Here we focused on salt stress for two reasons. First, earlier studies show that responses to salt stress have a strong spatial component [1–4] and we reinforced this notion with an expanded analyses (dotted rectangles, **Figure 1A**). Second, we know many of the TFs and the CREs that are involved in the regulation of transcription during salt stress [1,21–23], which gives us a basis to verify some of our results. We focused on up-regulation because our ultimate goal was to model gene expression, and previous attempts to model down-regulation using CREs have been unsuccessful, which may be indicate that post-transcriptional regulation has a larger role in down-regulation of genes under stress [12].

To test the hypothesis that there are significant differences in the functional categories of genes up-regulated by salt stress in the *A. thaliana* root and shoot, we performed an enrichment analysis on the Gene Ontology (GO) terms of genes up-regulated during salt stress (see **Methods**). Enrichment tests were performed on three sets of significantly up-regulated genes after 3 hours of salt stress: (1) "globally up-regulated": genes up-regulated in both the root and the shoot, 247 genes; (2) "root specifically up-regulated": genes only up regulated in the root, 1853 genes; and (3) "shoot specifically up-regulated": genes only up-regulated in the shoot, 277 genes. Dividing the genes up in this manner allowed us to look at the related and distinct parts of up-regulated genes in the root and shoot. There were 48 GO terms that were significantly over/underrepresented in at least one of the gene sets defined above (**Figure 1B**). There was only one term (thylakoid) that was only overrepresented among shoot specifically up-regulated genes. The overrepresentation of this term in the shoot is consistent with the fact that roots do not have plastids with thylakoid membranes and photosynthesis is significantly impacted by salt stress conditions [25]. The remaining 47 enriched terms are at least enriched among the root specifically and/or globally up-regulated genes. One of the terms overrepresented in all three gene sets was signal transduction (**Figure 1B**). Since these three gene sets are mutually exclusive, this result suggests that the root and the shoot are up-regulating signaling pathway genes unique to each organ, as well as pathways that are globally necessary for salt stress response. This result is supported by work on the SOS pathway [20,26], where the root and shoot have alternative calcium binding proteins (SOS3 and SCaBP8) involved in regulating the SOS2 genes. The SOS pathway is involved in signaling for the cell to extrude sodium from the cytosol [26], and it involves components that are common to both organs as well as specific to the root and shoot [20,26]. Some terms were only enriched amongst the root specifically and globally up-regulated genes. For example, "sequence-

specific DNA binding transcription factor activity" and "DNA binding" are both overrepresented in root specifically and globally up-regulated genes but not in the shoot specifically up-regulated genes, suggesting that there is a set of global TFs, and another set specific to the root (**S1 Table**). Thus, genes up-regulated in the root may be regulated by both a global and specific set of TFs, while genes up-regulated in the shoot may be regulated primarily by a global TF set. This possibility is explored further in later sections.

To summarize, a variety of functional categories were found to be enriched in the genes up-regulated by salt stress. In some instances, root specifically, shoot specifically and globally up-regulated genes have the same enriched functional categories. These common enriched terms suggest that roots and shoots are up-regulating similar types of genes, but that the specific genes up-regulated are different. In addition, we found evidence to suggest that there are global TFs up-regulated in both roots and shoots, while the roots have their own specific set of up-regulated TFs. The root specifically TFs may help to explain the differences in expression pattern that we see between the root and shoot under salt stress. Thus, the root specifically up-regulated genes may be controlled by the root specific TFs. Because TF differ in the CREs they bind, it is reasonable to hypothesize that genes up-regulated by salt in the root may have a specific set of CREs on their promoters.

Known TF binding motifs contribute to a better than random performing model for salt stress up-regulation prediction

The differences in the high salinity transcriptional response between root and shoot suggest that there are significant differences in the mechanisms regulating organ specific responses to high salinity. Additionally, our GO analysis suggested that the difference may be due to which TFs are present in each organ during salt stress. Because TFs exert their regulatory roles by binding to

CREs, we expected that the global and organ specific activities of TFs will be reflected in what CREs are located in the regulatory regions of globally, root-specific and shoot-specific up-regulated genes. We hypothesize that each organ has a different set of CREs regulating its salt stress up-regulated genes and these CREs can be used to construct CRCs that are models for predicting stress responsive gene expression [12]. To test these hypotheses, we first collected a comprehensive dataset of 355 TF binding motifs (TFBMs) from the CisBP database (**S2 Table**; [16]). Because TFs from the same family tend to have highly similar binding profiles [16] and the 355 TFBMs are derived from 27 of the 47 *A. thaliana* TF families defined in Weirauch et al. [16], we expected that this dataset might capture much of the "*cis*-regulatory space" for constructing reasonable CRCs for predicting two types of salt up-regulated genes: (1) "root up-regulated genes": genes up-regulated in the root, (2) the "shoot up-regulated genes": genes up-regulated in the shoot. Note that the root up-regulated genes are the union of the root specifically and globally up-regulated genes, and the shoot up-regulated genes are the union of the shoot specifically and globally up-regulated genes used for the GO analysis in this study.

We first tested if there is a difference in which TFBMs were significantly overrepresented on the putative 1kb promoter regions of root and shoot up-regulated genes. We found that 22 and 25 TFBMs are overrepresented on the promoters of root and shoot up-regulated genes, respectively (**S2 Table**). Among the overrepresented TFBMs, 81% and 88% for the root and shoot, respectively, were binding motifs for either bHLH or bZIP TFs. However, the overrepresented TFBMs for root and shoot up-regulated genes had 20 TFBMs in common, raising the question whether the TFBMs differentially enriched between up-regulated genes in root and shoot could explain the differences in root and shoot response. To address this question, a machine learning method, Support Vector Machine (SVM, [27,28]), was used to establish models for predicting

whether a gene was up-regulated or non-responsive to salt stress in one of the organs based on the presence and absence of TFBM sites in the putative promoter regions (see **Methods**). We used the "Area under the receiver operating characteristic" (AUC-ROC) as the performance measure – where an AUC-ROC=1 indicates a model that makes perfect predictions and AUC-ROC=0.5 indicates that a model performs as well as random predictions (**Figure 2A**). An alternative way to visualize performance, the precision-recall curve, was also provided where precision is the proportion of predicted genes that are truly up-regulated in an organ and recall is the proportion of truly up-regulated genes in an organ that are correctly predicted (**Figure 2B,C**). The precision-recall curve of a model focuses on the prediction of up-regulated genes, where curves tending more towards the upper-right corner of the graph represent better models.

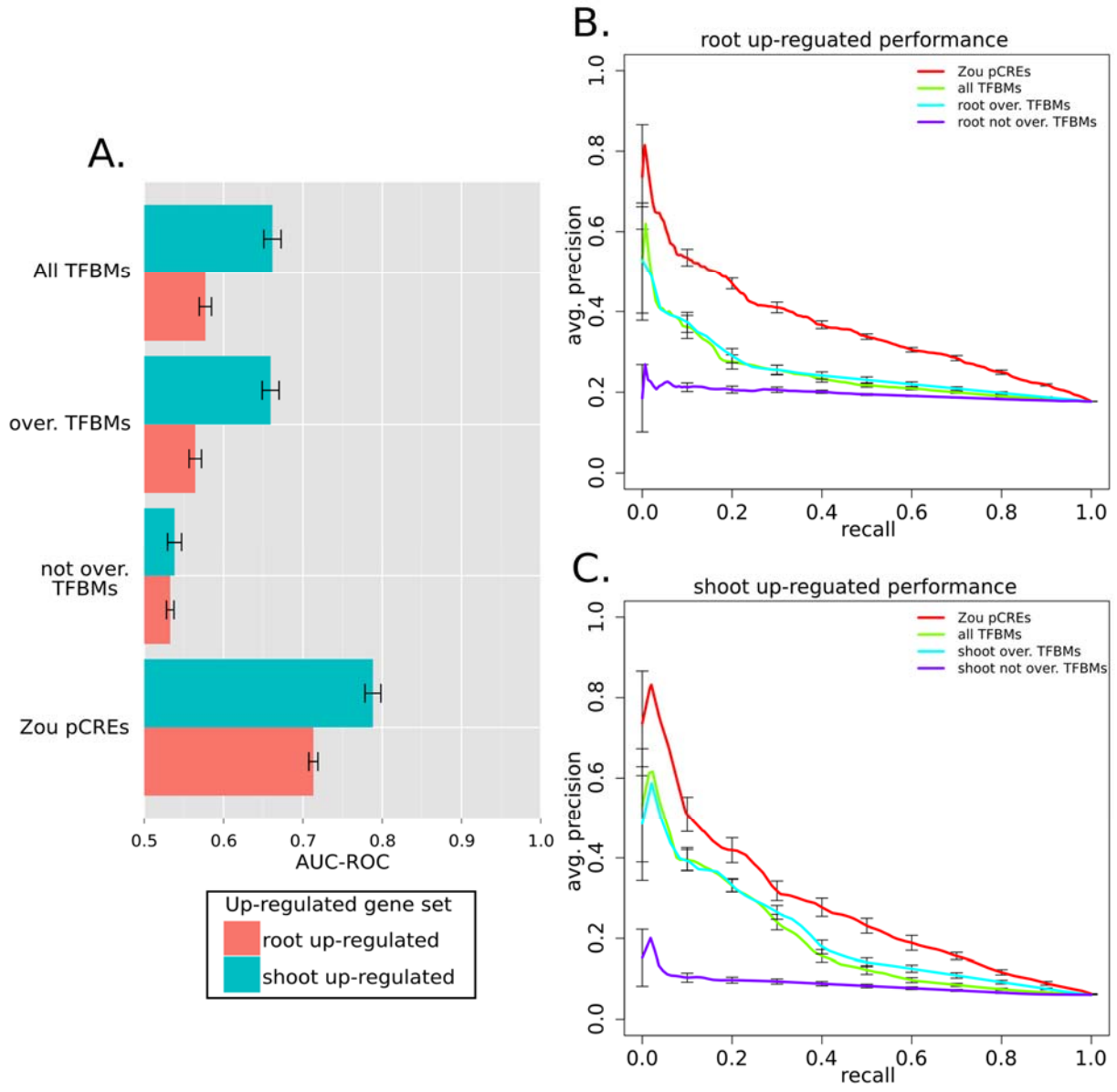

Figure 2. Performance of high salinity up-regulation prediction models using known TFBMs and the Zou pCREs.

(A) Bar plots showing the AUC-ROC values for predicting root (red) and shoot (blue) salt up-regulated gene with models using all TFBMs from CisBP [16], overrepresented ("over.") TFBMs, TFBMs that are not over-represented ("not over") and the Zou pCREs [12]. Error bar: standard error of AUC-ROCs from 10 fold cross validation in each model. **(B)** Precision-call curves for models predicting

Figure 2. (Cont'd)

root up-regulated genes. **(C)** Precision-recall curves for models predicting shoot up-regulated genes.

SVM models of salt stress up-regulation in root and shoot were first established using the root and shoot overrepresented TFBM sets in the promoters of root and shoot up-regulated genes. The models based on these overrepresented TFBMs led to slightly better than random prediction in the root (AUC-ROC=0.57) and shoot (AUC-ROC=0.66) (**Figure 2A**). Predictions based on a model where promoters and gene up-regulation class were shuffled made near random predictions (AUC-ROC=0.51 and 0.53 for root and shoot up-regulated genes, respectively), confirming that the overrepresented TFBM based models were performing better than random guessing. Note that models based on using all 355 TFBMs resulted in the same AUC-ROC values (**Figure 2A**) and similar precision-recall curves (**Figure 2B,C**), suggesting that only the overrepresented TFBMs are relevant for modeling root and shoot up-regulation. Because the TFBM dataset was available for 34% of the known *A. thaliana* TFs, many relevant CREs might not be included in the TFBM-based models. One approach that could potentially overcome this limitation in coverage is to search for sequence motifs among co-expressed genes [11–13,18]. Consistent with this possibility, 1,215 pCREs ("Zou pCREs") identified based on co-expression across shoot biotic and abiotic stress response data [12] led to a substantially better performing model of shoot salt stress up-regulation (AUC-ROC=0.78) than the models based on the TFBMs (AUC-ROC=0.66) (**Figure 2A,C**). This has two implications. First, the TFBMs currently available do not sufficiently represent the *cis*-regulatory space to model salt-stress up-regulation. Second, pCREs identified from co-expression clusters can, at least partially, provide complementary *cis*-regulatory information important for controlling salt stress responses in both the root and shoot.

It is worth noting that performance of modeling root up-regulated genes is not as good as modeling salt up-regulated genes in shoots with the Zou pCRE set (**Figure 2**). The lower performance on root expression predictions may be because the Zou pCREs were identified using

co-expression clusters based only on shoot expression data [12]. Thus to improve upon our understanding of what CREs are associated with and how these CREs may influence salt stress up-regulation in the root and shoot, it is likely important to identify CREs using both root and shoot expression data. This approach may help to capture the regulatory differences between the root and shoot during salt stress. Thus we next asked the question of how the regulatory logic differs between the root and shoot salt up-regulation.

pCREs derived from co-expression clusters are similar, but not identical, to the known binding motifs of TFs

We have shown that modeling up-regulation under salt stress with the Zou pCRE [12] set generated from co-expression clustering are better than using TFBMs derived from *in vitro* binding data [16]. We hypothesized that the inclusion of root expression data to generate co-expression clusters may contribute to an expanded pCRE set that can lead to better machine learning models explaining salt stress up-regulation in the root and shoot. To test this hypothesis, a previously established pCRE identification pipeline [12] was applied to root and shoot abiotic stress fold-change data. The application of this pipeline resulted in 1,894 pCREs ("the full pCRE set") significantly overrepresented in the promoters of root and/or shoot salt up-regulated *A. thaliana* genes (see **Methods, S3 Table, Figure 3A**).

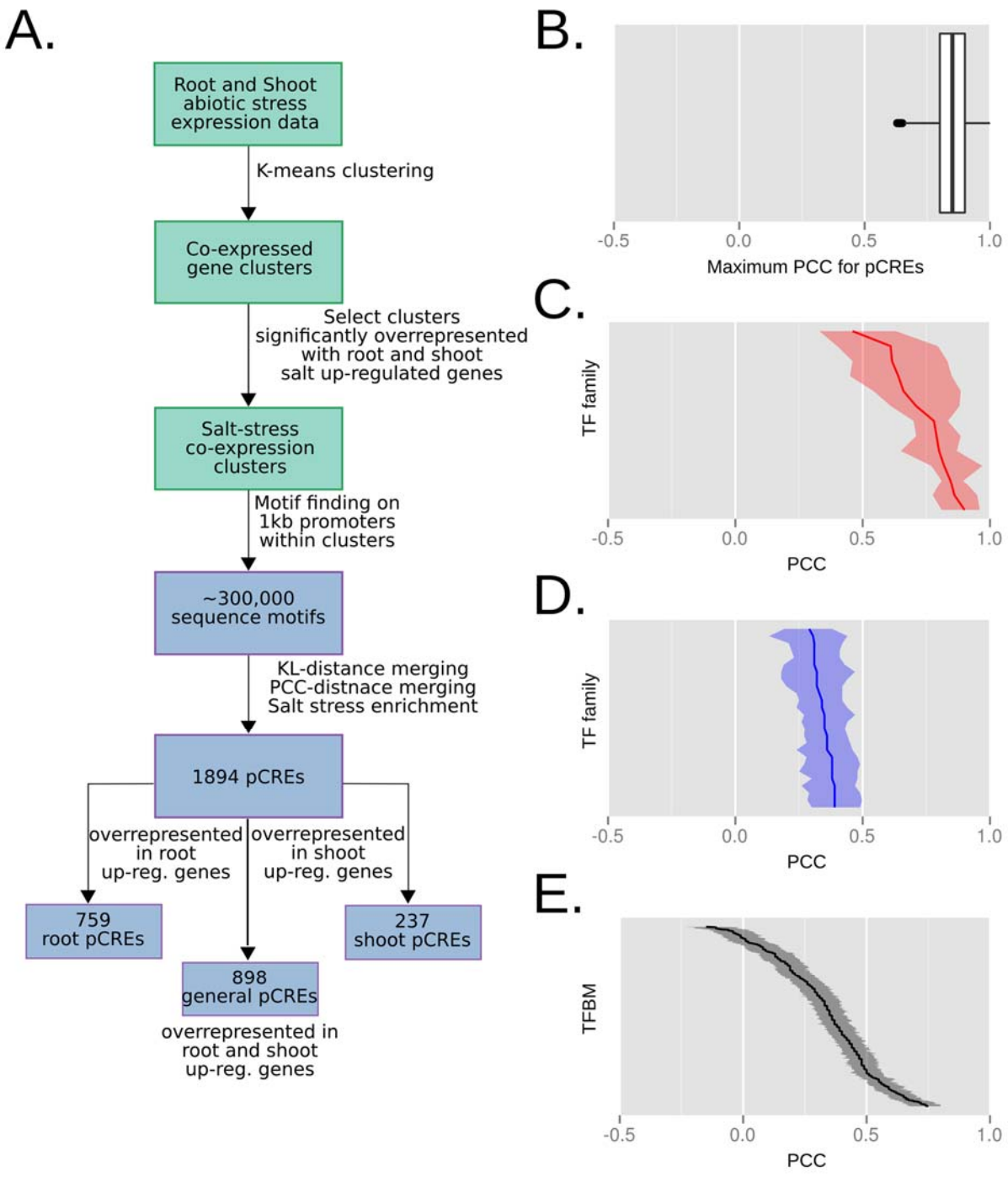

Figure 3. CRE identification pipeline and pCRE comparison to TFBMs.

(A) Overview of the pCRE identification pipeline. Please see **Methods** for complete details. **(B)** The boxplot represents the highest PCC for a pCRE compared to the TFBMs [16]. **(C)** The y-axis

Figure 3. (Cont'd)

represents TF families, Red line represents the median PCC between pairs of TFBMs within the same family, and the shaded ribbon represents the 25th and 75th percentile of PCCs. **(D)** Same format as **(C)**, but the blue line and ribbon summarizes comparisons between TFBMs of one family to TFBMs of all other families. **(E)** Has the same format as **(D)** and **(C)**, but the y-axis represents individual TFBMs, and the black line and ribbons represent comparisons between the TFBM and random motifs.

To determine whether the full pCRE set is a significant expansion in *cis*-regulatory space that can explain salt responsive up-regulation in the root and shoot compared to the 355 TFBMs [16], we first asked if the motifs from the full pCRE set were similar to known TFBMs. To answer this question, we calculated the PCC values of the position weight matrices (PWMs) of all motif pairs between pCRE and TFBM sets to find the best matching pCRE-TFBM pairs. A pCRE that is highly similar to a TFBM has a PCC of 1, with lower PCC values indicating diminishing similarity (**Figure 3B**. Only two experimentally determined TFBMs for two bZIP TFs, ATBZIP63 (AT5G28770) and ABF3 (AT4G34000), had a PCC of 1.0 with at least one of the pCREs. ABF3 regulates genes involved in abscisic acid (ABA) signaling [29]. ATBZIP3 is involved in the regulation of ABA synthesis genes during abiotic stress [24]. To determine if the PCC value of a pCRE-TFBM pair is high enough to indicate that the pCRE is likely bound by the same TF as the TFBM in question, a background PCC distribution was established between all possible TFBM pairs within each TF family (**Figure 3C**, see **Methods**). The rationale is that if the PCC value of a pCRE-TFBM pair is significantly higher (at the 5% level) than a random pair of TFs from the same family, the pCRE is likely bound by the same TF specified by the TFBM. Based on this criteria none of the pCREs was significantly similar to any TFBM. Thus, much of the full pCRE set do not resemble the top binding sites of TFs as represented by the TFBM dataset.

Next we asked whether a particular pCRE is likely bound by TF(s) from a particular family. To test this, a background PCC distribution was established based upon comparing the TFBMs across TF families (blue line, **Figure 3D**, see **Methods**). We found that 25% of motifs in the full pCRE sets had significant matches to ≥1 TFBMs from 24 of the 27 TF families represented in the TFBM dataset ($p \leq 0.05$; **S2 Table**). That is, these pCREs were more similar to a TFBM from a particular TF family than among TFBMs from different TF families. The pCREs with a significant

TF family match were over-represented in four TF families (EIN3, bZIP, bHLH, CxC; **S3 Figure**). Consistent with the relevance of some of these pCREs in regulating salt response, it is known that the *ein3* mutant had decreased salt stress tolerance [30], and a double mutant of two EIN3 family members, *ein3* and *eil1*, had reduced growth under salt stress [31]. While 25% of the organ pCREs are significantly similar to ≥1 TFBMs, what should be made of the remaining 75% of full pCREs? It is possible that these pCREs are a TF binding motifs without a good representative in the TFBM dataset. To test this possibility, we asked if the pCREs are more similar to a TFBM than to sequences motifs randomly drawn from the genome and we compared the PCCs between pCREs and their best matching TFBMs to a distribution of PCCs between 1,894 randomly generated motifs and TFBMs (see **Methods**; black line, **Figure 3E**). The null hypothesis for this resulting distribution was that PCCs between pCREs and TFBMs were the same as PCC values between a TFBM and random sequences. Contrary to this, we found that 60% of the full pCREs were in the 100th percentile of random PCC (**Figure 3A**). Furthermore, all of the pCREs had a TFBM match that is higher than the 96th percentile of the random motif PCC distribution (**Figure 3A**). These findings suggests that our full pCREs are not simply random, meaningless sequences pulled from the genome. Instead, the full pCRE set contained motifs that were more similar to TFBMs that are *bona fide* binding motifs for transcription factors, although these pCREs did not appear to be bound by the same TFs.

Our results from the GO analysis led us to hypothesize that the shoot and root up-regulated genes may be regulated by a core set of CREs, associated with the global TFs, while the root up-regulated genes may have an additional set of CREs associated with root TFs. To explore this possibility, we attempted to identify organ associated subsets of the pCREs. Among 1,894 motifs in the full pCRE set, there are three motif subsets that were overrepresented in the promoters of

salt up-regulated genes in the root ("root pCREs", 759), in the shoot ("shoot pCREs", 237), as well as in both root and shoot ("general pCREs", 898). We will call the union of the root pCREs and shoot pCREs “organ pCREs”. This finding suggests that there are different pCREs involved in regulating salt response in each organ.

In summary, we identified 1,894 pCREs associated with salt up-regulated genes in the root and shoot of *A. thaliana*. Only two pCREs were likely bound by TFs with known TFBMs, and 75% of the pCREs did not have sufficient similarity to a TFBM to suggest they were likely bound by TFs from the same family. Our limited ability to pair pCREs to TFBMs could be because the TFBMs are only covering a subset of the TF in *A. thaliana*. Alternatively, the pCREs may include sequence motifs that better reflect the binding of TF *in vivo*. Nonetheless, for pCREs with significant matches, they belong to TF families with known regulatory roles in salt stress responses. Furthermore, all the organ pCREs are more similar to ≥1 TFBMs than would be expected if the pCREs were random motifs generated from the genome. Together with the fact that motifs in the full pCRE set were overrepresented among root and/or shoot up-regulated genes, these findings suggest that the co-expression-based analysis contributed to an expanded set of motifs that are relevant for salt stress up-regulating genes in the root and shoot.

Motifs in the full pCRE set further improve salt up-regulation prediction in a spatially specific manner

To assess if the full pCRE set are an improved motif set compared to the TFBMs [16] and the Zou pCRE set [12], we modeled salt induced expression using the full pCRE set with the same SVM method. Salt up-regulation prediction models based on the full pCRE set had better prediction performance for both root up-regulated genes (AUC-ROC=0.76, **Figure 4A**) and shoot up-regulated genes (AUC-ROC=0.80, **Figure 4B**) than the models based on TFBMs (root AUC-

ROC=0.57, shoot AUC-ROC=0.66 , **Figure 2A**) or the Zou pCRE set (root AUC-ROC=0.71 and shoot AUC-ROC=0.78, **Figure 2A**). This improvement in modeling supports our hypothesis that using pCREs discovered from both root and shoot co-expression clusters results in better spatially specific prediction models of salt up-regulation. Considering that the Zou pCRE set was derived from co-expression clusters discovered via a combination of shoot abiotic and biotic stress data [12], it is not surprising that the full pCRE set-based model based on shoot and root abiotic stress data can better predict the root up-regulated genes. However, the improvement seen for the shoot up-regulated genes compared with the Zou pCRE based model was unexpected. This may suggest that the incorporation of root expression data when attempting to discover pCREs help to further refine pCREs discovery associated with shoot salt gene up-regulation.

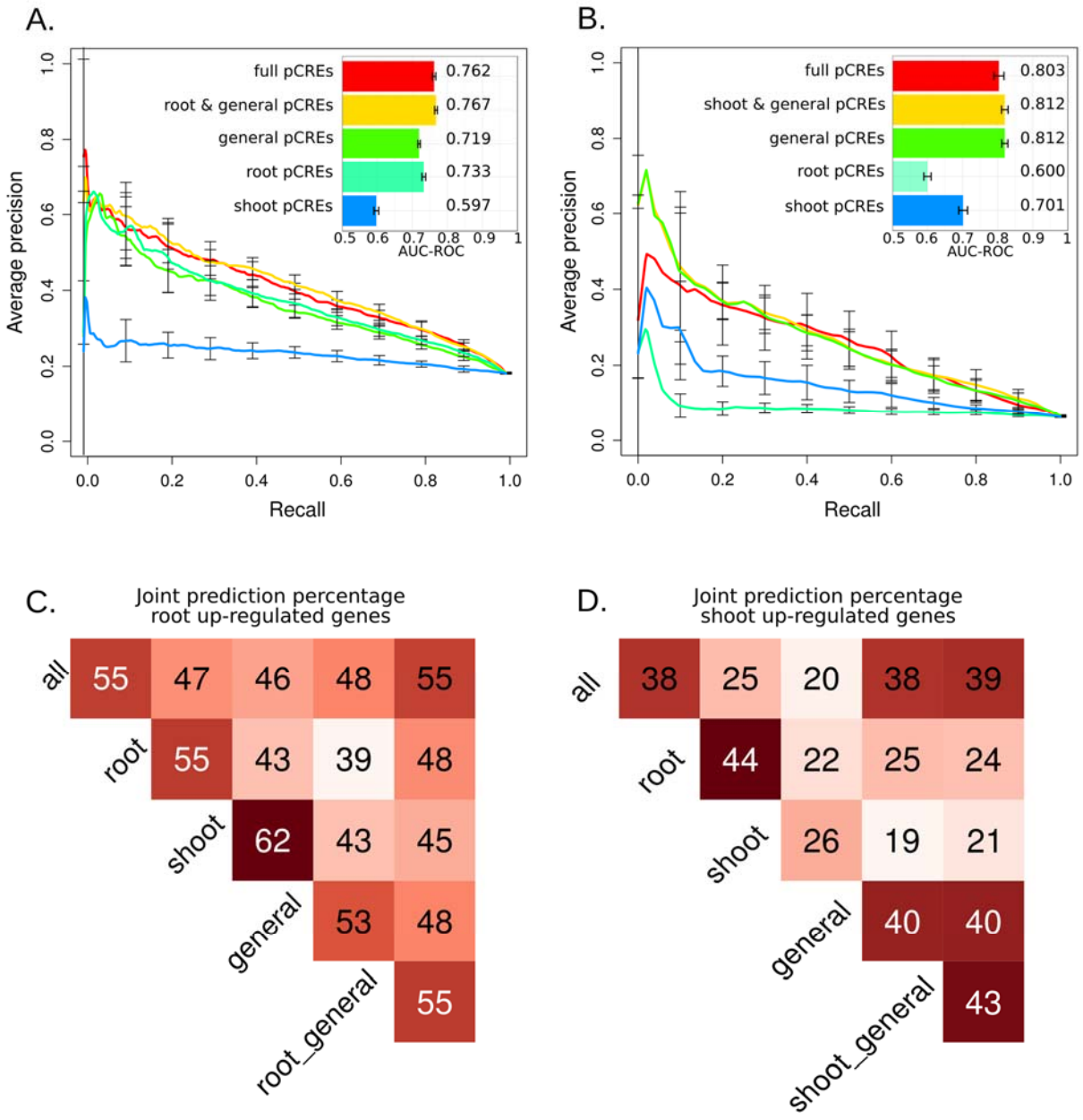

Figure 4. Up-regulated gene prediction performance based on the pCREs.

(A) Prediction performance of models using different subsets of the full pCREs for root up-regulated genes. The line graphs are precision-recall curves for each pCRE set, while the inset bar

Figure 4. (Cont'd)

charts summarize the AUC-ROC. The colors of the curves correspond to the colors of the models on the bar chart **(B)** Performance of models predicting shoot up-regulated genes. **(C)** The median percent overlap for root up-regulated gene predictions of different model pairs. The percent overlap is defined as the percentage of up-regulated genes that are correctly predicted by two models. **(D)** The median percent overlap for shoot up-regulated gene predictions of different model pairs. See **S7 Table** for a summary of statistical comparisons between genes predicted by different models.

Note that the full pCRE set has three subsets: root pCREs, shoot pCREs, and general pCREs that were enriched amongst root, shoot, and root and/or shoot salt up-regulated genes, respectively. One explanation for the presence of these pCRE subsets is that the root and shoot pCREs might be more critical to controlling expression for the root specifically and shoot specifically up-regulated genes, respectively, while the general pCREs might be critical for globally up-regulated genes. To test this hypothesis, salt up-regulation prediction models were established using various combinations of the root, shoot, and general pCREs. For predicting root up-regulate genes, we found that a model based on root pCREs (AUC-ROC=0.73) was much better than a model based on shoot pCREs (AUC-ROC=0.60; **Figure 4A**). Similarly, a model based on shoot pCREs better predicted shoot salt up-regulated genes (AUC-ROC=0.70) than a model based on root pCREs (AUC-ROC=0.60; **Figure 4B**). In other words, the root and shoot pCRE sets were better at predicting up-regulated genes in the organs for which they were associated, demonstrating they are relevant to spatially specific up-regulated genes. Consistent with this notion, for root up-regulated gene prediction, the model based on the union of the general and the root pCREs performed better (AUC-ROC=0.77) than the model based only on the general pCREs (AUC-ROC=0.72; **Figure 4A**), indicating that the root and general pCREs work additively. In addition, combining the general pCREs with the root pCRE set resulted in a model (AUC-ROC=0.77) that performed as well as the model using the full pCRE set (AUC-ROC=0.76; **Figure 4A**). This suggests that shoot pCREs provide no additional information to predict root up-regulated genes. In contrast, for shoot up-regulated genes, although the model based only on shoot pCRE performed reasonably well (AUC-ROC=0.70; **Figure 4B**), it is not as good as the model based only on the general pCREs (AUC-ROC=0.81; **Figure 4B**). More surprisingly, the general pCRE-based model performed as well as a model using both the general pCREs and the shoot pCREs (AUC-

ROC=0.81; **Figure 4B**). That is, adding the shoot pCRE set may not provide additional regulatory information for salt stress up-regulation in the shoots that is not already provided by the general pCREs. This further supports the notion that shoot up-regulated genes may be regulated by a global set of TFs (**Figure 1B**) that bind to set of general pCREs.

The increase in performance when both root pCREs and general pCREs were combined to produce a model of root up-regulated genes suggested that each model captured a distinct subset of the root up-regulated genes. In addition, because AUC-ROC only measures how well but not which non-responsive and up-regulated genes are being correctly predicted, it is possible that similar levels of performance seen in the shoot pCRE-based and the general pCRE-based models in predicting shoot up-regulation could have resulted from the prediction of different sets of genes. To assess if the models are predicting distinct or similar sets of genes, a binary prediction (up-regulated or non-responsive) was generated for each gene using the models based on the different pCRE subsets (see **Methods**). The models will assign a class to each gene, which may or may not be correct. Each pair of models was compared by looking at the percentage of root or shoot up-regulated genes that are correctly predicted by both models (percent overlap, **Figure 4C-D**). Models of root up-regulated genes based on either root pCREs or general pCREs had comparable performance in terms of their AUC-ROCs (0.73 and 0.72, respectively) but their median percent overlap in prediction was only 39%, significantly smaller than the percent overlaps from multiple runs of the root pCRE-based model (median=55%, Mann-Whitney p=3.59e-22) or the general pCRE-based model (median=53%, Mann-Whitney p=3.59e-22; **Figure 4C**). This finding suggests that the root pCRE model may be predicting root specifically up-regulated genes. To test this, we asked if the correctly predicted up-regulated genes from the root pCRE-based model were overrepresented with root specifically up-regulated genes. We found that the root pCRE-based

model was overrepresented with root specifically up-regulated genes compared with other pCRE-based models (**Table 1**). These findings suggest that the root pCREs and the general pCREs are predicting partly different subsets of genes, and explain why combining root and general pCREs leads to an improved model. For shoot up-regulated genes prediction, the shoot pCRE and general pCRE-based models also had significantly lower percent prediction overlap (median=19%) than for multiple runs of either the shoot pCRE based models (median=26%, Mann-Whitney p=1.08e-12) or general pCRE based model (median=40%, Mann-Whitney p=3.42e-22; **Figure 4D**), indicating that the shoot pCRE set does have a unique contribution to predicting shoot up-regulated genes. However, the general pCRE based model correctly predicted a greater number of shoot up-regulated genes at a higher level of precision (**Figure 4B**). Additionally, for shoot up-regulation prediction the overlap in predicted gene sets between the full pCRE-based and the general pCRE-based models (median=38%) was not significantly different from the overlaps between repeated runs of the general pCRE-based model (median=38%, p=0.16), indicating that much of the shoot up-regulation can be explained by the general pCREs alone. This further supports that the general pCREs are likely to be the best set for modeling shoot up-regulated genes, as was suggested from the overall AUC-ROC for models based on the shoot and general pCREs (**Figure 4B**).

Table 1. Prediction of root and shoot specifically up-regulated genes using pCRE based models.

Fisher's Exact Test (FET) on whether the root pCRE based model is biased towards predicting root specifically up-regulated genes and the shoot pCRE based model is biased towards predicting shoot specifically up-regulated genes. Each row represents one test of the root pCRE based model against a model based on another pCRE set. The numbers represent how many of the correctly predicted up-regulated genes (TP) are root specifically up-regulated (Organ specific) and how many are globally up-regulated.

Gene set to predict	**pCRE set used in model**		**Model 1 TP**		**Model 2 TP**		**FET *p***
	Model 1	Model 2	Organ specific	Globally	Organ specific	Globally	
root up-regulated	root	All	1,172	116	1121	189	1.87E-05
	root	Shoot	1,172	116	1111	181	7.52E-05
	root	General	1,172	116	1049	200	1.03E-07
	root	root&general	1,172	116	1030	168	8.88E-05
shoot up-regulated	shoot	All	62	57	105	140	0.116385
	shoot	Root	62	57	100	102	0.728924
	shoot	General	62	57	82	126	0.028346
	shoot	shoot&general	62	57	89	132	0.039873

Taken together, these results demonstrate that the identification of the full pCRE set using stress expression data from both the root and shoot can lead to improvements in modeling gene expression over known TFBMs and the Zou pCRE set identified from shoot expression data alone. This supports our hypothesis that the incorporation of data from both root and shoot would improve CRE discovery. We also found that salt stress up-regulated genes in the root and the shoot may be regulated by different subsets of motifs in the full pCRE set. Genes up-regulated by salt stress in the root can be best predicted with a model considering both the root and the general pCRE sets without considering shoot pCREs. However, the shoot up-regulated genes likely were regulated primarily by general pCREs, as seen in the equivalent performance of the general pCRE model and the full pCRE model of shoot up-regulated genes.

pCREs work best in combinatorial rules

We have shown that the pCREs identified from root and shoot stress co-expression clusters resulted in improved models for predicting salt stress gene up-regulation in an organ-specific manner. We have also shown how the incorporation of root and shoot expression data can further improve organ salt stress up-regulation prediction models compared to using shoot data alone. In addition to identifying individual CREs that may up-regulate genes in organs, the combinations of the pCREs may be important for regulating expression under stress conditions [12,14], as well as determining tissue specific expression [17,18]. Currently there is no comprehensive study of how CRE combinations may regulate both the spatial pattern (in our case the organs) and the environmental response (stress condition) of gene expression. Therefore, we asked: (1) whether pCRE combinations are important for salt stress up-regulated genes in the root and/or the shoots, (2) what the important pCRE combinations are and what types of pCREs are involved with the

combinations, and (3) if combinatorial rules important in root expression are also important for shoot expression, or *vice versa*.

To identify full pCRE combinations relevant to the up-regulation of genes under salt stress, we used the Classification by Association method (CBA; see **Methods**). Due to consideration of computational complexity, we restricted our analysis to combinatorial rules where the presence of two pCREs predicts up-regulation (pCRE A + pCRE B → up-regulation in organ of interest). Rule sets were generated for both the root salt up-regulated genes (2,838 "root rules", **Figure 5A**) and shoot salt up-regulated genes (363 "shoot rules", **Figure 5B**) using all 1,894 pCREs. We wanted to see if the rules tended to be composed of a general pCRE and an organ pCRE (general pCRE + organ pCRE → up-regulation) as opposed to two general pCREs or two organ pCREs (general pCRE + general pCRE or organ pCRE + organ pCRE), which might suggest that the subsets of pCREs depend on each other in the combinatorial rules. We found that there was a significant difference in the distribution of these three categories of combinatorial rules for the shoot rules (Chi-squared, p=6e-06). The shoot rules tended to have more general pCRE + general pCRE rules than expected (odds-ratio=1.5), and fewer organ pCRE + organ pCRE rules than expected (odds-ratio=0.52). This aligns with the notion that the general pCREs are more important for the regulation of shoot up-regulated genes. The root rules also had a significantly different distribution of rule types (Chi-squared, p=0.01), but the effect sizes were generally low (range 0.89-1.1) suggesting that there is only a small difference in the types of rules. Thus it does not appear that rules for root up-regulated genes are composed of general pCRE with a pCRE from one of the organ sets. Additionally, none of the root rules overlap with the shoot rules or *vice versa*, consistent with our expectation that rules for one organ are not relevant for predicting expression in another. Most importantly, models based on the combinatorial rule sets improved predictions for both root

(AUC=0.83, **Figure 5C**) and shoot (AUC=0.85, **Figure 5D**) up-regulated genes compared to the models based on presence/absence of single pCREs (AUC=0.76 and 0.80 for root and shoot, respectively). These modeling results confirm our hypothesis that pCRE combinatorial rules are important to the salt stress up-regulated genes in root and shoot. One unexpected result was that predictions using a model based on the pCREs found in at least one rule performed as well as models based on all motifs in the full pCRE set (**S4 Figure**) for shoot up-regulated genes. This suggests that the pCREs that work in combinations are the most influential in making predictions of gene shoot up-regulation. Models of root up-regulated genes based only on the pCREs found in root rules were not as good as using all the available pCREs, but they still performed well (**S4 Figure**). Another observation of the combinatorial rules is that shoot rules contained root pCREs and root rules contained shoot pCREs. Finding pCREs from one organs set in the rule set of another organ suggests that the combinations of pCREs, compared to the presence and absence of single pCREs, are more critical for specifying spatial response to stress. The greater importance of combinatorial rules aligns well with what is already known in mammals, where individual CREs are important for expression in multiple tissues, but CRE combinations are more relevant in controlling tissue-specific expression [17,18].

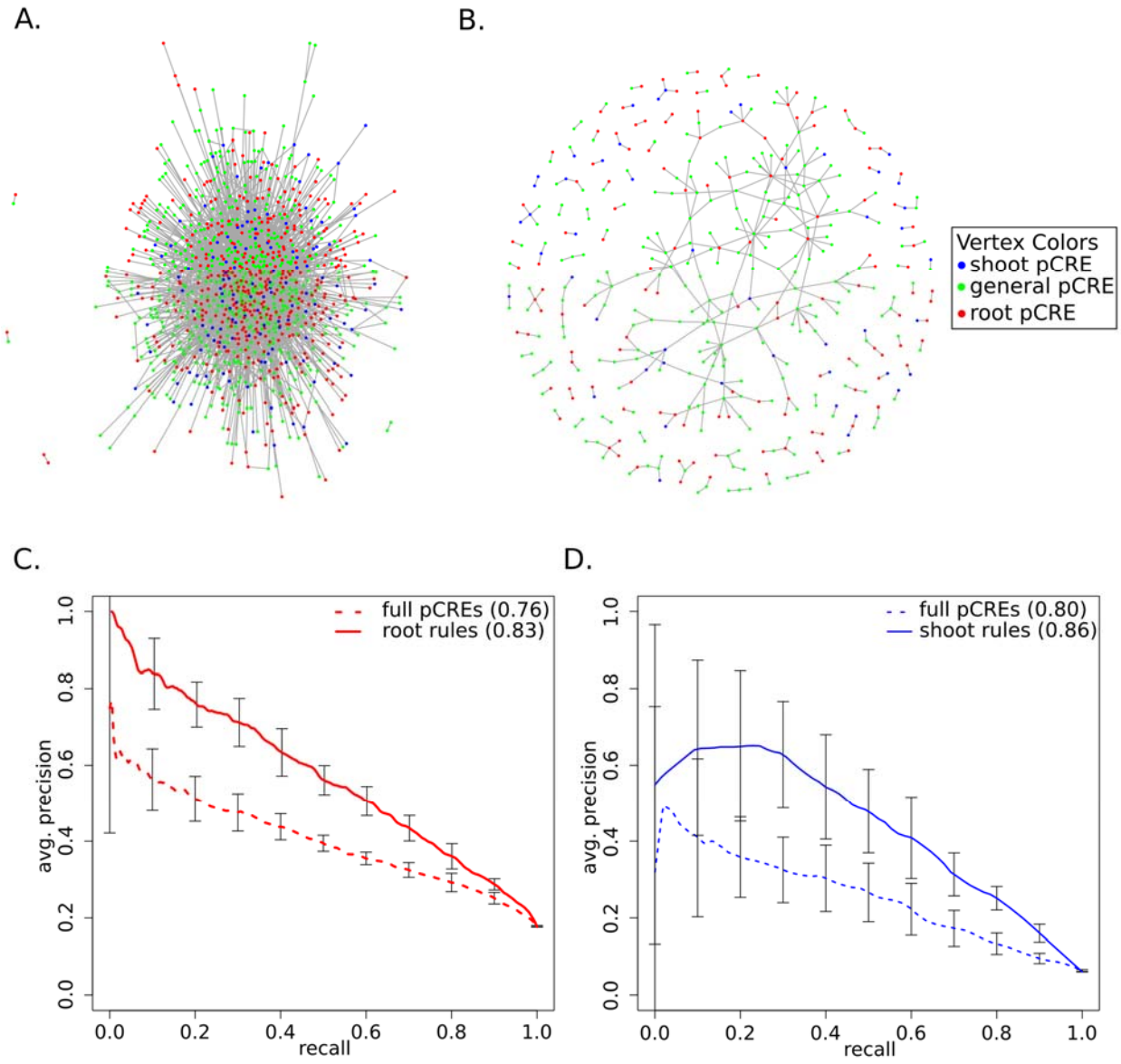

Figure 5. Summary of root and shoot combinatorial pCRE rules and model performance.

(A) Networks summarize the rules identified for salt stress up-regulated genes in the root**. (B)** Network for the rules identified for shoot up-regulated genes. Nodes represent a single pCRE color coded by the subset the pCRE belongs to (red=root pCREs, blue=shoot pCREs, green=general pCREs). Edges indicate that two pCREs are joined together in a rule. **(C)** Precision-recall curves comparing models based on combinatorial rules and the full pCRE set for root salt up-regulated

Figure 5. (Cont'd)

genes. **(D)** Precision-recall curves comparing model based on combinatorial rules and the full pCRE set for shoot salt up-regulated genes.

One may expect that sites of pCREs involved in the rule that were closer together might be better predictors of spatial stress response. Therefore, we examined the distance distributions of the instances of root rules (**S5A Figure**) and shoot rules (**S5B Figure**). Salt up-regulated genes with instances of the rules in their promoters do not tend to have rule pCREs that are closer together compared with non-responsive genes (**S5C,D Figure**). In addition, the model based only on the rules that have pCREs that significantly closer than expected (root up-regulated AUC-ROC=0.67; shoot up-regulated AUC-ROC=0.64) are not necessarily better than models based on the rules with pCRE pairs that are significantly further apart than expected (root up-regulated genes AUC-ROC=0.67; shoot up-regulated genes AUC-ROC=0.60). While we are only looking at the 1kb promoter, similar analysis in humans where the median promoter length was 1.4kb found that the significantly close CREs were predictive of tissue specific expression [18]. One possible interpretation for our result is that CRE combinations without a possible constraint on pCRE distance are still important for up-regulation of genes. Taken together, our findings suggest that the organ pCREs work best in combinations. Both root rules and shoot rules incorporate pCREs from the full set of organ pCREs, but there is no overlap in the two sets of rules. This suggests that the pCREs need to be considered in terms of the combinations of the pCREs. While some of the rules have motif pairs that are closer together than we would expect by chance, the majority of the rules do not have a significantly shorter distance between motifs. This may indicate that rules without a strong constraint on the distance between pCREs may still be important for spatial salt stress up-regulation.

Considering cell type expression performs as well as organ expression

Note: The analysis and the original draft of this section, was done by Sahra Uygun as part of the preparation of a manuscript for publication. It has been revised by Alexander Seddon in this thesis.

The models based on the full pCRE set were able to predict salt up-regulation in the root and shoot. In addition, the models were further improved when combinatorial rules of pCREs were examined. Given that each plant organ consists of multiple tissue/cell types, another possible way to improve the model is to focus on gene expression at a finer spatial resolution, namely, the differential expression of genes within individual cell-types. In human studies, cell-type specific CREs were identified using gene expression data [32] and these motifs were used to predict cell-type specific gene expression [33]. We hypothesized that we may find cell-type specific CREs in *A. thaliana* as well. In *A. thaliana*, root cell-type expression data are available for both control and salt stress treatment conditions [2]. We wanted to use this data to see whether the full pCRE set was sufficient to predict the root cell-type salt induced expression. Given that cell-type expression data clustered with the whole root salt expression data (**Figure 1A**, dotted rectangles II and III), our hypothesis was that full pCREs may be able to predict salt stress up-regulated genes expression to a cell-type level resolution as well as predictions for the whole root,. To test our hypothesis, we used full pCRE-based models to predict cell-type salt up-regulated genes. As we had expected, we found that the full pCRE set based models performed well in predicting salt up-regulated genes in each of the cell-types (AUC-ROC=0.70-0.78; **Fig 6A**). Because the model based on the union of the root and general pCRE set was found to predict whole root salt up-regulated genes as well as the full pCRE set, we expected this set to perform the best at predicting root cell-type salt up-regulated genes. As expected, we found this to be the case. However, we unexpectedly found that

the general pCRE based models also performed as well as the full pCRE based models. This suggests that the root pCREs may not be contributing unique information to the prediction of salt up-regulated genes in the root cell-types. This indicates that we may still be missing some cell-type specific CREs necessary to improve the model to predict the salt up-regulation in the cell-types.

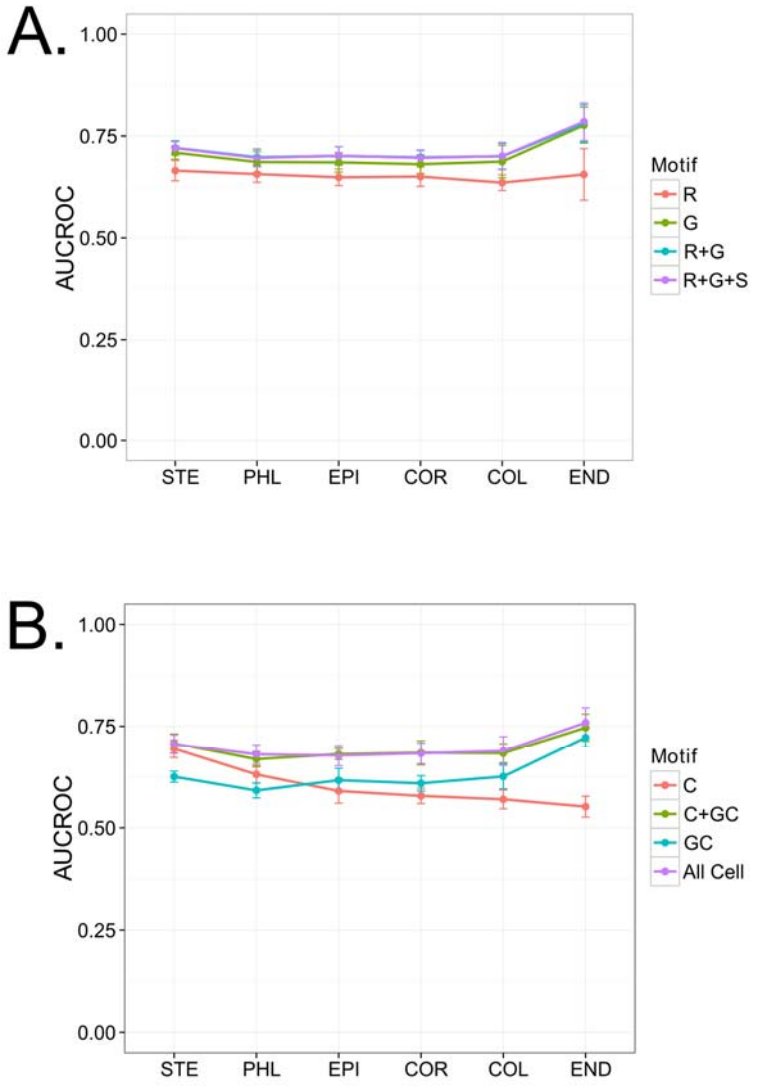

Figure 6. Summary of cell-type specific salt up-regulated gene models.

(A) AUC-ROCs for prediction of salt stress up-regulated genes in six cell-types: stele (STE), proto-phloem (PHL), epidermis-lateral (EPI), cortex (COR), columella (COL), and endodermis-quiescent center (END). Models were based on following sets of pCREs: root and general pCREs ("R+G"), general pCREs ("G"), root pCREs ("R"), and the full pCRE set ("R+G+S"). **(B)** AUC-ROC predicting genes up-regulated by salt in six cell-types using models based on pCREs specific

Figure 6. (Cont'd)

to that cell type ("C"), general to all cell-types ("GC"), the union of cell-type specific and cell-type general ("C+GC") and the full set of cell-type salt pCREs ("All Cell").

So far, we have shown that models based on the full pCREs can predict genes up-regulated by salt in different types of root cells. However, we found that the general pCREs were sufficient for making predictions, which raises the question of whether there are cell-type specific CREs that could improve our models of salt up-regulation. To this end, we used the CRE identification pipeline (see **Methods**) to identify pCREs that are cell-type specific. We incorporated root cell-type differential expression, for salt induced expression in stele, proto-phloem, columnar, cortex, endodermis-quiescent center, and epidermis-lateral root cell types [2], along with root abiotic stress data [4] to identify co-expression clusters that are over-represented with salt stress up-regulated genes in different root cell types. According to the enrichment of pCRE sites in the promoters of the genes in these clusters, 583 pCREs were classified as root cell-type general pCREs and 734-2828 pCREs were considered specific to a particular cell-type or found to be overrepresented in multiple cell-type induced genes (**S6 Figure**). Because we hypothesized that we would find new cell-type pCREs, we expected to have motifs that were distinct from full pCREs. Additionally, we also wanted to know to what extent each pCRE subset has similar pCREs. In order to address these questions, we calculated the average PCC among the pCREs within each pCRE subset. We found that within a pCRE set, we did not necessarily have the most similar pCREs. For example cell-type general pCREs had the highest average correlation (r=0.78), but epidermis-lateral specific and stele specific pCREs had average PCCs ≤ 0.4 within the same subset of pCREs. We also found that the full pCRE set and cell-type pCRE set did not have high similarity across subsets (r=0.37-0.47) supporting the notion that the pCREs we identified from genes up-regulated in different cell-type were distinct from the full set of pCREs identified using data from whole root and shoot (**S7 Figure**).

We have shown that cell-type pCREs were distinct from the full pCRE set, suggesting that novel motifs may be important in driving salt induced expression among root cell-types. We predicted cell-type salt up-regulated genes using models based on the cell-type pCREs. Contrary to the performance of general pCREs, cell-type general pCREs did not outperform the other cell-type motifs (AUC-ROC=0.60-0.72 vs AUC-ROC =0.68-0.76; **Figure 6B**). Overall, the performance of cell-type pCREs was similar to the full set of pCREs in predicting up-regulation in the various cell types. Even though we have seen similar performances of salt up-regulated gene prediction, potentially the full pCRE set and cell-type pCREs could predict different sets of genes to be responsive. To test this, we compared the sets of genes predicted by models based on the full pCRE set and cell-type pCREs, focusing on the stele cell up-regulated genes as an example. We found that only ~50% of the true positives predictions were the same from the two models (**S8 Figure**). Ten-percent of the salt responsive genes were correctly predicted by only the organ pCREs and 14% were predicted correctly by only the cell-type pCREs. This result implies that we were able to predict an additional set of the salt responsive genes using cell-type pCREs. Overall, we were able to improve the prediction of salt responsive gene expression to a finer resolution in *A. thaliana.*

CONCLUSION

In this study we identified a set of pCREs that are associated with the salt stress up-regulation of genes in the root and shoot of *A. thaliana*. Many of the pCREs in the full pCRE set are similar to the binding motifs for TFs in various families in *A. thaliana*. Models of salt stress up-regulation based on the full pCRE set are significantly better at modeling gene expression than *in vitro* derived TFBMs [16]. This improvement in modeling suggests that the full pCRE set may contain a more complete set of CREs that are essential for salt stress up-regulation of genes in the root and shoot. We found that genes up-regulated by salt in the root may need both a general pCRE set and a root pCRE set, while the shoot salt up-regulated genes rely primarily on a general pCRE set. Finding that root up-regulated genes are better modeled with an additional root pCRE set may be explained by the fact that root specifically up-regulated genes were enriched with TF genes. We also found that models of root up-regulation based only on the root pCREs were significantly overrepresented with root specifically up-regulated genes. Thus, a possible mechanisms for root up-regulation is that the root pCREs are found on the promoter of root specifically up-regulated genes, which are bound by root specific TF.

Furthermore, the full pCRE set works best when combinations of the pCREs are considered. One interpretation of these results is that the organ expression patterns seen under salt stress are the result of unique combinations of CREs. Thus, it may not be appropriate to label a CRE as a "salt stress" CRE or a "root" CRE, as the CREs themselves may play multiple roles in gene regulation. This finding is congruent with studies of tissue specific transcription in rice, where promoters of genes specifically transcribed in root tissues were enriched for two dehydration stress associated CREs – the binding sites of MYB and MYC TFs [19]. This indicates that the same CRE may play roles in tissue specific expression as well as stress expression. Combinatorial rules have

also been found to be important for tissue specific expression in humans [17,18], where it was found that pairs of CREs on a promoter are more critical to tissue specific expression, while individual TFs are often associated with expression in multiple tissues. It should be noted that CRE combinations in Yu et al. [18] were discovered by looking for pairs of CREs that are significantly closer together on a promoter than expected. Our method of discovering combinations of CREs did not consider the distance between CRE pairs, and when we generated models based on CRE pairs that were closer together than randomly expected, the models did not perform as well as using all the combinatorial rules, including those that were significantly further apart. This finding suggests that combinations of CREs important for the up-regulation of do not necessarily need to be constrained by the distance between pairs of CREs in a rule.

In addition to predicting salt up-regulation at the organ level, the full set of pCREs can be used to predict the differential expression of genes in root cell-types under salt stress, and the performance is equal to predicting expression using a distinct set of cell-type pCREs. This may indicate that cell-type specific expression is based on the same code as expression seen in the whole roots. This conclusion is supported by our expression clustering analysis, which found that the root-cell type expression under salt stress clustered with the whole root salt stress data, and by the finding of an organ-tissue-cell type expression clustering hierarchy in rice [19]. Along this line of thinking, regulation by CREs and TFs may be responsible for shaping the organ level of the expression hierarchy, while the cell-type level of the hierarchy may be further shaped by mechanisms other than *cis*-regulation. Thus, future research may need to focus on studying additional layers of regulation to understand cell-type gene regulation in the context of stress. It is also possible that we were not able to identify a substantially expanded set of cell-type pCREs because our expression matrix used for co-expression clustering was predominantly from whole

root and shoot stress expression data. The similarity in the expression matrix may have resulted in clusters that are not substantially different from clusters used to discover the full pCRE set. Without substantially different clusters, our method would not be able to identify new pCREs. Thus, our approach may be limited by the data that is available for identifying co-expressed genes.

Our observations with the cell-type pCREs point out one of the biggest limitations of our CRE identification method. We are limited by the expression data that we have to identify pCREs. The expression data that we use will influence the co-expression clusters identified, which ultimately changes the promoter sequences on which motif finding is performed. Thus, CRE identification may be enhanced by cell-type specific data under additional stress conditions other than salt stress. However, on the organ level, the primary strength of our method is that it uses expression data to infer pCREs, even though it is still limited by the type of data included. Using co-expression means we are not limited to looking for the CREs of every TF in *A. thaliana*. Thus, our pCREs can be seen as a complement to valuable resources that use *in vitro* derived CREs, such as the TFBM data used in this study [16].

To summarize, the results in this study indicated that the organ and cell-type specific regulation of genes during stress involves a nuanced CRC. This CRC incorporates pCREs that may be bound by a wide range of TF. The root and shoot up-regulated genes differ in which pCREs are necessary for regulation under salt stress, with the root requiring an additional set of root pCREs that might be regulating the root specific up-regulated genes. We found that the CRC was further improved by finding that combinatorial rules of pCREs may be more influential to gene up-regulation than looking at individual pCREs, and that these combinations of pCREs do not necessarily need to be constrained by distance. Finally, we found evidence that CRC regulating genes across the whole root might be the same CRC regulating cell-type specific expression.

However, we could not rule out the possibility that this was due to a limitation in the expression data used to identify pCREs. Our results show that co-expression based CRE identification methods are a promising method for globally assessing spatial gene regulation in the context of stress. This approach may have possible applications in engineering plants that can respond to stresses. Use of native and tissue specific and inducible promoters to engineer plants is promising, but it is limited by the promoters that are already available in nature [34]. The methods we used here may help to identify combinations of CREs that can be used to synthesize promoters to drive tissue specific expression in the context of stress.

METHODS

Expression data processing

A. thaliana abiotic stress expression data for the root and shoot [35] and biotic stress data from the shoot was downloaded from the Weigel World website (http://www.weigelworld.org/resources/microarray/AtGenExpress/). The data came preprocessed and normalized. Log2 fold changes and associated *p*-values were calculated for each stress condition and its corresponding control at each time point using limma [36] in the R environment [37], and the *p*-values were adjusted using the Benjamini-Hochberg method [38] to control for the False Discovery Rate. Root cell type expression data [2] under salt stress was download as CEL files from GEO (GSE7641). The CEL files were pre-processed with Robust Multi-array Average (RMA) [39] and quantile normalized using the affy package [40]. Log2 fold changes and *p*-values between salt stress and controls were calculated as described above for the whole root and shoot data. Genes were considered up-regulated if their log2 fold-change values ≥ 1 and their adjusted *p*-values≤ 0.05). As mentioned in **Results and Discussion**, the root up-regulated genes refer to genes up-regulated by salt at 3 hours in the root, and the same criteria is true for the shoot up-regulated genes, and any of the genes up-regulated in the root cell-types. Genes were considered non-responsive if they were not differentially expressed (either up or down-regulated) under any stress at any time point within the root or the shoot. Each organ had its own set of non-responsive genes ("root non-responsive" and "shoot non-responsive"). This stringent definition of non-responsive genes was chosen to reduce the noise from genes that are differentially expressed in related stress conditions.

Expression correlation calculation, clustering and Gene Ontology analysis

Log fold-change data was compiled from the stress data sets described in the ***Expression data processing*** section. To assess the relationship of differential expression in the root, shoot, and root cell-types during different stress conditions, the Pearson's Correlation Coefficients (PCCs) of log2 fold change were calculated pairwise for all differential data sets. A heatmap of the PCC values was generated using the gplots package [41]. To identify the functional categories of genes up-regulated in the root and shoot at 3 hours of salt stress, we looked at the enrichment of Plant GO slim categories (http://www.geneontology.org/ontology/subsets/goslim_plant.obo) containing over/underrepresented numbers of genes in each of these three gene sets identified based on Fisher's exact tests, as implemented in SciPy [42]. The p-values were adjusted using the Benjamini-Hochberg method.

Collection of Known Transcription Factor Binding Site Motifs and Zou pCREs

Position Frequency Matrices (PFMs) were obtained from the Cis-BP database website [16]. These PFMs are based on either protein binding microarray data or publicly available TRANSFAC motifs [16]. A bulk download of all the binding motifs for *A. thaliana* transcription factors was performed. The PFMs were converted to position weight matrices (PWMs) adjusted for the background AT-CG content of *A. thaliana* (AT=0.33 and GC=0.17) using the TAMO package in Python [43]. This resulted in a final set of 355 TFBMs. The Zou pCREs discovered in Zou et al. [12] were collected from the paper in a TAMO formatted file. The 1kb promoter sequence of all genes in *A. thaliana* was downloaded from The Arabidopsis Information Resource (TAIR; ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/TAIR10_blastsets/upstream_sequences/TAIR10_upstream_1000_20101104). The PWMs for the TFBM and Zou pCREs were mapped to these promoter sequences using a Python script with the motility [44] package. We identified

TFBMs that were overrepresented on the promoter of the root up-regulated and shoot up-regulated genes by performing a Fisher's Exact Test against the root non-responsive and shoot non-responsive genes, respectively.

Prediction of gene up-regulation using Support Vector Machine (SVM)

Our goal was to model salt up-regulation of genes in the root and shoot as a classification problem involving two classes: salt up-regulated genes (in either root of shoot) and genes that are not responsive as defined in ***Expression data processing***. The Support Vector Machine (SVM, [28]) method was used to perform this modeling. Every SVM model in this paper had two components: 1) a set of genes, each of which is classified as up-regulated or non-responsive ("expression class") and 2) a set of TFBMs, pCREs or pCRE combinatorial rules and their presence/absence on the promoter of each gene ("promoter features"). In this setup, SVM generated a model that best separates the genes from the two expression classes using the presence or absence of the promoter features. See **S5 table** for a complete listing of the up-regulated and non-responsive gene sets as well as the promoter features for all models generated in this study. We used the LIBSVM implementation of SVM [27] through a wrapper written for Weka [45]. Grid-searches were used to find the best combination of the following three parameters: (1) the ratio of non-responsive to up-regulated genes, (2) the parameter of the soft margin, and (3) the gamma parameter of the Radial Basis Function (RBF) kernel. The latter two parameters are part of the SVM method itself. The ratio of negative to positive examples was achieved using the Weka class "weka.filters.supervised.instance.SpreadSubsample", which subsamples the non-responsive genes to achieve the desired ratio of up-regulated to non-responsive genes. A grid-search runs a model for each possible combination of the three parameters (see **S6 Table** for model parameters used in the grid search). We used 10-fold cross validation as implemented in Weka , and the

average AUC-ROC from all 10 cross validation runs was calculated using the ROCR package [46]. The parameter combination with the maximum average AUC-ROC were taken as the best parameters for each model, and this maximum AUC-ROC is what we report for each model. Precision-recall curves were plotted using the output from the model with the maximum AUC-ROCs.

CRE Identification

To identify pCREs associated with salt up-regulated genes in the root and shoot, we used a CRE identification pipeline from an earlier publication with modifications [12]. The stress expression data in the form of a fold-change expression matrix (see ***Expression data processing***) was used to identify co-expression clusters using repeated rounds of k-means clustering – implemented in R – such that all clusters were 60 genes or less, while clusters smaller than 10 genes were excluded. Clusters enriched in salt up-regulated genes in any time point in either roots or shoots were analyzed further. Six motif finding programs were used to uncover 6-18bp pCREs in the putative promoter regions (1kb upstream to transcriptional starts) of genes in each cluster: AlignACE [47], MDScan [48], MEME [49], Motif Sampler [50], Weeder [51], and YMF [52]. In the motif finding step, ~300,000 sequence motifs were identified, many of which were redundant or potentially irrelevant to salt up-regulation.

Two rounds of pCRE merging-enrichment testing were performed. In the first round, the ~300,000 motifs were merged if their consensus sequences shared the same IUPAC codes and/or if they were highly similar to each other based on clustering together in a Kullback-Leibler (KL) distance based cluster as described in [12]. In the enrichment step, these merged pCREs were mapped to the 1kb promoter regions of genes in *A. thaliana* using motility [44], and we kept mappings with a p < 1e-06. The pCREs were further analyzed if their mapped sites were

significantly overrepresented (Fisher's Exact Test, Benjamini-Hochberg adjusted $p \leq 0.05$) in promoters of root and/or shoot salt up-regulated genes. In the second round, we merged enriched motifs based on Pearson's Correlation Coefficient-distance (PCC distance=1-PCC) of the motif PWMs. Using the PCC distance matrix, motifs were clustered hierarchically and distinct motif clusters were demarcated with a PCC distance threshold of 0.10, which was previously found to be the first percentile of PCC distances for non-redundant motifs in the JASPER CORE dataset [12]. Within each cluster, a single pCRE was chosen based on having the most significant degree of enrichment for genes up-regulated under salt stress in roots and/or shoots. To identify organ or cell-type specific motifs, a final round of enrichment analyses testing which motifs were significantly over-represented ($p<0.05$) only in the root salt up-regulated genes ("root pCREs"), only in shoot salt up-regulated genes ("shoot pCREs"), and among genes up-regulated in both organs (general pCREs). In the end, 1,984 shoot, root, and general pCREs were identified.

The same logic is applied to identify cell-type-specific pCREs, except that we clustered on an expression matrix incorporating root abiotic stress data, and replaces the whole root data at for salt at 3 hours with the cell-type data [2]. No shoot expression data was used for the co-expression clustering.

PCC comparison of pCREs and TFBMs.

To assess the similarity between the full pCREs identified in this paper and the TFBMs, the PCC between all pairwise combinations of pCREs and TFBMs was calculated using the same method for calculating PCC distance in the ***CRE identification*** section, except the PCC was not subtracted from 1. PCC was calculated for all possible pairs between the full pCREs and the TFBMs, as well as all pairs within the TFBM set. To assess the significance of the correlation between a full pCRE-TFBM pair, we used the distribution of each TFBM to TFBMs within its

own TF family as a background model. Using the within family distribution was chosen because it allowed us to test whether or not the pCRE was more similar to the TFBM than TFBMs in the same family as the TFBM. For each TFBM family, maximum likelihood has used to fit normal and beta distribution functions to the distributions of PCCs comparing TFBMs within the same family. Fitting was performed using the MASS package [53] in R. The distribution with the maximum log-likelihood was chosen as the representative distribution for that family. Every PCC between a pCRE and a TFBM was compared to the cumulative density function of the fitted within family distribution to get a *p*-value for the significance of that PCC. All *p*-values from the pairwise comparisons were adjusted for multiple testing within the same family using the Benjamini-Hochberg method.

To assess the likely families of TFs that might bind the pCREs, PCC distributions were estimated for outside of family comparisons using the same maximum-likelihood method described above. Thus, each TFBM had a distribution of PCC values for comparing the family members to TFBMs of other families. We compared the PCC for each pCRE-TFBM pair using the between family distributions to generate a *p*-value, which were adjusted using the Benjamini-Hochberg method. We set an adjusted *p*-value of 0.05 as the threshold to say that the pCRE may be bound by the same family as the TFBM.

To assess if the pCREs are more similar to a TFBM like sequence than to random genomic sequences, 1894 random PWMs with the same length distribution of sequence length as the full pCREs were generated. To make a random PWM of length k, 15 random k-mers using the background distribution of AT-GC in *A. thaliana*, and these were consolidated into a PWM using the TAMO function MotifTools.Motif_from_counts. The random PWMs were compared to the TFBMs using PCC. Each TFBM then had a distribution of PCC to randomized PWMs. For all

pCRE-TFBM PCC values, we asked what percentile this PCC lies on in the random PWM-TFBM distribution.

Binary prediction of root and shoot up-regulated genes

While the AUC-ROC is a good measure of the overall performance of an SVM model, it does not indicate how well individual genes are predicted. Thus, it is possible that two models can have similar levels of performance as measured by AUC-ROC resulting from the correct predictions of different sets of genes. To assess which genes were predicted by a particular models based on different pCRE sets, and to see if different models correctly predict different sets of genes, the Weka program CrossValidationAddPredictions was used to identify whether a gene was correctly predicted as up-regulated or non-responsive during salt stress. This program performs SVM as described in the section ***Prediction of gene up-regulation using SVM***, but it keeps track of the prediction of each gene. We used the best parameter combination identified from the original SVM grid-search as the basis for the binary prediction run. Because 10-fold cross validation uses a different set of randomly selected training examples to generate an SVM model, the predictions of genes from one run of 10-fold cross validation may be different from the prediction from another run. Thus, 10-fold cross validation was performed 10 times, resulting in 10 predictions for each gene. The SVM score (probability estimates of the classification) from each round of 10-fold cross validation had an and a SVM score threshold chosen at the maximum F-measure (harmonic mean of precision and recall, calculated using ROCR) to create binary predictions of for each gene.

We assessed the overlap of correctly predicted up-regulated genes (True Positives, "TP") based on models using different pCRE sets by looking at the percentage of the up-regulated genes correctly predicted by two different models. Because each model was run 10 times, we compared all 10 runs from one model pairwise with all 10 runs from the other. Thus, two models had 100

overlap percentage values, and the median of these percentages was reported in **Figure 4C-D**. To test if two models were significantly different in the TP genes they predicted, we compared the overlap percentages between the two models to the overlap percentages within multiple runs using the same pCRE based model using a Mann-Whitney test. TP gene sets were considered significantly different if they had a lower median percentage overlap between the multiple runs then seen with one of the pCRE based models as determined by the Mann-Whitney test

Combinatorial motif rule discovery

To test if the combinations of specific pCREs were predictive of up-regulation in the root or shoot, the Classification by Association (CBA) [54] method was used to identify combinatorial rules of the form pCRE A + pCRE B → up-regulation were selected from the CBA output. This method is useful for identifying rules where some combinations of features are associated with a particular class. The features in our case were the presence or absence of pCREs on a genes promoter and the class is root or shoot up-regulation (as was the case for SVM). The root or shoot up-regulated and non-responsive genes were broken up into different subsamples. Each of these subsamples was run through CBA using multiple values for minimum confidence (percentage of genes where "pCRE A + pCRE B → up-regulation" out of all the instances of "pCRE A + pCRE B") and support (percentage of genes that with the rule "pCRE A + pCRE B → up-regulation"). Rules for shoot up-regulated genes were discovered using a minimum support 0.5% and a minimum confidence of 60.0%, with a non-responsive to up-regulated ratio of 2:1. We went through several rounds of CBA to discover root rules using different values of support, confidence and non-responsive to up-regulated ratios (**S7 Table**). We ended up using a minimum support of 0.1%, a minimum confidence of 60%, and subsamples with 976 non-responsive genes to 488 responsive genes, which are the same numbers of genes used to generate the shoot rules. These

parameters were chosen because the rules generated gave an appreciable gain in the AUC-ROC when performing SVM. Due to the limitation of using a GUI version of CBA, we were not able to do an extensive exploration of the best CBA parameter values. Thus, it is possible that there is a more optimal parameter set that will yield a greater performance gain.

To see if the rules had a bias in which pCRE subsets were involved in the rules, we categorized each rule as general pCRE + general pCRE, organ pCRE + general pCRE and organ pCRE + organ pCRE. We performed a Chi-square test for each rule set comparing the observed numbers of each rule category to what would be expected if pCREs were randomly paired together as a rule.

The distance between pairs of pCREs in a rule were calculated for all instances of the rules in the putative promoters. The minimal distance between the closest ends of two pCREs were determined. To determine if the minimal distances were significantly different than randomly expected, background distributions of pCREs was generated by modeling the frequency of distances between two random pCREs of the same lengths as the pCREs in the rule pair based on an earlier approach [18]. The only difference in our method was that we compared our observed distance distributions to the background distribution using a Mann-Whitney test instead of a Kolmogorov-Smirnov test, as the Mann-Whitney test can more directly test whether one distribution has higher or lower distances than the other distribution.

BIBLIOGRAPHY

BIBLIOGRAPHY

1. Kreps JA, Wu Y, Chang H-S, Zhu T, Wang X, Harper JF. Transcriptome Changes for Arabidopsis in Response to Salt, Osmotic, and Cold Stress. Plant Physiol. 2002;130: 2129–2141. doi:10.1104/pp.008532

2. Dinneny JR, Long TA, Wang JY, Jung JW, Mace D, Pointer S, et al. Cell Identity Mediates the Response of Arabidopsis Roots to Abiotic Stress. Science. 2008;320: 942–945. doi:10.1126/science.1153795

3. Geng Y, Wu R, Wee CW, Xie F, Wei X, Chan PMY, et al. A Spatio-Temporal Understanding of Growth Regulation during the Salt Stress Response in Arabidopsis. Plant Cell. 2013;25: 2132–2154. doi:10.1105/tpc.113.112896

4. Kilian J, Whitehead D, Horak J, Wanke D, Weinl S, Batistic O, et al. The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses: AtGenExpress global abiotic stress data set. Plant J. 2007;50: 347–363. doi:10.1111/j.1365-313X.2007.03052.x

5. Hahn A, Kilian J, Mohrholz A, Ladwig F, Peschke F, Dautel R, et al. Plant Core Environmental Stress Response Genes Are Systemically Coordinated during Abiotic Stresses. Int J Mol Sci. 2013;14: 7617–7641. doi:10.3390/ijms14047617

6. Munns R, Tester M. Mechanisms of Salinity Tolerance. Annu Rev Plant Biol. 2008;59: 651–681. doi:10.1146/annurev.arplant.59.032607.092911

7. Munns R. Comparative physiology of salt and water stress. Plant Cell Environ. 2002;25: 239–250. doi:10.1046/j.0016-8025.2001.00808.x

8. Qin F, Shinozaki K, Yamaguchi-Shinozaki K. Achievements and Challenges in Understanding Plant Abiotic Stress Responses and Tolerance. Plant Cell Physiol. 2011;52: 1569–1582. doi:10.1093/pcp/pcr106

9. Stockinger EJ, Gilmour SJ, Thomashow MF. Arabidopsis thaliana CBF1 encodes an AP2 domain-containing transcriptional activator that binds to the C-repeat/DRE, a cis-acting DNA regulatory element that stimulates transcription in response to low temperature and water deficit. Proc Natl Acad Sci U S A. 1997;94: 1035–1040.

10. Beer MA, Tavazoie S. Predicting Gene Expression from Sequence. Cell. 2004;117: 185–198. doi:10.1016/S0092-8674(04)00304-6

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

12. Zou C, Sun K, Mackaluso JD, Seddon AE, Jin R, Thomashow MF, et al. Cis-regulatory code of stress-responsive transcription in Arabidopsis thaliana. Proc Natl Acad Sci. 2011;108: 14992–14997. doi:10.1073/pnas.1103202108

13. Priest HD, Filichkin SA, Mockler TC. Cis-regulatory elements in plant cell signaling. Curr Opin Plant Biol. 2009;12: 643–649. doi:10.1016/j.pbi.2009.07.016

14. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, et al. Transcriptional regulatory code of a eukaryotic genome. Nature. 2004;431: 99–104. doi:10.1038/nature02800

15. Franco-Zorrilla JM, López-Vidriero I, Carrasco JL, Godoy M, Vera P, Solano R. DNA-binding specificities of plant transcription factors and their potential to define target genes. Proc Natl Acad Sci. 2014;111: 2367–2372. doi:10.1073/pnas.1316278111

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

17. Hu Z, Gallo SM. Identification of interacting transcription factors regulating tissue gene expression in human. BMC Genomics. 2010;11: 49. doi:10.1186/1471-2164-11-49

18. Yu X, Lin J, Zack DJ, Qian J. Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. Nucleic Acids Res. 2006;34: 4925–4936. doi:10.1093/nar/gkl595

19. Jiao Y, Lori Tausta S, Gandotra N, Sun N, Liu T, Clay NK, et al. A transcriptome atlas of rice cell types uncovers cellular, functional and developmental hierarchies. Nat Genet. 2009;41: 258–263. doi:10.1038/ng.282

20. Zhu J-K. Salt and Drought Stress Signal Transduction in Plants. Annu Rev Plant Biol. 2002;53: 247–273. doi:10.1146/annurev.arplant.53.091401.143329

21. Golldack D, Lüking I, Yang O. Plant tolerance to drought and salinity: stress regulating transcription factors and their functional significance in the cellular transcriptional network. Plant Cell Rep. 2011;30: 1383–1391. doi:10.1007/s00299-011-1068-0

22. Mizoi J, Shinozaki K, Yamaguchi-Shinozaki K. AP2/ERF family transcription factors in plant abiotic stress responses. Biochim Biophys Acta BBA - Gene Regul Mech. 2012;1819: 86–96. doi:10.1016/j.bbagrm.2011.08.004

23. Nakashima K, Takasaki H, Mizoi J, Shinozaki K, Yamaguchi-Shinozaki K. NAC transcription factors in plant abiotic stress responses. Biochim Biophys Acta BBA - Gene Regul Mech. 2012;1819: 97–103. doi:10.1016/j.bbagrm.2011.10.005

24. Matiolli CC, Tomaz JP, Duarte GT, Prado FM, Bem LEVD, Silveira AB, et al. The Arabidopsis bZIP Gene AtbZIP63 Is a Sensitive Integrator of Transient Abscisic Acid and Glucose Signals. Plant Physiol. 2011;157: 692–705. doi:10.1104/pp.111.181743

25. Chaves MM, Flexas J, Pinheiro C. Photosynthesis under drought and salt stress: regulation mechanisms from whole plant to cell. Ann Bot. 2009;103: 551–560. doi:10.1093/aob/mcn125

26. Ji H, Pardo JM, Batelli G, Van Oosten MJ, Bressan RA, Li X. The Salt Overly Sensitive (SOS) Pathway: Established and Emerging Roles. Mol Plant. 2013;6: 275–286. doi:10.1093/mp/sst017

27. Chang C-C, Lin C-J. LIBSVM: A Library for Support Vector Machines. ACM Trans Intell Syst Technol. 2011;2: 27:1–27:27. doi:10.1145/1961189.1961199

28. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20: 273–297. doi:10.1007/BF00994018

29. Kang J, Choi H, Im M, Kim SY. Arabidopsis Basic Leucine Zipper Proteins That Mediate Stress-Responsive Abscisic Acid Signaling. Plant Cell Online. 2002;14: 343–357. doi:10.1105/tpc.010362

30. Achard P, Cheng H, Grauwe LD, Decat J, Schoutteten H, Moritz T, et al. Integration of Plant Responses to Environmentally Activated Phytohormonal Signals. Science. 2006;311: 91–94. doi:10.1126/science.1118642

31. Lei G, Shen M, Li Z-G, Zhang B, Duan K-X, Wang N, et al. EIN2 regulates salt stress response and interacts with a MA3 domain-containing protein ECIP1 in Arabidopsis. Plant Cell Environ. 2011;34: 1678–1692. doi:10.1111/j.1365-3040.2011.02363.x

32. Chen C, Zhang S, Zhang X-S. Discovery of cell-type specific regulatory elements in the human genome using differential chromatin modification analysis. Nucleic Acids Res. 2013;41: 9230–9242. doi:10.1093/nar/gkt712

33. Natarajan A, Yardımcı GG, Sheffield NC, Crawford GE, Ohler U. Predicting cell-type–specific gene expression from regions of open chromatin. Genome Res. 2012;22: 1711–1722. doi:10.1101/gr.135129.111

34. Potenza C, Aleman L, Sengupta-Gopalan C. Targeting transgene expression in research, agricultural, and environmental applications: Promoters used in plant transformation. Vitro Cell Dev Biol - Plant. 2004;40: 1–22. doi:10.1079/IVP2003477

35. Kilian J, Whitehead D, Horak J, Wanke D, Weinl S, Batistic O, et al. The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses: AtGenExpress global abiotic stress data set. Plant J. 2007;50: 347–363. doi:10.1111/j.1365-313X.2007.03052.x

36. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; doi:10.1093/nar/gkv007

37. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. R Foundation for Statistical Computing; 2012. Available: http://www.R-project.org

38. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Stat. 2001;29: 1165–1188. doi:10.1214/aos/1013699998

39. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4: 249–264. doi:10.1093/biostatistics/4.2.249

40. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20: 307–315. doi:10.1093/bioinformatics/btg405

41. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T, et al. gplots: Various R Programming Tools for Plotting Data [Internet]. 2015. Available: http://cran.r-project.org/web/packages/gplots/index.html

42. Jones E, Oliphant T, Peterson P, et al. SciPy: Open Source Scientific Tools for Python. 2001.

43. Gordon DB, Nekludova L, McCallum S, Fraenkel E. TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs. Bioinformatics. 2005;21: 3164–3165. doi:10.1093/bioinformatics/bti481

44. Brown T. motility: A C++/Python toolkit for sequence motif searching [Internet]. [cited 6 Apr 2015]. Available: http://cartwheel.caltech.edu/motility/intro.html

45. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA Data Mining Software: An Update. SIGKDD Explor Newsl. 2009;11: 10–18. doi:10.1145/1656274.1656278

46. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: Visualizing the Performance of Scoring Classifiers [Internet]. 2015. Available: http://cran.r-project.org/web/packages/ROCR/index.html

47. Roth FP, Hughes JD, Estep PW, Church GM. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotechnol. 1998;16: 939–945. doi:10.1038/nbt1098-939

48. Liu XS, Brutlag DL, Liu JS. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nat Biotechnol. 2002;20: 835–839. doi:10.1038/nbt717

49. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol ISMB Int Conf Intell Syst Mol Biol. 1994;2: 28–36.

50. Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouzé P, et al. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinforma Oxf Engl. 2001;17: 1113–1122.

51. Pavesi G, Mereghetti P, Zambelli F, Stefani M, Mauri G, Pesole G. MoD Tools: regulatory motif discovery in nucleotide sequences from co-regulated or homologous genes. Nucleic Acids Res. 2006;34: W566–W570. doi:10.1093/nar/gkl285

52. Sinha S, Tompa M. A statistical method for finding transcription factor binding sites. Proc Int Conf Intell Syst Mol Biol ISMB Int Conf Intell Syst Mol Biol. 2000;8: 344–354.

53. Ripley B, Venables B, Bates DM, 1998) KH (partial port ca, 1998) AG (partial port ca, Firth D. MASS: Support Functions and Datasets for Venables and Ripley's MASS [Internet]. 2015. Available: http://cran.r-project.org/web/packages/MASS/index.html

54. Ma BLWHY. Integrating classification and association rule mining. Proceedings of the 4th. 1998. Available: http://www.aaai.org/Papers/KDD/1998/KDD98-012.pdf