Cis-Regulatory Code for Predicting Plant Cell-Type Transcriptional Response to High Salinity1[OPEN]

Sahra Uygun,a,2,3 Christina B. Azodi,b,2 and Shin-Han Shiu,a,b,c,4,5

aGenetics Program, Michigan State University, East Lansing, Michigan 48824
bDepartment of Plant Biology, Michigan State University, East Lansing, Michigan 48824
cDepartment of Computational, Mathematics, Science, and Engineering, Michigan State University, East Lansing, Michigan 48824

ORCID IDs: 0000-0003-0863-0384 (S.U.); 0000-0002-6097-606X (C.B.A.); 0000-0001-6470-235X (S.-H.S.).

Multicellular organisms have diverse cell types with distinct roles in development and responses to the environment. At the transcriptional level, the differences in the environmental response between cell types are due to differences in regulatory programs. In plants, although cell-type environmental responses have been examined, it is unclear how these responses are regulated. Here, we identify a set of putative cis-regulatory elements (pCREs) enriched in the promoters of genes responsive to high-salinity stress in six Arabidopsis (Arabidopsis thaliana) root cell types. We then use these pCREs to establish cis-regulatory codes (i.e. models predicting whether a gene is responsive to high salinity for each cell type with machine learning). These pCRE-based models outperform models using in vitro binding data of Arabidopsis transcription factors. Surprisingly, organ codes (i.e. models predicting whether a gene is responsive to high salinity for each cell type with machine learning). These pCRE-based models outperform models using in vitro binding data of 758 Arabidopsis transcription factors. Surprisingly, organ codes also suggest broad applicability of the approach to any species, particularly those with little or no trans-regulatory data.

The identification of different types of cells and the characteristics that make them unique in multicellular organisms has fascinated and challenged biologists since Anton van Leeuwenhoek’s invention of the microscope in the late 17th century (Trapnell, 2015). These distinct cell types carry out, to various degrees, specialized functions that contribute greatly to organismal complexity. One of the crucial components that allows for these specialized functions is differences in transcription regulatory mechanisms, which allow for cell-type-specific gene expression during development as well as in response to changing environmental conditions. To study cell-type-specific gene expression profiles, isolation of individual cell types is required because the gene expression levels might not reflect per cell-type changes if a whole organ is analyzed (Benfey et al., 2010). Two prominent approaches for isolating distinct cell types are fluorescence-activated cell sorting and laser-capture microdissection, both of which have been applied to multiple metazoan species including nematode, fruit fly (Drosophila melanogaster), mouse (Mus musculus) and human (Homo sapiens). Schaffner et al., 1997; Bryant et al., 1999; Neira and Azen, 2002; Southall et al., 2013; Spencer et al., 2014; Yuelling et al., 2014, as well as plants (Birnbaum et al., 2003; Dinneny et al., 2008; Carter et al., 2013; Slane et al., 2014). In plants, roots are an ideal system in which to study cell types because they have radial organization with layers of distinct cell types and undergo continuous development following their derivation from stem cells (Benfey and Schiefelbein, 1994). In addition, the cell-sorting-based approaches have been developed to study cell-type-specific expression in Arabidopsis (Arabidopsis thaliana) root development (Birnbaum et al., 2003; Brady et al., 2007) and nitrogen/high-salinity responses among root cell types (Dinneny et al., 2008; Gifford et al., 2008; Geng et al., 2013). These studies of root cell types substantially advance our understanding of how individual cell types differ in gene expression over time and in response to different

---

1This work was supported by a Fulbright Science and Technology Award to S.U.; the U.S. National Science Foundation (IOS-1546617 and DEB-1655386) and U.S. Department of Energy Great Lakes Bioenergy Research Center (BER DE-SC0018409) to S.-H.S.; and a National Science Foundation Graduate Research Fellowship (Fellow ID: 2015196719) to C.B.A.

2These authors contributed equally to the article.

3Present address: Agendia, Inc., Irvine, CA 92618.

4Address for contact: shius@msu.edu.

5Senior author.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Shin-Han Shiu (shius@msu.edu).

S.U. and S.-H.S. conceived the original experimental design and research plan; S.U. and C.B.A. carried out computational experiments and wrote the article; S.-H.S. supervised the project and edited the article.

Articles can be viewed without a subscription. www.plantphysiol.org/cgi/doi/10.1104/pp.19.00653
environmental conditions, including high soil salinity, which results in reduced yield in crops (Hirt and Shinozaki, 2004).

While there is an understanding of how root cell types differ in their transcriptional response to high salinity (Dinneny et al., 2008), it remains a major question how such cell-type-specific response is regulated via cis-regulatory elements (CREs), transcription factors (TFs), cofactors, and chromatin-remodeling complexes (Narlikar and Ovcharenko, 2009). At the cis-regulatory level, multiple studies have utilized cell-type-specific data to globally identify CREs underlying differential gene expression across cell types in metazoans (Wenick and Hobert, 2004; Nègre et al., 2011; Shen et al., 2012; Stergachis et al., 2013; Kellis et al., 2014). A similar study in plants is now feasible for two reasons. First is the availability of Arabidopsis global in vitro TF-binding data generated with protein-binding array and DNA affinity purification sequencing (DAP-seq; Weirauch et al., 2014; O’Malley et al., 2016). Second is the availability of computational methods for identifying putative cis-regulatory elements (pCREs; Rombauts et al., 2003; Gao et al., 2013; Kumari and Ware, 2013; Koryachko et al., 2015; Banf and Rhee, 2017), which have facilitated the identification of stress-related pCREs and those contributing to organ-specific stress response (Zou et al., 2011; Uygun et al., 2017). These CREs can be used further to establish a stress cis-regulatory code with machine learning (Zou et al., 2011; i.e. supervised learning models that address how and to what extent a set of CREs collectively control transcriptional response under a stress condition). Our recent studies on Arabidopsis provided spatial cis-regulatory codes of stress-responsive gene expression at the organ level (root versus shoot; Uygun et al., 2017). Currently, there is no cis-regulatory code available that explains stress responses at the individual cell-type level.

While efforts to map gene regulatory networks at the cell-type level have suggested that regulatory modules may be similar across cell types (Walker et al., 2017), the cis-regulatory mechanisms responsible for plant cell-type-specific responses to external environmental stressors remain largely unknown (Heinz et al., 2015). In this study, we aimed to investigate the cis-regulatory code of high-salinity-responsive gene expression (particularly up-regulation) in six root cell types using an existing data set (Dinneny et al., 2008) with a machine learning approach (Fig. 1). Using a machine learning framework (Samuel, 1959), models were trained to

---

**Figure 1.** Graphical description of data and the machine learning approach. A, Description of the input labels and features used in this study. Labels (i.e. expression response to predict) indicate up-regulation or not under high salinity in six root cell types. Root figures were modified from Dinneny et al. (2008). Colors in the root figures and labels tables correspond to the cell types: epidermis (EPI; purple), cortex (COR; green), endodermis (END; orange), stele (STE; blue), phloem (PHL; yellow), and columella root cap (COL; red). Feature sets (i.e. predictive variables) were included if known transcription factor binding motifs (TFBMs; teal), organ pCREs derived from shoot/root high-salinity expression data (Uygun et al., 2017; pink), and cell-type pCREs derived from the cell-type high-salinity expression data (Dinneny et al., 2008; orange) were present (1) or absent (0) in the proximal promoter of each gene. B, Overview of the machine learning workflow using known TFBM features to predict the up-regulation response in COL as an example. Models were trained on 90% of the genes and validated on the remaining 10% using a 10-fold cross-validation scheme.

---

1740 Plant Physiol. Vol. 181, 2019

Uygun et al.

Downloaded on May 2, 2021. - Published by https://plantphysiol.org
Copyright (c) 2020 American Society of Plant Biologists. All rights reserved.
predict if a gene is up-regulated under high-salinity conditions in a specific cell type (i.e. the label; Fig. 1A) using the presence (1) or absence (0) of many potential cis-regulatory sequences (i.e. the features) in the proximal promoter of each gene. Models were trained and validated using a 10-fold cross-validation scheme (Fig. 1B), such that models were trained with training genes and applied to a set of genes that had been withheld from training (i.e. validation genes) to assess how well the predicted response (purple; Fig. 1B) aligned with the actual response measured experimentally (red; Fig. 1B). Three sets of potential cis-regulatory sequences were used as features: in vitro TF-binding information (teal; Fig. 1A; Weirauch et al., 2014; O’Malley et al., 2016), previously identified whole-root and whole-shoot organ-specific pCREs (pink; Fig. 1A; Uygun et al., 2017), and cell-type pCREs identified in this study (orange; Fig. 1A). We assessed models using these three sets of features to predict and validate root cell-type-specific high-salinity up-regulation and identified pCREs that likely regulate high-salinity up-regulation in each cell type and established cell-type cis-regulatory codes.

RESULTS

Comparison of Organ and Cell-Type Transcriptional Response to High Salinity

In Arabidopsis, differential gene expression in organs and cell types has been studied in a genome-wide manner across developmental stages as well as in response to a variety of environmental stresses (Brady et al., 2007; Kilian et al., 2007; Dinneny et al., 2008). Here, we used Arabidopsis cell-type transcriptome data under high-salinity stress (Dinneny et al., 2008) to dissect the cis-regulatory code driving cell-type response to stress. The six root cell types were COL, COR, STE, PHL, EPI, and END. First, we asked to what extent the root cell-type high-salinity responses differed from the whole-organ high-salinity response (Uygun et al., 2017). We used whole-organ (shoot or root) gene expression data focusing on abiotic and biotic stress treatments over multiple time points (Kilian et al., 2007). To compare global gene expression between samples, we calculated the between-sample Pearson correlation coefficient (PCC; Fig. 2A; Supplemental File S1). The overall expression patterns of genes in high-salinity-treated root cell types (box 1, Fig. 2A), except COR, were more similar to each other than to those of

Figure 2. Gene expression correlation across stress data sets of root, shoot, and root cell types. A, Heatmap of Pearson’s correlation coefficient (PCC) calculated using expression values of all sample pairs. The colors represent PCC values from low (lighter red) to high (darker red). PCC values less than the 95th percentile of all pairwise sample PCCs (0.42) are in white. Boxes with black outline are clusters of similar treatments (e.g. root high-salinity treatment samples clustering with root osmotic treatment samples). Boxes with blue outline and blue text indicate root cell-type samples including COL, COR, STE, PHL, EPI, and END. Boxes with dashed black outline and gray dotted lines emphasize the relationships between root cell-type high-salinity treatment samples with whole-root abiotic stress treatment samples. B, GO categories with overrepresented numbers of high-salinity up-regulated genes in root, shoot, and/or root cell types. Dotted lines separate categories enriched in different numbers of organs/cell types. Shades of red indicate significant overrepresentation with \( q \leq 0.05 \); blue indicates \( q > 0.05 \). ABA, Abscisic acid; JA, jasmonic acid.
whole root and shoot (Fig. 2A), consistent with earlier findings (Dinneny et al., 2008). Additionally, a subset of the high-salinity-treated root cell types was significantly and positively correlated (PCC > 0.42) with the high-salinity/osmotic stress-treated whole-root treatments (samples in boxes 2 and 3, Fig. 2A). This was not the case for other treatments (boxes 4 and 5, Fig. 2A). Given that the cell types examined are a subset of cells examined in the whole-root samples, the observed correlation (PCC > 0.42) between root cell type and whole-root responses is expected. However, we should emphasize that, confirming earlier studies (Dinneny et al., 2008), there is clearly information captured in cell-type data that cannot be obtained from whole-organ data. This difference was also apparent based on an assessment of which Gene Ontology (GO) terms tend to be found among the high-salinity up-regulated genes in root cell types compared with whole-root and shoot up-regulated genes (Fig. 2B; Supplemental Table S1). We found that, despite the substantial differences in their transcriptional programs (Fig. 2A), similar biological processes relevant to high-salinity stress responses are activated regardless of the organ or cell type. That said, there were also substantial differences in enriched GO terms specific to a subset of or to each organ/cell-type response (Fig. 2B; Supplemental Table S1).

Predicting Cell-Type High-Salinity Up-Regulation with Large-Scale in Vitro TF-Binding Data

Given that the regulatory program responsible for controlling cell-type response to stress, not just under high salinity, remains largely unknown, we first focused on identifying TFs likely regulating high-salinity up-regulation in each root cell type. Extensive binding data for 758 Arabidopsis TFs are available from two large-scale in vitro TF-binding studies, Catalogue of Inferred Sequence Preferences of DNA-Binding Proteins (CIS-BP; Weirauch et al., 2014) and DNA Affinity Purification-sequencing (DAP-seq; O’Malley et al., 2016). We first tested which TFs might control cell-type gene expression under high-salinity stress by identifying TFs with overrepresented numbers of DAP-seq binding sites in the promoters of high-salinity up-regulated genes in each root cell type. Among cell types, zero to 140 binding sites of TFs were enriched (Fisher’s exact test, q ≤ 0.05; Supplemental Table S2). For COR and PHL, no sites were significantly enriched. On the other hand, COL and END cell types had the most TF-binding sites enriched (26 and 140, respectively). Those TFs with enriched sites are likely important for regulating high-salinity cell-type responses.

To further investigate the extent to which the current knowledge of large-scale TF-binding data can explain high-salinity up-regulation in these root cell types, we built machine learning models using the presence of CIS-BP (Weirauch et al., 2014) or DAP-seq (O’Malley et al., 2016) sites in the promoter of a gene as predictors of whether the gene in question would be up-regulated in a particular root cell type or not (Fig. 3). This approach allowed us to integrate information from all in vitro TFs into one model that could identify informative TFs that were predictive of expression patterns. The machine learning models were trained on a training data set and tested using an independent validation data set using a cross-validation approach to avoid overfitting (see “Materials and Methods”; Fig. 1B). Note that the training and validation data sets were both from the root cell-type expression data set. Therefore, the predictions of high-salinity response in the validation data set generated by applying the trained model could be directly compared with the experimentally derived expression patterns of genes in the validation set to assess model performance.

Here, the machine learning model performance is measured using the area under the curve-receiver operating characteristic (AUC-ROC), which jointly considers false-positive and true-positive rates, where AUC-ROC = 1 indicates a perfect model and AUC-ROC = 0.5 indicates that the model is no better than random guessing. Consistent with the interpretation that a subset of the TFs is likely involved in root cell-type high-salinity up-regulation, models based on CIS-BP (AUC-ROC = 0.63–0.71) or DAP-seq (AUC-ROC = 0.58–0.68) were better than randomly expected for all six cell-type predictions (Fig. 3A). Among these predictions, the model predicting END response had the best AUC-ROC score (for an alternative measure of performance using precision-recall curves, see Supplemental Fig. S1A). We hypothesized that END genes may be predicted better because they have a stronger response to high salinity. To test this, we compared the level of differential expression in each of the cell types but did not find END genes to be expressed more highly than genes up-regulated in other cell types (Supplemental Fig. S1B). For comparison with known TFBMs, similar models were also established using pCREs derived from whole-organ (Fig. 3B) or cell-type transcriptome. They will be discussed in later sections.

Next, we asked what the most important TFs were for predicting cell-type high-salinity response based on the importance scores of machine learning models (see “Materials and Methods”; Fig. 1B). While some TFs were important across multiple cell types, others were important for only one cell type in predicting high-salinity response (Fig. 3C). For example, TFs belonging to the bZIP and bHLH TF families were repeatedly identified as important for multiple cell types. In contrast, CSD and AT-hook family TFs were important for predicting PHL only, suggesting their roles in cell-type-specific regulation. We should emphasize that, although the root cell-type-responsive gene expression can be predicted better than random by the in vitro TF-binding data, the AUC-ROCs are low, indicating that there is still substantial room for improvement (Fig. 3A; Supplemental Fig. S1A). One potential reason is that the binding data account for ~50% of known Arabidopsis
TFs, even though it is the most extensive for any plant species. Thus, some TFs and their associated binding sites important for regulating cell-type high-salinity up-regulated response may be missed by this approach.

Predicting Cell-Type High-Salinity Up-Regulation with pCREs Identified at the Organ Level

To determine if accounting for TFs with no available in vitro binding data will further improve cell-type high-salinity up-regulation prediction, we used a set of organ (root and shoot) pCREs identified based on gene coexpression under high salinity, as well as another stress treatment expression data set (Uygun et al., 2017), as input for building the high-salinity up-regulation prediction models (organ pCREs were used instead of TFBMs in Fig. 3A). Three sets of organ pCREs were considered: (1) general organ pCREs, with sites that are enriched among both root and shoot high-salinity up-regulated genes; (2) whole root pCREs, with sites enriched among genes up-regulated in root only; and (3) whole-shoot pCREs, with sites enriched among genes up-regulated in shoot only (Uygun et al., 2017; Fig. 1A). Considering that the high-salinity-treated root cell types had a similar expression profile to the high-salinity and osmotic stress-treated whole root (PCCs $95^{th}$ percentile PCC from all pairwise sample comparisons; Fig. 2), we expected that whole-root pCREs and general organ pCRE together might explain root cell-type-specific high-salinity up-regulation.

Using the presence of these previously described organ pCREs as features for machine learning, we found that the models based on whole-root + general
organ pCREs outperformed models using in vitro TF-binding data in predicting cell-type high-salinity up-regulated genes (AUC-ROC = 0.68–0.78; cyan, Fig. 3B; example precision-recall curves are shown in Supplemental Fig. S1C). Interestingly, there was little improvement over the TF-binding models using only whole-root pCREs (red, Fig. 3B). Instead, the major contributors were the general organ pCREs, as those pCREs substantially improved model performance when used alone. Finally, the addition of the third sets of pCREs, the whole-shoot pCREs, to the whole-root + general organ pCREs did not improve performance further.

We hypothesized that general organ pCREs were better predictors of cell-type high-salinity response than whole-root pCREs for two reasons. First, multiple representatives and derivatives of known stress-responsive elements, such as the abscisic acid-responsive element (ACGTGG/T), were among the general organ pCREs because these elements are associated with TFs that regulate high-salinity responses across organ types (Uygun et al., 2017). Therefore, these pCREs would be useful predictors regardless of the cell type of interest. Second, because the whole-root pCREs were identified using the whole-root expression data set, any useful signals from specific root cell types would likely be muted or lost, indicating the need to identify novel pCREs based on individual root cell-type gene expression data.

Identifying and Characterizing Root Cell-Type pCREs Associated with High-Salinity Up-Regulation

In earlier studies, human cell-type-specific CREs were identified for expression prediction using cell-type gene expression data and other information (Natarajan et al., 2012; Chen et al., 2013). Considering that the whole-root pCREs were identified without considering responses of the cell types, we might miss regulatory information that was only available by examining cell-type-level data. To identify root cell-type pCREs that might be involved in Arabidopsis high-salinity stress, we used existing root cell-type high-salinity response data from COL, COR, END, EPI, PHL, and STE (Dinneny et al., 2008) and whole-root abiotic stress data (Kilian et al., 2007). We identified 3,095 pCREs from putative promoters of genes in high-salinity clusters (see “Materials and Methods”; Supplemental Files S2 and S3), with high-salinity clusters defined as coexpression clusters with an overrepresented number of high-salinity up-regulated genes. For each pCRE X, if its sites were enriched among high-salinity up-regulated genes of a cell-type Y, we refer to X as a cell-type pCRE for Y (Supplemental Table S3). The number of pCREs for each cell type was correlated with the number of high-salinity up-regulated genes from each cell type (PCC = 0.95, P = 0.005), reflecting a potential relationship between the cis-regulatory complexity and the extent of high-salinity up-regulation in different cell types. We classified a cell-type pCRE as a (1) cell-type specific, (2) multiple cell-type, and (3) general cell-type pCRE based on if the pCRE was identified for only one, two to five, and all six cell types, respectively (Fig. 4A). We found 583 general cell-type pCREs (note that these are distinct from the general organ pCREs discussed earlier; Fig. 3B). In addition, between seven and 360 cell-type-specific pCREs were identified for each cell type (Fig. 4A), suggesting that the extent of cell-type-specific regulation of high-salinity up-regulation may differ by cell type.

Before using these cell-type pCREs as features in our predictive models, we first determined if they are similar to known TFBMs and how similar they are to cell-type pCREs from other cell types. Best-matching TFBMs for the cell-type pCRE most significantly enriched (P < 10^{-6}) in that cell type are shown in Figure 4B. For some of these cell-type pCREs, the most significantly enriched (P < 10^{-6}) had high sequence similarity with a known TFBM. For example, the most enriched PHL pCRE was a perfect match (PCC = 1) to the WRKY50 TFBM, suggesting that this TF could be important in regulating PHL high-salinity-responsive gene expression. However, others, such as the most enriched COR and EPI pCREs, matched poorly to the

---

**Figure 4.** Classification of cell-type pCRE sets. A, Heatmap of over-represented pCREs. Each row is a pCRE, and red color is for overrepresentation of that pCRE in the cell-type high-salinity up-regulated genes. Top numbers indicate cell types and numbers of cell-type pCREs. B, Sequence logo of the most highly enriched pCRE in each of the six cell types. For the general cell-type pCREs (GEN), an example among the highly enriched motifs is given. Fisher’s exact test q values are as follows: COL, 1.1 × 10^{-6}; COR, 2.94 × 10^{-11}; END, 5.11 × 10^{-11}; EPI, 1.24 × 10^{-12}; PHL, 6.08 × 10^{-14}; STE, 2.25 × 10^{-14}; and GEN, less than 10^{-20}.
known TFBMs (Fig. 4B) and could represent previously not characterized cis-regulatory sequences. To determine which enriched pCREs represent known and which represent putative TFBMs, we developed three tests of statistical significance. In the first test, we asked which cell-type pCREs had greater sequence similarity to a specific TFBM than 95% of TFBMs in the same TF family (orange bars, Fig. 5A). We found that 825 cell-type pCREs (27%) passed this test and were considered representatives of specific known TFBMs. In the second test, we asked which cell-type pCREs had greater sequence similarity to a TFBM in a TF family than 95% of TFBMs from other TF families (dark green bars, Fig. 5A). We found that 73.3% of pCREs qualified and considered these pCREs novel but belonging to a known TF family. Only one pCRE did not meet either of the thresholds defined in the above two tests; however, it was more similar to a known TFBM than 95% of random k-mers (the third test), suggesting that it may also be a novel TFBM but is likely from a known family. Overall, these findings suggest that while some cell-type pCREs likely represent specific known TFBMs, the rest may be novel TFBMs within these TF families.

To determine how similar cell-type pCREs (cell-type specific, multiple cell-type, and general cell-type) were to organ pCREs (whole root, whole shoot, and general

**Figure 5.** Characterization of cell-type pCREs. A, PCCs between cell-type pCREs and the best matching, known TFBM from the DAP-Seq and CIS-BP databases (blue dots). Three significance thresholds are shown: orange, significantly similar to a specific TFBM (i.e. PCC between the pCRE in question and a specific TFBM; greater than 95th percentile of PCCs between TFBMs of TFs within the same family); dark green, significantly similar to a TF family (i.e. greater than 95th percentile of PCCs between TFBMs of TFs from different TF families); light green, significantly similar to a TFBM compared with random 6-mers (i.e. greater than 95th percentile of PCCs between TFBMs of TFs from a family and 1,000 randomly generated 6-mers). TF family names with asterisks indicate two or fewer TFBMs available for that TF family, so within TF family distribution (orange) could not be calculated. B, Heatmap of similarity (PCCs between PWMs) among pCRE sets. Root/shoot, pCREs enriched among up-regulated genes from whole root/shoot data. Boxes with thicker outlines represent self-self comparisons; red and blue outlined boxes are described in the text.
organ), we determined the average similarity of pCREs (PCC) within (diagonal) and between (upper triangle) these different pCRE sets (Fig. 5B). The similarities of pCREs within four cell-type-specific sets (COL, COR, END, and PHL) is higher (PCC = 0.56–0.63) than the similarities across sets (average PCC = 0.41), indicating that high-salinity up-regulation in these cell types involves distinct types of cis-regulatory sequences. When we considered cross-set comparisons, one notable finding is that the general cell-type set and the whole-root set (PCC = 0.42; cyan rectangle, Fig. 5B) are not any more similar to each other than they are to the other sets (average PCC of all cross-set comparisons = 0.44). This is also true when we compare the similarities between each cell-type-specific set with the root-specific set (magenta rectangle, Fig. 5B). These findings further highlight the differences between whole-root and root cell-type response.

Contributions of Different pCRE Sets to Models Predicting High-Salinity Up-Regulation in Different Cell Types

The sequence differences between the cell-type pCREs and the organ pCREs led us to hypothesize that the motifs that we identified could be novel motifs important for driving high-salinity up-regulation among root cell types. To assess this, for each cell type, we first used all 3,095 cell-type pCREs as predictors to build a machine learning model (all-cell-type) for predicting high-salinity up-regulated genes in that cell type. The AUC-ROC for the all-cell-type pCRE models ranged from 0.68 for EPI to 0.76 for END (purple, Fig. 6A; Supplemental Table S4). Because only a subset of the 3,095 cell-type pCREs were enriched in high-salinity up-regulated genes in each cell type, we also used just the pCREs enriched in genes up-regulated under high salinity in each cell-type X. The cell-type X models (cyan, Fig. 6A) performed just as well as those using all-cell-type pCREs. Although not surprising, this serves as a quality control of our approach, in that adding pCREs that may be important for high-salinity regulation in other cell types does not improve model performance for that cell type.

Because the cell-type X models are based on genes up-regulated in cell-type X that may also be up-regulated in one or more other cell types, such cell-type X models use a combination of three pCRE sets: (1) general cell-type, (2) multiple cell-type, and (3) cell-type-specific pCREs. To distinguish the contributions of these three sets of pCREs in the model, we next builds models using only general cell-type or only cell-type-specific pCREs for each cell type. For COL, COR, END, and EPI, the general cell-type pCREs (yellow, Fig. 6A; Supplemental Fig. S2A) were better predictors than the cell-type-specific pCREs (red, Fig. 6A), indicating that in these cell types, high-salinity response tends to be controlled by a general cis-regulatory code. On the other hand, PHL and STE high-salinity up-regulation were predominantly driven by cell-type-specific pCREs (Fig. 6A; Supplemental Fig. S2A), highlighting the differences in general versus cell-type-specific controls between root cell types. In addition, cell-type pCREs can be used to predict cell-type high-salinity up-regulated genes better than using in vitro TF-binding data alone (e.g., CIS-BP; red lines, Figs. 3A and 6A). This indicates that the cell-type pCREs further improve our knowledge of the root cell-type cis-regulatory program.

We next assessed if combining general organ and cell-type pCREs would further improve our ability to predict cell-type high-salinity up-regulation. Surprisingly, the performance of cell-type pCRE-based models was not better than the models based on the combination of general organ and whole-root pCREs in predicting up-regulation in various cell types (blue line, Figs. 3B and 6A). Furthermore, when we used all-organ and all-cell-type pCREs to build a model for each cell type (Fig. 6B; Supplemental Fig. S2B), our ability to predict high-salinity up-regulated genes was not improved compared with using just the organ pCREs.
(blue line, Figs. 3B and 6A). This was unexpected because these cell-type pCREs were derived directly from the cell-type expression data sets and were different compared with organ pCREs (Fig. 5B).

Finally, we assessed if incorporating additional information about the cell-type pCREs would improve their predictive power. We considered conserved non-coding sequence (CNS), DNase I hypersensitivity (DHS), and known TF binding (DAP), using high-salinity response in the END cell type as an example. We found that the cell-type pCRE sites were more likely to overlap with conserved (based on CNS), accessible (based on DHS), and bound (based on DAP) regions when they were in END high-salinity up-regulated gene promoters as opposed to nonresponsive gene promoters (Supplemental Fig. S3A). To further assess if incorporating this information would allow improvement of our prediction models, we next considered only cell-type pCRE sites that overlapped with CNS, DHS, and DAP regions in our machine learning models. For models based on pCRE sites overlapping with CNS (AUC-ROC = 0.63) or DHS (0.69), they have substantially decreased performance compared with using all pCRE sites (0.76; Supplemental Fig. S3B). In addition, models using only pCREs that overlapped with DHS regions only resulted in marginal improvement in model performance (0.77; Supplemental Fig. S3B). Thus, we did not include CNS, DHS, and DAP data in further analyses.

**Characteristics of High-Salinity Up-Regulated Genes Predicted Correctly by Models Based on Organ and Cell-Type pCREs**

One explanation for the similar performances of organ and cell-type pCRE-based models was that they correctly predicted different sets of high-salinity up-regulated genes. Using the STE high-salinity up-regulated genes as an example, we found that the all-organ pCRE-based models and the cell-type pCRE-based models had very similar true-positive rates at 61% and 63%, respectively. However, 12% of the STE high-salinity up-regulated genes were only correctly predicted by the organ pCREs and 14% were only predicted correctly by cell-type pCREs. While genes predicted correctly by both cell-type and organ pCREs had significantly higher STE expression than those not predicted correctly by either (Kolmogorov–Smirnov test, \( P = 5e-4 \); Fig. 6C), the effect size was small. In addition, there were no significant expression level differences observed in genes correctly predicted by only the cell-type or only the organ pCRE sets (Kolmogorov–Smirnov test, \( P = 0.15–0.46 \); Fig. 6C). Together, these findings suggest that expression levels in STE are not responsible for the differences in which STE high-salinity up-regulated genes were correctly predicted by models using organ versus cell-type pCREs.

To shed more light on why some of these genes were or were not correctly predicted, we also looked for differences in expression patterns and functions between these sets of genes. Focusing on differences in expression patterns, we found that STE genes correctly predicted by both the cell-type and organ pCRE-based models tend to belong to expression clusters highly up-regulated at all time points under osmotic stress and at later time points under cold stress (clusters 2 and 8, Fig. 6D; Supplemental Fig. S4; Supplemental Table S5). Notably, genes not correctly predicted by either model tend to be highly up-regulated under heat stress and at later time points under osmotic stress (cluster 7, Fig. 6D). In addition to significant differences in expression patterns, STE genes predicted by both organ and cell-type pCRE models were enriched for genes in GO categories including response to water deprivation, abscisic acid, and cold compared with STE genes not predicted by either pCRE set (Fisher’s exact test; \( q = 7–9e-3 \); Supplemental Table S6). Together, these findings suggest that both organ and cell-type pCREs were better able to identify STE high-salinity up-regulated genes that were also up-regulated at the organ level under similar conditions.

Because genes predicted by different pCRE sets were not enriched (Supplemental Table S7) for genes with similar pCRE profiles (Supplemental Fig. S5), it remains unclear what cis-regulatory information explains the difference in predictive ability. The lack of improvement in the models including both organ and cell-type pCREs (Fig. 6B) is likely due to overfitting, where increasing the number of predictors (pCREs) does not make the model better because there is no corresponding increase in observations (high-salinity up-regulated and nonresponsive genes). Nonetheless, we should emphasize that, despite the caveats, the organ set and the cell-type sets contain complementary cis-regulatory information. Jointly, they provide a comprehensive picture of the complexities of cis-regulatory control at the cell-type level under an environmental perturbation.

**DISCUSSION**

The existing root high-salinity transcriptomic data (Dinney et al., 2008) provide a rich resource to not only dissect the extent of gene expression in distinct environments across different cell types but also provide insights into the molecular mechanisms regulating cell-type transcriptional response. In this study, we identified pCREs likely responsible for cell-type high-salinity up-regulation in Arabidopsis. Taking it a step further, we established cell-type cis-regulatory codes with these pCREs that reveal the relative importance of different pCREs in regulating high-salinity up-regulation in different cell types. By contrasting the cis-regulatory codes governing whole-organ (root or shoot) expression (Uygun et al., 2017) with those for the cell types used in this study, we found that cell-type and whole-root pCREs regulate only a partially overlapping set of high-salinity up-regulated genes. We also demonstrated
that the pCRE-based models perform better than existing in vitro TF-binding data in predicting high-salinity up-regulation. Our findings demonstrate the feasibility of using computationally identified pCREs to establish machine learning models that can predict genome-wide transcriptional changes to specific environmental conditions at a cell-type level of resolution. In addition, our work shows how these pCREs can be studied in order to determine the extent to which known versus novel regulatory elements are predictive of cell-type-specific responses to high salinity or any other stress of interest. Furthermore, using this machine learning-based framework, we demonstrate how one can quantitatively assess how much the regulatory information available tells us about gene regulation.

With the importance noted, we also become aware of a few limitations through this study. The first limitation is related to our approach in identifying pCREs from clusters of coexpressed genes. Although the approach has been fruitful, the relationship between coexpression and coregulation is far from perfect (Allocco et al., 2004). In addition, not all regulatory sequences among coregulated genes can be efficiently identified by motif finders, mainly due to the discovery that the three-dimensional structure adopted by the regulatory sequences can be more important than the primary sequences (Parker et al., 2009; Tsai et al., 2015). The second limitation is related to how pCREs are used for modeling. The identified pCREs are in the form of PWMs that are mapped to the Arabidopsis genome. To counter the high false-positive rate in site mapping using PWMs (Wasserman and Sandelin, 2004), we have set a relatively stringent threshold mapping P value that err on the side of missing relevant sites. In addition, in this study, the cis-regulatory code is built on relatively simple regulatory logic: how the presence or absence of pCRE sites in the proximal promoter region may predict up-regulation. Given the complexity of gene regulatory networks, future studies incorporating hypothesized regulatory network motif information and/or considering combinatorial relations (Uygun et al., 2017) and copy numbers (Ezer et al., 2014) of pCREs may further improve prediction.

The third limitation relates to the types of information considered in the model. The current model is built on cell-type gene expression data with only one time point. Recently, a data set consisting of multiple high-salinity treatment time points across four root cell types (COL, COR, EPI, PHL, and STE) has become available (Geng et al., 2013). It is anticipated that the time-course data will allow better clustering of coregulated genes and should therefore be considered in future studies. Furthermore, it was expected that incorporating additional regulatory information with a cell-type-level resolution (e.g. TF binding, chromatin accessibility, and sequence conservation) would lead to further improvement of the regulatory models. However, when we incorporated this information into our END cell-type high-salinity up-regulation prediction model, performance did not improve. This could be due to one or more of a number of limitations in the data we had available. These limitations include (1) TF-binding data only being available for ~38% of known TFs in Arabidopsis and (2) TF-binding and DNase I hypersensitivity data were not generated at the cell-type-specific level or under high-salinity conditions and therefore may not be representative of the TF binding and chromatin landscape in these cells.

Apart from the limitations noted above, our study provides a comprehensive cis-regulatory code controlling transcription at the cell-type level in response to a stressful environment. This is an important step forward beyond earlier predictions based on organ-level transcriptional response to high salinity (Uygun et al., 2017). The computational models provide estimates on how well cis-regulatory sequences alone may account for the regulatory information necessary to control cell-type transcriptional response to a stressor. The models also provide mechanistic insight on how cell-type transcriptional responses are regulated by cis-regulatory sequences. Our study represents an important step in establishing detailed, statistical models of stress-responsive gene expression of plant cell types. While our focus is on high-salinity conditions, this approach could be used to study responses to any environment condition or treatment or to model changes in gene expression at different stages of development. With future infusion of additional regulatory information, we anticipate that the machine learning approaches used here will allow for more accurate models of spatial gene regulation in diverse environmental contexts.

MATERIALS AND METHODS

Gene Expression Data Sets and Their Processing

The root cell-type high-salinity stress expression data set (Dinneny et al., 2008) was downloaded from the Gene Expression Omnibus (GSE7641). This expression data set consists of control and high-salinity stress conditions (150 mM NaCl treatment for 1 h) for the following cell types: COL, COR, END, EPI, PHL, and STE. The Affymetrix CEL files were preprocessed and quantile normalized using the Bioconductor affy package in the R environment (https://bioconductor.org/packages/release/bioc/html/affy.html). For differential gene expression, log2 fold changes and associated P values were calculated using high-salinity stress treatment and corresponding control samples for each cell type with the limma package from Bioconductor (Ritchie et al., 2015). The P values were adjusted for multiple testing (Benjamini and Hochberg, 1995). The whole-root abiotic stress data set from AtGenExpress (https://www.arabidopsis.org/portals/expression/microarray/ATGenExpress.jsp) was also used and processed according to a previous study (Zou et al., 2011). A gene was considered up-regulated in a cell type or in the whole root if its log2 fold-change value was ≥1 and the adjusted P was ≤0.05. Nonresponsive genes were defined as genes that were neither up- nor down-regulated under any stress at any time point in any sample from the AtGenExpress data as defined previously (Uygun et al., 2017).

GO Enrichment Analyses

To find functional categories that were significantly overrepresented or underrepresented in the organ and root cell-type up-regulated genes, GO-SLIM terms were retrieved (http://www.geneontology.org/ontology/subsets/goslim_plant.obo). The genes annotated to each GO term that were high-salinity up-regulated (in root, shoot, or one of the six root cell types) were...
Gene Coexpression Analyses

To find coexpressed gene clusters, the root cell-type high-salinity stress expression data set was combined with the root stress expression data set from AtGenExpress (Kilian et al., 2007). Genes in the combined data set were classified into coexpression clusters using c means (Hathaway et al., 1996) in the R environment. Among the resulting clusters (Supplemental File S4), those with less than 10 genes were excluded from further motif-finding analyses because the motifs identified would have limited statistical support. The clusters with more than 60 genes were further divided using c means, resulting in 536 clusters included for further analysis, each with 10 to 60 genes (Supplemental File S4).

This range of number of genes in a cluster was required for efficiently running the motif finders (Zou et al., 2011). Fisher’s exact test was used to identify the clusters with overrepresented numbers of high-salinity up-regulated genes in each root cell type compared with the rest of the genome (multiple testing $q \leq 0.05$; Storey, 2003).

Analysis of TF-Binding Data

To identify whether existing in vitro TF-binding data could explain root cell-type high-salinity-responsive gene expression, two sets of TF-binding data sets were obtained. These data sets included position frequency matrices from the CIS-BP database (Weirauch et al., 2014) and DAP-seq peaks (https://github.com/ctb/motility). DAP-seq peaks (a 200 bp long) obtained from the Arabidopsis (Arabidopsis thaliana) cistrome study (O’Malley et al., 2016). The handling of the CIS-BP and DAP-seq TF-binding data was as described previously (Uygun et al., 2017). CIS-BP position frequency matrices were converted to PWMs adjusted for the background AT (0.33) and GC (0.17) contents of Arabidopsis promoter and mapped to promoter sequences using Motility (https://github.com/ctb/motility). DAP-seq peaks (a 200 bp long) that correspond to Arabidopsis promoters and with the fraction of reads in peaks $> 5\%$ were used in predictions. The promoters were defined as the regions $1,000$ bp upstream of transcription start sites.

Identification of pCREs Regulating Cell-Type High-Salinity Response

To identify pCREs relevant to high-salinity up-regulation in a cell type from the promoter regions, a previously established pipeline (Zou et al., 2011) was applied to each coexpression cluster enriched in high-salinity up-regulated genes in a given cell type. This pipeline tested if a pCRE, $X$, was significantly more likely to be found, at least once, in the promoter regions of genes up-regulated in a cell type, $C$, compared with the promoters of nonresponsive genes, using Fisher’s exact test. To evaluate the impact of threshold significance levels in calling a pCRE as overrepresented, we applied two threshold $q$ values, at 0.05 and $10^{-5}$, that lead to 7,417 and 3,095 pCREs (Supplemental Files S2 and S3) enriched in at least one cell type, respectively. As the prediction performances (see “Predictive Models of Cell-Type High-Salinity Up-Regulation” below) were similar using 7,417 (AUC-ROC $= 0.71 - 0.79$) and 3,095 (AUC-ROC $= 0.68 - 0.76$; Supplemental Table S4) pCREs, the pCRE set with a more stringent enrichment threshold ($q < 10^{-5}$) was used in further analyses. To assess similarity between pCREs and between pCREs and known TFBSs, a PCC was calculated using the PWMs of a pCRE pair or a PCRE-known TFBM pair as described previously (Zou et al., 2011).

To determine if considering only cell-type pCREs that are conserved or are important pCREs from each set, as determined by our predictive models, were predicted correctly by different sets of pCREs. Only the top 100 most important pCREs from each of the eight clusters using the Fisher’s exact test and multiple testing correction described above. The same approach was used to identify differences in the presence or absence of organ and cell-type pCREs between STE genes that were predicted correctly by different sets of pCREs. Only the top 100 most important pCREs from each set, as determined by our predictive models, were used for clustering.

Predictive Models of Cell-Type High-Salinity Up-Regulation

To establish the cis-regulatory code for genes up-regulated by high-salinity treatment in each cell type, we built a machine learning model capable of predicting whether a gene would be up-regulated or nonresponsive for each of the cell types of interest. A machine learning approach, a form of predictive modeling, was taken in order to allow the models to learn the relationship between putative regulatory elements and the up-regulation response from the data, rather than biasing the model with a predetermined structure. The SVM (Cortes and Vapnik, 1995) machine learning algorithm, implemented in the Waikato Environment for Knowledge Analysis (Frank et al., 2004), was used in this study.

To find the optimal parameters for classification, grid searches were performed. The parameters for SVM were: (1) the ratio of nonresponsive to up-regulated genes, (2) the soft margin, and (3) the gamma parameter of the radial basis function kernel. A standard 10-fold cross-validation scheme was used to prevent model overfitting and assess the predictive ability of our models on data not used to train them. In this scheme, the genes were randomly split into 10 sets, then sets 1 to 9 were used to train the model and the trained model was then applied to set 10, then sets 1 to 8 and 10 were used to train and the model was applied to set 9, etc. (Fig. 1B). The performance metrics were then calculated using only the predictions from the testing set. Two measures were used to evaluate the prediction performance. The first was the AUC-ROC measure, where a perfect model would have AUC-ROC $= 1$ and random predictions would lead to AUC-ROC $= 0.5$. The second approach was plotting the precision-recall curve, where precision was the ratio of true-positive predictions to the overall number of genes that were predicted as positive and recall was the ratio of true-positive predictions to the total number of positive class (high-salinity up-regulated genes in a cell type). The models with satisfactory classification would have precision-recall curves toward the top-right corner of the graph, and the models with random predictions would be no better than the background of the ratio of positive to negative class. To assign importance scores to the features, SVM models were built using Scikit-Learn (Pedregosa et al., 2011) and the absolute value of the coefficients assigned to each feature over 100 replicates was averaged. A gene was considered correctly classified by the model if its median predicted probability score over the 100 replicates was greater than the decision threshold.

Subclustering of STE High-Salinity Up-Regulated Genes

To determine if high-salinity up-regulated genes that were predicted correctly by different sets of pCREs had different expression patterns, we used STE genes as an example by clustering them into eight clusters based on their pattern of differential expression in response to a range of stress conditions using $k$-means (Ipc R package; https://cran.r-project.org/web/packages/ipc/index.html) in the R environment. Next, genes that were correctly predicted by both, either, or neither cell-type and organ pCREs were tested for enrichment of genes from only the eight clusters using the Fisher’s exact test and multiple testing correction described above. The same approach was used to identify differences in the presence or absence of organ and cell-type pCREs between STE genes that were predicted correctly by different sets of pCREs. Only the top 100 most important pCREs from each set, as determined by our predictive models, were used for clustering.

Supplemental Data

The following supplemental materials are available.

Supplemental Figure S1. Precision/recall of high-salinity up-regulation prediction models using TF-binding data and organ pCREs.

Supplemental Figure S2. Precision/recall of high-salinity up-regulation models using cell-type pCREs and cell-type + organ pCREs.
Supplemental Figure S3. Overlap of conservation, accessibility, and TF-binding regions with END cell-type pCREs and impact on model performance.

Supplemental Figure S4. Expression profile clusters of STE high-salinity up-regulated genes.

Supplemental Figure S5. Principal component analysis of STE genes based on the pattern of organ or cell-type pCREs.

Supplemental Table S1. GO-SLIM enrichment results.

Supplemental Table S2. DAP-seq enrichment results.

Supplemental Table S3. Cell-type pCRE enrichment results.

Supplemental Table S4. AUC-ROC and \( \Phi \) values of prediction results.

Supplemental Table S5. Enrichment of STE expression clusters in STE genes predicted by different sets of pCREs.

Supplemental Table S6. GO-SLIM enrichment results for STE genes predicted by different sets of pCREs.

Supplemental Table S7. Enrichment of STE organ cell-type pCRE profile clusters in STE genes predicted by different sets of pCREs.

Supplemental File S1. PCC values between expression samples.

Supplemental File S2. Cell-type pCREs.

Supplemental File S3. Cell-type pCRE PWMs.

Supplemental File S4. Coexpression clusters (\( n = 538 \)) used for the enrichment test.

ACKNOWLEDGMENTS

We thank Alexander Seddon for helping with expression data processing and programming in establishing the analysis pipeline during the initial phase of the project and Ronan O’Malley for advice on the use of DAP-seq data. We also thank members of the Shi lab for their valuable suggestions to our project. Received May 31, 2019; accepted September 11, 2019; published September 24, 2019.

LITERATURE CITED


Kumari S, Ware D (2013) Genome-wide computational prediction and analysis of core promoter elements across plant monocots and dicots. PLoS ONE 8: e79011


1750 Plant Physiol. Vol. 181, 2019

Copyright (c) 2020 American Society of Plant Biologists. All rights reserved.


Tsai ZTY, Shiu SH, Tsai HK (2015) Contribution of sequence motif, chromatin state, and DNA structure features to predictive models of transcription factor binding in yeast. PLOS Comput Biol 11: e1004418


