doi: 10.1093/nar/gkaa264

Prediction of condition-specific regulatory genes using machine learning

Qi Song[1], Jiyoung Lee[1], Shamima Akter[2], Matthew Rogers[3], Ruth Grene[1,2] and Song Li[1,2,*]

[1]Graduate program in Genetics, Bioinformatics and Computational Biology. Virginia Tech., Blacksburg, VA 24061, USA, [2]School of Plant and Environmental Sciences. Virginia Tech., Blacksburg, VA 24061, USA and [3]Department of Statistics. Virginia Tech., Blacksburg, VA 24061, USA

Received November 06, 2019; Revised February 19, 2020; Editorial Decision March 31, 2020; Accepted April 20, 2020

ABSTRACT

Recent advances in genomic technologies have generated data on large-scale protein–DNA interactions and open chromatin regions for many eukaryotic species. How to identify condition-specific functions of transcription factors using these data has become a major challenge in genomic research. To solve this problem, we have developed a method called ConSReg, which provides a novel approach to integrate regulatory genomic data into predictive machine learning models of key regulatory genes. Using Arabidopsis as a model system, we tested our approach to identify regulatory genes in data sets from single cell gene expression and from abiotic stress treatments. Our results showed that ConSReg accurately predicted transcription factors that regulate differentially expressed genes with an average auROC of 0.84, which is 23.5–25% better than enrichment-based approaches. To further validate the performance of ConSReg, we analyzed an independent data set related to plant nitrogen responses. ConSReg provided better rankings of the correct transcription factors in 61.7% of cases, which is three times better than other plant tools. We applied ConSReg to Arabidopsis single cell RNA-seq data, successfully identifying candidate regulatory genes that control cell wall formation. Our methods provide a new approach to define candidate regulatory genes using integrated genomic data in plants.

INTRODUCTION

Understanding transcriptional regulation in plants is crucial to the improvement of crop productivity under adverse environmental conditions (1,2). Over the past decades, thousands of expression profiles have been generated to investigate how environmental perturbations and developmental cues regulate gene expressions in plants (3). Protein–DNA interaction assays such as large-scale chromatin immunoprecipitation sequencing (ChIP-seq) (4), protein binding microarrays (5), enhanced yeast one-hybrid (6–8), and DNA affinity purification sequencing (DAP-seq) (9,10) have generated millions of candidate TF–target gene interactions. ATAC-seq (assay for transposase-accessible chromatin using sequencing) and DNase hypersensitive assays have enabled profiling of active chromatin regions under specific conditions or tissue types (11–15). With this large amount of regulatory-genomic data becoming available, a current major computational challenge is the integration of data from protein-DNA interactions, active chromatin region measurements, and gene expression assays to discover novel and key regulators that operate under specific conditions in plants.

Regulatory mechanisms have been revealed by constructing genetic regulatory networks (GRN) that contain thousands of TF–target interactions. Many approaches have been developed to construct GRNs by combining different types of genomic data. Early attempts explored the use of unsupervised methods, in which known TF–target information was not considered. For example, relevance network and other mutual information-based approaches have been developed to infer interactions (16–19). Other unsupervised methods have also been developed including those based on partial correlation (20), weighted co-expression networks (21) and ensemble approaches (22). By contrast, supervised machine learning approaches which take known interactions as prior knowledge, have also been applied. Several commonly used supervised models can infer GRNs from expression data, including support vector machine (23,24), least angle regression (25), least absolute shrinkage and selection operator (26,27) and elastic net (28). Some of these approaches need gene expression data from multiple samples such as those from a time course experiment (26–30) or multiple tissue- or cell-types (16–19). Such experiments

[*]To whom correspondence should be addressed. Tel: +1 540 231 2756; Email: songli@vt.edu
Present address: Qi Song, Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School., Boston, MA 02115, USA.

are typically time consuming and are still not available for many plant species (see Supplementary table S1 for a summary of published methods). For these methods, each inferred interaction represents an association between a given TF and its target genes across many observations. However, since regulatory interactions sometimes are characterized by condition-specific binding of TFs to *cis*-regulatory elements (31), such binding events might not be reflected across all data points used by an inference algorithm.

Other methods for the inference of interactions focus on data types that represent direct binding events. Binding site data have received much attention in recent years in plant research as evidenced by databases such as PlantTFDB (32), AGRIS (33) and Grassius (34), which have accumulated substantial amounts of data documenting experimentally identified binding sites. Previous studies have identified a considerable number of binding sites from data obtained *in vivo* related to different environmental perturbations in plants. For example, binding sites were screened to construct regulatory networks in response to far red light (35,36), hormones (37–39) and fungal infection (40) in *Arabidopsis thaliana*. Based on available binding site data, several web-based tools have been developed to prioritize the targets of specific TFs for a group of genes using enrichment analysis. Some examples include TF2Network (41) and Cistome (42), which compute enrichment of binding sites for corresponding TFs based on large collections of documented binding sites in Arabidopsis. PlantPAN 3.0 (43) identifies enriched combinations of TFs for multiple plant species. The direct evidence available for binding site identification also has some limitations. For example, due to the cost of ChIP-seq experiments, typically only a few TFs have been screened under any specific condition. Compared to ChIP-seq, DAP-seq can identify possible targets of thousands of TFs efficiently (9). However, DAP-seq is an *in vitro* technique (10), and some binding sites detected by DAP-seq may not be available for binding under a given environmental perturbation. Therefore, integration of binding site and expression data is key to improving prediction accuracy under specific conditions or cell types.

In this study, we developed the condition specific regulatory network inference engine (ConSReg), a machine learning approach which integrates expression data, TF–DNA binding data and open chromatin data to infer condition-specific regulatory genes. In ConSReg, lists of differentially expressed genes (DEGs) were supplied to machine learning models to perform binary classification with feature selection by regularization. This procedure can prioritize and select the most relevant TFs for a specific environmental perturbation. We performed cross-validation for ConSReg using a compendium of expression data sets obtained under different environmental perturbations from 26 different publications (see Supplementary Table S2). The evaluation result shows that the features of the integrated representation can accurately predict the expression of target genes (average area under curve for receiver operating characteristic curve, AUC–ROC = 0.84) and is significantly better than alternative approaches.

Our results highlight several important discoveries that provide new insights into the regulation of gene expression in plants. First, the appropriate selection of negative training data sets is crucial for the improvement of model performance, specifically, undetected genes (UDGs) are better negative training data than non-differentially expressed genes (NDEGs). Second, we demonstrated that including ATAC-seq data significantly improves model performance regardless of the experimental conditions, whereas prior publications of plant data only demonstrated enrichment of binding sites or regulatory motifs in ATAC-seq peaks (11,12,44,45). Third, we found that the length of promoter regions contributes to model performance. Although published studies show that stress-regulated motifs are enriched in 500 bp upstream of the TSS of target genes (46), our analysis showed that using 3KB upstream of TSS + 0.5KB downstream of TSS as promoter provides better performance across all data sets, which is consistent with another independent report (41). The results of ConSReg are compared with multiple published approaches (41–43) using an independent validation data set (47). We have found that ConSReg consistently provide better ranking for true regulatory genes than other existing approaches. Finally, we applied ConSReg to two recently published single cell sequencing data from plants to infer regulatory networks at the single cell level and identified candidate regulatory genes for cell wall formation in the endodermis. ConSReg is implemented as an open source python package (GitHub repository: https://github.com/LiLabAtVT/ConSReg) with flexible parameter settings such that it can be used in other eukaryotic species, which is demonstrated as we applied this method to a recently published maize DAP-seq and ATAC-seq data (15).

MATERIALS AND METHODS

Single cell RNA-seq expression data preprocessing

The single cell data set was a combination of two separate data sets from individual experiments (Gene Expression Omnibus (GEO) with accession number GSE122687 (48) and GSE123013 (49)). The R package Seurat v3.0 was used to apply basic quality control procedures in order to remove outliers in feature counts and to ensure no contaminated cells were included in our analyses (50). Once the data sets were cleaned, they were normalized using Seurat's built-in 'NormalizeData' function which applies a log-normalization to the feature counts across the rows of the data set. The factor for each normalization process was kept at 10,000 for each data set. Seurat's built-in 'merge' function was used to combine the data sets after they were normalized individually. The resulting data set was a normalized combination of the two original data sets. Expression matrix was filtered by selecting cells that have a minimum of two expressed genes and genes that are expressed in more than one cell. The normalized expressions were used to cluster cells by a graph-based clustering approach in the Seurat package. The identified clusters were assigned with known Arabidopsis root cell types by computing index of cell identity (ICI) scores (51). Next, we used clusters identified as endodermis, cortex and quiescent center (QC) cells to compute fold change of each gene. DEsingle package (52) was used to identify differentially expressed genes between these three cell types by pairwise comparisons.

Evaluation data set A and evaluation data set B

Details of bulk RNA-seq data processing, DAP-seq data and ATAC-seq data processing are provided in the supplementary text. We constructed different evaluation data sets. The reason for using different evaluation data sets is to provide sufficient positive/negative training genes for machine learning and feature selection methods. For all expression experiments, we selected differential contrasts (evaluation data set A) which can provide >500 positive and 500 negative genes for all three types of negative genes (NDEGs, LEGs, UDGs, see supplementary table S2). After we determined that UDGs are the best negative training sets, we selected differential contrasts (evaluation data set B) which provide more than 500 positive and 500 negative genes for only UDGs. This data set was then used to evaluate the performance of integrating ATAC-seq data and the performance of different types of DAP-seq data.

Feature construction

Based on expression data, we constructed differential contrasts between replicate groups of control and treatment samples. Each replicate group typically includes expression data from multiple samples and each differential contrast produces a list of genes with fold change, mean expression value and FDR adjusted *P*-values for differential expression. Supplementary table S2 provides more details regarding the replicate group for each sample, and treatment and control information for each differential contrast. Next, we generated a feature matrix for each differential contrast by two steps. First, for each differential contrast, we generated a list of DEGs as positive training samples and sampled equal number of negative samples from the genome. The feature matrix X is a n by m matrix where n is the sum of number of positive samples and negative samples, and m is the number of TFs. In the second step, information from expression data, DAP-seq data and ATAC-seq data were integrated to construct X. Each entry X_{ij} in the feature matrix is computed by the following equations:

$$X_{ij} = F_j w\left(i, j\right) \tag{1}$$

$$w\ \left(i, j\right) = \sum_p \sum_q \frac{len\left(O\left(D_{ijp},\ A_{iq}\right)\right) \cdot S_{iq}^{(A)} \cdot S_{ijp}^{(D)}}{len\left(D_{ijp}\right)} \tag{2}$$

where j denotes jth TF and i denotes ith gene (either positive or negative gene). F_j is the $\log_2$ fold change value of TF j. In equation (1), $w(i, j)$ is the weight for each X_{ij}. In equation (2), D_{ijp}denotes the pth DAP-seq peak region of TF j found in the promoter region of gene i. We evaluated each DAP-seq peak region D_{ijp} by information from ATAC-seq, which was done by searching overlapping regions between each DAP-seq peak on promoter of gene i and all open chromatin regions on promoter of gene i. For qth overlapped ATAC-seq region (denoted by A_{iq}) on promoter of gene i, its importance is weighted by both ATAC-seq peak signal score (denoted by $S_{iq}^{(A)}$) and DAP-seq peak signal score (denoted by $S_{ijp}^{(D)}$). When peak signal scores were not used in the model, the default setting is $S_{iq}^{(A)} = 1$ and $S_{ijp}^{(D)} = 1$. This integration method will give higher weight $w(i, j)$, if DAP-seq peaks for a TF j have more overlapping regions with the open chromatin regions found on promoter gene i. The weight, $w(i, j)$, equals zero if (1) no DAP-seq peaks of TF j can be found on promoter of gene i, (2) no ATAC-seq peaks can be found on promoter of gene i, or (3) no overlapping regions were detected between them.

To efficiently search for all overlaps, we constructed an interval tree for ATAC-seq peaks in each chromosome then iterated over each DAP-seq peak to find all overlaps between DAP-seq peak and ATAC-seq peaks. Python package Intervaltree (https://github.com/chaimleib/intervaltree) was used to perform the search. While our current analysis only explored the use of DAP-seq interaction data and ATAC-seq open chromatin region data, other types of interaction data and chromatin feature data can be easily integrated into Equations (1) and (2). We will leave this to future exploration.

To construct feature matrices with only DAP-seq data, we marked each entry X_{ij} by '1' if binding site(s) of TF j are found in promoter region of gene i and '0' if not. We normalized the feature matrices by min-max normalization. Each X_{ij} was normalized by:

$$X'_{ij} = \frac{X_{ij} - \min\left(|X|\right)}{\max\left(|X|\right) - \min\left(|X|\right)}$$

where min(|*X*|) is the smallest absolute value in feature matrix X and max(|*X*|) is the largest absolute value in feature matrix X. During cross-validation, we computed min(|*X*|) and max(|*X*|) from training feature matrix and used them to normalize validation feature matrix and testing feature matrix.

Machine learning models and feature selection

We tested several machine learning methods for classification, including logistic regression (LR), support vector machine (SVM), random forest (RF) and deep neural network (DNN). To perform feature selection, we applied different regularization techniques to each classifier. The details of classification for LRLASSO and DNN and feature selection methods are described below. Other methods are described in the supplementary text.

LRLASSO. This method is logistic regression with lasso penalty, which uses L1-regularization for feature selection (53). LRLASSO minimizes the following loss function:

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} -L\left(y_i,\ \widehat{y_i}\right) + \lambda \sum_{j=1}^{m} \left|\beta_j\right|$$

where y_i and $\widehat{y_i}$ are the true label and predicted label for each training sample, respectively. $\widehat{y_i}$ is estimated by the logistic function:

$$\widehat{y_i} = \frac{1}{1 + e^{\sum_j^m X_{ij}\beta_j}}$$

$L(y_i,\ \widehat{y_i})$ is the log likelihood function and $\lambda \sum_{j=1}^{m} |\beta_j|$ is the L1 penalty term. β_j is the coefficient for feature j

(In our analysis, TF j). $L(y_i, \widehat{y_i})$ is calculated by the cross-entropy loss function:

$$L\,(y_i,\ \widehat{y_i}) =\ y_i \log(\widehat{y_i}) + (1 - y_i)\log(1 - \widehat{y_i})$$

To perform feature selection, we tuned the L1 penalty parameter λ for this model using the R package gglasso (54). Given a sequence of ordered λ values, gglasso computes the solution for each λ iteratively. The computed solution for current λ will be used as the initial value for next λ in the sequence. For each round of cross-validation, we used a sequence of 100 λ values which ranged from min(λ) to max(λ) and were spaced evenly on a log scale. max(λ) is the smallest λ value that shrinks all coefficients to zero. And min(λ) = η * max(λ), where η is a factor specified by user. For more details, see documentation for gglasso (54) and online documentation of the R package (https://cran.r-project.org/web/packages/gglasso/gglasso.pdf). In this way, each λ generates a LRLASSO model by training on training data set and the model was evaluated using validation data set to determine which λ gave the best prediction accuracy. Then the λ and the model with best prediction accuracy was again evaluated by the test data set.

DNN. This method is deep neural network with L1 regularization for feature selection. The use of regularized DNN for genomic feature selection has been investigated in a previous publication (55). The authors added a one-to-one layer between the input layer and hidden layers. L1 and L2 regularization were applied to the one-to-one layer to select features. Due to the high computational cost of tuning hyperparameters of DNN, we chose to use only L1 regularization in the one-to-one layer and hidden layers. We used a similar DNN architecture as in the previous publication (55). In the input layer, there are 387 neurons and this number is equal to the number of input features (TFs). In the second layer (one-to-one layer), the same number of neurons are used, and each is connected to one neuron from the input layer. Then we added two hidden layers which have 32 and 16 neurons after the one-to-one layer. The first hidden layer is fully connected with one-to-one layer and second hidden layer is fully connected with the first hidden layer. The last layer is an output layer which only has one neuron. Batch normalization was applied to one-to-one layer and each hidden layer to accelerate the training process. See (55) for more details about using DNN to select features.

For hyperparameter tuning, we tuned L1 regularization parameter λ for DNN model. We used a sequence of 10 λ values which range from 10^{-6} to 10^3 and are evenly spaced on a log scale. Adam optimizer was used to train the DNN model and learning rate α was fixed as 0.1. We compiled and trained DNN model using Keras library (https://keras.io/) with CUDA GPU acceleration. Training, hyperparameter tuning and testing was performed in the same way as described in other methods.

Evaluation strategy

Evaluating different conditions. To evaluate the effect of selecting negative training samples, we tested three different methods: (i) non-significantly differentially expressed genes (NDEGs), which have P-value > 0.05; (ii) low-expressed genes (LEGs), which have mean expression values between 0 and 0.5; (iii) undetected genes (UDGs), which have mean expression values equal to zero. To evaluate the effect of promoter region length, we constructed feature matrices using three different promoter lengths, which are (a) 5 kb upstream of TSS to 1 kb downstream of TSS; (b) 3 kb upstream of TSS to 0.5 kb downstream of TSS and (c) 0.5 kb upstream of TSS to TSS. The promoter region length is passed as an input argument to the ChIPseeker package to search for corresponding genes for each DAP-seq peak. To evaluate the effect of regular DAP-seq peaks and merged DAP-seq peaks, we constructed the feature matrices using the DAP-seq peaks from regular DAP-seq (methylated DAP-seq peaks) and the DAP-seq peaks from merged DAP-seq peaks. The performance of two methods were then compared.

Cross-validation. For each feature matrix, we randomly split the matrix into three subsets: 60% for training, 20% for validation (hyperparameter tuning) and 20% for testing. We trained the machine learning models on training data set and found the optimal set of hyperparameters by evaluating the trained model on validation data set (Figure 1B). Then the final performance of model with optimal hyperparameters was evaluated using the test data set. We used AUC–ROC and AUC–PRC as the metrics for evaluation. This process was repeated five times for each feature matrix to obtain the mean and standard deviation of AUC–ROC and AUC–PRC.

Compare to enrichment-based method. We compared our methods to enrichment-based method. Similar to the approach used in TF2Network (41), we computed the statistical significance of enrichment for each individual TF by hypergeometric test. The probability mass function is defined as:

$$P\ (x\ =\ i) = \frac{\binom{N}{i}\binom{M-N}{n-i}}{\binom{M}{N}}$$

where each parameter is explained below:

i is the number of DEGs that have DAP-seq peak(s) of the current TF.
N is the total number of DEGs in the current differential contrast.
n is the total number of protein-coding genes that have DAP-seq peak(s) of the current TF.
M is the total number of protein-coding genes.

P-values for all TFs were then computed by hypergeometric test and corrected by Benjamini–Hochberg correction (56). We used the same training set of positive genes and negative genes to compare LRLASSO with enrichment-based method. For each condition, this training set is the same feature matrix we used to evaluate machine learning models

To calculate AUC–ROC value for the enrichment-based method, we first ranked all TFs by ascending order using corrected P-values. Then we iterated over the ranked list of

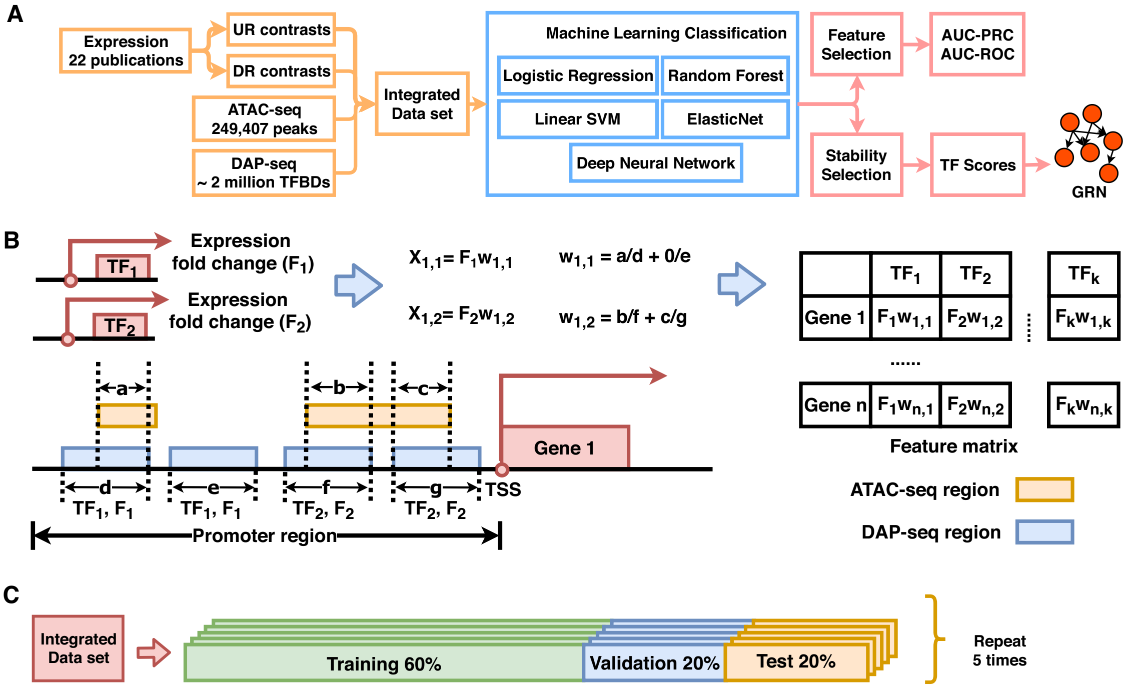


Figure 1. Flowchart of ConSReg pipeline. (**A**) Analysis workflow. (**B**) Genomic data integration strategy. DAP-seq and ATAC-seq regions were intersected and the weight for each intersected region was computed, and then summed up as the final weight for each TF–gene pair. The product of TF fold change and final weight is filled into corresponding entry of the feature matrix (see Materials and Methods for more details). parameters a, b, c, d, e, f, g are lengths of corresponding regions. (**C**) Cross-validation strategy. Final AUC–ROC values were computed from the 20% test data. We repeated this analysis five times for each integrated data set and calculated average and standard deviation of AUC–ROC values.

TFs. In each iteration, we used top kTFs as predictors and k is increased by one in next iteration until all 387 available TFs were included as predictors. Gene is considered as predicted positive if it has any predictor' peak regions in its promoter region and predicted negative if not. Therefore, false positive genes are those predicted as positive but are negative in the training set and false negative genes are those predicted as negative but are positive in the training set. We calculated false positive rate $\frac{FP}{N}$ and false negative rate $\frac{FN}{P}$ in each iteration and then all points of $(\frac{FP}{N}, \frac{FN}{P})$ were put together to construct ROC curve for computing AUC–ROC value.

Since hold-out test will not be applicable for enrichment-based method, for both LRLASSO and enrichment-based method, training and testing were performed using the same training set to have fair comparison.

Ranking TFs by stability selection

Since coefficients generated by LRLASSO model do not reflect the importance of each TF and the selected set of TFs would be slightly different when coefficients are initialized randomly, we applied stability selection (57) to generate robust feature selection result from LRLASSO.

Randomized lasso was proposed as an implementation of stability selection for the lasso method. The difference between randomized lasso and regular lasso is that subsampling of training samples and random perturbations for features are introduced into the feature selection process (57). Briefly, a subset of training samples were selected, and their features were randomly perturbed. Then a lasso model was trained using the perturbed subset of the original training data set. This process was then repeated multiple times. The idea is that important features will be selected more often than the unimportant ones during this randomized process. When used with LRLASSO, the objective function of randomized lasso can be written as:

$$\min_{\beta} -L(y_i,\ \widehat{y_i}) + \lambda \sum_{j=1}^{n} \frac{|\beta_j|}{w_j}$$

where $L(y_i,\ \widehat{y_i})$ is the log likelihood function as described previously in Materials and Methods. β_j is the coefficient for feature j. $\lambda \sum_{j=1}^{n} \frac{|\beta_j|}{w_j}$ can be considered as the penalty term for randomized lasso, similar to L1 penalty term for regular lasso model. The only difference here is that random perturbation is introduced by w_j, a scaling factor sampled from the range (0,1). For simplicity of implementation, features can be rescaled to have the same effect with rescaling the coefficients (57).

In our analysis, we randomly sampled half of the training samples from a feature matrix. Features were randomly perturbed by a scaling factor randomly sampled from (0,1). Randomized lasso was performed n times for each feature matrix. For each feature, the final importance score was calculated as number of times the feature gets non-zero coefficient divided by n. In our analysis, we set $n = 200$.

RESULTS

Analysis overview

In this work, we focused on using protein-DNA interaction data and open chromatin data to predict the combinations of TFs that can best explain observed differential gene expression under different environmental perturbations or cell types. To achieve this goal, we have tested multiple machine learning methods in combination with different feature selection techniques to determine the optimal parameters and training strategies. Our pipeline consists of two major steps (see Figure 1A). The first step is to integrate genomic data sets including interaction data generated from DAP-seq, open chromatin region data from ATAC-seq and expression data from RNA-seq/microarray experiments. This step produces training, validation, and testing data set for machine learning models. The second step is to perform binary classification with sparse feature selection methods. The input feature matrix for classification was constructed from binding site information and activated chromatin regions for a list of differentially expressed genes. These genes were obtained by standard statistical approaches (see Supplementary Text) using a contrast between a replicate group of treated samples and a replicate group of control samples (58).

For each gene, the feature matrix consists of all interactions between TFs and their target genes specified by DAP-seq experiments. Open chromatin regions were used to set a weight on the feature matrix (see Figure 1B). Our method can also incorporate peak heights from DAP-seq and ATAC-seq when calculating the weight of the feature matrix, but the improvement is small (Supplementary figure S1 and Supplementary Table S3). Up- and down-regulated genes were analyzed separately to train up-regulated (**UR**) and down-regulated (**DR**) models. Performance of machine learning models was evaluated by AUC–ROC and AUC–PRC computed from cross-validation. To prioritize important TFs for each condition, we assigned an importance score to each TF by performing stability selection (57). We performed multiple analyses to identify optimal settings for machine learning models. Our analyses include tests of (i) different machine learning approaches, (ii) different types of negative training samples, (iii) lengths of promoter region, (iv) combinations of data types and (v) difference between regular DAP-seq and amp-DAP-seq, where effects of DNA-methylation were removed by amplification.

Evaluation of different negative training samples and different machine learning approaches

The choice of negative training samples has been shown to significantly affect the performance of machine learning models (59). We evaluated three methods to select negative training samples: (i) non significantly differentially expressed genes (**NDEGs**), which have *P*-value > 0.05; (ii) low-expressed genes (**LEGs**), which have average expression between 0 and 0.5 FPKM and (iii) undetected genes (**UDGs**), which have a mean expression value equal to zero FPKM. The three methods were tested using evaluation data set A (see Materials and Methods), where we constructed both an up-regulated feature matrix and a down-regulated feature matrix for each differential contrast. Machine learning models tested in this analysis include logistic regression (LR), support vector machine (SVM), random forest (RF) and deep neural networks (DNN). See Methods and Supplementary Text for more details about the machine learning models.

Figure 2A shows a boxplot of all AUC-ROC values computed from the six machine learning approaches. These results show that AUC–ROC values of UDGs are significantly higher than NDEGs and LEGs (Wilcoxon signed-rank tests, *P*-values < 0.001). This suggests that machine learning classifiers perform better when UDGs are used as negative training samples. However, we did not find obvious differences for the number of selected TFs among the three types of negative training genes (embedded plot in Figure 2A). We further compared the performance of different machine learning approaches and found the six machine learning approaches achieved similar AUC–ROC values (Figure 2B, C for UR and DR respectively). However, the numbers of selected TFs obtained from different machine learning models are quite different (Figure 2B, C). LRLASSO consistently selected fewer TFs than other methods.

In recent years, the deep neural network (DNN) method has been extensively applied in the field of genomics to model gene regulation (60–62). We further explored whether DNN can bring better performance than LRLASSO. A previous study has introduced a DNN-based feature selection method (55). We used a similar strategy in our analysis (see Materials and Methods) to prioritize TFs and compare the result to LRLASSO. DNN usually needs large number of training samples to estimate model parameters. However, most of the expression data sets used in this study have fewer than 2000 genes available for training (see Supplementary table S2). Therefore, a comparison using multiple data sets in evaluation data set A or evaluation data set B (see Materials and Methods) resulted in an only poorly fitted DNN model. We selected a differential contrast which has the largest number of training samples (8948 genes for UR and DR feature matrices, respectively). For the UR feature matrix, the performance of LRLASSO is significantly better than DNN (Figure 2D, Wilcoxon rank-sum test, *P*-value < 0.01), whereas the performance does not show significant difference for DR feature matrix (Figure 2D, Wilcoxon rank-sum test, *P*-value > 0.05). At the same time, LRLASSO selected fewer TFs (embedded plot in Figure 2D) and has smaller variation in the number of selected TFs than DNN.

Although positive training genes in this study reflect condition-specific activities, it is unclear whether negative training genes are also condition specific. One possibility is that all negative training genes are not detected under any tested condition. We checked whether UDGs are different under different environmental perturbations. For each differential contrast in each environmental perturbation, we computed the percentage of UDGs that are detected (FPKM > 0) in other perturbations. Then the percentages were averaged for each environmental perturbation. We found that this average percentage ranges from 72.54% to 91.76%, suggesting that UDGs in one condition are typically expressed under other environmental perturbation(s). Therefore, a large portion of UDGs are inactive in one or

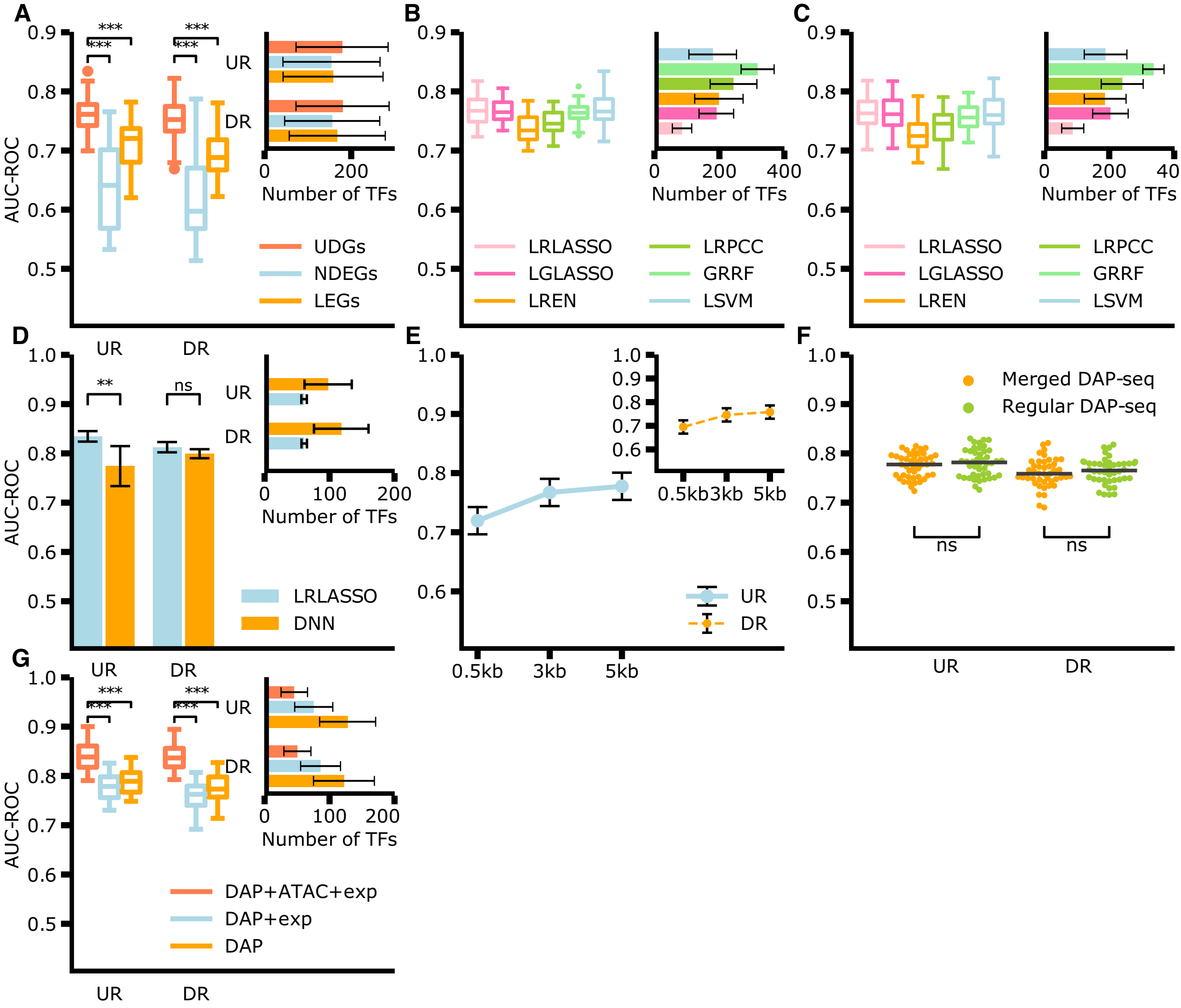


Figure 2. A comprehensive evaluation of model performance under different conditions. (**A**) Evaluation of different negative samples. UDGs: undetected genes, NDEGs: non-differentially expressed genes, LEGs: low-expression genes. Box plot demonstrates AUC–ROC for different negative samples (three boxes on the left: UR models, three boxes on the right: DR models). The embedded bar plot shows the number of TFs obtained using different negative data sets. (**B**, **C**) Evaluation of different classifiers. LRLASSO: logistic regression with LASSO penalty, LGLASSO: logistic group LASSO, LREN: logistic regression with an elastic net penalty. LRPCC: logistic regression with Pearson correlation coefficient, GRRF: Guided regularized random forest, LSVM: linear support vector machine. B shows the AUC–ROC values for UR model and C shows the AUC–ROC values for DR model. In both B and C, box plots show AUC-ROC values and embedded bar plots show the number of selected TFs. (**D**) Comparison between LRLASSO and DNN. (**E**) evaluation for different promoter region lengths. A curve in the major plot area shows AUC–ROC values for UR model and curve in the embedded plot area shows AUC-ROC values for DR model. (**F**) comparison between merged DAP-seq and regular DAP-seq. Medians were marked by black bars. (**G**) Evaluation of different integration strategies. The box plot shows a comparison of AUC–ROC values and an embedded bar plot shows a comparison of the number of selected TFs. **P*-value < 0.05; ***P*-value < 0.01; ****P*-value < 0.001; ns: not significant. *P*-value was computed from the Wilcoxon signed-rank test.

multiple specific environmental perturbations (Supplementary Figure S2).

Choice of promoter region length affects model performance

TFs regulate expressions of target genes by binding to regulatory elements located in the promoter regions of these genes. It has been shown that binding sites located within the 5 kb upstream region of transcription start sites (TSS) can better explain any regulatory effects on the target genes than shorter regions (41). To test the effect of promoter length on model performance, we set the promoter region length up to 5 kb upstream of TSS and 1 kb downstream of TSS in feature construction step. We tested three types of promoter regions: (i) 5 kb upstream of TSS to 1 kb downstream of TSS; (ii) 3 kb upstream of TSS to 0.5 kb downstream of TSS and (iii) 0.5 kb upstream of TSS. Figure 2E shows AUC–ROC values for three types of promoter regions evaluated on evaluation data set B (see Materials & Methods). We observed consistent improvements when promoter region length was extended from 0.5 to 3 kb upstream + 0.5 kb downstream. When the promoter region was further extended to 5 kb upstream + 1 kb downstream, no significant improvement was found. Additional promoter lengths were also tested, and the results are shown in Supplementary Table S3 and Supplementary figure S1. As shown in Figure 2E, these results are consistent between UR models and DR models.

In addition to the length of promoter regions, we also checked model performance when first intron sequences were included, because early molecular results suggest that first introns are important for regulating expression for some genes in plants (63,64). Our results do not show significant changes when first introns are included (Supplementary Figure S1 and Supplementary Table S3). For molecular validations of promoter functions, using intergenic sequences instead of fixed sequence length is very common (65,66). We tested our model performance by using intergenic sequences which have a specific length for each gene depending on the upstream gene location. We also did not see significant changes in the model performance (Supplementary Figure S1 and Supplementary Table S3). In summary, our findings suggest that most of the binding sites predictive of gene expressions were successfully captured within 3 kb upstream + 0.5 kb downstream region.

Types of DAP-seq experiments do not significantly affect model performance

As described previously (9), DAP-seq can be performed in two ways: (i) sequence regular genomic DNA (gDNA), (ii) sequence gDNA libraries in which methyl-cytosines were removed by PCR. The former is regular DAP-seq and the latter is called 'ampDAP-seq' (9). We tested the performance of two sets of binding sites: (a) using all available DAP-seq binding sites, which is the merged set of regular DAP-seq binding sites and ampDAP-seq binding sites (b) using only regular DAP-seq binding sites. It was reported that many DAP-seq binding sites (∼180 000) are occluded by DNA methylation, which is likely to affect the binding of TFs. However, our result shows that, compared to using regular DAP-seq binding sites, the merged set of DAP-seq binding sites does not provide better prediction result (Figure 2F). In particular, for reproductive tissues where gene expressions were known to be significantly impacted by DNA methylation (67), we do not see a significant difference in model performance using different types of DAP-seq data (Supplementary Table S3).

ATAC-seq data significantly improves model performance

Since all DAP-seq binding sites are detected *in vitro*, and some of the binding sites *in vitro* might not be accessible in living cells, it has been suggested that this limitation can be overcome by integrating DAP-seq data with open chromatin data (10,68). Therefore, we encoded open chromatin information from ATAC-seq data into the feature matrices (see Figure 1B and Materials and Methods). To assess the impact of chromatin accessibility, the feature matrices were constructed either with, or without, integrating ATAC-seq data. We then compared the model performance of ATAC-seq included feature matrices to ATAC-seq free feature matrices using evaluation data set B (see Materials and Methods). For both UR and DR genes, there are consistent improvements when ATAC-seq data were included in the feature matrices (Figure 2G). The other noticeable advantage of including ATAC-seq data is that fewer TFs were selected (embedded plot in Figure 2G). We further investigated whether including condition-specific expression and ATAC-seq data can better predict expression than using DAP-seq binding site information alone. Our results show that including all three types of data has consistently improved model performance (Figure 2G).

ConSReg outperforms a simple enrichment test

Enrichment tests have been applied in recent studies to identify candidate regulatory TFs given a set of input genes (41,42,69,70). We compared our prediction pipeline to a simple enrichment-test-based method (see Materials and Methods) and computed AUC–ROC values using evaluation data set B. As shown in Figure 3A, enrichment tests achieved an average AUC–ROC of 0.68 and 0.67 for UR and DR genes respectively. In contrast, ConSReg achieved average AUC–ROC of 0.84 and 0.84 for UR and DR genes respectively. AUC–ROC values for ConSReg are significantly higher than enrichment test (Wilcoxon rank-sum test, P-value < 0.001 for both UR and DR feature matrices).

ConSReg recovered TFs known to be involved in nitrogen response

Although ConSReg shows consistent better performance in AUC–ROC and AUC-PRC, it is unknown whether the TFs selected by ConSReg are key regulators of actual underlying biological processes. This is a challenging problem because there is no gold standard data set to evaluate such predictions. Traditional molecular genetic approaches typically involve the study of one or a few TFs at a time and, consequently, cannot rule out the possibility that other TFs are also involved in the same process. ChIP-seq or DAP-seq can only detect binding events but it is unclear whether any specific TF–target interaction is indeed actively regulating gene expression. To evaluate whether ConSReg can recall known TFs involved in a specific environmental perturbation, we applied ConSReg to a recently published study of TARGET (transient assay reporting genome-wide effects of transcription factors) which provides an ideal validation data set (47). This study used an updated TARGET assay (71) to evaluate how nitrogen (N) response TFs can impact the gene expressions of their target genes. In this study, 33 TFs were selected, and the TARGET system was designed such that genes were differentially expressed only due to the effect of each of these 33 TFs. We applied ConSReg to this RNA-seq data set and evaluated how many of the selected TFs can be recovered by ConSReg. Among 33 selected TFs, 17 TFs were also found in DAP-seq data. We therefore used these 17 TFs for our evaluation (see Figure 3D and Figure 3E for TF gene names). We re-analyzed the published TARGET data using DESeq2 (58) and generated differential contrasts (see Supplementary Table S2) as input for ConSReg.

We first compared importance scores of these TFs under two conditions: (i) cycloheximide (CHX) and N-treated

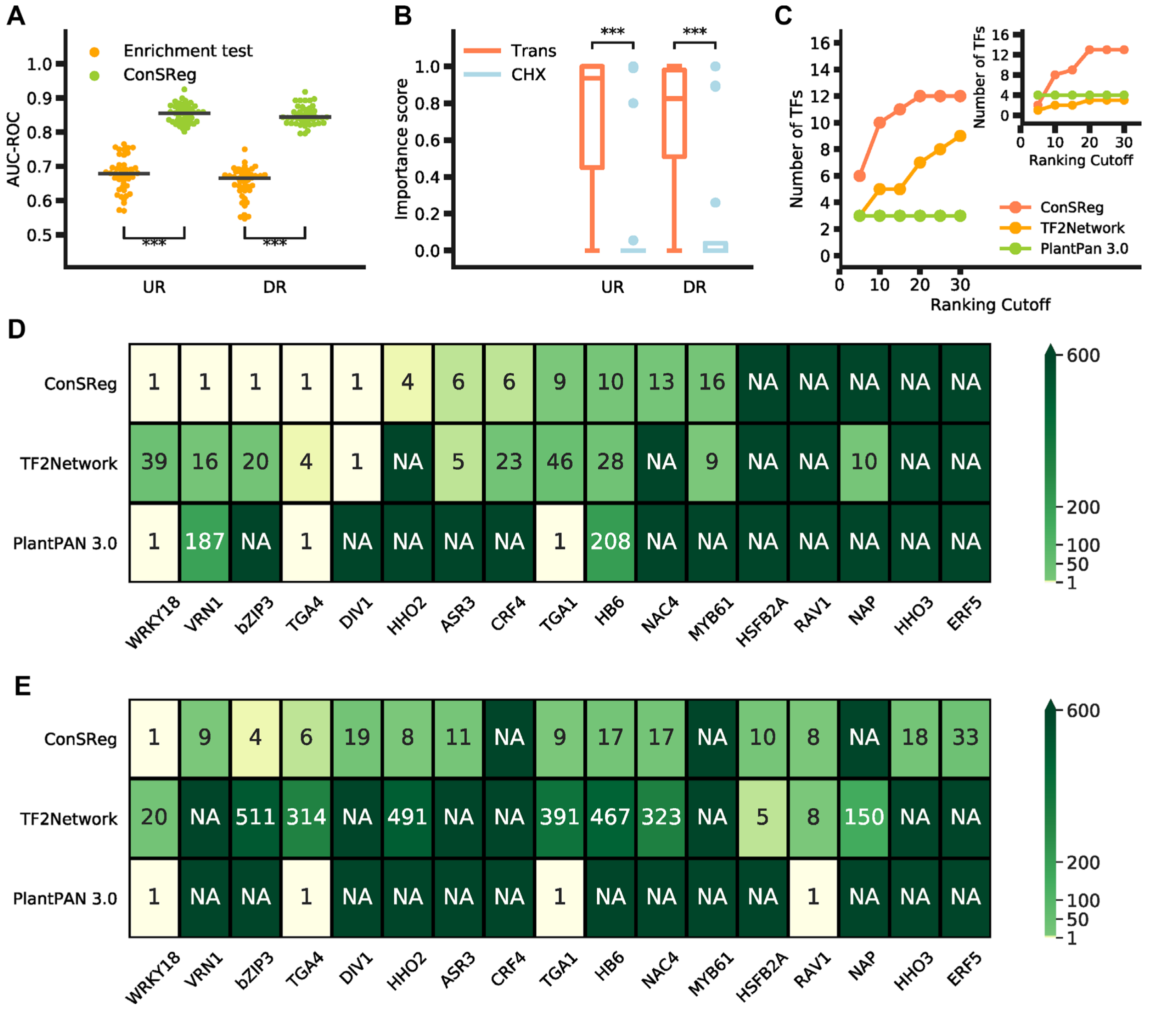


Figure 3. Comparison of different computational methods. (**A**) AUC-ROC for enrichment test and ConSReg. Two clusters on the left represent AUC-ROC values for UR models and two clusters on the right represent AUC–ROC values for DR models. (**B**) Importance scores of the 17 TFs for TF transfected root cells and CHX treated root cells. Two boxes on the left represent importance scores for UR models and two boxes on the right represent importance scores for DR models. (**C**) Number of recovered TFs for ConSReg, TF2Network, and PlantPAN 3.0 in different ranking cutoffs. Results predicted from UR models are plotted in the major plot area and results predicted from DR models are plotted in the embedded plot. (**D, E**) Ranking for each of the 17 nitrogen response TFs predicted by ConSReg, TF2Network, PlantPAN 3.0. Ranking for each TF was mapped to a color scale represented by a color bar on the right. A lighter color indicates better ranking. (D) The results predicted by UR model and (E) shows the results predicted by DR model. **P*-value < 0.05; ***P*-value < 0.01; ****P*-value < 0.001; ns: not significant. *P*-value was computed from the Wilcoxon signed-rank test.

TF transfected root cells VS empty vector transfected root cells and (ii) CHX and N treated VS N treated EV transfected root cells. CHX was used to block downstream regulation of secondary TF targets (47). For the first condition, we expected that DEGs are mainly direct targets for each of these TFs, whereas for the second condition, the DEGs are not induced by any of these TFs specifically. We generated DEGs (see Supplementary Table S2) and obtained importance scores for all 17 TFs from these DEGs in both conditions. For both UR and DR genes, the importance scores of these TFs from the first condition were significantly higher than the second condition (Figure 3B, Wilcoxon signed-rank test, *P*-value < 0.001 for both), suggesting that ConSReg can generate higher importance scores for the true regulatory TFs as compared to EV control experiments.

We then compared the result obtained using the 17 TFs to the result generated from TF2Network, and plantPAN 3.0, methods that can infer regulators for a given list of target genes. We set different cutoffs for ranking and counted how many TFs can be recovered at different ranking threshold (Figure 3C, Supplementary data file 1). As an example, when the ranking cutoff is set to the top 30 predicted TFs, ConSReg can recover 12/17 nitrogen-response TFs from UR models, which is better than the recovery rates of TF2Network (11/17), and PlantPAN 3.0 (9/17). For DR models, ConSReg was able to recover 14/17 nitrogen-response TFs from the top 30 predicted TFs, compared to the recovery rate of TF2Network (3/17), and PlantPAN3.0 (4/17). ConSReg provided better or had the same ranking for the correct TFs than other methods in 61.7% of all cases (10 TFs in UR models and 11 TFs in DR models). In contrast, TF2Network and PlantPAN 3.0 both provided better ranking for 20.6% of all cases. This result shows that ConSReg performs three times better than alternative methods in selecting regulatory genes in this testing data set.

As shown in Figure 3D and Figure 3E, there is a considerable overlap of TFs (10 TFs) between UR and DR models predicted by ConSReg and this number is higher than TF2Network (6 TFs) and PlantPAN 3.0 (3 TFs). This observation is consistent with the previously reported results that some of these TFs can act as both an inducer and a repressor of target genes (47). Detailed ranking results showed that for many recovered TFs, ConSReg assigned better rankings compared to other tools. For example, five recovered UR model TFs (WRKY18, VRN1, bZIP3, TGA4 and DIV1) were ranked as top 1 by ConSReg and these rankings are better than the other two tools (see Figure 3D). Notably, a few TFs predicted by PlantPAN 3.0 achieved ranking of top 1, while others predicted by PlantPAN 3.0 were assigned very low rankings (187 for VRN1, 208 for HB6, see Figure 3D). This is not surprising because many TFs predicted by PlantPAN 3.0 have identical support values. These TFs will therefore share identical ranking. For example, although WRKY18 was ranked as top 1 by PlantPAN 3.0 in UR models, there are 187 other TFs which were assigned the same ranking (see Supplementary data file 1). ConSReg only predicted one other TF that shared the same ranking as WRKY18. Compared to PlantPAN 3.0, this result is more specific.

Importance score can indicate predictive power of TF

To evaluate the predictive power of highly ranked TFs and verify whether these TFs can be more predictive of gene expressions than other TFs, we performed simulation of perturbation to TFs with high importance scores (importance score > 0.5). In this simulation, we compiled three sets of TFs in each differential contrast: (i) all TFs with importance scores > 0.5; (ii) replace the top five TFs in (i) using five lowest ranked TFs; and (iii) replace the top ten TFs in (i) using ten lowest ranked TFs. We evaluated the performance of the three sets of TFs using the same cross-validation strategy shown in Figure 1B. The results are shown in Figure 4. The reported AUC-ROC and AUC-PRC values for (i) are significantly higher than other two sets of TFs (Wilcoxon signed-rank test,). This can be observed clearly in Figure 4A and B, where performance of UR models was evaluated. A similar pattern was not apparent for DR feature matrices (Figure 4C and D), suggesting that DR regulatory processes are more difficult to be modeled than UR. Taken together, we concluded that for modeling UR genes, TFs with higher importance scores can be more predictive of gene expressions.

Case studies of using ConSReg in selecting candidate TFs

With the improved performance of ConSReg, we demonstrate here several examples of how ConSReg can be used to generate hypotheses based on integration of genomic data. These hypotheses are novel, data-driven hypotheses that were not generated by the original publication of these regulatory genes, or by expression data or any type of regulatory genomic data alone.

Case 1. New hypothetical function of a stress related transcription factor ZAT10

We performed a comprehensive investigation of TFs that are active under multiple abiotic environmental perturbations, encompassing nine common environmental perturbations: cold, heat, drought, salt, wounding, osmotic stress, red light, blue light and high light. For each differential contrast under each environmental condition, we assigned an importance score to each TF by ConSReg. The highest importance score was selected as a representative score for each TF under each environmental condition. This analysis can identify many known stress regulated genes, for example, MYB and ERF protein families are known for regulating many abiotic stress responsive genes (72,73). In our top 20 candidates generated from UR feature matrices, five TFs from the MYB/MYB related family (AT1G18330, AT3G50060, AT1G49010, AT5G67300 and AT1G74650) and two TFs from the ERF family (AT2G31230 and AT4G16750) have been identified. Results of the analysis are provided in Supplementary Tables S4 and S5.

Among these top candidate genes, we found that ZAT10 (AT1G27730) was predicted to be a top candidate regulator for all abiotic perturbations tested. Although our prediction is solely based on integration of DAP-seq, ATAC-seq and RNA-seq data, the predicted role of ZAT10 has been supported by detailed molecular characterization. That is, ZAT10 was reported to be involved in high light (74,75), heat (76), cold (77), dehydration (76,77) and salt responses (78). However, among numerous studies of ZAT10, no study reported the effect of ZAT10 in blue-light or red-light responses as predicted by ConSReg. Additional published molecular interactions suggest that ZAT10 might be indeed a regulator of blue/red light responses. For example, a previous study identified ZAT10 as the substrate of Mitogen-Activated Protein Kinase (MAPK) and showed that ZAT10 can directly interact with two MAPKs: MPK3 and MPK6 (79). It has been reported that MPK3 and MPK6 can be activated by blue light (80) or red light (81,82) in plants. However, the regulatory mechanism involved has not been well characterized. Therefore, we further investigated whether ZAT10 is related to MPK3 and MPK6 under blue light treatment. We computed PCC to quantify co-expressions of ZAT10 with a gene that encodes MPK3 protein (AT3G45640), and with another gene encoding MPK6 protein (AT2G43790). Significance of co-expression was computed by Fisher's Z-transformation as described in (83). Expression data used were from a GSE data set (GSE59699) generated under blue light treatment. Our result shows that ZAT10 has exhibited a significantly high co-expression with MPK3 (PCC = 0.943, P-value = 4.232×10^{-12}). However, ZAT10 was not significantly co-expressed with MPK6 (PCC = 0.190, P-value = 0.228). Given the evidences above, we hypothesize that ZAT10 is a candidate transcription factor that affects the blue light response by interacting with MPK3.

Case 2, ConSReg uncovers combinatorial regulations

TFs are known to modulate expression of target genes by combinatorial regulation in plants (84,85) through forming

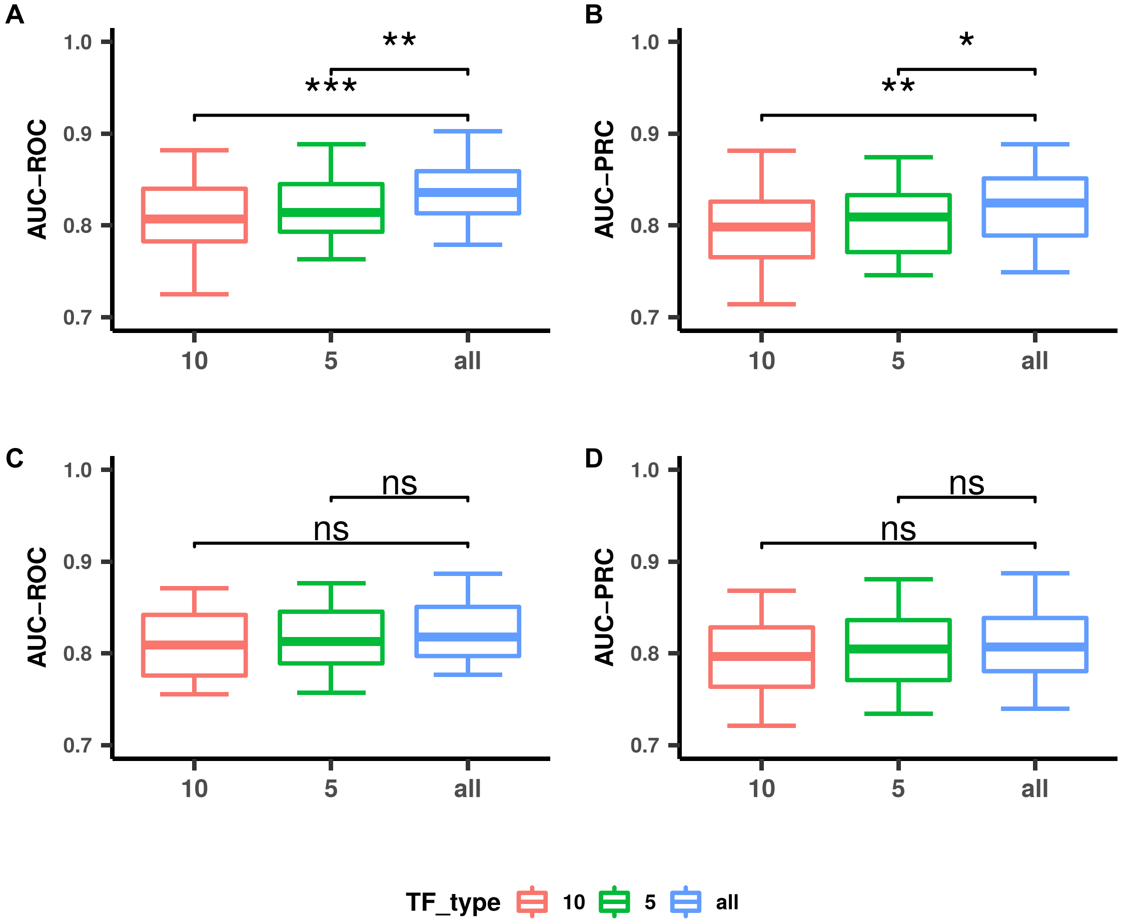


Figure 4. Simulation of perturbation for TFs. We performed simulation to perturb TFs with high importance scores (importance score > 0.5). Results shown here were generated from evaluation data set B. For each differential contrast in evaluation data set B, we used TFs with importance scores > 0.5 to construct three sets of TFs for testing: (i) all TFs with importance scores > 0.5 (marked by 'all' in the figure); (ii) replace the top five TFs in 1) using five lowest ranked TFs (marked by '5' in the figure); (iii) replace the top ten TFs in (i) using ten lowest ranked TFs (marked by '10' in the figure). (**A, B**) AUC–ROC and AUC–PRC for UR models. (**C, D**) AUC–ROC and AUC–PRC for DR feature matrices. Significance level was marked by stars over the boxes. *P-value < 0.05; **P-value < 0.01; ***P-value < 0.001; ns: not significant. P-value was computed from Wilcoxon signed-rank test.

protein complexes between TFs, or indirect interactions between TFs (85). ConSReg was used to identify TFs with a high importance score (> 0.5) for three environmental perturbations: cold, heat, drought which are known to regulate similar sets of genes. Sub-networks of TFs for cold, heat and drought were clustered using a simulation-annealing-based algorithm (86), and the results were visualized (Supplementary Figures S3, S4 and Supplementary Tables S6, S7).

We identified co-regulating modules of TFs from each GRN using a previously published tool (87), to identify common co-regulating TFs across the three conditions. While no common co-regulating TFs can be found for UR GRNs, we found a pair of co-regulating TFs for DR GRNs: MYB77 (AT3G50060) and MYB44 (AT5G67300) (Figure 5). Despite the fact that the two TFs are not differentially expressed under the three abiotic stresses tested (cold, heat, drought), they both have high importance scores under other abiotic stresses (see Supplementary table S5). Several published molecular and genetic studies support the hypothesis that these two transcription factors are regulated by the stress hormone ABA and that they regulate auxin responsive genes (88–91). These results support the predictions made by ConSReg. Although this finding was not observed based solely on any single type of data, the importance score generated by ConSReg was able to provide insight into putative regulatory roles for MYB77 and MYB44. Taken together, we conclude that combinatorial regulation between MYB77 and MYB44 confers abiotic stress tolerance to plants.

Case 3, Identification of regulatory genes in root using single cell gene expression

Single cell RNA-seq (scRNA-seq) is an emerging technology which has been successfully applied to characterize gene expression in Arabidopsis roots (48,49,92–94). To use scRNA-seq data to generate new discoveries of regulatory genes in plants, we applied ConSReg to two published scRNA-seq data sets (GSE122687 and GSE123013) of Arabidopsis roots (48,49). Cell types were identified by index of cell identity (ICI) scores as described in a previous publication (51). For simplicity, we focused on three cell types in

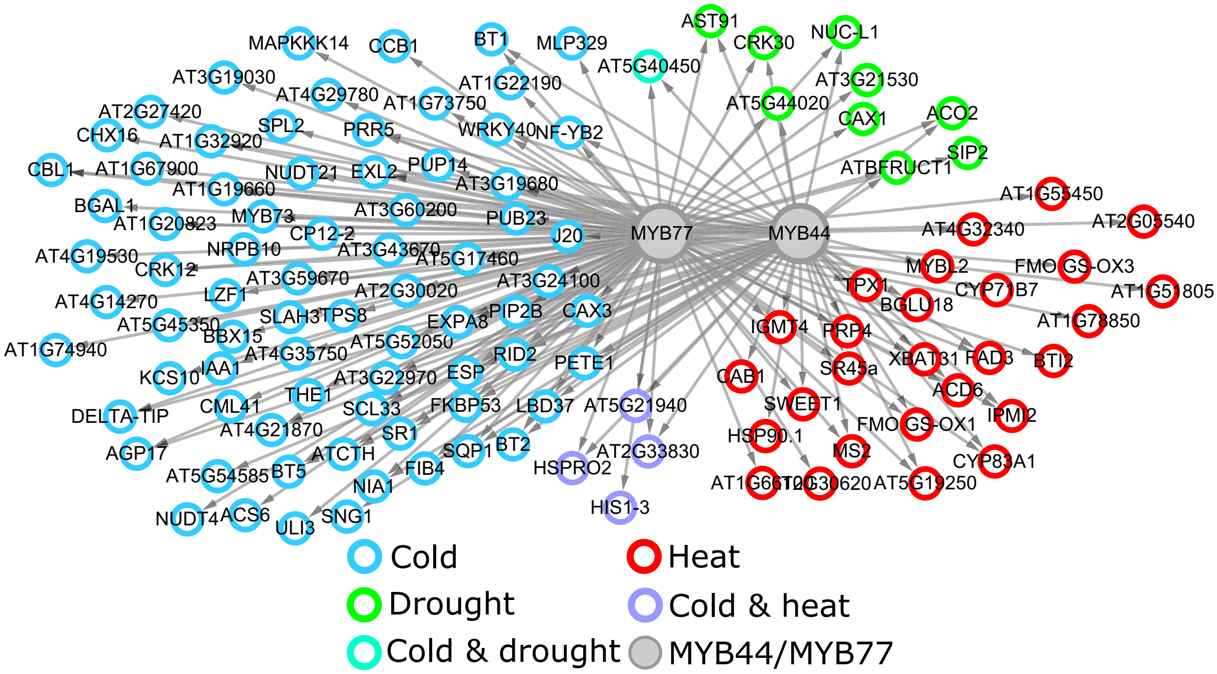


Figure 5. Combinatorial regulation between MYB44 and MYB77. Plotted in the center are MYB44 and MYB77 which regulate many common target genes under different abiotic stresses. The 20 top DEGs in each differential contrast were selected to be plotted in the figure. Edge list of this network can be found in Supplementary Table S9.

this study: endodermis, cortex and quiescent center (QC). UR and DR feature matrices were generated by comparing cortex cells and endodermis cells to QC cells respectively and importance scores for transcription factors were computed (see Materials and Methods for details).

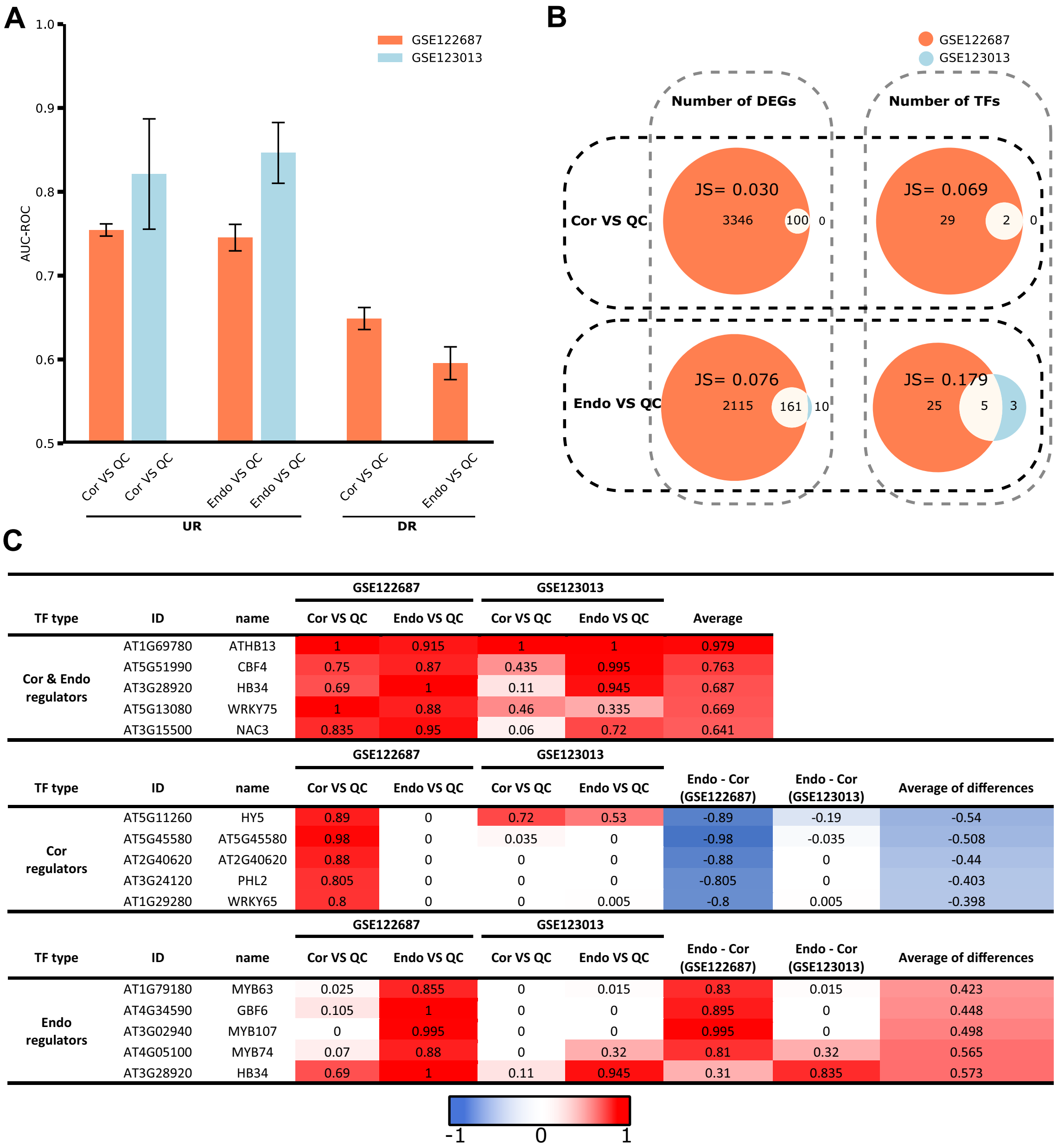
Among all the comparisons that we analyzed, the results from UR genes in endodermis versus QC and cortex versus QC (Figure 6A) provided the highest AUC–ROC and AUC–PRC values. We focus on these comparisons for the following analysis. ConSReg predicted more regulators for the data set of GSE122687 than the data set of GSE123013 regardless of the types of comparisons (Figure 6B). This is because our processing pipeline generated more DEGs from GSE122687 (4366 DEGs) than from GSE123013 (515 DEGs). The correlation of importance scores between GSE122687 and GSE123013 is 0.46. We found that predicted regulators have higher similarity as measured by Jaccard similarity (JS) than the differentially expressed genes. We found one gene (AT1G69780, ATHB13) that is consistently predicted as a regulator in both endodermis and cortex (Figure 6C). This gene is a known negative regulator of primary root length, suggesting a role in both endodermis and cortex cells (95). We further examined functions for predicted regulators only in one cell type but not in the other. We found MYB107 and MYB63, two genes that are known to be regulators of secondary cell wall formation (96), in particular, suberin biosynthesis (97,98). Interestingly, both genes are only predicted as regulators in endodermis. This is highly consistent with the biological function of endodermis where a water non-permeable layer is developed to limit water flow in and out of vascular tissue (99–102). In summary, ConSReg has led to the discovery of key regulatory genes that perform cell-type-specific function using single cell RNA-seq data.

DISCUSSION

Choice of negative training data and interpretation of TFs selected by the model

We tested three types of background gene sets: (i) NDEGs, which are non-significantly differentially expressed genes; (ii) LEGs, which are low-expressed genes; and (iii) UDGs, which are undetected genes. We showed that the average model performance using UDGs as negative training data is better than using NDEGs and LEGs. One possible explanation for lower performance of NDEGs is that some genes with *P* values that are marginally smaller than 0.05 are included in NDEGs. Similarly, LEGs are lowly expressed, but some LEGs could have fold changes close to significant DEGs. The binding sites in these gene groups could compromise the performance of the model when used as a negative training set.

For all the machine learning methods that we tested in this work, we also included a feature selection step (Figure 2B and C), where we found that LRLASSO consistently selected the fewest features. We chose LRLASSO for downstream analysis because it can provide fewer candidate genes for biological validation. Using fewer explanatory variables also aligns with other well-established model selection methods (103,104). However, caution should be taken in specific biological situations. For example, if two TFs are highly homologous and have redundant functions, feature selection methods may select only one with a slightly higher performance. In this case, inspecting the raw data for



TF type	ID	name	GSE122687 Cor VS QC	GSE122687 Endo VS QC	GSE123013 Cor VS QC	GSE123013 Endo VS QC	Average
Cor & Endo regulators	AT1G69780	ATHB13	1	0.915	1	1	0.979
	AT5G51990	CBF4	0.75	0.87	0.435	0.995	0.763
	AT3G28920	HB34	0.69	1	0.11	0.945	0.687
	AT5G13080	WRKY75	1	0.88	0.46	0.335	0.669
	AT3G15500	NAC3	0.835	0.95	0.06	0.72	0.641

TF type	ID	name	GSE122687 Cor VS QC	GSE122687 Endo VS QC	GSE123013 Cor VS QC	GSE123013 Endo VS QC	Endo - Cor (GSE122687)	Endo - Cor (GSE123013)	Average of differences
Cor regulators	AT5G11260	HY5	0.89	0	0.72	0.53	-0.89	-0.19	-0.54
	AT5G45580	AT5G45580	0.98	0	0.035	0	-0.98	-0.035	-0.508
	AT2G40620	AT2G40620	0.88	0	0	0	-0.88	0	-0.44
	AT3G24120	PHL2	0.805	0	0	0	-0.805	0	-0.403
	AT1G29280	WRKY65	0.8	0	0	0.005	-0.8	0.005	-0.398

TF type	ID	name	GSE122687 Cor VS QC	GSE122687 Endo VS QC	GSE123013 Cor VS QC	GSE123013 Endo VS QC	Endo - Cor (GSE122687)	Endo - Cor (GSE123013)	Average of differences
Endo regulators	AT1G79180	MYB63	0.025	0.855	0	0.015	0.83	0.015	0.423
	AT4G34590	GBF6	0.105	1	0	0	0.895	0	0.448
	AT3G02940	MYB107	0	0.995	0	0	0.995	0	0.498
	AT4G05100	MYB74	0.07	0.88	0	0.32	0.81	0.32	0.565
	AT3G28920	HB34	0.69	1	0.11	0.945	0.31	0.835	0.573

Figure 6. Analysis for two single cell data sets using ConSReg. Cell types for these cells were mapped by ICI. Specifically, we focused on the analysis of expression in endodermis (Endo) and cortex (Cor) cells. Differential contrasts were generated by comparing Endo cells to quiescent center (QC) cells, and Cor cells to QC cells. (**A**) AUC-ROC values for different comparisons in two data sets. DR results for GSE123013 were missing because there were no DR negative training genes that satisfied the selection criterion ($0 <$ mean FPKM <0.5, $-0.5 < \log_2$ fold change < 0). (**B**) Venn diagram for the overlap of DEGs and overlap of selected TFs. JS stands for Jaccard similarity score. (**C**) Predicted regulatory TFs ranked using importance score. Top table: regulators common for Endo and Cor; Middle table: regulators specific to Cor; Bottom table: regulators specific to Endo. Only results from UR models were shown here.

these similar TFs can be used to check the similarity of the binding profiles of these TFs for a set of DEGs.

Apply ConSReg to other species and Arabidopsis single cell expression data

The ConSReg pipeline is very flexible and uses standard input data formats. To demonstrate the ability of using ConSReg in a different species, we applied ConSReg to recently published DAP-seq, ATAC-seq and RNA-seq data from maize (15). The maize genome is 2.3Gb and is substantially larger than Arabidopsis. We performed ConSReg analysis using promoter length as long as 100Kb. The resulting average AUC-ROC is lower than that of Arabidopsis (Supplementary Table S8), and is likely due to the smaller number (32 TFs) of DAP-seq data that are available for maize as compared to Arabidopsis (387 TFs). This will change in the near future when more DAP-seq or other protein-DNA interaction data become available for maize. Next, we explored the application of ConSReg to single cell expression data. Recent advances in single cell sequencing technology has enabled the investigation of gene expression in individual cells in plants. We have demonstrated that ConSReg can identify transcriptional regulators using single cell data to define cell type-specific functional TFs. Our results also showed that the predicted regulators are more similar between different data sets than the differentially expressed genes (Figure 6B). It is expected that experimental noise and technical variations may lead to identification of DEGs that are not due to biological signals for single cell data. For example, a large portion of zero read counts may arise from technical noise or biological variability between single cells (105). Our results suggest that using predicted regulators may provide a better interpretation of the single cell sequencing results through identifying common regulatory genes. In previous work, attempts were made to address stochastic dropout by modeling it as a three component mixture model (106), two-component mixture linear model (107) or exponential function of expected expression (108). Our results suggest that dropout events might be compensated by incorporating regulatory network information into the model of single cell sequencing data.

Potential future improvement with condition specific data

While ConSReg achieved good performance (average ROC–AUC = 0.84), we think the results can be further improved by including data types that indicate dynamic regulation. Open chromatin regions have been reported to be both cell-type-specific (109,110) and condition-specific (111) as revealed by the distribution of DNaseI hypersensitive sites (DHSs). In our analysis, expression data and ATAC-seq data were not generated under the same conditions nor from the same tissue type. This is because data from roots and seedlings only are currently available for Arabidopsis (11). We merged all open chromatin regions detected in two tissue types to maximize the discovery of potential interactions. This could introduce false positives, which can be reduced by integrating open chromatin data and expression data generated under the same conditions and same tissue type. In our analysis, we did not include any quantitative proteomics data or protein activities due to post-translational modifications. A possible future improvement of ConSReg could also incorporate such information into feature matrices.

Additional possible functionalities

To better understand how condition- or cell-type-specific regulation changes across different condition or cell types, networks inferred by ConSReg can be compared. For example, when applied to single cell expression data, or bulk expression data with many time points, network comparisons can identify different regulation patterns that occur at different time points, resulting in inferences on how a given network dynamically changes over a time series. This will allow the capture of transient and dynamic regulatory mechanisms. For cell-type-specific expression data, an effective strategy might be to investigate the specificity of network module(s) for each cell type or a group of cell types. Modules that are shared by many cell types (112) may reveal fundamental pathways. Modules that are common to a limited number of cell types may play unique functional roles.

In summary, we have developed a novel computational tool, ConSReg. We have performed comprehensive analyses to identify the factors that affect the performance of machine learning models and the optimal settings for constructing a feature matrix. We have performed a systematic recovery of nitrogen-response TFs using ConSReg, TF2Network, and PlantPAN 3.0 and showed that ConSReg generated better ranking results and recovered more known nitrogen-responsive TFs compared to other computational tools. Network analysis for the GRNs inferred by ConSReg revealed new roles for ZAT10 in blue light regulation, and a novel combinatorial regulation between MYB44 and MYB77 in response to cold, heat and drought stresses. We applied ConSReg to Arabidopsis scRNA-seq data of root cell types and successfully identified cell type-specific regulators of cell wall formation which are supported by existing publications. In conclusion, ConSReg has the potential to transform any published gene expression data into condition-specific gene regulatory networks which will provide a system level overview of transcriptional regulation in plants.

DATA AVAILABILITY

ConSReg is implemented as an open source python package and is freely available at GitHub repository: https://github.com/LiLabAtVT/ConSReg.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Advanced Research Computing at Virginia Tech for providing the cluster computer system used in this research. We thank Dr Carol Huang for her input on how to generate peak scores and how to integrate peak scores into the machine learning model for DAP-seq data. Q.S. would like to thank the student support provided by graduate program in Genetic, Bioinformatics and Computational Biology (GBCB) program at Virginia Tech. Author contributions: Q.S., and S.L. conceived the original idea of the project. Q.S., J.L., S.A., M.R. analyzed the published RNAseq and single cell RNAseq data. Q.S. developed the code for machine learning algorithms. Q.S., M.R., R.G. and S.L. wrote the manuscript.

FUNDING

Jeffress Trust Awards Program in Interdisciplinary Research (to J.L., M.R., S.L.); United States Department of Energy Funding [DE-SC0020358 to S.L.]; Hatch Programs from United States Department of Agriculture (to S.L. and R.G.).

Conflict of interest statement. None declared.

REFERENCES

1. Krasensky,J. and Jonak,C. (2012) Drought, salt, and temperature stress-induced metabolic rearrangements and regulatory networks. *J. Exp. Bot.*, **63**, 1593–1608.
2. Golldack,D., Lüking,I. and Yang,O. (2011) Plant tolerance to drought and salinity: stress regulating transcription factors and their functional significance in the cellular transcriptional network. *Plant Cell Rep.*, **30**, 1383–1391.
3. Athar,A., Füllgrabe,A., George,N., Iqbal,H., Huerta,L., Ali,A., Snow,C., Fonseca,N.A., Petryszak,R., Papatheodorou,I. *et al.* (2019) ArrayExpress update - from bulk to single-cell expression data. *Nucleic Acids Res.*, **47**, D711–D715.
4. Song,L., Huang,S.C., Wise,A., Castanon,R., Nery,J.R., Chen,H., Watanabe,M., Thomas,J., Bar-Joseph,Z. and Ecker,J.R. (2016) A transcription factor hierarchy defines an environmental stress response network. *Science*, **354**, 598.
5. Franco-Zorrilla,J.M., López-Vidriero,I., Carrasco,J.L., Godoy,M., Vera,P. and Solano,R. (2014) DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 2367–2372.
6. Taylor-Teeples,M., Lin,L., de Lucas,M., Turco,G., Toal,T.W., Gaudinier,A., Young,N.F., Trabucco,G.M., Veling,M.T., Lamothe,R. *et al.* (2014) An Arabidopsis gene regulatory network for secondary cell wall synthesis. *Nature*, **517**:571–575.
7. Gaudinier,A., Rodriguez-Medina,J., Zhang,L., Olson,A., Liseron-Monfils,C., Bågman,A.-M., Foret,J., Abbitt,S., Tang,M., Li,B. *et al.* (2018) Transcriptional regulation of nitrogen-associated metabolism and growth. *Nature*, **563**:259–264.
8. Sparks,E.E.E., Drapek,C., Gaudinier,A., Li,S., Ansariola,M., Shen,N., Hennacy,J.H., Zhang,J., Turco,G., Petricka,J.J. *et al.* (2016) Establishment of expression in the shortroot-scarecrow transcriptional cascade through opposing activities of both activators and repressors. *Dev. Cell*, **39**, 585–596.
9. O'Malley,R.C., Huang,S., Song,L., Lewsey,M.G., Bartlett,A., Nery,J.R., Galli,M., Gallavotti,A. and Ecker,J.R. (2016) Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell*, **166**, 1598.
10. Bartlett,A., O'Malley,R.C., Huang,S.C., Galli,M., Nery,J.R., Gallavotti,A. and Ecker,J.R. (2017) Mapping genome-wide transcription-factor binding sites using DAP-seq. *Nat. Protoc*, **12**, 1659–1672.
11. Lu,Z., Hofmeister,B.T., Vollmers,C., DuBois,R.M. and Schmitz,R.J. (2016) Combining ATAC-seq with nuclei sorting for discovery of cis-regulatory regions in plant genomes. *Nucleic Acids Res.*, **45**, e41.
12. Maher,K.A., Bajic,M., Kajala,K., Reynoso,M., Pauluzzi,G., West,D.A., Zumstein,K., Woodhouse,M., Bubb,K., Dorrity,M.W. *et al.* (2018) Profiling of accessible chromatin regions across multiple plant species and cell types reveals common gene regulatory principles and new control modules. *Plant Cell*, **30**, 15–36.
13. Cumbie,J.S., Filichkin,S.A. and Megraw,M. (2015) Improved DNase-seq protocol facilitates high resolution mapping of DNase I hypersensitive sites in roots in Arabidopsis thaliana. *Plant Methods*, **11**, 42.
14. Zhang,W., Zhang,T., Wu,Y. and Jiang,J. (2012) Genome-Wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in arabidopsis. *Plant Cell*, **24**, 2719–2731.
15. Ricci,W.A., Lu,Z., Ji,L., Marand,A.P., Ethridge,C.L., Murphy,N.G., Noshay,J.M., Galli,M., Mejía-Guerra,M.K., Colomé-Tatché,M. *et al.* (2019) Widespread long-range cis-regulatory elements in the maize genome. *Nat. Plants*, **5**, 1237–1249.
16. Butte,A.J. and Kohane,I.S. (2000) Mutual information relevance networks:functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.*, **426**, 418–429.
17. Margolin,A.A., Nemenman,I., Basso,K., Wiggins,C., Stolovitzky,G., Favera,R. and Califano,A. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7**, S7.
18. Faith,J.J., Hayete,B., Thaden,J.T., Mogno,I., Wierzbowski,J., Cottarel,G., Kasif,S., Collins,J.J. and Gardner,T.S. (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.*, **5**, e8.
19. Meyer,P.E., Kontos,K., Lafitte,F. and Bontempi,G. (2007) Information-theoretic inference of large transcriptional regulatory networks. *EURASIP J. Bioinforma. Syst. Biol.*, **2007**, 79879.
20. Yuan,Y., Li,C.T. and Windram,O. (2011) Directed partial correlation: inferring large-scale gene regulatory network through induced topology disruptions. *PLoS One*, **6**, e16835.
21. Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.
22. Redekar,N., Pilot,G., Raboy,V., Li,S. and Saghai Maroof,M.A. (2017) Inference of transcription regulatory network in low phytic acid soybean seeds. *Front. Plant Sci.*, **8**, 2029.
23. Mordelet,F. and Vert,J.P. (2008) SIRENE: supervised inference of regulatory networks. *Bioinformatics*, **24**, i76–82.
24. Ni,Y., Aghamirzaie,D., Elmarakeby,H., Collakova,E., Li,S., Grene,R. and Heath,L.S. (2016) A machine learning approach to predict gene regulatory networks in seed development in arabidopsis. *Front. Plant Sci.*, **7**, 1936.
25. Haury,A.-C., Mordelet,F., Vera-Licona,P. and Vert,J.-P. (2012) TIGRESS: trustful inference of gene regulation using stability selection. *BMC Syst. Biol.*, **6**, 145.
26. Liu,L.Z., Wu,F.X. and Zhang,W.J. (2014) A group LASSO-based method for robustly inferring gene regulatory networks from multiple time-course datasets. *BMC Syst. Biol.*, **8**, S1.
27. Omranian,N., Eloundou-Mbebi,J.M.O., Mueller-Roeber,B. and Nikoloski,Z. (2016) Gene regulatory network inference using fused LASSO on multiple data sets. *Sci Rep.*, **6**, 20533.
28. Altarawy,D., Eid,F.-E. and Heath,L.S. (2017) PEAK: integrating curated and noisy prior knowledge in gene regulatory network inference. *J. Comput. Biol.*, **24**, 863–873.
29. de Luis Balaguer,M.A., Fisher,A.P., Clark,N.M., Fernandez-Espinosa,M.G., Möller,B.K., Weijers,D., Lohmann,J.U., Williams,C., Lorenzo,O. and Sozzani,R. (2017) Predicting gene regulatory networks by combining spatial and temporal gene expression data in Arabidopsis root stem cells. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, E7632–E7640.
30. Desai,J.S., Sartor,R.C., Lawas,L.M., Jagadish,S.V.K. and Doherty,C.J. (2017) Improving gene regulatory network inference by incorporating rates of transcriptional changes. *Sci. Rep.*, **7**, 17244.
31. Varala,K., Marshall-Colón,A., Cirrone,J., Brooks,M.D., Pasquino,A.V., Léran,S., Mittal,S., Rock,T.M., Edwards,M.B., Kim,G.J. *et al.* (2018) Temporal transcriptional logic of dynamic regulatory networks underlying nitrogen signaling and use in plants. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, 6494–6499.
32. Jin,J., Tian,F., Yang,D.-C., Meng,Y.-Q., Kong,L., Luo,J. and Gao,G. (2016) PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.*, **45**, D1040–D1045.
33. Davuluri,R.V., Sun,H., Palaniswamy,S.K., Matthews,N., Molina,C., Kurtz,M. and Grotewold,E. (2003) AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis -regulatory elements and transcription factors. *BMC Bioinformatics*, **4**, 25.
34. Yilmaz,A., Nishiyama,M.Y., Fuentes,B.G., Souza,G.M., Janies,D., Gray,J. and Grotewold,E. (2009) GRASSIUS: a platform for comparative regulatory genomics across the grasses. *Plant Physiol.*, **149**, 171–180.
35. Chen,F., Li,B., Li,G., Charron,J.-B., Dai,M., Shi,X. and Deng,X.W. (2014) Arabidopsis phytochrome a directly targets numerous promoters for individualized modulation of genes in a wide range of pathways. *Plant Cell*, **26**, 1949–1966.
36. Chen,F., Li,B., Demone,J., Charron,J.-B., Shi,X. and Deng,X.W. (2014) Photoreceptor partner FHY1 has an independent role in gene modulation and plant development under far-red light. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 11888–11893.
37. Fan,M., Bai,M.-Y., Kim,J.-G., Wang,T., Oh,E., Chen,L., Park,C.H., Son,S.-H., Kim,S.-K., Mudgett,M.B. *et al.* (2014) The bHLH transcription factor HBI1 mediates the trade-off between growth and pathogen-associated molecular pattern-triggered immunity in arabidopsis. *Plant Cell*, **26**, 828–841.
38. Song,L., Huang,S.C., Wise,A., Castanon,R., Nery,J.R., Chen,H., Watanabe,M., Thomas,J., Bar-Joseph,Z. and Ecker,J.R. (2016) A transcription factor hierarchy defines an environmental stress response network. *Science*, **354**, aag1550.

39. Shani,E., Salehin,M., Zhang,Y., Sanchez,S.E., Doherty,C., Wang,R., Mangado,C.C., Song,L., Tal,I., Pisanty,O. *et al.* (2017) Plant stress tolerance requires auxin-sensitive Aux/IAA transcriptional repressors. *Curr. Biol.*, **27**, 437–444.
40. Liu,S., Kracher,B., Ziegler,J., Birkenbihl,R.P. and Somssich,I.E. (2015) Negative regulation of ABA signaling by WRKY33 is critical for Arabidopsis immunity towards Botrytis cinerea 2100. *Elife*, **4**, e07295.
41. Kulkarni,S.R., Vaneechoutte,D., Van de Velde,J., Vandepoele,K., Van de Velde,J. and Vandepoele,K. (2018) TF2Network: predicting transcription factor regulators and gene regulatory networks in Arabidopsis using publicly available binding site information. *Nucleic Acids Res.*, **46**, e31.
42. Austin,R.S., Hiu,S., Waese,J., Ierullo,M., Pasha,A., Wang,T.T., Fan,J., Foong,C., Breit,R., Desveaux,D. *et al.* (2016) New BAR tools for mining expression data and exploring Cis-elements in Arabidopsis thaliana. *Plant J.*, **88**, 490–504.
43. Chow,C.N., Lee,T.Y., Hung,Y.C., Li,G.Z., Tseng,K.C., Liu,Y.H., Kuo,P.L., Zheng,H.Q. and Chang,W.C. (2019) PlantPAN3.0: a new and updated resource for reconstructing transcriptional regulatory networks from ChIP-seq experiments in plants. *Nucleic Acids Res.*, **47**, D1155–D1163.
44. Sijacic,P., Bajic,M., McKinney,E.C., Meagher,R.B. and Deal,R.B. (2018) Changes in chromatin accessibility between Arabidopsis stem cells and mesophyll cells illuminate cell type-specific transcription factor networks. *Plant J.*, **94**, 215–231.
45. Tannenbaum,M., Sarusi-Portuguez,A., Krispil,R., Schwartz,M., Loza,O., Benichou,J.I.C. and Hakim,O. (2018) Regulatory chromatin landscape in Arabidopsis thaliana roots uncovered by coupling INTACT and ATAC-seq. *Plant Methods.*, **14**, 113.
46. Wang,D., Rendon,A., Ouwehand,W. and Wernisch,L. (2012) Transcription factor co-localization patterns affect human cell type-specific gene expression. *BMC Genomics*, **13**, 263.
47. Brooks,M.D., Cirrone,J., Pasquino,A.V., Alvarez,J.M., Swift,J., Mittal,S., Juang,C.L., Varala,K., Gutiérrez,R.A., Krouk,G. *et al.* (2019) Network Walking charts transcriptional dynamics of nitrogen signaling by integrating validated and predicted genome-wide interactions. *Nat. Commun.*, **10**, 1569.
48. Shulse,C.N., Cole,B.J., Ciobanu,D., Lin,J., Yoshinaga,Y., Gouran,M., Turco,G.M., Zhu,Y., O'Malley,R.C., Brady,S.M. *et al.* (2019) High-throughput single-cell transcriptome profiling of plant cell types. *Cell Rep.*, **27**, 2241–2247.
49. Ryu,K.H., Huang,L., Kang,H.M. and Schiefelbein,J. (2019) Single-cell RNA sequencing resolves molecular relationships among individual plant cells. *Plant Physiol.*, **179**, 1444–1456.
50. Butler,A., Hoffman,P., Smibert,P., Papalexi,E. and Satija,R. (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, **36**, 411–420.
51. Efroni,I., Ip,P.L., Nawy,T., Mello,A. and Birnbaum,K.D. (2015) Quantification of cell identity from single-cell gene expression profiles. *Genome Biol.*, **16**, 9.
52. Miao,Z., Deng,K., Wang,X. and Zhang,X. (2018) DEsingle for detecting three types of differential expression in single-cell RNA-seq data. Berger B, editor. *Bioinformatics.*, **34**, 3223–3224.
53. Lee,S.S., Lee,H., Abbeel,P. and Ng,A.Y.A. (2006) Efficient L1 regularized logistic regression. Twenty-first natl. conf. artif. intell. eighteenth innov. In: Appl. Artif. Intell. Conf. Boston. pp. 401–408.
54. Yang,Y. and Zou,H. (2015) A fast unified algorithm for solving group-lasso penalize learning problems. *Stat. Comput.*, **25**, 1129–1141.
55. Li,Y., Chen,C.Y. and Wasserman,W.W. (2015) Deep feature selection: theory and application to identify enhancers and promoters. Lect. Notes Comput. Sci.(including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). Vol. **23**, pp. 205–217.
56. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B.*, **57**, 289–300.
57. Meinshausen,N. and Bühlmann,P. (2010) Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **72**, 417–473.
58. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
59. Natarajan,A., Yardimci,G.G., Sheffield,N.C., Crawford,G.E. and Ohler,U. (2012) Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Res.*, **22**, 1711–1722.
60. Alipanahi,B., Delong,A., Weirauch,M.T. and Frey,B.J. (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–838.
61. Zhou,J. and Troyanskaya,O.G. (2015) Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods*, **12**, 931–934.
62. Singh,R., Lanchantin,J., Robins,G. and Qi,Y. (2016) DeepChrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics*, **32**, i639–i648.
63. Vasil,V., Clancy,M., Ferl,R.J., Vasil,I.K. and Hannah,L.C. (1989) Increased gene expression by the first intron of maize shrunken-1 locus in grass species. *Plant Physiol.*, **91**, 1575–1579.
64. Rose,AB. (2002) Requirements for intron-mediated enhancement of gene expression in Arabidopsis. *RNA*, **8**, 1444–1453.
65. Lee,J.-Y., Colinas,J., Wang,J.Y., Mace,D., Ohler,U. and Benfey,P.N. (2006) Transcriptional and posttranscriptional regulation of transcription factor expression in Arabidopsis roots. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 6055–6060.
66. Wang,R.-S., Pandey,S., Li,S., Gookin,T.E., Zhao,Z., Albert,R. and Assmann,S.M. (2011) Common and unique elements of the ABA-regulated transcriptome of Arabidopsis guard cells. *BMC Genomics*, **12**, 216.
67. Loraine,A.E., McCormick,S., Estrada,A., Patel,K. and Qin,P. (2013) RNA-seq of Arabidopsis pollen uncovers novel transcription and alternative splicing. *Plant Physiol.*, **162**, 1092–1109.
68. Kulkarni,S.R., Jones,D.M. and Vandepoele,K. (2019) Enhanced maps of transcription factor binding sites improve regulatory networks learned from accessible chromatin data. *Plant Physiol.*, **181**, 412–425.
69. Chow,C.N., Zheng,H.Q., Wu,N.Y., Chien,C.H., Huang,H.-D., Lee,T.Y., Chiang-Hsieh,Y.-F., Hou,P.-F., Yang,T.-Y. and Chang,W.-C. (2016) PlantPAN 2.0: An update of Plant Promoter Analysis Navigator for reconstructing transcriptional regulatory networks in plants. *Nucleic Acids Res.*, **44**, D1154–D1164.
70. Reimand,J., Arak,T., Adler,P., Kolberg,L., Reisberg,S., Peterson,H. and Vilo,J. (2016) g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.*, **44**, W83–W89.
71. Bargmann,B.O.R.R., Marshall-Colon,A., Efroni,I., Ruffel,S., Birnbaum,K.D., Coruzzi,G.M. and Krouk,G. (2013) TARGET: a transient transformation system for genome-wide transcription factor target discovery. *Mol. Plant.*, **6**, 978–980.
72. Fujita,M., Fujita,Y., Noutoshi,Y., Takahashi,F., Narusaka,Y., Yamaguchi-Shinozaki,K. and Shinozaki,K. (2006) Crosstalk between abiotic and biotic stress responses: a current view from the points of convergence in the stress signaling networks. *Curr. Opin. Plant Biol.*, **9**, 436–442.
73. Müller,M. and Munné-Bosch,S. (2015) Ethylene response factors: a key regulatory hub in hormone and stress signaling. *Plant Physiol.*, **169**, 32–41.
74. Rossel,J.B., Wilson,P.B., Hussain,D., Woo,N.S., Gordon,M.J., Mewett,O.P., Howell,K.A., Whelan,J., Kazan,K. and Pogson,B.J. (2007) Systemic and intracellular responses to photooxidative stress in arabidopsis. *Plant Cell*, **19**, 4091–4110.
75. Gordon,M.J., Carmody,M., Albrecht,V. and Pogson,B. (2013) Systemic and local responses to repeated HL stress-induced retrograde signaling in arabidopsis. *Front. Plant Sci.*, **3**, 303.
76. Mittler,R., Kim,Y., Song,L., Coutu,J., Coutu,A., Ciftci-Yilmaz,S., Lee,H., Stevenson,B. and Zhu,J.-K. (2006) Gain- and loss-of-function mutations in Zat10 enhance the tolerance of plants to abiotic stress. *FEBS Lett.*, **580**, 6537–6542.
77. Sakamoto,H., Maruyama,K., Sakuma,Y., Meshi,T., Iwabuchi,M., Shinozaki,K. and Yamaguchi-Shinozaki,K. (2004) Arabidopsis Cys2/His2-type zinc-finger proteins function as transcription repressors under drought, cold, and high-salinity stress conditions. *Plant Physiol.*, **136**, 2734–2746.
78. Xie,Y., Mao,Y., Lai,D., Zhang,W. and Shen,W. (2012) H2 enhances arabidopsis salt tolerance by manipulating ZAT10/12-mediated antioxidant defence and controlling sodium exclusion. Blazquez MA, editor. *PLoS One.*, **7**, e49800.

79. Nguyen,X.C., Kim,S.H., Lee,K., Kim,K.E., Liu,X.-M.M., Han,H.J. *et al.* (2012) Identification of a C2H2-type zinc finger transcription factor (ZAT10) from Arabidopsis as a substrate of MAP kinase. *Plant Cell Rep.*, **31**, 737–745.
80. Sethi,V., Raghuram,B., Sinha,A.K. and Chattopadhyay,S. (2014) A mitogen-activated protein kinase cascade module, MKK3-MPK6 and MYC2, is involved in blue light-mediated seedling development in arabidopsis. *Plant Cell.*, **26**, 3343–3357.
81. Xin,X., Chen,W., Wang,B., Zhu,F., Li,Y., Yang,H., Li,J. and Ren,D. (2018) Arabidopsis MKK10-MPK6 mediates red-light-regulated opening of seedling cotyledons through phosphorylation of PIF3. *J. Exp. Bot.*, **69**, 423–439.
82. Zhao,Y., Zhou,J. and Xing,D. (2014) Phytochrome B-mediated activation of lipoxygenase modulates an excess red light-induced defence response in Arabidopsis. *J. Exp. Bot.*, **65**, 4907–4918.
83. Weirauch,M.T. (2011) Gene co-expression networks for the analysis of DNA microarray data. *Appl. Stat. Netw. Biol. Methods Syst. Biol.* **1**,215–250.
84. Singh,K.B. (1998) Transcriptional regulation in plants: the importance of combinatorial control. *Plant Physiol.*, **118**, 1111–1120.
85. Kaufmann,K., Pajoro,A. and Angenent,G.C. (2010) Regulation of transcription in plants: mechanisms controlling developmental switches. *Nat. Rev. Genet.*, **11**, 830–842.
86. Gerstein,M.B., Kundaje,A., Hariharan,M., Landt,S.G., Yan,K.-K., Cheng,C., Mu,X.J., Khurana,E., Rozowsky,J., Alexander,R. *et al.* (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature*, **489**, 91–100.
87. Song,Q., Grene,R., Heath,L.S. and Li,S. (2017) Identification of regulatory modules in genome scale transcription regulatory networks. *BMC Syst. Biol.*, **11**, 140.
88. Shin,R., Burch,A.Y., Huppert,K.A., Tiwari,S.B., Murphy,A.S., Guilfoyle,T.J. and Schachtman,DP. (2007) The arabidopsis transcription factor MYB77 modulates auxin signal transduction. *Plant Cell*, **19**, 2440–2453.
89. Xing,L., Zhao,Y., Gao,J., Xiang,C. and Zhu,J.-K. (2016) The ABA receptor PYL9 together with PYL8 plays an important role in regulating lateral root growth. *Sci. Rep.*, **6**, 27177.
90. Jaradat,M.R., Feurtado,J., Huang,D., Lu,Y. and Cutler,A.J. (2013) Multiple roles of the transcription factor AtMYBR1/AtMYB44 in ABA signaling, stress responses, and leaf senescence. *BMC Plant Biol.*, **13**, 192.
91. Zhao,Y., Xing,L., Wang,X., Hou,Y.-J., Gao,J., Wang,P., Duan,C.-G., Zhu,X. and Zhu,J.-K. (2014) The ABA receptor PYL8 promotes lateral root growth by enhancing MYB77-dependent transcription of auxin-responsive genes. *Sci. Signal.*, **7**, ra53.
92. Denyer,T., Ma,X., Klesen,S., Scacchi,E., Nieselt,K. and Timmermans,M.C.P.P. (2019) Spatiotemporal developmental trajectories in the arabidopsis root revealed using high-throughput single-cell RNA sequencing. *Dev. Cell*, **48**, 840–852.
93. Jean-Baptiste,K., McFaline-Figueroa,J.L., Alexandre,C.M., Dorrity,M.W., Saunders,L., Bubb,K.L., Trapnell,C., Fields,S., Queitsch,C. and Cuperus,J.T. (2019) Dynamics of gene expression in single root cells of arabidopsis thaliana. *Plant Cell*, **31**, 993–1011.
94. Zhang,T.-.Q.Q., Xu,Z.-.G.G., Shang,G.-.D.D. and Wang,J.-.W.W. (2019) A single-cell RNA sequencing profiles the developmental landscape of arabidopsis root. *Mol. Plant*, **12**, 648–660.
95. Silva,A.T., Ribone,P.A., Chan,R.L., Ligterink,W. and Hilhorst,H.W.M. (2016) A predictive coexpression network identifies novel genes controlling the seed-to-seedling phase transition in arabidopsis thaliana. *Plant Physiol.*, **170**, 2218–2231.
96. Zhou,J., Lee,C., Zhong,R. and Ye,Z.-.H. (2009) MYB58 and MYB63 are transcriptional activators of the lignin biosynthetic pathway during secondary cell wall formation in arabidopsis. *Plant Cell*, **21**, 248–266.
97. Gou,M., Hou,G., Yang,H., Zhang,X., Cai,Y., Kai,G. and Liu,C.-J. (2017) The MYB107 transcription factor positively regulates suberin biosynthesis. *Plant Physiol.*, **173**, 1045–1058.
98. Lashbrooke,J., Cohen,H., Levy-Samocha,D., Tzfadia,O., Panizel,I., Zeisler,V., Massalha,H., Stern,A., Trainotti,L., Schreiber,L. *et al.* (2016) MYB107 and MYB9 homologs regulate suberin deposition in angiosperms. *Plant Cell*, **28**, 2097–2116.
99. Thomas,R., Fang,X., Ranathunge,K., Anderson,T.R., Peterson,C.A. and Bernards,M.A. (2007) Soybean root suberin: anatomical distribution, chemical composition, and relationship to partial resistance to phytophthora sojae. *Plant Physiol.*, **144**, 299–311.
100. Barberon,M. (2017) The endodermis as a checkpoint for nutrients. *New Phytol.*, **213**, 1604–1610.
101. Robbins,N.E., Trontin,C., Duan,L. and Dinneny,J.R. (2014) Beyond the barrier: communication in the root through the endodermis. *Plant Physiol.*, **166**, 551–559.
102. Schreiber,L., Hartmann,K., Skrabs,M. and Zeier,J. (1999) Apoplastic barriers in roots: chemical composition of endodermal and hypodermal cell walls. *J. Exp. Bot.*, **50**, 1267–1280.
103. Kass,R.E. and Raftery,A.E. (1995) Bayes factors. *J. Am. Stat. Assoc*, **90**, 773–795.
104. Burnham,K.P. and Anderson,D.R. (2004) Multimodel inference. *Sociol. Methods Res.*, **33**, 261–304.
105. Vallejos,C.A., Risso,D., Scialdone,A., Dudoit,S. and Marioni,J.C. (2017) Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat. Methods*, **14**, 565–571.
106. Kharchenko,P.V., Silberstein,L. and Scadden,D.T. (2014) Bayesian approach to single-cell differential expression analysis. *Nat. Methods*, **11**, 740–742.
107. Finak,G., McDavid,A., Yajima,M., Deng,J., Gersuk,V., Shalek,A.K., Slichter,C.K., Miller,H.W., McElrath,M.J., Prlic,M. *et al.* (2015) MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.*, **16**, 278.
108. Pierson,E. and Yau,C. (2015) ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.*, **16**, 241.
109. Boyle,A.P., Davis,S., Shulha,H.P., Meltzer,P., Margulies,E.H., Weng,Z., Furey,T.S. and Crawford,G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322.
110. Thurman,R.E., Rynes,E., Humbert,R., Vierstra,J., Maurano,M.T., Haugen,E., Sheffield,N.C., Stergachis,A.B., Wang,H., Vernot,B. *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75–82.
111. Hesselberth,J.R., Chen,X., Zhang,Z., Sabo,P.J., Sandstrom,R., Reynolds,A.P., Thurman,R.E., Neph,S., Kuehn,M.S., Noble,W.S. *et al.* (2009) Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods*, **6**, 283–289.
112. Li,S., Pandey,S., Gookin,T.E., Zhao,Z., Wilson,L. and Assmann,S.M. (2012) Gene-sharing networks reveal organizing principles of transcriptomes in Arabidopsis and other multicellular organisms. *Plant Cell*, **24**, 1362–1378.