Improving Gene Regulatory Network Inference by Incorporating Rates of Transcriptional Changes

Jigar S. Desai[1], Ryan C. Sartor[2], Lovely Mae Lawas[3,4], SV Krishna Jagadish[5], Colleen J. Doherty[1]

[1]Department of Molecular and Structural Biochemistry, North Carolina State University, Raleigh, NC, United States of America, [2]Department of Biological Sciences, University of California, San Diego, La Jolla, CA, United States of America, [3]International Rice Research Institute DAPO Box 7777, Metro Manila, Philippines, [4] Central Infrastructure Group Genomics and Transcript Profiling, Max Planck Institute of Molecular Plant Physiology Potsdam, Germany, [5]Department of Agronomy, Kansas State University, Manhattan, KS, United States of America.

*To whom correspondence should be addressed.

Abstract

Motivation: Transcriptional regulatory networks (TRNs) are hierarchies of regulatory factors that control the expression levels of target genes. The signal transduction pathways of these networks are often determined by experimental analysis. Computational inference of the connections between regulators and target genes using transcriptional assays can identify high confidence candidate regulator-target relationships. Often, expression experiments designed for this purpose are performed in a time series. Most TRN identifying algorithms, however, do not take full advantage of that temporal data. We developed a new approach, ExRANGES, which utilizes both the rate of change in expression and the absolute expression level to identify TRN connections.

Results: Our novel strategy, ExRANGES improves the ability to computationally infer TRN from time series expression data by emphasizing the comparison between regulator and target at time points where there is a significant change in expression. ExRANGES combines the rate of change in expression with the absolute expression level and improves the ability to accurately identify known targets of transcriptional regulators. We evaluated ExRANGES in four large data sets from different model systems and in one sparse data set using two different network construction approaches. ExRANGES improved the identification of experimentally validated transcription factor targets for all species even in unevenly spaced and sparse data sets. This improved ability to predict known regulator-target relationships in model species enhances the utility of network inference approaches in non-model species where experimental validation is challenging.

Availability: ExRANGES has been implemented as an R package and is available http://github.com/DohertyLab/ExRANGES

To install the package type: devtools::install_github("DohertyLab/ExRANGES")

Contact: colleen_doherty@ncsu.edu

Supplementary information is available on-line.

1 Introduction

Transcriptional regulatory networks (TRN) provide a framework for understanding how signals propagate through a molecular network and result in transcriptomic changes. These regulatory networks are biological computational modules that carry out decision-making processes and, in many cases, determine the ultimate response of an organism to a stimulus (Balázsi *et al.* 2011). Understanding the regulatory networks that drive responses of an organism to the environment provide access points to modulate these responses through breeding or genetic modifications. The first step in constructing such networks is to identify the primary

relationships between regulators such as transcription factors (TFs) and the target genes they control.

Experimental approaches such as Chromatin Immunoprecipitation followed by sequencing (ChIP-Seq) can identify direct targets of transcriptional regulators. However, ChIP-Seq must be optimized to each specific TF and specific antibodies must be used that recognize either the native TF or a tagged version of the protein. This can present a technical challenge particularly for TFs where the tag interferes with function, for species that are not easily transformable, or for tissues that are limited in availability (Park 2009). Since global transcript levels are comparatively easy to measure in most species and tissues, several approaches have been developed to identify connections between regulators and their targets by examining the changes in transcription levels across many samples (Qian *et al.* 2003; Bonneau *et al.* 2006; Margolin *et al.* 2006; Faith *et al.* 2007). The assumption of these approaches is that the regulatory relationship between a regulator TF and its targets can be discerned from a correspondence between the RNA levels of the regulator gene and its targets. If this is true, then given sufficient variation in expression, the targets of a given factor can be predicted based on associated changes in expression. Initial approaches designed to do this focused on the correlation between regulators and targets, assuming that activators are positively correlated and repressors are negatively correlated with their target expression levels (Eisen *et al.* 1998). For almost two decades, these approaches successfully identified relationships between regulators and targets. Updates to this simple idea have included preclustering of transcript data, modifying regression analysis, incorporating training classifier models, and incorporating prior biological knowledge or additional experimental data. Each of these has improved the ability to identify connections between regulators and targets, even in sparse and noisy data sets (Bonneau *et al.* 2006; Margolin *et al.* 2006; Faith *et al.* 2007; Huynh-Thu *et al.* 2010; Li *et al.* 2011; Marbach *et al.* 2012; Wilkins *et al.* 2016). In 2010, the DREAM5 challenge evaluated the ability of different methods to identify TRN from gene expression datasets (Marbach *et al.* 2012). One of the top performing methods was GENIE3 (Huynh-Thu *et al.* 2010). This method uses the machine learning capabilities of random forest to identify targets for selected regulators (Breiman 2001; Liaw and Wiener 2002). Other successfully implemented approaches include SVM (Qian *et al.* 2003), CLR (Faith *et al.* 2007), CSI (Penfold *et al.* 2012, 2015), ARACNE (Margolin *et al.* 2006), Inferelator (Bonneau *et al.* 2006), and DELDBN (Li *et al.* 2011). Common to these methods is the use of transcript abundance levels to evaluate the relationship between a regulator and its putative targets. Experiments performed in time series can provide additional kinetic information useful for associating regulators and targets. Many approaches have been developed that take advantage of the additional information available from time series data (Reviewed in Bar-Joseph *et al.*, 2012; Thompson *et al.*, 2015). However, the steady state transcript level as measured by most high-throughput transcriptional assays such as RNA-Seq is a measure of both transcriptional activity and mRNA stability. Therefore, correlation between expression levels alone may not provide a direct assessment of transcriptional regulation as it can be confounded by the RNA stability of the target. Further complicating the identification of regulator relationships is the fact that a single gene can be regulated by different transcription factors in response to different stimuli.

Here we present an approach that extends current approaches to TRN construction by emphasizing the relationship between regulator and targets at the time points where there is a significant change in the rate of expression. We demonstrate that: 1) Focusing on the rate of change captured previously unrecognized characteristics in the data, identifying experimentally validated regulatory relationships not detected by the standard approaches. 2) Combining expression level and the rate of change resulted in improved identification of experimentally validated regulatory relationships.

We first developed a method, RANGES (RAte Normalized in a GEne Specific manner) that evaluates the significance of the rate changes at each consecutive time point on a per-gene basis. We then combined the expression level and significance of this rate change in ExRANGES (Expression by RANGES) to prioritize the correlation between regulators and targets at time points where there is significant change in gene expression. ExRANGES improved the ability to identify experimentally validated TF targets in microarray and RNA-Seq data sets across multiple experimental designs, and in several different species. We demonstrate that this approach improves the identification of experimentally validated TF targets for GENIE3 [8] and INFERELATOR [4], and anticipate that it will offer a similar benefit to when combined with other network inference algorithms.

2 Methods

2.1 Identifying consecutive time points with significant changes in expression

The first step of ExRANGES is to identify time points where active regulation is observed based on changes in RNA levels. Our method examines the change in expression between two consecutive time points on a per gene basis. Significance is determined by comparing these expression changes to the bootstrapped background variance observed across the dataset for that gene (Supplemental Figure 1).

For each $gene_i$, the background variance is derived from the change in expression of that gene at all consecutive time steps in all samples across all experiments from a given data set. The change in expression between two consecutive time points is evaluated against this background. For example, if we consider the mammalian circadian data set available from CircaDB (Pizarro *et al.* 2013), the data set consists of time series experiments from 12 different tissues, sampled every 2 h for 48 h (288 samples). The change in expression levels between time t and time t +1 was determined for each consecutive time point. Since this data is cyclical, the interval between the last time point and the first time point is also included. For the CircaDB data set, the background of each consecutive time interval across the entire time series consists of 288 slopes (12 tissues x 2 h for 48). At each time step, t the slope between t and t + 1 was compared to a bootstrapped version of this background generated by sampling 10,000 times with replacement. For each gene the resulting p-value, calculated by using an empirical cumulative distribution function from the R stats package. This p-value was transformed to the $-\log_{10}$ and the sign of the change in slope was preserved (R script provided). This significance of the change at each time interval is the rate change, or "RANGES" value.

2.2 Combining EXPRESSION and Rate Change using ExRANGES

RANGES identifies time points that show significant changes in expression for a given gene. ExRANGES adjusts the expression level at each time point by the rate of change in the following time interval. In doing this, the ExRANGES value of the time point preceding a significant change in expression is higher than the value of a time point when the following expression remains unchanged. This modified expression level is used in subsequent network analysis so that the time points when the rate is changing the most, and thus when regulators are likely to be active, are emphasized. ExRANGES multiplies the Expression level at time t with the RANGES value from time t to $t+1$ (Supplemental Figure 1B). This ExRANGES value was used in lieu of the expression level to generate a TRN using GENIE3 or INFERELATOR as described below (Bonneau *et al.* 2006; Huynh-Thu *et al.* 2010).

2.3 Network Inference using GENIE3

To predict regulatory interaction between transcription factor and target gene, GENIE3 was used. The GENIE3 script was taken from http://www.montefiore.ulg.ac.be/~huynh-thu/software.html on June 14, 2016 (Huynh-Thu *et al.* 2010). GENIE3 was modified by to be usable with *parLapply* from the R parallel package (R Core Team 2016). We compared ExRANGES to the standard approach of using expression levels alone (hereinafter after called EXPRESSION). The EXPRESSION network was built by providing the expression values across all samples for both TFs and targets. In contrast, for ExRANGES, we used the ExRANGES value for both TFs and targets, to emphasize the time points where expression is changing. For example, for the CircaDB data, we considered 1690 murine TFs as the regulators (Zhang *et al.* 2012a). The EXPRESSION network was built by providing the expression values across the 288 samples for each of the 1690 TFs as regulators and the 35,556 potential target genes across the same samples. To generate the ExRANGES network for the CircaDB data, the ExRANGES value was used as the input for the 35,556 targets and the 1690 regulators. For both approaches all TFs were also included in the target list to identify regulatory connections between TFs.

To implement GENIE3, we used 2000 trees for random forest for all data sets except the viral data set. For the viral data set, due to the size, we limited it to 100 trees. The importance measure from the random forest was calculated using the mean decrease in accuracy upon random permutation of individual features. In GENIE3, this measure is used as the prediction score for TF-Target relationships.

2.4 Network Inference using INFERELATOR

TF-target interactions were calculated from both EXPRESSION and ExRANGES for the Rice dataset. TF and targets labels are identical to those used as GENIE3 input. Time information in the form of the time step between each sample was added to satisfy time course conditions as a parameter, default values were used for all other parameters. Only confidence scores of TF-target interactions greater than 0 were evaluated against ChIP-Seq standards. The confidence scores were used as the prediction score to evaluate against the targets identified for each TF from experimental ChIP-Seq data.

2.5 ROC Calculation

ROC values were determined by the ROCR package in R (Sing *et al.* 2005). The random forest importance measures were used as the prediction score and the targets from the respective experimental validation (ChIP-Seq, protein binding array, or DAP-Seq) were used as the metric to evaluate the performance function. The area under the ROC curve (AUC) is presented to summarize the accuracy.

3 Results

3.1 ExRANGES Improves Identification of Circadian TF Targets in a Circadian Data Set

The assumption behind using correlation in gene expression to identify relationships between TFs and their targets is that there is a predictable relationship between the expression of the TF regulator and its corresponding targets. For transcriptional activators, the target will accumulate as the TF regulator accumulates. Conversely, targets of repressors will decrease in expression as the repressor TF increases. Current approaches evaluate the correspondence in expression between the regulator TF and targets across all time points equally. We tested whether incorporating the rate of change via ExRANGES identified different targets than EXPRESSION alone and if ExRANGES improves the overall ability to identify experimentally validated regulatory relationships.

To evaluate the ability of the ExRANGES or EXPRESSION approaches to correctly identify targets of the TFs, we applied both approaches to the CircaDB (Pizarro *et al.* 2013) data using GENIE3. We compared the results of each approach to the targets identified experimentally using ChIP-Seq for five TFs involved in circadian regulation: PER1, CLOCK, NPAS2, NR1d2, and ARNTL (Koike *et al.* 2012; Takahashi *et al.* 2015). Targets identified by each computational approach that were also considered significant targets in these published ChIP-Seq experiments were scored as true positive results.

We calculated the ROC AUC for the five circadian TFs to compare the identification of true targets attained with GENIE3 using EXPRESSION values to the combination of expression and p-values using ExRANGES. We observed that for all five TFs ExRANGES improved the identification of ChIP-Seq validated targets (Figure 1A).

Incorporation of a delay between regulator expression and target expression has previously been shown to improve the ability to identify regulatory networks (Huynh-Thu 2012). A modification of GENIE3 incorporates this approach to identify transcriptional changes in the regulator that precedes the effects on the target by a defined time step. We compared ExRANGES to this modified implementation of GENIE3 that includes the time delay step (Supplemental Figure 2A). As previously reported, we observe that the time step delay improved target identification for some TFs, compared to EXPRESSION alone, although in this data set, target identification for CLOCK, PER1, and NR1D2 TFs did not improve. However, for all five TFs, ExRANGES outperformed both the EXPRESSION and time-delay approaches in identifying the true positive targets of each TF; although for CLOCK, this improvement was minimal.

3.2 ExRANGES Improves Target Identification for TFs That Are Not Components of the Circadian Clock

To evaluate the performance of ExRANGES on TFs that are not core components of the circadian clock, we compared the ability to identify targets of additional TFs validated by ChIP-Seq. We selected seven TFs in our regulator list with ChIP-Seq data available from at least two experimental replicates performed in epithelial cells, a tissue not included in the CircaDB dataset. The seven TFs are: ESR1, STAT5A, STAT5B,

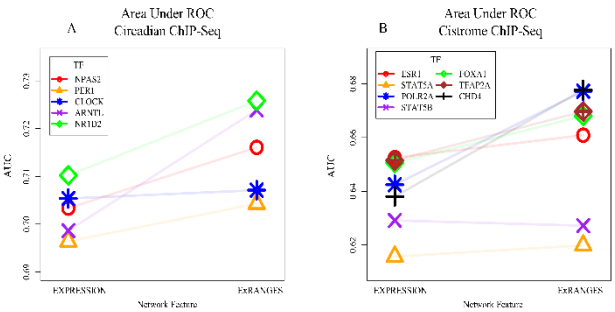


Figure 1: ExRANGES outperforms EXPRESSION in identifying targets for select TFs. A) ROC AUC for targets identified with GENIE3 using EXPRESSION or ExRANGES on five circadian TFs. The targets identified computationally were validated against the ChIP-Seq identified targets (Koike *et al.* 2012; Takahashi *et al.* 2015). B) ROC AUC for targets computationally identified by GENIE3 analysis using EXPRESSION or ExRANGES for seven TFs not known to be components of the circadian clock. Experimentally validated targets for these TFs were identified by ChIP-Seq in epithelial cells, a tissue not included in the expression data (Qin *et al.* 2012)**.**

POL2A, FOXA1, TFAP2A, and CHD4 (Qin *et al.* 2012). Combining expression and rate change information using ExRANGES improved the AUC curve for five of the seven TFs (ESR1, POL2A, FOXA1, TFAP2A, and CHD4) (Figure 1B, Supplemental Figure 2B). As we observed above for CLOCK, STAT5A and STAT5B performed equally well, but did not show significant improvement. STAT5A and STAT5B are known to be activated post-transcriptionally perhaps indicating why evaluating the change in expression of these TFs did not lead to improved identification of targets (Darnell *et al.* 1994; Liu *et al.* 1998; Horvath 2000; Bromberg and Chen 2001; Stark and Darnell 2012).

3.3 ExRANGES Identified Targets have Less Variation Across the Time Series

The targets identified by ExRANGES or EXPRESSION approaches show moderate overlap in the ranked score of predicted targets (r^2= 0.53); however, each network identifies different targets (Figure 1 and Supplemental Figure 3). To understand the difference in targets identified by EXPRESSION and ExRANGES we examined the variance in expression for the top 1000 predicted targets of the 12 TFs identified by EXPRESSION or by ExRANGES across all 288 samples in the CircaDB data set. The targets identified by ExRANGES showed overall lower variation in expression across all samples compared to targets identified by EXPRESSION (Figure 2). The experimentally identified targets from ChIP-Seq showed low average variation in expression. The ability of ExRANGES to identify targets with lower variation than EXPRESSION may account for some of the improved identification of such the True Positive Targets.

ExRANGES is a combination of rate change and expression. To evaluate the contribution of the rate change component compared to the expression values in the target identification, we generated a rate-based network using only the *p*-values of the rate change at each time step as our network feature. The rate change alone did not improve the overall identification of true positive targets (Supplemental Figure 4). However, the targets identified in the rate-based network did show lower overall variation in expression compared to the EXPRESSION identified targets. The CircaDB data consists of individual time series experiments from different tissues. Using rate change alone may enhance the identification of targets that have within tissue variation driven by changes across time compared to the larger overall variation between tissues observed in this dataset. In contrast expression identified targets may favor those with large changes in expression between tissues. To evaluate how expression and rate identified targets compared in variation within each time series in a single tissue versus between tissues, we compared the between tissue and within tissue standard deviation for the top 1000 targets identified by using EXPRESSION or rate change. The targets identified by EXPRESSION showed more variation between tissue types (Supplemental Figure 5A). In contrast, the targets identified by rate change alone showed increased variation within each tissue time series compared to the EXPRESSION identified targets (Supplemental Figure 5B). We also compared the mean intensity level of the top 1000 predicted targets of the rate change and EXPRESSION approaches. We observed that the top 1000 targets of PER1 identified by EXPRESSION had higher intensity levels compared to the distribution of expression of all transcripts on the microarray (Supplemental Figure 6A). In contrast, the top 1000 predicted targets of PER1 identified by rate change resembled the background distribution of intensity for all the transcripts on the array (Supplemental Figure S6B). Likewise, the hybridization intensity of the genes identified as the top 1000 targets identified by EXPRESSION of

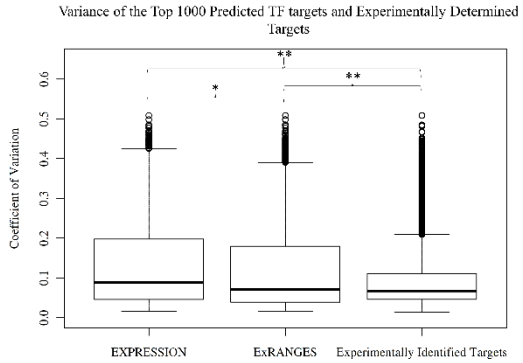


Figure 2: Target identified by ExRANGES and EXPRESSION have different variation across CircaDB dataset. Box plot showing the coefficient of variation for the expression levels of the top 1000 targets of each TF (ARNTL, CLOCK, NPAS2, NR1D2, PER1, ESR1, POL2A, FOXA1, TFAP2A, CHD4) predicted by GENIE3 using EXPRESSION or ExRANGES and the experimentally identified targets from ChIP-Seq. Experimentally determined targets showed the lowest variation. Targets identified using EXPRESSION show a greater variation in expression across all samples compared to targets predicted with ExRANGES values (* *p*-value < $2.5e^{-8}$, ** *p*-value < $2e^{-16}$).

all 1690 TFs considered as regulators was shifted higher compared to the background distribution levels (Supplemental Figure 6C). The top 1000 targets of all 1690 TFs identified by rate change reflected the background distribution of hybridization intensity (Supplemental Figure 6D). While hybridization intensity cannot directly be translated into expression levels, these observations suggest that there are features of the targets identified by rate change that are distinct from those identified by EXPRESSION.

3.4 ExRANGES Improves Identification of TF Targets in Unevenly Spaced Time Series Data

Circadian and diel time series experiments are a rich resource providing temporal variance, which can be used to identify regulatory relationships. However, most available experimental data is not collected with this design. Often sample collection cannot be controlled precisely to attain evenly spaced time points. For example, in human studies, the subject may not be available for consistent sampling. To evaluate the ability of ExRANGES to identify true targets of TFs across unevenly spaced and heterogeneous genotypes, we analyzed expression studies of viral infections in various individuals ("Respiratory Viral DREAM Challenge - Synapse ID syn5647810"; Liu *et al.* 2016) using both ExRANGES and EXPRESSION approaches. This data set consists of a series of blood samples from human patients taken over a seven to nine day period, depending on the specific study. Sampling was not evenly spaced between time points. Seven studies that each sampled multiple individuals before and after respiratory infection are included. In total 2372 samples were used, providing a background of 2231 consecutive time steps. Overall, the variance between samples was lower for this study than the circadian study examined above (Supplemental Figure 7). The significance of a change in expression for each gene at each time step was compared to a background distribution of change in expression across all patients and time steps (2231 total slope changes). We observed an overall improvement in the detection of ChIP-Seq identified targets for the 83 TFs on the HGU133 Plus 2.0 microarray (Affymetrix, Santa Clara, CA) with ChIP-Seq data from blood tissue (Liu *et al.* 2011), (Figure 3A). The improvement varies by TF (Figure 3B).

3.5 ExRANGES Improves Functional Cohesion of Identified Targets

ChIP-Seq targets are one method to identify true targets of a TF. Another approach is to look at functional enrichment. The true targets of a TF are likely to be involved in the same functional pathways and therefore true targets would be enriched for the same functional categories as measured by enrichment of GO terms. Comparison of functional enrichment of TF targets identified by each approach enables the evaluation of how each approach performs on identifying targets for TFs without available ChIP-Seq data. We compared the functional enrichment of the top 1000 targets of each TF predicted by either approach using Homo sapiens GO slim annotation categories. We evaluated the 930 TFs on the HGU133 microarray (Zhang *et al.* 2012a). Of these, the targets identified by ExRANGES for the majority of the TFs (590) showed improved functional enrichment compared to the targets identified by EXPRESSION (Figure 5A and B). Likewise, when focusing on the 83 TFs with available ChIP-Seq data from blood, the majority of TF targets predicted by ExRANGES were more functionally cohesive compared to EXPRESSION targets as evaluated by GO slim (Figure 4C). We observed that the improvement ranking of ExRANGES over EXPRESSION varies between the two validation approaches. For example, targets of the TF JUND identified by ExRANGES show no improvement over EXPRESSION when validated by ChIP-Seq identified targets, yet showed improved functional cohesion (Supplemental Table ST1).

3.6 ExRANGES Improves TF Target Identification from RNA-Seq Data and Validated by Experimental Methods Other Than ChIP-Seq

The previous evaluations of ExRANGES were performed on expression data obtained from microarray-based studies. To evaluate the performance of ExRANGES compared to EXPRESSION for RNA-Seq data we applied each approach to an RNA-Seq data set from *Saccharomyces cerevisiae* (Vardi *et al.* 2014). This data set consisted of samples collected from six different genotypes every fifteen minutes for six hours after transfer to media lacking phosphate. The slope background was calculated from 144 time steps. To evaluate the performance of ExRANGES compared to EXPRESSION approaches we calculated the AUC for the identified targets using GENIE3 for each of the 52 TFs using the TF targets identified by protein binding microarray analysis as the gold standard (Zhu *et al.* 2009). For most TFs, the AUC was improved using ExRANGES compared to EXPRESSION (Figure 5A).

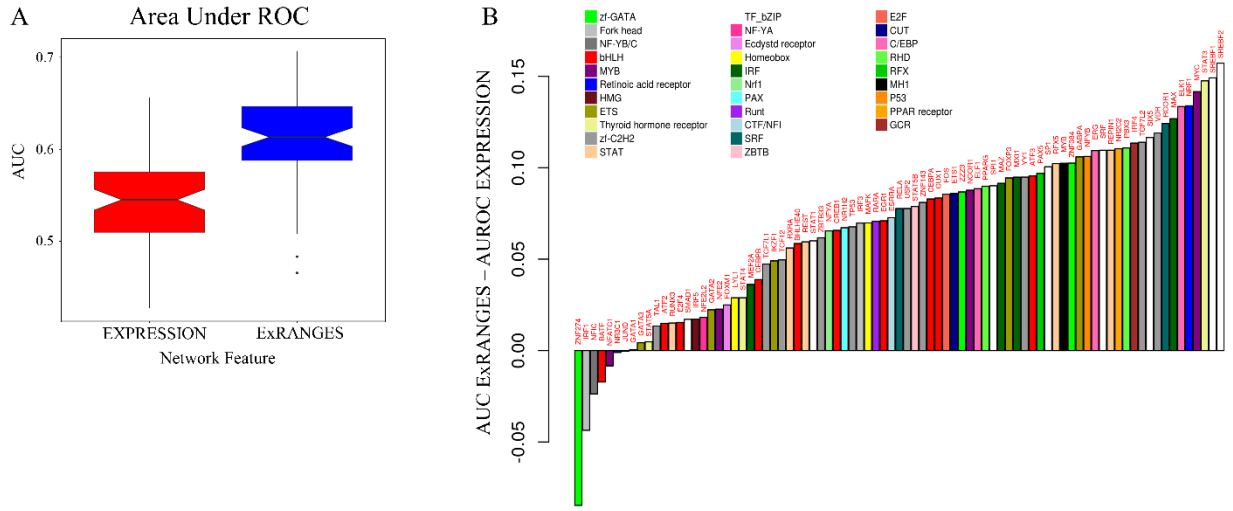


Figure 3: ExRANGES Improves Identification of Targets for most TFs from Unevenly Spaced Time Series Data. We compared targets of 83 TFs where ChIP-Seq data is available from Cistrome (Qin *et al.* 2012) identified using GENIE3 with either EXPRESSION or ExRANGES using expression data from the viral data set. A) Box plot of ROC AUC for the GENIE3 analysis for all 83 TFs using either EXPRESSION or ExRANGES compared to ChIP-Seq identified targets. B) The difference between the ROC AUC of ExRANGES and EXPRESSION predicted targets is plotted individually for each of the 83 TFs tested, in ascending order. TFs are colored by TF family.

A GO Enrichment Score Table

	Mean	Max	TFs with Higher Enrichment
ExRANGES	165.72	738.63	595
EXPRESSION	131.60	505.78	335

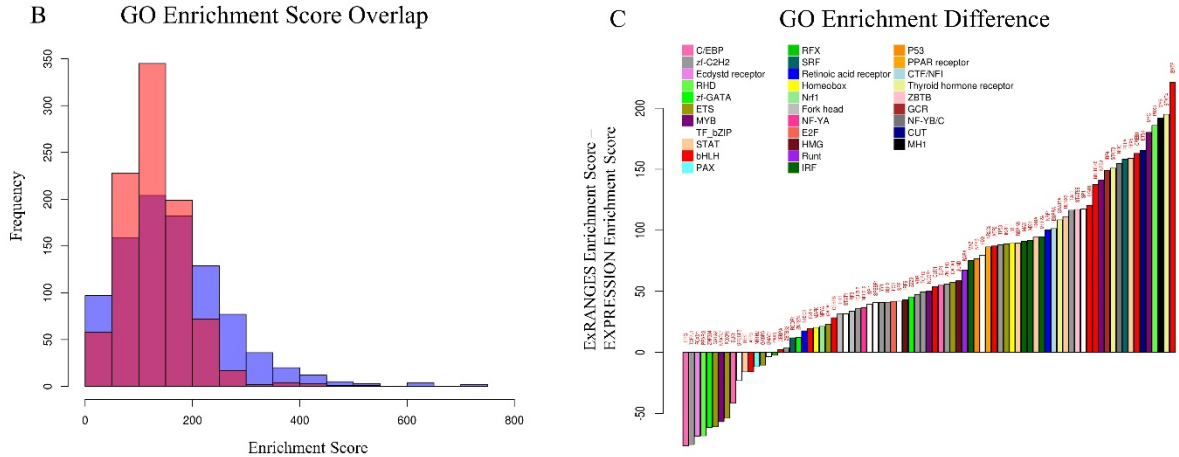


Figure 4: ExRANGES improves Functional Cohesion of Identified Targets. Gene Ontology term enrichment was calculated for the top 1000 predicted targets of 930 TFs using GENIE3 with either ExRANGES or EXPRESSION. Enrichment score is the sum of the $-\log_{10}$ of the *p*-value of each GO category. A) Summary of the enrichment scores for the top 1000 targets of all TFs on the microarray. B) The distribution of enrichments scores from EXPRESSION identified targets (red) and ExRANGES identified targets (blue). C) Difference in the enrichment score for the 83 TFs with available ChIP-Seq data (Fig 3). Positive values indicate TF targets with a higher enrichment score in ExRANGES compared to EXPRESSION.

We next evaluated the performance of EXPRESSION and ExRANGES on a set of data from Arabidopsis consisting of 144 samples collected every four hours for two days in 12 different growth conditions (Harmer *et al.* 2000; Smith *et al.* 2004; Bläsing *et al.* 2005; Edwards *et al.* 2006; Michael *et al.* 2008). Even though fewer ChIP-Seq data sets are available to validate the predicted targets in Arabidopsis, we were able to evaluate the performance of the algorithms for five TFs with available ChIP-Seq or ChIP-Chip identified targets performed in at least two replicates (Lee *et al.* 2007; Yant *et al.* 2010; Chang *et al.* 2013; Liu *et al.* 2013; Nagel *et al.* 2015). We observed that for all five TFs, ExRANGES showed improved identification of the ChIP-based true positive TF targets (Figure 5B). To evaluate a larger range of targets we compared our predicted targets by EXPRESSION or ExRANGES to 307 TFs targets identified by DAP-Seq (O'Malley *et al.* 2016). We observed that ExRANGES also showed an improved ability to identify targets as validated by DAP-Seq compared to EXPRESSION (Figure 5C).

3.7 Application of ExRANGES to Smaller Data Sets with Limited Validation Resources

Time series data offers several advantages; however, it also increases the experimental costs. We have shown that using ExRANGES improves performance of GENIE3 on large data sets as validated by ChIP-Seq (228 samples in mouse, 2372 in human, and 144 in arabidopsis) (Figure 6). Since our interest is to develop a tool that can assist with the identification of regulatory networks in non-model species, we wanted to determine if ExRANGES could also improve identification of TF targets in more sparsely sampled data sets where there is only limited validation data available.

To determine the effectiveness of the ExRANGES approach for experiments with limited time steps, we evaluated the targets identified by ExRANGES and EXPRESSION for a single unpublished time series consisting of 32 samples from eight unevenly sampled time points of field grown rice panicles. ChIP-Seq with replicates has only been performed for one transcription factor in rice, OsMADS1 (Khanday *et al.* 2016). Therefore, we compared the ability of ExRANGES and EX-

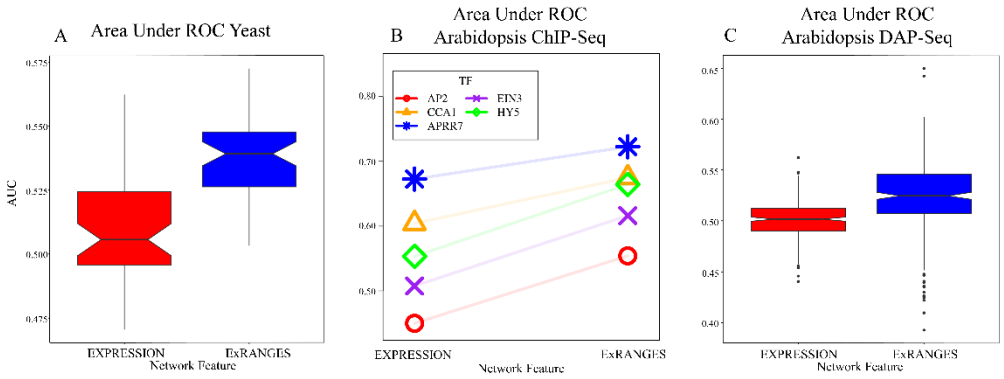


Figure 5: ExRANGES improves identification of TF targets validated by different methods. A) Box plots of the ROC AUC for targets identified for 52 yeast TFs by ESPRESSION or ExRANGES and validated against experimentally identified targets from protein binding microarray data (Zhu *et al.* 2009). B) ROC AUC for targets identified using GENIE3 with either EXPRESSION or ExRANGES for five Arabidopsis TFs validated against ChIP-Seq data. C). Box plot of AUC for targets identified for 307 Arabidopsis TFs by EXPRESSION and ExRANGES validated against DAP-Seq identified targets (O'Malley *et al.* 2016).

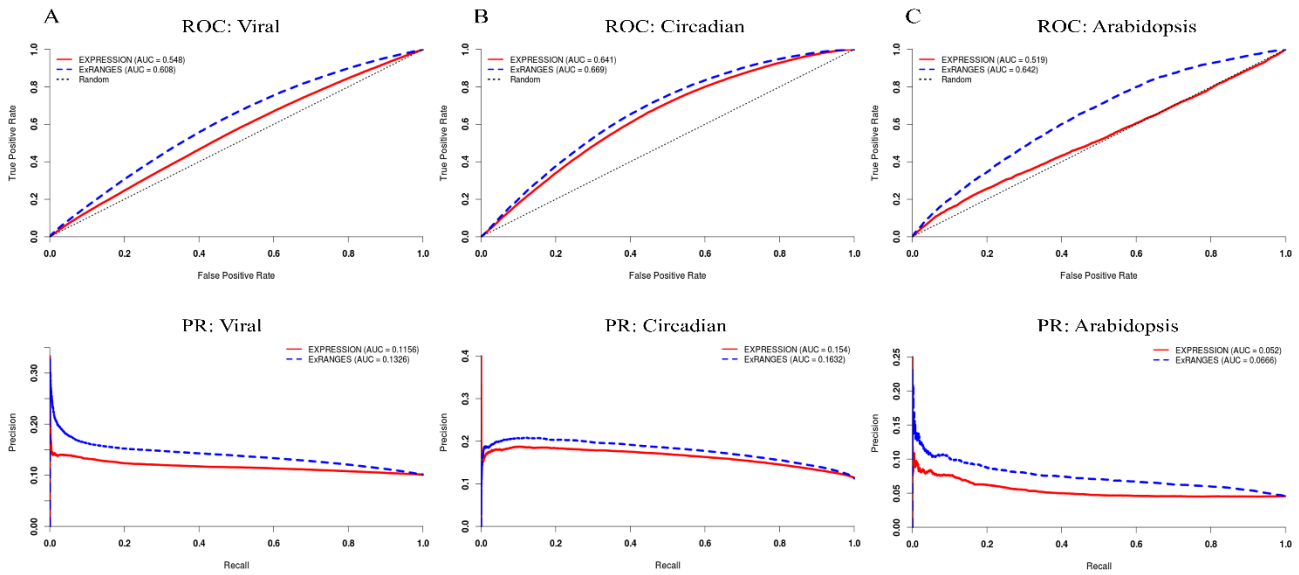


Figure 6: Summary of ExRANGES improvement across three data sets from different species. ROC and Precision recall (PR) curves for targets of all ChIP-Seq validated TFs as identified using GENIE3 with either EXPRESSION (solid) or ExRANGES (dotted) for A) CircaDB dataset from mouse tissues B) Human viral data set C) Arabidopsis circadian dataset across different environmental variables.

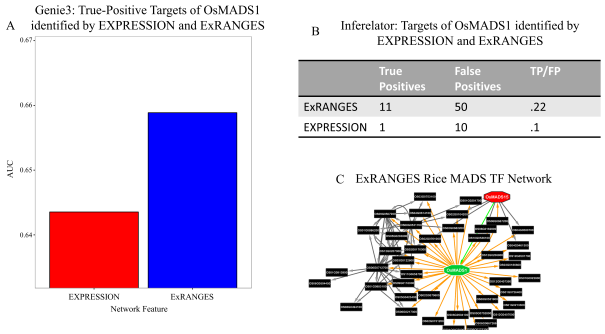
PRESSION to identify the OsMADS1 targets identified by L. Khanday et al. Of the 3112 OsMADS1 targets identified by ChIP-Seq, ExRANGES showed an improved ability to identify these targets (Figure 7) compared to EXPRESSION alone.

4 Discussion

Computational approaches that identify candidate targets of regulators can advance research. Many approaches have been developed to identify regulator targets, but most of these use expression values. We have demonstrated that combining the expression levels and rate of change improves the ability to predict true targets of TFs across a range of species and experimental designs. This approach improves the identification of targets as determined by ChIP-Seq and protein binding microarray across many different collections of time series data including experiments with replicates and without, with time series that have unevenly sampled time points, and even for time series with limited number of samples. ExRANGES provides improvement in TF target identification over EXPRESSION values alone for time series performed with both microarray and RNA-Seq measurements of expression.

Expression analysis performed in time series, such as experiments evaluating the transcriptional changes throughout a circadian cycle, provide rich resources for identifying relationships between individual transcripts. Since in many species the majority of transcripts show variation in expression levels throughout the day (Michael *et al.* 2008; Doherty and Kay 2011; Pizarro *et al.* 2013) circadian and diel data sets provide a snapshot of the potential ranges in expression that a regulator can attain. The associated changes in target expression levels can be analyzed to identify potential regulatory relationships that may be enhanced in response to other perturbations such as stress. However, for some targets, the daily variation in expression may be dwarfed by the large variation in expression between tissues. Here, we show that using ExRANGES, data sets that combine circadian time series in multiple tissues can be a powerful resource for identifying regulatory relationships between TFs and their targets not just for circadian regulators, but also for regulators that are not components of the circadian clock. Targets identified using EXPRESSION as the features showed large variance between tissues, while targets identified using rate change showed larger variance within each time series (Figure 2, and Supplemental Figure 5). ExRANGES takes advantage of both sources of variation and improves the identification of TF targets for most regulators tested, including for TF-target relationships in tissues not included in the transcriptional analysis. Additionally, ExRANGES simplifies incorporation of replicate samples.

As implemented, ExRANGES improves the ability to identify regulator targets, however, there are many aspects that could be further optimized. For example, we tested ExRANGES with the network inference algorithms GENIE3 and INFERELATOR demonstrating that it improves the performance of these algorithms. The ExRANGES method can be ap-



	True Positives	False Positives	TP/FP
ExRANGES	11	50	.22
EXPRESSION	1	10	.1

Figure 7: ExRANGES Retains Performance Improvement over EXPRESSION on Small Data Sets. A) ROC AUC for the top 1000 targets of OsMADS1 identified by GENIE3 using EXPRESSION or ExRANGES and validated against the OsMADS1 ChIP-Seq data (Khanday *et al.* 2016). B) Comparison of targets identified by EXPRESSION and ExRANGES using INFERELATOR. ExRANGES scores higher in the ratio of True Positive (TP) to False Positives (FP). C) Interactions predicted by ExRANGES of OsMADS1 (center, green) with other MADS TFs. Orange arrows indicate ExRANGES predicted targets of OsMADS1. ExRANGES predicts that OsMADS15 (red) regulates OsMADS1 (green arrow). Interactions between other MADS TFs predicted by ExRANGES are indicated by black arrows.

plied to most other machine learning applications such as Bayesian networks, mutual information networks, or even supervised machine learning tools. In addition, we showed that ExRANGES outperformed a one-step time delay. Conceptually, our method essentially increases the weight of the time point before a major change in expression level. ExRANGES could be further modified to adjust where that weight is placed, a step or more in advance, depending on the time series data. Such incorporation of a time delay optimization into the ExRANGES approach could lead to further improvement for identification of some TF targets, although it would increase the computational cost.

Here, we compared ExRANGES based features to EXPRESSION based features by validating against TF targets identified by ChIP-Seq and ChIP-Chip. While these experimental approaches identify potential TF targets in a genome-wide manner, systemic bias in ChIP could bias the comparisons (Teytelman *et al.* 2013). For example, we observed that ChIP-seq identified targets in the CircaDB dataset showed lower variation in expression than computationally identified targets (Figure 2). The use of ExRANGES as a network input also outperformed the use of EXPRESSION alone when validated against DAP-Seq, and protein binding microarray. Although ChIP-Seq may not be an ideal gold standard, it is the most available experimental resource for benchmarking computational approaches to identifying TF targets. Unfortunately, high quality ChIP-Seq data is not available in most organisms for more than a handful of TFs. For example, validation of this approach in rice was limited to one recently published ChIP-Seq dataset. This lack of experimentally identified targets is a severe hindrance to advancing research in these species. New experimental approaches such as DAP-Seq may provide alternatives for TF target identification in species recalcitrant to ChIP-Seq analysis (O'Malley *et al.* 2016). Additionally, O'Malley et al. improved their recall of ChIP-Seq identified targets by selecting targets that were also supported by DNase-Seq sensitivity assays (Zhang *et al.* 2012b; Sullivan *et al.* 2014). Likewise, distinguishing between direct and indirect targets predicted computationally could be enhanced by incorporation of DNase-Seq or motif occurrence information for the targets. Incorporation of such a priori information on regions of open chromatin and occurrence of cis-regulatory elements leads to improved network reconstruction (Greenfield *et al.* 2013; Wilkins *et al.* 2016). Use of ExRANGES could lead to improvement for these integrated approaches. Although approaches such as DAP-Seq are more global in analyses than individual ChIP-Seq assays, these genome-wide approaches still require a significant investment from the community in the development of an expressed TF library collection. For non-model systems, computational identification of TF targets can provide an economical first pass that can be followed up by experimental analysis of predicted targets, accepting the fact that there will be false positives in the validation pipeline. In this strategy, a small improvement in the ability to identify true targets of a given TF can translate into a reduced number of candidates to test and fewer experiments that must be performed. We hope that the improvements to regulatory network algorithms provided by the ExRANGES approach can facilitate research in species where identification of TF targets is experimentally challenging. Additionally, we hope that our finding of how gene expression values are incorporated in a network has a significant effect on the ability to identify regulatory relationships will stimulate evaluation of new approaches that use alternative methods to incorporate time signals into regulatory network analysis.

In summary, we demonstrate that consideration of how expression data is incorporated can contribute to the success of TRN reconstruction. ExRANGES is a first step at evaluating different approaches for how features are supplied to regulatory network inference algorithms. We anticipate that further optimization and other novel methods for integrating expression information will lead to improvements in network reconstruction that ultimately will accelerate biological discovery.

Acknowledgements

We would like to thank Dahlia Nielsen, Katie Greenham, and Erin Slabaugh for critical suggestions on the manuscript preparation. Additionally, we thank Steve Briggs for sharing the time, expertise, and helpful discussions of his research group. This is contribution no. 17-389-J of the Kansas Agricultural Experiment Station.

Funding

This work was supported by funding from USDA NIFA 2014-04051.

Conflict of Interest: none declared.

References

Anders S., McCarthy D. J., Chen Y., Okoniewski M., Smyth G. K., *et al.*, 2013 Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat. Protoc. 8: 1765–1786.

Balázsi G., Oudenaarden A. Van, Collins J. J., 2011 Cellular decision making and biological noise: From microbes to mammals. Cell 144: 910–925.

Bar-Joseph Z., Gitter A., Simon I., 2012 Studying and modelling dynamic biological processes using time-series gene expression data. Nat Rev Genet 13: 552–564.

Bläsing O. E., Gibon Y., Günther M., Höhne M., Morcuende R., *et al.*, 2005 Sugars and Circadian Regulation Make Major Contributions to the Global Regulation of Diurnal Gene Expression in Arabidopsis. Plant Cell Online 17: 3257–3281.

Bonneau R., Reiss D. J., Shannon P., Facciotti M., Hood L., *et al.*, 2006 The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. Genome Biol. 7: R36.

Breiman L., 2001 Random forests. Mach. Learn. 45: 5–32.

Bromberg J., Chen X., 2001 STAT proteins: Signal tranducers and activators of transcription. In: Enzymology BT-M in (Ed.), *Regulators and Effectors of Small GTPases, Part G*, Academic Press, pp. 138–151.

Chang K. N., Zhong S., Weirauch M. T., Hon G., Pelizzola M., *et al.*, 2013 Temporal transcriptional response to ethylene gas drives growth hormone cross-regulation in Arabidopsis. Elife 2: e00675.

Darnell J. E., Kerr I. M., Stark G. R., 1994 Jak-STAT pathways and transcriptional activation in response to IFNs and other extracellular signaling proteins. Science (80-.). 264: 1415 LP-1421.

Doherty C. J., Kay S. A., 2011 Circadian Control of Global Gene Expression Patterns. Annu. Rev. Genet. 44: 419–444.

Edwards K. D., Anderson P. E., Hall A., Salathia N. S., Locke J. C. W., *et al.*, 2006 FLOWERING LOCUS C Mediates Natural Variation in the High-Temperature Response of the Arabidopsis Circadian Clock. Plant Cell Online 18: 639–650.

Eisen M. B., Spellman P. T., Brown P. O., Botstein D., 1998 Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. 95: 14863–14868.

Faith J. J., Hayete B., Thaden J. T., Mogno I., Wierzbowski J., *et al.*, 2007 Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles. PLOS Biol 5: e8.

Greenfield A., Hafemeister C., Bonneau R., 2013 Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. Bioinformatics 29: 1060–1067.

Harmer S. L., Hogenesch J. B., Straume M., Chang H. S., Han B., *et al.*, 2000 Orchestrated transcription of key pathways in Arabidopsis by the circadian clock. Science 290: 2110–2113.

Horvath C. M., 2000 STAT proteins and transcriptional responses to extracellular signals. Trends Biochem. Sci. 25: 496–502.

Huynh-Thu V. A., Irrthum A., Wehenkel L., Geurts P., 2010 Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. PLoS One 5: e12776.

Huynh-Thu V. A., 2012 Machine learning-based feature ranking: Statistical interpretation and gene network inference.

Jin J., He K., Tang X., Li Z., Lv L., *et al.*, 2015 An Arabidopsis Transcriptional Regulatory Map Reveals Distinct Functional and Evolutionary Features of Novel Transcription Factors. Mol. Biol. Evol. 32: 1767–1773.

Khanday I., Das S., Chongloi G. L., Bansal M., Grossniklaus U., *et al.*, 2016 Genome-wide targets regulated by the OsMADS1 transcription factor reveals its DNA recognition properties. Plant Physiol. .

Kim D., Pertea G., Trapnell C., Pimentel H., Kelley R., *et al.*, 2013 TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14.

Koike N., Yoo S.-H., Huang H.-C., Kumar V., Lee C., *et al.*, 2012 Transcriptional Architecture and Chromatin Landscape of the Core Circadian Clock in Mammals. Science (80-.).

Lee J., He K., Stolc V., Lee H., Figueroa P., *et al.*, 2007 Analysis of Transcription Factor HY5 Genomic Binding Sites Revealed Its Hierarchical Role in Light Regulation of Development. Plant Cell 19: 731–749.

Li Z., Li P., Krishnan A., Liu J., 2011 Large-scale dynamic gene regulatory network inference combining differential equation models with local dynamic Bayesian network analysis. Bioinformatics 27: 2686–2691.

Liaw A., Wiener M., 2002 Classification and regression by randomForest. R news 2: 18–22.

Liu K. D., Gaffen S. L., Goldsmith M. A., 1998 JAK/STAT signaling by cytokine receptors. Curr. Opin. Immunol. 10: 271–278.

Liu T., Ortiz J. A., Taing L., Meyer C. A., Lee B., *et al.*, 2011 Cistrome: an integrative platform for transcriptional regulation studies. Genome Biol. 12: R83.

Liu T., Carlsson J., Takeuchi T., Newton L., Farré E. M., 2013 Direct regulation of abiotic responses by the Arabidopsis circadian clock component PRR7. Plant J.: n/a-n/a.

Liu T.-Y., Burke T., Park L. P., Woods C. W., Zaas A. K., *et al.*, 2016 An individualized predictor of health and disease using paired reference and target samples. BMC Bioinformatics 17: 47.

Marbach D., Costello J. C., Küffner R., Vega N. M., Prill R. J., *et al.*, 2012 Wisdom of crowds for robust gene network inference. Nat. Methods 9: 796–804.

Margolin A. A., Nemenman I., Basso K., Wiggins C., Stolovitzky G., *et al.*, 2006 ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. BMC Bioinformatics 7: 1–15.

Matsumoto T., Wu J. Z., Kanamori H., Katayose Y., Fujisawa M., *et al.*, 2005 The map-based sequence of the rice genome. Nature 436: 793–800.

Michael T. P., Mockler T. C., Breton G., McEntee C., Byer A., *et al.*, 2008a Network discovery pipeline elucidates conserved time-of-day–specific cis-regulatory modules. PLoS Genet. 4: e14.

Michael T. P., Mockler T. C., Breton G., McEntee C., Byer A., *et al.*, 2008b Network Discovery Pipeline Elucidates Conserved Time-of-Day-Specific cis-Regulatory Modules. PLoS Genet. 4.

Nagel D. H., Doherty C. J., Pruneda-Paz J. L., Schmitz R. J., Ecker J. R., *et al.*, 2015 Genome-wide identification of CCA1 targets uncovers an expanded clock network in Arabidopsis. Proc. Natl. Acad. Sci. 112: E4802–E4810.

O'Malley R. C., Huang S. C., Song L., Lewsey M. G., Bartlett A., *et al.*, 2016 Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell 165: 1280–1292.

Park P. J., 2009 ChIP-seq: advantages and challenges of a maturing technology. Nat. Rev. Genet. 10: 669–80.

Penfold C. A., Buchanan-Wollaston V., Denby K. J., Wild D. L., 2012 Nonparametric Bayesian inference for perturbed and orthologous gene regulatory networks. Bioinformatics 28: 233–241.

Penfold C. A., Shifaz A., Brown P. E., Nicholson A., Wild D. L., 2015 CSI: A nonparametric Bayesian approach to network inference from multiple perturbed time series gene expression data. Stat. Appl. Genet. Mol. Biol. 14: 307–310.

Pizarro A., Hayer K., Lahens N. F., Hogenesch J. B., 2013 CircaDB: a database of mammalian circadian gene expression profiles. Nucleic Acids Res. 41: D1009–D1013.

Qian J., Lin J., Luscombe N. M., Yu H., Gerstein M., 2003 Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. Bioinformatics 19: 1917–1926.

Qin B., Zhou M., Ge Y., Taing L., Liu T., *et al.*, 2012 CistromeMap: a knowledgebase and web server for ChIP-Seq and DNase-Seq studies in mouse and human. Bioinformatics 28: 1411–1412.

R Core Team, 2016 R: A Language and Environment for Statistical Computing.

Respiratory Viral DREAM Challenge - syn5647810,

Sing T., Sander O., Beerenwinkel N., Lengauer T., 2005 ROCR: visualizing classifier performance in R. Bioinformatics 21: 7881.

Smith S. M., Fulton D. C., Chia T., Thorneycroft D., Chapple A., *et al.*, 2004 Diurnal Changes in the Transcriptome Encoding Enzymes of Starch Metabolism Provide Evidence for Both Transcriptional and Posttranscriptional Regulation of Starch Metabolism in Arabidopsis Leaves. Plant Physiol. 136: 2687–2699.

Stark G. R., Darnell J. E., 2012 The JAK-STAT Pathway at Twenty. Immunity 36: 503–514.

Sullivan A. M., Arsovski A. A., Lempe J., Bubb K. L., Weirauch M. T., *et al.*, 2014 Mapping and Dynamics of Regulatory DNA and Transcription Factor Networks in A. thaliana. Cell Rep. 8: 2015–2030.

Takahashi J. S., Kumar V., Nakashe P., Koike N., Huang H.-C., *et al.*, 2015 ChIP-seq and RNA-seq methods to study circadian control of transcription in mammals. Methods Enzymol. 551: 285–321.

Teytelman L., Thurtle D. M., Rine J., Oudenaarden A. van, 2013 Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. Proc. Natl. Acad. Sci. 110: 18602–18607.

Thompson D., Regev A., Roy S., 2015 *Comparative Analysis of Gene Regulatory Networks: From Network Reconstruction to Evolution*.

Vardi N., Levy S., Gurvich Y., Polacheck T., Carmi M., *et al.*, 2014 Sequential Feedback Induction Stabilizes the Phosphate Starvation Response in Budding Yeast. Cell Rep. 9: 1122–1134.

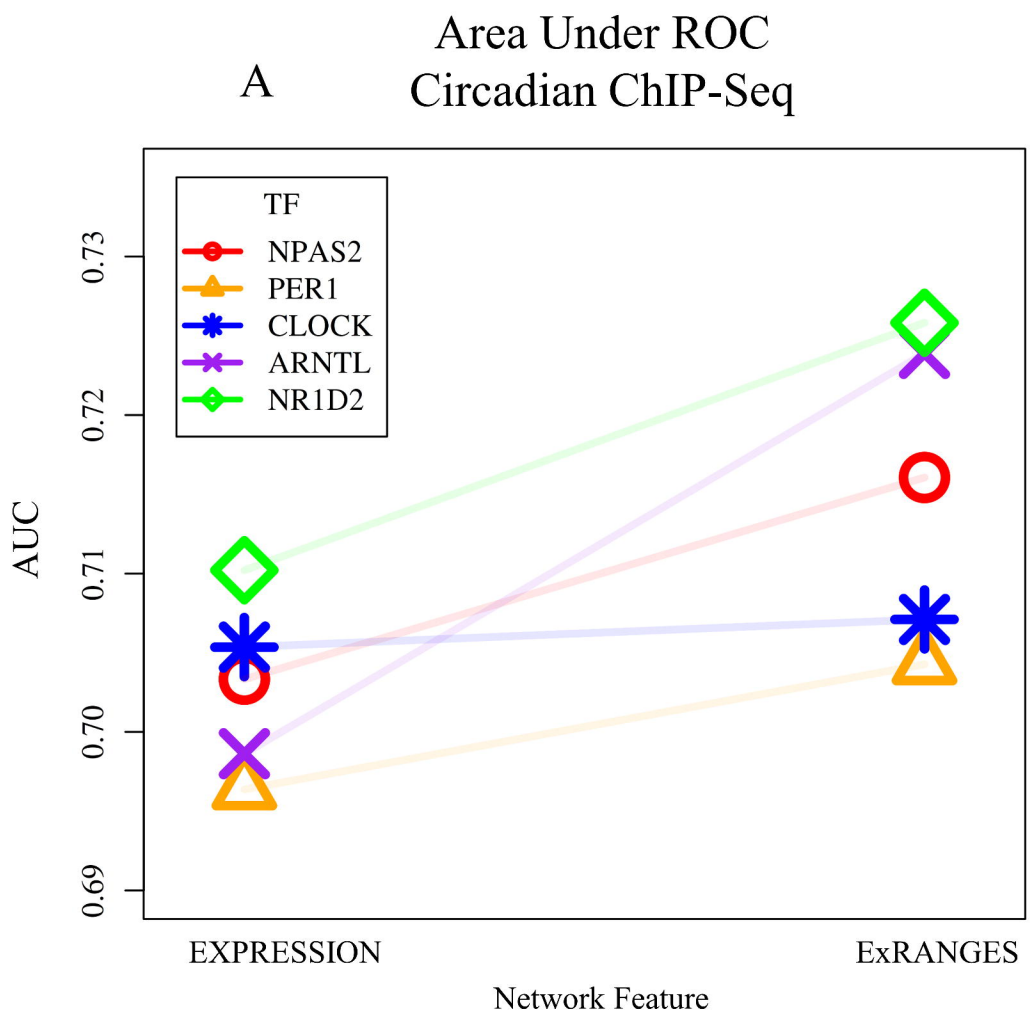
Wilkins O., Hafemeister C., Plessis A., Holloway-Phillips M.-M., Pham G. M., *et al.*, 2016 EGRINs (Environmental Gene Regulatory Influence Networks) in Rice That Function in the Response to Water Deficit, High Temperature, and Agricultural Environments. Plant Cell: tpc.00158.2016.

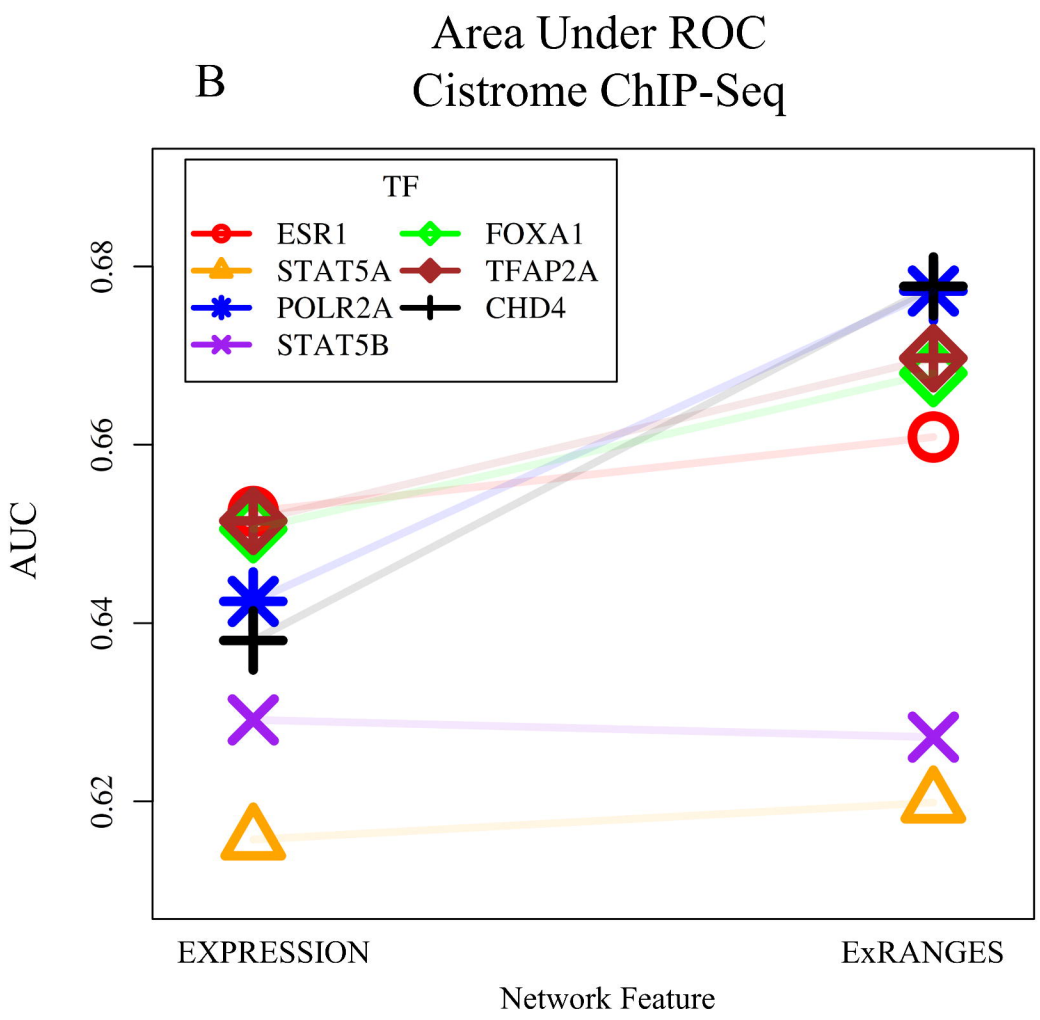
Yant L., Mathieu J., Dinh T. T., Ott F., Lanz C., *et al.*, 2010 Orchestration of the Floral Transition and Floral Development in Arabidopsis by the Bifunctional Transcription Factor APETALA2. Plant Cell Online 22: 2156–2170.

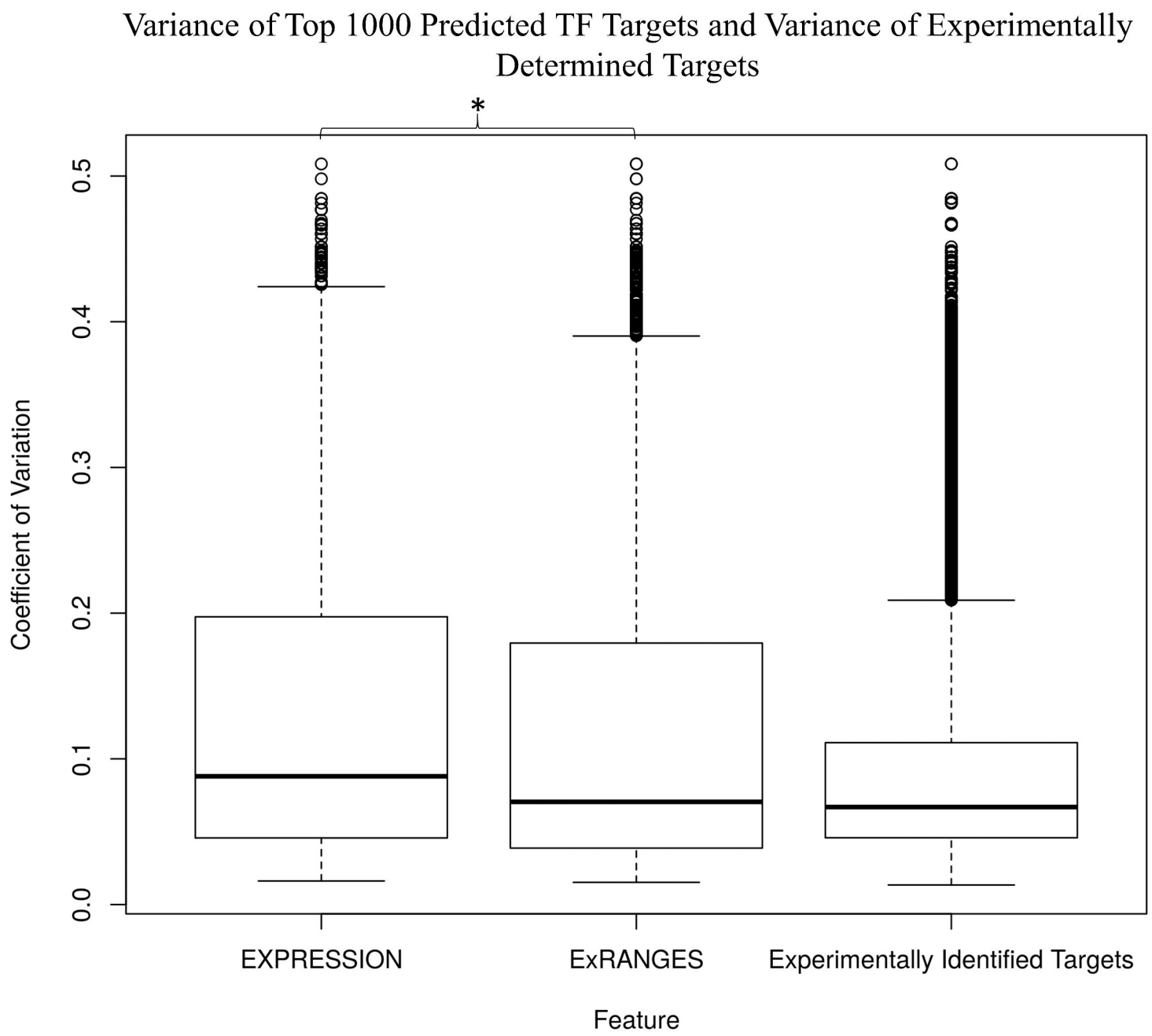
Zhang H.-M., Chen H., Liu W., Liu H., Gong J., *et al.*, 2012a Animal TFDB: a comprehensive animal transcription factor database. Nucleic Acids Res. 40: D144–D149.

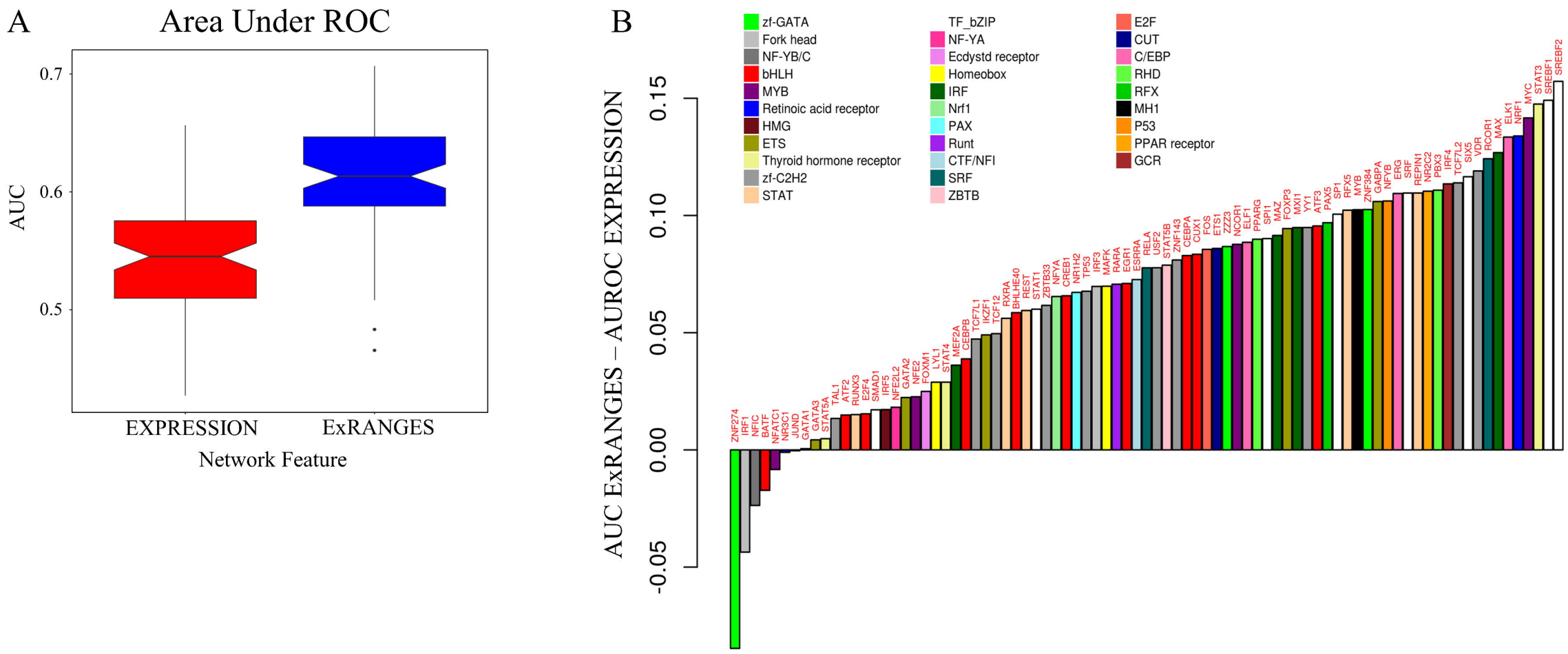
Zhang W., Zhang T., Wu Y., Jiang J., 2012b Genome-Wide Identification of Regulatory DNA Elements and Protein-Binding Footprints Using Signatures of Open Chromatin in Arabidopsis[C][W][OA]. Plant Cell 24: 2719–2731.

Zhu C., Byers K. J. R. P., McCord R. P., Shi Z., Berger M. F., *et al.*, 2009 High-resolution DNA-binding specificity analysis of yeast transcription factors. Genome Res. 19: 556–566.


A
Area Under ROC
Circadian ChIP-Seq
TF
NPAS2
PER1
CLOCK
ARNTL
NR1D2
AUC
0.69
0.70
0.71
0.72
0.73
EXPRESSION
ExRANGES
Network Feature


B
Area Under ROC
Cistrome ChIP-Seq
TF
ESR1
STAT5A
POLR2A
STAT5B
FOXA1
TFAP2A
CHD4
AUC
0.62
0.64
0.66
0.68
EXPRESSION
ExRANGES
Network Feature


Variance of Top 1000 Predicted TF Targets and Variance of Experimentally Determined Targets
*
Coefficient of Variation
0.0
0.1
0.2
0.3
0.4
0.5
EXPRESSION
ExRANGES
Experimentally Identified Targets
Feature


A
Area Under ROC
AUC
0.7
0.6
0.5
EXPRESSION
ExRANGES
Network Feature
B
AUC ExRANGES – AUROC EXPRESSION
0.15
0.10
0.05
0.00
-0.05
zf-GATA
Fork head
NF-YB/C
bHLH
MYB
Retinoic acid receptor
HMG
ETS
Thyroid hormone receptor
zf-C2H2
STAT
TF_bZIP
NF-YA
Ecdystd receptor
Homeobox
IRF
Nrf1
PAX
Runt
CTF/NFI
SRF
ZBTB
E2F
CUT
C/EBP
RHD
RFX
MH1
P53
PPAR receptor
GCR
ZNF274
IRF1
NFIC
BATF
NFATC1
NR3C1
JUND
GATA1
GATA3
STAT5A
TAL1
ATF2
RUNX3
E2F4
SMAD1
IRF5
NFE2L2
GATA2
NFE2
FOXM1
LYL1
STAT4
MEF2A
CEBPB
TCF7L1
IKZF1
TCF12
RXRA
BHLHE40
REST
STAT1
ZBTB33
NFYA
CREB1
NR1H2
TP53
IRF3
MAFK
RARA
EGR1
ESRRA
RELA
USF2
STAT5B
ZNF143
CEBPA
CUX1
FOS
ETS1
ZZZ3
NCOR1
ELF1
PPARG
SPI1
MAZ
FOXP3
MXI1
YY1
ATF3
PAX5
SP1
RFX5
MYB
ZNF384
GABPA
NFYB
ERG
SRF
REPIN1
NR2C2
PBX3
IRF4
TCF7L2
SIX5
VDR
RCOR1
MAX
ELK1
NRF1
MYC
STAT3
SREBF1
SREBF2

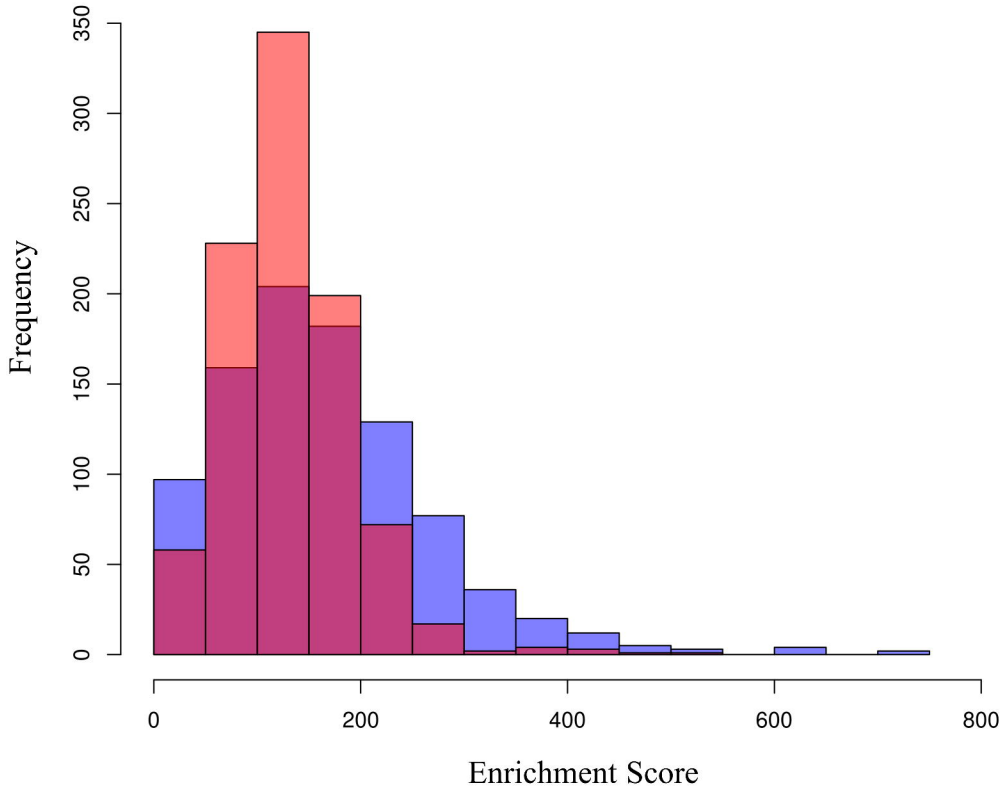
A

GO Enrichment Score Table

	Mean	Max	TFs with Higher Enrichment
ExRANGES	165.72	738.63	595
EXPRESSION	131.60	505.78	335

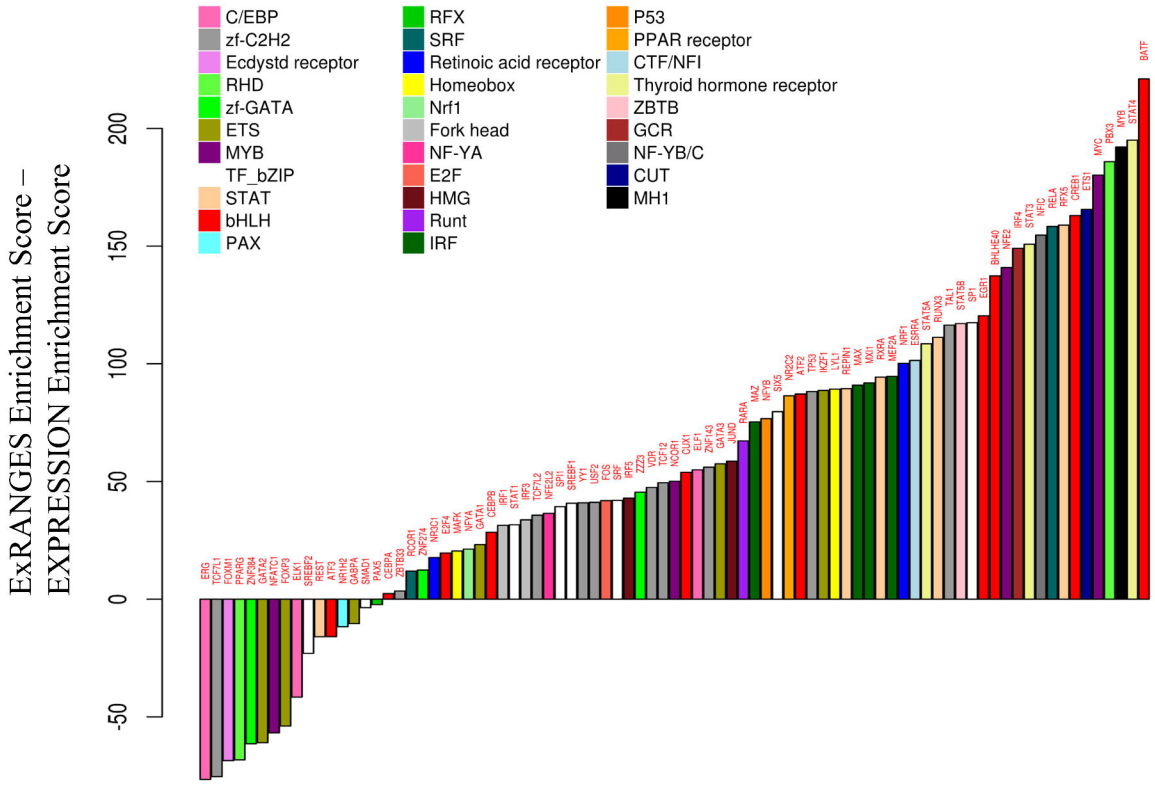
B

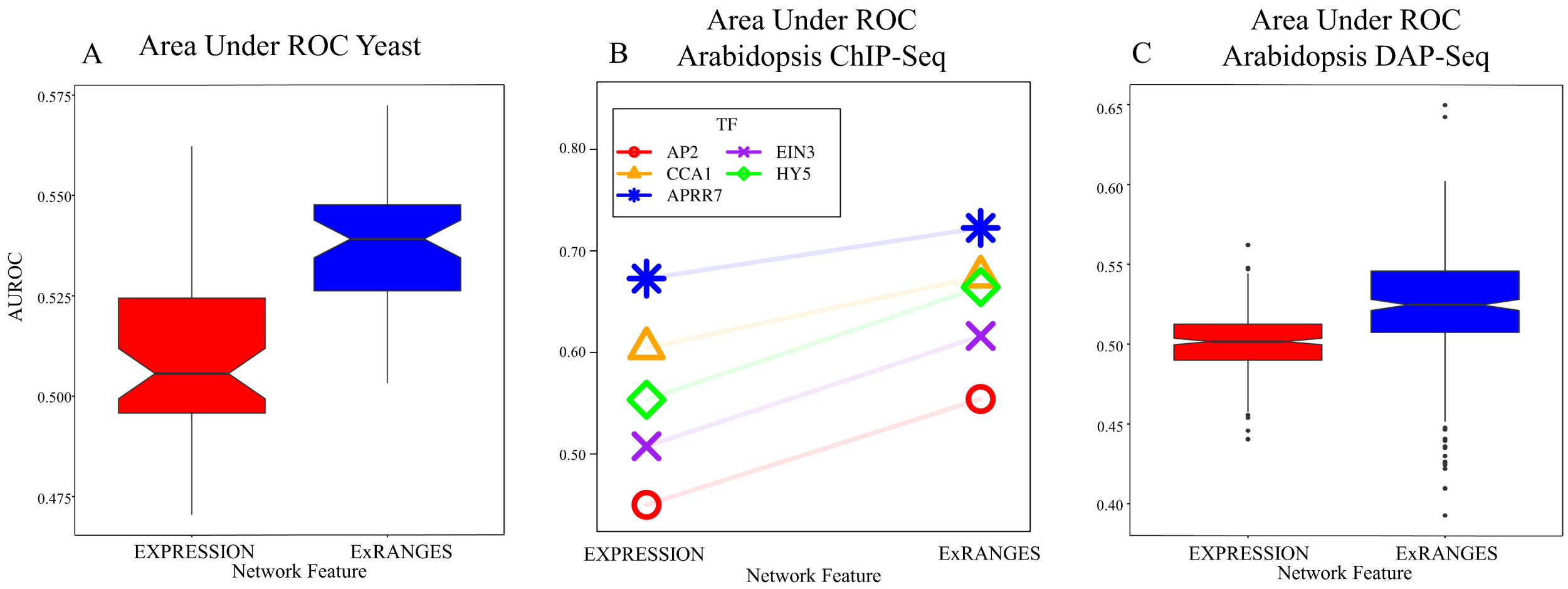
GO Enrichment Score Overlap

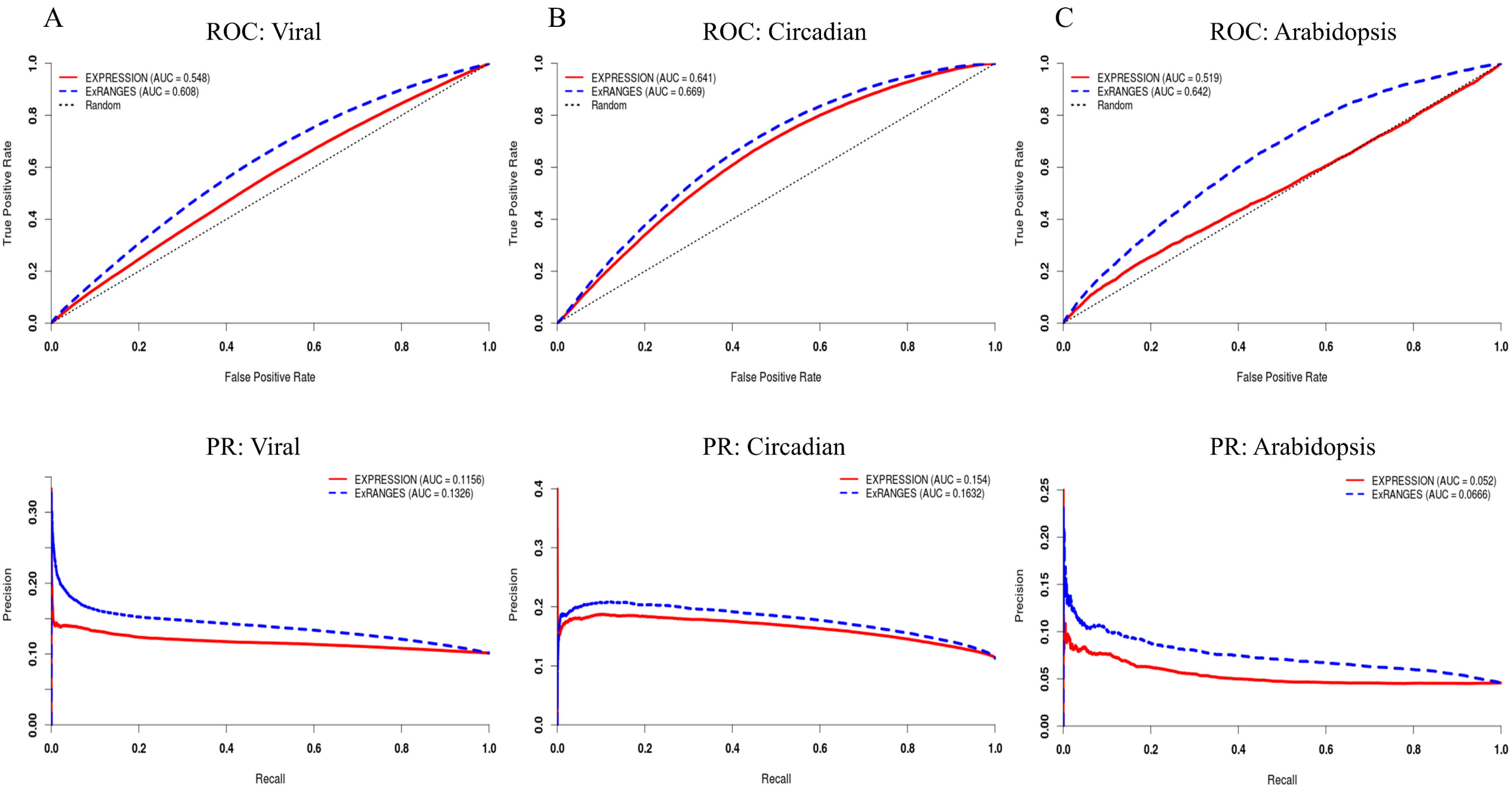


C

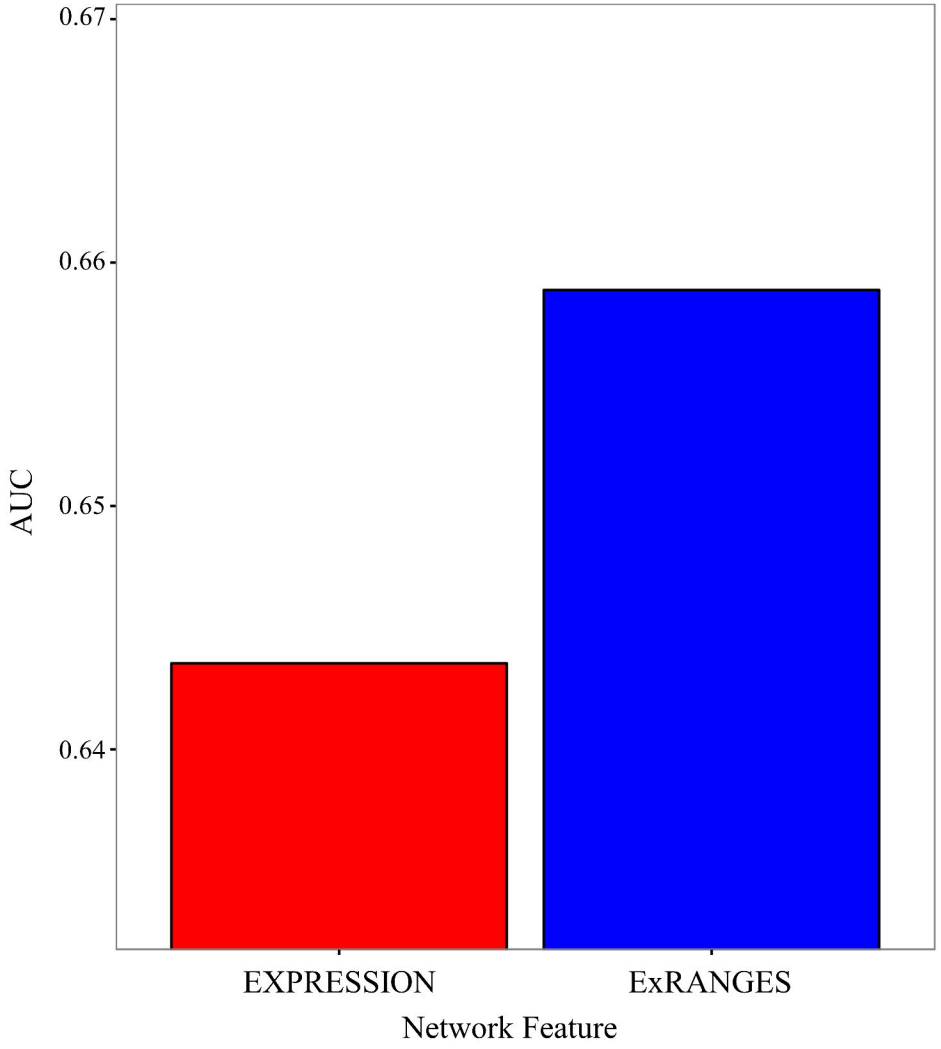
GO Enrichment Difference



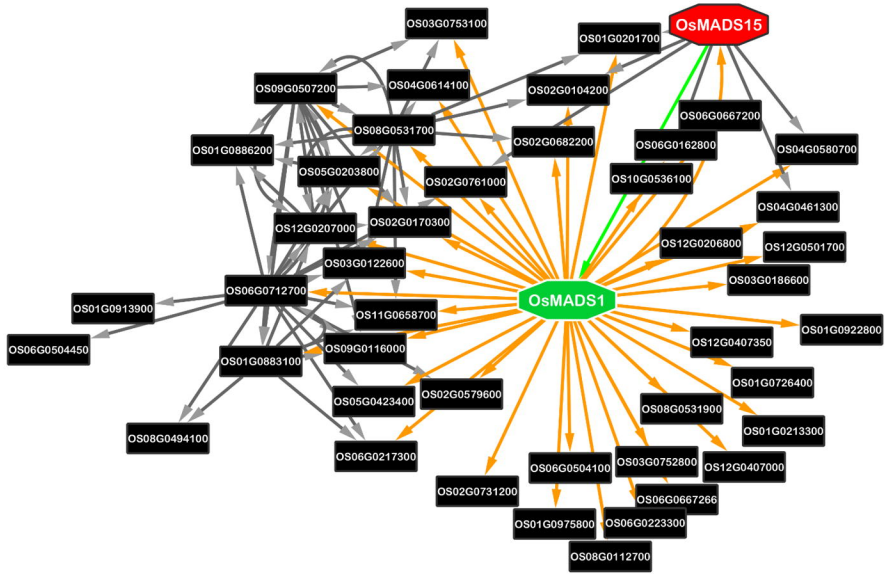

A
Area Under ROC Yeast
AUROC
0.575
0.550
0.525
0.500
0.475
EXPRESSION
ExRANGES
Network Feature
B
Area Under ROC
Arabidopsis ChIP-Seq
TF
AP2
CCA1
APRR7
EIN3
HY5
0.80
0.70
0.60
0.50
EXPRESSION
ExRANGES
Network Feature
C
Area Under ROC
Arabidopsis DAP-Seq
0.65
0.60
0.55
0.50
0.45
0.40
EXPRESSION
ExRANGES
Network Feature


A
ROC: Viral
EXPRESSION (AUC = 0.548)
ExRANGES (AUC = 0.608)
Random
True Positive Rate
False Positive Rate
B
ROC: Circadian
EXPRESSION (AUC = 0.641)
ExRANGES (AUC = 0.669)
Random
True Positive Rate
False Positive Rate
C
ROC: Arabidopsis
EXPRESSION (AUC = 0.519)
ExRANGES (AUC = 0.642)
Random
True Positive Rate
False Positive Rate
PR: Viral
EXPRESSION (AUC = 0.1156)
ExRANGES (AUC = 0.1326)
Precision
Recall
PR: Circadian
EXPRESSION (AUC = 0.154)
ExRANGES (AUC = 0.1632)
Precision
Recall
PR: Arabidopsis
EXPRESSION (AUC = 0.052)
ExRANGES (AUC = 0.0666)
Precision
Recall

A Genie3: True-Positive Targets of OsMADS1 identified by EXPRESSION and ExRANGES



B Inferelator: Targets of OsMADS1 identified by EXPRESSION and ExRANGES

	True Positives	False Positives	TP/FP
ExRANGES	11	50	.22
EXPRESSION	1	10	.1

C ExRANGES Rice MADS TF Network