Incorporating the Rate of Transcriptional Change Improves Construction of Gene Regulatory Networks

Short Title: Incorporating Rate Change into Gene Regulatory Networks

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

Department of Molecular and Structural Biochemistry, North Carolina State University, Raleigh, NC, United States of America

Department of Biological Sciences, University of California, San Diego, La Jolla, CA, United States of America

Central Infrastructure Group Genomics and Transcript Profiling, Max Planck Institute of Molecular Plant Physiology Potsdam, Germany

International Rice Research Institute (IRRI), DAPO Box 7777, Metro Manila, Philippines

Department of Agronomy, Kansas State University, Manhattan, KS, United States of America

* Corresponding author

E-mail: colleen_doherty@ncsu.edu (CJD)

Abstract

Transcriptional regulatory networks (TRNs) can be developed by computational approaches that infer regulator-target gene interactions from transcriptional assays. Successful algorithms that generate predictive, accurate TRNs enable the identification of regulator-target relationships in conditions where experimentally determining regulatory interactions is a challenge. Improving the ability of TRNs to successfully predict known regulator-target relationships in model species will enhance confidence in applying these approaches to determine regulator-target interactions in non-model species where experimental validation is challenging. Many transcriptional profiling experiments are performed across multiple time points; therefore we sought to improve regulator-target predictions by adjusting how time is incorporated into the network**.** We created **ExRANGES,** which incorporates **Expression in a Rate-Normalized GEne Specific** manner that adjusts how expression data is provided to the network algorithm. We tested this on a two different network construction approaches and found that ExRANGES prioritizes targets differently than traditional expression and improves the ability of these networks to accurately predict known regulator targets. ExRANGES improved the ability to correctly identify targets of transcription factors in large data sets in four different model systems: mouse, human, Arabidopsis, and yeast. Finally, we examined the performance of ExRANGES on a small data set from field-grown *Oryza sativa* and found that it also improved the ability to identify known targets even with a limited data set.

Author Summary

In model organisms, the ability to identify direct targets of transcription factors (**TF**s) via high throughput experimental assays has advanced our understanding of transcriptional regulatory networks and how organisms regulate gene expression. However, for non-model organisms, it remains a challenge to identify TF–target relationships through experimental approaches such as ChIP-Seq, thus limiting the ability to understand regulatory control is limited. Computational approaches to identify regulator-target relationships *in silico* from easily attainable transcriptional data offer a solution. Most algorithms for identifying gene regulatory networks from time series data weigh the relationship between regulators and putative targets at all time points equally. However, many regulators may control a single target in response to different inputs. Our approach, ExRANGES, focuses on time points where there is a significant change in expression to identify the association between regulators and targets. ExRANGES essentially weights the expression value of each time point by the slope change after that time point, thereby emphasizing the relationship between regulators and targets at the time points when the transcript levels are changing. We show that this change to the way expression data is included into gene regulatory network algorithms improves the identification of regulator-target interactions and we hope this will improve *in silico* identification of regulatory relationships in many species.

Introduction

Transcriptional regulatory networks provide a framework for understanding how signals are propagated throughout the transcriptome of an organism. 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 [1]. 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 transcription factor (TF) regulators and the target genes they control.

Experimental approaches such as 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 [2]. 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 [3–6]. The assumption of these approaches is that there is a correspondence between the expression of the regulator gene and its targets that can be discerned from RNA levels. Therefore, given sufficient variation in expression, the targets of a given factor can be predicted based on associated changes in expression. Initial approaches focused on the correlation between regulators and targets such that activators are positively correlated and repressors are negatively correlated with their target expression levels. These approaches have been successful in identifying some relationships [7].

More recent methods improved the ability to identify connections between regulators and targets even in sparse and noisy data sets [4–6,8–10]. The DREAM5 challenge compared many methods for their ability to identify transcriptional regulatory networks from gene expression datasets [11]. One of the top performing methods was GENIE3 [8]. This method identifies targets for selected regulators by taking advantage of the regressive capabilities of the random forest machine learning algorithm [12] and [13]. Other successfully implemented approaches include SVM [3], CLR [6], CSI [14,15], ARACNE [5], Inferelator [4], and DELDBN [9]. Common to these methods is the use of the transcript abundance levels to evaluate the relationship between a regulator and its putative targets. However, correlation between expression levels alone may not utilize all information available in time series data. Many approaches have been developed that take advantage of the additional information available from time series data [reviewed in [16,17].

Here we present an approach that expands upon these existing algorithms by using the rate of change between consecutive time points to emphasize the relationships between regulator and targets at times when expression is significantly changing. We predict that: 1) Focusing on the rate of change will utilize different characteristics in the data and identify different regulatory relationships than using the expression values. 2) Combining expression level and the rate of change will result in improved identification of true regulatory relationships.

We first evaluated the effects of incorporating the rate of change, and developed RANGES RAte Normalized in a GEne Specific manner to evaluate the significance of the rate changes at each consecutive time point. This approach has a similar recall rate to using expression values alone, but identifies a distinct set of true-positive targets. We then combined the expression and slope change in ExRANGES (Expression by RANGES) to emphasize the connections between regulators and targets at time points before a significant change in gene

expression. ExRANGES improves 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], but anticipate that it will offer a similar benefit to when combined with other network inference algorithms.

Results

RANGES Identifies Significant Changes in Rate of Expression

We hypothesized that for experiments measuring RNA levels across multiple time points incorporating the rate of change between consecutive time points would identify regulator -target relationships missed by comparing expression values alone. If a gene is changing in expression at only a few time points across a data series these time points may be more important samples for considering the relationship between potential regulators of that gene than time points where the target is expressed at a stable level. Therefore, we developed an approach that evaluates the rate of change of target genes across all consecutive time points and weights the change between each consecutive time point based on the background variance observed across the dataset for each gene. We predict that this approach focuses the comparison between regulatory factors and their targets to the time points where the effects of active regulation can be observed based on changes in RNA levels and will therefore identify regulatory relationships not detected by comparing expression values alone.

The first step in incorporating the rate of change into the identification of regulatory networks is to distinguish significant rate changes from normal variation between time points caused by sampling or measurement error. Our method determines the significance of the

change in expression between two consecutive time points on a per gene basis enabling us to assess the significance of the change at each time step for a given gene. For each gene, we quantified the significance of the change in expression at a given time point by estimating a p-value for the change in expression between the consecutive time points under evaluation against the background of all possible time steps. The background was constructed from the change in expression at all consecutive time steps in all samples across all experiments from a given data set (Fig 1A). For example, if we consider the mammalian circadian data set available from CIRCADB [18], the data set consists of time series experiments from 12 different tissues, sampled every 2 h for 48 h (288 samples). Therefore, the change in expression levels between time t and time $t+1$ can be 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. We defined the background as the change in expression for each consecutive time interval across the entire time series. For this data set, the background consists of 288 slopes (12 tissues x 24 time points) for each gene. At each time step, t the slope between t and $t+1$ was compared to this background a p-value is estimated. This was done for each gene and the resulting p-value was transformed to the negative log 10 and the sign of the change in slope was preserved (R script provided). We call this value RANGES (for RAte Normalized in a GEne Specific manner). The RANGES value was used in lieu of the expression in generation of a regulatory network using GENIE3 [8]. We considered 1690 TFs as the regulators [19]. To determine the potential for the rate of change to identify targets of each TF we compared RANGES to the standard approach of using expression values (hereinafter after called EXPRESSION). For the EXPRESSION approach, the input into the regression analysis included the expression values across the 288 samples for each of the 1690 TFs as regulators and the expression values of 35,556 genes as

168 potential targets across the same samples. For the RANGES approach, the $-\log_{10}$ of the p-value
169 for the significance of each change in time across the 24 time steps was used as the input for both
170 the 35,556 targets and 1690 regulators. For both approaches all TFs were also included in the
171 target list to identify regulatory connections between TFs.

172 To evaluate the ability of each approach to correctly identify targets of the TFs, we
173 compared the resulting targets of each TF identified by either the RANGES or EXPRESSION
174 approach with the targets identified by ChIP-Seq for five TFs involved in circadian regulation
175 where three replicates of each ChIP-Seq experiment were performed: PER1, CLOCK, NPAS2,
176 NR1d2, and ARNTL [20,21]. Targets identified by each approach that were considered
177 significant targets by these published ChIP-Seq experiments were scored as true positive targets
178 of that TF.

179

180 **RANGES and EXPRESSION Values Identify Different Sets of True Positive Targets**

181 We compared the targets identified by using RANGES to those identified using EXPRESSION.
182 For PER1 both approaches identified true targets more than would be expected by chance (ROC
183 curve, Fig S1). EXPRESSION showed a larger area under the ROC curve, indicating higher
184 accuracy in identifying true positive targets of PER1. However, there was little overlap in the
185 top true positive targets identified by each approach (Fig 1B). Many genes that were scored
186 strongly by RANGES as PER1 targets, including many true positive targets of PER1, had low
187 scores when evaluated using the EXPRESSION approach. Likewise, several of the top scoring
188 true positive targets by EXPRESSION had low RANGE scores. This difference in the targets
189 identified by each approach, including true positives, was also observed for the other four TFs
190 we evaluated (Fig S2). These results indicate that information contained in the relationship

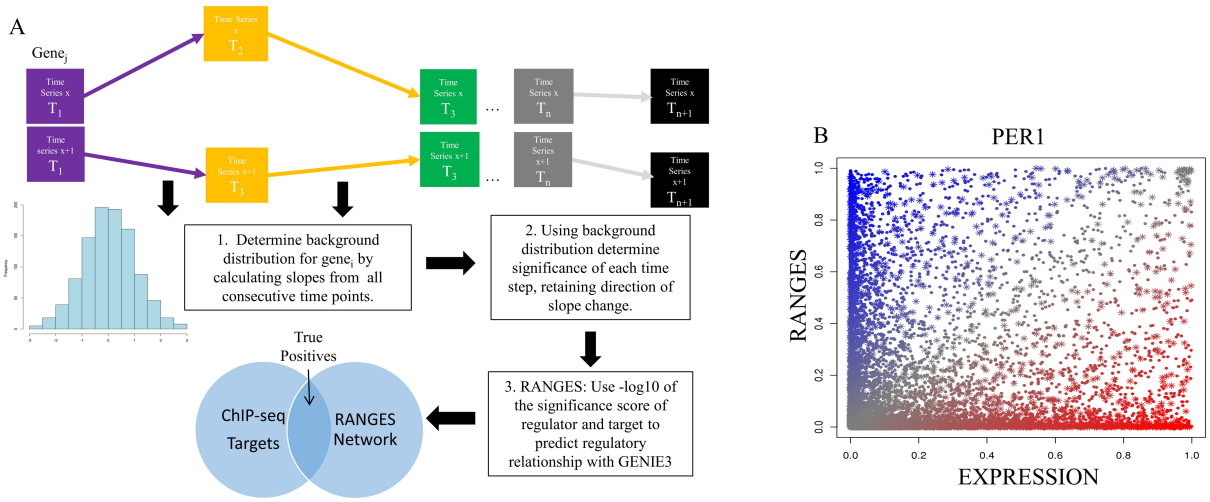


Figure 1: A) Overview of RANGES approach. For each Gene$_i$, the slope is calculated for all possible consecutive changes time points. From this background distribution of changes in expression the significance of each time point change is calculated. The $-\log_{10}$ of the p-value is calculated and the sign change of direction is preserved. In the RANGES approach, this significance value is used as the input into network inference using GENIE3 [8] for both the transcription factor (TF) regulators and targets. For the EXPRESSION approach, the expression values at each time point are provided for both the regulator and target to GENIE3. The predictive ability of each approach was compared to the targets experimentally identified for each TF by ChIP-Seq. B) Targets identified by RANGES and EXPRESSION approaches show little overlap. Scatter plot of targets of PER1 as identified by EXPRESSION or RANGES approaches. PER1 targets identified with similar rank by both approaches are shown in grey. PER1 targets identified as high ranking by RANGES are shown in blue and those ranking higher by EXPRESSION are red. PER1 targets identified by ChIP-Seq [20] are marked as stars. Genes identified as PER1 targets by each approach that were not identified in the ChIP-Seq identified targets are plotted as points.

between the rate of change of the TF and target identifies TF-target relationships missed by analyzing expression levels alone.

Rate Change Identified Samples with Lower Variation Between Tissues

To understand why some targets are identified by EXPRESSION only and others by RANGES only we compared the expression of the top predicted PER1 targets for each method (Fig 2A).

We observed that the top hits identified by EXPRESSION showed more variation between each tissue than in those identified by RANGES. We therefore examined the variance between each tissue by calculating the variance of the mean expression for each of the 12 tissue samples for the top 1000 targets for all five of the TFs with ChIP-Seq data available (Fig 2B) [20]. As observed for the top PER1 targets, the targets identified by EXPRESSION generally showed more variation between tissues than the targets identified by RANGES. We also examined the within tissue variation to evaluate how well each approach identified targets that show a range of expression throughout the day within each time series (Fig 2C). The targets identified by RANGES showed more variation in the time series within each tissue suggesting that this approach might be more sensitive to changes that are dependent on the rate of expression as we would expect for this rate-based approach. To evaluate if the increased variance within each tissue observed for top TF targets identified by the RANGES approach is limited to circadian associated TFs, we compared the between tissue and within tissue standard deviation for the top 1000 targets identified by EXPRESSION or RANGES for all 1690 TF regulators (Figs 2D and E). As we observed for the circadian TFs, the targets identified by EXPRESSION showed more variation between tissue types (Fig 2D). The RANGES approach was able to identify targets with increased variation within each tissue time series compared to the EXPRESSION approach (Fig 2E).

We also compared the mean intensity level of the top 1000 predicted targets of the RANGES 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 (Fig S3A). In contrast, the top 1000 predicted targets of PER1 identified by RANGES resembled the background distribution of intensity for

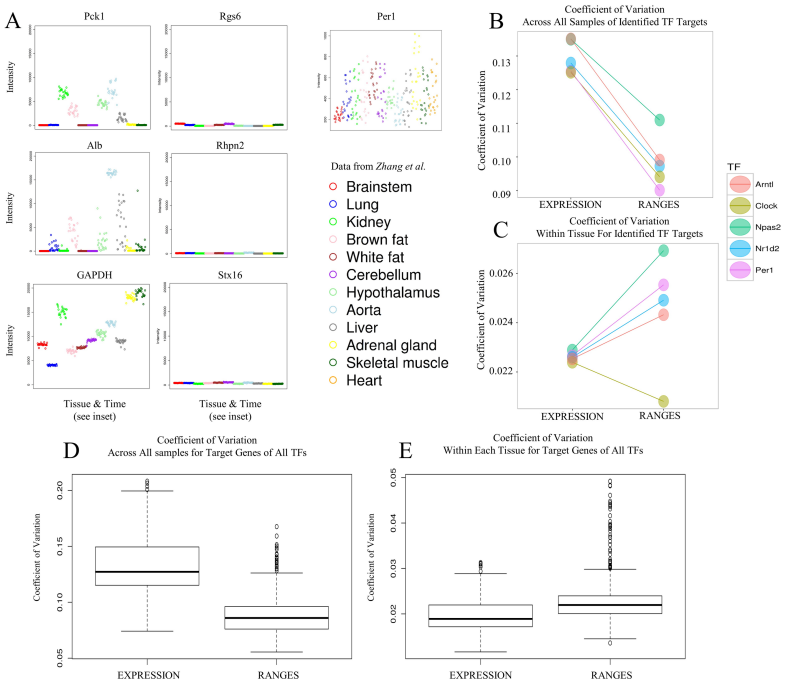


Figure 2: A) Top targets for RANGES and EXPRESSION show different expression features. The expression values of the top three targets of PER1 identified by EXPRESSION (left; Pck1, Alb, GAPDH) and RANGES (right; RGS6, Rhpn2, Stx16) across the two day time series performed in twelve tissues [18]. The order of the tissues are: Brainstem, Lung, Kidney, Brown fat, White fat, Cerebellum, Hypothalamus, Aorta, Liver, Adrenal gland, Skeletal muscle, and Heart. Each tissue is plotted side by side in different colors. The points within a tissue represent expression levels every 2 hours over 48 hours. PER1 expression is shown in the center for comparison. B) Targets of the circadian TFs identified by EXPRESSION show higher standard deviation in expression levels across all samples than targets identified by RANGES. The standard deviation across all samples for the top 1000 targets of each circadian TF (ARNTL, CLOCK, NPAS, NR1D2, and PER1) identified by either the EXPRESSION or RANGES approach. C) Most circadian TF targets identified by RANGES show higher within tissue standard deviation. The standard deviation across the time series for each individual tissue was calculated for the top 1000 targets of each circadian TF (Arntl Clock, Npas, Nr1d2, and Per1) identified by either the EXPRESSION or RANGES approach. The mean of these within tissue standard deviations is plotted. D) EXPRESSION identified TF targets show greater variation in expression across all samples. Box plot showing the standard deviation of the top 1000 targets of all TFs identified by either EXPRESSION or RANGES. E) RANGES identified TF targets show greater within tissue variation. The standard deviation was calculated for each time series in each tissue for the top 1000 targets of all TFs identified by EXPRESSION or RANGES. Boxplot showing the mean standard deviation for each tissue for these top targets.

all the transcripts on the array (Fig S3B). Likewise, the hybridization intensity of the genes identified as the top 1000 targets identified by EXPRESSION of all 1690 TFs considered as

222 regulators was shifted higher compared to the background distribution levels (Fig S3C). While
223 the top 1000 targets of all 1690 TFs identified by RANGES reflected the background distribution
224 of hybridization intensity (Fig S3D). While hybridization intensity cannot directly be translated
225 into expression levels, these observations suggest that there are features of the targets identified
226 by RANGES that are distinct from those identified by EXPRESSION. We hypothesized that
227 combining these two approaches would improve the overall ability to detect true positive targets
228 of each regulator.
229

230 **ExRANGES Combines Rate Change with Expression Levels**

231 Since many of the true positive targets of the TFs we evaluated identified by RANGES were not
232 identified by EXPRESSION and visa versa, we hypothesized that combining these two would
233 improve the overall ability to predict true positive targets. To combine these approaches we took
234 the product of the expression at time point t by the RANGES p-values for the change in
235 expression from time point t to $t+1$ for each target (ExRANGES) (Fig 3A). This adjusts each
236 time point by the rate of change in the following time interval. Therefore, the value of the time
237 point preceding a significant change in expression is higher than the value of a time point when
238 the following expression remains unchanged. We anticipate that this will enhance the signal
239 between the regulator and target for the time points where regulation is occurring, thus
240 improving the ability to correctly identify targets of each TF. For the regulators, only the
241 expression value of the TF was provided. For all targets, this ExRANGES value was provided to
242 GENIE3. All TFs were also considered as potential targets and the ExRANGES value was used
243 in the target matrix for all TFs.

Using the identified ChIP-Seq targets as true positives from Koike et al. [20], we calculated the area under the ROC curve to compare the identification of true targets attained by EXPRESSION to the combination of expression and p-values using ExRANGES. We observed that for all five TFs there was an improvement in the ability to identify ChIP-Seq targets (Fig 3B).

A modification of GENIE3 uses a time delay to identify transcriptional changes in the regulator that precedes the effects on the target by a defined time step as incorporation of a delay between regulator expression and target expression has previously been shown to improve the ability to identify regulatory networks [22]. We compared our approach to this modified implementation of GENIE3 that includes the time delay step. As previously reported, we observed that the time step delay improved target identification for some transcription factors, 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 very small (Fig 3B).

The ExRANGES Approach 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. To test ExRANGES performance across tissue types, we selected seven TFs in our regulator list that have available ChIP-Seq data from at least two experimental replicates performed in epithelial cells, a tissue not included in the circadian time series samples. The

267 seven TFs that we tested are: ESR1, STAT5A, STAT5B, POL2A, FOXA1, TFAP2A, and CHD4
268 [23]. We observed improvement of the area under the ROC curve for five of the seven TFs
269 (ESR1, POL2A, FOXA1, TFAP2A, and CHD4) by combining expression and rate change
270 information using ExRANGES (Fig 3C). As we observed above for CLOCK, STAT5A and
271 STAT5B performed equally well, but did not show significant improvement. STAT5A and
272 STAT5B are known to be activated post-transcriptionally perhaps indicating why evaluating the
273 change in expression of these TFs did not lead to improved identification of targets [24–29].
274 This suggests that for TFs that show little variation in expression throughout the day in each time
275 series the addition of the RANGES component may not offer much improvement. (Fig S4).
276

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

278 Although circadian and diel time series experiments are a rich resource providing substantial
279 variance for identifying regulatory relationships, most available experimental data is not
280 collected with this design. Often sample collection cannot be controlled precisely to attain
281 evenly spaced time points. For example, in human studies, the subject may not be available for
282 consistent sampling. To evaluate the ability of ExRANGES to identify true targets of TFs across
283 unevenly spaced and heterogeneous genotypes, we analyzed expression studies of viral
284 infections in various individuals [30,31] using both ExRANGES and EXPRESSION approaches.
285 This data set consists of a series of blood samples from human patients taken over a seven to
286 nine day period, depending on the specific study. Sampling was not evenly spaced between time
287 points. Seven studies that each sampled multiple individuals before and after respiratory
288 infection are included. In total 2372 samples were used, providing a background of 2231
289 consecutive time steps. Overall, the variance between samples was lower for this study than the

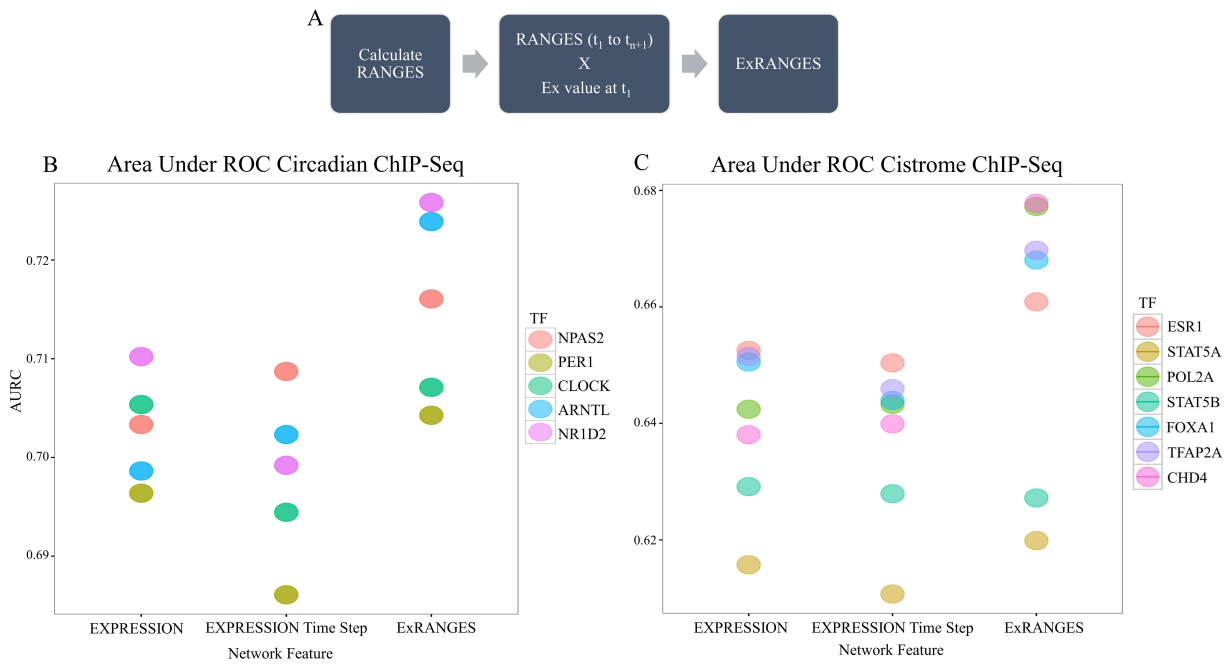


Figure 3: ExRANGES combines EXPRESSION and RANGES approach. A) Schematic of how ExRANGES combines expression value and slope change. B) ExRANGES outperforms EXPRESSION. The targets of the circadian TFs (ARNTl CLOCK, NPAS, NR1D2, and PER1) identified by EXPRESSION or RANGES were validated against the ChIP-Seq identified targets for these TFs [20]. Area under the ROC Curve (AURC) is plotted for targets identified by EXPRESSION, EXPRESSION where the GENIE3 algorithm included a time step, and ExRANGES (without a time step). C) ExRANGES improves identification of ChIP-Seq validated targets in TFs that are not core components of the circadian clock. The EXPRESSION and RATE identified targets of seven TFs with ChIP-Seq identified targets available from CISTROME that are in our list of TFs and ChIP-Seq performed in epithelial cells which is a tissue not sampled in the circadian time series [23] were compared and the area under the ROC curve (AURC) is plotted. ExRANGES showed increased AURC for five of the TFs (ESR1, POL2A, FOXA1, TFAP2A, and CHD4) over EXPRESSION or EXPRESSION including a time step. For STAT5A and STAT5B ExRANGES did not increase the AURC.

circadian study examined above (Figure 4A). 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). For the 83 TFs on the HGU133 Plus 2.0 microarray (Affymetrix, Santa Clara, CA) with ChIP-Seq data from blood tissue [32], we observed an overall improvement in the detection of ChIP-Seq identified targets (Fig 4B). The improvement varies by TF (Fig 4C).

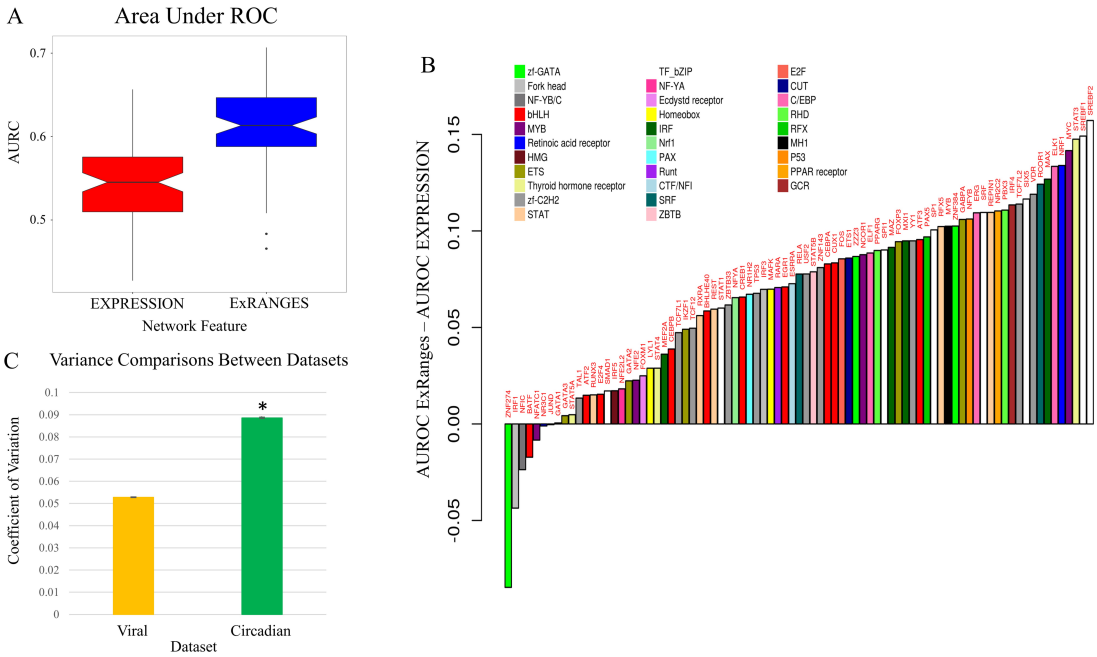


Figure 4: A) ExRANGES improves identification of TF targets in unevenly sampled and heterogeneous data. Targets of 83 TFs where ChIP-Seq data is available from Cistrome [23] were compared for EXPRESSION and ExRANGES. Predictions of targets from EXPRESSION and ExRANGES were compared to ChIP-Seq identified targets and the results for all 83 TFs are presented as a box plot of area under the ROC curve (AURC). B) Variance comparison of the viral and circadian data sets. The index of dispersion is calculated by dividing the variance of each gene by its mean expression level and taking the mean of these values over all genes in the dataset. The circadian data set showed a significantly higher Index of Variation than the viral data set (Student's t-test, p-value < 10^{-15}). Improvement observed in ExRANGES identified targets varies across the 83 TFs tested. The difference between area under ROC curve of ExRANGES and EXPRESSION is plotted in ascending order for the 83 TFs tested. TFs are colored by TF family.

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 of predicted targets for a given regulator. 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-

303 Seq data. We compared the functional enrichment of the top 1000 targets of each TF predicted
304 by either approach using *Homo sapiens* GO slim annotation categories. We evaluated the 930
305 TFs on the HGU133 microrarray [19]. Of these, the targets identified by ExRANGES for the
306 majority of the TFs (590) showed improved functional enrichment compared to the targets
307 identified by EXPRESSION (Fig 5A and B). Likewise, when focusing on the 83 TFs with
308 available ChIP-Seq data from blood, the majority of TF targets predicted by ExRANGES were
309 more functionally cohesive compared to EXPRESSION targets as evaluated by GO slim (Fig
310 5C). We observed that the improvement ranking of ExRANGES over EXPRESSION varies
311 between the two validation approaches. For example, targets of the TF JUND identified by
312 ExRANGES show no improvement over EXPRESSION when validated by ChIP-Seq identified
313 targets, yet showed improved functional cohesion (Supplemental Table ST1).
314

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

317 The previous evaluations of ExRANGES were performed on expression data obtained from
318 microarray-based studies and true positives were based on ChIP-Seq identified targets of each
319 TF. To evaluate the performance of ExRANGES compared to EXPRESSION for RNA-Seq data
320 we applied each approach to an RNA-Seq data set performed in *Saccharomyces cerevisiae.* This
321 data set consisted of samples collected from six different genotypes every fifteen minutes for six
322 hours after transfer to media lacking phosphate. The slope background was calculated from 144
323 time steps. To evaluate the performance of ExRANGES compared to EXPRESSION approaches
324 we calculated the area under the ROC curve for the identified targets for each of the 52 TFs
325 using the TF targets identified by protein binding microarray analysis as true positives [33]. For

A GO Enrichment Score Table

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

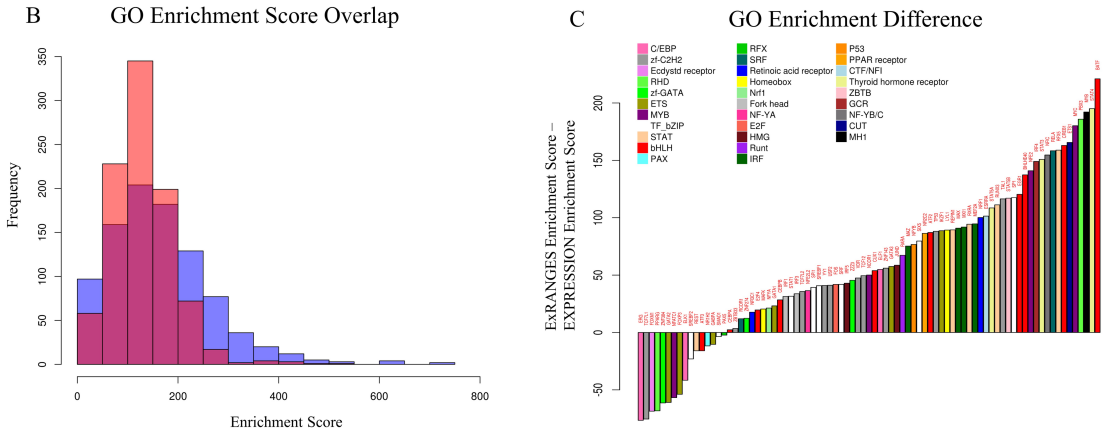


Figure 5: Functional Enrichment of ExRANGES identified targets. Gene Ontology enrichment was calculated using *Homo sapiens* GO slim annotations for the top 1000 targets of each TF predicted by either ExRANGES or EXPRESSION. The background annotations were limited the genes present on the HGU133 microarray. Enrichment score is the sum of the $-\log_{10}$ of the p-value of each GO category. A) Summary table of the enrichment scores for the top targets of all 930 TFs on the microarray. B) The distribution of enrichments scores from EXPRESSION targets (red) and ExRANGES targets (blue). C) Enrichment score difference of the 83 TFs with available ChIP-Seq data (Fig 4). Positive values indicate TF targets with a higher enrichments score in ExRANGES compared to EXPRESSION.

most TFs, the AUC was improved by the use of ExRANGES compared to EXPRESSION (Fig 6A).

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. 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 [34–38]. We observed that for all five TFs ExRANGES showed improved

identification of the ChIP-based true positive TF targets (Fig 6B). To evaluate a larger range of targets we compared our predicted targets by EXPRESSION or ExRANGES to 307 TFs targets identified by DAP-Seq [39]. We observed that ExRANGES also showed an improved ability to identify targets as validated by DAP-Seq compared to EXPRESSION (Fig 6C).

Application of ExRANGES to Smaller Data Sets with Limited Validation Resources

Time series data offers several advantages, however the expense is also significantly increased. We have shown that using ExRANGES in conjunction with GENIE3 improves performance on large data sets as validated by ChIP-Seq (228 samples in mouse, 2372 in human, and 144 in arabidopsis) (Fig 7). We also compared the use of the ExRANGES approach to EXPRESSION

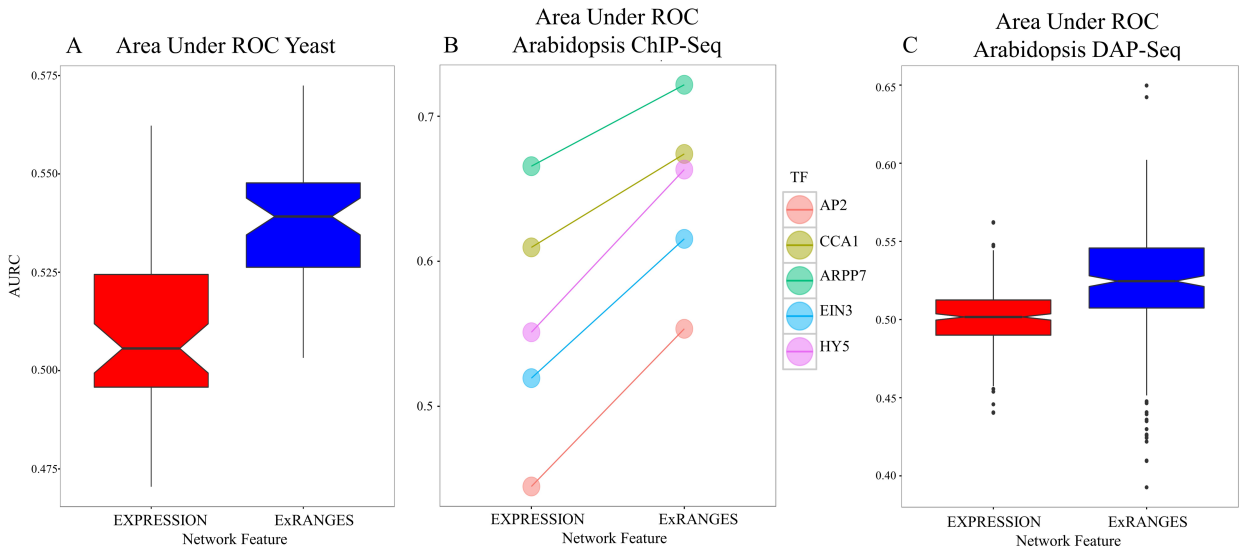


Figure 6: ExRANGES improves identification of TF targets validated by different methods. A) Targets identified for 52 yeast TFs by ESPRESSION (red) and ExRANGES (blue) were validated against the targets identified for each TF using a protein binding microarray [33] and boxplots generated from the area under the ROC curve (AURC). B) AURC for targets of the five Arabidopsis TFs with replicated ChIP-Seq data available for EXPRESSION and ExRANGES identified targets. C). AURC for targets identified for 307 TFs by EXPRESSION (red) and ExRANGES (blue) as validated against DAP-Seq identified targets [39].

alone with the INFERELATOR algorithm, although ExRANGES showed an improved AUROC in all three data sets; the largest increase observed was in the Arabidopsis data set, which has the lowest sample number (Fig S4). 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 time series consisting of 28 samples from seven unevenly sampled time points of field grown rice data. ChIP-Seq has only been performed for one transcription factor in rice, OsMADS1 [40]. Therefore, we compared the ability of ExRANGES and EXPRESSION to identify the

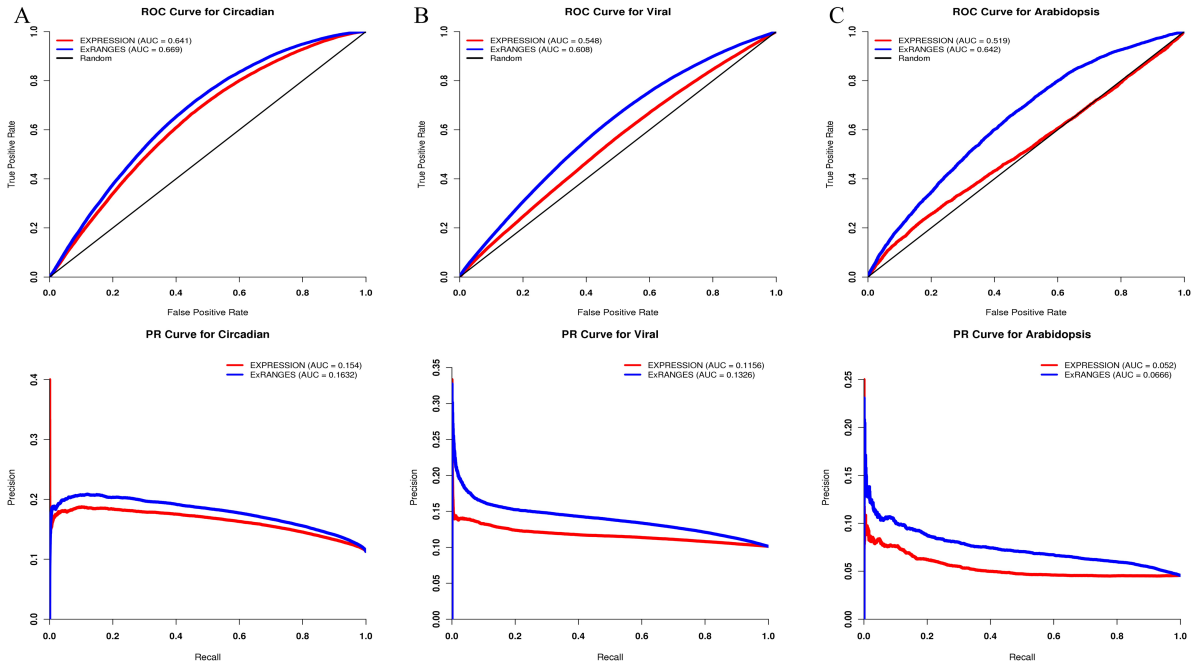
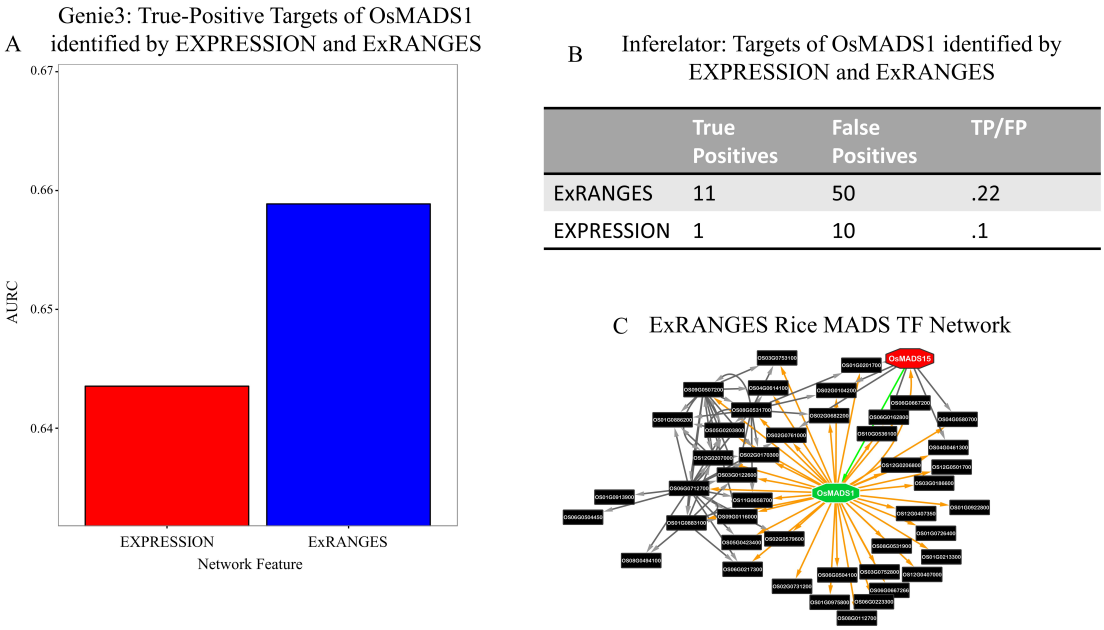


Figure 7: Summary of ExRANGES improvement across three data sets from different species. ROC and Precision recall (PR) curves for targets of all ChIP-Seq validated TFs as identified by EXPRESSION (red) or ExRANGES (blue) for A) Circadian dataset from different mouse tissue samples B) Viral data set C) Circadian dataset from Arabidopsis across different environmental variables.



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

Figure 8: ExRANGES retains performance improvement on small data sets. A) Area under the ROC curve (AURC) for the top 1000 targets of OsMADS1 identified by EXPRESSION (red) or ExRANGES (blue) using GENIE3 and validated against the OsMADS1 ChIP-Seq data. B) Comparison of targets identified by EXPRESSION and ExRANGES using INFERELATOR. True positives indicated the number of OsMADS1 targets identified by each approach that were also detected by ChIP-Seq, False positives indicated the number predicted by each approach that were not identified as targets of OsMADS1 by ChIP-Seq. C) Network of MADS TFs predicted by ExRANGES. Interactions with OsMADS1 (green) determined by ExRANGES with other MADS TFs in rice are visualized as an interaction network. ExRANGES predicted targets of OsMADS1 are indicated in orange [40]. OsMADS15 (red) is predicted to regulate OsMADS1 by ExRANGES (green arrow). Interactions between other MADS TFs predicted by ExRANGES are indicated by black arrows.

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 (Fig 8) compared to EXPRESSION.

Discussion

Computational approaches that can identify candidate targets of regulators can advance research. Many approaches have been developed to identify regulator targets, but most of these use

362 expression values. We have demonstrated that combining the expression levels and rate of
363 change improves the ability to predict true targets of TFs across a range of species and
364 experimental designs. This approach improves the identification of targets as determined by
365 ChIP-Seq and protein binding microarray across many different collections of time series data
366 including experiments with replicates and without, with time series that have unevenly sampled
367 time points, and even for time series with limited number of samples. ExRANGES provides
368 improvement in TF target identification over EXPRESSION values alone for time series
369 performed with both microarray and RNA-Seq measurements of expression.

370 Expression analysis performed in time series, such as experiments evaluating the
371 transcriptional changes throughout a circadian cycle, provide rich resources for identifying
372 relationships between individual transcripts. Since in many species the majority of transcripts
373 show variation in expression levels throughout the day [18,41,42] circadian and diel data sets
374 provide a snapshot of the potential ranges in expression that a regulator can attain. The
375 associated changes in target expression levels can be analyzed to identify potential regulatory
376 relationships that may be enhanced in response to other perturbations such as stress. Here, we
377 show that data sets that combine circadian time series in multiple tissues can be a powerful
378 resource for identifying regulatory relationships between TFs and their targets not just for
379 circadian regulators, but also for regulators that are not components of the circadian clock.
380 Targets identified using EXPRESSION as the features were those that showed large variance
381 between tissue, while RANGES identified targets that showed larger variance within each time
382 series. ExRANGES takes advantage of both sources of variation and improves the identification
383 of TF targets for most regulators tested, including for TF-target relationships in tissues not

384 included in the transcriptional analysis. Additionally, ExRANGES simplifies incorporation of
385 replicate samples.

386 As implemented, ExRANGES improves the ability to identify regulator targets, however,
387 there are many aspects that could be further optimized. For example, we tested ExRANGES
388 with the network inference algorithm GENIE3 and demonstrated that it improved the
389 performance of this algorithm. The ExRANGES method can be applied to most other machine
390 learning applications such as Bayesian networks, mutual information networks, or even
391 supervised machine learning tools. In addition, we showed that ExRANGES outperformed a
392 one-step time delay. Conceptually, our method essentially increases the weight of the time point
393 before a major change in expression level. ExRANGES could be further modified to adjust
394 where that weight is placed, a step or more in advance, depending on the time series data. Such
395 incorporation of a time delay optimization into the ExRANGES approach could lead to further
396 improvement for identification of some TF targets, although it would increase the computational
397 cost.

398 Here, we compared ExRANGES based features to EXPRESSION based features by
399 validating against TF targets identified by ChIP-Seq, ChIP-Chip, DAP-Seq, and protein binding
400 microarray. While these experimental approaches identify potential TF targets in a genome-wide
401 manner, they are not perfect as gold-standards for validation of transcriptional regulatory
402 networks. If there are systematic errors in target identification by ChIP-Seq, ExRANGES may
403 perform better than indicated here. Although ChIP-Seq may not be an ideal gold standard, it
404 does provide a benchmark for comparing computational approaches to identifying TF targets.
405 Unfortunately, high quality ChIP-Seq data is not available in most organisms for more than a
406 handful of TFs. For example, validation of this approach in rice was limited to one recently

407 published ChIP-Seq dataset. This lack of experimentally identified targets is a severe hindrance
408 to advancing research in these species. New experimental approaches such as DAP-Seq may
409 provide alternatives for TF target identification in species recalcitrant to ChIP-Seq analysis [39].
410 Additionally, the authors of this paper improved their recall of ChIP-Seq identified targets by
411 selecting targets that were also supported by DNase-Seq sensitivity assays [43,44]. Likewise,
412 distinguishing between direct and indirect targets predicted computationally could be enhanced
413 by incorporation of DNase-Seq or motif occurrence information for the targets. Incorporation of
414 such *a priori* information on regions of open chromatin and occurrence of *cis*-regulatory
415 elements leads to improved network reconstruction [10,45]. Use of ExRANGES could lead to
416 improvement for these integrated approaches. Although approaches such as DAP-Seq are more
417 global in analyses than individual ChIP-Seq assays, these genome-wide approaches still require a
418 significant investment from the community in the development of an expressed TF library
419 collection. For non-model systems, computational identification of TF targets can provide an
420 economical first pass that can be followed up by experimental analysis of predicted targets,
421 accepting the fact that there will be false positives in the validation pipeline. In this strategy, a
422 small improvement in the ability to identify true targets of a given TF can translate into a
423 reduced number of candidates to test and fewer experiments that must be performed. We hope
424 that the modest improvements to regulatory network algorithms provided by the ExRANGES
425 approach can facilitate research in species where identification of TF targets is experimentally
426 challenging. Additionally, we hope that our finding of how gene expression values are
427 incorporated in a network has a significant effect on the ability to identify regulatory
428 relationships will stimulate evaluation of new approaches that use alternative methods to
429 incorporate time signals into regulatory network analysis.

430 In summary, we demonstrate that consideration of how expression data is incorporated
431 can contribute to the success of transcriptional regulatory network reconstruction. ExRANGES is
432 a first step at evaluating different approaches for how features are supplied to regulatory network
433 inference algorithms. We anticipate that further optimization and other novel methods for
434 integrating expression information will lead to improvements in network reconstruction that
435 ultimately will accelerate biological discovery.
436

437 **Materials and Methods**

438 **Sources for Expression Data Sets**

439 Circadian Data Set

440 Normalized expression data from murine sources was downloaded from CircaDB [18].
441 Microarray-based expression levels from 288 samples were used in this study. The data
442 available was from twelve different tissues that were sampled every 2 h for 48 h.

443 Viral Data Set

444 The expression data used for the viral experimental analysis was downloaded from GEO
445 GSE73072. The data is composed of seven studies of individuals sampled before and after
446 respiratory infection. Expression is data from blood samples of approximately twenty
447 individuals taken over a seven to nine day period depending on the individual study. Sampling
448 was not evenly spaced between time points. In total data from 2372 microarrays were used. The
449 expression datasets used for the analyses described in this manuscript were contributed by Drs.
450 Ephraim Tsalik and Geoffrey Ginsburg from Duke University and the Durham VA Medical
451 Center. They were obtained as part of The Respiratory Viral DREAM Challenge through
452 Synapse ID syn5647810 [31].

453 S. *cerevisiae* RNA-Seq Data

454 RNA-Seq based expression data from *S. cerevisiae* was downloaded from GEO GSE61668 [46].
455 This data set evaluates phosphate starvation in six genotypes of *S. cerevisiae*. Transcript
456 expression was measured by RNA-Seq every 15m for six hours after transfer to reduced
457 phosphate media (150 samples total).

458 Arabidopsis Circadian Data

459 Normalized microarray expression data for Arabidopsis was obtained from
460 www.mocklerlab.org/diurnal [47]. This data set consisted of Arabidopsis plants of various ages
461 grown in 12 different environmental conditions sampled every 4 h for 48 h for a total of 144
462 samples.

463 *Oryza sativa* Diel Data

464 Rice variety IR64 was grown in the field at the International Rice Research Institute (,
465 Philippines). When the plants reached 50% flowering, panicle tissue was harvested at dawn,
466 dawn + 3.5h, dawn + 7h, dawn + 10.5h, dusk, dawn + 14h, dawn + 17.5h, and dawn + 21h. Four
467 replicates were harvested for each of these eight time points for a total of 32 samples. The third
468 rachis of the panicle was ground in liquid nitrogen with a metal pestle. The tissue was then
469 lyophilized at -60C°overnight. Total RNA was isolated using RNeasy Plant Mini Kit (Qiagen,
470 Germany) with the recommended RLT lysis buffer. The RNA extraction protocol was modified
471 to include and additional incubation with DNaseI. mRNA was isolated from 2 μg of total RNA
472 using magnetic oligo(dT) (NEB, Ipswich, MA). Directional RNA-Seq libraries were prepared
473 from isolated mRNA. Libraries were quantified using a 2100 Bioanalyzer (Agilent, Santa Clara,
474 CA) RNA-Seq was performed on a HiSeq 2500 (Illumina, San Diego, CA). Reads were trimmed
475 using seqtk (https://github.com/lh3/seqtk). Samples were aligned with Tophat2 to the IRGSP-1.0

476 genome [48,49]. Counts per gene were identified by HTSeq Count [50]. Data is available
477 through GSE92302.
478

479 **Selection of Regulators**

480 Transcription factors used as regulators for the murine circadian data and human viral data were
481 obtained from http://www.bioguo.org/AnimalTFDB/index.php [19]. Arabidopsis transcription
482 factor list were obtained from http://planttfdb.cbi.pku.edu.cn/ [51]. *S. cerevisiae* transcription
483 factors were obtained from [33].

484 **Sources for Validation Resources**

485 The direct targets for the five circadian TFs from murine data were obtained from the
486 supplementary information provided in [20]. Targets for additional TFs and the validation of the
487 viral data from human expression data were obtained from the cistrome project
488 (http://cistrome.org/Cistrome/Cistrome_Project.html) [23]. Eighty-three TFs were selected as
489 regulators that were labeled as evaluated blood tissue and present on the HGU133 microarray.
490 ChIP-Seq targets were determined by BETA (http://cistrome.org/BETA/). If multiple ChIP-Seq
491 were provided they were combined. The Arabidopsis ChIP-Seq validations were obtained from
492 multiple sources [34–38]. The validation for the yeast analysis was obtained from a TF-DNA
493 binding array from Zhu et al. [33].

494 **Slope and p-value Calculation for RANGES and ExRANGES**

495 The R package ExRANGES has been prepared and is available
496 http://github.com/DohertyLab/ExRANGES. Briefly the package performs the following
497 modifications to expression data. The slope was calculated as $\frac{Expression_{n+1} - Expression_n}{Timepoint_{n+1} - Timepoint_n}$.
498 *Sample* from the R base package was used to sample 10,000 with replacement for the slopes

499 calculated for each gene. The sampling population is dependent on the time series length (i.e.
500 the circadian data has 48 data points to sample from) P-values of the actual slope compared to
501 the distribution of the background slopes were calculated using *ecdf* from the R stat package
502 [52]. To preserve direction a duplicate version of the p-values matrix is created. The tails are
503 switched in this duplicated matrix by subtracting 1 from the matrix. The original matrix and the
504 switched matrix are both transformed by $-\log_{10}$. The matrices are then combined by taking the
505 higher value of the two matrices if the switched version is taken the sign is changed
506 (ifelse(matrix.up<matrix.down, -(matrix.down), matrix.up)). ExRANGES values were
507 determined for each gene by multiplying the expression at t_n by the weighted rate calculated
508 from t_n to t_{n+1}. See R package provided in http://github.com/DohertyLab/ExRANGES and linked
509 in OMICS tools https://omictools.com/expression-in-a-rate-normalized-gene-specific-tool.

510 **Network Inference using GENIE3**

511 To predict regulatory interaction between transcription factor and target gene, GENIE3 was used.
512 GENIE3 script was taken from http://www.montefiore.ulg.ac.be/~huynh-thu/software.html on
513 June 14, 2016 [8]. GENIE3 was modified by to be usable with *parLapply* from the R parallel
514 package [52]. We used 2000 trees for random forest for all data sets except the viral data set.
515 For the viral data set we limited it to 100 trees due to the size. The importance measure from
516 random forest was used to calculate the area under ROC.

517

518 **Network Inference using INFERELATOR**

519 TF-target interactions were calculated from both EXPRESSION and ExRANGES for the
520 Circadian, Viral, Arabidopsis, and rice datasets. TF and targets labels are identical to those used
521 as GENIE3 input. Time information in the form of the time step between each sample was added

522 to satisfy time course conditions as a parameter, default values were used for all other
523 parameters. Only confidence scores of TF-target interactions **greater than 0** were evaluated
524 against ChIP-Seq standards. The confidence scores were used as the prediction score to evaluate
525 against the targets identified for each TF from experimental ChIP-Seq data.
526

527 **ROC Calculation**

528 ROC values were determined by the ROCR package in R [53]. The importance measures were
529 used as the prediction score and the targets from the respective experimental validation (ChIP-
530 Seq, protein binding array, or DAP-Seq) were used as the metric to evaluate the performance
531 function. The area under the ROC is presented to summarize the accuracy.
532

533 **Acknowledgments**

534 This work was supported by funding from USDA NIFA 2014-04051. We would like to thank
535 Katie Greenham and Erin Slabaugh for critical suggestions on the manuscript preparation.
536 Additionally, we thank Steve Briggs for sharing the time, expertise, and helpful discussions of
537 his research group.

538 **References**

539

540 1. Balázsi G, Van Oudenaarden A, Collins JJ. Cellular decision making and biological noise:
541 From microbes to mammals. Cell. 2011;144: 910–925. doi:10.1016/j.cell.2011.01.030
542 2. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet.
543 2009;10: 669–680.
544 3. Qian J, Lin J, Luscombe NM, Yu H, Gerstein M. Prediction of regulatory networks:
545 genome-wide identification of transcription factor targets from gene expression data.
546 Bioinformatics. 2003;19: 1917–1926. doi:10.1093/bioinformatics/btg347
547 4. Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, Baliga NS, et al. The Inferelator:
548 an algorithm for learning parsimonious regulatory networks from systems-biology data
549 sets de novo. Genome Biol. 2006;7: R36. doi:10.1186/gb-2006-7-5-r36
550 5. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, et al.
551 ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a
552 Mammalian Cellular Context. BMC Bioinformatics. 2006;7: 1–15. doi:10.1186/1471-
553 2105-7-S1-S7
554 6. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, et al. Large-Scale
555 Mapping and Validation of Escherichia coli Transcriptional Regulation from a
556 Compendium of Expression Profiles. PLOS Biol. 2007;5: e8.
557 doi:10.1371/journal.pbio.0050008
558 7. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-
559 wide expression patterns. Proc Natl Acad Sci. 1998;95: 14863–14868.
560 8. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring Regulatory Networks from

561 Expression Data Using Tree-Based Methods. PLoS One. 2010;5: e12776.
562 9. Li Z, Li P, Krishnan A, Liu J. Large-scale dynamic gene regulatory network inference
563 combining differential equation models with local dynamic Bayesian network analysis.
564 Bioinformatics. 2011;27: 2686–2691. doi:10.1093/bioinformatics/btr454
565 10. Wilkins O, Hafemeister C, Plessis A, Holloway-Phillips M-M, Pham GM, Nicotra AB, et
566 al. EGRINs (Environmental Gene Regulatory Influence Networks) in Rice That Function
567 in the Response to Water Deficit, High Temperature, and Agricultural Environments.
568 Plant Cell. 2016; tpc.00158.2016. doi:10.1105/tpc.16.00158
569 11. Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, et al. Wisdom of
570 crowds for robust gene network inference. Nat Methods. 2012;9: 796–804.
571 doi:10.1038/nmeth.2016
572 12. Breiman L. Random forests. Mach Learn. Springer; 2001;45: 5–32.
573 13. Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002;2: 18–
574 22.
575 14. Penfold CA, Buchanan-Wollaston V, Denby KJ, Wild DL. Nonparametric Bayesian
576 inference for perturbed and orthologous gene regulatory networks. Bioinformatics.
577 2012;28: 233–241. doi:10.1093/bioinformatics/bts222
578 15. Penfold CA, Shifaz A, Brown PE, Nicholson A, Wild DL. CSI: A nonparametric Bayesian
579 approach to network inference from multiple perturbed time series gene expression data.
580 Stat Appl Genet Mol Biol. 2015;14: 307–310. doi:10.1515/sagmb-2014-0082
581 16. Bar-Joseph Z, Gitter A, Simon I. Studying and modelling dynamic biological processes
582 using time-series gene expression data. Nat Rev Genet. Nature Publishing Group, a
583 division of Macmillan Publishers Limited. All Rights Reserved.; 2012;13: 552–564.

584 17. Thompson D, Regev A, Roy S. Comparative Analysis of Gene Regulatory Networks:
585 From Network Reconstruction to Evolution. Annual Review of Cell and Developmental
586 Biology. 2015. doi:10.1146/annurev-cellbio-100913-012908
587 18. Pizarro A, Hayer K, Lahens NF, Hogenesch JB. CircaDB: a database of mammalian
588 circadian gene expression profiles. Nucleic Acids Res. Oxford University Press; 2013;41:
589 D1009–D1013. doi:10.1093/nar/gks1161
590 19. Zhang H-M, Chen H, Liu W, Liu H, Gong J, Wang H, et al. AnimalTFDB: a
591 comprehensive animal transcription factor database. Nucleic Acids Res. Oxford
592 University Press; 2012;40: D144–D149. doi:10.1093/nar/gkr965
593 20. Koike N, Yoo S-H, Huang H-C, Kumar V, Lee C, Kim T-K, et al. Transcriptional
594 Architecture and Chromatin Landscape of the Core Circadian Clock in Mammals. Science
595 (80-). 2012;
596 21. Takahashi JS, Kumar V, Nakashe P, Koike N, Huang H-C, Green CB, et al. ChIP-seq and
597 RNA-seq methods to study circadian control of transcription in mammals. Methods
598 Enzymol. 2015;551: 285–321. doi:10.1016/bs.mie.2014.10.059
599 22. Huynh-Thu VA. Machine learning-based feature ranking: Statistical interpretation and
600 gene network inference. Université de Liège, Liège, Belgium. 2012.
601 23. Qin B, Zhou M, Ge Y, Taing L, Liu T, Wang Q, et al. CistromeMap: a knowledgebase
602 and web server for ChIP-Seq and DNase-Seq studies in mouse and human.
603 Bioinformatics. 2012;28: 1411–1412. doi:10.1093/bioinformatics/bts157
604 24. Darnell JE, Kerr IM, Stark GR. Jak-STAT pathways and transcriptional activation in
605 response to IFNs and other extracellular signaling proteins. Science (80-). 1994;264: 1415
606 LP-1421.

607 25. Darnell JE. STATs and Gene Regulation. Science (80-). 1997;277: 1630 LP-1635.
608 26. Liu KD, Gaffen SL, Goldsmith MA. JAK/STAT signaling by cytokine receptors. Curr
609 Opin Immunol. 1998;10: 271–278. doi:http://dx.doi.org/10.1016/S0952-7915(98)80165-9
610 27. Horvath CM. STAT proteins and transcriptional responses to extracellular signals. Trends
611 Biochem Sci. 2000;25: 496–502. doi:http://dx.doi.org/10.1016/S0968-0004(00)01624-8
612 28. Bromberg J, Chen X. STAT proteins: Signal tranducers and activators of transcription. In:
613 Enzymology BT-M in, editor. Regulators and Effectors of Small GTPases, Part G.
614 Academic Press; 2001. pp. 138–151. doi:http://dx.doi.org/10.1016/S0076-6879(01)33052-
615 5
616 29. Stark GR, Darnell JE. The JAK-STAT Pathway at Twenty. Immunity. 2012;36: 503–514.
617 doi:10.1016/j.immuni.2012.03.013
618 30. Liu T-Y, Burke T, Park LP, Woods CW, Zaas AK, Ginsburg GS, et al. An individualized
619 predictor of health and disease using paired reference and target samples. BMC
620 Bioinformatics. London: BioMed Central; 2016;17: 47. doi:10.1186/s12859-016-0889-9
621 31. Respiratory Viral DREAM Challenge - syn5647810 [Internet]. [cited 8 Dec 2016].
622 Available: https://www.synapse.org/#!Synapse:syn5647810/wiki/399103
623 32. Liu T, Ortiz JA, Taing L, Meyer CA, Lee B, Zhang Y, et al. Cistrome: an integrative
624 platform for transcriptional regulation studies. Genome Biol. BioMed Central Ltd;
625 2011;12: R83. doi:10.1186/gb-2011-12-8-r83
626 33. Zhu C, Byers KJRP, McCord RP, Shi Z, Berger MF, Newburger DE, et al. High-
627 resolution DNA-binding specificity analysis of yeast transcription factors. Genome Res.
628 2009;19: 556–566. doi:10.1101/gr.090233.108
629 34. Lee J, He K, Stolc V, Lee H, Figueroa P, Gao Y, et al. Analysis of Transcription Factor

630 HY5 Genomic Binding Sites Revealed Its Hierarchical Role in Light Regulation of
631 Development. Plant Cell. 2007;19: 731–749. doi:10.1105/tpc.106.047688
632 35. Yant L, Mathieu J, Dinh TT, Ott F, Lanz C, Wollmann H, et al. Orchestration of the Floral
633 Transition and Floral Development in Arabidopsis by the Bifunctional Transcription
634 Factor APETALA2. Plant Cell Online. 2010;22: 2156–2170.
635 36. Chang KN, Zhong S, Weirauch MT, Hon G, Pelizzola M, Li H, et al. Temporal
636 transcriptional response to ethylene gas drives growth hormone cross-regulation in
637 Arabidopsis. Elife. 2013;2: e00675.
638 37. Liu T, Carlsson J, Takeuchi T, Newton L, Farré EM. Direct regulation of abiotic
639 responses by the Arabidopsis circadian clock component PRR7. Plant J. 2013; n/a-n/a.
640 doi:10.1111/tpj.12276
641 38. Nagel DH, Doherty CJ, Pruneda-paz JL, Schmitz RJ, Ecker JR. Genome-wide
642 identification of CCA1 targets uncovers an expanded clock network in Arabidopsis.
643 doi:10.1073/pnas.1513609112
644 39. O'Malley RC, Huang SC, Song L, Lewsey MG, Bartlett A, Nery JR, et al. Cistrome and
645 Epicistrome Features Shape the Regulatory DNA Landscape. Cell. Elsevier Inc.;
646 2016;165: 1280–1292. doi:10.1016/j.cell.2016.04.038
647 40. Khanday I, Das S, Chongloi GL, Bansal M, Grossniklaus U, Vijayraghavan U. Genome-
648 wide targets regulated by the OsMADS1 transcription factor reveals its DNA recognition
649 properties. Plant Physiol . 2016; doi:10.1104/pp.16.00789
650 41. Michael TP, Mockler TC, Breton G, McEntee C, Byer A, Trout JD, et al. Network
651 Discovery Pipeline Elucidates Conserved Time-of-Day-Specific cis-Regulatory Modules.
652 PLoS Genet. 2008;4.

653 42. Doherty CJ, Kay SA. Circadian Control of Global Gene Expression Patterns. Annu Rev
654 Genet. 2011;44: 419–444.
655 43. Zhang W, Zhang T, Wu Y, Jiang J. Genome-Wide Identification of Regulatory DNA
656 Elements and Protein-Binding Footprints Using Signatures of Open Chromatin in
657 Arabidopsis[C][W][OA]. Plant Cell. 2012;24: 2719–2731.
658 44. Sullivan AM, Arsovski AA, Lempe J, Bubb KL, Weirauch MT, Sabo PJ, et al. Mapping
659 and Dynamics of Regulatory DNA and Transcription Factor Networks in A. thaliana. Cell
660 Rep. 2014;8: 2015–2030. doi:10.1016/j.celrep.2014.08.019
661 45. Greenfield A, Hafemeister C, Bonneau R. Robust data-driven incorporation of prior
662 knowledge into the inference of dynamic regulatory networks. Bioinformatics. 2013;29:
663 1060–1067. doi:10.1093/bioinformatics/btt099
664 46. Vardi N, Levy S, Gurvich Y, Polacheck T, Carmi M, Jaitin D, et al. Sequential Feedback
665 Induction Stabilizes the Phosphate Starvation Response in Budding Yeast. Cell Rep.
666 2014;9: 1122–1134. doi:10.1016/j.celrep.2014.10.002
667 47. Mockler TC, Michael TP, Priest HD, Shen R, Sullivan CM, Givan SA, et al. The diurnal
668 project: Diurnal and circadian expression profiling, model-based pattern matching, and
669 promoter analysis. Cold Spring Harb Symp Quant Biol. 2007;72: 353–363.
670 48. Matsumoto T, Wu JZ, Kanamori H, Katayose Y, Fujisawa M, Namiki N, et al. The map-
671 based sequence of the rice genome. Nature. 2005;436: 793–800. doi:10.1038/nature03895
672 49. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg S. TopHat2: accurate
673 alignment of transcriptomes in the presence of insertions, deletions and gene fusions.
674 Genome Biol. 2013;14. doi:10.1186/gb-2013-14-4-r36
675 50. Anders S, Pyl PT, Huber W. HTSeq - A Python framework to work with high-throughput

676 sequencing data. Bioinformatics. 2015;31. doi:10.1093/bioinformatics/btu638
677 51. Jin J, He K, Tang X, Li Z, Lv L, Zhao Y, et al. An Arabidopsis Transcriptional Regulatory
678 Map Reveals Distinct Functional and Evolutionary Features of Novel Transcription
679 Factors. Mol Biol Evol . 2015;32: 1767–1773. doi:10.1093/molbev/msv058
680 52. R Core Team. R: A Language and Environment for Statistical Computing. Vienna,
681 Austria; 2016.
682 53. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier
683 performance in R. Bioinformatics. 2005;21: 7881.
684
685
686
687
688 **Supporting Information**
689
690 FIGURES
691 Figure 1: A) Overview of RANGES approach. For each $Gene_i$, the slope is calculated for all
692 possible consecutive changes time points. From this background distribution of changes in
693 expression the significance of each time point change is calculated. The $-\log_{10}$ of the p-value is
694 calculated and the sign change of direction is preserved. In the RANGES approach, this
695 significance value is used as the input into network inference using GENIE3 [8] for both the
696 transcription factor (TF) regulators and targets. For the EXPRESSION approach, the expression
697 values at each time point are provided for both the regulator and target to GENIE3. The
698 predictive ability of each approach was compared to the targets experimentally identified for

each TF by ChIP-Seq. B) Targets identified by RANGES and EXPRESSION approaches show little overlap. Scatter plot of targets of PER1 as identified by EXPRESSION or RANGES approaches. PER1 targets identified with similar rank by both approaches are shown in grey. PER1 targets identified as high ranking by RANGES are shown in blue and those ranking higher by EXPRESSION are red. PER1 targets identified by ChIP-Seq [20] are marked as stars. Genes identified as PER1 targets by each approach that were not identified in the ChIP-Seq identified targets are plotted as points.

Figure 2: A) Top targets for RANGES and EXPRESSION show different expression features. The expression values of the top three targets of PER1 identified by EXPRESSION (left; Pck1, Alb, GAPDH) and RANGES (right; RGS6, Rhpn2, Stx16) across the two day time series performed in twelve tissues [18]. The order of the tissues are: Brainstem, Lung, Kidney, Brown fat, White fat, Cerebellum, Hypothalamus, Aorta, Liver, Adrenal gland, Skeletal muscle, and Heart. Each tissue is plotted side by side in different colors. The points within a tissue represent expression levels every 2 hours over 48 hours. PER1 expression is shown in the center for comparison. B) Targets of the circadian TFs identified by EXPRESSION show higher standard deviation in expression levels across all samples than targets identified by RANGES. The standard deviation across all samples for the top 1000 targets of each circadian TF (ARNTL, CLOCK, NPAS, NR1D2, and PER1) identified by either the EXPRESSION or RANGES approach. C) Most circadian TF targets identified by RANGES show higher within tissue standard deviation. The standard deviation across the time series for each individual tissue was calculated for the top 1000 targets of each circadian TF (Arntl Clock, Npas, Nr1d2, and Per1) identified by either the EXPRESSION or RANGES approach. The mean of these within tissue standard deviations is plotted. D) EXPRESSION identified TF targets show greater variation in expression across all samples. Box plot showing the standard deviation of the top 1000 targets of all TFs identified by either EXPRESSION or RANGES. E) RANGES identified TF targets show greater within tissue variation. The standard deviation was calculated for each time series in each tissue for the top 1000 targets of all TFs identified by EXPRESSION or RANGES. Boxplot showing the mean standard deviation for each tissue for these top targets.

707

708 Figure 3: ExRANGES combines EXPRESSION and RANGES approach. A) Schematic of how
709 ExRANGES combines expression value and slope change. B) ExRANGES outperforms
710 EXPRESSION. The targets of the circadian TFs (ARNTl CLOCK, NPAS, NR1D2, and PER1)
711 identified by EXPRESSION or RANGES were validated against the ChIP-Seq identified targets
712 for these TFs [20]. Area under the ROC Curve (AURC) is plotted for targets identified by
713 EXPRESSION, EXPRESSION where the GENIE3 algorithm included a time step, and
714 ExRANGES (without a time step). C) ExRANGES improves identification of ChIP-Seq
715 validated targets in TFs that are not core components of the circadian clock. The EXPRESSION
716 and RATE identified targets of seven TFs with ChIP-Seq identified targets available from
717 CISTROME that are in our list of TFs and ChIP-Seq performed in epithelial cells which is a
718 tissue not sampled in the circadian time series [23] were compared and the area under the ROC
719 curve (AURC) is plotted. ExRANGES showed increased AURC for five of the TFs (ESR1,
720 POL2A, FOXA1, TFAP2A, and CHD4) over EXPRESSION or EXPRESSION including a time
721 step. For STAT5A and STAT5B ExRANGES did not increase the AURC.
722
723 Figure 4: A) ExRANGES improves identification of TF targets in unevenly sampled and
724 heterogeneous data. Targets of 83 TFs where ChIP-Seq data is available from Cistrome [23]
725 were compared for EXPRESSION and ExRANGES. Predictions of targets from EXPRESSION
726 and ExRANGES were compared to ChIP-Seq identified targets and the results for all 83 TFs are
727 presented as a box plot of area under the ROC curve (AURC). B) Variance comparison of the
728 viral and circadian data sets. The index of dispersion is calculated by dividing the variance of
729 each gene by its mean expression level and taking the mean of these values over all genes in the
730 dataset. The circadian data set showed a significantly higher Index of Variation than the viral

data set (Student's t-test, p-value < 10^{-15}). Improvement observed in ExRANGES identified targets varies across the 83 TFs tested. The difference between area under ROC curve of ExRANGES and EXPRESSION is plotted in ascending order for the 83 TFs tested. TFs are colored by TF family.

Figure 5: Functional Enrichment of ExRANGES identified targets. Gene Ontology enrichment was calculated using *Homo sapiens* GO slim annotations for the top 1000 targets of each TF predicted by either ExRANGES or EXPRESSION. The background annotations were limited the genes present on the HGU133 microarray. Enrichment score is the sum of the $-\log_{10}$ of the p-value of each GO category. A) Summary table of the enrichment scores for the top targets of all 930 TFs on the microarray. B) The distribution of enrichments scores from EXPRESSION targets (red) and ExRANGES targets (blue). C) Enrichment score difference of the 83 TFs with available ChIP-Seq data (Fig 4). Positive values indicate TF targets with a higher enrichments score in ExRANGES compared to EXPRESSION.

Figure 6: ExRANGES improves identification of TF targets validated by different methods. A) Targets identified for 52 yeast TFs by ESPRESSION (red) and ExRANGES (blue) were validated against the targets identified for each TF using a protein binding microarray [33] and boxplots generated from the area under the ROC curve (AURC). B) AURC for targets of the five Arabidopsis TFs with replicated ChIP-Seq data available for EXPRESSION and ExRANGES identified targets. C). AURC for targets identified for 307 TFs by EXPRESSION (red) and ExRANGES (blue) as validated against DAP-Seq identified targets [39].

754 Figure 7: Summary of the improvement observed by using ExRANGES with GENIE3 across
755 three data sets from different species. ROC and Precision recall (PR) curves for targets of all
756 ChIP-Seq validated TFs as identified by EXPRESSION (red) or ExRANGES (blue) with
757 GENIE3 for A) Circadian dataset from different mouse tissue samples B) Viral data set C)
758 Circadian dataset from Arabidopsis across different environmental variables.

759

760 Figure 8: ExRANGES retains performance improvement on small data sets. A) Area under the
761 ROC curve (AURC) for the top 1000 targets of OsMADS1 identified by EXPRESSION (red) or
762 ExRANGES (blue) and validated against the OsMADS1 ChIP-Seq data. B) Network of MADS
763 TFs predicted by ExRANGES. Interactions with OsMADS1 (green) determined by ExRANGES
764 with other MADS TFs in rice are visualized as an interaction network. ExRANGES predicted
765 targets of OsMADS1 are indicated in orange [40]. OsMADS15 (red) is predicted to regulate
766 OsMADS1 by ExRANGES (green arrow). Interactions between other MADS TFs predicted by
767 ExRANGES are indicated by black arrows.

768

769 **S1 Figure. ROC and precision recall (PR) curves.** ROC and PR curves for A) EXPRESSSION
770 B) RANGES and C) ExRANGES identified targets of the five circadian TFs based on true
771 positives identified with ChIP-Seq data [20].

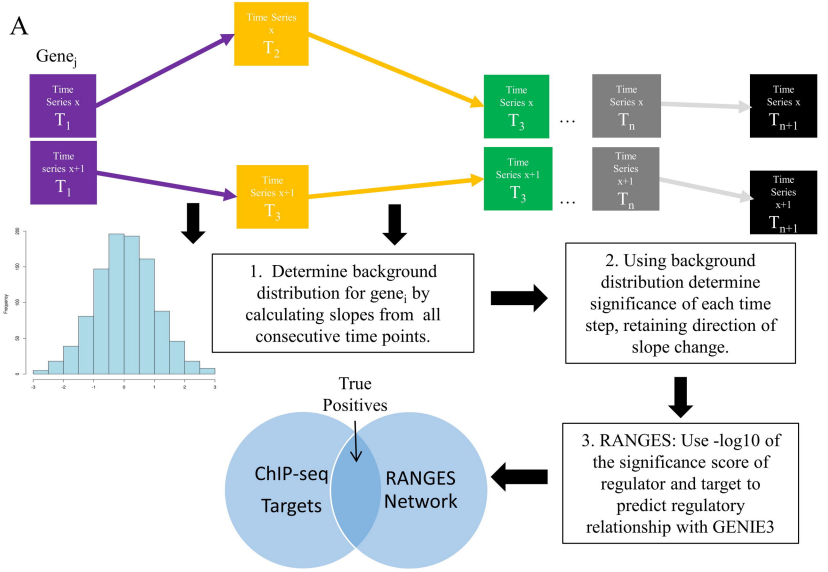
772 **S2 Figure. TF Targets identified differ when using EXPRESSION or EXPRESSION**
773 **approaches as features.** Scatter plots showing targets for the TFs A) NPAS2 B) CLOCK C)
774 NR1D1 and D) ARNTL. TF targets identified with similar rank by both approaches are shown
775 in black. Targets identified as high ranking by RANGES are shown in blue and those identified
776 by EXPRESSION are red. TF targets identified by ChIP-Seq [20] are marked as stars. Genes

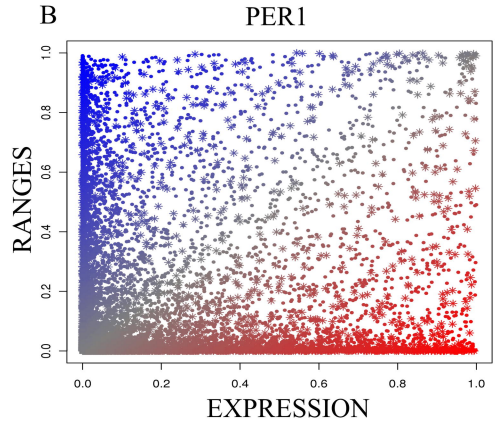
identified as TF targets by each approach that were not in the ChIP-Seq identified targets are plotted as points.

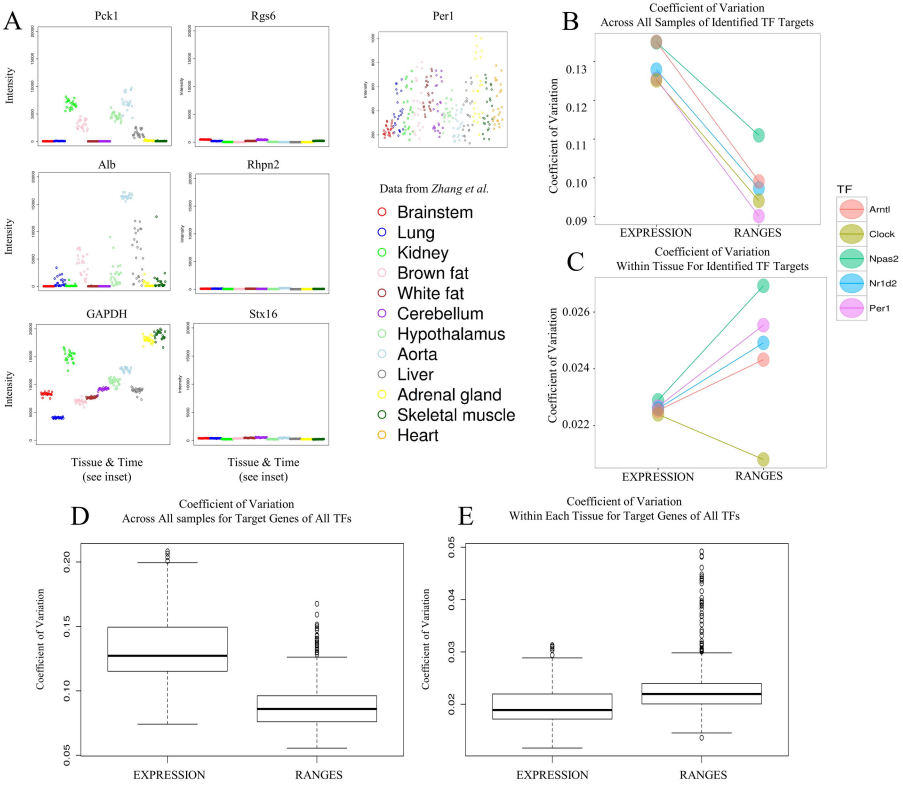
S3 Figure. Targets identified using EXPRESSION or RANGES show different distributions of hybridization intensity. Histogram showing the top 1000 PER1 targets identified by A) EXPRESSION (red) have a higher distribution of expression as measured by hybridization intensities compared to the background distribution of all genes (grey). B) RANGES (blue) identified targets show a similar expression distribution to the background genes. C) The distribution of expression levels of the top 1000 targets identified by EXPRESSION (red) for all TFs is higher than the background gene expression (grey). D) The distribution of expression levels for the top 1000 targets of each TF identified by RANGES (blue) is similar to the distribution of the expression levels from all genes (grey).

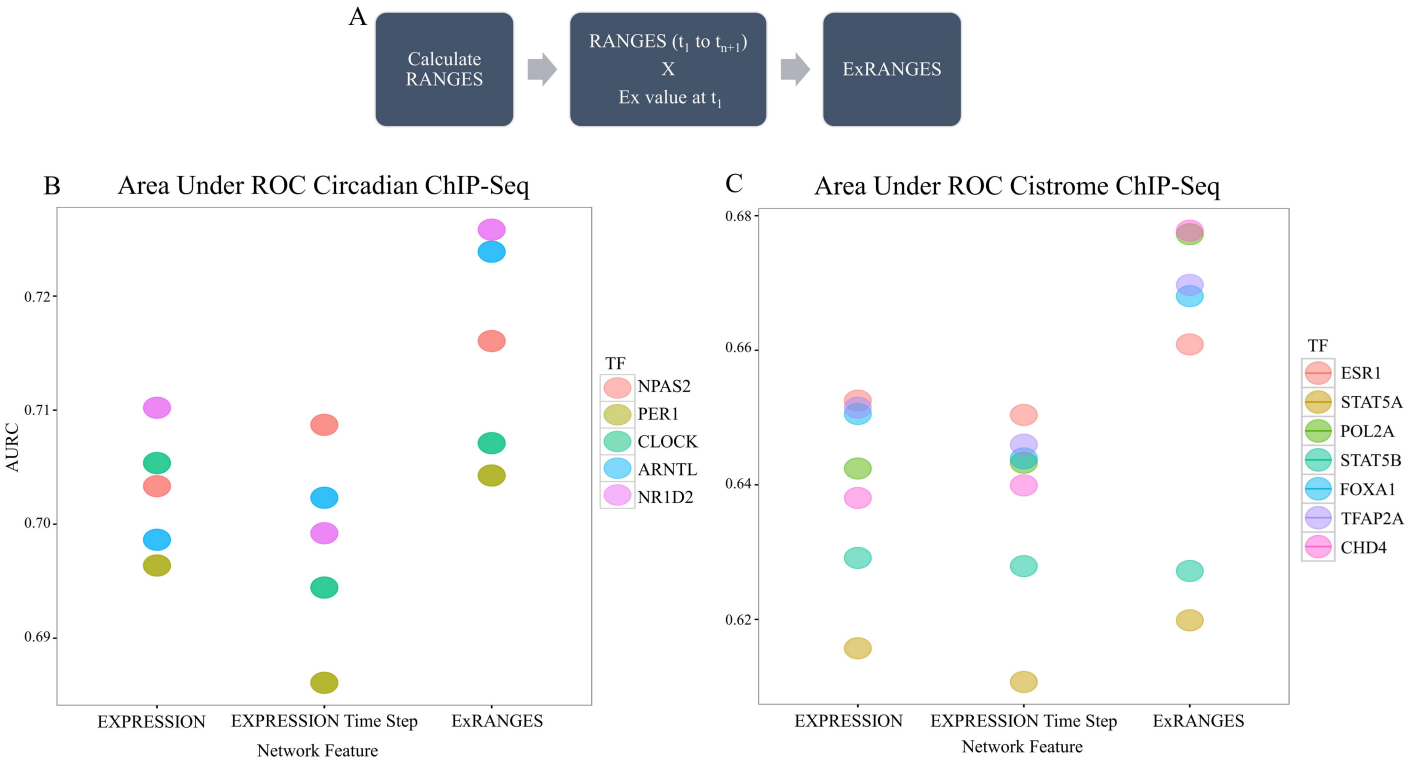
S4 Figure. Using the INFERELATOR algorithm, ExRANGES shows the greatest improvement in identifying TF targets in the Arabidopsis data set. ROC and Precision recall (PR) curves for targets of all ChIP-Seq validated TFs as identified by EXPRESSION (red) or ExRANGES (blue) using INFERELATOR for A) Circadian dataset from different mouse tissue samples B) Viral data set C) Circadian dataset from Arabidopsis across different environmental variables.

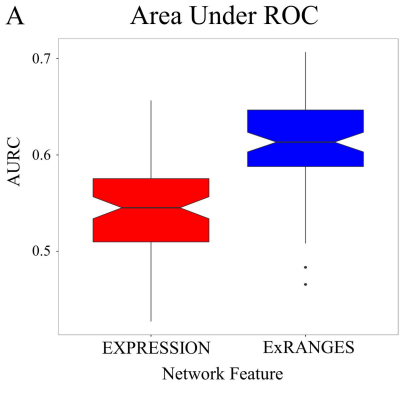
ST1 Table. GO Enrichment for JUND. GO categories enriched in expression in top 1000 JUND targets identified by either EXPRESSION or ExRANGES (FDR adjusted p-value <0.01) show more target genes per category in ExRANGES top targets (32 categories) than in EXPRESSION top targets (8 categories). The vacuolar transport category showed no change in the number of gene annotated in that category in EXPRESSION or ExRANGES targets.

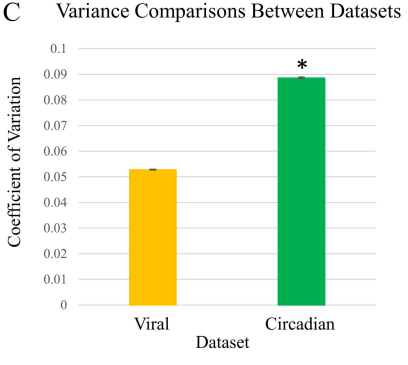

A
Gene$_j$
Time Series x
T_1
Time series x+1
T_1
Time Series x
T_2
Time Series x+1
T_3
Time Series x
T_3
Time Series x+1
T_3
Time Series x
T_n
Time Series x+1
T_n
Time Series x
T_{n+1}
Time Series x+1
T_{n+1}
1. Determine background distribution for gene$_i$ by calculating slopes from all consecutive time points.
2. Using background distribution determine significance of each time step, retaining direction of slope change.
3. RANGES: Use -log10 of the significance score of regulator and target to predict regulatory relationship with GENIE3
True Positives
ChIP-seq Targets
RANGES Network

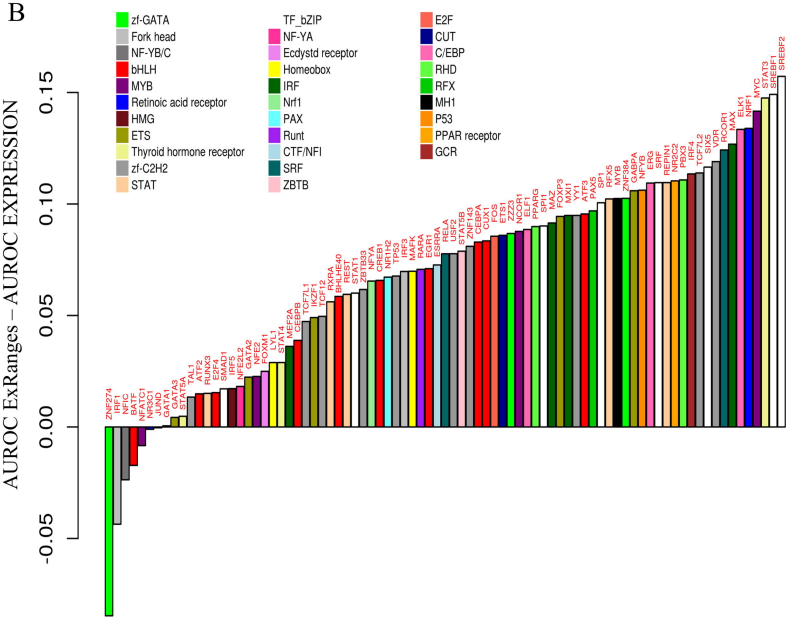

B
PER1
RANGES
EXPRESSION
0.0
0.2
0.4
0.6
0.8
1.0


A
Pck1
Rgs6
Per1
Alb
Rhpn2
GAPDH
Stx16
Intensity
Tissue & Time
(see inset)
Data from Zhang et al.
Brainstem
Lung
Kidney
Brown fat
White fat
Cerebellum
Hypothalamus
Aorta
Liver
Adrenal gland
Skeletal muscle
Heart
B
Coefficient of Variation
Across All Samples of Identified TF Targets
Coefficient of Variation
EXPRESSION
RANGES
TF
Arntl
Clock
Npas2
Nr1d2
Per1
C
Coefficient of Variation
Within Tissue For Identified TF Targets
D
Coefficient of Variation
Across All samples for Target Genes of All TFs
E
Coefficient of Variation
Within Each Tissue for Target Genes of All TFs


A
Calculate RANGES
RANGES (t_1 to t_{n+1})
X
Ex value at t_1
ExRANGES
B
Area Under ROC Circadian ChIP-Seq
AURC
0.72
0.71
0.70
0.69
EXPRESSION
EXPRESSION Time Step
ExRANGES
Network Feature
TF
NPAS2
PER1
CLOCK
ARNTL
NR1D2
C
Area Under ROC Cistrome ChIP-Seq
0.68
0.66
0.64
0.62
EXPRESSION
EXPRESSION Time Step
ExRANGES
Network Feature
TF
ESR1
STAT5A
POL2A
STAT5B
FOXA1
TFAP2A
CHD4


A
Area Under ROC
AURC
0.7
0.6
0.5
EXPRESSION
ExRANGES
Network Feature


C
Variance Comparisons Between Datasets
Coefficient of Variation
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
*
Viral
Circadian
Dataset


B
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
AUROC ExRanges – AUROC EXPRESSION
0.15
0.10
0.05
0.00
-0.05

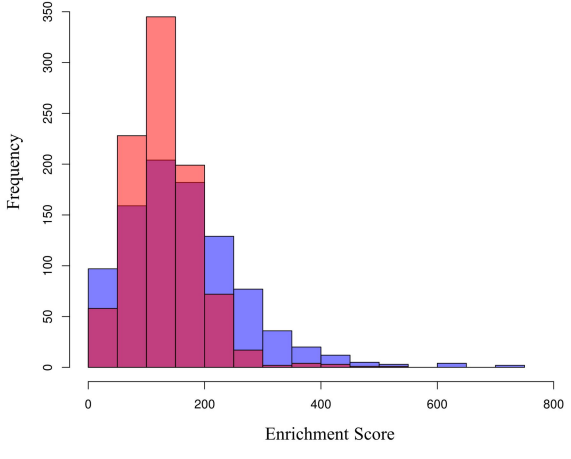
A

GO Enrichment Score Table

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

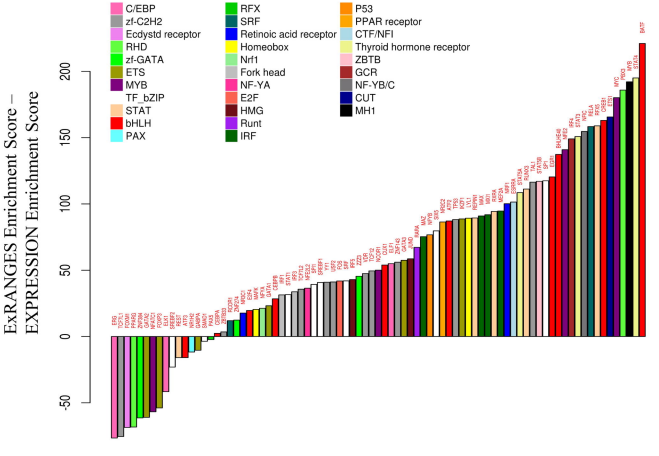
B

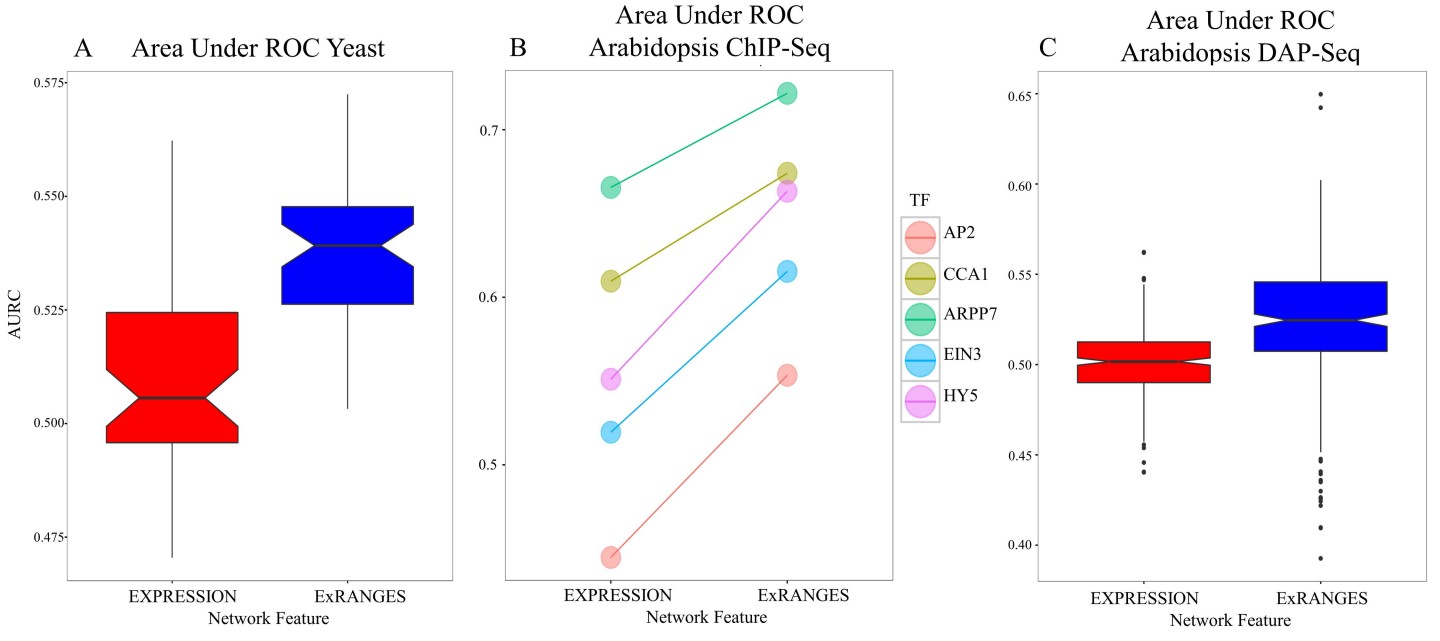
GO Enrichment Score Overlap

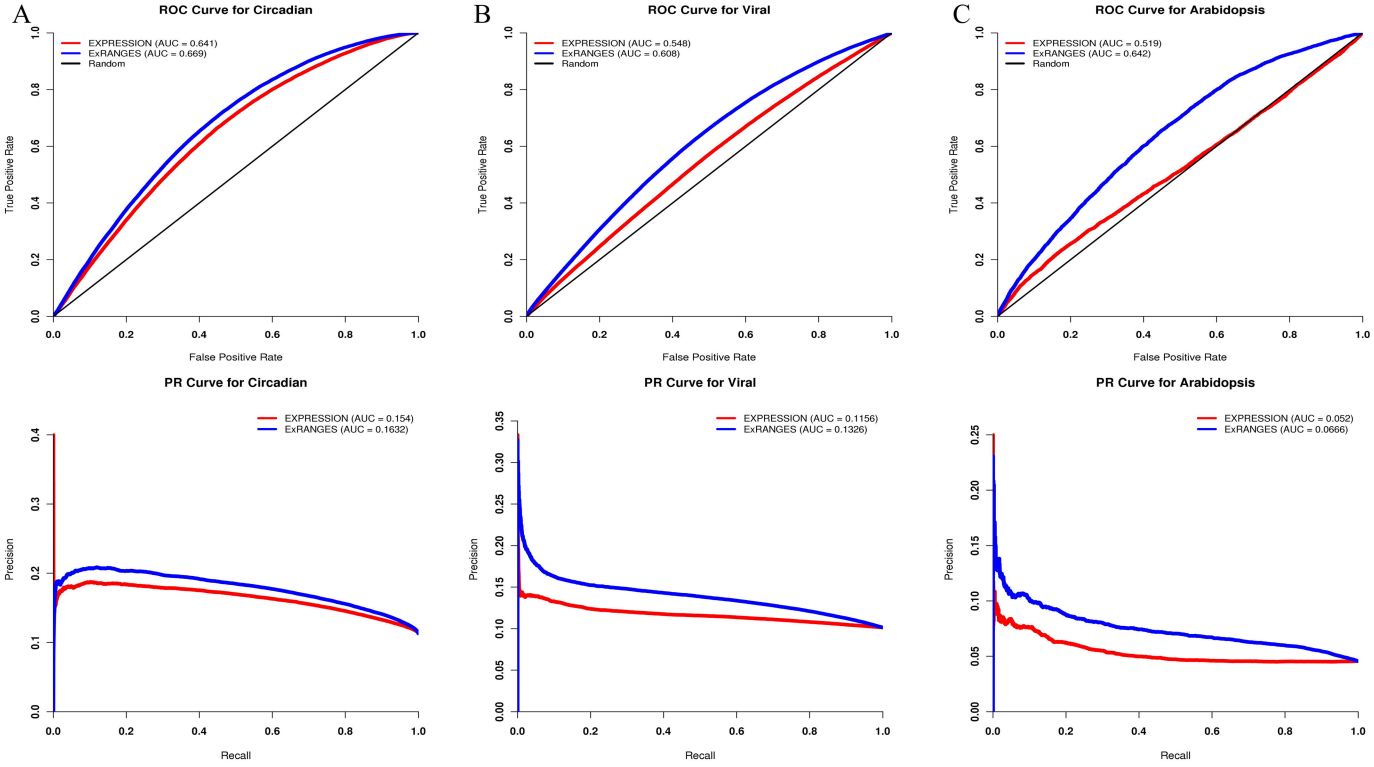


C

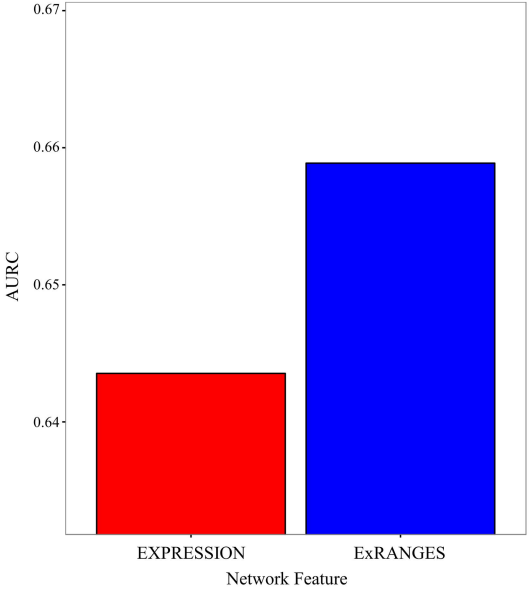
GO Enrichment Difference



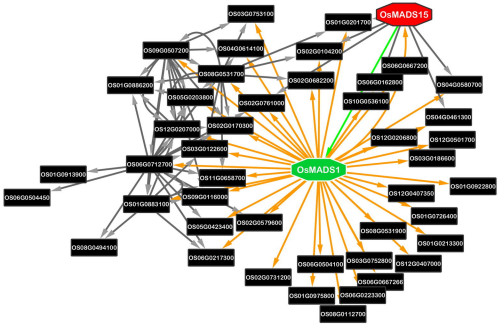

A
Area Under ROC Yeast
AURC
0.575
0.550
0.525
0.500
0.475
EXPRESSION
ExRANGES
Network Feature
B
Area Under ROC
Arabidopsis ChIP-Seq
0.7
0.6
0.5
EXPRESSION
ExRANGES
Network Feature
TF
AP2
CCA1
ARPP7
EIN3
HY5
C
Area Under ROC
Arabidopsis DAP-Seq
0.65
0.60
0.55
0.50
0.45
0.40
EXPRESSION
ExRANGES
Network Feature


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