Incorporating the Rate of Transcriptional Change Improves Construction of Transcription Factor 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 random-forest based network construction approach 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

To understand how organisms can turn a collection of genes into a physiological response, we need to understand how certain genes are turned on and off. Transcription factors (TFs) have a huge role in gene regulation and control how genes coordinate daily activities, developmental processes, stress responses, and many other cellular processes. In model organisms, the ability to identify direct targets of TFs via ChIP-Seq in a high-throughput manner has advanced our understanding of transcriptional regulatory networks and how organisms regulate gene expression. However, for non-model organisms, the challenge of assessing TF–target relationships remains a limiting factor to understanding regulatory control. Computational approaches to identify regulator-target relationships *in silico* from easily attainable transcriptional data offer a solution. We observed that transcriptional assays are often performed as a time series and that most computational approaches do not take full advantage of the information available in time series data. Therefore, we have implemented this approach, which modifies the input into network identification algorithms and improves their ability to predict regulator-target interactions. Our approach essentially weights the expression value of each time point by the slope change after that time point so that relationships between regulators and targets are emphasized at the time points when the transcript levels are changing. We hope this improvement will increase the ability to identify regulators of interest in non-model 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 elemental 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], 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. Here, we observed that this created a bias that preferentially identified highly expressed targets. We reasoned that using the rate of change between two time points would reduce this bias and improve the ability to identify targets for each regulator. We have developed a method that incorporates the temporal change in expression levels and have implemented this with the GENIE3 algorithm. We observe that our approach combining Expression by RAte Normalized in a GEne Specific manner (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.

Results

Developing a Method to Incorporate Rate of Expression Change

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 [14], 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 2 h for 48) for each gene. At each time step, *t* the slope between *t* and *t + 1* was compared to a permutation of this background generated by sampling 10,000 times with replacement and 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 [15]. 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 potential targets across the same samples. For the RANGES approach, the $-\log_{10}$ of the p-value for the significance of each change in time across the 24 time steps was used as the input for both the 35,556 targets and 1690 regulators. For both approaches all TFs were also included in the target list to identify regulatory connections between TFs.

To evaluate the ability of each approach to correctly identify targets of the TFs, we compared the resulting targets of each TF identified by either the RANGES or EXPRESSION approach with the targets identified by ChIP-Seq for five TFs involved in circadian regulation where three replicates of each ChIP-Seq experiment were performed: PER1, CLOCK, NPAS2, NR1d2, and ARNTL [16,17]. Targets identified by each approach that were considered significant targets by these published ChIP-Seq experiments were scored as true positive targets of that TF.

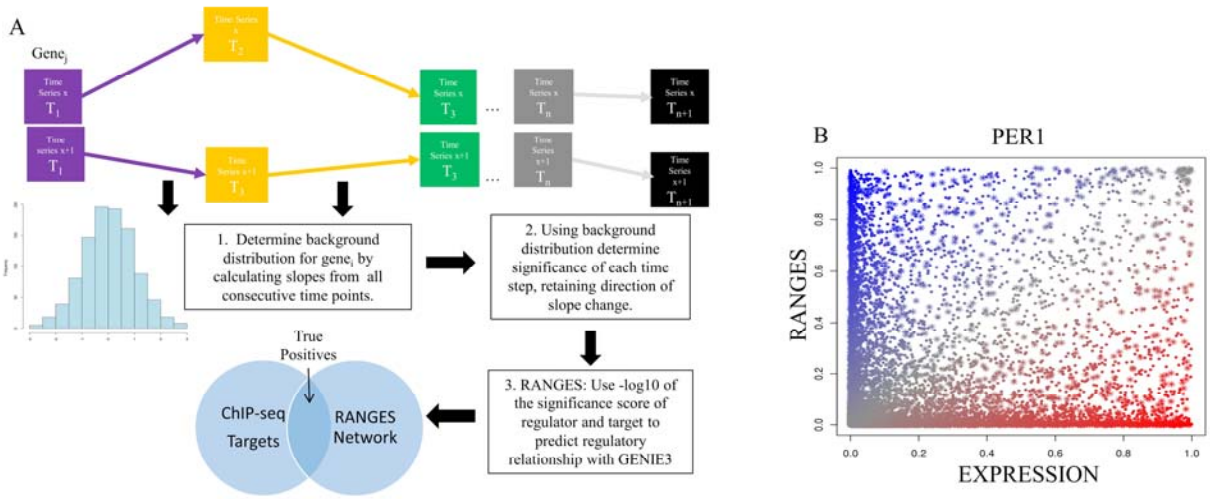

Figure 1: A) Overview of RANGES approach. For each Gene$_i$, the slope is calculated for all possible consecutive changes time points. From this background distribution of changes in expression the significance of each time point change is calculated. The $-\log_{10}$ of the p-value is calculated and 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 [16] 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.

RANGES and EXPRESSION Values Identify Different Sets of True Positive Targets

We compared the targets identified by using RANGES to those identified using EXPRESSION. For PER1 both approaches identified true targets more than would be expected by chance (ROC curve, Fig S1). EXPRESSION showed a larger area under the ROC curve, indicating higher accuracy in identifying true positive targets of PER1. However, there was little overlap in the top true positive targets identified by each approach (Fig 1B). Many genes that were scored

strongly by RANGES as PER1 targets, including many true positive targets of PER1, had low scores when evaluated using the EXPRESSION approach. Likewise, several of the top scoring true positive targets by EXPRESSION had low RANGE scores. This difference in the targets identified by each approach, including true positives, was also observed for the other four TFs we evaluated (Fig S2). These results indicate that information contained in the relationship 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) [16]. 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 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 regulators was shifted higher compared to the background distribution levels (Fig S3C). While the top 1000 targets of all 1690 TFs identified by RANGES reflected the background distribution of hybridization intensity (Fig S3D). While hybridization intensity cannot directly be translated into expression levels, these observations suggest that there are features of the targets identified by RANGES that are distinct from those identified by EXPRESSION. We hypothesized that combining these two approaches would improve the overall ability to detect true positive targets of each regulator.

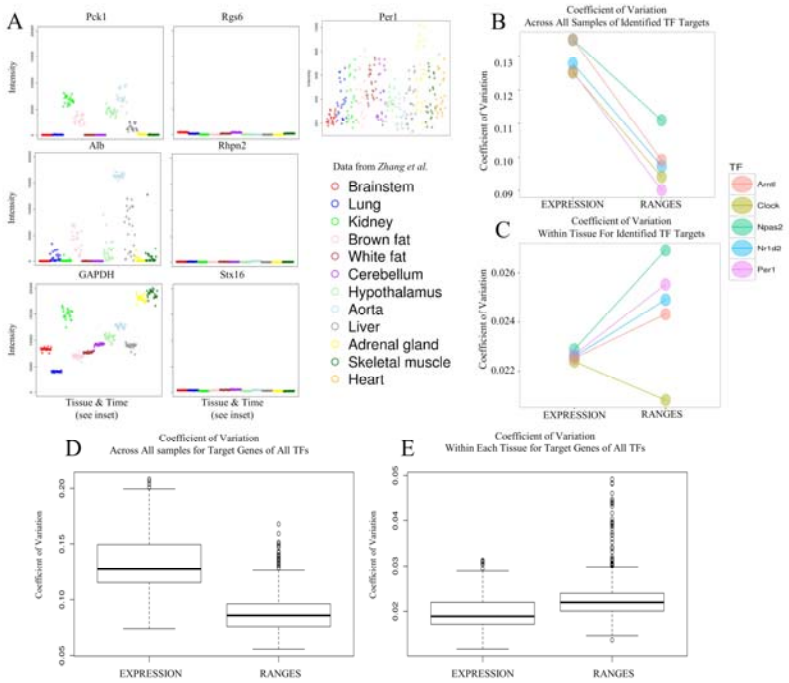

Figure 2: A) Top targets for RANGES and EXPRESSION show different expression features. The expression values of the top three targets of PER1 identified by EXPRESSION (left; Pck1, Alb, GAPDH) and RANGES (right; RGS6, Rhpn2, Stx16) across the two day time series performed in twelve tissues [14]. The order 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.

Combining Rate Change with Expression Levels

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

Using the identified ChIP-Seq targets as true positives from Koike et al. [16], 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 [18]. 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).

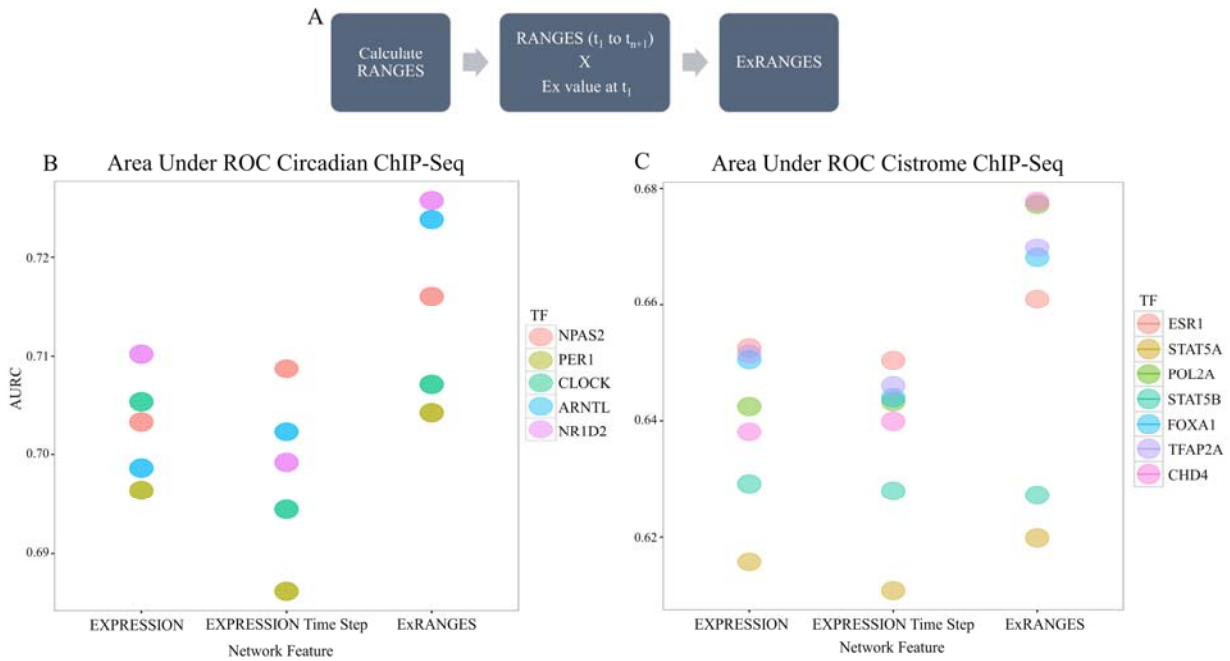

Figure 3: ExRANGES combines EXPRESSION and RANGES approach. A) Schematic of how ExRANGES combines expression value and slope change. B) ExRANGES outperforms EXPRESSION. The targets of the circadian TFs (ARNTl CLOCK, NPAS, NR1D2, and PER1) identified by EXPRESSION or RANGES were validated against the ChIP-Seq identified targets for these TFs [16]. 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 [19] 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.

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 seven TFs that we tested are: ESR1, STAT5A, STAT5B, POL2A, FOXA1, TFAP2A, and CHD4 [19]. We observed improvement of the area under the ROC curve for five of the seven TFs (ESR1, POL2A, FOXA1, TFAP2A, and CHD4) by combining expression and rate change information using ExRANGES (Fig 3C). 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 [20–25]. This suggests that for TFs that show little variation in expression throughout the day in each time series the addition of the RANGES component may not offer much improvement. (Fig S4).

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

Although circadian and diel time series experiments are a rich resource providing substantial variance for identifying regulatory relationships, 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 [26,27] 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 (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 [28], we observed an overall improvement in the detection of ChIP-Seq identified targets (Fig 4B). The improvement varies by TF (Fig 4C).

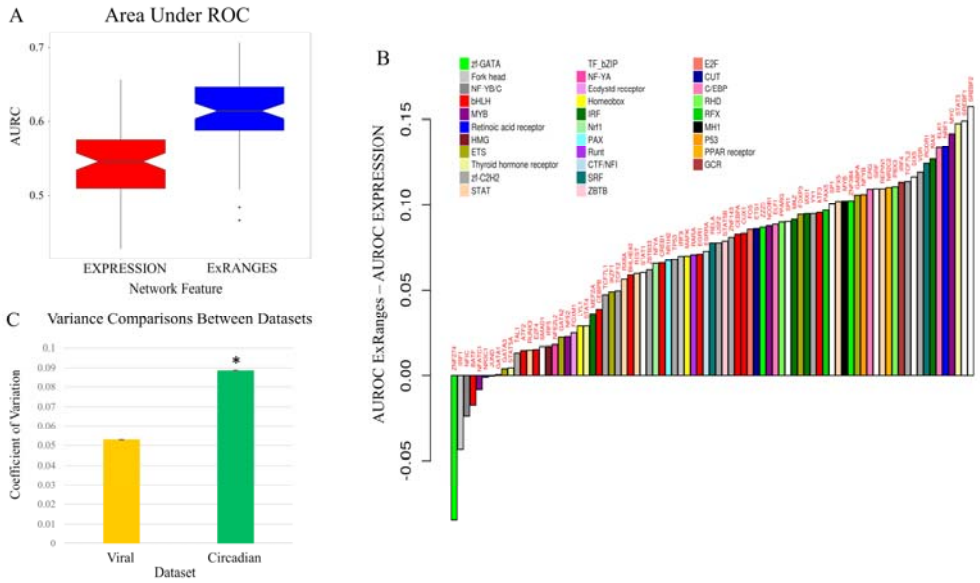

Figure 4: A) ExRANGES improves identification of TF targets in unevenly sampled and heterogeneous data. Targets of 83 TFs where ChIP-Seq data is available from Cistrome [19] were compared for EXPRESSION and ExRANGES. Predictions of targets from EXPRESSION and ExRANGES were compared to ChIP-Seq identified 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. 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 microrarray [15]. 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 (Fig 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 (Fig 5C). 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).

A GO Enrichment Score Table

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

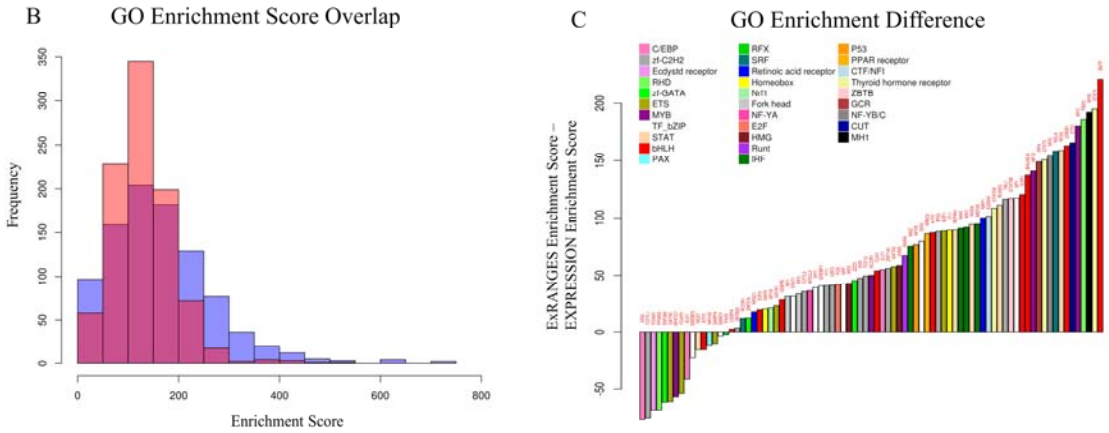

Figure 5: Functional Enrichment of ExRANGES identified targets. Gene Ontology enrichment was calculated using *Homo sapiens* GO slim annotations for the top 1000 targets of each TF predicted by either ExRANGES or EXPRESSION. The background annotations were limited the genes present on the HGU133 microarray. 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.

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 and true positives were based on ChIP-Seq identified targets of each TF. To evaluate the performance of ExRANGES compared to EXPRESSION for RNA-Seq data

we applied each approach to an RNA-Seq data set performed in *Saccharomyces cerevisiae.* 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 area under the ROC curve for the identified targets for each of the 52 TFs using the TF targets identified by protein binding microarray analysis as true positives [29]. For most TFs, the AUC was improved by the use of ExRANGES compared to EXPRESSION (Fig 6A).

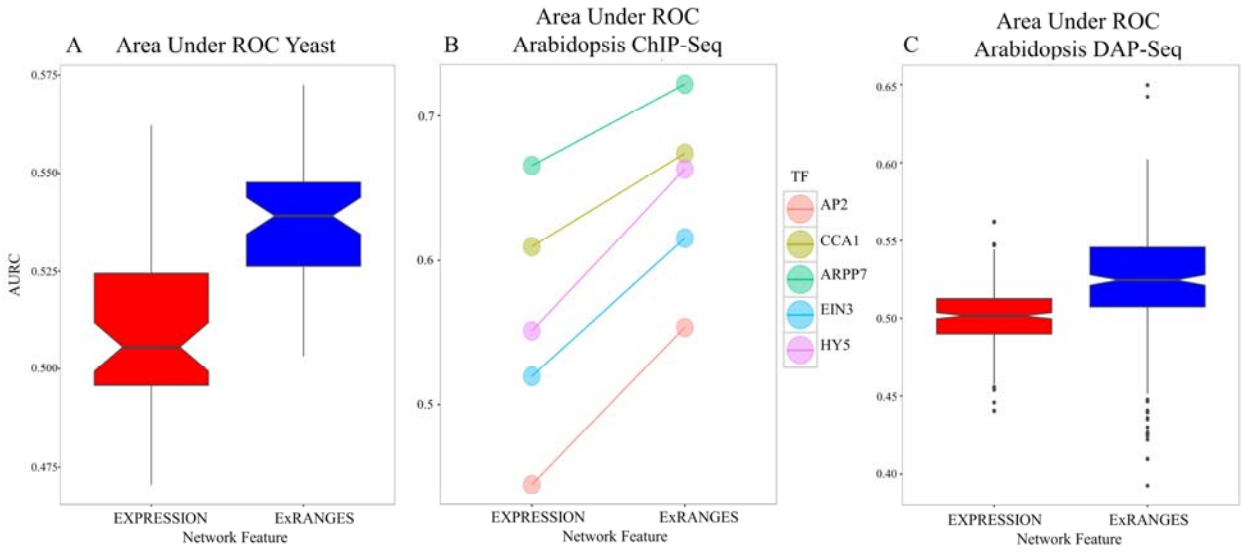

Figure 6: ExRANGES improves identification of TF targets validated by different methods. A) Targets identified for 52 yeast TFs by ESPRESSION (red) and ExRANGES (blue) were validated against the targets identified for each TF using a protein binding microarray [29] and boxplots generated from the area 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 [35].

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 [30–34]. 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 [35]. 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 ExRANGES improves performance on large data sets (228 mouse, 2372 human, and 150 yeast samples) (Fig 7). 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.

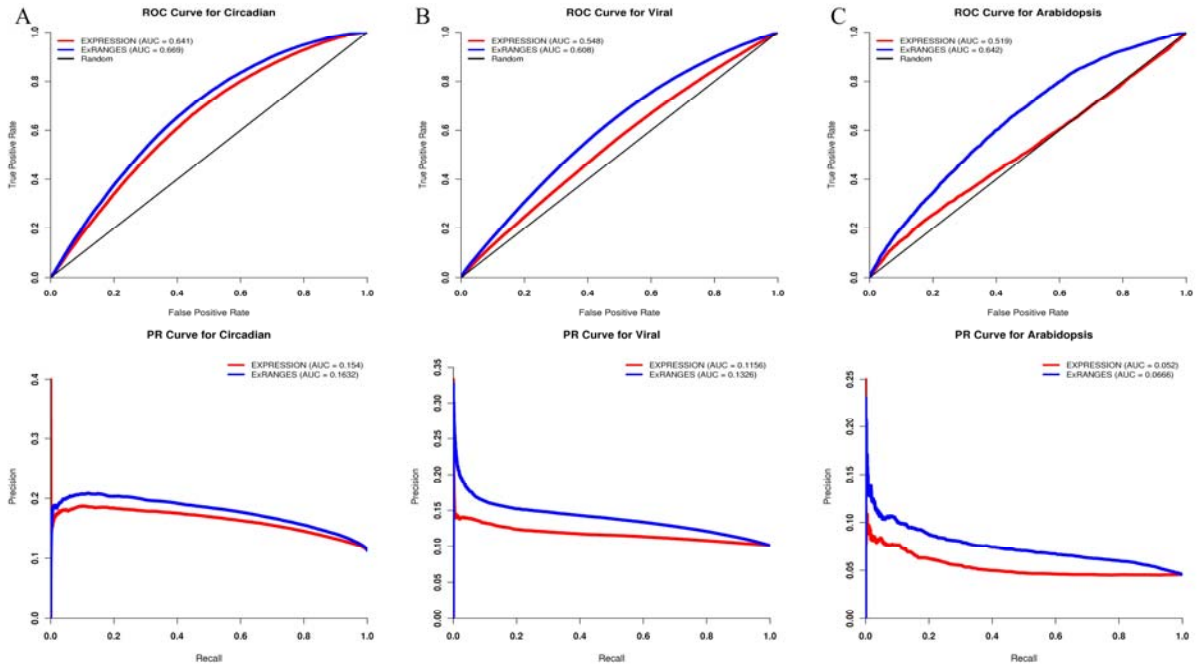

Figure 7: Summary of ExRANGES improvement across three data sets from different species. ROC and Precision recall (PR) curves for targets of all ChIP-Seq validated TFs as identified by EXPRESSION (red) or ExRANGES (blue) for A) Circadian dataset from different mouse tissue samples B) Viral data set C)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

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 [36]. Therefore, we compared the ability of ExRANGES and EXPRESSION 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 (Fig 8) compared to EXPRESSION.

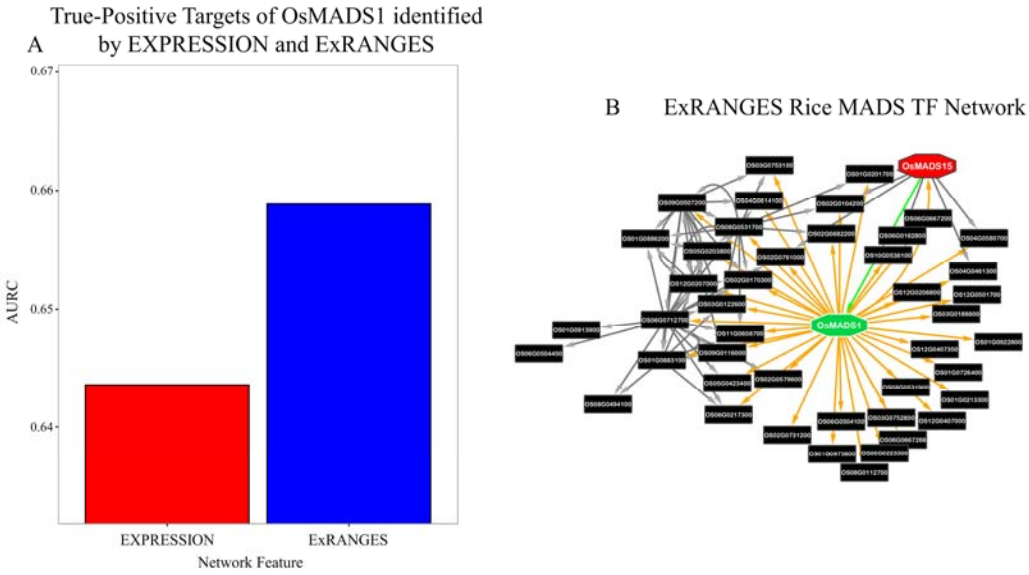

Figure 8: ExRANGES retains performance improvement on small data sets. A) Area under the ROC curve (AURC) for the top 1000 targets of OsMADS1 identified by EXPRESSION (red) or ExRANGES (blue) and validated against the OsMADS1 ChIP-Seq data. B) Network of MADS TFs predicted by ExRANGES. Interactions with 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 [36]. OsMADS15 (red) is predicted to regulate OsMADS1 by ExRANGES (green arrow). Interactions between other MADS TFs predicted by ExRANGES are indicated by black arrows.

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 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 [14,37,38] 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. Here, we show that 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 were those that showed large variance between tissue, while RANGES identified targets that showed larger variance within each time series. 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 algorithm GENIE3 and demonstrated that it improved the performance of this algorithm. The ExRANGES method can be applied 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, ChIP-Chip, DAP-Seq, and protein binding microarray. While these experimental approaches identify potential TF targets in a genome-wide manner, they are not perfect as gold-standards for validation of transcriptional regulatory networks. If there are systematic errors in target identification by ChIP-Seq, ExRANGES may perform better than indicated here. Although ChIP-Seq may not be an ideal gold standard, it does provide a benchmark for comparing 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 [35]. Additionally, the authors of this paper improved their recall of ChIP-Seq identified targets by selecting targets that were also supported by DNase-Seq sensitivity assays [39,40]. 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 [10,41]. 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 modest 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 transcriptional regulatory network 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.

Materials and Methods

Sources for Expression Data Sets

Circadian Data Set

Normalized expression data from murine sources was downloaded from CircaDB [14]. Microarray-based expression levels from 288 samples were used in this study. The data available was from twelve different tissues that were sampled every 2 h for 48 h.

Viral Data Set

The expression data used for the viral experimental analysis was downloaded from GEO GSE73072. The data is composed of seven studies of individuals sampled before and after respiratory infection. Expression is data from blood samples of approximately twenty individuals taken over a seven to nine day period depending on the individual study. Sampling was not evenly spaced between time points. In total data from 2372 microarrays were used. The expression datasets used for the analyses described in this manuscript were contributed by Drs. Ephraim Tsalik and Geoffrey Ginsburg from Duke University and the Durham VA Medical Center. They were obtained as part of The Respiratory Viral DREAM Challenge through Synapse ID syn5647810 [27].

S. *cerevisiae* RNA-Seq Data

RNA-Seq based expression data from *S. cerevisiae* was downloaded from GEO GSE61668 [42]. This data set evaluates phosphate starvation in six genotypes of *S. cerevisiae*. Transcript expression was measured by RNA-Seq every 15m for six hours after transfer to reduced phosphate media (150 samples total).

Arabidopsis Circadian Data

Normalized microarray expression data for Arabidopsis was obtained from www.mocklerlab.org/diurnal [43]. This data set consisted of Arabidopsis plants of various ages grown in 12 different environmental conditions sampled every 4 h for 48 h for a total of 144 samples.

Oryza sativa Diel Data

Rice variety IR64 was grown in the field at the International Rice Research Institute (, Philippines). When the plants reached 50% flowering, panicle tissue was harvested at dawn, dawn + 3.5h, dawn + 7h, dawn + 10.5h, dusk, dawn + 14h, dawn + 17.5h, and dawn + 21h. Four replicates were harvested for each of these eight time points for a total of 32 samples. The third rachis of the panicle was ground in liquid nitrogen with a metal pestle. The tissue was then lyophilized at -60C°overnight. Total RNA was isolated using RNeasy Plant Mini Kit (Qiagen, Germany) with the recommended RLT lysis buffer. The RNA extraction protocol was modified to include and additional incubation with DNaseI. mRNA was isolated from 2 μg of total RNA using magnetic oligo(dT) (NEB, Ipswich, MA). Directional RNA-Seq libraries were prepared from isolated mRNA. Libraries were quantified using a 2100 Bioanalyzer (Agilent, Santa Clara, CA) RNA-Seq was performed on a HiSeq 2500 (Illumina, San Diego, CA). Reads were trimmed using seqtk (https://github.com/lh3/seqtk). Samples were aligned with Tophat2 to the IRGSP-1.0 genome [44,45]. Counts per gene were identified by HTSeq Count [46]. Data is available through GSE92302.

Selection of Regulators

Transcription factors used as regulators for the murine circadian data and human viral data were obtained from http://www.bioguo.org/AnimalTFDB/index.php [15]. Arabidopsis transcription

factor list were obtained from http://planttfdb.cbi.pku.edu.cn/ [47]. *S. cerevisiae* transcription factors were obtained from [29].

Sources for Validation Resources

The direct targets for the five circadian TFs from murine data were obtained from the supplementary information provided in [16]. Targets for additional TFs and the validation of the viral data from human expression data were obtained from the cistrome project (http://cistrome.org/Cistrome/Cistrome_Project.html) [19]. Eighty-three TFs were selected as regulators that were labeled as evaluated blood tissue and present on the HGU133 microarray. ChIP-Seq targets were determined by BETA (http://cistrome.org/BETA/). If multiple ChIP-Seq were provided they were combined. The Arabidopsis ChIP-Seq validations were obtained from multiple sources [30–34]. The validation for the yeast analysis was obtained from a TF-DNA binding array from Zhu et al. [29].

Slope and p-value Calculation for RANGES and ExRANGES

The R package ExRANGES has been prepared and is available http://github.com/DohertyLab/ExRANGES. Briefly the package performs the following modifications to expression data. The slope was calculated as $\frac{Expression_{n+1} - Expression_n}{Timepoint_{n+1} - Timepoint_n}$. *Sample* from the R base package was used to sample 10,000 with replacement for the slopes calculated for each gene. The sampling population is dependent on the time series length (i.e. the circadian data has 48 data points to sample from) P-values of the actual slope compared to the distribution of the background slopes were calculated using *ecdf* from the R stat package [48]. To preserve direction a duplicate version of the p-values matrix is created. The tails are switched in this duplicated matrix by subtracting 1 from the matrix. The original matrix and the switched matrix are both transformed by $-\log_{10}$. The matrices are then combined by taking the

higher value of the two matrices if the switched version is taken the sign is changed (ifelse(matrix.up<matrix.down, -(matrix.down), matrix.up)). ExRANGES values were determined for each gene by multiplying the expression at t_n by the weighted rate calculated from t_n to t_{n+1}. See R package provided in http://github.com/DohertyLab/ExRANGES.

Network Inference using GENIE3

To predict regulatory interaction between transcription factor and target gene, Genie3 was used. Genie3 script was taken from http://www.montefiore.ulg.ac.be/~huynh-thu/software.html on June 14, 2016 [8]. Genie3 was modified by to be usable with *parLapply* from the R parallel package [48]. We used 2000 trees for random forest for all data sets except the viral data set. For the viral data set we limited it to 100 trees due to the size. The importance measure from random forest was used to calculate the area under ROC.

ROC Calculation

ROC values were determined by the ROCR package in R [49]. The 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 is presented to summarize the accuracy.

Acknowledgments

This work was supported by funding from USDA NIFA 2014-04051. We would like to thank 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.

References

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

Expression Data Using Tree-Based Methods. PLoS One. 2010;5: e12776.

9. Li Z, Li P, Krishnan A, Liu J. Large-scale dynamic gene regulatory network inference combining differential equation models with local dynamic Bayesian network analysis. Bioinformatics. 2011;27: 2686–2691. doi:10.1093/bioinformatics/btr454

10. Wilkins O, Hafemeister C, Plessis A, Holloway-Phillips M-M, Pham GM, Nicotra AB, et al. EGRINs (Environmental Gene Regulatory Influence Networks) in Rice That Function in the Response to Water Deficit, High Temperature, and Agricultural Environments. Plant Cell. 2016; tpc.00158.2016. doi:10.1105/tpc.16.00158

11. Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, et al. Wisdom of crowds for robust gene network inference. Nat Methods. 2012;9: 796–804. doi:10.1038/nmeth.2016

12. Breiman L. Random forests. Mach Learn. Springer; 2001;45: 5–32.

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

14. Pizarro A, Hayer K, Lahens NF, Hogenesch JB. CircaDB: a database of mammalian circadian gene expression profiles. Nucleic Acids Res. 2013;41: D1009–D1013. doi:10.1093/nar/gks1161

15. Zhang H-M, Chen H, Liu W, Liu H, Gong J, Wang H, et al. AnimalTFDB: a comprehensive animal transcription factor database. Nucleic Acids Res. 2012;40: D144–D149. doi:10.1093/nar/gkr965

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

17. Takahashi JS, Kumar V, Nakashe P, Koike N, Huang H-C, Green CB, et al. ChIP-seq and RNA-seq methods to study circadian control of transcription in mammals. Methods Enzymol. 2015;551: 285–321. doi:10.1016/bs.mie.2014.10.059

18. Huynh-Thu VA. Machine learning-based feature ranking: Statistical interpretation and gene network inference. Université de Liège, Liège, Belgium. 2012.

19. Qin B, Zhou M, Ge Y, Taing L, Liu T, Wang Q, et al. CistromeMap: a knowledgebase and web server for ChIP-Seq and DNase-Seq studies in mouse and human. Bioinformatics. 2012;28: 1411–1412. doi:10.1093/bioinformatics/bts157

20. Darnell JE, Kerr IM, Stark GR. Jak-STAT pathways and transcriptional activation in response to IFNs and other extracellular signaling proteins. Science (80-). 1994;264: 1415 LP-1421.

21. Darnell JE. STATs and Gene Regulation. Science (80-). 1997;277: 1630 LP-1635.

22. Liu KD, Gaffen SL, Goldsmith MA. JAK/STAT signaling by cytokine receptors. Curr Opin Immunol. 1998;10: 271–278. doi:http://dx.doi.org/10.1016/S0952-7915(98)80165-9

23. Horvath CM. STAT proteins and transcriptional responses to extracellular signals. Trends Biochem Sci. 2000;25: 496–502. doi:http://dx.doi.org/10.1016/S0968-0004(00)01624-8

24. Bromberg J, Chen X. STAT proteins: Signal tranducers and activators of transcription. In: Enzymology BT-M in, editor. Regulators and Effectors of Small GTPases, Part G. Academic Press; 2001. pp. 138–151. doi:http://dx.doi.org/10.1016/S0076-6879(01)33052-5

25. Stark GR, Darnell JE. The JAK-STAT Pathway at Twenty. Immunity. 2012;36: 503–514. doi:10.1016/j.immuni.2012.03.013

26. Liu T-Y, Burke T, Park LP, Woods CW, Zaas AK, Ginsburg GS, et al. An individualized

predictor of health and disease using paired reference and target samples. BMC Bioinformatics. London: BioMed Central; 2016;17: 47. doi:10.1186/s12859-016-0889-9

27. Respiratory Viral DREAM Challenge - syn5647810 [Internet]. [cited 8 Dec 2016]. Available: https://www.synapse.org/#!Synapse:syn5647810/wiki/399103

28. Liu T, Ortiz JA, Taing L, Meyer CA, Lee B, Zhang Y, et al. Cistrome: an integrative platform for transcriptional regulation studies. Genome Biol. BioMed Central Ltd; 2011;12: R83. doi:10.1186/gb-2011-12-8-r83

29. Zhu C, Byers KJRP, McCord RP, Shi Z, Berger MF, Newburger DE, et al. High-resolution DNA-binding specificity analysis of yeast transcription factors. Genome Res. 2009;19: 556–566. doi:10.1101/gr.090233.108

30. Lee J, He K, Stolc V, Lee H, Figueroa P, Gao Y, et al. Analysis of Transcription Factor HY5 Genomic Binding Sites Revealed Its Hierarchical Role in Light Regulation of Development. Plant Cell. 2007;19: 731–749. doi:10.1105/tpc.106.047688

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

32. Chang KN, Zhong S, Weirauch MT, Hon G, Pelizzola M, Li H, et al. Temporal transcriptional response to ethylene gas drives growth hormone cross-regulation in Arabidopsis. Elife. 2013;2: e00675.

33. Liu T, Carlsson J, Takeuchi T, Newton L, Farré EM. Direct regulation of abiotic responses by the Arabidopsis circadian clock component PRR7. Plant J. 2013; n/a-n/a. doi:10.1111/tpj.12276

34. Nagel DH, Doherty CJ, Pruneda-paz JL, Schmitz RJ, Ecker JR. Genome-wide

identification of CCA1 targets uncovers an expanded clock network in Arabidopsis. doi:10.1073/pnas.1513609112

35. O'Malley RC, Huang SC, Song L, Lewsey MG, Bartlett A, Nery JR, et al. Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell. Elsevier Inc.; 2016;165: 1280–1292. doi:10.1016/j.cell.2016.04.038

36. Khanday I, Das S, Chongloi GL, Bansal M, Grossniklaus U, Vijayraghavan U. Genome-wide targets regulated by the OsMADS1 transcription factor reveals its DNA recognition properties. Plant Physiol . 2016; doi:10.1104/pp.16.00789

37. Michael TP, Mockler TC, Breton G, McEntee C, Byer A, Trout JD, et al. Network Discovery Pipeline Elucidates Conserved Time-of-Day-Specific cis-Regulatory Modules. PLoS Genet. 2008;4.

38. Doherty CJ, Kay SA. Circadian Control of Global Gene Expression Patterns. Annu Rev Genet. 2011;44: 419–444.

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

40. Sullivan AM, Arsovski AA, Lempe J, Bubb KL, Weirauch MT, Sabo PJ, et al. Mapping and Dynamics of Regulatory DNA and Transcription Factor Networks in A. thaliana. Cell Rep. 2014;8: 2015–2030. doi:10.1016/j.celrep.2014.08.019

41. Greenfield A, Hafemeister C, Bonneau R. Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. Bioinformatics. 2013;29: 1060–1067. doi:10.1093/bioinformatics/btt099

42. Vardi N, Levy S, Gurvich Y, Polacheck T, Carmi M, Jaitin D, et al. Sequential Feedback

Induction Stabilizes the Phosphate Starvation Response in Budding Yeast. Cell Rep. 2014;9: 1122–1134. doi:10.1016/j.celrep.2014.10.002

43. Mockler TC, Michael TP, Priest HD, Shen R, Sullivan CM, Givan SA, et al. The diurnal project: Diurnal and circadian expression profiling, model-based pattern matching, and promoter analysis. Cold Spring Harb Symp Quant Biol. 2007;72: 353–363.

44. Matsumoto T, Wu JZ, Kanamori H, Katayose Y, Fujisawa M, Namiki N, et al. The map-based sequence of the rice genome. Nature. 2005;436: 793–800. doi:10.1038/nature03895

45. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg S. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14. doi:10.1186/gb-2013-14-4-r36

46. Anders S, Pyl PT, Huber W. HTSeq - A Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31. doi:10.1093/bioinformatics/btu638

47. Jin J, He K, Tang X, Li Z, Lv L, Zhao Y, et al. An Arabidopsis Transcriptional Regulatory Map Reveals Distinct Functional and Evolutionary Features of Novel Transcription Factors. Mol Biol Evol . 2015;32: 1767–1773. doi:10.1093/molbev/msv058

48. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2016.

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

Supporting Information

FIGURES

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 [16] 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 [14]. 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.

674

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 [16]. 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 [19] 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.

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 [19] 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.

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 [29] 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 [35].

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.

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) and validated against the OsMADS1 ChIP-Seq data. B) 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 [36]. OsMADS15 (red) is predicted to regulate OsMADS1 by ExRANGES (green arrow). Interactions between other MADS TFs predicted by ExRANGES are indicated by black arrows.

S1 Figure. ROC and precision recall (PR) curves. ROC and PR curves for A) EXPRESSSION B) RANGES and C) ExRANGES identified targets of the five circadian TFs based on true positives identified with ChIP-Seq data [16].

S2 Figure. TF Targets identified differ when using EXPRESSION or EXPRESSION approaches as features. Scatter plots showing targets for the TFs A) NPAS2 B) CLOCK C) NR1D1 and D) ARNTL. TF targets identified with similar rank by both approaches are shown in black. Targets identified as high ranking by RANGES are shown in blue and those identified by EXPRESSION are red. TF targets identified by ChIP-Seq [16] 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).

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.