BMC Bioinformatics

RESEARCH

Open Access

Conservation of dynamic characteristics of transcriptional regulatory elements in periodic biological processes

Francis C. Motta[1*†], Robert C. Moseley[2†], Bree Cummins[3], Anastasia Deckard[4] and Steven B. Haase[2]

*Correspondence:
fmotta@fau.edu
†Francis C. Motta and Robert C. Moseley equal contributor
[1] Department of Mathematical Sciences, Florida Atlantic University, 777 Glades Rd, Boca Raton, FL 33431, USA
Full list of author information is available at the end of the article

Abstract

Background: Cell and circadian cycles control a large fraction of cell and organismal physiology by regulating large periodic transcriptional programs that encompass anywhere from 15 to 80% of the genome despite performing distinct functions. In each case, these large periodic transcriptional programs are controlled by gene regulatory networks (GRNs), and it has been shown through genetics and chromosome mapping approaches in model systems that at the core of these GRNs are small sets of genes that drive the transcript dynamics of the GRNs. However, it is unlikely that we have identified all of these core genes, even in model organisms. Moreover, large periodic transcriptional programs controlling a variety of processes certainly exist in important non-model organisms where genetic approaches to identifying networks are expensive, time-consuming, or intractable. Ideally, the core network components could be identified using data-driven approaches on the transcriptome dynamics data already available.

Results: This study shows that a unified set of quantified dynamic features of high-throughput time series gene expression data are more prominent in the core transcriptional regulators of cell and circadian cycles than in their outputs, in multiple organism, even in the presence of external periodic stimuli. Additionally, we observe that the power to discriminate between core and non-core genes is largely insensitive to the particular choice of quantification of these features.

Conclusions: There are practical applications of the approach presented in this study for network inference, since the result is a ranking of genes that is enriched for core regulatory elements driving a periodic phenotype. In this way, the method provides a prioritization of follow-up genetic experiments. Furthermore, these findings reveal something unexpected—that there are shared dynamic features of the transcript abundance of core components of unrelated GRNs that control disparate periodic phenotypes.

Keywords: Cell cycle, Circadian rhythms, Gene regulatory networks, Transcription factors, Network inference

Background

Periodic phenotypes span nearly the entire tree of life and include such fundamental processes as the cell-division cycle, circadian rhythms, and developmental cycles. Probing the genetic mechanisms that give rise to these dynamic activities is not only crucial to our fundamental understanding of life and its evolution, it may also add to the current collection of synthetic biology components and principles of design, and may reveal novel treatments for disease and infection. A vast body of experimental evidence, gathered over years of targeted experimentation (e.g. gene knock-outs) has uncovered the existence of endogenous circadian clocks: complex GRNs—comprised mostly of interacting transcription factors (TFs)—within cyanobacteria, fungi, plants and mammals [1–3]. Moreover, a GRN also appears to control the timing of cell-cycle events in budding yeast [4–8]. To understand the complex dynamic functions of these GRNs, experimentalists and computational scientists have developed a variety of approaches to infer the structure of GRNs. An essential first step is to identify, from among an expansive set of candidate genes, those *core* gene products controlling the dynamics of the associated program of gene expression. We conceptualize core nodes as interacting in a strongly connected subnetwork of mutual activation and repression. The core then drives the dynamics of "output" or "effector" nodes that do not feed back into the core but rather transmit the dynamic expression pattern to downstream target genes (Fig. 1).

Identifying core nodes is especially daunting for organisms where genetic experiments are largely intractable. Moreover, functional redundancy, and complex GRN mechanisms, such as accessory feedback loops, can complicate the discovery of core nodes. Here we identify distinguishing characteristics of the dynamics of gene expression that are conserved across organisms that are separated by hundreds of millions of years of evolution, in vastly different biological processes, and across data-collection modalities. We discover that a combination of dynamic features provides a rank ordering of all genes such that core nodes are generally highly ranked, even among the many genes which exhibit these features. Moreover, we find that, in general, a combination of dynamic features more accurately distinguishes core transcriptional regulators than individual features on their own. Our findings support the use of quantified dynamic characteristics of gene expression to identify core regulatory elements and show that there are common features in the dynamic gene expression of core regulatory variables that drive a variety of biological processes.

Results and discussion

Understanding the function of GRNs requires a specification of the control variables and their interactions. Accurate inferences have generally required substantial genetic perturbation and physical localization studies and thus has been confined to experimentally tractable model systems. However, previous work has indicated that interactions between GRN nodes can be inferred directly from transcriptome dynamics data [9]. Here we investigated whether the core nodes themselves could also be identified from time series transcriptomics. We determined that quantifiable features from time-series gene expression measurements can be used to identify

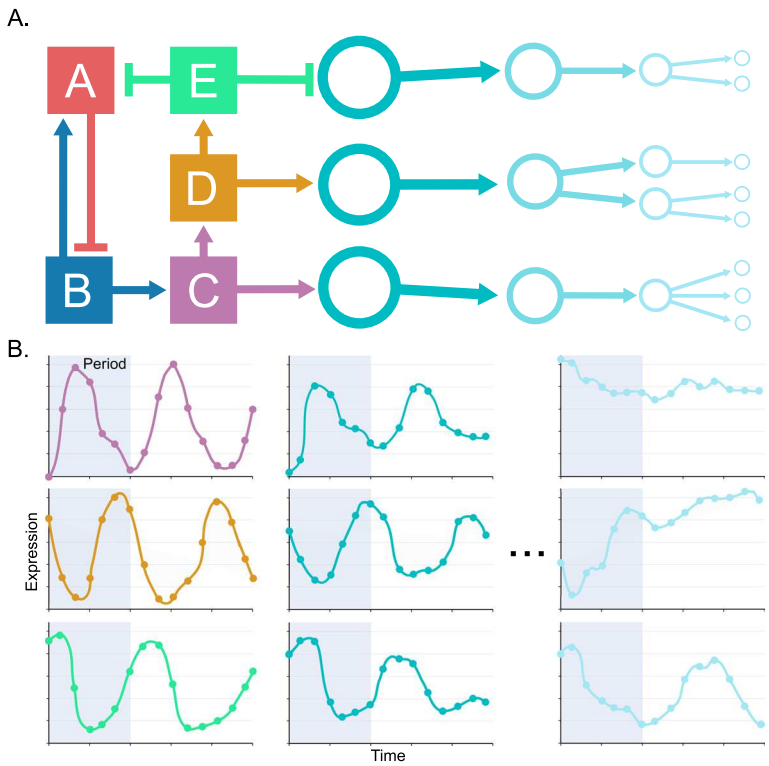

Fig. 1 Conceptual model of core regulatory elements. **A** Conceptual model of a transcriptional regulatory network with core nodes (squares) operating in a strongly-connected subnetwork of mutual activation (arrows) and repression (short bars), together with outputs of the core (circles). Output nodes transmit the transcriptional signal that is generated by the core, but which diminishes as it moves away from core nodes. **B** Illustrations of transcript abundance profiles exhibited by the core and its output nodes, with core nodes having oscillations that have a precise match to a specified period (shaded region) and large variations in expression

experimentally-inferred core nodes from model systems across taxa (yeast cell cycle, mouse circadian cycle, plant circadian cycle).

We consider two quantifiable characteristics of dynamic transcript abundance profiles, measured in multiple ways, and assess the capacity of each to differentiate core from non-core regulatory elements. Because the dynamic phenotypes of interest are rhythmic, e.g. sleep-wake cycles, cell division, etc., it is natural to ask to what extent, relative to all genes, will the core elements driving these processes be endowed with periodicity that matches the observed cycling at the level of their transcript abundance? Moreover, since the core elements are by definition those TFs governing the dynamics of gene expression, to what extent will the strength of the regulatory signal be reflected in the dynamics of transcript abundance?

Dynamic transcript abundance features identify regulatory elements in core networks

We first examined the list of dynamic features, used both individually and in various combinations (see Table 3) to distinguish core TFs from among all TFs. To provide a unified measure of performance across datasets, we considered the average precision (AP) of each metric's ranking of transcripts. When restricting to TFs, using both periodicity and regulation strength features together yields significantly higher AP scores than

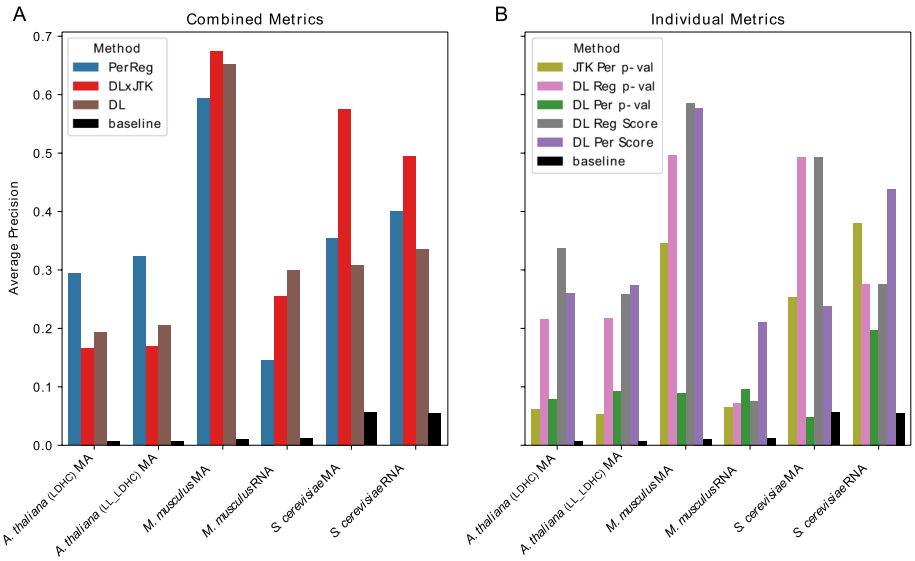

Fig. 2 Identifying core genes among transcription factors. Average precision of classifiers identifying core from non-core TFs among all TFs by combined metrics (**A**) and individual metrics (**B**) (Table 3) as well as the baseline average precision of a random classifier, for each dataset (Table 4)

the baseline for each of the six datasets examined (Fig. 2A). Even using just one of the two types of dynamic features, we see remarkable improvement over baseline, although generally lower AP scores, than the combined metrics, across all six datasets (Fig. 2B). Notably, the datasets considered in this study represent organisms from three different kingdoms, undergoing two ostensibly mechanistically distinct periodic dynamic processes. The complete set of metrics scoring all genes in all datasets are available in Additional file 4: Gene Rankings and the complete precision-recall curves for all datasets and all metrics are available in Additional file 5: Figs. S1–S6.

From the viewpoint of an experimentalist interested in understanding the entirety of a core network, it is encouraging to observe the enrichment of the top 25 TFs with core genes. Among the top 25 TFs ranked by the measure DL×JTK, 13 (12) of the possible 17 *S. cerevisiae* core genes are identified using the microarray (RNASeq) data. Similarly, 10 (4) core *M. musculus* genes from the possible list of 15 (14) core genes, are among the top 25 transcription factors as ranked by DL×JTK using microarray (RNASeq) data. Finally, *A. thaliana* LDHC and LL_LDHC datasets contain 4 and 5 core genes, respectively, from among the 11 possible core, in the top 25. Strikingly, 9 of the top 10 *M. musculus* TFs and 6 of the top 10 *S. cerevisiae* TFs are core when the high temporal resolution microarray datasets are ranked using DL×JTK. These results are given in Table 1.

We emphasize the skill of dynamic gene expression features to identify core TFs in Fig. 3, which gives the distribution of core TF DL×JTK ranks among all TFs for *S. cerevisiae* (see also Additional file 5: Table S1) and heatmaps of microarray gene expression grouped by DLxJTK rankings. The top 25 genes are clearly seen to robustly oscillate at approximately the specified period (94 min) and among these are 13 of the 17 core genes.

The recall of core genes by DL × JTK among the top 25 TFs is as much as 76.5% of the core yeast cell-cycle transcriptional regulatory network, up to 66.67% for the

Table 1 Top 25 transcription factors ranked by DL×JTK metric

Rank	*S. cerevisiae*		*M. Musculus*		*A. thaliana*	
	MA	RNA	MA	RNA	LDHC	LL_LDHC
1	*SWI5**	*TOS4**	*ARNTL**	*DBP**	*COL1*	*STH*
2	*YOX1**	*HST4*	*DBP**	*NPAS2**	*HB-12*	*AT1G26790*
3	*HST3*	*HST3*	*NPAS2**	*CDX4*	*TGA3*	*CCA1**
4	*ASF1*	*SWI5**	*NR1D1**	*ARNTL**	*RVE1*	*BBX18*
5	*ACE2**	*YOX1**	*NR1D2**	*EGR1*	*MYBL2*	*COL1*
6	*RTT107*	*RTT107*	*BHLHE41**	*GM14401*	*LHY**	*CDF1*
7	*STB1**	*WTM2*	*CLOCK**	*GM14305*	*CO*	*COL2*
8	*HCM1**	*ASH1**	*NFIL3**	*POU4F1*	*PIL6*	*CDF3*
9	*RME1*	*FKH1**	*RFXANK*	*EN2*	*AT2G28200*	*AT2G28200*
10	*FKH1**	*ASF1*	*RORC**	*DMRTA2*	*COL2*	*RVE1*
11	*PLM2**	*ACE2**	*TEF**	*LHX1*	*CCA1**	*LHY**
12	*SWI4**	*POG1*	*CREM*	*GM20422*	*PRR7**	*COL5*
13	*NDD1**	*SWI4**	*EGR1*	*GM14444*	*HYH*	*PIF4*
14	*ASH1**	*RME1*	*PPARD*	*OVOL2*	*BBX18*	*PIL6*
15	*YHP1**	*PLM2**	*ZBTB21*	*GM4969*	*RVE8**	*BBX16*
16	*TOS4**	*RLF2*	*NFIC*	*HOXC4*	*PRE1*	*LUX**
17	*EDS1*	*NDD1**	*AHCTF1*	*FOXO6*	*BZS1*	*PRR7**
18	*RIF1*	*HCM1**	*ATF5*	*MESP1*	*EPR1*	*CDF2*
19	*SIP4*	*GAT1*	*LITAF*	*AI854703*	*CDF3*	*LZF1*
20	*FHL1**	*TEC1*	*KLF10*	*NR1D1**	*RVE2*	*HB-12*
21	*NUT1*	*STB1**	*KLF13*	*BNC2*	*AT1G26790*	*RVE8**
22	*ASG1*	*YHP1**	*ESR1*	*NPAS3*	*BBX16*	*ATCTH*
23	*TBF1*	*RPI1*	*STAT5B*	*2210418O10RIK*	*COL9*	*MYBL2*
24	*SNF5*	*MTH1*	*SREBF1*	*HOXC6*	*LZF1*	*ARF11*
25	*WTM2*	*RIF1*	*MAFB*	*TBX1*	*ARF10*	*RL6*
Recall	76.5%	70.6%	66.7%	28.6%	36.4%	45.5%

LL_LDHC: Constant light and temperature; LDHC: 24 hour cycling light and temperature; MA: Microarray; RNA: RNAseq

*Core transcription factors in Additional file 2—Core Genes

mouse circadian clock with well-sampled data, and 45.45% for the core plant circadian network under circadian conditions. Meaning, by using only the dynamics of transcript abundance and a list of TFs, an experimentalist would identify three-quarters of the known core cell-cycle TFs in yeast, two-thirds of the core circadian TFs in mice, and almost half of the core circadian TFs in plants from among the top 25 TFs when ranked using a combined measure of periodicity and regulation strength. Other combined measures perform skillfully when examining the top 25 ranked TFs, although not as consistently well across all the datasets as DL × JTK (Additional file 5: Tables S2 and S3).

The ability of dynamic characteristics to identify core TFs from among all TFs may depend on the data collection modality and will certainly depend on the number of time points per cycle collected. This is made apparent by comparing the *S. cerevisiae* RNASeq and microarray datasets and, separately, *M. musculus* RNASeq and microarray datasets. We expect that the reduced DL × JTK classifier performance is largely due to the sensitivity of the JTK algorithm to the number of timepoints per cycle [10], although we cannot conclusively rule out the impact of the data type.

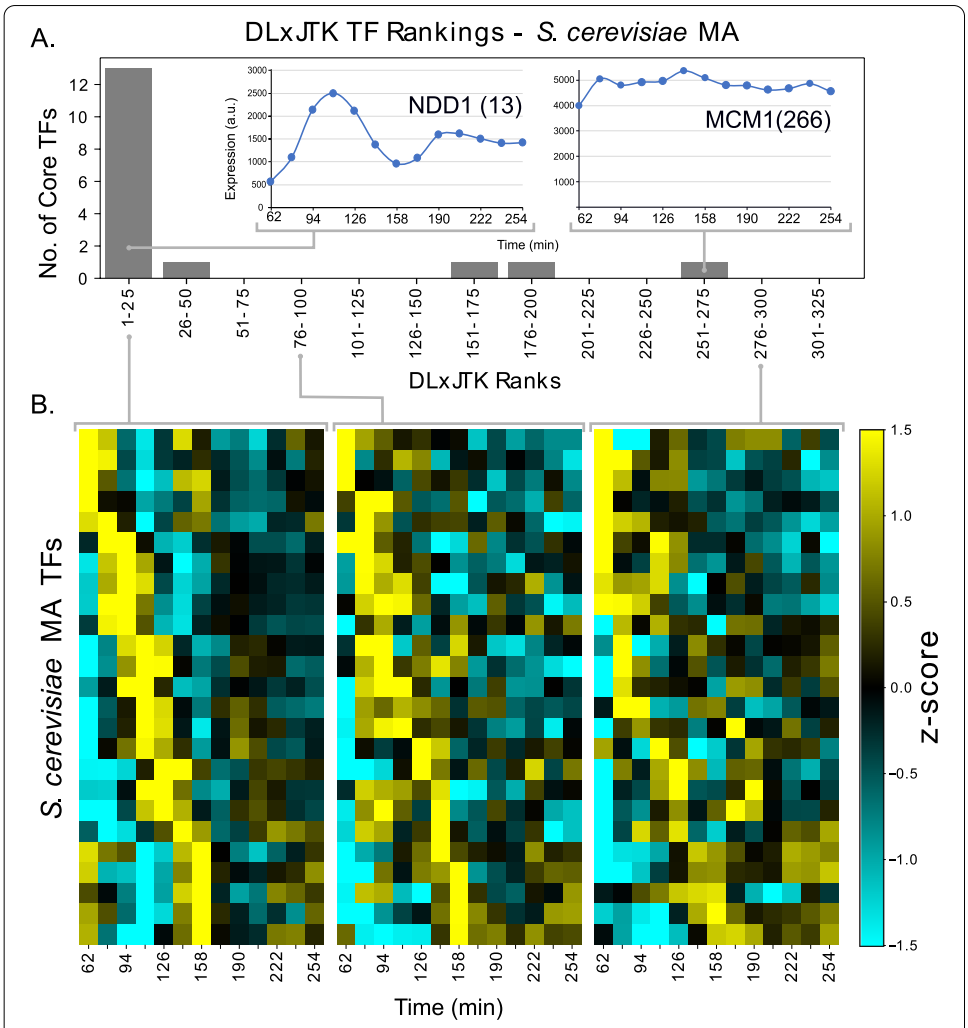

Fig. 3 Transcript abundance dynamics across DL × JTK rankings of transcription factors. **A** Distribution of DL × JTK ranks of core *S. cerevisiae* TFs among all TFs and time series expression of two core TFs: NDD1, which is highly ranked (rank 13), and MCM1, which is not highly ranked (rank 266). NDD1 and MCM1 act in a complex to regulate downstream targets. **B** Heatmaps of standardized gene expression profiles of the genes ranked (left) 1–25, (middle) 76–100, and (right) 276–300 by DL × JTK. Within each subpanel, genes are ranked by peak expression

At the same time, quantitative measures of rhythmicity in transcript abundance and strength of regulation both independently improve the skill of a classifier above random. Thus, the functional regulatory elements driving very different biological processes exhibit common characteristics in the dynamics of their transcript expression.

Dynamic transcript abundance characteristics remain adept at identifying core regulatory elements, even in the absence of prior knowledge of transcription factors

The organisms chosen for this study are model organisms in mammalian, plant, and fungi research which have been extensively studied. Thus, for these organisms, there are reliable annotations of gene function and comprehensive lists of TFs. If studying a non-model organism, evidence of gene function may be much weaker, for example relying

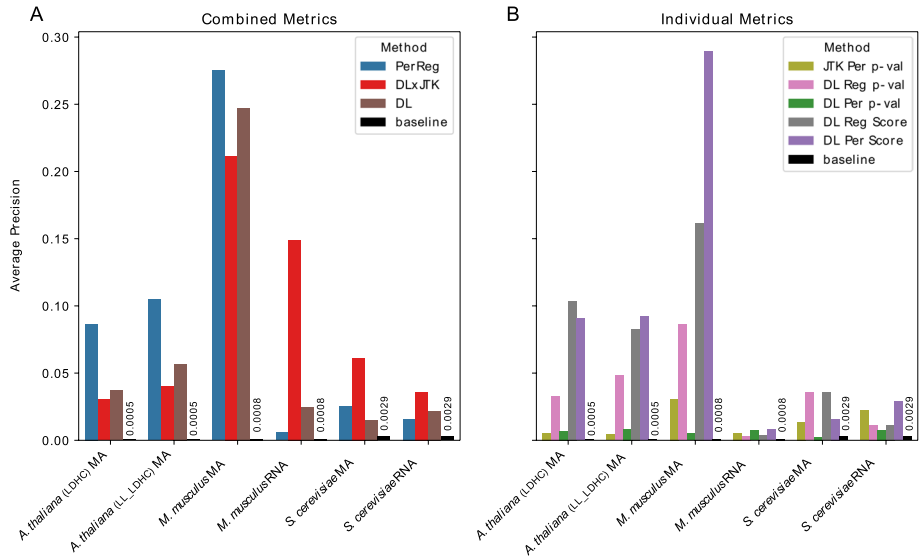

Fig. 4 Identifying core genes among all genes. Average precision of classifiers identifying core from non-core TFs among all genes by **A** combined metrics and **B** individual metrics (Table 3) as well as the baseline average precision of a random classifier, for each dataset (Table 4)

on sequence-based inferences. We ask, to what extent do the dynamic characteristics of transcript abundance that distinguish core TFs from non-core TFs continue to distinguish core from all genes? In this way, we assess the capacity for gene expression dynamics to reduce hypothesis space in the absence of any prior biological knowledge. Note, this is an extremely lofty goal given the minuscule fraction of these genomes occupied by core transcriptional regulator elements.

For each dataset in Table 4 we ranked all transcript abundance profiles using the methods in Table 3. We have chosen to be very conservative in our labelling of core genes: only 17 out of nearly 6000 transcripts in *S. cerevisiae*, 14 out of close to 20,000 genes in *M. Musculus*, and 11 of over 22,000 genes in *A. thaliana*. As expected, AP scores are greatly reduced across all datasets. However, the APs remain significantly above baseline in most cases (Fig. 4). Examining the top 25 genes ranked by the measure DL $\times$ JTK, at least one core TF remained in the top 25 for all datasets, except the *A. thaliana* LDHC microarray dataset (Additional file 4—Gene Rankings). Remarkably, six of the 15 core mouse circadian TFs (recall of 40%) are identified among the top 25 genes ranked by DL $\times$ JTK in the *M. Musculus* liver microarray dataset.

The dynamic transcript abundance characteristics of core regulatory elements are not overrepresented among transcription factors

It is certainly possible that the dynamic features under investigation are characteristic of TFs themselves, and thus filtering on TFs selects for these features. To investigate the possibility that the dynamic metrics in this study are overrepresented in TFs and not just core transcriptional regulatory elements, we assessed the ability of the dynamic characteristics of transcript abundance to identify TFs from among all transcripts. In line with our hypothesis, all methods listed in Table 3 performed poorly as each method's AP dropped to near or below the AP baseline (Fig. 5). Said another way, TFs within these organisms are effectively randomly distributed in the rankings of all genes by periodicity

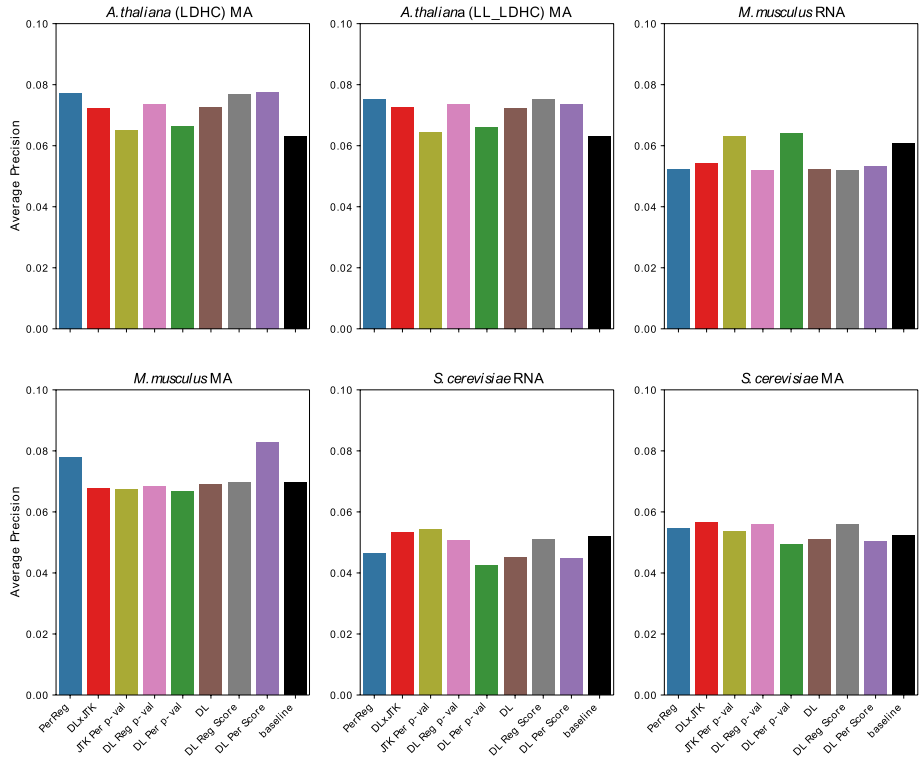

Fig. 5 Identifying transcription factors among all genes. Average precision of classifiers identifying TFs from non-TFs among all genes by combined metrics and individual metrics (Table 3) as well as the baseline average precision of a random classifier, for each dataset (Table 4)

and variability of transcript abundance. The inability of the methods to identity TFs in each dataset demonstrates that these dynamic features are not characteristic of TFs in general, although they are indicative of core regulatory elements in disparate circadian systems and in the yeast cell-cycle.

Statistical significance measures are not required to skillfully rank core genes

A major concern with the DL methods for determining significance is that they require the generation of empirical null distributions derived from the periodicity and regulator metrics of many synthetic expression profiles generated by repeated sampling of the experimental data. As the number of genes and/or the number of time points increases, the background distributions of potential random synthetic abundance profiles grows rapidly. As a result, in general, many more synthetic profiles must be generated and characterized to improve estimates of these p-values. If too few random curves are analyzed, there may be ambiguity in the final rankings due to repeated p-values caused by the resulting coarse discretization of possible estimates. This is potentially an issue since ambiguous p-value rankings could, in principle, overstate the quality of the metric. In the worst case scenario an experimentalist would have to test all genes in a block with the same p-value, since one cannot prioritize by this method alone one gene over another. Additionally, the choice of a background distribution has a large impact on statistical significance [11] and gives poor results when assumptions of the background

distribution do not match the reality of the data (see the discussion of the malaria dataset in [12]).

It should also be noted that, unlike some test statistics, the DL Per Scores need not rank genes in exactly the same way as the corresponding DL Per *p*-values. Thus we ask, is it necessary to compute a significance value in order to skillfully rank core TFs? We address this question by ranking genes according to DL's "naive" measurements for periodicity and regulation, individually (DL Per Score and DL Reg Score in Table 3, respectively) and in combination (PerReg). These naive measurements are calculated quickly with no permutations or random sampling required, and thus greatly reduce the computational time required to rank genes. When used individually, the naive DL measurements perform equally well or better than the empirical *p*-values at identifying core, as measured by AP (Fig. 2B). Indeed, there is a striking difference across all datasets in the ranking of core genes using DL's naive periodicity score rather than its associated empirical *p*-value, which is particularly expensive to compute for large gene sets.

When combined, the naive measures also skillfully rank genes well above baseline across all datasets. In fact, there is a notable increase in AP over the other combined metrics, which are derived from *p*-values, for the *A. thaliana* data in both conditions (Fig. 2A). We expect that this, along with the generally lower performance of these metrics on *A. thaliana* data compared to the other datasets, may be due to the fact that the *A. thaliana* transcript abundance profiles reflect gene expression in multiple tissue types, making it difficult to collect accurate empirical *p*-values.

Much like DL × JTK, PerReg shows skillful recall at identifying core genes among the top 25 TFs (Additional file 5: Table S3), identifying at least 4 and at most 10 core TFs among the top 25, across all datasets considered in this study.

Several high ranking non-core genes display regulatory relationships with core genes

The lists of core TFs used in this study are conservative since (1) a lack of strong evidence supporting a gene as a core regulator is not proof that it is not core and (2) many functional regulators are also known to be transcriptional co-regulators and post-transcriptional modifiers; we labelled the latter as non-core to ensure fair assessment of the performance of the ranking methods. Thus, our binary labels may contain false negatives (core labeled as non-core) due to a lack of strong experimental evidence, and certainly contain false negatives due to our restriction to TFs. We ask, what are the identities of the most highly ranked non-core TFs, and does there exist any evidence that they target the activity of and/or are targeted by our core TFs?

Utilizing the curated list of regulatory relationships in YEASTRACT [13] and PlantTFDB [14], as well as a literature search for *M. musculus* TF interactions, we indeed observe evidence that several yeast, plant, and mouse genes among the top 25 TFs ranked by the measure DL × JTK target core and/or are targeted by core (Table 2). For example, we find that among the top 25 *S. cerevisiae* TFs ranked by DL × JTK in either MA or RNASeq datasets, that 40% (9/23) of the genes have existing evidence of both regulating and being regulated by core. This observation suggests that genes that appear highly ranked by our combined measures, but were not labeled as core due to a lack of existing evidence, may in fact be core nodes.

Table 2 Interaction relationships* between core TFs and non-core that appear in the top 25 TFs as ranked by DL × JTK†

S. cerevisiae			*M. musculus*			*A. thaliana*		
Gene	**Targeted**	**Targets**	**Gene**	**Targeted**	**Targets**	**Gene**	**Targeted**	**Targets**
ASG1	FHL1	*NDD1*	*EGR1*	ARNTL [15]	*ARNTL* [16]	*EPR1*	RVE4	*PRR5*
EDS1	FHL1	*TOS4*	*KLF10*	ARNTL [17]	*ARNTL* [18]	*PIF4*	CCA1	*LHY*
GAT1	ACE2	*ACE2*	*NFIC*	HLF [19]		*PIL6*	CCA1	*LHY*
MTH1	FHL1	*STB1*	*ATF5*	CLOCK [20]		*ARF11*	CHE	
RME1	ACE2	*ASH1*	*ESR1*	CLOCK [21]		*CO*	CCA1	
RPI1	FHL1	*NDD1*	*SREBF1*		*BHLHE40* [22]	*COL1*	CHE	
SIP4	FHL1	*STB1*			*BHLHE41* [22]	*COL9*	CHE	
TEC1	SWI4	*ASH1*				*MYBL2*	CHE	
WTM2	ACE2	*STB1*				*CDF2*		*LHY*
ASF1	SWI4					*RVE1*		*PRR5*
HST3	FKH1					*RVE2*		*CCA1*
HST4	MBP1							
POG1	MCM1							
RLF2	MBP1							
RTT107	MCM1							
SNF5	ACE2							
TBF1	FHL1							

**S. cerevisiae* and *A. thaliana* interactions determined respectively by database searches of [13] and [14] and represent a range of direct and indirect evidence types, including the presence of binding motifs in regulatory regions and response to TF over-expression. *M. musculus* interactions determined by evidence gathered in the associated citation

†*M. musculus* non-core TFs drawn from MA dataset only, while non-core *S. cerevisiae* and *A. thaliana* TFs were drawn from the unions of each pair of analyzed datasets

Table 3 Quantitative metrics of periodicity and regulation strength used in this study to rank genes

Name	Function	Type	Description
DL Per Score	Per(G)	Periodicity	A measure of abundance profile periodicity as defined by Eq. (3)*
DL Per *p*-val	$p_{per}(G)$	Periodicity	An empirical *p*-value measuring the probability that a random abundance profile will exhibit a DL Per Score larger than the actual gene's expression pattern
JTK Per *p*-val	$p_{jtk}(G)$	Periodicity	An analytic *p*-value introduced in [41] measuring the correlation in the discrete up-down patterns of expression between a gene and a sinusoidal template
DL Reg Score	Reg(G)	Regulation	A measure of the variability of transcript abundance about its mean expression level as defined by Eq. (2)*
DL Reg *p*-val	$p_{reg}(G)$	regulation	An empirical *p*-value measuring the probability that a random abundance profile will exhibit a DL Reg Score larger than the actual gene's
PerReg		Combined	The product of DL Per and DL Reg Scores
DL		Combined	The original periodicity measure introduced in [42] and defined according to Eq. (1)*
DL × JTK		Combined	A modified version of the original periodicity measure introduced by [42], defined according to Eq. (1)* with $p_{per}(G)$ replaced by $p_{jtk}(G)$

*Refer to Additional file 5: Supplementary Information for equation definitions

Within the top 25 of all genes, as ranked by DL × JTK, we observe a number of regulatory elements that are known to be essential to produce the given periodic program of gene expression, but which are not strictly TFs, and therefore do not qualify in our definition as a core gene. Examples include the mouse transcriptional

Table 4 Time series transcript abundance datasets used in this study

Organism	***S. cerevisiae***		***M. musculus* (liver)**		***A. thaliana* (whole leaf)**	
Synch. in	Cell cycle	Cell cycle	Circadian	Circadian	Diurnal	Circadian
Technology	RNASeq	Microarray	Microarray	RNASeq	Microarray	Microarray
Period	75 min*	94 min*	24 h	24 h	24 h	24 h
Duration	245 min	254 min	48	42	48	48
Frequency	5 min	16 min	2 h	6 h	4 h	4 h
Timepoints/cycle	15	5.875	12	4	6	6
Reference	[53]	[6]	[29]	[29]	[26] (LL_LDHC)	[26] (LDHC)
No. of genes†	5910	5718	19,750	18,388	22,484	22,484
No. of TFs†	304	307	1373	1118	1415	1415
No. of core	17	17	15	14	11	11

LL_LDHC: Constant light and temperature; LDHC: 24 hour cycling light and temperature

*Cell-cycle period length was taken from the respective publication, which estimated period length using the CLOCCS algorithm [54]

† Counts are based on post-processed datasets (see Materials and Methods)

co-regulators Period 3 (PER3) [23] and Cryptochrome 1 (CRY1) [24] and the plant post-transcriptional gene Gigantea (GI) [25] (Table 2), which are known or proposed to be transcriptional co-regulators and post-transcriptional elements. This supports our conclusion that core elements, even beyond the TFs, can be identified by quantifiable features in their transcript abundance dynamics. Improvement in the annotation of non-TF regulatory elements is needed before we can reliably quantify the extent to which these dynamic characteristics are exhibited by all nodes of these networks at the level of transcript abundance.

External periodic signals do not significantly alter the skill of transcript abundance dynamics at identifying core genes

Implicit in the definitions of the core transcriptional regulatory networks considered in this study is that they are free-running and can support rhythmic oscillations in the absence of external periodic stimuli due to their mutual regulatory interactions with other core elements. Is it necessary to collect time series transcriptomics in the absence of external circadian stimuli to skillfully identify core regulatory elements?

To address this question, we compared the skill of dynamic expression features to identify the core TFs for *A. thaliana* in (1) periodically fluctuating light and temperature (diurnal) conditions (LDHC) and (2) constant light, (circadian) conditions (LL_LDHC). For the details on the precise experimental setup see [26].

One might expect that the transcript dynamics of diurnal non-core genes—those that are strictly driven by periodic light-dark and/or temperature cycles—would reduce the capacity of dynamic gene expression features to distinguish core regulatory elements. We find that the signal of core genes is not degraded in the presence of external periodic stimuli in these experiments, since all combined quantitative measures show nearly identical skill at identifying core genes across both conditions (Fig. 2A). Even more striking is the consistency in the individual ranks of core genes across diurnal and circadian conditions, as shown for DL × JTK in Additional file 5: Table S1.

Conclusions

Elucidating the underlying GRNs driving dynamic biological processes, such as cell-division and sleep-wake cycles, is crucial if we are to leverage existing control mechanisms for synthetic biology applications, understand the evolution of biological networks, and inform experiments to discover new drug targets. However, experimentally identifying the core regulatory elements of these gene networks can be costly, time consuming and daunting, even for the simplest organisms, due to the large hypothesis space. We have shown that many core transcriptional regulators governing different periodic processes, appearing in evolutionarily distinct organisms, share common features in their transcript abundance dynamics. These findings indicate that cell and circadian cycle GRNs share functionally and/or evolutionarily conserved features. We demonstrated the use of several metrics that quantify and combine these dynamic features. The outcome is a substantial reduction in hypothesis space: a prioritization of gene targets for experimental validation, which may accelerate the discovery of the core control variables of gene regulatory networks.

High degrees of periodicity and strong regulation signals appear to be characteristic features of many core TFs involved in generating periodic biological processes. However, not all known core regulatory TFs strongly exhibit the dynamic features quantified here at the level of their transcript abundance. For instance, the abundance profile of the core *S. cerevisiae* TF *NDD1* is highly periodic with a precise match to cell-cycle period and exhibits large dynamic range, but *MCM1* does not show convincing oscillations at the cell-cycle period (Fig. 3A). *MCM1* is the only core TF to not rank in the top 70 TFs in at least one of the two *S. cerevisiae* datasets using DL × JTK (Additional file 5: Table S1). However, MCM1 acts in complex with other rhythmically-expressed genes like NDD1 [27, 28], so it can still be part of a highly periodic TF complex without itself exhibiting highly periodic signatures in transcript abundance. It is enticing to imagine there may be other features captured in the gene or protein expression profiles, as well as features not related to gene expression, such as sequence-based and protein interaction features that could be used to more accurately capture all core genes, including those identified in TF complexes.

It is known in the circadian field that several core clock genes have tissue-specific periodic properties in mice [29]. Thus, we expect not all core genes will rise to the top of our rankings in every tissue. For example, within the three retinoid-related orphan receptors (RORs) TFs, *RORA*, *RORB*, and *RORC*, only *RORC* is known to display periodic gene expression in mouse liver [30]. Indeed, only *RORC* was ranked in the top 25 TFs ranked by DL × JTK (Table 1) in the mouse liver microarray dataset. Another example is the mouse core clock gene *ARNTL2*, which is not ranked highly in the mouse liver datasets. Most studies suggest *ARNTL2* has brain-specific circadian expression with lower levels of expression in the liver in mammals [31–33]. There is also growing evidence for genes to exhibit tissue-specific dynamics in plants [34].

Our ability to identify plant core genes appears generally lower than the other organisms we considered. This may be due to the fact that samples were taken from the whole leaf and thus contained a mixture of multiple tissue types such as mesophyll, epidermis, and vasculature [26]. The abundance and periodicity of any particular transcript might therefore appear muted as genes are likely expressed differentially across tissues.

Consistent with this hypothesis, several studies have shown that tissue-specific clocks in plants can be asymmetrically coupled [35], have different period lengths [36], or have different levels of gene expression for core components [37, 38]. Naturally it is more difficult to identify a core component whose observed dynamics is either a convolution of multiple dissimilar abundance profiles derived from multiple tissues or has specificity to an under-represented tissue in a mixture of tissue types. Interestingly, the dominant tissue type in whole leaf samples is mesophyll, and morning-expressed clock genes (*CCA1*, PRRs, and *LHY*) are highly expressed in the mesophyll [35, 39]. These morning-expressed genes are mostly the only plant core genes ranked highly in this study (Table 1).

Broadly speaking, our findings suggest that even naive measures of periodicity and regulatory strength can be used to skillfully rank genes. We speculate that other methods that quantify and combine these two features will show similar skill at ranking core above non-core genes. With the availability of proper experimental controls across organism, platform, sampling density, etc., it might be possible to compare the various metrics to make a more prescriptive recommendation of which particular method to use for a given dataset.

The use of naive metrics rather than empirical *p*-values does not suffer from ambiguous rankings caused by insufficient sampling of the null distribution, as may be the case with DL's method of measuring significance. It is possible to reduce the ambiguity of a ranking by increasing the sampling of the null distribution at the cost of increased compute time. The disambiguation of empirical regulator *p*-values computed by the DL metric through increased sampling is visualized in Additional file 5: Fig. S7. Similarly, combining several *p*-values derived from different dynamic characteristics into combined metrics can eliminate ambiguous rankings that may be present in one of these features.

We have demonstrated the importance of reliable genome annotation of TF genes, but many organisms of interest currently lack comprehensive gene annotations. Thus it is desirable to have methods that can leverage high-throughput technologies to provide evidence of gene function. Additional evidence such as identifying DNA-binding domains and/or orthology to known TFs in other organisms are two such methods that could be used to provide putative TF lists for poorly-annotated genomes.

Here we demonstrate that dynamic features of periodic transcriptomes appear to be conserved across kingdoms and networks that appear to serve disparate functions such as cell-cycle or circadian clocks. It is possible that the conservation of these features results from a fundamental property of these GRNs, where a transcriptional signal is developed within a core set of nodes and that the signal degrades as it is propagated through effector nodes that control downstream gene expression. Alternatively, the conservation of features could reflect an evolutionary conservation of network topologies that produce rhythmic behaviors during circadian and cell cycles.

Methods

Dynamic curve features

We focused on two dynamic curve features of transcript abundance profiles: (1) periodicity at a specified period and (2) amplitude. Although amplitude has been suggested as

a feature of core genes in mouse circadian GRN [40], to the best of our knowledge, this feature has not been articulated for core nodes of cell-cycle or plant circadian GRNs.

Several algorithms have been published that quantify one or both of these two features [41–48] and several studies have performed benchmarking of the metrics used by these algorithms to quantify the dynamic features [10, 49, 50]. The consensus of the benchmarking studies is that there is no one best metric, as individual metrics each have various underlying definitions of these two dynamic features. Furthermore, when selecting a metric, one must take into account the characteristics of their dataset, e.g., noise, number of cycles, sampling frequency, etc., and whether these characteristics fit the algorithm's definitions. Of the numerous algorithms to choose from, we selected JTK-CYCLE (JTK) [41] and de Lichtenberg (DL) [42] as they have been shown to perform reasonably well across datasets with diverse characteristics [10].

JTK's metric for measuring how well a transcript abundance profile fits to a specified period is based on correlating the profile to that of a reference curve that oscillates at the specified period, and then computing the significance of the correlation, using a non-parametric test that can capture non-linear correlations. DL's metric for measuring periodicity of a transcript abundance profile combines statistical measures of fit to a specified period and strength of regulation. DL's strength of regulation is a measure of variability within the transcript abundance profile, and can be thought of as a measure of amplitude. To reduce any potential confusion between this study and any studies that also use DL, we use "strength of regulation" as opposed to "amplitude" as this is the same language used in the original DL study. The JTK and DL metrics used in this study are summarized in Table 3. Detailed descriptions of the algorithms used to compute these metrics are available in Additional file 5.

Performance of gene ranking metrics

The problem of identifying the core regulatory elements within an organism's genome is fundamentally a question of binary classification of gene function: is a gene core or not? In practice, this decision task amounts to ranking all genes by some quantitative metric or "score" in the hope that the ranking is enriched with core genes, so as to reduce the expected effort required to gather additional experimental evidence of gene function through, for example, knock-out experiments.

To assess the capacity of each ranking metric given in Table 3 to rank core genes above non-core genes, we compute the precision-recall (PR) curves of the gene rankings. PR curves track the precision (the fraction of true core genes among all genes ranked above some score threshold) across all levels of recall (the fraction of true core genes appearing above the chosen threshold). From each PR curve we compute the average precision (AP), which summarizes with a single number a ranking's performance across all recall levels. See Additional file 5—Supplementary Information for a precise definition of PR curves, precision, recall and AP.

Any ranking can achieve a perfect recall of 1 if the threshold is chosen so permissively as to label all genes as core. However, given the goal to reduce hypothesis space and limit the amount of experimental validation needed to identify core regulatory elements, a permissive choice of threshold is of little practical utility. Thus, in this context, a meaningful measure of classifier skill is the precision at a given recall. For example,

the precision at a recall of 10% characterizes how many knock-out experiments one would expect to perform, in accordance with a given algorithm's ranking of genes, before 10% of the core regulatory elements are identified. It is this interpretation that may be of particular value to a researcher interested in using a ranking algorithm to prioritize experiments.

In some rare cases, if a scoring algorithm is particularly discriminating between two classes, the scores may be bimodally distributed and well-separated, allowing a data-driven justification of a choice of threshold. Usually, this is not the case, and a threshold must be chosen arbitrarily. Moreover, it is known that periodicity scores produced by the methods used in this study depend on attributes of the data that may vary from one experiment to the next, e.g. number of time points per cycle [10], and that there is no universal threshold to distinguish periodic from non-periodic genes [51]. Thus, better measures of classifier performance, such as AP, assess the ranking itself, quantifying the skill of the classifier to rank the members of the true class (core) above the members of the other class (non-core).

A perfect ranking of genes is one in which all core genes are ranked higher than all non-core genes. In this way, an experimentalist prioritizing hypotheses using the gene ranking would encounter all core genes before testing any non-core. The AP of a perfect ranking will be 1. At the other extreme is an uninformative ranking which assigns scores to genes at random. The average precision achieved for a random classifier is C/N [52], where C is the number of core genes and N is the number of all genes. Moreover, the expected PR curve for such an algorithm is a horizontal line at precision level C/N across all recall levels, as seen in Additional file 5: Figs. S1–S6. Thus, performance of each classifier, as measured by its PR curve and its AP, should be compared against the (non-universal) baseline performance of a random classifier. In other words, precision-recall points above the baseline reflect the skill of a metric, over the random classifier, to rank genes in a way which enriches the top of a list with core genes.

Gene expression datasets

Data processing

The normalized transcriptomic datasets used in this analysis were taken from the references presented in Table 4. The datasets were adjusted to account for possible technical and biological variations between samples by the authors of the studies that generated them. For the specific normalization applied to each dataset, we refer the reader to the references cited in Table 4. Before deriving dynamic features, transcript abundances were processed to remove unreliable data. For the *M. musculus* and *S. cerevisiae* RNAseq datasets, genes were removed that had less than 1 FPKM normalized transcript level in more than half of the measured time points and were not considered in any part of this analysis.

Authors of [6] produced the *S. cerevisiae* microarray dataset from *S. cerevisiae* cells that were synchronized via centrifugal elutriation. It is known that elutriation impacts the transcription of many genes and that a brief recovery period is needed after elutriation. The resulting transcript abundance dynamics early in the time series, which are not related to cell-cycle transcript abundance dynamics, can impact periodicity analyses [54]. Therefore, prior to any analysis, [6] eliminated data determined to be associated

with the elutriation recovery period. We adopted the same method of eliminating the first two time points from the *S. cerevisiae* microarray dataset.

In the *S. cerevisiae* mircoarray dataset and both *A. thaliana* datasets, some genes were associated with multiple probes, causing some genes to have more than one transcript abundance profile. The *A. thaliana* core gene, *RVE8*, was one such gene. Having two transcript abundance profiles for *RVE8* resulted in inaccurate performance metrics. To remedy this issue, we applied a filtering step to the *S. cerevisiae* mircoarray dataset and both *A. thaliana* datasets after quantifying dynamic features using the methods in Table 3. For genes with multiple abundance profiles, we kept the profile with the highest average abundance, resulting in the elimination of 96 and 326 profiles from the *S. cerevisiae* mircoarray dataset and both *A. thaliana* datasets, respectively. All time series data can be found in Additional file 1—Gene Expression Data.

Curation of Core Regulatory Elements

In order to evaluate the ability of each method given in Table 3 to identify core TFs driving a periodic program of gene expression, we consider data derived from well-studied organisms for which there is significant experimental evidence of gene function. Core cell-cycle TFs in yeast are described as genes functioning in an autoregulatory transcriptional network that robustly maintains a large program of periodic gene expression [4–6, 8]. A list of yeast core cell-cycle TFs based on this definition was compiled in [9] for evaluating the transcriptonal oscillator underlying the yeast cell cycle. Therefore, the core TF list defined in [9] was used in this study as the ground truth for *S. cerevisiae* (Additional file 2—Core Genes). Similarly, core circadian clock TFs are described as genes functioning in an autoregulatory transcriptional feedback loop, maintaining circadian-like transcript abundances under constant light or dark conditions and are necessary components for generation and regulation of circadian rhythms [1, 55, 56]. The literature evidence supporting our labeling of plant and mammalian genes as core are listed in Additional file 2—Core Genes. Although the core networks are known to include non-TF regulatory elements that control functional activity, such as kinases and ubiquitin ligases [1, 56, 57], we limit our definition of core to TFs since these are more reliably annotated in the genomes we consider. This ensures our conclusions are conservative by not unfairly inflating the core list with known core post-transcriptional modifiers while not simultaneously including all non-core members of these gene categories.

Curation of transcription factors

In this study, we define a TF as a gene that has the ability for sequence-specific DNA binding alone or in a complex and is capable of activating and/or repressing gene expression. This definition excludes genes that are also known to affect gene expression, such as chromatin-related genes like chromatin remodeling factors, histone demethylases, and histone acetyltransferases. To ensure the lists of TFs are consistent across strains, we used curated TF databases that use the given TF definition. In particular, TFs used in this study (Additional file 3—Transcription Factors) were retrieved from Animal TF Database 3.0 [58], Plant TF Database 4.0 [14], and YEASTRACT [13] for *M. musculus*, *A. thaliana*, and *S. cerevisiae*, respectively. Each species list of TFs was inspected for presence of the respective species core regulatory elements. Upon inspection of the *A. thaliana* TF list, it was discovered

that the core regulatory elements from the pseudo-response regulator (PRR) family were not present. Therefore, we added *PRR5*, *PRR7*, *PRR9*, and *PRR1* (*TOC1*) to *A. thaliana* list of TFs, which are known as core regulatory elements of the plant circadian clock [59–61].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-022-04627-9.

Additional file 1. Gene expression data: an EXCEL file containing gene expression profiles for each dataset used in this study.

Additional file 2. Core genes: an EXCEL file containing lists of core genes for all organisms.

Additional file 3. Transcription factors: an EXCEL file containing lists of transcription factors for all organisms.

Additional file 4. Gene rankings: an EXCEL file containing the rankings of all genes by each metric for all datasets.

Additional file 5. Supplementary information: A PDF document with additional files including mathematical details, Figs. S1–S7 and Tables S1–S3.

Acknowledgements
Not applicable.

Authors' contributions
We adopt the CASRAI Contributor Roles Taxonomy (https://casrai.org/credit/) to specify author contributions. FCM and RCM contributed equally to this work. All authors read and approved the final manuscript. FCM: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing—original draft, Writing—review and editing. RCM: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing—original draft, Writing—review and editing. BC: Conceptualization, Writing—original draft, Writing—review and editing AD: Conceptualization, Writing—review and editing SBH: Conceptualization, Data curation, Funding acquisition, Supervision, Writing—original draft, Writing—review and editing. All authors read and approved the final manuscript.

Funding
FCM, AD, BC and SBH were supported by the Defense Advanced Research Projects Agency award number D12AP00025. RCM, BC and SBH were supported by the National Institutes of Health award number 1R01GM126555-0, and the National Science Foundation award number DMS-1839299. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials
The datasets supporting the conclusions of this article are available in the Gene Expression Omnibus repositories, [62–64], and the lab Mockler lab ftp site [65]. The datasets supporting the conclusions of this article are also included within the article and its additional files. To ensure the reproducibility of results, the datasets analyzed during the current study and the code used to analyze these datasets and to generate figures and tables are also available in the Gitlab repository [66].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Mathematical Sciences, Florida Atlantic University, 777 Glades Rd, Boca Raton, FL 33431, USA. [2]Department of Biology, Duke University, 130 Science Drive, Durham, NC 27708, USA. [3]Department of Mathematical Sciences, Montana State University, P.O. Box 172400, Bozeman, MT 59717, USA. [4]Geometric Data Analytics, 343 W Main St, Durham, NC 27701, USA.

Received: 13 July 2021 Accepted: 1 March 2022
Published online: 17 March 2022

References

1. Harmer SL. The circadian system in higher plants. Annu Rev Plant Biol. 2009;60(1):357–77. https://doi.org/10.1146/annurev.arplant.043008.092054 (**PMID: 19575587.**).

2. Brunner M, Schafmeier T. Transcriptional and post-transcriptional regulation of the circadian clock of cyanobacteria and Neurospora. Genes Dev. 2006;20:1061–74. https://doi.org/10.1101/gad.1410406.
3. Panda S, Hogenesch J, Kay S. Circadian rhythms from flies to human. Nature. 2002;417:329–35. https://doi.org/10.1038/417329a.
4. Bristow SL, Leman AR, Kovacs LAS, Deckard A, Harer J, Haase SB. Checkpoints couple transcription network oscillator dynamics to cell-cycle progression. Genome Biol. 2014;15(9):446. https://doi.org/10.1186/s13059-014-0446-7.
5. Simmons-Kovacs L, Mayhew M, Orlando D, Jin Y, Li Q, Huang C, et al. Cyclin-dependent kinases are regulators and effectors of oscillations driven by a transcription factor network. Mol Cell. 2012;45(5):669–79.
6. Orlando DA, Lin CY, Bernard A, Wang JY, Socolar JE, Iversen ES, et al. Global control of cell-cycle transcription by coupled CDK and network oscillators. Nature. 2008;453(7197):944.
7. Haase SB, Reed SI. Evidence that a free-running oscillator drives G1 events in the budding yeast cell cycle. Nature. 1999;401:394–7. https://doi.org/10.1038/43927.
8. Cho CY, Kelliher CM, Haase SB. The cell-cycle transcriptional network generates and transmits a pulse of transcription once each cell cycle. Cell Cycle. 2019;18(4):363–78.
9. McGoff KA, Guo X, Deckard A, Kelliher CM, Leman AR, Francey LJ, Edge The Local. Machine: inference of dynamic models of gene regulation. Genome Biol. 2016;17(1):214. https://doi.org/10.1186/s13059-016-1076-z.
10. Deckard A, Anafi RC, Hogenesch JB, Haase SB, Design Harer J. Analysis of large-scale biological rhythm studies: a comparison of algorithms for detecting periodic signals in biological data. Bioinformatics. 2013;29(24):3174–80. https://doi.org/10.1093/bioinformatics/btt541.
11. Futschik ME, Herzel H. Are we overestimating the number of cell-cycling genes? The impact of background models on time-series analysis. Bioinformatics. 2008;24(8):1063–9. https://doi.org/10.1093/bioinformatics/btn072.
12. Kallio A, Vuokko N, Ojala M, Haiminen N, Mannila H. Randomization techniques for assessing the significance of gene periodicity results. BMC Bioinform. 2011;12(1):330. https://doi.org/10.1186/1471-2105-12-330.
13. Teixeira MC, Monteiro PT, Palma M, Costa C, Godinho CP, Pais P, et al. YEASTRACT: an upgraded database for the analysis of transcription regulatory networks in Saccharomyces cerevisiae. Nucleic Acids Res. 2017;46(D1):D348–53. https://doi.org/10.1093/nar/gkx842.
14. Jin J, Tain F, Yang DC, Meng YQ, Kong L, Luo J, et al. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Res. 2016;45(D1):D1040–5. https://doi.org/10.1093/nar/gkw982.
15. Lee JH, Sancar A. Circadian clock disruption improves the efficacy of chemotherapy through p73-mediated apoptosis. Proc Natl Acad Sci. 2011;108(26):10668–72.
16. Riedel CS, Georg B, Jørgensen HL, Hannibal J, Fahrenkrug J. Mice lacking EGR1 have impaired clock gene (BMAL1) oscillation, locomotor activity, and body temperature. J Mol Neurosci. 2018;64(1):9–19. https://doi.org/10.1007/s12031-017-0996-8.
17. Guillaumond F, Gréchez-Cassiau A, Subramaniam M, Brangolo S, Peteri-Brünback B, Staels B, et al. Krüppel-Like Factor KLF10 is a link between the circadian clock and metabolism in liver. Mol Cell Biol. 2010;30(12):3059–70.
18. Hirota T, Kon N, Itagaki T, Hoshina N, Okano T, Fukada Y. Transcriptional repressor TIEG1 regulates Bmal1 gene through GC box and controls circadian clockwork. Genes Cells. 2010;15(2):111–21. https://doi.org/10.1111/j.1365-2443.2009.01371.x.
19. Wahlestedt M, Ladopoulos V, Hidalgo I, Castillo MS, Hannah R, Säwén P, et al. Critical modulation of hematopoietic lineage fate by hepatic leukemia factor. Cell Rep. 2017;21(8):2251–63.
20. Lemos DR, Goodspeed L, Tonelli L, Antoch MP, Ojeda SR, Urbanski HF. Evidence for circadian regulation of activating transcription factor 5 but not tyrosine hydroxylase by the chromaffin cell clock. Endocrinology. 2007;148(12):5811–21.
21. Yoshitane H, Ozaki H, Terajima H, Du NH, Suzuki Y, Fujimori T, et al. CLOCK-controlled polyphonic regulation of circadian rhythms through canonical and noncanonical E-boxes. Mol Cell Biol. 2014;34(10):1776–87.
22. Lecomte V, Meugnier E, Euthine V, Durand C, Freyssenet D, Nemoz G, et al. A new role for sterol regulatory element binding protein 1 transcription factors in the regulation of muscle mass and muscle cell differentiation. Mol Cell Biol. 2010;30(5):1182–98.
23. Zhang L, Hirano A, Hsu PK, Jones CR, Sakai N, Okuro M, et al. A PERIOD3 variant causes a circadian phenotype and is associated with a seasonal mood trait. Proc Natl Acad Sci USA. 2016;113(11):E1536–44. https://doi.org/10.1073/pnas.1600039113 (**PMID: 26903630.**).
24. van der Horst GTJ, Muijtjens M, Kobayashi K, Takano R, Kanno SI, Takao M. Mammalian Cry1 and Cry2 are essential for maintenance of circadian rhythms. Nature. 1999;398:627–30. https://doi.org/10.1038/19323.
25. Mishra P, Panigrahi KC. GIGANTEA: an emerging story. Front Plant Sci. 2015;6:8. https://doi.org/10.3389/fpls.2015.00008 (**PMID: 25674098.**).
26. Mockler TC, Michael TP, Priest HD, Chen 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;72:353–63. https://doi.org/10.1101/sqb.2007.72.006.
27. Kelliher CM, Foster MW, Motta FC, Deckard A, Soderblom EJ, Moseley MA, et al. Layers of regulation of cell-cycle gene expression in the budding yeast *Saccharomyces cerevisiae*. Mol Biol Cell. 2018;29(22):2644–55. https://doi.org/10.1091/mbc.E18-04-0255 (**PMID: 30207828.**).
28. Koranda M, Schleiffer A, Endler L, Ammerer G. Forkhead-like transcription factors recruit Ndd1 to the chromatin of G2/M-specific promoters. Nature. 2000;406:94–8. https://doi.org/10.1038/35017589.
29. Zhang R, Lahens NF, Ballance HI, Hughes ME, Hogenesch JB. A circadian gene expression atlas in mammals: implications for biology and medicine. Proc Natl Acad Sci. 2014;111(45):16219–24.
30. Ueda HR, Chen W, Adachi A, Wakamatsu H, Hayashi S, Takasugi T, et al. A transcription factor response element for gene expression during circadian night. Nature. 2002;418:534–9. https://doi.org/10.1038/nature00906.
31. Ikeda M, Yu W, Hirai M, Ebisawa T, Honma S, Yoshimura K, et al. cDNA cloning of a novel bHLH-PAS transcription factor superfamily gene, BMAL2: its mRNA expression, subcellular distribution, and chromosomal localization. Biochem Biophys Res Commun. 2000;275(2):493–502.

32. Maemura K, de la Monte SM, Chin MT, Layne MD, Hsieh CM, Yet SF, et al. CLIF, a novel cycle-like factor, regulates the circadian oscillation of plasminogen activator inhibitor-1 gene expression. J Biol Chem. 2000;275(47):36847–51.
33. Hogenesch JB, Gu YZ, Moran SM, Shimomura K, Radcliffe LA, Takahashi JS, et al. The basic helix-loop-helix-PAS protein MOP9 is a brain-specific heterodimeric partner of circadian and hypoxia factors. J Neurosci. 2000;20(13):RC83–RC83.
34. Inoue K, Araki T, Endo M. Oscillator networks with tissue-specific circadian clocks in plants. Semin Cell Dev Biol. 2018;83:78–85. https://doi.org/10.1016/j.semcdb.2017.09.002.
35. Endo M, Shimizu H, Nohales MA, Araki T, Kay SA. Tissue-specific clocks in Arabidopsis show asymmetric coupling. Nature. 2014;515:419–22. https://doi.org/10.1038/nature13919.
36. Yakir E, Hassidim M, Melamed-Book N, Hillman D, Kron I, Green RM. Cell autonomous and cell-type specific circadian rhythms in Arabidopsis. Plant J. 2011;68(3):520–31.
37. Para A, Farré EM, Imaizumi T, Pruneda-Paz JL, Harmon FG, Kay SA. PRR3 is a vascular regulator of TOC1 stability in the arabidopsis circadian clock. Plant Cell. 2007;19(11):3462–73.
38. Edwards J, Martin AP, Andriunas F, Offler CE, Patrick JW, McCurdy DW. GIGANTEA is a component of a regulatory pathway determining wall ingrowth deposition in phloem parenchyma transfer cells of Arabidopsis thaliana. Plant J. 2010;63(4):651–61.
39. Endo M, Shimizu H, Araki T. Rapid and simple isolation of vascular, epidermal and mesophyll cells from plant leaf tissue. Nat Protoc. 2016;11:1388–95. https://doi.org/10.1038/nprot.2016.083.
40. Anafi RC, Lee Y, Sato TK, Venkataraman A, Ramanathan C, Kavakli IH, et al. Identify machine learning helps, CHRONO as a circadian clock component. PLOS Biol. 2014;12(4):1–18. https://doi.org/10.1371/journal.pbio.1001840.
41. Hughes ME, Hogenesch JB, Kornacker K. JTK_CYCLE: an efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. J Biol Rhythms. 2010;25(5):372–80. https://doi.org/10.1177/0748730410379711 (**PMID: 20876817.**).
42. de Lichtenberg U, Jensen LJ, Fausbøll A, Jensen TS, Bork P, Brunak S. Comparison of computational methods for the identification of cell cycle-regulated genes. Bioinformatics. 2005;21(7):1164–71. https://doi.org/10.1093/bioinformatics/bti093.
43. Straume M. DNA microarray time series analysis: automated statistical assessment of circadian rhythms in gene expression patterning. In: Numerical computer methods, Part D. vol. 383 of methods in enzymology. Academic Press; 2004. p. 149–166. Available from: http://www.sciencedirect.com/science/article/pii/S0076687904830076.
44. Michael TP, Mockler TC, Breton G, McEntee C, Byer A, Trout JD, et al. Time-of-day-specific network discovery pipeline elucidates conserved, cis-regulatory modules. PLOS Genet. 2008;4(2):1–17. https://doi.org/10.1371/journal.pgen.0040014.
45. 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 Harbor Symposia on Quant Biol. 2007;72:353–63.
46. Scargle JD. Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data. Astrophys J. 1982;263:835–53.
47. Lomb NR. Least-squares frequency analysis of unequally spaced data. Astrophys Space Sci. 1976;39(2):447–62. https://doi.org/10.1007/BF00648343.
48. Cohen-Steiner D, Edelsbrunner H, Harer J, Mileyko Y. Lipschitz functions have Lp-stable persistence. Found Comput Math. 2010;10(2):127–39. https://doi.org/10.1007/s10208-010-9060-6.
49. Hutchison AL, Maienschein-Cline M, Chiang AH, Tabei SMA, Gudjonson H, Bahroos N, et al. Sensitivity improved statistical methods enable greater, in rhythm detection for genome-wide data. PLOS Comput Biol. 2015;11(3):1–29. https://doi.org/10.1371/journal.pcbi.1004094.
50. Wu G, Zhu J, Yu J, Zhou L, Huang JZ, Zhang Z. Evaluation of five methods for genome-wide circadian gene identification. J Biol Rhythms. 2014;29(4):231–42. https://doi.org/10.1177/0748730414537788 (**PMID: 25238853.**).
51. Hughes ME, Abruzzi KC, Allada R, Anafi R, Arpat AB, Asher G, et al. Guidelines for genome-scale analysis of biological rhythms. J Biol Rhythms. 2017;32(5):380–93. https://doi.org/10.1177/0748730417728663 (**PMID: 29098954.**).
52. Saito T, Rehmsmeier M. The precision-recall plot is more informative, than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE. 2015;10(3):1–21. https://doi.org/10.1371/journal.pone.0118432.
53. Kelliher CM, Leman AR, Sierra CS, Haase SB. Investigating conservation of the cell-cycle-regulated transcriptional program in the fungal pathogen *Cryptococcus neoformans*. PLOS Genet. 2016;12(12):1–23. https://doi.org/10.1371/journal.pgen.1006453.
54. Orlando D, Lin CY, Bernard A, Iversen ES, Hartemink AJ, Haase SB. A probabilistic model for cell cycle distributions in synchrony experiments. Cell Cycle. 2007;6(4):478–88. https://doi.org/10.4161/cc.6.4.3859 (**PMID: 17329975.**).
55. Lowrey PL, Takahashi JS. Mammalian circadian biology: elucidating genome-wide levels of temporal organization. Annu Rev Genomics Hum Genet. 2004;5(1):407–41. https://doi.org/10.1146/annurev.genom.5.061903.175925 (**PMID: 15485355.**).
56. Takahashi J. Transcriptional architecture of the mammalian circadian clock. Nat Rev Genet. 2017;18:167–97.
57. Haase SB, Wittenberg C. Topology and control of the cell-cycle-regulated transcriptional circuitry. Genetics. 2014;196(1):65–90.
58. Hu H, Miao YR, Jia LH, Yu QY, Zhang Q, Guo AY. AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. Nucleic Acids Res. 2018;47(D1):D33–8. https://doi.org/10.1093/nar/gky822.
59. Nakamichi N, Kiba T, Henriques R, Mizuno T, Chua NH, Sakakibara H. Pseudo-response regulators 9, 7, and 5 are transcriptional repressors in the arabidopsis circadian clock. Plant Cell. 2010;22(3):594–605.
60. Kim H, Kim HJ, Vu QT, Jung S, McClung CR, Hong S, et al. Circadian control of ORE1 by PRR9 positively regulates leaf senescence in Arabidopsis. Proc Natl Acad Sci. 2018;115(33):8448–53.
61. Gendron JM, Pruneda-Paz JL, Doherty CJ, Gross AM, Kang SE, Kay SA. Arabidopsis circadian clock protein, TOC1, is a DNA-binding transcription factor. Proc Natl Acad Sci. 2012;109(8):3167–72.
62. A circadian gene expression atlas in mammals: Implications for biology and medicine. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE54652.

63. Investigating conservation of the cell-cycle-regulated transcriptional program in the fungal pathogen, Cryptococcus neoformans. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE80474.
64. Global control of cell cycle transcription by coupled CDK and network oscillators. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE8799.
65. Mockler Lab FTP—diurnal data. http://diurnal.mocklerlab.org/diurnal_data_finders/new.
66. Dynamic Features Analysis. Operating system: platform independent. Programming language: python. Requirements: python ≥ 3.6.0 via Anaconda, MPI, cmake. License: GNU GPL v3. https://gitlab.com/bertfordley/dynamic_features_analysis.

Publisher's Note