



SUPPLEMENTARY MATERIALS

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


Francis C. Motta¹^{*}, Robert C. Moseley², Bree Cummins³, Anastasia Deckard⁴, and Steven B. Haase²

1 Department of Mathematical Sciences, Florida Atlantic University, Boca Raton, FL, USA

2 Biology, Duke University, Durham, NC, USA

3 Department of Mathematical Sciences, Montana State University, Bozeman, MT, USA

4 Geometric Data Analytics, Durham, NC, USA

 These authors contributed equally to this work.

*Corresponding Author: fmotta@fau.edu

1 Methods

Dynamic Curve Features

De Lichtenberg Measures

In [1] the authors introduced a periodicity detection algorithm designed to identify genes in yeast that oscillated with the cell division cycle. The de Lichtenberg algorithm (DL) measures how periodic a signal is at a specified period by quantifying and combining statistical measures of gene expression periodicity and strength of regulation (Eq. 1). For each gene expression profile two empirical p -values, p_{reg} and p_{per} , are independently computed. Respectively, these p -values estimate the probabilities that the observed fold-change variability (Eq. 2) and rhythmicity at a specified period (Eq. 3) of the expression profile occurred, in some sense, at random. These statistics are then combined in a manner which accentuates expression profiles that are simultaneously highly-periodic and highly-variable. Explicitly, let $G \in \mathcal{G}$ be the gene expression profile, measured at timepoints $T = (t_1, \dots, t_n)$, corresponding to gene G in the set of all measured gene expression profiles in a collection, \mathcal{G} . Then

$$DL(G) := p_{reg}(G)p_{per}(G) \left[1 + \left(\frac{p_{reg}(G)}{0.001} \right)^2 \right] \left[1 + \left(\frac{p_{per}(G)}{0.001} \right)^2 \right]. \quad (1)$$

The p -value $p_{reg}(G)$ is meant to be the probability that a “random” curve appears more highly regulated than G . This variability metric, denoted $Reg(G)$, and defined by

$$Reg(G) = \text{std}(\log_{10}(G/\text{mean}(G))) \quad (2)$$

captures the deviation of the time series about its mean, with a small value indicating little variation over time about the mean expression level. Thus, $Reg(G)$ may be interpreted as the magnitude of regulation of the gene G . The rationale of this study is rooted in the expectation that $Reg(G)$ will actually be largest among those genes which

are primarily responsible for generating an observed program of dynamic transcript abundance.

In practice, to estimate $p_{\text{reg}}(G)$, an empirical null distribution of curve variability metrics is generated by first creating a large number of random expression curves through selection at each experimental time point the expression value of a transcript chosen from \mathcal{G} uniformly at random. The variability metric (Reg) of each random curve is then computed and $p_{\text{reg}}(G)$ is taken as the fraction of random curves whose regulation score is larger than $\text{Reg}(G)$.

Likewise, for each gene G , many simulated curves are generated by randomly permuting the expression values of G and comparing the periodicity metric (called a Fourier score in [1]) of the randomized curve to that of G . The p -value $p_{\text{per}}(G)$ is the fraction of simulated curves whose periodicity score (Per) is larger than that of G . This periodicity metric is taken to be

$$\text{Per}(G) := \sqrt{(G \cdot \cos(\omega T))^2 + (G \cdot \sin(\omega T))^2}, \quad (3)$$

where ω is a specified period, and therefore reflects the magnitude of the Fourier coefficient at that period. A new implementation of the DL algorithm was written for use in this analysis and is available as a Python 3 module under the MIT License [2].

JTK-CYCLE Measures

JTK-CYCLE (JTK) was developed as a periodicity-detection algorithm to identify circadian rhythmically expressed genes in mice [3]. Since then it has been successfully applied to time series transcriptomics data collected from many other species exhibiting rhythmic phenotypes. JTK correlates a gene’s expression profile to that of a reference curve with known periodicity properties, and computes a significance of that correlation. Usually, and in this analysis, sinusoidal template curves are generated with user-specified periods and at various phase shifts determined by the sampling times of the expression profiles. A pattern of “ups” and “downs” is computed—by comparing the expression level at each time with all subsequent times—for both the expression curve G and the equivalently sampled sinusoidal template curves. Then the total number of agreements (concordances) and disagreements (discordances) in the up-down pattern of G and that of the known periodic curves are computed, giving the Kendall rank correlation coefficient between the curves. By precomputing the exact null distribution of Kendall’s tau correlation [4] using the Harding algorithm [5], an exact Bonferroni-adjusted p -value is rapidly computed for each gene, G , and is denoted here by $p_{\text{jtk}}(G)$. An implementation of JTK in Python by Alan Hutchinson [6] was modified for use in this analysis and is available as a Python 3 module under the MIT License [7].

Precision-Recall Curves & Average Precision

Given a ranking of N genes by some metric that is meant to discriminate between core and non-core, for some choice of threshold, we examine the true positives (TP) compared to false positives (FP) and/or false negatives (FN). The *precision* of the classifier’s ranking at a given threshold is the fraction of true core genes among all genes ranked above that threshold, i.e. $\text{TP} / (\text{TP} + \text{FP})$. The *recall* is the fraction of true core genes appearing above the chosen threshold: $\text{TP} / (\text{TP} + \text{FN})$. 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 percent characterizes how many knock-out experiments one would expect to perform, in accordance with a given algorithm's ranking of genes, before 10 percent 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 threshold. Usually, this is not the case, and a threshold must be chosen arbitrarily. Thus, better measures of classifier performance incorporate algorithm performance across all thresholds. These measures 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). One such measure is the so-called *average precision* (AP) of the *precision-recall* (PR) curve defined to be $\sum_{n=1}^N P_n \Delta R_n$, where ΔR_n is the change in the recall level caused by moving the decision threshold from the $(n - 1)$ st gene to the n th gene in the list of genes ranked by some metric, and P_n is the precision of the metric when the threshold is set to the metric of the n th gene in the ranked list.

2 Tables

Table S1. Ranks among all transcription factors of core genes using the DL×JTK score.

<i>S. cerevisiae</i>			<i>M. Musculus</i>			<i>A. thaliana</i>		
Gene	MA	RNA	Gene	MA	RNA	Gene	LDHC	LLLDHC
<i>TOS4</i>	16	1	<i>DBP</i>	2	1	<i>CCA1</i>	11	3
<i>SWI5</i>	1	4	<i>NPAS2</i>	3	2	<i>LHY</i>	6	11
<i>YOX1</i>	2	5	<i>ARNTL</i>	1	4	<i>LUX</i>	30	16
<i>ASH1</i>	14	8	<i>NR1D1</i>	4	20	<i>PRR7</i>	12	17
<i>FKH1</i>	10	9	<i>NR1D2</i>	5	45	<i>RVE8</i>	15	21
<i>ACE2</i>	5	11	<i>RORC</i>	10	67	<i>PRR9</i>	49	30
<i>SWI4</i>	12	13	<i>NFIL3</i>	8	71	<i>TOC1</i>	44	34
<i>PLM2</i>	11	15	<i>TEF</i>	11	87	<i>PRR5</i>	48	45
<i>NDD1</i>	13	17	<i>BHLHE41</i>	6	123	<i>CHE</i>	97	138
<i>HCM1</i>	8	18	<i>HLF</i>	69	164	<i>RVE4</i>	63	222
<i>STB1</i>	7	21	<i>CLOCK</i>	7	419	<i>RVE6</i>	453	908
<i>YHP1</i>	15	22	<i>BHLHE40</i>	195	501			
<i>FKH2</i>	49	28	<i>ARNTL2</i>	982	671			
<i>FHL1</i>	20	40	<i>RORA</i>	283	823			
<i>SWI6</i>	176	50	<i>RORB</i>	1210	NA			
<i>MBP1</i>	151	66						
<i>MCM1</i>	266	146						
No. of TFs[†]	304	307		1373	1118		1415	1415

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

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

Table S2. The top 25 highest DL-ranked genes among all transcription factors.

Rank	<i>S. cerevisiae</i>		<i>M. Musculus</i>		<i>A. thaliana</i>	
	MA	RNA	MA	RNA	LDHC	LL.LDHC
1	ASH1*	<i>RME1</i>	ARNTL*	NPAS2*	LHY*	<i>RVE1</i>
2	<i>RME1</i>	YOX1*	DBP*	DBP*	<i>BBX19</i>	<i>PIF4</i>
3	SWI5*	ASH1*	NPAS2*	<i>CDX4</i>	<i>COL2</i>	<i>BBX18</i>
4	<i>HST4</i>	TOS4*	NR1D1*	<i>GM14444</i>	RVE8*	<i>COL5</i>
5	<i>NUT1</i>	<i>KAR4</i>	BHLHE41*	ARNTL*	TOC1*	LHY*
6	<i>HST3</i>	<i>RTT107</i>	NR1D2*	NR1D1*	<i>COL1</i>	<i>STH</i>
7	<i>CIN5</i>	SWI5*	NFIL3*	<i>CREB5</i>	<i>COL9</i>	PRR7*
8	ACE2*	SWI4*	<i>EGR1</i>	<i>PPARD</i>	<i>STH</i>	RVE8*
9	<i>MET4</i>	<i>TEC1</i>	TEF*	<i>NPAS3</i>	<i>TGA3</i>	<i>COL1</i>
10	SWI4*	<i>HST4</i>	<i>PPARD</i>	<i>POU4F1</i>	<i>RVE1</i>	<i>COL2</i>
11	YOX1*	<i>ASF1</i>	RORC*	<i>FOXO6</i>	<i>HYH</i>	<i>AT2G28200</i>
12	<i>ISW2</i>	<i>HST3</i>	<i>MAFB</i>	<i>DMRTA2</i>	<i>BBX18</i>	<i>PIL6</i>
13	<i>RGT2</i>	<i>RLF2</i>	<i>KLF10</i>	<i>EGR1</i>	<i>AT1G26790</i>	<i>MYBL2</i>
14	<i>MET18</i>	<i>WTM1</i>	<i>RFXANK</i>	<i>GM20422</i>	PRR7*	CCA1*
15	<i>RGT1</i>	ACE2*	CLOCK*	<i>EGR3</i>	<i>HB-12</i>	<i>BBX8</i>
16	<i>SIP4</i>	<i>ZNF1</i>	<i>TSC22D3</i>	<i>TBX1</i>	<i>PRE1</i>	<i>HSFA8</i>
17	TOS4*	STB1*	<i>KLF13</i>	<i>MAFF</i>	<i>STO</i>	<i>CDF3</i>
18	<i>RTT107</i>	<i>WTM2</i>	<i>NR0B2</i>	<i>GM6710</i>	<i>PIL6</i>	<i>ABF1</i>
19	<i>ESC2</i>	<i>CSE2</i>	<i>ZFP3611</i>	<i>EN2</i>	CCA1*	TOC1*
20	<i>ASF1</i>	<i>GAT1</i>	<i>ATF5</i>	<i>ZBTB7C</i>	<i>EPR1</i>	<i>EPR1</i>
21	<i>WTM2</i>	<i>OTU1</i>	<i>SREBF1</i>	<i>MESP1</i>	<i>LZF1</i>	<i>AT1G70000</i>
22	<i>HAP5</i>	HCM1*	<i>STAT5B</i>	NR1D2*	<i>BZS1</i>	<i>STO</i>
23	<i>HPA2</i>	<i>PHD1</i>	<i>GTF2IRD1</i>	<i>ZFP987</i>	<i>ASG4</i>	<i>BBX16</i>
24	<i>EDS1</i>	<i>MAL13</i>	<i>ZBTB21</i>	<i>GM14401</i>	<i>CDF3</i>	<i>ATCTH</i>
25	<i>RTG2</i>	<i>RGT2</i>	<i>SOX9</i>	<i>GM14305</i>	<i>CO</i>	LUX*
Recall	35.3%	47.1%	66.7%	35.7%	45.5%	54.5%

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

* Core transcription factors in Dataset S3.

Table S3. The top 25 highest PerReg-ranked genes among all transcription factors.

Rank	<i>S. cerevisiae</i>		<i>M. Musculus</i>		<i>A. thaliana</i>	
	MA	RNA	MA	RNA	LDHC	LL.LDHC
1	<i>RME1</i>	<i>YOX1*</i>	<i>NR1D1*</i>	<i>CDX</i>	<i>RVE1</i>	<i>MYBL2</i>
2	<i>ASH1*</i>	<i>TOS4*</i>	<i>NPAS2*</i>	<i>NPAS2*</i>	<i>LHY*</i>	<i>CCA1*</i>
3	<i>SWI5*</i>	<i>RME1</i>	<i>DBP*</i>	<i>GM14444</i>	<i>CCA1*</i>	<i>LHY*</i>
4	<i>HST3</i>	<i>ASH1*</i>	<i>ARNTL*</i>	<i>FOXO6</i>	<i>COL2</i>	<i>COL2</i>
5	<i>ACE2*</i>	<i>SWI5*</i>	<i>EGR1</i>	<i>NR1D1*</i>	<i>STH</i>	<i>PRR7*</i>
6	<i>SWI4*</i>	<i>KAR4</i>	<i>NR1D2*</i>	<i>POU4F1</i>	<i>MYBL2</i>	<i>PRR5*</i>
7	<i>YOX1*</i>	<i>HST3</i>	<i>MYC</i>	<i>GM20422</i>	<i>AT1G26790</i>	<i>PIF4</i>
8	<i>HST4</i>	<i>ZNF1</i>	<i>NFIL3*</i>	<i>DMRTA2</i>	<i>PRR7*</i>	<i>STH</i>
9	<i>SIP4</i>	<i>RTT107</i>	<i>BHLHE41*</i>	<i>NPAS3</i>	<i>HYH</i>	<i>COL5</i>
10	<i>ASF1</i>	<i>MGA1</i>	<i>TEF*</i>	<i>EGR1</i>	<i>PRR5*</i>	<i>RVE1</i>
11	<i>TOS4*</i>	<i>SWI4*</i>	<i>RORC*</i>	<i>DBP*</i>	<i>BBX16</i>	<i>CDF3</i>
12	<i>RTT107</i>	<i>SIP4</i>	<i>ZBTB16</i>	<i>CREB5</i>	<i>LZF1</i>	<i>LZF1</i>
13	<i>STB1*</i>	<i>ACE2*</i>	<i>PPARD</i>	<i>TBX1</i>	<i>BBX18</i>	<i>LUX*</i>
14	<i>EDS1</i>	<i>PHD1</i>	<i>MAFB</i>	<i>EGR3</i>	<i>CDF2</i>	<i>CDF1</i>
15	<i>GAT1</i>	<i>MAL33</i>	<i>KLF10</i>	<i>MESP1</i>	<i>EPR1</i>	<i>COL1</i>
16	<i>HAP4</i>	<i>HST4</i>	<i>NR0B2</i>	<i>GM6710</i>	<i>CDF1</i>	<i>ATCTH</i>
17	<i>HPA2</i>	<i>TEC1</i>	<i>RFXANK</i>	<i>ARNTL*</i>	<i>COL1</i>	<i>CDF2</i>
18	<i>GAL3</i>	<i>RLF2</i>	<i>BCI6</i>	<i>EN2</i>	<i>BBX19</i>	<i>BBX18</i>
19	<i>WTM1</i>	<i>ASF1</i>	<i>ZFP931</i>	<i>ZBTB7C</i>	<i>LUX*</i>	<i>ABF1</i>
20	<i>RLF2</i>	<i>FKH1*</i>	<i>NR3C2</i>	<i>GM14401</i>	<i>PIF4</i>	<i>COL9</i>
21	<i>NDT80</i>	<i>MSS11</i>	<i>CLOCK*</i>	<i>GM14305</i>	<i>PRR9*</i>	<i>EPR1</i>
22	<i>XBP1</i>	<i>MTH1</i>	<i>ONECUT1</i>	<i>BHLHE22</i>	<i>CDF3</i>	<i>AT1G26790</i>
23	<i>TEC1</i>	<i>RPI1</i>	<i>KLF9</i>	<i>PPARD</i>	<i>RVE2</i>	<i>BBX16</i>
24	<i>MSN4</i>	<i>STB1*</i>	<i>ZBTB7C</i>	<i>OVO12</i>	<i>RVE8*</i>	<i>AT5G44260</i>
25	<i>HCM1*</i>	<i>HCM1*</i>	<i>FOXQ1</i>	<i>FOXP3</i>	<i>TGA3</i>	<i>HYH</i>
Recall	47.1%	52.9%	66.7%	28.6%	63.6%	45.5%

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

* Core transcription factors in Dataset S3.

3 Figures

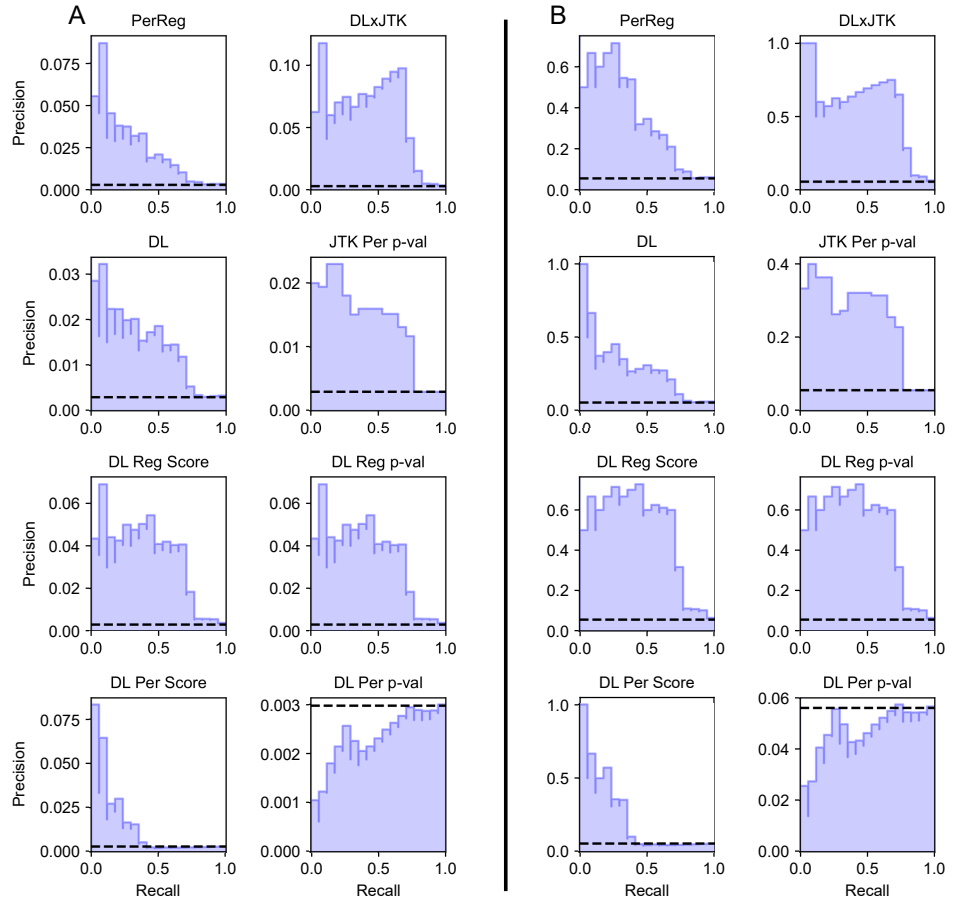


Fig S1. Precision-recall curves of classifiers identifying core from non-core TFs among all genes (A) and among only TFs (B) in *S. cerevisiae* microarray dataset. Between changes in recall, $R_{n-1} < R_n$, precision is plotted as a constant equal to the precision P_n at the minimum decision-threshold rank giving a recall of R_n , so that the area under the curve is equal to average precision. The horizontal dashed line indicates the baseline average precision of a random classifier.

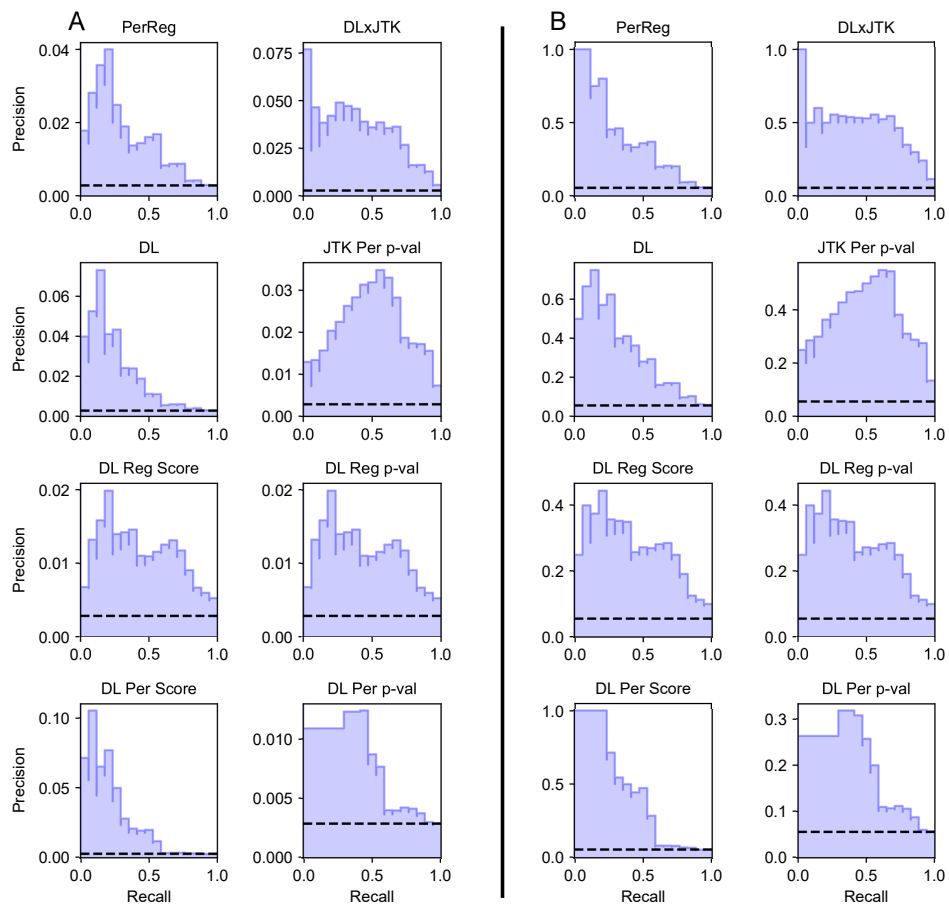


Fig S2. Precision-recall curves of classifiers identifying core from non-core TFs among all genes (A) and among only TFs (B) in *S. cerevisiae* RNAseq dataset. Between changes in recall, $R_{n-1} < R_n$, precision is plotted as a constant equal to the precision P_n at the minimum decision-threshold rank giving a recall of R_n , so that the area under the curve is equal to average precision. The horizontal dashed line indicates the baseline average precision of a random classifier.

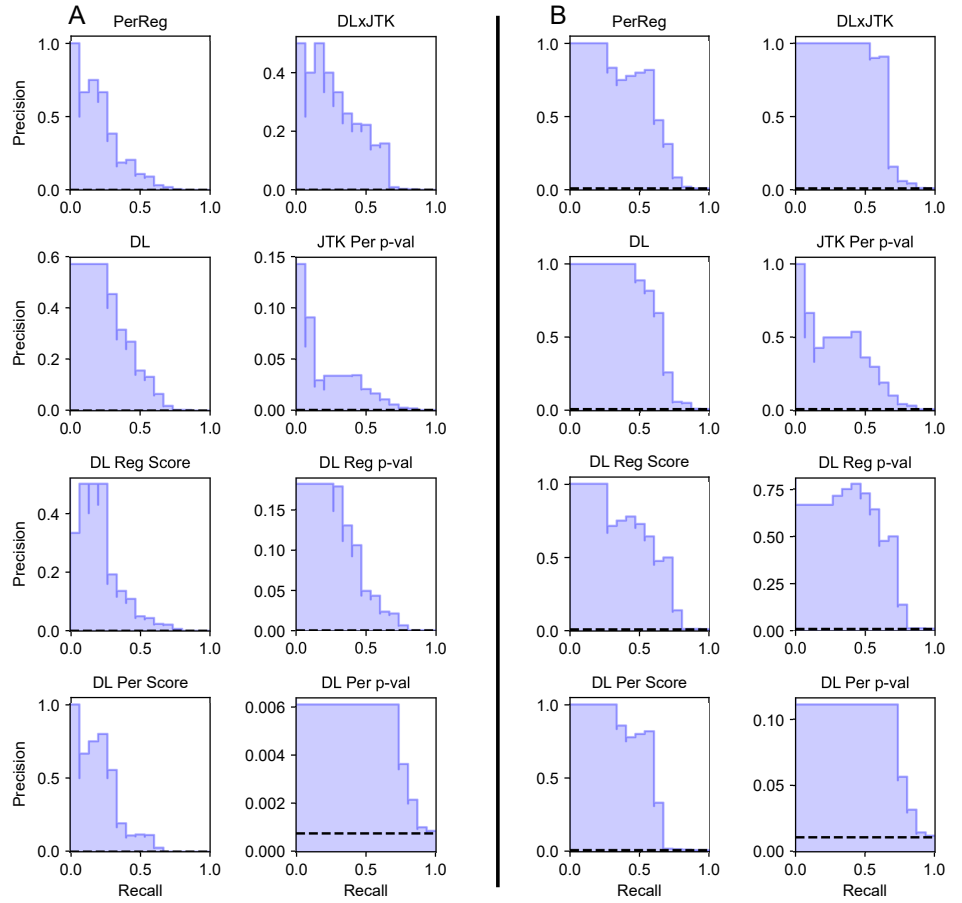


Fig S3. Precision-recall curves of classifiers identifying core from non-core TFs among all genes (A) and among only TFs (B) in *M. Musculus* microarray dataset. Between changes in recall, $R_{n-1} < R_n$, precision is plotted as a constant equal to the precision P_n at the minimum decision-threshold rank giving a recall of R_n , so that the area under the curve is equal to average precision. The horizontal dashed line indicates the baseline average precision of a random classifier.

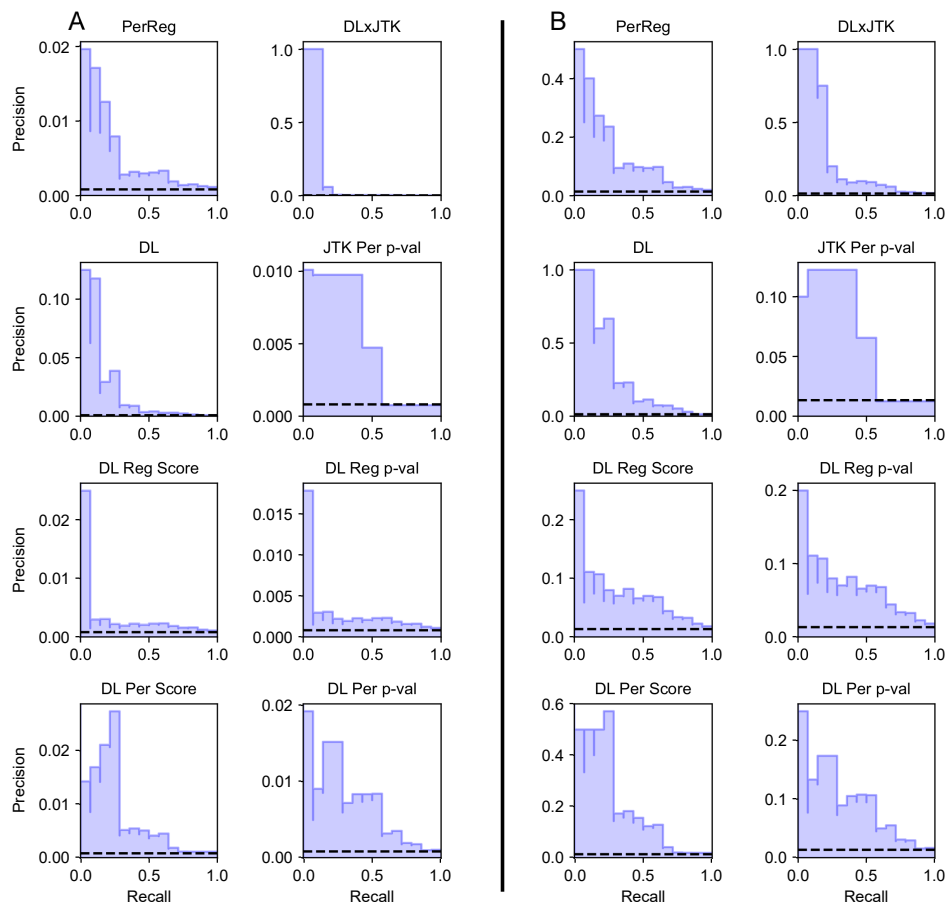


Fig S4. Precision-recall curves of classifiers identifying core from non-core TFs among all genes (A) and among only TFs (B) in *M. Musculus* RNaseq dataset. Between changes in recall, $R_{n-1} < R_n$, precision is plotted as a constant equal to the precision P_n at the minimum decision-threshold rank giving a recall of R_n , so that the area under the curve is equal to average precision. The horizontal dashed line indicates the baseline average precision of a random classifier.

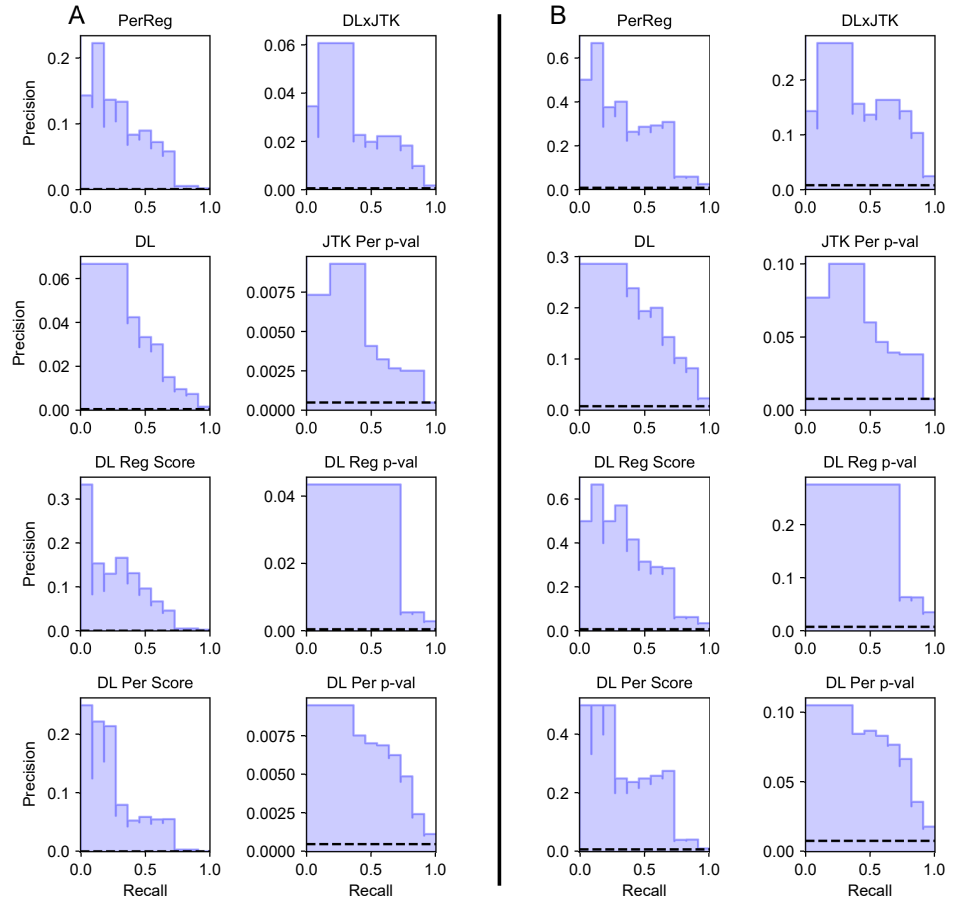


Fig S5. Precision-recall curves of classifiers identifying core from non-core TFs among all genes (A) and among only TFs (B) in *A. thaliana* microarray LDHC dataset. Between changes in recall, $R_{n-1} < R_n$, precision is plotted as a constant equal to the precision P_n at the minimum decision-threshold rank giving a recall of R_n , so that the area under the curve is equal to average precision. The horizontal dashed line indicates the baseline average precision of a random classifier.

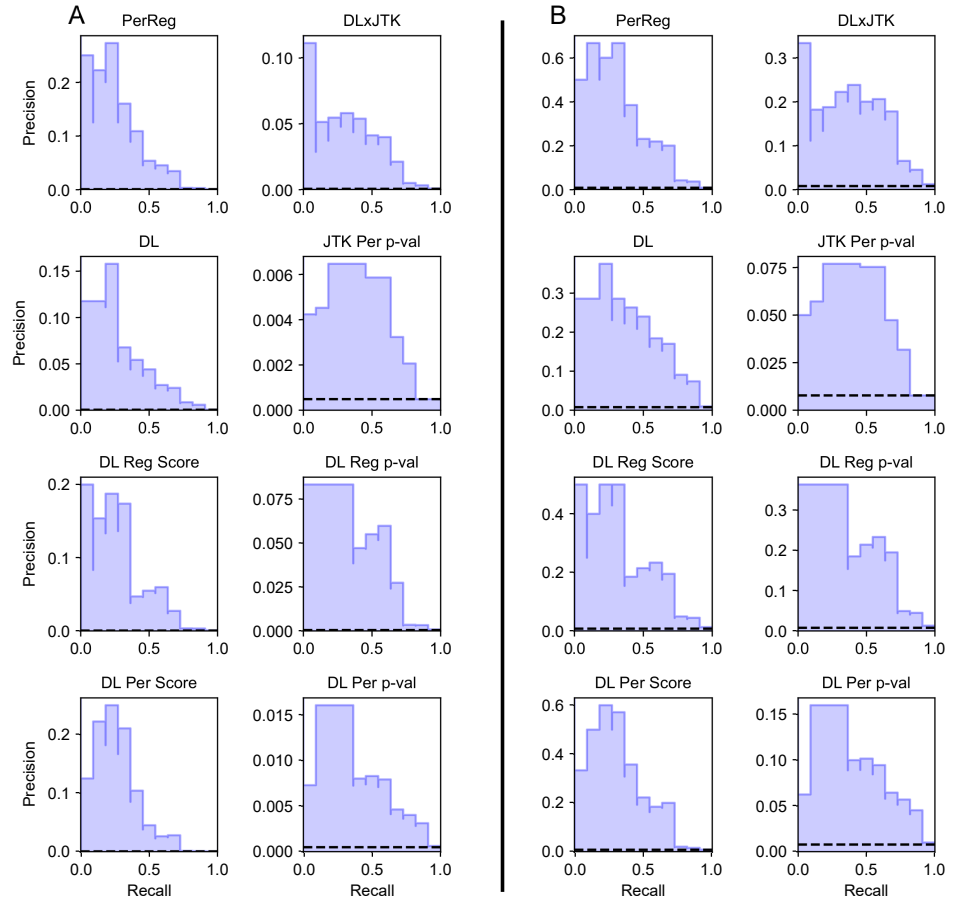


Fig S6. Precision-recall curves of classifiers identifying core from non-core TFs among all genes (A) and among only TFs (B) in *A. thaliana* microarray LL_LDHC dataset. Between changes in recall, $R_{n-1} < R_n$, precision is plotted as a constant equal to the precision P_n at the minimum decision-threshold rank giving a recall of R_n , so that the area under the curve is equal to average precision. The horizontal dashed line indicates the baseline average precision of a random classifier.

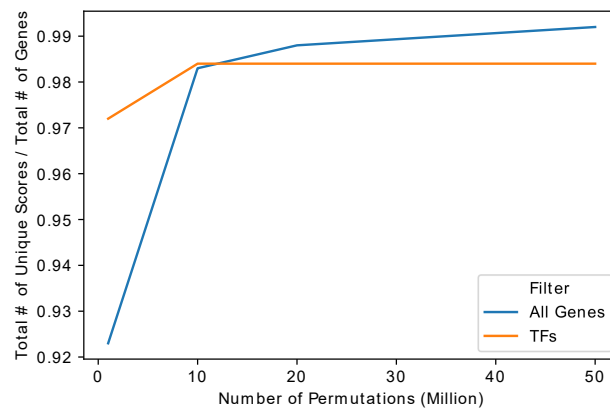


Fig S7. The number of permutations for computing DL Reg p -val vs. the number of unique DL Reg p -values. The number of unique values of DL Reg p -val per total number of genes before filtering for transcription factors (blue line) and after filtering for transcription factors (orange line) for the *A. thaliana* LDHC dataset as a function of the number of permutations chosen for the DL algorithm.

References

1. 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):1164–1171. doi:10.1093/bioinformatics/bti093.
2. Motta F, Deckard A, Moseley RC. pyDL; 2019. <https://gitlab.com/biochron/pydl>.
3. Hughes ME, Hogenesch JB, Kornacker K. JTK_CYCLE: An Efficient Nonparametric Algorithm for Detecting Rhythmic Components in Genome-Scale Data Sets. *Journal of Biological Rhythms*. 2010;25(5):372–380. doi:10.1177/0748730410379711.
4. Kendall MG. A NEW MEASURE OF RANK CORRELATION. *Biometrika*. 1938;30(1-2):81–93. doi:10.1093/biomet/30.1-2.81.
5. Harding EF. An Efficient, Minimal-Storage Procedure for Calculating the Mann-Whitney U, Generalized U and Similar Distributions. *Journal of the Royal Statistical Society Series C (Applied Statistics)*. 1984;33(1):1–6.
6. Hutchison A. pyJTK: Python implementation of the JTK_CYCLE statistical test; 2013. <https://github.com/alanlhutchison/pyJTK>.
7. Hutchinson A, Motta F, Deckard A. pyJTK; 2019. <https://gitlab.com/biochron/pyjtk>.